Proceedings of the 17th International Conference on Computational and Mathematical Methods in Science and Engineering

Costa Ballena, Rota, Cádiz (Spain)

July 4th-8th, 2017

CMMSE 2017

VOLUMES I-VI

Editor: J. Vigo-Aguiar

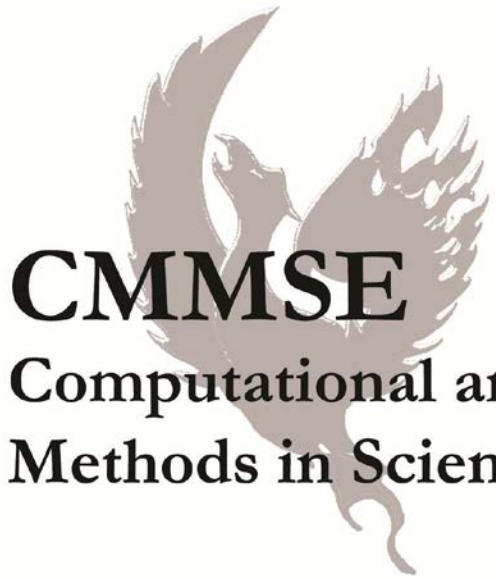
Associate Editors

J. Medina, M.E. Cornejo, W. Sprößig, T. Sheng,
P. Gill, E. Venturino, I. P. Hamilton, J.A. Alvarez-Bermejo, H Ramos

Proceedings of the 17th International Conference on Computational and Mathematical Methods in Science and Engineering CMMSE-2017

Costa Ballena (Rota), Cádiz, Spain

July 4-8, 2017



CMMSE

Computational and Mathematical Methods in Science and Engineering

Editors

J. Vigo-Aguiar

Associate Editors

J. Medina, M. E. Cornejo, W. Sprößig, T. Sheng,
P. Gill, E. Venturino, I. P. Hamilton, J.A. Álvarez-Bermejo, H. Ramos

ISSN: 2312-0177
ISSN-L:2312-0177

ISBN: 978-84-617-8694-7

Printed on acid-free paper

Cover: A 1754 painting by H.J. Detouche shows Galileo Galilei displaying his telescope to Leonardo Donato and the Venetian Senate.

Preface

The European Commission (EC) has highlight in different occasions the importance of mathematics in the resolution of the current societal challenges. Concerning the current EU Framework Programme for Research and Innovation, Horizon2020, in 2016 the EC launched an online consultation on mathematics in order to nourish the next Horizon2020 Work Programme (2018-20) with innovative mathematical content.

As a consequence of this importance, the diverse calls in the future Horizon2020 Work Programme demand the inclusion of mathematical partners. This fact is not mandatory, but the consortia with this component are more valuable. For example, the report of the online consultation includes the following paragraph:

"*Whereas mathematics is not a prerequisite in Horizon2020, the various areas covered by the programme rely on its development and its use; HPC, Big data, Quantum computing just to name a few. Without mathematical tools, future research will be severely hampered."*

This trend is being reflected in national and regional calls. Therefore, the Applied Mathematics is in a really key position. The traditional interaction with other important areas, such as Engineering, Physics, Chemistry, Computer Science, etc., has been fundamental for the consideration of the mathematics nowadays. For example, the use of mathematical tools in the well-known and actual Big Data problem, which has all kind of companies - small, medium and large - already plays and will continue playing a fundamental role in the resolution of this big problem of the XXI century. This is determinant in (ciber)security, recommendation in social networks, in productive processes, etc.

Thus, in this great framework, it is a pleasure to welcome you to the **17th International Conference on Computational and Mathematical Methods in Science and Engineering** (CMMSE 2017), which will be held at Rota, Cádiz (Spain), July 4^{th}-8^{th}, 2017. After seventeen editions, this is a consolidate international conference, which provide a comfortable environment in which researchers from different disciplines interact and create synergies focus on the achievement of new challenges and based on really breakthrough ideas. The current EU framework brings important funding opportunities for European partners, but also for important international research groups outside of Europe, in which the ideas developed during this conference can be funded.

The proceedings of CMMSE 2017, which contains the extended abstracts and papers accepted to the conference, is a perfect seed for this collaboration. In this edition, it has six volumes, the first five correspond to the articles typeset in LaTeX and the sixth to articles typeset in Word. The invited speeches given by renowned researchers will also be a starting point for the establishment of new challenges.

CMMSE 2017 covers a wide range of disciplines which have been divided into forty-five different symposia. High quality papers have been collected in each of these symposia. The first symposium entitled *High-Performance Computing (HCP)* focuses on computational problems appearing in many scientific and engineering applications. The second one, *mathematically modeling the future Internet and developing future Internet security technology*, is a self-explanatory session. The symposium *Computational finance*

deals with recent advances on modeling and computation in quantitative finance. *New educational methodologies supported by new technologies* is a session to discuss the use of new technologies and the possibility of developing new resources to improve the teaching and learning activities. In what regards *Mathematical models and information-intelligent systems on transport* session, it is dedicated to flow-modeling of particles framework whose applications are closely related to traffic flows, pedestrian flows, ecology, etc. *Computational Methods for Linear and Nonlinear Optimization* addresses some recent techniques and efficient methods to solve different type of optimization problems. New trends in the field of iterative methods for nonlinear problems are presented in *Numerical Methods for Solving Nonlinear Problems.* The 8th symposium entitled *Bio-mathematics* presents theoretical and practical applications associated with population dynamics, eco-epidemiology, epidemiology of infectious diseases and molecular and antigenic evolution. Modeling interesting problems arisen in Computer Science, considering algebraic and computational (fuzzy) techniques, is the main goal of *Mathematical Models for Computer Science*. The aim of the symposium *Mathematics meets Chemistry-Theorical Models at the Nanoscale* is to obtain a consistent description of the transition from clusters to the solid state, which is a major challenge in computational chemistry and physics.

The session *Hypercomplex methods in mathematics and Applied Sciences* includes new methods for modeling and solving of boundary value problems in fluid mechanics, elasticity theory and other related fields. Fixed-point theory has an enormous applicability in mathematics, engineering, chemistry, biology, economics, computer science, which justifies the great interest in *fixed point theory in various abstract spaces and related applications*. The 13th symposium *Industrial Mathematics* presents mathematical and computational researches related to corporate or government applications and problems arising from different economic sectors. C*omputational methods for fluid flow* are considered in this conference in order to discuss recent problems and advances in this area. A wide variety of topics related to pattern formation in animate and inanimate, patterns stationary, travelling or disordered in both space and time, is included in *Pattern formation in spatially extended systems.* Different research fields of fractal and fractional dynamics in theoretical and practical studies of Mathematics, Physics, and Engineering will be addressed in *Fractional calculus and applications.* The aim of *New Trends on Boundary Value Problems* is to present and discuss new trends in related fields such as variational methods and topological methods. *Estimation and control for stochastic systems: theory and applications* is a forum to discuss the latest approaches on this research topic from both theoretical and application perspectives. With respect to *Numerical Methods in Mathematics and Mechanics*, it is a session focused on the study of numerical methods and software capabilities related to multi-body system dynamics and equations of motion, structural and nonlinear control and modern vibrational methods. Machine Learning techniques are increasingly being used to tackle problems in bioinformatics and computational biology, hence the importance of the session *Mechanical Learning Techniques in Bioinformatics.*

The 21th symposium will put particular emphasis on algorithmic issues in the treatment of curves and surfaces from the symbolic and also from the numeric point of view: *Numerical Problems on Algebraic Curves and Surfaces.* The session *New advances in statistical methodologies* is intended to show theoretical, applied or computational techniques, which separately or in combination, have provide interesting achievements to a broad spectrum of statistical methods. Taking into account that *Non Newtonian Calculus, Theory and Applications* is a very novel research topic, the goal of this session is to gather the scientific community working on this framework in order to discuss and exchange

ideas. Different techniques to determine effective properties in elasticity and conductivity of heterogeneous materials such as homogenization and micromechanic are given in the symposium: *Homogenization of Partial Differential Equations. Micromechanics. Elasticity and conductivity of composite materials.* The stability, growth and prosperity of an economy as well as the wellbeing of society is affected by the increasing frequency and severity of natural and human-caused disasters. This fact justifies the great significance of the symposium entitled *Mathematical modeling of Man-made Natural disasters: forest fire & environmental pollution.* The session, *Recent trends in the analysis and computations of nonlinear partial differential equations and systems,* will pay special attention the investigation of analytical features of the solutions of nonlinear problems and the analysis of approximation techniques to simulate them. *Control and estimation for Cyberphysical and Distributed parameter systems* aims at processing problems corresponding to this topic in a unified manner and to propose generic and widely applicable solutions for robotics, transportation and socioeconomic systems, etc.

Current theoretical and computational advances concerning the numerical treatment of rank structured matrices and their applications will be introduced in *Rank structured matrices: recent developments and new perspectives.* Novel trends in the field of computational intelligence methods such as genetic algorithms, evolutionary strategies and ant colony optimization, among others, will be analysed in the session *Computational intelligence methods for solving complex optimization problems.* An interesting discussion about the latest findings in nonlinear wave theory will be carried out during the symposium *Dynamics and stability of nonlinear wave patterns.* New research developments to push flow control understanding and applications, as well as the study different devices capable of interacting with the fluid boundary layer, will be presented in the conference session entitled *Flow Control, Active/Passive*. The main purpose of the symposium *Numerical Linear Algebra Methods for Large Scale Scientific Computing* is to collect methods for solving linear algebra problems derived from the discretization of partial differential equations and optimization problems. *Mathematics, Science and Engineering Education* analyzes the use of new ICT tools in teaching methods corresponding to different educational levels. *Recent trends in the development of meshless or meshfree methods and applications* is a session which offers the opportunity to discuss current progresses in this area. The 35th symposium entitled *Contemporary Approaches in Multivariate Representations and Approximation Methods for Discrete and Continuous Mathematical Objects* is designed to focus on mostly multivariate problems encountered in science and engineering.

Lie Symmetry Analysis and Conservation Laws for Nonlinear Differential Equations and Applications covers the modelization of real world problems by means of nonlinear differential equations. *Statistical Modeling and Applications* are considered in this conference in order to show the new theoretical steps on this research line and on its applications. All topics related to orthogonal polynomials and their use in solving real-life problems will be covered in *Orthogonal Polynomials and Applications*. *Complex Networks* covers general phenomena which can be modelled as networks where a number of individuals interact over a complex networked structure. *Data Analysis and Modeling Science and Engineering: Numerical Methods and Computational Approaches* aims to discussing recent trends in data analysis, forecast and modeling focusing on the numerical aspects and computational approaches for the analysis and classification of large data sets. *Multiscale modeling of solid/fluid systems with focus on structures, thermodynamic properties and industrial applications* presents new results and developments within this fascinating area. The session *Theoretical and computational of the free boundary problems* considers results obtained from the both theoretical and computational view of a free

boundary problems which include the stationary and evolutionary variational, quasi-variational inequalities and Hamilton-Jacobi-Bellman. The 43rd symposium *Processing, modeling, and describing time series* offers an international forum for the presentation of original results in this field. *Data mining and engineering* aims at joining the contemporary innovations about data mining and the related area of data engineering. Finally, the last symposium, *Metaheuristics in science and engineering* deals with contemporary approaches in the context of metaheuristics focusing on applications in real problems or simulated scenarios.

We would like to finish this preface expressing our deepest gratitude to the plenary speakers for their outstanding contributions to research and leadership in their respective fields. We are also most grateful to the special session organizers and scientific committee members. Working with an efficient and capable team has been fundamental in the execution of CMMSE 2017. Finally, it is important to mention that this conference would not have been possible without the interest and enthusiasm of all participants. Therefore, we can only conclude this preface giving them thanks and a cordial welcome. We hope you enjoy the conference.

Costa Ballena, Rota, Cádiz (Spain), July 3rd, 2017

I. P. Hamilton, J. Medina & J. Vigo-Aguiar

CMMSE 2017 Mini-symposia

Session Title	Type of Session/Organizers
High Performance Computing (HPC)	
Computational Finance	Luis Ortiz-García & Iñigo Arregui
Mathematical Models and Information-Intelligent Systems on Transport	Valery V. Kozlov & Alexander P. Buslaev
Numerical Methods for Solving Nonlinear Problems	Juan R. Torregrosa & A. Cordero
Bio-mathematics	Ezio Venturino & Maíra Aguiar
Mathematical Models for Computer Science	Jesús Medina & Manuel Ojeda-Aciego
Mathematics meets Chemistry-Theorical Models at the Nanoscale	Ian Hamilton & Peter Schwerdtfeger & Ottorino Ori & Jerzy Cioslowski
Hypercomplex methods in mathematical and Applied Sciences	Klaus Gürlebe & Helmuth Malonek
Fixed Point Theory in various abstract spaces and related applications	Antonio Francisco Roldán López de Hierro & Juan Martínez Moreno & Erdal Karapinar
Computational methods for fluid flow	Zhenquan Li & Tiejin Wang
New Trends on Boundary Value Problems	Feliz Minhós & João Fialho
Estimation and control for stochastic systems: theory and applications	Raquel Caballero Águila & Josefa Linares Pérez
Numerical Methods in Mathematics and Mechanics	Mihai Dupac
New advances in statistical methodologies	Filipe J. Marques & Carlos A. Coelho
Non Newtonian Calculus. Theory and Applications	Fernando Córdova-Lepe
Processing, modeling and describing time series	Juan Moreno-Garcia & Luis Rodriguez-Benitez

CMMSE 2017 Mini-symposia

Session Title	Type of Session/Organizers
Recent trends in the analysis and computations of nonlinear partial differential equations and systems	Matthias Ehrhardt & J. E. Macías-Díaz & Tim Sheng
Rank structured matrices: recent developments and new perspectives	Aceto Lidia & Boito Paola & Gemignani Luca
Flow Control, Active/Passive	Ivette Rodriguez & Fernando Mellibovsky & Manel Soria & Josep M Bergadà,
Numerical Linear Algebra Methods for Large Scale Scientific Computing	Luca Bergamaschi & Angeles Martinez
Mathematics, Science and Engineering Education	Ángel Alberto Magreñán & Nuria Arís Redó

Recent trends in the development of meshless or meshfree methods and applications	Luis Gavete & Juan José Benito & Francisco Ureña
Contemporary Approaches in Multivariate Representations and Approximation Methods for Discrete and Continuous Mathematical Objects	Metin Demiralp
Lie Symmetry Analysis and Conservation Laws for Nonlinear Differential Equations and Applications	C. M. Khalique & M. L. Gandarias & M. S. Bruzón
Statistical Modeling and Applications	M. Virtudes Alba-Fernández & M. Dolores Jiménez-Gamero.

Acknowledgements

We would like to express our gratitude to the University of Cádiz, especially to the group of professor Jesús Medina of the department of Mathematics, professor José Antonio Álvarez Bermejo - University of Almeria and the group of HPC (Prof. Jose Ranilla & R. Cortina) and Linear Algebra (Prof. Pedro Alonso) of University of Oviedo.

We also would like to thank all of the local organizers for their efforts devoted to the success of this conference:

- Pedro Alonso, Universidad Oviedo
- José Ranilla, Universidad Oviedo
- Raquel Cortina, Universidad Oviedo
- Carmelo Clavero, Universidad Zaragoza
- Higinio Ramos, Universidad Salamanca
- José Antonio Álvarez Bermejo - Universidad de Almería,
- María Eugenia Cornejo Piñero - Universidad de Cádiz,
- Juan Carlos Díaz Moreno - Universidad de Cádiz,
- Eloísa Ramírez Poussa - Universidad de Cádiz,
- Juan Martinez Moreno Universidad de Jaén,
- Antonio Francisco Roldan, Universidad de Jaén,

CMMSE 2017 Plenary Speakers

- Wolfgang Sprößig - TU Bergakademie Freiberg, **Germany**
- Tim Sheng - Baylor University, **USA**
- Klaus Guerlebeck - University of Weimar, **Germany**
- Peter Gill - Australian National University, **Australia**
- Ezio Venturino - University of Turin, **Italy**
- Metin Demiralp, University of Istanbul, **Turkey**

Contents:

Volume I

Contents:

Volume II

Contents:

Volume III

Contents:

Volume IV

Contents:

Volume V

Contents:

Volume VI

Contents:

Addendum

Volume I

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

Linear-time solvers for linear systems with sparse and structured matrices of interest in applications.

J. Abderramán Marrero[1]

[1] *Department of Mathematics Applied to Information Technology, School of Telecommunication Engineering, UPM - Technical University of Madrid (Spain)*

emails: jc.abderraman@upm.es

Abstract

Some recent results on linear-time solvers, e.g. based on incomplete or full Givens reduction, for determined linear systems with matrix coefficient having a particular sparse or low-rank structure are pointed out. These results have motivated its possible application to matrices of related interest.

Key words: Linear-time solvers, large-order linear systems, Givens reduction, low-rank structured matrices.

1 Extended abstract

Fast algorithms have been proposed recently for both linear systems and inversion methods, where particular kinds of large-order matrices arising in applications are involved. More precisely, for general k-tridiagonal [1, 2], bordered tridiagonal [3], bordered k-tridiagonal [4], opposite-bordered tridiagonal [5], and comrade matrices [6], where different approaches were applied.

A sequential numerical algorithm based on tridiagonal solvers to manage comrade linear systems was introduced in [6]. In [1], a direct algorithm was design for inverting k-tridiagonal matrices based on symbolic computation. This symbolic approach is based on the following known fact. Let $A(\lambda)$ be the matrix obtained from A by changing the entry $a_{i,j}$ by $a_{i,j}(\lambda)$, such that, when $\lambda \to 0$, $a_{i,j}(\lambda) \to a_{i,j}$. Thus $\lim_{\lambda\to 0} A(\lambda) = A$, component-wise. Therefore, if A is nonsingular, there exists δ, with $0 < |\lambda| < \delta$, satisfying $lim_{\lambda\to 0}A^{-1}(\lambda) = A^{-1}$, component-wise. Also, such symbolic strategy has been applied for inverting other related matrices and to solve linear systems, e.g. the opposite-bordered tridiagonal ones [5].

 ISBN: 978-84-617-8694-7

The solution of the determined linear system, formulated by using the vector linear equation

$$Ax = b; \quad \text{with } A \in \mathcal{GL}_n(\mathbb{R}), \text{ and } \quad x, b \in \mathbb{R}^n, \tag{1}$$

with $\mathcal{GL}_n(\mathbb{R})$ the general linear group of the $(n \times n)$ real matrices and the matrix A nonsingular bordered tridiagonal, shortly GNBT, was computed in [3] using a sequential numerical solver. It was based on an incomplete or full Givens reduction of the GNBT matrix A, depending on whether the tridiagonal submatrix associated with A is nonsingular or singular, respectively. Therefore, the GNBT linear system (1) was transformed in an equivalent, but simpler, nonsingular bordered triangular system and the vector solution x was computed using the useful Sherman-Morrison formula or an adapted back substitution. Using a similar procedure, this numerical procedure was also extended on the bordered k-tridiagonal, shortly GNBkT, matrices in [4].

As a result, when the tridiagonal (k-tridiagonal) matrix, associated with the GNBT (GNBkT) matrix A, is singular, then a linear-time solver for the corresponding linear system was obtained independently of the LU factorizations of such matrices. Nevertheless, as was advised in [3], when such associated matrix is singular, the Givens reduction of A can give a dense triangular matrix, or having rows with full nonzero entries. It can generate numerical instability. Infrequently, the Sherman-Morrison formula can generate also instabilities. To avoid these drawbacks, specialized subroutines should be used, e.g. the basic linear algebra subroutines, BLAS.

The (unique) QR factorization of a nonsingular k-tridiagonal matrix [1, 2], a particular type of GNBkT matrix, is sparse, independently of the existence of its LU factorization. Therefore, the determined linear systems (1) with such matrix coefficient can be solved in linear time using Givens reduction of the k-tridiagonal matrix coefficient and an adapted back substitution [2].

Since the speedup and reliability of the numerical solvers from [2], [3], and [4], when the correspondent matrices, or associated submatrices, have a sparse and unique QR factorization, we ask about their applications on vector, or matrix, equations involving opposite-bordered tridiagonal [5], comrade [6], and other sparse and low-rank structured matrix coefficients arising in applications. Some preliminary results on opposite-bordered tridiagonal and also on comrade linear systems have given support for a positive answer.

The search of necessary as well as sufficient conditions to obtain linear-time solvers for these classes of linear systems, related with the structure of the matrix coefficients, or that of some of its associated submatrices, and its correspondent QR (or LU) factorization, can be a challenging line.

This numerical approach should be extended to find fast algorithms for linear systems with dense but low-rank structured matrices, e.g. on particular subclasses of generalized Hessenberg matrices [7].

 ISBN: 978-84-617-8694-7

References

[1] J. Jia, S. Li, *Symbolic algorithms for the inverses of general k-tridiagonal matrices*, Comput. Math. Appl. **70** (2015) 3032–3042.

[2] J. Abderramán Marrero, *Fast algorithms for solving general k-tridiagonal matrix linear equations*, Proceedings of the 16-th International Conference on Computational and Mathematical Methods in Science and Engineering, CMMSE 2016 Vol **1** (2016) 25–28.

[3] J. Abderramán Marrero, *A numerical solver for general bordered tridiagonal matrix equations*, Comput. Math. Appl. **72** (2016) 2731–2740.

[4] J. Abderramán Marrero, V. Tomeo, *A fast and reliable numerical solver for general bordered k-tridiagonal matrix linear equations*, J. Comput. Appl. Math. **318** (2017) 211–219.

[5] J. Jia, T. Sogabe, S. Li, *A generalized symbolic Thomas algorithm for the solution of opposite-bordered tridiagonal linear systems*, J. Comput. Appl. Math. **290** (2015) 423–432.

[6] T. Sogabe, *Numerical algorithms for solving comrade linear systems based on tridiagonal solvers*, Appl. Math. Comput. **197** (2008) 117–122.

[7] J. Abderramán Marrero, M. Rachidi, *Inverses of generalized Hessenberg matrices*, Linear Multilinear Algebra, **63** (2015) 559–570.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

OpenCL Code Generation for Mobile Devices

Sergio Afonso[1], Alejandro Acosta[1] and Francisco Almeida[1]

[1] *Department of Computer Engineering, Universidad de La Laguna, Spain*

emails: safonsof@ull.es, aacostad@ull.es, falmeida@ull.es

Abstract

The efficient exploitation of the computational capabilities of mobile devices is still a challenge. The heterogeneity of Systems on Chip (SoC) requires specific knowledge of their hardware in order to harness their full potential. OpenCL is a standard for cross-platform access to accelerator devices. By following an annotation-based approach we can automatically translate a program written in a high-level language into OpenCL. This helps to reduce the cost of development of OpenCL code for mobile devices. With our approach, the programmer can select from different target languages the one that is better suited to implement each section of an application. Our automatically-generated OpenCL code can outperform Renderscript in many cases.

Key words: parallelizing compiler, source-to-source translation, annotation based, opencl, android, heterogeneous architecture

1 Introduction

Modern SoCs now integrate technologies previously only available in servers and desktop computers, such as multicore CPUs, GPUs and DSPs, in a reduced form factor and under higher energy consumption constraints. The architecture of these devices is composed of a set of heterogeneous processors and a single memory that is shared between them. Some of the most relevant SoCs in the market today are Samsung Exynos, Qualcomm Snapdragon, Apple Ax and Nvidia Tegra, which are all based on ARM technology.

Due to these features, a software developer looking for the fastest or most energy-efficient way of running an algorithm on mobile devices, usually needs to develop parallel code to run on the best fitting processor available. Even though companies behind the most widely used mobile operating systems, such as Android or iOS, provide development frameworks that ease the creation of applications for their platforms, their support for high

 ISBN: 978-84-617-8694-7

performance computing is still limited. They provide programming languages, like Renderscript and Metal, that provide a higher-level view on the hardware and a way to accelerate parallel algorithms. However, learning these languages and developing applications with them require a non-trivial effort from programmers.

Lower-level computing APIs, such as CUDA [4] or OpenCL [5], can provide the best performance because of the finer control they provide, but that requires taking the specific hardware architecture of each device into consideration and spending a much larger amount of time fine-tuning the code. In the context of mobile architectures, the existing heterogeneity makes this prohibitely expensive. The key to achieving high performance at a low development cost has to come from powerful abstractions that let an automated system make use of these lower-level APIs in a cross-platform and efficient way.

Paralldroid [1, 2, 3] is a development framework that allows the automatic generation of Native C and Renderscript code for Android devices. This system defines a reduced set of annotations that, applied to the main components of a Java class, indicate the required memory movements and parallel methods of that class. It always creates a Java class with the API of the original class and some additions, enabling a seamless integration with the rest of the application. It is an evolution over other annotation-based approaches such as OpenACC or OpenMP, because it defines higher level anotations, and it is better suited to the object-oriented programming paradigm.

The main contributions of this paper are:

- The OpenCL programming model, whose importance is well known in desktop systems, is introduced to mobile devices by allowing transparent execution from Android Java applications.
- OpenCL support opens the possibility to develop accelerated Java applications for platforms other than Android using Paralldroid. The only requirements for these platforms are Java and OpenCL support.
- Our new approach lets high-level Java developers take advantage of more efficient GPU executions without modifying the annotated Java code. Performance improvements come from the use of a lower-level library for heterogeneous computing and, therefore, an increase in the code generation process complexity.

2 Methodology

In Android, applications are written mainly in Java, but it is also possible to use the Java Native Interface (JNI) to implement methods in C/C++, and Renderscript to accelerate parallel algorithms.

Paralldroid is embedded in the compilation process of OpenJDK, and is able to process the Abstract Syntax Tree (AST) of the Java code detecting a custom set of annotations,

 ISBN: 978-84-617-8694-7

summarized in Table 1. Depending on the target language selected for each class, a set of translator classes will be created and used to process them. These translators turn the original code and annotations into another AST that will be converted into code.

Annotation	Applied to	Parameters	Scope
@Target	Classes	value	—
@Map	Fields, method parameters	value	@Target
@Declare	Fields, methods	size	@Target
@Parallel	Methods	—	@Target
@Input	Method parameters	—	@Parallel
@Output	Method parameters	—	@Parallel
@NumThreads	Methods, method parameters	field	@Parallel
@Index	Method parameters	—	@Parallel

Table 1: Paralldroid annotations

For OpenCL support we implemented three translators. The Java translator generates a modified Java code that interfaces with the Java application, keeping the same interface as the original class, while forwarding the code execution to a native context. The native translator generates C code that handles communications with the Java context through JNI and with the OpenCL context, acting as the OpenCL host code that sets up memory buffers and enqueues commands. The OpenCL C translator generates the parallel kernels and support functions that get embedded as a string constant into the native source code and compiled at runtime. A native support library implements error handling routines and manages shared OpenCL resources, such as the context, platform, command queue, etc.

This way, developers do not have to modify the rest of the application when testing different backends, and they can improve the generated code by hand if necessary. This greatly simplifies the process of accelerating parallel algorithms in Android.

3 Computational results

In Figure 1 we show the results we obtained by running a set of image processing algorithms implemented in Java and Paralldroid-generated Renderscript and OpenCL, in the Sony Xperia Z (labelled SXZ) and Odroid XU3 (labelled XU3) devices. In the XU3, we found that OpenCL execution was slower than Renderscript due to our limitation to use only one GPU, whereas in this device the GPU is partitioned and reported as two separate GPUs by the OpenCL driver. Furthermore, all Renderscript executions are performed by the CPU, which turned out to be faster than the GPU in this platform. However, when Renderscript and OpenCL run on the same hardware, we find that our OpenCL code runs faster in most cases.

 ISBN: 978-84-617-8694-7

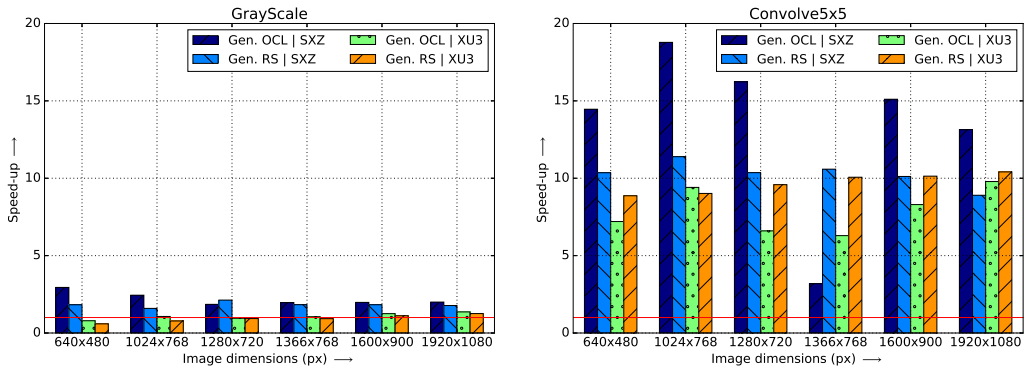


Figure 1: Speed-up obtained with respect to the sequential Java version

Acknowledgement

This work was supported by the EC (ERDF), the NESUS IC1315 COST Action, the Spanish Ministry of Economy, Industry and Competitiveness through the TIN2016-78919-R project, and the CAPAP-H network.

References

[1] Acosta, A., Afonso, S., and Almeida, F. Extending paralldroid with object oriented annotations. *Parallel Computing 57* (2016), 25 – 36.

[2] Acosta, A., and Almeida, F. Towards a unified heterogeneous development model in android. In *Eleventh International Workshop HeteroPar'2013: Algorithms, Models and Tools for Parallel Computing on Heterogeneous Platforms* (2013).

[3] Acosta, A., and Almeida, F. The particle filter algorithm: parallel implementations and performance analysis over android mobile devices. *Concurrency and Computation: Practice and Experience 28*, 3 (2016), 788–801.

[4] Nvidia Corporation. *CUDA C Programming Guide*, 2017.

[5] Stone, J. E., Gohara, D., and Shi, G. OpenCL: A parallel programming standard for heterogeneous computing systems. *Computing in Science Engineering 12*, 3 (2010), 66–73.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

Energetic BEM for soft and hard scattering of 2D damped waves by open arcs

Alessandra Aimi[1], Mauro Diligenti[1] and Chiara Guardasoni[1]

[1] *Dept. of Mathematical, Physical and Computer Sciences, University of Parma, Italy*

emails: alessandra.aimi@unipr.it, mauro.diligenti@unipr.it,
chiara.guardasoni@unipr.it

Abstract

The analysis of damping phenomena, that occur in many physics and engineering problems, reformulated in terms of boundary integral equations and solved via the boundary element method is a novelty. In this context, we develop here an extension of the so-called energetic boundary element method, recently proposed in literature, for soft and hard scattering of 2D damped waves by open arcs, directly in space-time domain. The presented numerical results confirm accuracy and stability of the proposed technique, already proved for the numerical treatment of undamped wave propagation problems in several dimensions and for the 1D damped case.

Key words: Damped wave equation, scattering, energy, boundary element method
MSC 2000: 65M38

1 Introduction

Wave propagation and scattering are two of the most studied physics phenomena that can be well described by mathematical models, leading to the problem of solving linear hyperbolic partial differential equations (PDEs) in two or three dimensional space [1, 2]. These problems are normally considered in an unbounded homogeneous domain and a method to tackle them can be the reformulation of the PDE as a boundary integral equation (BIE) on the usually bounded boundary of the domain, which then is numerically solved using the boundary element method (BEM) [3].
In the context of wave propagation, while the elastic forces tends to maintain the oscillatory motion, the transient effect dies out because of energy dissipations. The process of energy

 ISBN: 978-84-617-8694-7

dissipation is generally referred to as damping. The analysis of damping phenomena that occur, for example, in fluid dynamics, in kinetic theory and in semiconductors, is of particular interest: the dissipation is generated by the interaction between the waves and the propagation medium and can be also closely related to the dispersion, as in the interactions between water streams and surface waves or in ferromagnetic materials. On the other side, in mechanical systems, in general, damping has the effect of reducing the amplitude of vibration and, therefore, it is desirable to have some amount of damping in order to achieve stability in a faster way. Hence, damping is whether an unavoidable presence in physical reality or a desired characteristic in design.

The use of advanced numerical techniques to solve the related PDEs, such as finite element (FEM) and the finite difference (FDM) methods is well established and it is standard in this framework, even if the research of a numerical method that could reproduce the expected damping decay is an actual argument in literature [4, 5, 6]. On the other hand, in the context of BEMs, the analysis of dissipation through damped wave equation rewritten as a BIE is a relatively new topic, because it has been scarcely investigated until now. Since wave propagation phenomena are often observed in infinite media (domain) where Sommerfeld radiation condition holds, a suitable numerical method has to ensure that this condition is not violated. For example, FEMs need the application of special techniques to fulfill this condition that, on the contrary, is implicitly fulfilled by BEM.

In principle, both frequency-domain and time-domain BEM can be used for hyperbolic initial-boundary value problems [7, 8, 9, 10]. Space-time BEM has the advantage that it directly yields the unknown time-dependent quantities. In this last approach, the construction of the BIEs, via representation formula in terms of single and double layer potentials, uses the fundamental solution of the hyperbolic partial differential equation and jump relations [11, 12]. The mathematical background of time-dependent boundary integral equations is summarized by M. Costabel in [13].

For the numerical solution of the damped wave equation in 1D unbounded media, we have already considered in [14, 15, 16] the extension of the so-called energetic BEM, introduced for the undamped wave equation in several space dimensions [17, 18, 19]. Energetic BEM is based on a weak formulation directly expressed in the space-time domain, thus avoiding the use of the Laplace transform and of its inversion suggested in [11].

The analysis carried out for 1D damped wave propagation problems allowed to fully understand the approximation technique for what concerns marching on time, avoiding space integration with BEM singular kernels and it was considered as a touchstone for the extension to higher space dimensions, which is done here for the 2D case. In particular, we will treat soft and hard scattering of damped waves by open arcs. Also in this context, the energetic approach gives consistent approximations and accurate numerical solutions.

The paper is structured as follows: at first, we present the differential model problem outside an open arc in the plane, equipped by Dirichlet boundary conditions (soft scattering),

 ISBN: 978-84-617-8694-7

and we introduce the corresponding boundary integral weak formulation. Then, we illustrate the consequent energetic BEM discretization. Further, to complete the presentation, we describe an analogous model problem equipped by Neumann boundary condition (hard scattering), always discretized by energetic BEM. At last, some numerical benchmarks are introduced and discussed, showing, from a numerical point of view, stability and accuracy of the obtained numerical solutions.

2 Model problem for soft scattering and its weak boundary integral formulation

We consider a Dirichlet problem in a bounded time interval $[0,T]$ for the damped wave equation exterior to an obstacle, given by an open arc $\Gamma \subset \mathbb{R}^2$:

$$\left[\Delta u - \frac{1}{c^2}\, u_{tt} - \frac{2D}{c^2} u_t - \frac{P}{c^2} u\right](\mathbf{x},t) = 0, \qquad \mathbf{x} \in \mathbb{R}^2 \setminus \Gamma,\ t \in (0,T] \tag{1}$$

$$u(\mathbf{x},0) = u_t(\mathbf{x},0) = 0, \qquad \mathbf{x} \in \mathbb{R}^2 \setminus \Gamma, \tag{2}$$

$$u(\mathbf{x},t) = \bar{u}(\mathbf{x},t), \qquad \mathbf{x} \in \Gamma,\ t \in (0,T]\,, \tag{3}$$

where c is the propagation velocity of a perturbation inside the domain, D and P are the viscous and material damping coefficients, respectively. Equation (1) appears in different contexts, from electric transmission to sound propagation, from primary or secondary seismic wave propagation in presence of intrinsic attenuation to quantum field theory.
When $D = P = 0$ the given PDE collapses to the classic wave equation. The boundary datum $\bar{u}(\mathbf{x},t)$ represents the value of the excitation field over Γ.
Since we want to discretize the above problem using BEM, we have to rewrite it in a boundary integral form. This can be done using classical arguments and the knowledge of the fundamental solution of the 2D damped wave operator. Hence, we start writing the single-layer representation of the solution of (1)-(3):

$$u(\mathbf{x},t) = \int_\Gamma \int_0^t G(\mathbf{x}-\boldsymbol{\xi}, t-\tau)\phi(\boldsymbol{\xi},\tau)\, d\tau\, d\gamma_{\boldsymbol{\xi}}, \qquad \mathbf{x} \in \mathbb{R}^2 \setminus \Gamma, t \in (0,T], \tag{4}$$

where the unknown density $\phi = \left[\dfrac{\partial u}{\partial \mathbf{n}}\right]_\Gamma$ represents the time history of the jump of the normal derivative of u along Γ and

$$G(\mathbf{x},t) = \begin{cases} \frac{c}{2\pi}\, e^{-Dt}\, \dfrac{\cos\left(\frac{\sqrt{P-D^2}}{c}\sqrt{c^2t^2-\|\mathbf{x}\|^2}\right)}{\sqrt{c^2t^2-\|\mathbf{x}\|^2}}\, H[c\,t-\|\mathbf{x}\|]\,, & P \geq D^2 \\ \\ \frac{c}{2\pi}\, e^{-Dt}\, \dfrac{\cosh\left(\frac{\sqrt{D^2-P}}{c}\sqrt{c^2t^2-\|\mathbf{x}\|^2}\right)}{\sqrt{c^2t^2-\|\mathbf{x}\|^2}}\, H[c\,t-\|\mathbf{x}\|]\,, & P \leq D^2 \end{cases} \tag{5}$$

 ISBN: 978-84-617-8694-7

is the forward fundamental solution of the 2D damped wave operator, with $H[\cdot]$ the Heaviside distribution. Definition (5) switches from $\cos(\cdot)$ to $\cosh(\cdot)$ depending on the reciprocal magnitude of P and D^2: when $P > D^2$ we are in the so-called *underdamping* configuration, when $P < D^2$ we are in *overdamping* configuration, while the separation state $P = D^2$, referred to the vanishing of both $\cos(\cdot)$ and $\cosh(\cdot)$ arguments, is called *critical damping*. Note that in the limit for D, P tending to 0, $G(\mathbf{x}, t)$ tends to the fundamental solution of the 2D undamped wave operator.
Now, it is clear that if we want to recover the solution of the differential problem at any point outside the obstacle and at any time instant, we have to proceed with a post-processing phase provided that we know the density function $\phi(\mathbf{x}, t)$. Hence, since the extension of u in (4) for $\mathbf{x}$ tending to Γ is continuous, the unknown density ϕ can be determined via the assigned Dirichlet boundary condition (3). This results in the space-time BIE

$$\int_\Gamma \int_0^t G(\mathbf{x} - \boldsymbol{\xi}, t - \tau)\phi(\boldsymbol{\xi}, \tau)\, d\tau\, d\gamma_{\boldsymbol{\xi}} = \bar{u}(\mathbf{x}, t), \qquad \mathbf{x} \in \Gamma, t \in [0, T], \tag{6}$$

which can be written with the compact notation

$$V\phi = \bar{u}. \tag{7}$$

The energetic weak formulation of problem (7) is defined similarly as in [18] and it can be deduced observing that, multiplying the PDE (1) by u_t, integrating over $[0, T] \times (\mathbb{R}^2 \setminus \Gamma)$ and using integration by parts in space, one obtains that the energy $\mathcal{E}(u, T)$ of the solution u at the final time of analysis T, defined by

$$\frac{1}{2}\int_{\mathbb{R}^2\setminus\Gamma} \Big[\|\nabla_{\mathbf{x}} u(\mathbf{x}, T)\|^2 + \frac{1}{c^2}u_t^2(\mathbf{x}, T) + \frac{P}{c^2}u^2(\mathbf{x}, T) + \frac{4D}{c^2}\int_0^T u_t^2(\mathbf{x}, t)dt\Big] d\gamma_{\mathbf{x}} \tag{8}$$

can be rewritten as

$$\mathcal{E}(u, T) = \int_\Gamma \int_0^T u_t(\mathbf{x}, t)\left[\frac{\partial u}{\partial \mathbf{n}}\right]_\Gamma (\mathbf{x}, t)\, dt\, d\gamma_{\mathbf{x}}\,. \tag{9}$$

Hence, remembering the nature of our BIE which is expressing the time history of u over Γ, we can derive it w.r.t. time and write down the energetic weak problem associated to (7) as:
find $\phi \in L^2([0, T]; H^{-1/2}(\Gamma))$ *such that*

$$\int_\Gamma \int_0^T (V\phi)_t(\mathbf{x}, t)\psi(\mathbf{x}, t)\, dt\, d\gamma_{\mathbf{x}} = \int_\Gamma \int_0^T \bar{u}_t(\mathbf{x}, t)\psi(\mathbf{x}, t)\, dt\, d\gamma_{\mathbf{x}}\,, \tag{10}$$

where ψ is a suitable test function, belonging to the same functional space of ϕ.
With an integration in the sense of distributions, we can equivalently write

$$\int_\Gamma \int_0^T (V\phi)(\mathbf{x}, t)\psi_t(\mathbf{x}, t)\, dt\, d\gamma_{\mathbf{x}} = \int_\Gamma \int_0^T \bar{u}(\mathbf{x}, t)\psi_t(\mathbf{x}, t)\, dt\, d\gamma_{\mathbf{x}}\,. \tag{11}$$

 ISBN: 978-84-617-8694-7

The theoretical analysis of the quadratic form coming from the left-hand side of (10) was carried out for $P = D = 0$ in [18] where, under suitable hypothesis, coercivity was proved with some technicalities. This allowed us to deduce stability and convergence of the related Galerkin approximate solution, which here, for the case of non-trivial damping coefficients, will be verified from a numerical point of view.

3 Energetic BEM discretization

We consider on the obstacle Γ, a boundary mesh constituted by $M_{\Delta x}$ straight elements $\{e_1, \cdots, e_{M_{\Delta x}}\}$, with $length(e_i) \leq \Delta x$, $e_i \cap e_j = \emptyset$ if $i \neq j$ and such that $\bigcup_{i=1}^{M_{\Delta x}} \overline{e}_i$ coincides with $\overline{\Gamma}$ if the obstacle is (piece-wise) linear, or is a suitable approximation of $\overline{\Gamma}$, otherwise. The functional background compels one to choose space shape functions belonging to $L^2(\Gamma)$, although higher degree shape functions can be used. Hence we use standard piece-wise constant polynomial boundary element functions $w_j(\mathbf{x})$, $j = 1, \cdots, M_{\Delta x}$, suitably defined in relation to the introduced mesh over Γ.
For time discretization we consider a uniform decomposition of the time interval $[0, T]$ with time step $\Delta t = T/N_{\Delta t}\,, N_{\Delta t} \in \mathbb{N}^+$, generated by the $N_{\Delta t} + 1$ instants

$$t_k = k\,\Delta t, \quad k = 0, \cdots, N_{\Delta t}\,, \tag{12}$$

and we choose piece-wise constant time shape functions. Note that, for this particular choice, our shape functions, denoted with $v_k(t)$, $k = 0, \cdots, N_{\Delta t} - 1$, will be defined as

$$v_k(t) = H[t - t_k] - H[t - t_{k+1}]\,. \tag{13}$$

Hence, the approximate solution of the problem at hand will be expressed as

$$\phi(\mathbf{x}, t) \simeq \sum_{k=0}^{N_{\Delta t}-1} \sum_{j=1}^{M_{\Delta x}} \alpha_j^{(k)}\, w_j(\mathbf{x})\, v_k(t). \tag{14}$$

The Galerkin BEM discretization coming from energetic weak formulation (11) produces the linear system

$$A\,\alpha = b\,, \tag{15}$$

of order $M_{\Delta x} \cdot N_{\Delta t}$, where matrix A has a block lower triangular Toeplitz structure. Each block has dimension $M_{\Delta\, x}$. If we indicate with $A^{(\ell)}$ the block obtained when $t_h - t_k = \ell\,\Delta t$, $\ell = 0, \ldots, N_{\Delta t} - 1$, the linear system can be written as

$$\begin{pmatrix} A^{(0)} & 0 & 0 & \cdots & 0 \\ A^{(1)} & A^{(0)} & 0 & \cdots & 0 \\ A^{(2)} & A^{(1)} & A^{(0)} & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & 0 \\ A^{(N_{\Delta t}-1)} & A^{(N_{\Delta t}-2)} & \cdots & A^{(1)} & A^{(0)} \end{pmatrix} \begin{pmatrix} \alpha^{(0)} \\ \alpha^{(1)} \\ \alpha^{(2)} \\ \vdots \\ \alpha^{(N_{\Delta t}-1)} \end{pmatrix} = \begin{pmatrix} b^{(0)} \\ b^{(1)} \\ b^{(2)} \\ \vdots \\ b^{(N_{\Delta t}-1)} \end{pmatrix} \tag{16}$$

 ISBN: 978-84-617-8694-7

where

$$\alpha^{(\ell)} = \left(\alpha_j^{(\ell)}\right) \quad \text{and} \quad b^{(\ell)} = \left(b_j^{(\ell)}\right), \quad j = 1, \ldots, M_{\Delta x}\,. \tag{17}$$

The solution of (16) is obtained with a block forward substitution, i.e. at every time instant $t_\ell = \ell\,\Delta t$, $\ell = 0, \cdots, N_{\Delta t} - 1$, we solve a reduced linear system of the type

$$A^{(0)}\alpha^{(\ell)} = b^{(\ell)} - (A^{(1)}\alpha^{(\ell-1)} + \cdots + A^{(\ell)}\alpha^{(0)}). \tag{18}$$

Procedure (18) is a time-marching technique, where the only matrix to be inverted is the symmetric positive definite diagonal block $A^{(0)}$, while all the other blocks are used to update at every time step the right-hand side. Owing to this procedure we can construct and store only the blocks $A^{(0)}, \cdots, A^{(N_{\Delta t}-1)}$ with a considerable reduction of computational cost and memory requirement.
Further, matrix elements are generated by suitable quadrature schemes, already used in [18], taking into account kernel singularity.

4 The case of hard scattering

Here, the scattered wave $u(\mathbf{x}, t)$ satisfies problem (1)-(2), equipped now by Neumann boundary conditions:

$$q(\mathbf{x}, t) := \frac{\partial u}{\partial \mathbf{n}}(\mathbf{x}, t) = \bar{q}(\mathbf{x}, t), \qquad \mathbf{x} \in \Gamma,\ t \in (0, T]\,, \tag{19}$$

where the datum $\bar{q}$ represents the opposite of the normal derivative of the incident wave along Γ, i.e. $\bar{q} = -\frac{\partial u^I}{\partial \mathbf{n}}$.
In order to obtain a boundary integral reformulation of the problem at hand, we consider the double-layer representation of the solution :

$$u(\mathbf{x}, t) = \int_\Gamma \int_0^t \frac{\partial G}{\partial \mathbf{n}_{\boldsymbol{\xi}}}(\mathbf{x} - \boldsymbol{\xi}, t - \tau)\varphi(\boldsymbol{\xi}, \tau)\, d\tau\, d\gamma_{\boldsymbol{\xi}}, \qquad \mathbf{x} \in \mathbb{R}^2 \setminus \Gamma, t \in (0, T], \tag{20}$$

where the unknown density $\varphi = [u]_\Gamma$ represents the time history of the jump of u along Γ and G is given in (5).
Again, it is clear that if we want to recover the solution of the differential problem at any point outside the obstacle and at any time instant, we have to proceed with a post-processing phase provided that we know the density function $\varphi(\mathbf{x}, t)$, which can be determined applying the normal derivative w.r.t. $\mathbf{x}$ to (20) and using the assigned Neumann boundary condition (19). This results in the hyper-singular space-time BIE

$$\int_\Gamma \int_0^t \frac{\partial^2 G}{\partial \mathbf{n}_{\mathbf{x}} \partial \mathbf{n}_{\boldsymbol{\xi}}}(\mathbf{x} - \boldsymbol{\xi}, t - \tau)\varphi(\boldsymbol{\xi}, \tau)\, d\tau\, d\gamma_{\boldsymbol{\xi}} = \bar{q}(\mathbf{x}, t), \qquad \mathbf{x} \in \Gamma, t \in [0, T], \tag{21}$$

 ISBN: 978-84-617-8694-7

which can be written with the compact notation

$$D\phi = \bar{q}. \tag{22}$$

For this model problem, the energy (8) can be rewritten as

$$\mathcal{E}(u,T) = \int_\Gamma \int_0^T [u_t]_\Gamma(\mathbf{x},t)\, \frac{\partial u}{\partial \mathbf{n}}(\mathbf{x},t)\, dt\, d\gamma_{\mathbf{x}}\,. \tag{23}$$

Hence, remembering the nature of our BIE which is expressing the time history of $\frac{\partial u}{\partial \mathbf{n}}$ over Γ, we can write down the related energetic weak problem as:
find $\varphi \in H^1([0,T]; H_0^{1/2}(\Gamma))$ *such that*

$$\int_\Gamma \int_0^T (D\varphi)(\mathbf{x},t)\psi_t(\mathbf{x},t)\, dt\, d\gamma_{\mathbf{x}} = \int_\Gamma \int_0^T \bar{q}(\mathbf{x},t)\psi_t(\mathbf{x},t)\, dt\, d\gamma_{\mathbf{x}}\,. \tag{24}$$

where ψ is a suitable test function, belonging to the same functional space of φ.

Also in this case, the theoretical analysis of the quadratic form coming from the left-hand side of (24) was carried out for $P = D = 0$ in [21] where, under suitable hypothesis, coercivity was proved with some technicalities. This allowed us to deduce stability and convergence of the related Galerkin approximate solution, which here, for the case of non-trivial damping coefficients, will be verified from a numerical point of view.

For what concerns the energetic BEM discretization, we consider on the obstacle Γ, the boundary mesh already described at the beginning of Section 3. The functional background compels one to choose space shape functions belonging to $H_0^1(\Gamma)$, although higher degree shape functions can be used. Hence we use standard piece-wise linear polynomial boundary element functions $w_j(\mathbf{x})$, $j = 1, \cdots, M_{\Delta x} - 1$, suitably defined in relation to the introduced mesh over Γ.
For time discretization we consider a uniform decomposition of the time interval $[0,T]$ with time step $\Delta t = T/N_{\Delta t}\,, N_{\Delta t} \in \mathbb{N}^+$, generated by the $N_{\Delta t} + 1$ instants (12) and we choose piece-wise linear time shape functions. Note that, for this particular choice, our shape functions, denoted with $v_k(t)$, $k = 0, \cdots, N_{\Delta t} - 1$, will be defined as

$$v_k(t) = R(t - t_k) - 2\, R(t - t_{k+1}) + R(t - t_{k+2})\,, \tag{25}$$

where $R(t - t_k) = \frac{t - t_k}{\Delta t}\, H[t - t_k]$ is the ramp function. Hence, the approximate solution of the problem at hand will be expressed as

$$\varphi(\mathbf{x},t) \simeq \sum_{k=0}^{N_{\Delta t}-1} \sum_{j=1}^{M_{\Delta x}-1} \alpha_j^{(k)}\, w_j(\mathbf{x})\, v_k(t). \tag{26}$$

The Galerkin BEM discretization coming from energetic weak formulation (24) produces a linear system analogous to (15) and with the same properties. Matrix elements are generated by suitable quadrature schemes, already used in [21], taking into account kernel singularity.

 ISBN: 978-84-617-8694-7

5 Numerical results

In the following, we will present and discuss numerical results obtained by energetic BEM applied to for 2D Dirichlet or Neumann damped wave propagation exterior problems.
At first, we consider the problem (1)-(2)-(3) with $\Gamma = \{\mathbf{x} = (x, 0) \,|\, x \in [-0.5, 0.5]\}$ and the Dirichlet boundary datum

$$\bar{u}(\mathbf{x}, t) = -H[t] f(t)\, x, \quad \text{where} \quad f(t) = \begin{cases} \sin^2(4\pi\, t)\,, & \text{if} \quad 0 \le t \le \dfrac{1}{8} \\ 1\,, & \text{if} \quad t \ge \dfrac{1}{8}\,, \end{cases} \tag{27}$$

taken from [18]. The velocity is fixed as $c = 1$.
We choose a uniform decomposition of Γ in 40 straight elements ($\Delta x = 0.025$) and we set $\Delta t = 0.025$. In Figure 1, we show the time history of the approximate solutions $\phi(x, t)$ of BIE (6), on the straight element e_{10} and on the time interval $[0, T] = [0, 2]$, for $P = 0$ varying $D = 0, 1, 2, 5, 10, 20$ (left) and for $D = 0$ varying $P = 0, 1, 10, 20$ (right). Note the effects of increasing viscous and material damping, which substantially change the aspect of the solution related to the classical wave equation, visible in the graphs for trivial parameters.

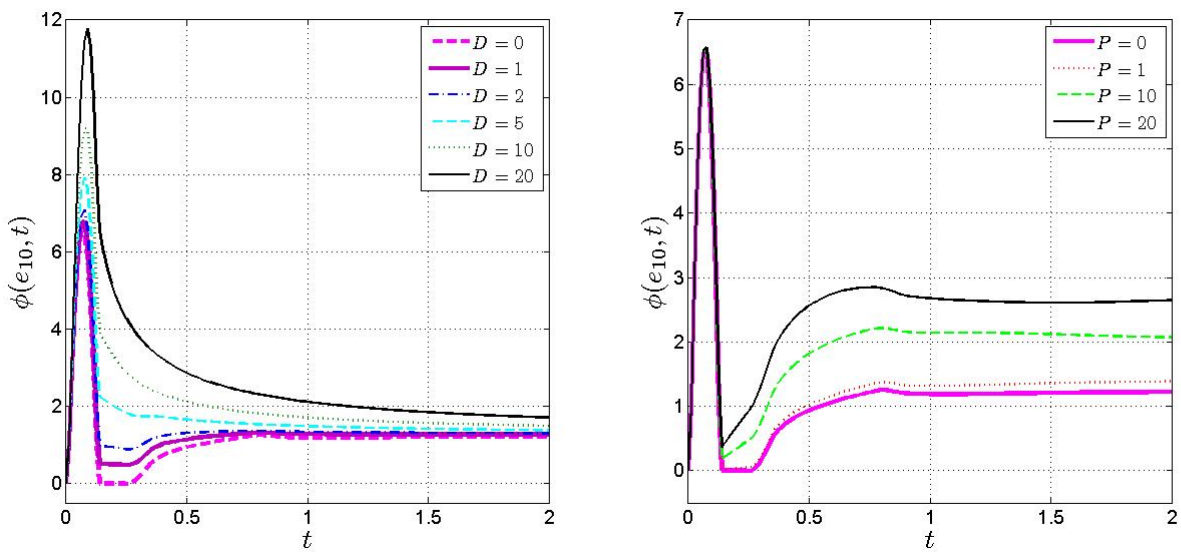


Figure 1: Time history of $\phi(x, t)$ on element e_{10}, for $P = 0$ (left) and $D = 0$ (right).

Since the Dirichlet datum becomes independent of time, it has been numerically checked on the time interval $[0, T] = [0, 60]$ that, for $P = 0$, the BIE transient solution $\phi(x, t)$ on

 ISBN: 978-84-617-8694-7

Γ tends to the stationary one $\phi_\infty(x)$, i.e. the solution of the BIE related to the following Dirichlet problem for the Laplace equation:

$$\begin{array}{ll} \Delta u_\infty(\mathbf{x}) = 0, & \mathbf{x} \in \mathbb{R}^2 \setminus \Gamma \\ u_\infty(\mathbf{x}) = -x, & \mathbf{x} \in \Gamma \\ u_\infty(\mathbf{x}) = O(1), & \|\mathbf{x}\| \to \infty \end{array} \tag{28}$$

while, for $P > 0$, the transient solution $\phi(x,t)$ on Γ tends to the stationary one $\phi_{\infty,k}(x)$, i.e. the solutions of the BIE related to the following Dirichlet problem for the Helmholtz equation:

$$\begin{array}{ll} \Delta u_{\infty,k}(\mathbf{x}) + k^2 u_{\infty,k}(\mathbf{x}) = 0, & \mathbf{x} \in \mathbb{R}^2 \setminus \Gamma \\ u_{\infty,k}(\mathbf{x}) = -x, & \mathbf{x} \in \Gamma \\ u_{\infty,k}(\mathbf{x}) = O(\|\mathbf{x}\|^{-1}), & \|\mathbf{x}\| \to \infty \end{array} \tag{29}$$

with wave number $k = \sqrt{-P}/c^2$.

For the case of hard scattering, we consider the model problem (1)-(2)-(19) fixing $\Gamma = \{\mathbf{x} = (x,0) \,|\, x \in [0,1]\}$, $c = 1$, $[0,T] = [0,5]$ and Neumann boundary datum coming from an incident plane linear wave $u^I(\mathbf{x},t)$ propagating in direction $\mathbf{k} = (\cos\theta, \sin\theta)$, i.e.

$$\bar{q}(\mathbf{x},t) = -\frac{\partial}{\partial \mathbf{n}_\mathbf{x}} f(t - \mathbf{k}\cdot\mathbf{x})\Big|_\Gamma. \tag{30}$$

with $f(t) = 0.5\, t\, H[t]$. In this case, the Neumann datum (30) tends to the constant value $\bar{q}_\theta = 0.5 \sin\theta$, when t tends to infinity, so we expect that the approximate transient solution $\varphi(x,t)$ of BIE (21) on Γ will tend to the BIE solution of the corresponding Laplace ($P = 0$) or Helmholtz ($P > 0$) problems, equipped with the same Neumann datum $\bar{q}_\theta$.

For an incident angle of $\pi/3$, in Figure 2 on the left we present for $D = 1$ and $P = 0$, the transient solution $\varphi(x,T)$ on Γ at the final time instant of analysis, which can be compared with the corresponding static solution shown in the same Figure on the right. A similar comparison can be done looking at Figure 3, having fixed $\pi/3$, $D = 0$ and $P = 1$.
In both cases, the approximate solutions obtained by energetic BEM with discretization parameters fixed as $\Delta x = 0.05$ and $\Delta t = 0.05$ show an optimal accuracy.

Acknowledgements

This work has been partially supported by INdAM-GNCS 2017 Project "Nuove tecniche numeriche per la risoluzione di problemi evolutivi mediante il metodo degli elementi di contorno".

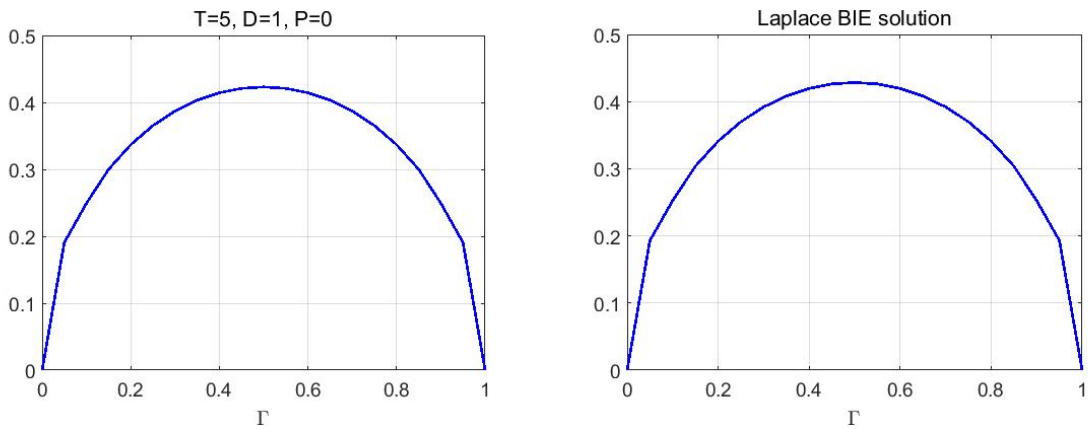


Figure 2: $\varphi(x,T)$ on Γ, for $D=1$, $P=0$ (left) and corresponding static solution (right).

References

[1] Colton, D. and Kress, R., *Inverse Acoustic and Electromagnetic Scattering Theory*, Springer, 2012.

[2] Leung, T., Jin A. K., Kung-Han, D. and Chi O. A., *Scattering of Electromagnetic waves, Numerical Simulations*, John Wiley & Sons., 2004.

[3] Banerjee, P. K. and Butterfield, P. K., *Boundary Element Methods in Engineering*, McGraw-Hill U.K. Ltd., 1981.

[4] Münch, A. and Pazoto, A. F., *Uniform stabilization of a viscous numerical approximation for a locally damped wave equation*, ESAIM Control Optim. Calc. Var. **13**(2) (2007) 265–293.

[5] Tébou, T., Roder, L. and Zuazua, E., *Uniform exponential long time decay for the space semi-discretization of a locally damped wave equation via an artificial numerical viscosity*, Numer. Math. **95**(3) (2003) 563–598.

[6] Zuazua, E., *Propagation, observation, and control of waves approximated by finite difference methods*, SIAM Rev. **47**(2) (2005) 197–243.

[7] Bamberger, A. and T. Ha Duong, *Formulation variationnelle espace-temps pour le calcul par potentiel retardé de la diffraction d'une onde acoustique. I*, Math. Methods Appl. Sci. **8**(3) (1986) 405–435.

 ISBN: 978-84-617-8694-7

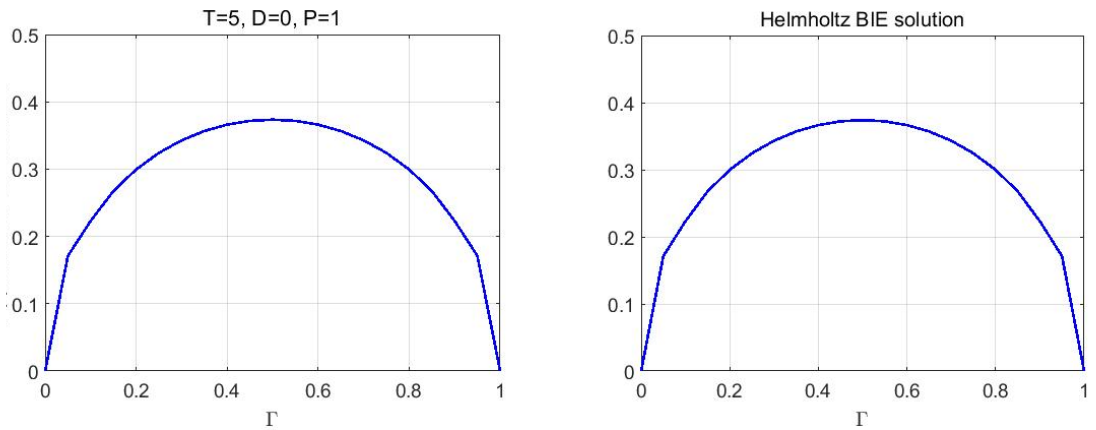


Figure 3: $\varphi(x,T)$ on Γ, for $D=0$, $P=1$ (left) and corresponding static solution (right).

[8] Bamberger, A. and T. Ha Duong, *Formulation variationnelle pour le calcul de la diffraction d'une onde acoustique par une surface rigide*, Math. Methods Appl. Sci. **8**(4) (1986) 598–608.

[9] Lubich, C., *Convolution quadrature and discretized operational calculus I*, Numer. Math. **52** (1988) 129–145.

[10] Lubich, C., *Convolution quadrature and discretized operational calculus II*, Numer. Math. **52** (1988) 413–425.

[11] Ha Duong, T., *On retarded potential boundary integral equations and their discretization*, in P. Davies et al. Eds.: Topics in computational wave propagation. Direct and inverse problems, Springer-Verlag, 2003, 301–336.

[12] Vick, A. and West, R.L., *Analysis of Damped Wave Using the Boundary Element Method*, Trans. Model. Simulation **15** (1997) 265–278.

[13] Costabel, M., *Time-dependent problems with the boundary integral equation method*, in E. Stein et al. Eds.: Encyclopedia of Computational Mechanics, John Wiley and Sons, 2004, 1–28.

[14] Aimi, A. and Panizzi, S., *BEM-FEM coupling for the 1D Klein-Gordon equation*, Numer. Methods Partial Differential Equations **30**(6) (2014) 2042–2082.

[15] Aimi, A. and Diligenti, M. and Guardasoni, C., *Energetic BEM-FEM coupling for the numerical solution of the damped wave equation*, Adv. Comput. Math. (2017) in press.

 ISBN: 978-84-617-8694-7

[16] Aimi, A. and Diligenti, M. and Guardasoni, C., *Comparison between numerical methods applied to the damped wave equation*, J. Integral Equations Appl. **29**(1) (2017) 5–40.

[17] Aimi, A. and Diligenti, M., *A new space-time energetic formulation for wave propagation analysis in layered media by BEMs*, Int. J. Numer. Meth. Engng. **75** (2008) 1102–1132.

[18] Aimi, A. and Diligenti, M. and Guardasoni, C. and Mazzieri, I. and Panizzi, S., *An energy approach to space-time Galerkin BEM for wave propagation problems*, Int. J. Numer. Meth. Engng. **80**(9) (2009) 1196–1240.

[19] Aimi, A. and Diligenti, M. and Frangi, A. and Guardasoni, C., *Neumann exterior wave propagation problems: Computational aspects of 3D energetic Galerkin BEM*, Comput. Mech. **51**(4) (2013) 475–493.

[20] Antes, H. and Beer, G. and Moser, W., *Soil-structure interaction and wave propagation problems in 2D by a Duhamel integral based approach and the convolution quadrature method*, Comput. Mech. **36**(6) (2005) 431–443.

[21] A. Aimi, M. Diligenti and S. Panizzi, *Energetic Galerkin BEM for wave propagation Neumann exterior problems*, CMES **1**(1) (2009) 1–33.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

Numerical Pricing of Geometric Asian Options with Barriers

A. Aimi[1] **and C. Guardasoni**[1]

[1] *Department of Mathematical Physical and Computer Sciences, University of Parma*

emails: alessandra.aimi@unipr.it, chiara.guardasoni@unipr.it

Abstract

A new method, the so-called SABO (Semi-Analytical method for pricing of Barrier Options), recently introduced in literature for the evaluation of barrier options, is here described. The method, already applied in the context of European options, is now extended to Asian options with geometric mean.

Key words: Boundary Element Method, Geometric Asian Options, Barrier Options.
MSC 2000: 91G60, 65M38

1 Introduction

The variety of financial products in recent years grew much quicker than ever before. The recent financial crisis has highlighted the need for a more scientific approach to the problem of pricing of these products, taking advantage of more advanced statistical and mathematical skills and of the availability of numerical techniques and faster computer systems.
In particular in this paper we will illustrate a new algorithm, the so-called SABO (Semi-Analytical method for pricing of Barrier Options), to evaluate Asian options with barriers. Asian options are derivative contracts whose payoff at maturity depends on the (geometric or arithmetic) average value of an underlying asset over some time interval; in the case of "barrier option", these contracts get into existence or extinguish when the underlying asset reaches a certain barrier value. With respect to European vanilla option, the buyer has a reasonable protection against inconvenient fluctuations in the underlying price and the issuer can attain a better forecasting of the terminal position.
For standard Asian options with geometric mean equipped with floating or fixed strike price, closed formula solutions are available [7] but if the contract involves non standard

  ISBN: 978-84-617-8694-7

payoffs or arithmetic mean or barriers, numerical techniques are unavoidable. The pricing is then traditionally based on Monte Carlo methods [7] or on domain methods, such as Finite Volume Methods [9] and Finite Difference methods [2], but Monte Carlo methods are affected by high computational costs and inaccuracy due to their slow convergence and domain methods have some troubles particularly in unbounded domains.
Barrier options are largely exchanged as they are good products for hedging and investment and they are cheaper than vanilla options but, for Asian options, we found in literature only the analysis of [1] which provides rigorous bounds in the arithmetic mean case.
SABO has been recently introduced for the computation of European barrier options within various differential models ([4], [5] and[6]). This new approach, based on Boundary Element Method turns out to be stable and efficient especially when the differential problem is defined in an unbounded domain and the data are assigned on a limited boundary (which is the case of the "barrier options"). The method is particularly advantageous for its high accuracy, for the implicit satisfaction of the far-field behavior of the solution and for the low discretization costs. Moreover it provides a straight hedging computation. The essential requisite, that makes it not as general as other numerical methods is that, for its application, we need the knowledge of the transition probability density related to the vanilla option problem at least in an approximated form.
The method is here extended to Asian barrier options evaluated with geometric mean that, although not common among practitioners, give some information also about the evaluation of Asian barrier options with arithmetic mean. This work can be conceived as an intermediate and preparatory step in this last direction.

2 The model problem

A geometric Asian option V is an option depending on the evolution of the stock price S_t and on the geometric average of the stock price over some time interval

$$A_t := \int_0^t \log(S_t)dt\,.$$

If the stochastic process S_t is modeled by the usual geometric Brownian motion

$$dS_t = rS_t dt + \sigma S_t dW_t$$

where r denotes the risk free interest rate, σ the volatility and W_t a standard Wiener process, then, A_t is a lognormal stochastic process too.
With the classical hedging arguments applied in the Black-Scholes framework, it is possibile to conclude that the Asian option value $V(S, A, t)$ solves the following partial differential equation (PDE):

$$\frac{\partial V}{\partial t} + \frac{\sigma^2}{2}S^2\frac{\partial^2 V}{\partial S^2} + rS\frac{\partial V}{\partial S} + \log(S)\frac{\partial V}{\partial A} - rV = 0 \qquad S \in \mathbb{R}^+,\, A \in \mathbb{R},\, t \in [0, T)\,. \tag{1}$$

 ISBN: 978-84-617-8694-7

Different final boundary conditions (payoffs) define different types of contract, such as:

$$\begin{aligned}
\text{floating strike call} \quad & V(S,A,T) = \max\left(S - \exp\left(\frac{A}{T}\right), 0\right) \\
\text{floating strike put} \quad & V(S,A,T) = \max\left(\exp\left(\frac{A}{T}\right) - S, 0\right) \\
\text{fixed strike call} \quad & V(S,A,T) = \max\left(\exp\left(\frac{A}{T}\right) - E, 0\right) \\
\text{fixed strike put} \quad & V(S,A,T) = \max\left(E - \exp\left(\frac{A}{T}\right), 0\right)
\end{aligned}$$

for $S \in \mathbb{R}^+$, $A \in \mathbb{R}$ and E the strike price.
Some boundary conditions are implicitly satisfied by V through its payoff behavior and they are such to assure existence and uniqueness of the Cauchy partial differential problem solution. The exact solution can be defined by an integral form as payoff expected value and it can be therefore employed also with more general payoff contracts:

$$V(S,A,t) = \int_{-\infty}^{+\infty} \int_{0}^{+\infty} V(\widetilde{S},\widetilde{A},T) G(S,A,t;\widetilde{S},\widetilde{A},T) d\widetilde{S}\, d\widetilde{A}\,. \tag{2}$$

The function $G(S,A,t;\widetilde{S},\widetilde{A},\widetilde{t})$ is the *transition probability density function* (PDF), also known as *Green's function* or *fundamental solution* of the partial differential operator. As a function of $(S,A,t) \in \mathbb{R}^+ \times \mathbb{R} \times [0,T)$ the PDF solves (1) and, as a function of $(\widetilde{S},\widetilde{A},\widetilde{t})$, it solves the *backward Kolmogorov equation* adjoint of (1):

for each $(S,A,t) \in \mathbb{R}^+ \times \mathbb{R} \times [0,T)$

$$\begin{cases}
-\dfrac{\partial G}{\partial \widetilde{t}} + \dfrac{\sigma^2}{2}\widetilde{S}^2 \dfrac{\partial^2 G}{\partial \widetilde{S}^2} + (2\sigma^2 - r)\widetilde{S}\dfrac{\partial G}{\partial \widetilde{S}} - \log(\widetilde{S})\dfrac{\partial G}{\partial \widetilde{A}} + (\sigma^2 - 2r)G = 0 & \widetilde{S} \in \mathbb{R}^+,\ \widetilde{A} \in \mathbb{R},\ \widetilde{t} > t \\
G(S,A,t;\widetilde{S},\widetilde{A},t) = \delta(S-\widetilde{S})\delta(A-\widetilde{A}) & \widetilde{S} \in \mathbb{R}^+,\ \widetilde{A} \in \mathbb{R}
\end{cases} \tag{3}$$

where $\delta(\cdot,\cdot)$ represents the Dirac distribution. The solution of problem (3) must satisfy opportune boundary conditions assuring that the Green identity is verified.

 ISBN: 978-84-617-8694-7

We can prove, thanks to results in [8], that the closed form solution of problem (3) is

$$\begin{aligned} G(S,A,t;\widetilde{S},\widetilde{A},\widetilde{t}) &= \frac{\sqrt{3}\,H[\widetilde{t}-t]}{\pi\sigma^2(\widetilde{t}-t)^2}\exp\left\{-\frac{2}{\sigma^2(\widetilde{t}-t)}\log^2\left(\frac{S}{\widetilde{S}}\right)\right. \\ &+ \frac{6}{\sigma^2(\widetilde{t}-t)^2}\log\left(\frac{S}{\widetilde{S}}\right)\left(A-\widetilde{A}+(\widetilde{t}-t)\log(S)\right) \\ &- \frac{6}{\sigma^2(\widetilde{t}-t)^3}\left(A-\widetilde{A}+(\widetilde{t}-t)\log(S)\right)^2 \\ &- \left.\left(\frac{2r+\sigma^2}{2\sqrt{2}\sigma}\right)^2(\widetilde{t}-t)\right\}\left(\frac{\widetilde{S}}{S}\right)^{\frac{2r-\sigma^2}{2\sigma^2}}\frac{1}{\widetilde{S}} \end{aligned} \tag{4}$$

denoting by $H[\cdot]$ the Heaviside step function.
When considering fixed strike options or floating strike options, the exact solution can be evaluated also by other more efficient closed-formulas of Black-Scholes type [10].
To Asian options we can apply some barriers, as often done with European options, in order to reduce their price and to ward against excessive fluctuations of strike price. For this kind of geometric Asian options neither closed form solutions are available nor we have found some analysis in literature.
As example, a geometric Asian up-and-out barrier call option is an option that is extinguished when the price of the underlying asset grows up enough to breach an assigned upper barrier B before the expiry date T. Its value is modeled by the differential boundary value problem:

$$\frac{\partial V}{\partial t}+\frac{\sigma^2}{2}S^2\frac{\partial^2 V}{\partial S^2}+rS\frac{\partial V}{\partial S}+\log(S)\frac{\partial V}{\partial A}-rV=0 \quad S\in(0,B),\ A\in\mathbb{R},\ t\in[0,T) \tag{5}$$

$$V(S,A,T) \quad \text{assigned} \quad S\in(0,B),\ A\in\mathbb{R} \tag{6}$$

$$V(B,A,t)=0 \quad A\in\mathbb{R},\ t\in[0,T) \tag{7}$$

$$\text{asymptotic conditions of vanilla option} \quad \{(S,A): S=0\vee A\to-\infty\vee A\to+\infty\}. \tag{8}$$

The method, that we will illustrate in the following section for the solution of (5)-(8), is rather flexible; it therefore can be easily extended also to Asian call options with other types of barrier, that widen or contract (moving barriers), and to put options, too.

3 The SABO approach

SABO is the acronym of **S**emi-**A**nalytical method for the pricing of **B**arrier **O**ptions and substantially it is the application of Boundary Element Method (BEM) to barrier option problems. The method is based on the below listed steps.

 ISBN: 978-84-617-8694-7

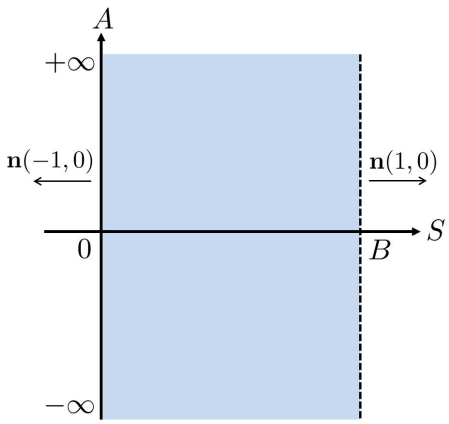


Figure 1: $(S, A) \in \Omega := (0, B) \times \mathbb{R}$, spatial domain of the differential problem (5)-(8) modeling an Asian option with up-and-out barrier.

3.1 The integral representation formula in the domain of the differential problem

Let us start from the differential problem (5) for an up-and-out barrier call option: the domain of investigation for $V(S, A, t)$ is now $\Omega \times [0, T)$, defining $\Omega := (0, B) \times \mathbb{R}$, as represented in Fig. 1.
Recalling some theory in [3], we proved that a new integral formulation holds: it is obtained adding one more term to the basic integral formulation (2) restricted to the actual domain, i.e.

$$\begin{aligned} V(S,A,t) &= \int_{-\infty}^{+\infty}\int_{0}^{B} V(\widetilde{S},\widetilde{A},T)G(S,A,t;\widetilde{S},\widetilde{A},T)d\widetilde{S}\,d\widetilde{A} \\ &+ \int_{t}^{T}\int_{-\infty}^{+\infty}\frac{\sigma^2}{2}B^2\frac{\partial V}{\partial \widetilde{S}}(B,\widetilde{A},\widetilde{t})G(S,A,t;B,\widetilde{A},\widetilde{t})d\widetilde{A}\,d\widetilde{t}\,. \end{aligned} \tag{9}$$

3.2 The boundary integral equation (BIE)

In the integral formula (9), $\frac{\partial V}{\partial \widetilde{S}}(B, \widetilde{A}, \widetilde{t})$ is unknown. If we succeed in computing it, formula (9) gives us the solution of problem (5)-(8) over the whole domain $\Omega \times [0, T)$.
With this purpose, we take the limit for $S \to B$ in (9) and, using boundary condition (7), we obtain

$$\begin{aligned} 0 = V(B,A,t) &= \int_{-\infty}^{+\infty}\int_{0}^{B} V(\widetilde{S},\widetilde{A},T)G(B,A,t;\widetilde{S},\widetilde{A},T)d\widetilde{S}\,d\widetilde{A} \\ &+ \int_{t}^{T}\int_{-\infty}^{+\infty}\frac{\sigma^2}{2}B^2\frac{\partial V}{\partial \widetilde{S}}(B,\widetilde{A},\widetilde{t})G(B,A,t;B,\widetilde{A},\widetilde{t})d\widetilde{A}\,d\widetilde{t}\,. \end{aligned} \tag{10}$$

 ISBN: 978-84-617-8694-7

in the sole unknown $\frac{\partial V}{\partial \widetilde{S}}(B, \widetilde{A}, \widetilde{t})$.
The idea implemented by SABO is to approximate $\frac{\partial V}{\partial \widetilde{S}}(B, \widetilde{A}, \widetilde{t})$ numerically solving (10) and then, inserting it in the representation formula (9), to recover the solution V at every desired point of the domain Ω and at every desired time instant.

3.3 The numerical approximation of the BIE solution

The approximation of the BIE unknown $\frac{\partial V}{\partial \widetilde{S}}(B, \widetilde{A}, \widetilde{t})$ is found by collocation method as in [6] and it is structured as follows:
- introduction of a uniform decomposition in the time interval $[0, T]$

$$\Delta t := \frac{T}{N_t}, \quad N_t \in \mathbb{N}^+, \qquad t_k := k\Delta t, \quad k = 0, \ldots, N_t$$

and time representation of the BIE unknown by piecewise constant basis functions

$$\varphi_k(\widetilde{t}) := H[\widetilde{t} - t_{k-1}] - H[\widetilde{t} - t_k], \quad k = 1, \ldots, N_t\,;$$

- introduction of a uniform decomposition in the unbounded A-domain $\equiv \mathbb{R}$ suitably truncated by $[A_{\min}, A_{\max}]$

$$\Delta A := \frac{A_{\max} - A_{\min}}{N_A}, \quad N_A \in \mathbb{N}^+, \qquad A_h := A_{\min} + h\Delta A, \quad h = 0, \ldots, N_A$$

and representation of the BIE unknown in the independent variable A by piecewise constant basis functions

$$\psi_h(\widetilde{A}) := H[\widetilde{A} - A_{h-1}] - H[\widetilde{A} - A_h], \quad h = 1, \ldots, N_A\,;$$

- approximation of the BIE (10) unknown as

$$\frac{\partial V}{\partial \widetilde{S}}(B, \widetilde{A}, \widetilde{t}) \approx \sum_{k=1}^{N_t} \sum_{h=1}^{N_A} \alpha_h^{(k)} \psi_h(\widetilde{A}) \varphi_k(\widetilde{t})\,; \tag{11}$$

- definition of the collocation points: as usual when considering piecewise constant trial functions, they are the centers of intervals $[A_{i-1}, A_i]$ and $[t_{j-1}, t_j]$, i.e.

$$\overline{A}_i = \frac{A_i + A_{i-1}}{2}, \quad i = 1, \ldots, N_A\,; \qquad \overline{t}_j = \frac{t_j + t_{j-1}}{2}, \quad j = 1, \ldots, N_t\,;$$

- evaluation of (10) at the collocation points $(\overline{A}_i, \overline{t}_j)$ building a linear system of $N_A \times N_t$ equations:

 ISBN: 978-84-617-8694-7

for $i = 1, \dots, N_A$, $j = 1, \dots, N_t$

$$\begin{aligned} 0 &= \int_{-\infty}^{+\infty} \int_0^B V(\widetilde{S}, \widetilde{A}, T) G(B, \overline{A}_i, \bar{t}_j; \widetilde{S}, \widetilde{A}, T)\, d\widetilde{S}\, d\widetilde{A} \\ &+ \int_{\bar{t}_j}^T \int_{-\infty}^{+\infty} \frac{\sigma^2}{2} B^2 \sum_{k=1}^{N_t} \sum_{h=1}^{N_A} \alpha_h^{(k)} \psi_h(\widetilde{A}) \varphi_k(\widetilde{t}) G(B, \overline{A}_i, \bar{t}_j; B, \widetilde{A}, \widetilde{t})\, d\widetilde{A}\, d\widetilde{t}\,. \end{aligned}$$

- resolution of the linear system

$$\mathcal{A}\alpha = \mathcal{F} \tag{12}$$

whose unknowns are the coefficients of linear representation in (11)

$$\alpha = (\alpha^{(k)}|_{k=1,\dots,N_t}) = \big((\alpha_h^{(k)}|_{h=1,\dots,N_A})|_{k=1,\dots,N_t}\big).$$

The rhs entries are:

for $i = 1, \dots, N_A$, $j = 1, \dots, N_t$

$$\mathcal{F}_i^{(j)} = -\int_{-\infty}^{+\infty} \int_0^B V(\widetilde{S}, \widetilde{A}, T) G(B, \overline{A}_i, \bar{t}_j; \widetilde{S}, \widetilde{A}, T)\, d\widetilde{S}\, d\widetilde{A}\,. \tag{13}$$

The matrix entries are:

for $i, h = 1, \dots, N_A$, $j, k = 1, \dots, N_{\Delta t}$

$$\begin{aligned} \mathcal{A}_{ih}^{(jk)} &= \frac{\sigma^2}{2} B^2 \int_{\bar{t}_j}^T \int_{-\infty}^{+\infty} \psi_h(\widetilde{A}) \varphi_k(\widetilde{t}) G(B, \overline{A}_i, \bar{t}_j; B, \widetilde{A}, \widetilde{t})\, d\widetilde{A}\, d\widetilde{t} \\ &= \frac{\sigma^2}{2} B^2 H[t_k - \bar{t}_j] \int_{\max(t_{k-1}, \bar{t}_j)}^{t_k} \int_{A_{h-1}}^{A_h} G(B, \overline{A}_i, \bar{t}_j; B, \widetilde{A}, \widetilde{t})\, d\widetilde{A}\, d\widetilde{t} \\ &= \frac{\sigma^2}{2} B^2 H[t_k - \bar{t}_j] \int_{\max(t_{k-1}, \bar{t}_j)}^{t_k} \int_{A_{h-1}}^{A_h} \frac{\sqrt{3}}{\pi \sigma^2 (\widetilde{t} - \bar{t}_j)^2 B} \\ &\quad \exp\left\{ -\frac{6\big(\overline{A}_i - \widetilde{A} + (\widetilde{t} - \bar{t}_j)\log(B)\big)^2}{\sigma^2 (\widetilde{t} - \bar{t}_j)^3} - \left(\frac{2r + \sigma^2}{2\sqrt{2}\sigma}\right)^2 (\widetilde{t} - \bar{t}_j) \right\} d\widetilde{A}\, d\widetilde{t} \end{aligned}$$

and, since equation (5) has constant parameters in time, they depend only on the difference between time instants so, defining $\ell = k - j$, $\ell = 0, \dots, N_t - 1$ and performing the change of variable $\widetilde{t} = \Delta t(\tau + k - 1)$, we get $\widetilde{t} - \bar{t}_j = \Delta t(\tau + k - j - 1/2) = \Delta t(\tau + \ell - 1/2)$ and

 ISBN: 978-84-617-8694-7

therefore

$$\begin{aligned}\mathcal{A}_{ih}^{(jk)} &= \frac{\sigma^2}{2}B^2\int_{\frac{1}{2}-\frac{1}{2}H[\ell]}^{1}\int_{A_{h-1}}^{A_h}\frac{\sqrt{3}}{\pi\sigma^2\Delta t(\tau+\ell-1/2)^2B}\\ &\quad\exp\left\{-\frac{6\big(\overline{A}_i-\widetilde{A}+\Delta t(\tau+\ell-1/2)\log(S)\big)^2}{\sigma^2\Delta t^3(\tau+\ell-1/2)^3}-\left(\frac{2r+\sigma^2}{2\sqrt{2}\sigma}\right)^2\Delta t(\tau+\ell-1/2)\right\}d\widetilde{A}\,d\tau\\ &=: \mathcal{A}_{ih}^{(\ell)}\,.\end{aligned} \tag{14}$$

Observe that, by consequence, $\mathcal{A}$ is a block upper triangular matrix with Toeplitz structure: the upper triangularity is due to the fact that the fundamental solution is defined by (3) only for $\widetilde{t}>t$ implying that the matrix entries are non trivial only for $k\geq j$; the Toeplitz structure is due to the dependence on time differences. Therefore it holds:

$$\mathcal{A}=\begin{bmatrix} A^{(0)} & A^{(1)} & A^{(2)} & \cdots & A^{(N_t-1)}\\ 0 & A_{(0)} & A^{(1)} & \cdots & A^{(N_t-2)}\\ 0 & 0 & A_{(0)} & \ddots & \vdots\\ \vdots & \vdots & \ddots & \ddots & A_{(1)}\\ 0 & 0 & \cdots & 0 & A_{(0)}\end{bmatrix}. \tag{15}$$

This allows to compute only last blocks column and to solve the linear system by block backward substitution.

3.4 The numerical approximation of option price

Once system (12) is solved, the knowledge of α, that determines the approximation of BIE solution, implies the possibility of computing $V(S,A,t)$ at any point $(S,A,t)\in\Omega\times[0,T)$, introducing (11) in the integral representation formula (9):

$$\begin{aligned}V(S,A,t) &\approx \int_{-\infty}^{+\infty}\int_0^B V(\widetilde{S},\widetilde{A},T)G(S,A,t;\widetilde{S},\widetilde{A},T)d\widetilde{S}\,d\widetilde{A}\\ &+ \frac{\sigma^2}{2}B^2\sum_{k=\mathrm{floor}[\frac{t}{\Delta t}]+1}^{N_t}\sum_{h=1}^{N_A}\alpha_h^{(k)}\int_{\max(t,t_{k-1})}^{t_k}\int_{A_{h-1}}^{A_h}G(S,A,t;B,\widetilde{A},\widetilde{t})d\widetilde{A}\,d\widetilde{t}\,,\end{aligned} \tag{16}$$

where floor$[\cdot]$ is the function that rounds its argument to the nearest integers towards minus infinity.
The main advantage of this method is that we can avoid to evaluate the solution over a grid, considering only the evaluation at the points of interest.

 ISBN: 978-84-617-8694-7

References

[1] C. Atkinson and S. Kazantzaki, *Double knock-out Asian barrier options which widen or contract as they approach maturity*, Quant. Finance **9**(3) (2009) 329–340.

[2] E. Barucci and S. Polidoro and V. Vespri, *Some results on partial differential equations and Asian options*, Math. Models Methods Appl. Sci. **11**(3) (2001) 475–497.

[3] A. Friedman, *Partial Differential Equations of Parabolic type*, Englewood Cliffs, N.Y.: Prentice-Hall Inc, 1964.

[4] C. Guardasoni, *Semi-Analytical method for the pricing of Barrier Options in case of time-dependent parameters (with Matlab codes)*, Commun. Appl. Ind. Math., submitted.

[5] C. Guardasoni and S. Sanfelici, *A Boundary Element approach to barrier option pricing in Black-Scholes framework*, Int. J. Comput. Math. **93**(4) (2016) 696–722.

[6] C. Guardasoni and S. Sanfelici, *Fast Numerical Pricing of Barrier Options under Stochastic Volatility and Jumps*, SIAM J. Appl. Math. **76**(1) (2016) 27–57.

[7] A. G. Z. Kemna and A. C. F. Vorst, *A pricing method for options based on average asset values*, J. Bank. Financ. **14**(1) (1990) 113–129.

[8] S. Polidoro, *Uniqueness and representation theorems for solutions of Kolmogorov-Fokker-Planck equations*, Rend. Mat. Appl. VII **15**(4) (1995) 535–560.

[9] R. Zvan and K. R. Vetzal and P. A. Forsyth, *PDE methods for pricing barrier options*, J. Econom. Dynam. Control **24**(11-12) (2000) 1563–1590.

[10] J. E. Zhang, *Theory of Continuously-sampled Asian Option Pricing*, Working Paper of East Asian Bureau of Economic Research (2004).

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

A test for the homogeneity of confusion matrices

M. V. Alba-Fernández[1] and F. J. Ariza-López[2]

[1] *Department of Statistics and O.R., University of Jaén, Spain*

[2] *Department of Cartographic engineering, Geodesy and Photogrammetry, University of Jaén, Spain.*

emails: mvalba@ujaen.es, fjariza@ujaen.es

Abstract

The confusion matrix is the standard way to report on the thematic accuracy of geographic information data. Two widely adopted indices for thematic accuracy controls upon error matrix are the overall accuracy and the Kappa coefficient. Both indices are global and do not allow for a category-wise control. Provided that a multinomial sampling, this work proposes a new method for testing the homogeneity of two independent thematic classifications which is based on whole error matrices. Specifically, a test function is proposed that uses as a test statistic the discrete Hellinger distance. A simulation experiment is carried out to evaluate the goodness of the proposal and an application to a real data set is included.

Key words: thematic accuracy, confusion matrix, hellinger distance

1 Introduction

Geographic Information (GI) supports decision making in several fields as climate change, crop forecasting, forest fires, national defense, civil protection or spatial planning. The quality of the GI is essential to ensure that decisions based on it are technically the best. There are different components to describe this quality. One of them is the thematic quality, as is established by the international standard ISO 19157. This thematic quality is usually quantitatively assessed by means of the so called confusion matrix or error matrix (i.e. when classification correctness has to be assessed). The confusion matrix is a contingency table established by the cross-reference of a set of categories. The true categories or reference data are located in columns, whereas in rows are located the categories for the classification

 ISBN: 978-84-617-8694-7

	Reference Data			
Observed Data	Category 1	Category 2	...	Category M
Category 1	n_{11}	n_{12}	...	n_{1M}
Category 2	n_{11}	n_{12}	...	n_{1M}
...	...	...	...	...
Category M	n_{M1}	n_{M2}	...	n_{MM}

Table 1: Contingency table associated with a confusion matrix.

process. So, the elements in the diagonal are correctly classified items, and the off diagonal elements contain the number of confusions, the errors due to omissions or commissions.

The control of a confusion matrix is usually carried out by using two widely adopted indices like the overall accuracy (OA) and the Kappa coefficient (κ). The first is the ratio between the account of elements that are correctly classified and the total amount of elements in the matrix. The Kappa coefficient is a measure based on the difference between the agreement indicated by the values of the main diagonal (OA) and the chance agreement estimated by the marginal values. If M represents the number of classes under consideration and $k = M \times M$ the number of cells in the error matrix, assuming a multinomial sampling model, n_{ij} represents the number of elements in the class j of the reference data that are classified in class i of the observed data and n stands for the total number of elements classified. Table 1 shows a general way to present a confusion matrix.

The kappa coefficient (Cohen, [3]) is defined as

$$\hat{\kappa} = \frac{OA - P_c}{1 - P_c}$$

where the overall accuracy is $OA = \frac{1}{n}\sum_{i=1}^{M} n_{ii}$ and $P_c = \sum_{i=1}^{M} p_{i+}p_{+i}$, represents the proportion of agreement expected by chance agreement; $p_{i+} = \frac{n_{i+}}{n}$, $p_{+i} = \frac{n_{+i}}{n}$ being the proportions in the ith row and in the ith column, respectively. Banerjee et al. [2] gave an estimate of the variance of $\hat{\kappa}$ by means of

$$\hat{V}(\hat{\kappa}) = \frac{P_c + P_c^2 - \sum_{i=}^{M} p_{+i}p_{i+}(p_{i+} + p_{+i})}{n(1 - P_c)^2}. \tag{1}$$

The Kappa coefficient is asymptotically normally distributed and this fact provides a means for testing the significance of the Kappa coefficient for a single confusion matrix in order to determine if the agreement between the observed data and the reference data is significantly greater than 0. However, some authors have criticized the use of OA and κ because they only use the account of elements that are correctly classified, the marginal values, and the total number of elements in the matrix (Gwet, [5], Stehman, [8], among many others), that is, they do not incorporate the off-diagonal information of the error matrix.

 ISBN: 978-84-617-8694-7

In the same way, it is possible to determine if two independent Kappa values, associated with two confusion matrices, are significantly different. So, it can be compare two analysts, two strategies, the same analyst over time, etc.

The null hypothesis $H_0 : \kappa_1 - \kappa_2 = 0$ can be rejected if $Z > Z_\alpha$, with

$$Z = \frac{|\hat{\kappa}_1 - \hat{\kappa}_2|}{\sqrt{\hat{\sigma}_1^2 + \hat{\sigma}_2^2}}$$

where $\hat{\kappa}_1$, $\hat{\kappa}_2$ are the estimate of kappa coefficient for both confusion matrices, $\hat{\sigma}_1^2$, $\hat{\sigma}_2^2$ their estimated variances, and Z_α is the $1-\alpha$ percentile of $N(0,1)$.

For example, for the confusion matrices A and B,

Matrix A	
50	0
20	30

Matrix B	
50	10
10	30

However, the use of the Kappa coefficient for this purpose can lead us to incoherent results. In this case, $Z = 0.12$ and at signification level of 5%, the null hypothesis can not be rejected, the conclusion should be that, globally, both confusion matrices reveal the same information although it is easily to observe that this assertion is not true for the matrices. Our proposal is to take advantage of the underlying sampling model and to deal with a confusion matrix as a multinomial distribution. This way, by testing the homogeneity of the associated confusion matrices, it is possible to determine if two photointerpretation works, two different classification works, or a test work and a product reach the same conclusion.

2 The homogeneity test

Let X, Y be two independent random vectors whose values have been grouped into $k = M \times M$ classes $C_1, C_2, \ldots, C_k$, or equivalently, taking values in $\Upsilon = (1, 2, \ldots, k)$ with probabilities $P = (p_1, p_2, \ldots, p_k)$ and $Q = (q_1, q_2, \ldots, q_k)$, respectively. Our task is to test the following null hypothesis

$$H_0 : P = Q. \tag{2}$$

Several test statistics can be chosen for testing H_0, but not all of them can be applied when empty cells are present, as in the confusion matrices. Here, we are going to consider a test statistic based on the Hellinger distance. So, let $(X_1, \ldots, X_n)$ and $(Y_1, \ldots, Y_m)$ be two independent random samples from X and Y, with sizes n and m, respectively. For testing (2), we consider the following test function

$$\Psi = \begin{cases} 1, & \text{if } T_{n,m} \geq t_{n,m,\alpha}, \\ 0, & \text{otherwise,} \end{cases}$$

 ISBN: 978-84-617-8694-7

where $T_n = 2(n+m)\sum_{i=1}^{k}(\sqrt{\hat{p}_i} - \sqrt{\hat{q}_i})^2$ and $t_{n,m,\alpha}$ is the $1-\alpha$ percentile of the null distribution of $T_{n,m}$.

A reasonable test for testing H_0 should reject the null hypothesis for large values of $T_{n,m}$. To decide when to reject H_0, that is, to calculate $t_{n,m,\alpha}$ or, equivalently, to calculate the p-value of the observed value of the test statistic, we need to know the null distribution of $T_{n,m}$, which is clearly unknown, so one has to approximate it. Asymptotically, the null distribution of T_n is a chi-square variate with $k-1$ degrees of freedom. However, for small and moderate sample sizes, the behaviour of the asymptotic null distribution is rather poor (Alba et al. [1], Kim et al. [6]). To overcome this problem, we approximate the null distribution of the test statistic by means of a bootstrap estimator. It is proved the bootstrap provides a consistent estimator of the null distribution of $T_{n,m}$.

3 Simulation study

The bootstrap approximation is valid for large samples. To evaluate the goodness of the proposal for small and moderate sample sizes, a simulation experiment is carried out. All computations have been performed using programs written in the R language [7].

To study the goodness of the bootstrap approximations to the null distribution of the test statistic $T_{n,m}$, we first generated two independent samples with equal sample sizes $n = m = 50$ from the multinomial with $P_1 = (0.5, 0.2, 0, 0.3)$ ($k = 2 \times 2$) and calculated $\hat{p}$ with $B = 1000$ replications. This was repeated 5000 times and we calculated the fraction of $\hat{p}$s less than or equal to 0.05 and 0.10 (f05 and f10 in tables), which are the estimated type I error probability for $\alpha = 0.05, 0.10$. The whole experiment is repeated for $n = 100, 300, 500$ and for $P_2 = (0.3, 0, 0.15, 0, 0.15, 0, 0.1, 0.15, 0.15)$ ($k = 3 \times 3$). Table 2 shows the results obtained. According to these results, it can be highlighted that the estimated type I error probabilities are quite close to the nominal values in all the tried cases. It means the proposal behaves properly when the null hypothesis is true.

Now, the simulation experiment is completed with a power study. Now, we generated two independent samples with sizes n and m from two multinomials with P_1 and $Q_1 = (0.50, 0.16, 0.02, 0.32)$ and calculated $\hat{p}$ in the same conditions as before. Now, the fraction of $\hat{p}$s less than or equal to 0.05 and 0.10 are the estimated powers associated with the nominal values $\alpha = 0.05, 0.10$, respectively. The whole experiment is repeated for $n = 50, 100, 200, 300$ and $Q_1^{\dagger} = (0.52, 0.16, 0.02, 0.30)$. The whole experiment was repeated for P_2, $Q_2 = (0.32, 0, 0.13, 0, 0.17, 0, 0.06, 0.15, 0.17)$ and $Q_2^{\dagger} = (0.3, 0, 0.15, 0.02, 0.17, 0, 0.06, 0.13, 0.17)$. Table 3 and 4 show the results obtained.

According to the results in tables 3 and 4, the procedure is able to detect if two photointerpretation works, two different classification works, or a test work and a product reach the same conclusion.

 ISBN: 978-84-617-8694-7

Sample sizes	P_1		P_2	
$n = m$	f05	f10	f05	f10
50	0.056	0.107	0.048	0.096
100	0.052	0.106	0.050	0.100
300	0.050	0.106	0.050	0.101
500	0.051	0.101	0.050	0.099

Table 2: Estimated type I error.

Sample sizes	Q_1		$Q_1^\dagger$	
$n = m$	f05	f10	f05	f10
50	0.103	0.189	0.096	0.195
100	0.225	0.377	0.227	0.365
300	0.861	0.931	0.855	0.933
500	0.988	0.997	0.989	0.996

Table 3: Estimated powers (k=4).

Sample sizes	Q_2		$Q_2^\dagger$	
$n = m$	f05	f10	f05	f10
50	0.077	0.142	0.109	0.198
100	0.127	0.214	0.227	0.362
300	0.332	0.454	0.856	0.924
500	0.526	0.649	0.989	0.997

Table 4: Estimated powers (k=9).

 ISBN: 978-84-617-8694-7

Acknowledgements

This work has been partially supported by grant CMT2015-68276-R MINECO/FEDER,UE.

References

[1] V. Alba-Fernández, M. D. Jiménez-Gamero, *Bootstrapping divergence statistics for testing homogeneity in multinomial populations*, Mathematics and Computers in Simulation **79** (2009) 3375-3384.

[2] M. Banerjee, M. Cappozzoli, L. McSweeney, D. Sinha, *Beyond kappa: a review of interrater agreement measures*, The Canadian Journal of Statistics **27:1** (1999) 3–23.

[3] J. Cohen, *A coefficient of agreement for nominal scales*, Educational and Psychological Measurement **20: 1** (1960) 37–46.

[4] R. G. Congalton, K. Green, *Assesing the accuracy of remote sensed data. Principles and practice*, CRC Press, 2009.

[5] K. L. Gwet, *Handbook of Inter-rater reliability*, Advanced Analytics, LLC, 2014.

[6] S. H. Kim, H. Choi, S. Lee, *Estimate-based goodness-of-fit test for large sparse multinomial distributions*, Computational Statistis and Data Analysis **53** (2009) 1122–1131.

[7] R Core Team *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria, 2015. URL http://www.R-project.org/.

[8] S. V. Stehman, *Selecting and interpreting measures of thematic classification accuracy*, Remote Sensing of Environment **62: 1** (1997) 77–89.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

Approaching the Rank Aggregation Problem by Local Search-based Metaheuristics

Juan A. Aledo[1], José A. Gámez[2] and David Molina[1]

[1] *Department of Mathematics, University of Castilla-La Mancha*

[2] *Department of Computer Systems, University of Castilla-La Mancha*

emails: juanangel.aledo@uclm.es, jose.gamez@uclm.es, david.molina@uclm.es

Abstract

Encouraged by the success of applying metaheuristics algorithms to other ranking-based problems (Kemeny ranking problem and parameter estimation for Mallows distributions), in this paper we deal with the rank aggregation problem (RAP), which can be viewed as a generalization of the Kemeny problem to arbitrary rankings.

In particular, we perform a comparative study among some local-based search metaheuristics: hill climbing (HC), iterated local search (ILS), variable neighborhood search (VNS) and greedy randomized adaptive search procedure (GRASP).

We provide a complete analysis of the experimental study regarding accuracy and number of iterations required to reach the best solution.

Key words: Ranking, Rank aggregation problem, Metaheuristic algorithm, Kendall distance, Permutation, Partial ranking.

1 Introduction

In the last decades, the *Rank Aggregation Problem* (RAP) [1] has gained popularity in several fields as statistics and machine learning because of its significant applications in many real-world problems, including information retrieval and recommender systems.

Rankings represent preferences in a natural way. Given a set of items $[n] = \{1, 2, \ldots, n\}$, a ranking π is an ordering of (some of) these items. The case when the rankings are permutations (complete rankings without ties) has received a great attention in the literature (see [2] and references therein). However, many real world problems usually deal with incomplete rankings, that is, those where only p items are ranked, $2 \leq p < n$, and/or with rankings where some items are equally preferred or tied.

Given a sample of arbitrary rankings, the solution to the RAP is the *consensus* permutation, that is, the permutation which best summarizes the rankings in the sample. When all the rankings in the sample are permutations, this problem is known as the *Kemeny ranking problem* [3]. Both are NP-hard problems when the number of rankings to aggregate is greater than 3.

The RAP has been widely studied and several proposals have arisen in the last years. However, they only focus in particular cases of the problem, incomplete rankings or rankings with ties, but not both of them simultaneously. Recently, metaheuristics algorithms have been used to approach several ranking-based problems [4, 5].

In this paper we deal with problem instances of the RAP containing any kind of rankings (complete, incomplete, with and without ties). In a recent paper [6], the authors have developed a version of the greedy Borda algorithm tailored to this problem, which clearly outperforms the standard Borda method. Now, we go one step further, and with the goal of obtaining better solutions (closer to the global optimum) we propose the use of more complex search engines. Since local search-based metaheuristics have shown a good trade-off between efficiency and accuracy in related problems (e.g. the traveling salesman problem, the routing-packing problem, the vertex separation problem, the linear ordering problem, etc.), in this study we focus on this family of metaheuristics. Specifically, we perform a comparative study among the following algorithms:

- *Hill climbing (HC)* algorithm. It is a local search algorithm which iteratively tries to improve a given solution by moving at each iteration to the neighbor representing the biggest increase (decrease) in the evaluation or objective function with respect to the current solution.

- *Iterated local search (ILS)* method. It is a multi-start local search algorithm based on the HC algorithm, which tries to escape from the local optimum by perturbing it, and use the resulting configuration to seed a new HC iteration.

- *Variable neighborhood search (VNS).* It is also a multi-start local search algorithm which, instead of modifying the starting point at each iteration, changes to a different neighborhood.

- *Greedy randomized adaptive search procedure (GRASP)* It is a multi-start local search algorithm which, at each iteration, constructs a randomized informed solution and locally improves it.

Thus, the main contribution of this paper is to provide a comparative study among the selected MHs. We consider different neighborhoods and allowed resources (number of fitness evaluations). The results show the influence of the starting point and the selected neighborhood in the performance of the (local) search algorithms. Regarding the trade-off between goodness of the solution and number of evaluations, we get that the VNS is the best

 ISBN: 978-84-617-8694-7

choice under limited resources, while the GRASP performs better when more evaluations are allowed.

Acknowledgements

This work has been partially funded by FEDER funds and the Regional Government of Castilla-La Mancha through project PEII-2014-049-P.

References

[1] F. Schalekamp and A. van Zuylen, *Rank aggregation: Together we're strong*, ALENEX (2009).

[2] J. Huang, *Probabilistic reasoning and learning on permutations: Exploiting structural decompositions of the symmetric group*, Ph.D. thesis, Carnegie Mellon University, San Francisco, 2011.

[3] A. Ali and M. Meila, *Experiments with Kemey ranking: What works when?*, Mathematical Social Sciences **64** (2012) 28–40.

[4] J. A. Aledo, J. A. Gámez and D. Molina, *Tackling the rank aggregation problem with evolutionary algorithms*, Applied Mathematics and Computation **222** (2013) 632–644.

[5] J. A. Aledo, J. A. Gámez and D. Molina, *Using metaheuristic algorithms for parameter estimation in generalized Mallows models*, Applied Soft Computing **38** (2016) 308–320.

[6] J. A. Aledo, J. A. Gámez and D. Molina, *Using extension sets to aggregate partial rankings in a flexible setting*, Applied Mathematics and Computation **290** (2016) 208–223.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

Hermite finite element method for nonlinear Black-Scholes equation governing European options

Rui M.P. Almeida[1], Teófilo D. Chihaluca[1] and José C.M. Duque[1]

[1] *Center of Mathematics and Applications, University of Beira Interior, Covilhã, Portugal*

emails: ralmeida@ubi.pt, teofilo.chihaluca@ubi.pt, jduque@ubi.pt

Abstract

We develop a numerical algorithm for solving a generalized Black-Scholes partial differential equation, which arise in European option pricing considering transaction costs. The Crank-Nicolson method is considered to discretize in the temporal direction and the Hermite cubic interpolation method in the spatial direction. The efficiency and accuracy of the proposed method are tested numerically, the results confirm the theoretical behaviour of the solutios, and they are with good agreement with the exact solution.

Key words: Non-linear Black-Scholes, Finite Element Method, Crank-Nicolson, Hermite Polynomials.

1 Introduction

The valuation of options based on stochastic processes dates back to 1877, when Charles Castelli wrote the book entitled "The Theory of Option in Stocks and Shares". Two decades later, Louis Bachelier, in his dissertation "Théorie de la spéculation", presented the first analytical way of calculating the price of an option. Subsequently, in 1955, in an unpublished manuscript entitled "Brownian Motion in the Stock Market", a professor at the Massachusetts Institute of Technology (MIT), Paul Samuelson, 1970 Nobel Prize in Economics, showed that the asset price can be modeled by a stochastic process called Brownian Geometric Motion. In 1962, A. James Boness presented a dissertation entitled "Theory and Measurement of Stock Option Value", where he announced an option evaluation model that represented a great step forward from his predecessors and served as the basis for the work later developed by Black and Scholes. The Black-Scholes [5] model is a well-known

 ISBN: 978-84-617-8694-7

popular model which is used to calculate the price of European options. Since its inception in 1973 by Fischer Black and Myron Scholes, it remains one of the most preferred models and provides the basis for the theory of financial options. The linear Black-Scholes equation is given by

$$0 = V_t + \frac{1}{2}\sigma^2 S^2 V_{SS} + rSV_S - rV, \qquad S > 0, \qquad t \in]0, T[, \tag{1}$$

where V is the option value, T the expiry date, S the underlying asset price, σ the volatility and r the riskless interest rate.
Equation (1) permit the evaluation of the price of a European option under the assumptions listed below:

- the value of the financial asset underlying the option can be modeled by a geometric Brownian motion;
- there are no transaction costs associated with the management of financial asset portfolios, nor fees payable in the market;
- the market does not allow arbitrage opportunities;
- you can do *short selling*;
- there is a risk-free rate that is constant throughout the life of the option, to which it is possible to lend and borrow at that same rate any financial asset;
- the volatility of the underlying asset is known and remains constant throughout the option lifespan;
- the transaction of the financial asset is made on a continuous basis and changes in its price are also on a continuous basis;
- fractional parts of an asset can be obtained;
- the financial asset does not pay dividends during the option lifespan.

The classical Black-Scholes model is notable for its explicit closed form solution of European style options (call and put options). Many researchers have attempted to obtain the solution of the Black-Scholes Equation analytically and/or numerically, thereby adopting and using various direct and iterative methods, respectively.

J. Ankudinova and M.Ehrhardt [3], make a comparative study between models with transaction costs and the linear model. The influence of transaction costs modeled by the volatilities given by the Leland, Barles and Soner and Krakta models was calculated by the Crank-Nicolson method in time and by the finite difference method in space.

 ISBN: 978-84-617-8694-7

In [11] a study of the linear Black-Scholes equation for European and American call options was made. The finite element method was used in the space dimension and explicit Euler method in the time discretization. The analytical solution was obtained by applying the Fourier transformation, while the numerical solution was obtained by the finite element method. The results were them compared using the finite difference method.

Almeida et al., in ([2],[1],[6]), established convergence, properties and error bounds for the fully discrete solutions of a class of nonlinear equations of reaction-diffusion nonlocal type, using a linearised Crank-Nicolson-Galerkin finite element method with polynomial approximations of arbitrary degree.

The remainder of this paper is organized as follows. In Section 2 the problem is described considering some transaction cost models for European options. In Section 3, we define the problem as a general non-linear partial differential equation in non divergent form. In Section 4, we construct the discretization in spatial direction with the Hermite interpolation method in a uniform mesh and, in Section 5, we discretize in temporal direction with the Crank-Nicolson method. In Section 6, we obtain and compare the approximate numerical solutions. Finally, in Section 7, we draw some conclusions.

2 Non-linear Black-Scholes Model

The Black-Scholes model requires a portfolio adjustment in order to protect a "hedge" risk free. In the presence of transaction cost, this adjustment is likely to be more expensive, since an infinite number of transactions is required [12]. But the hedger needs to find the balance between the transaction costs that are needed to rebalance the portfolio and the implicit costs of hedging errors. As a result of this "imperfect" coverage, the option can be overly underestimated, where the risk-free profit obtained by the arbitrator is offset by the transaction cost so that there is no single equilibrium price plus a viable price range. It has been demonstrated that in a transaction market there is no replicator portfolio for a European type call option and the portfolio is required to dominate rather than replicate the option value (see [4]). Soner, Shreve and Cvitanič have proved in [10] that the minimum coverage portfolio of a financial option is trivial, so efforts have been made to ease the condition coverage criterion to better replicate *pay-off* of derivative securities. Because of the presence of transaction costs (see [4], [5], [9]) the classical model results in a strongly or wholly non-linear and possibly degenerate parabolic type diffusion equation where the volatility σ may depend on the time t, the price S or on other derivatives of the option price V. In this work, we study the non-linear Black-Scholes equation with some transaction cost models for European options, with σ a non-constant modified volatility function

$$\tilde{\sigma}^2 := \tilde{\sigma}^2(t, S, V_S, V_{SS}).$$

 ISBN: 978-84-617-8694-7

In this way equation (1) becomes the following non-linear Black-Scholes equation, which we will consider for European options

$$0 = V_t + \frac{1}{2}\tilde{\sigma}^2(t, S, V_S, V_{SS})S^2V_{SS} + rSV_S - rV, \qquad S > 0, \qquad t \in (0, T). \tag{2}$$

A European call option allows buying an asset of value S for a value of K on the maturity date T, while a European put allows selling an asset of value S for a value K at the maturity date T. Since the option can only be exercised at maturity and, in order to avoid arbitrage, we complement equation (2) with the following conditions:

European call option:

$$V(S, T) = \max\{S - K, 0\}, \qquad \text{when } S \geq 0 \tag{3}$$

$$\lim_{S\to\infty} \frac{V(S, t)}{S - Ke^{-r(T-t)}} = 1, \qquad \text{for } t \in [0, T] \tag{4}$$

$$V(0, t) = 0, \qquad \text{for } t \in [0, T] \tag{5}$$

$$\lim_{S\to\infty} V_S(S, t) = 1, \qquad \text{for } t \in [0, T] \tag{6}$$

European put option:

$$V(S, T) = \max\{K - S, 0\}, \qquad \text{when } S \geq 0 \tag{7}$$

$$V(0, t) = Ke^{-r(T-t)}, \qquad \text{for } t \in [0, T] \tag{8}$$

$$\lim_{S\to\infty} V(S, t) = 0, \qquad \text{for } t \in [0, T] \tag{9}$$

$$\lim_{S\to\infty} V_S(S, t) = 0, \qquad \text{for } t \in [0, T] \tag{10}$$

2.1 Leland's model

In [9], Leland deduces that the option price is the solution of the non-linear Black-Scholes equation (2), with the modified volatility given by

$$\tilde{\sigma}^2 = \sigma^2\Big(1 + Le \times sign(V_{SS})\Big). \tag{11}$$

In equation (11) Le is Leland's number, which is given by

$$Le = \sqrt{\frac{2}{\pi}}\left(\frac{k}{\sigma\sqrt{\delta t}}\right) \tag{12}$$

 ISBN: 978-84-617-8694-7

where δt is the interval between two successive revisions of the portfolio, k is the round trip transaction cost per transacted monetary unit and σ represents the historical volatility. By a different process, Boyle and Vorst deduced a similar modified volatility, given by

$$\tilde{\sigma}^2 = \sigma^2 \left(1 + Le\sqrt{\frac{\pi}{2}} sign(V_{SS})\right). \tag{13}$$

However, δt in Leland's definition (12) represents the interval between two successive portfolio reconstructions and not the transaction frequency as in (13) (see [4]).

2.2 Barles and Soner's model

Barles and Soner obtained the most complex model. Following the Hedges and Neuberger [7] utility function approach, they propose, to simplify the calculations, the volatility model

$$\tilde{\sigma}^2 = \sigma^2 \left(1 + e^{r(T-t)} a^2 S^2 V_{SS}\right), \tag{14}$$

where σ is the historical volatility and $a = \frac{k}{\sqrt{\varepsilon}}$. Barles and Soner in [4] proved the existence of a viscosity solution for the European option with the volatility given by (14). Their numerical results indicate an economically significant price difference between the standard Black-Scholes model and the non-linear model with transaction costs.

2.3 Kratka's model

The model proposed by Krakta in [8] minimizes the sum of the rate of the transaction costs and the rate of the risk from an unprotected portfolio. In this way, the portfolio is well protected with the Risk Adjusted Pricing Methodoly (RAPM) and the modified volatility is given by

$$\tilde{\sigma}^2 = \sigma^2 \left(1 + 3(\frac{C^2 M}{2\pi} S V_{SS})^{\frac{1}{3}}\right), \tag{15}$$

where $M \geq 0$ is the measure of the transaction cost and $C \geq 0$ the risk premium.

It should be noted that the non-linear transaction cost models described above are all consistent with the linear model if the additional parameters for the transaction cost are zero.

3 Non-linear General Equation

In this work, we study equations of the form:

$$u_t = c_0 u_{xx} + c_1 u_x + c_2 u + f, \qquad \text{with} \qquad a < x < b, \qquad 0 < t < T, \tag{16}$$

 ISBN: 978-84-617-8694-7

under the initial and boundary conditions

$$u(x,0) = u_0(x), \qquad a < x < b, \tag{17}$$

$$\begin{cases} u(a,t) = g_1(t) \\ u(b,t) = g_2(t) \end{cases} \quad \text{and} \quad \begin{cases} u_x(a,t) = g_3(t) \\ u_x(b,t) = g_4(t) \end{cases} \quad 0 < t < T, \tag{18}$$

where $c_0 = c_0(x,t,u,u_x,u_{xx})$, $c_1 = c_1(x,t,u,u_x)$, $c_2 = c_2(x,t,u)$, $f = f(x,t)$, $g_1(t)$, $g_2(t)$,$g_3(t)$,$g_4(t)$ e $u_0(x)$ are known real bounded functions.
Note that (16)-(18) is a general model which includes the problem under study. The transformation $u(x,t) = V(S, T-t)$ transforms (2) into (16) with $c_0 = \frac{1}{2}\tilde{\sigma}^2 x^2$, $c_1 = rx$, $c_2 = -r$ and $f = 0$, and the initial condition becomes

$$u(x,0) = \max\{K - x, 0\} \quad \text{or} \quad u(x,0) = \max\{x - K, 0\}$$

by (3) and (7).
For a call option, condition (5) is satisfied considering $g_1(t) = 0$. For b sufficient large conditions (4) and (6) can be approximated by $g_2(t) = b - Ke^{-rt}$ and $g_4(t) = 1$.
Since we need another condition, motivated by the behavior of the solution for the linear equation, we consider $g_3(t) = 0$.
For a put option, (8), (9) and (10) implies that $g_1(t) = Ke^{-rt}$, $g_2(t) = 0$, $g_3(t) = 1$ and $g_4(t) = 0$.

Let w be a test function. Multiplying (16) by w and integrating in]a,b[, we obtain

$$\int_a^b u_t w\, dx - \int_a^b c_0 u_{xx} w\, dx - \int_a^b c_1 u_x w\, dx - \int_a^b c_2 u w\, dx = \int_a^b f w\, dx. \tag{19}$$

Since c_0 depends on u_{xx}, the integration by parts is useless. For relation (19) to make sense, u, u_t, u_x and $u_{xx} \in L_2(a,b)$, that is, u must be in $C^1(a,b)$, for $t \in]0,T]$. According to the conditions in (18), we choose the test function space to be

$$V_0 = \{w, w_x, w_{xx} \in L_2(a,b) : w(a) = w(b) = w_x(a) = w_x(b) = 0\},$$

and for the space solution we consider

$$V = \{u, u_t, u_x, u_{xx} \in L_2(a,b) : v(a,t) = g_1(t), v(b,t) = g_2(t), v_x(a,t) = g_3(t), v_x(b,t) = g_4(t) \\ \forall t \in [0,T]\},$$

 ISBN: 978-84-617-8694-7

4 Discretization in Space

Considering the discretization $a = x_0 < x_1 < \cdots < x_{m+1} = b$ of $[a, b]$, with spacing h and requiring continuity in C^1, we define for each node x_i, two Hermite interpolation polynomials, $\varphi_i(x)$ and $\psi_i(x)$. The Hermite interpolation polynomials have support $[x_{i-1}, x_{i+1}]$, and are defined by

$$\varphi_i(x) = \begin{cases} 2\left(\frac{x-x_i}{h}\right)^3 - 3\left(\frac{x-x_i}{h}\right)^2 + 1, & x \in [x_{i-1}, x_i], \\ \frac{(x-x_i)^3}{h^2} - \frac{2(x-x_i)^2}{h} + (x - x_i), & x \in [x_i, x_{i+1}] \end{cases} \tag{20}$$

and

$$\psi_i(x) = \begin{cases} -2\left(\frac{x-x_i}{h}\right)^3 + 3\left(\frac{x-x_i}{h}\right)^2 + 1, & x \in [x_{i-1}, x_i], \\ \frac{(x-x_i)^3}{h^2} - \frac{(x-x_i)^2}{h}, & x \in [x_i, x_{i+1}] \end{cases} \tag{21}$$

The Hermite cubic polynomials satisfy the following interpolation properties:

$$\varphi_j(x_i) = \begin{cases} 1 & , i = j \\ 0 & , i \neq j \end{cases}, \quad \varphi_j'(x_i) = 0, \tag{22}$$

$$\psi_j(x_i) = 0, \quad \psi_j'(x_i) = \begin{cases} 1 & , i = j \\ 0 & , i \neq j \end{cases}, i, j = 0, \ldots, m+1 \tag{23}$$

and hence they satisfy the required continuity conditions.
Let $T > 0$ and $\mathcal{H} = \langle \varphi_0, \psi_0, \varphi_1, \psi_1, \cdots, \varphi_{m+1}, \psi_{m+1} \rangle$, the vector subspace generated by the $2m + 2$ elements of the Hermite base. Let us consider the test function in

$$V_{0m} = \{w_m(x) \in \mathcal{H} : w_m(a) = w_m(b) = (w_m)_x(a) = (w_m)_x(b) = 0\} \subset V_0$$

and the approximate solution in

$$\begin{aligned} V_m = \{u_m(x,t) \in \mathcal{H} : u_m(a,t) = g_1(t),\ u_m(b,t) = g_2(t), \\ (u_m)_x(a,t) = g_3(t),\ (u_m)_x(b,t) = g_4(t), \quad \forall t \in [0,t]\} \subset V. \end{aligned}$$

A function $u_m \in V_m$ is said to be an approximate solution of (19) if, for each $t \in]0, T]$, it satisfies

$$\begin{aligned} &\int_a^b (u_m)_t w_m \, dx - \int_a^b c_0 (u_m)_{xx} w_m \, dx - \int_a^b c_1 (u_m)_x w_m \, dx \\ &- \int_a^b c_2 (u_m) w_m \, dx = \int_a^b f w_m \, dx, \qquad \forall w_m \in V_{0m}. \end{aligned} \tag{24}$$

 ISBN: 978-84-617-8694-7

Any function $w_m \in V_{0m}$ can be written as

$$w_m(x) = \sum_{i=1}^{m} W_i \varphi_i(x) + Z_i \psi_i(x), \tag{25}$$

and any function $u_m \in V_m$ can be written as

$$\begin{aligned} u_m(x,t) \quad &= \varphi_0(x) g_1(t) + \psi_0(x) g_3(t) + \sum_{i=1}^{m} \varphi_i(x) U_i(t) + \psi_i(x) V_i(t) \\ &+ \varphi_{m+1}(x) g_2(t) + \psi_{m+1}(x) g_4(t), \end{aligned} \tag{26}$$

Substituting (25) and (26) in equation (24) and simplifying the expressions, we obtain a system of ordinary differential equations, which can be written in matrix form:

$$MU(t)' - A(U(t))U(t) - B(U(t))U(t) - C(U(t))U(t) = F(t) + D(t), \tag{27}$$

with the unknown

$$U(t) = [U_1, ..., U_m, V_1, ..., V_m]^T.$$

In general, the solution $U(t)$ is not explicitly known for all $t \geq 0$, so it is necessary to use a numerical method to obtain an approximate solution.

5 Discretization in Time

Let us now consider the partition $0 = t_0 < t_1 < \cdots < t_N = T$, with step δ, of $[0, T]$. By the Cranck-Nicolson method, evaluating (27) in $t_{n+\frac{1}{2}} = \frac{t_n + t_{n+1}}{2}$ and using the approximations

$$U'(t_{n+\frac{1}{2}}) \approx \frac{U(t_{n+1}) - U(t_n)}{\delta} = \frac{U_{n+1} - U_n}{\delta} \tag{28}$$

and

$$U(t_{n+\frac{1}{2}}) \approx \frac{U(t_{n+1}) + U(t_n)}{2} = \frac{U_{n+1} + U_n}{2}, \tag{29}$$

we obtain

$$\begin{aligned} &(2M + \delta\,(A_{n+1} + B_{n+1} + C_{n+1})U_{n+1} = (2M - \delta\,(A_{n+1} + B_{n+1} + C_{n+1}))U_n \\ &+2\delta F_{n+1/2} - 2\delta D_{n+1/2}, \quad n = 0, 1, \cdots, N-1. \end{aligned} \tag{30}$$

Since the function $f(x,t)$ is known, $F_{n+1/2} = F(x, t_{n+1/2})$ is also known for all n. Then, for each $n = 0, 1, \cdots, N-1$, we have to solve the nonlinear system (30) and thus determine the solution $u(x,t)$ at discrete times $t_n = n\delta$.

 ISBN: 978-84-617-8694-7

For the system of algebraic equations (30), we propose the fixed point scheme:

$$\begin{aligned}&\left[2M+\delta\left(A_{n+1}^{(k)}+B_{n+1}^{(k)}+C_{n+1}^{(k)}\right)\right]U_{n+1}^{(k+1)}=(2MU_n\\&-\delta\left(A_{n+1}^{(k)}+B_{n+1}^{(k)}+C_{n+1}^{(k)}\right))U_n+2\delta F_{n+1/2}-2\delta D_{n+1/2}\end{aligned}\tag{31}$$

$$U_{n+1}^{(0)}=U_n\quad and\quad n=1,2,\ldots,N\quad k=1,2,\ldots.$$

6 Numerical results

In this section, we present the results of a Matlab implementation of the theory. First we validate the code by simulating the linear equation and calculating the error. Then we compare the solutions of the nonlinear equation obtained with the different modified volatilities presented.

Example 1: The analytical solution of the linear Black-Scholes equation in (1), where r and σ are constant and satisfying conditions (3), is given by the well-known formula [5]

$$V(S,t)=SN(d_1)-Ke^{-r(T-t)}N(d_2)\tag{32}$$

where $N(x)=(1/2\pi)\int_{-\infty}^{x}e^{\frac{-y^2}{2}}dy$, $x\in\mathbb{R}$ is the cumulative distribution function of $N(0,1)$, $d_1=\{\log(S/K)+\left(r+\sigma^2/2\right)(t-T)\}/\left(\sigma\sqrt{t-T}\right)$ and $d_2=d_1-\sigma\sqrt{(t-T)}$. Knowing the exact explicit solution permitted us to calculate the exact error of the approximations. We simulated equation (16) using the parameters $T=1$, $r=0.1$, $\sigma=0.2$, $a=0.4$, $b=1$, $K=0.4$, $T=1$, $g_1(t)=0$, $g_2(t)=1-0.4e^{rt}$, $g_3(t)=0$, $g_4(t)=1$ and $u_0(x)=\max\{x-0.4,0\}$.
In Figure (1), we present the solution obtained with $h=0.01$ and $\delta=0.001$, in some instants. We may observe that the behaviour is similar to the behaviour of the exact solution. In figure (2), the convergence error for h is studied, that is, we establish a fixed value $\delta=0.001$ and different values of $h=0.1,0.01,0.001$. For each value of h, we calculated the error in the $L_2(a,b)$ norm and we collected the results in the graph represented.
From the figure presented we may conclude that the convergence is only of order 2. We suspect that this behaviour is due to the lack of regularity of the solution.

Example 2: In order to compare the behaviour of the solution for the different models presented, we simulated Equation (16) for each model. In Figure 3, we represent the solution and the first derivative of the non-linear Black-Scholes equation for the different transaction costs models at $t=0$. The parameters used are: $r=0.2$, $\sigma=0.2$, $Le=0.6$, $M=30$, $C=0.01$, $a^2=0.4$, $K=0.4$, $T=1$, $h=0.1$, $\delta=0.001$, $g_1(t)=0$, $g_2(t)=1-0.4e^{rt}$, $g_3(t)=0$, $g_4(t)=1$ and $u_0(x)=\max\{x-0.4,0\}$.

 ISBN: 978-84-617-8694-7

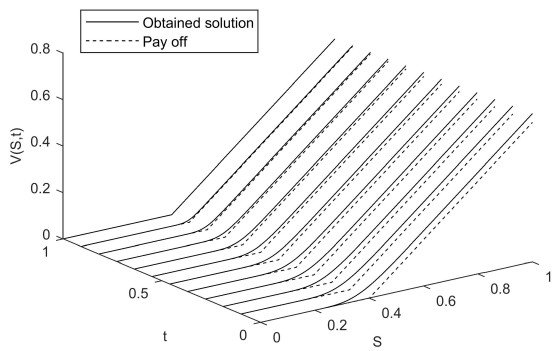


Figure 1: Obtained solution in example 1.

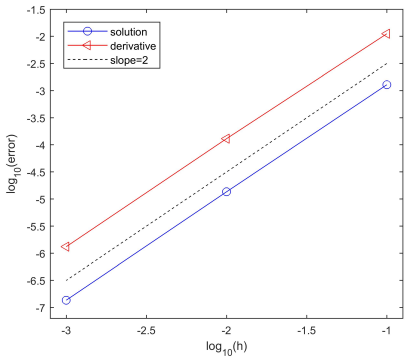


Figure 2: Convergence analysis for h in example 1.

The chart shows that the difference is not significant for all the transaction cost models. At this point, with the given parameters, the Leland model provides the highest price, followed by the Barles and Soner model, the Kratka model, and finally the linear model with constant volatility without transaction costs. An analysis from the initial to the maturity date permits us to conclude that the difference between the various models decreases as the expiration date approaches. This is an expected consequence of the decreasing necessity of portfolio adjustment and hence lower transaction costs closer to expiry. The difference is bigger at the beginning of the year, where the non-linear price is higher than the linear price.

7 Conclusions

A finite element method based on Hermite polynomials, to solve the non-linear problem in the non-divergent form, in a domain with fixed boundaries, was presented. The program resulting from the implementation of this method in Matlab code was tested with the linear equation. The error and convergence analysis was done in Example 1. The Crank-Nicolson method presents a convergence order of 2, for both solution and its derivative, while the finite element method has a convergence order of approximately 2 for the solution and 2 for its derivative, which does not fit the methods applied in this work. In Example 2, the solution and the derivative of the non-linear Black-Scholes equation were simulated with the different transaction cost models, taking into account the European-type call option. The study shows that the difference is not significant for all transaction cost models and it decreases the closer we are from the expiration date. As future work, we will carry out a similar study for American options.

 ISBN: 978-84-617-8694-7

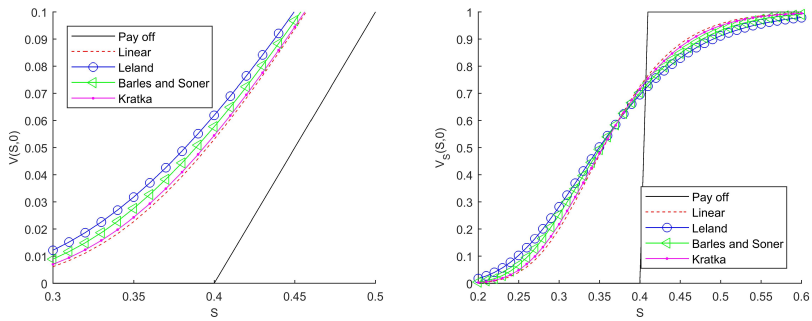


Figure 3: Solution (left) and its derivative (right) of the non-linear Black-Scholes equation with different transaction costs for European call options.

Acknowledgements

This work was partially supported by the research project: Grant N. UID/MAT/00212/2013 - financed by FEDER through the - Programa Operacional Factores de Competitividade, FCT - Fundação para a Ciência e a Tecnologia.

References

[1] R. M. Almeida, S. N. Antontsev, and J. C. Duque, *On the finite element method for a nonlocal degenerate parabolic problem*, Computers and Mathematics with Applications, 73 (2017), pp. 1724 – 1740.

[2] R. M. P. Almeida, J. C. M. Duque, J. Ferreira, and R. J. Robalo, *The Crank-Nicolson-Galerkin finite element method for a nonlocal parabolic equation with moving boundaries*, Numer. Methods Partial Differential Equations, 31 (2015), pp. 1515–1533.

[3] J. Ankudinova and M. Ehrhardt, *On the numerical solution of nonlinear BlackScholes equations*, Computers and Mathematics with Applications, 56 (2008), pp. 799 – 812. Mathematical Models in Life Sciences amp; Engineering.

[4] G. Barles and H. M. Soner, *Option pricing with transaction costs and a nonlinear BlackScholes equation*, Finance and Stochastics, 2 (1998), pp. 369–397.

[5] F. Black and M. Scholes, *The pricing of options and corporate liabilities*, J. Polit. Econ., 81 (1973), pp. 637–654.

 ISBN: 978-84-617-8694-7

[6] J. C. M. Duque, R. M. P. Almeida, S. N. Antontsev, and J. Ferreira, *The Euler-Galerkin finite element method for a nonlocal coupled system of reaction-diffusion type*, J. Comput. Appl. Math., 296 (2016), pp. 116–126.

[7] S. Hodges and A. Neuberger., *Optimal replication of contingent claims under transaction costs*, Review of Futures Markets, 8 (1989), pp. 222–239.

[8] M. Kratka, *No mystery behind the smile*, Risk, 9 (1998), pp. 67–71.

[9] H. E. Leland, *Option pricing and replication with transactions costs*, The Journal of Finance, 40 (1985), pp. 1283–1301.

[10] H. M. Soner, S. E. Shreve, and J. Cvitanic, *There is no nontrivial hedging portfolio for option pricing with transaction costs*, Appl. Probab, 5 (1995), pp. 327–355.

[11] M. K. S. Uddin, M. Ahmed, and S. K. Bhowmik, *A note on numerical solution of a linear Black-Scholes model*, Ganit, 33 (2013), pp. 103–115.

[12] Yue-Kuen, *Mathematical Models of Financial Derivatives*, Springer, 1998.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

A Fast Implementation of Matrix Trigonometric Functions Sine and Cosine

Pedro Alonso[1], Jesús Peinado[2], Javier Ibáñez[2], Jorge Sastre[3] and Emilio Defez[4]

[1] *Depto. de Sistemas Informaticos y Computación, Universitat Politècnica de València*

[2] *Instituto de Instrumentación para Imagen Molecular, Universitat Politècnica de València*

[3] *Instituto de Telecomunicaciones y Aplicaciones Multimedia (iTEAM), Universitat Politècnica de València*

[4] *Instituto de Matemática Multidisciplinar, Universitat Politècnica de València*

emails: palonso@upv.es, jpeinado@dsic.upv.es, jjibanez@dsic.upv.es, jsastrem@upv.es, edefez@imm.upv.es

Abstract

This paper describes the highlights of our algorithm implementation for the computation of the trigonometric functions sine and cosine of a matrix. This algorithm is based on a Taylor series approximation, which is characterized from the computational point of view because is very rich in matrix multiplications. We have used this feature to develop an implementation that can use one or two GPUs if available. In addition, our implementation can be easily used in the MATLAB environment.

Key words: Matrix Trigonometric functions, GPU computing, MATLAB, mex MATLAB.

1 Introduction

The trigonometric matrix functions *sine* and *cosine* arise in many engineering processes that are described by second order differential equations. Several state-of-the-art algorithms have been provided recently for computing these matrix functions using polynomial and rational approximations with scaling and recovering techniques [3, 5, 1]. One of the last and most competitive proposals in both accuracy and performance to obtain the cosine of a matrix

A ($\cos(A)$) is [2], which is a Taylor based algorithm. This paper addresses the challenge of performing the computation suggested in [2] as fast as possible on a computer composed by a multicore CPU and two NVIDIA GPUs. The challenge is twofold since the aim is also to develop an easy-to-use interface that allows the user to execute our algorithm from MATLAB.

2 Matrix Cosine Computation

The matrix cosine can be defined for all $A \in \mathbb{C}^{n\times n}$ by

$$\cos(A) = \sum_{i=0}^{\infty} \frac{(-1)^i A^{2i}}{(2i)!} ,$$

and let

$$T_{2m}(A) = \sum_{i=0}^{m} \frac{(-1)^i B^i}{(2i)!} \equiv P_m(B), \tag{1}$$

be the Taylor approximation of order $2m$ of $\cos(A)$, where $B = A^2$. We have implemented an "accelerated" version of the algorithm in [2] that computes a Taylor approximation $P_m(B)$ (1). The approximation of $\cos(A)$ is recovered by means of the double angle formula $\cos(2X) = 2\cos^2(X) - I$. By using the fact that $\sin(A) = \cos(A - \frac{\pi}{2}I)$, the same algorithm can also be used to compute the matrix sine.

3 The Accelerated Version

In our implementation, just like in [2], we have used a technique called Paterson-Stockmeyer's method [4] that allows to reduce the number of matrix products. But still the number of matrix products can be large. The fact that the computational cost is based on matrix products is a good reason to use GPUs since the matrix product is a very optimized operation in GPU. Using one or two GPUs is thus a logical step forward in reducing the execution time required to compute (1). Our implementation is efficient because it highly exploits the capabilities of the underlying existing GPUs.

The *accelerated* version was developed with the aim of being not only efficient but also easy to use and easy to modify. This is accomplished by keeping the MATLAB interface and syntax of the original algorithm [2]. The strategy consists of executing those parts of the MATLAB function that are good candidates to be accelerated using MATLAB *mex functions* implemented in CUDA and C++, and leaving the remaining operations to the host.

Data in both host and device memories must be persistent all along the time a MATLAB function is executing. A key point of our implementation is the use of only one `mex` function

 ISBN: 978-84-617-8694-7

to implement all the different operations because some data must remain in the device memory between consecutive calls, and this can be realized calling repeatedly to the same `mex` function. The `mex` function implemented, called `call_gpu`, has a string as its first argument (we name it the *command tag*) that labels the action to do. The rest of the arguments depend on the action to be performed. For example, `call_gpu('init',A))`, which actually is the first call that must be written in the code, allocates memory into the device to host some matrices like, e.g. matrix A and the resulting matrix, among others. Under this command, matrix A is also transferred from the MATLAB working space in the host memory to the device memory. Other operations implemented for the GPU are called with labels `'scale'`, `'unscale'`, `'power'`, `'norm1'`, and `'finalize'`.

This solution provides ease to the user. The original MATLAB code is modified with very few commands, i.e. only replacing few lines in the code with calls to the `mex` function with the appropriate *command tag*. This can be carry out by a non expert user on GPUs and/or on CUDA programming. However, the user who implements the MATLAB code must be aware that once matrix A has been uploaded to the GPU, this matrix and the derived ones, e.g. (B_1, B_2, ...), will be stored only into the device.

4 Experimental Results

The experimentation has been performed on a host computer, equipped with an Intel Quad-Core i7-3820 (3.6Ghz) processor, and with two NVIDIA GPU of type K20c (Kepler architecture) attached. Each GPU has 13 multiprocessors with 192 cores each, resulting in a total of 2496 CUDA cores. The two devices jointly make available to the user a total of 4992 CUDA cores for processing. Each device features 4800 MBytes of RAM memory.

The experiment shown in Figure 1 compares the execution time using the host CPU, one GPU or two GPUs. In general, it can be stated that the overhead incurred by data transference through the PCIe bus in implementations that use GPUs makes inadequate its use for small problem sizes. The raw MATLAB implementation for CPU is the best option when the problem size is less than $n \approx 520$ in case there are two GPUs available, or $n \approx 640$ when there exists just one GPU within reach. The use of two GPUs is always better than using only one for problems larger than ≈ 60.

5 Conclusions

This contribution describes the two key features of our implementation of the Taylor series approximation algorithm proposed in [2] to calculate the sine (or cosine) of a matrix. Our implementation can use one or two GPUs. The application is easy to use since the commands provided to the user can be easily integrated into a MATLAB code.

 ISBN: 978-84-617-8694-7

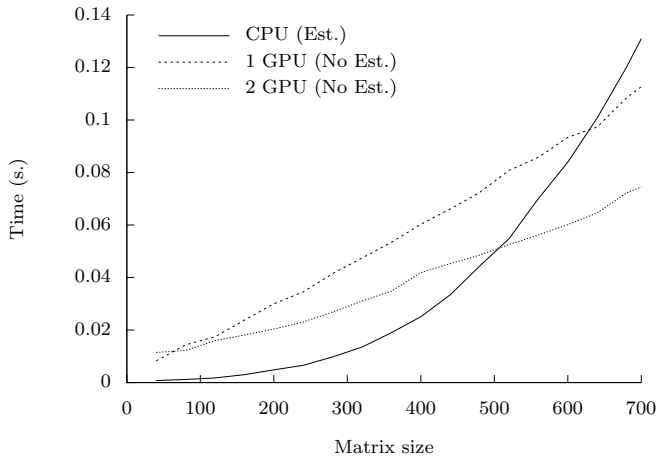


Figure 1: Execution time to obtain $\cos(A)$ using the CPU, 1 GPU, and 2 GPUs.

Our solution outperforms MATLAB in the sense that, though currently MATLAB provides with some functions that allow to use one GPU, it is not possible yet to use two GPUs at the same time.

Acknowledgements

This work has been supported by Spanish Ministerio de Economía y Competitividad and the European Regional Development Fund (ERDF) grants TIN2014-59294-P and TEC2015-67387-C4-1-R.

References

[1] Awad H. Al-Mohy, Nicholas J. Higham, and Samuel D. Relton. New algorithms for computing the matrix sine and cosine separately or simultaneously. *SIAM J. Sci. Comput.*, 37(1):A456–A487, 2015.

[2] P. Alonso, J. Ibáñez, J. Sastre, J. Peinado, and E. Defez. Efficient and accurate algorithms for computing matrix trigonometric functions. *J. Comput. Appl. Math.*, 309(1):325–332, January 2017.

[3] E. Defez, J. Sastre, Javier J. Ibáñez, and Pedro A. Ruiz. Computing matrix functions arising in engineering models with orthogonal matrix polynomials. *Math. Comput. Model.*, pages 1738–1743, 2011. . doi:10.1016/j.mcm.2011.11.022.

 ISBN: 978-84-617-8694-7

[4] Michael S. Paterson and Larry J. Stockmeyer. On the number of nonscalar multiplications necessary to evaluate polynomials. *SIAM J. Comput.*, 2(1):60–66, 1973.

[5] J. Sastre, J. Ibáñez, P. Ruiz, and E. Defez. Efficient computation of the matrix cosine. *Appl. Math. Comput.*, 219:7575–7585, 2013.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

Pivoting strategies and almost strictly sign regular matrices

P. Alonso[1], J.M. Peña[2] and M.L. Serrano[1]

[1] *Departamento de Matemáticas, Universidad de Oviedo, Spain*

[2] *Departamento de Matemática Aplicada, Universidad de Zaragoza, Spain*

emails: palonso@uniovi.es, jmpena@unizar.es, mlserrano@uniovi.es

Abstract

Several pivoting strategies for Neville elimination of almost strictly sign regular matrices are recalled and compared. They are also compared with Gauss elimination with partial pivoting. Some optimal properties of two-determinant pivoting are presented.

Key words: pivoting strategies, almost strictly sign regular matrices
MSC 2000: 65F05, 65F15, 65F35

1 Introduction

Numerical methods adapted to structured classes of matrices have been studied recently. A very important class of structured matrices due to its applications is the class of *sign regular* (SR) matrices. A matrix is SR if all its minors of the same order have the same sign. The importance of nonsingular SR matrices comes from their characterization as variation diminishing transformations. This property has played a crucial role in the applications to Statistics, Economy or Computer-Aided Geometric Design (see [5, 6, 15]).

A relevant subclass of SR matrices is formed by *totally positive* (TP) matrices, that is, matrices such that all their minors are nonnegative. The study of TP matrices began in 1930 with the work of Schoenberg (see [17]). The exhaustive research carried out in the field of TP matrices is reflected in books written several decades ago, such as Gantmacher and Krein, whose original was published in Russian in 1941 and has an English version of 2002 (see [9]), or Karlin's 1968 book (see [14]). There are also more recent texts such as that edited by Gasca and Michelli in 1996 (see [10]) and more recently Pinkus' book (see [16]) and Fallat and Johnson's book (see [8]). In contrast, the knowledge about the class

of SR matrices is much smaller. This is due, above all, to the much greater difficulties that arise from their study.

The goal of this work is the study of another subclass of SR matrices: the *almost strictly sign regular* (ASSR) matrices (see [1]). These matrices have all their nontrivial minors of the same order with the same strict sign and this subclass contains the nonsingular *almost strictly totally positive* (ASTP) matrices, introduced by Gasca, Miccheli and Peña (see [11]). A nonsingular matrix is ASTP if a minor with consecutive rows and columns is positive if and only if it has positive diagonal entries. Hurwitz matrices and B-splines collocation matrices are examples of ASTP matrices. In general, problems with ASSR matrices are much more difficult to deal with than the corresponding problems for ASTP matrices.

This work analyzes several aspects about the application of some pivoting strategies to ASSR matrices using Neville elimination (NE) or Gaussian elimination (GE). NE is an elimination procedure alternative to GE, very useful when dealing with SR matrices and their subclasses.

In [12] the scaled partial pivoting with respect to the l_∞-norm and Euclidean norm are studied for GE and NE applied to totally positive linear systems. It is proved that in exact arithmetic row exchanges are not necessary. In [2] a backward error analysis of Neville procedure is presented (with and without pivoting strategy). In the case of TP matrices, the error bounds are similar to those obtained previously by other authors for GE. In 2007 a pivoting strategy called two-determinant pivoting (see [7]) is analyzed. The application of NE with two-determinant pivoting strategy to ASSR matrices is studied in [3]. It is shown that this procedure preserves the almost strict sign regularity and that the associated Wilkinson-type growth factor is optimal.

This work is organized as follows. In Section 2 we present several pivoting strategies for NE. In Section 3 we compare, these pivoting strategies and GE with partial pivoting, using an illustrative example. Moreover, we announce the better properties of the two-determinant pivoting strategies under several points of view.

2 Some pivoting strategies for Neville elimination

In this section we briefly present the NE and several row pivoting strategy associated to this method.

NE is a very convenient procedure when working with ASSR matrices and other related types of matrices. If A is a nonsingular $n \times n$ matrix, NE consists of at most $n-1$ successive major steps, resulting in a sequence of matrices as follows:

$$A = \widetilde{A}^{(1)} \to A^{(1)} \to \cdots \to \widetilde{A}^{(n)} = A^{(n)} = U \tag{1}$$

where U is an upper triangular matrix.

 ISBN: 978-84-617-8694-7

For each t, $1 \leq t \leq n$, $A^{(t)} = \left(a_{ij}^{(t)}\right)_{1\leq i,j\leq n}$ has zeros in the positions $a_{ij}^{(t)}$, for $1 \leq j \leq t$, $j \leq i \leq n$. Besides it holds that

$$a_{it}^{(t)} = 0,\ i \geq t \Rightarrow a_{ht}^{(t)} = 0,\ \forall h \geq i. \tag{2}$$

The matrix $A^{(t)}$ is obtained from $\widetilde{A}^{(t)}$ reordering rows t, $t+1$, ..., n according to a row pivoting strategy that satisfies (2).

To obtain $\widetilde{A}^{(t+1)}$ from $A^{(t)}$ we produce zeros in the column t below the main diagonal by subtracting a multiple of the ith row from the $(i+1)$th, for $i = n-1,\ n-2,\ \ldots,\ t$, according to the following formula:

$$\widetilde{a}_{ij}^{(t+1)} = \begin{cases} a_{ij}^{(t)}, & 1 \leq i \leq t, \\ a_{ij}^{(t)} - \frac{a_{it}^{(t)}}{a_{i-1,t}^{(t)}} a_{i-1,j}^{(t)}, & \text{if } a_{i-1,t}^{(t)} \neq 0, \quad t+1 \leq i \leq n, \\ a_{ij}^{(t)}, & \text{if } a_{i-1,t}^{(t)} = 0, \quad t+1 \leq i \leq n, \end{cases} \tag{3}$$

for all $j = 1, 2, \ldots, n$.

The element

$$p_{ij} = a_{ij}^{(j)}, \quad 1 \leq j \leq i \leq n \tag{4}$$

is called the (i,j) *pivot* of NE of A and the number

$$m_{ij} = \begin{cases} \dfrac{a_{ij}^{(j)}}{a_{i-1,j}^{(j)}} \left(= \dfrac{p_{ij}}{p_{i-1,j}} \right), & \text{if } a_{i-1,j}^{(j)} \neq 0, \\ 0, & \text{if } a_{i-1,j}^{(j)} = 0, \end{cases} \tag{5}$$

the (i,j) *multiplier*. Note that $m_{ij} = 0$ if and only if $p_{ij} = 0$ and, by (2),

$$m_{ij} = 0 \implies m_{hj} = 0,\ \forall h > i. \tag{6}$$

Now we present some pivoting strategies for NE, using row exchanges with similar purposes to those of the pivoting strategies for GE. Recall that $A^{(t)}$ is obtained by reordering the rows of matrix $\widetilde{A}^{(t)}$ by an adequate pivoting strategy with a criterion for the choice of the pivots p_{ij}. GE with partial pivoting chooses the pivots so that all multipliers have absolute value not greater than 1. With a similar purpose, one can define *NE with partial pivoting*. We interchange the rows of $\widetilde{A}^{(t)}$ so that $A^{(t)}$ satisfies

$$|a_{tt}^{(t)}| \geq |a_{t+1,t}^{(t)}| \geq \cdots \geq |a_{nt}^{(t)}|$$

and from $A^{(t)}$ we construct $\widetilde{A}^{(t+1)}$ as in (3).

 ISBN: 978-84-617-8694-7

To improve the previous strategy we will incorporate an adequate scaling, thus arising the scaled partial pivoting, if we use the standard $\|\cdot\|_\infty$, or the Euclidean scaled partial pivoting if the norm used is Euclidean $\|\cdot\|_2$ (see [12]).

We will denote by $r_i^{(t)}$ to the ith row of the submatrix $A^{(t)}[1,\dots,n|t,\dots,n]$, with $i=t,t+1,\dots,n$ and $t=1,\dots,n-1$, and with $s_i^{(t)}$, $S_i^{(t)}$ to the amounts

$$\begin{aligned} s_i^{(t)} &= \|r_i^{(t)}\|_\infty = \max_{t\le j\le n} |a_{ij}^{(t)}|, \\ S_i^{(t)} &= \|r_i^{(t)}\|_2 = ((a_{it}^{(t)})^2+\cdots+(a_{in}^{(t)})^2)^{1/2}. \end{aligned} \tag{7}$$

In the *scaled partial pivoting for NE*, for each value of t, the rows of $A^{(t)}$, from the row tth to the nth are ordered according to a permutation $(i_1,i_2,\dots,i_{n-t+1})$ of the elements $(t,t+1,\dots,n)$, so that

$$\frac{|a_{i_1,t}^{(t)}|}{s_{i_1}^{(t)}} \ge \frac{|a_{i_2,t}^{(t)}|}{s_{i_2}^{(t)}} \ge \cdots \ge \frac{|a_{i_{n-t+1},t}^{(t)}|}{s_{i_{n-t+1}}^{(t)}}. \tag{8}$$

Then, $A^{(t)}$ has been obtained by substituting the tth row of $\widetilde{A}^{(t)}$ for the i_1th, $(t+1)$th for the i_2th and so on.

For the case of the Euclidean scaled partial pivoting strategy, we can obtain an analogous expression to (8) simply replacing in it $s_i^{(t)}$ for $S_i^{(t)}$.

In [7] a row pivoting strategy associated to NE for nonsingular SR matrices is introduced. It will be called two-determinant pivoting strategy due to the special role played by some 2×2 determinants of some matrices appearing along the Neville procedure. Besides, the authors also study the scaled partial pivoting in NE.

The criterion of the *two-determinant pivoting* strategy to obtain $A^{(t)}[t,\dots,n]$ from a reordering of the rows of $\widetilde{A}^{(t)}[t,\dots,n]$ is the following:

- If $\widetilde{a}_{tt}^{(t)} = 0$: then we reverse the ordering of the rows, that is, $A^{(t)}[t,\dots,n] := P_{n-t+1}\widetilde{A}^{(t)}[t,\dots,n]$.
- If $\widetilde{a}_{nt}^{(t)} = 0$: then we do not perform rows exchanges, that is, $A^{(t)} := \widetilde{A}^{(t)}$.
- If $\widetilde{a}_{tt}^{(t)} \neq 0$ and $\widetilde{a}_{nt}^{(t)} \neq 0$, then we compute the determinant $d_1 = \det \widetilde{A}^{(t)}[t,t+1]$.
  - If $d_1 > 0$ then $A^{(t)} := \widetilde{A}^{(t)}$.
  - If $d_1 < 0$ then $A^{(t)}[t,\dots,n] := P_{n-t+1}\widetilde{A}^{(t)}[t,\dots,n]$.
  - If $d_1 = 0$ then compute the determinant $d_2 = \det \widetilde{A}^{(t)}[n-1,n|t,t+1]$.
    - If $d_2 > 0$ then $A^{(t)} := \widetilde{A}^{(t)}$.
    - If $d_2 < 0$ then $A^{(t)}[t,\dots,n] := P_{n-t+1}\widetilde{A}^{(t)}[t,\dots,n]$.

 ISBN: 978-84-617-8694-7

In Section 3 of [7] it is shown that this pivoting strategy is well defined for nonsingular SR matrices. The computational cost of the NE without row exchanges for an $n \times n$ matrix coincide with the cost of GE without row exchanges. So it has a cost of $\frac{4n^3+3n^2-7n}{6} \simeq \frac{2n^3}{3}$ flops (floating-point operations). Using the two-determinant pivoting strategy, this cost is increased with at most $2n-2$ subtractions and $4n-4$ multiplications. Besides, by Theorem 4.1 of [7], for a nonsingular SR matrix, the two-determinant pivoting strategy for NE is a scaled partial pivoting strategy for any monotone vector norm.

Other pivoting strategies used in both GE and NE are called pairwise pivoting (see [18, 19]). These are suitable for implementations on parallel computers and reduce the communication cost considerably .

In contrast to row pivoting strategies described above (associated with (1)) pairwise pivoting strategies interchange consecutive rows in each step and then produce a zero. In fact *pairwise pivoting for NE* is defined as follows: to produce a zero at position $(n,1)$ in $\widetilde{A}^{(1)}$, one compares the element $\widetilde{a}_{n1}^{(1)}$ with $\widetilde{a}_{n-1,1}^{(1)}$. If $|\widetilde{a}_{n1}^{(1)}| > |\widetilde{a}_{n-1,1}^{(1)}|$ the corresponding rows are exchanged, in order that $|\widetilde{a}_{n1}^{(1)}/\widetilde{a}_{n-1,1}^{(1)}| \leq 1$. Then the elements of row n are updated producing a zero in the $(n,1)$ entry. We continue with the first column until producing a zero in the $(2,1)$ entry, and then we would continue with the second column and later columns, analogously to NE, until obtaining an upper triangular matrix U. For a nonsingular matrix A, NE with pairwise pivoting consists of al most $n(n-1)/2$ successive step, resulting in a sequence of matrices as follows:

$$A = \widetilde{A}^{(1)} \to A^{(1)} \to \cdots \to \widetilde{A}^{(\frac{n(n-1)}{2})} = A^{(\frac{n(n-1)}{2})} = U \tag{9}$$

where U is an upper triangular matrix.

3 Comparison of the pivoting strategies

In this section, some aspects regarding the application of pivoting strategies defined in the previous section to ASSR matrices are studied. We compare those pivoting strategies under different points of view.

Let A be an $n \times n$ matrix. The Wilkinson-type growth factor associated to NE with a certain pivoting strategy is given by

$$\rho(A) = \frac{\max_{i,j,t} \left|\widetilde{a}_{ij}^{(t)}\right|}{\max_{i,j} |a_{ij}|} \tag{10}$$

where $\widetilde{A}^{(t)} = \left(\widetilde{a}_{ij}^{(t)}\right)_{1\leq i,j\leq n}$ are the intermediate matrices of the elimination process. A similar Wilkinson-type growth factor for GE with partial pivoting can be defined (see Section 9.3 of [13]).

 ISBN: 978-84-617-8694-7

We shall define another growth factor, which is a normwise growth factor, in contrast to the componentwise growth factor defined in (10). We announce that the two-determinant pivoting strategy is optimal for both growth factors. It also has other important advantages over the other pivoting strategies defined in Section 2. In fact, two-determinant pivoting is zero-increasing and preserves the almost strictly sign regularity.

Given an ASSR matrix A we can observe that, NE with partial pivoting, GE with partial pivoting and pairwise pivoting for NE do not preserve the structure of this matrix. The following example illustrates this situation.

Let A be an ASSR matrix with signature $\varepsilon = (-1, 1, -1, 1, -1, -1)$

$$A = \widetilde{A}^{(1)} = \begin{pmatrix} -1 & -4 & 0 & 0 & 0 & 0 \\ -2 & -10 & -10 & -16 & -2 & 0 \\ 0 & -6 & -33 & -60 & -21 & 0 \\ 0 & -8 & -46 & -92 & -70 & -36 \\ 0 & 0 & -9 & -60 & -242 & -316 \\ 0 & 0 & -6 & -60 & -443 & -2823 \end{pmatrix}.$$

Taking into account that $|a_{21}| = 2 > |a_{11}| = 1$, the application of NE with partial pivoting (which coincides with GE with partial pivoting in the first step) implies that the rows 1 and 2 must be exchanged, so

$$A^{(1)} = \begin{pmatrix} -2 & -10 & -10 & -16 & -2 & 0 \\ -1 & -4 & 0 & 0 & 0 & 0 \\ 0 & -6 & -33 & -60 & -21 & 0 \\ 0 & -8 & -46 & -92 & -70 & -36 \\ 0 & 0 & -9 & -60 & -242 & -316 \\ 0 & 0 & -6 & -60 & -443 & -2823 \end{pmatrix}, \ \widetilde{A}^{(2)} = \begin{pmatrix} -2 & -10 & -10 & -16 & -2 & 0 \\ 0 & 1 & 5 & 8 & 1 & 0 \\ 0 & -6 & -33 & -60 & -21 & 0 \\ 0 & -8 & -46 & -92 & -70 & -36 \\ 0 & 0 & -9 & -60 & -242 & -316 \\ 0 & 0 & -6 & -60 & -443 & -2823 \end{pmatrix}$$

and $\widetilde{A}^{(2)}[2, \ldots, 6]$ matrix is not an ASSR matrix. Note that $\widetilde{A}^{(2)}$ has less zeros than A.

If pairwise pivoting with NE is considered, then

$$A = \widetilde{A}^{(1)} = A^{(1)} = \widetilde{A}^{(2)} = A^{(2)} = \widetilde{A}^{(3)} = A^{(3)} = \widetilde{A}^{(4)}$$

and the matrices

$$A^{(4)} = \begin{pmatrix} -2 & -10 & -10 & -16 & -2 & 0 \\ -1 & -4 & 0 & 0 & 0 & 0 \\ 0 & -6 & -33 & -60 & -21 & 0 \\ 0 & -8 & -46 & -92 & -70 & -36 \\ 0 & 0 & -9 & -60 & -242 & -316 \\ 0 & 0 & -6 & -60 & -443 & -2823 \end{pmatrix}, \ \widetilde{A}^{(5)} = \begin{pmatrix} -2 & -10 & -10 & -16 & -2 & 0 \\ 0 & 1 & 5 & 8 & 1 & 0 \\ 0 & -6 & -33 & -60 & -21 & 0 \\ 0 & -8 & -46 & -92 & -70 & -36 \\ 0 & 0 & -9 & -60 & -242 & -316 \\ 0 & 0 & -6 & -60 & -443 & -2823 \end{pmatrix}.$$

Observe that the property of almost strict sign regularity is not inherited by $\widetilde{A}^{(5)}[2, \ldots, 6]$. Note also that $\widetilde{A}^{(5)}$ has less zeros than A.

 ISBN: 978-84-617-8694-7

As for two-determinant pivoting strategy, we get the following matrices $A = \widetilde{A}^{(1)} = A^{(1)}$ and

$$\widetilde{A}^{(2)} = \begin{pmatrix} -1 & -4 & 0 & 0 & 0 & 0 \\ 0 & -2 & -10 & -16 & -2 & 0 \\ 0 & -6 & -33 & -60 & -21 & 0 \\ 0 & -8 & -46 & -92 & -70 & -36 \\ 0 & 0 & -9 & -60 & -242 & -316 \\ 0 & 0 & -6 & -60 & -443 & -2823 \end{pmatrix}.$$

In this case $\widetilde{A}^{(2)}[2, \ldots, 6]$ is ASSR and $\widetilde{A}^{(2)}$ has more zeros than A. These properties also hold by matrices $\widetilde{A}^{(t)}[t, \ldots, 6]$, for all $t = 3, \ldots, 6$, when we apply NE with two-determinant pivoting. Besides, $\rho(A) = 1$ and in this case the stability of the strategy is assured.

Acknowledgements

This work has been partially supported by the Spanish Research Grant MTM2015-65433-P and MTM2015-68805-REDT.

References

[1] P. Alonso, J.M. Peña, M.L. Serrano, *On the characterization of almost strictly sign regular matrices*, J. Comput. Appl. Math. **275** (2015), 480–488.

[2] P. Alonso, M. Gasca, J.M. Peña, *Backward error analysis of Neville elimination*, Appl Numer Math. **23** (1997) 193–204.

[3] P. Alonso, J.M. Peña, M.L. Serrano, *Almost strictly sign regular matrices and Neville elimination with two-determinant pivoting*, Appl. Math. Comput. **289** (2016) 426–434.

[4] P. Alonso, J.M. Peña, M.L. Serrano, *Almost strictly sign regular matrices and Neville elimination with two-determinant pivoting*, J. Comput. Appl. Math. **322** (2017) 71–80.

[5] T. Ando, *Total positive matrices*, Linear Algebra Appl. **90** (1987) 165–219.

[6] L.D. Brown, I.M. Johnstone, K.B. MacGibbon, *Variation Diminishing transformations: a direct approach to total positivity and its statistical applications*, J. Am. Stat. Assoc. **76** (1981) 824–832.

[7] V. Cortés, J.M. Peña, *Sign regular matrices and Neville elimination*, Linear Algebra Appl. **421** (2007) 53–62.

[8] S.M. Fallat, C.R. Johnson, *Totally Nonnegative Matrices*, Princeton University Press, 2011.

[9] F P. Gantmacher, M.G. Krein, *Oscillation Matrices and Kernels and Small Vibrations of Mechanical Systems*, AMS Chelsea, 2002.

[10] M. Gasca, C.A. Micchelli, *Total Positivity and its Applications*, Kluwer Academic Publishers, 1996.

[11] M. Gasca, C.A. Micchelli, J.M. Peña, *Almost strictly totally positive matrices*, Numer. Algor. **2** (1992) 225–236.

[12] M. Gasca, J.M. Peña, *Scaled pivoting in Gauss and Neville elimination for totally positive systems*, Appl. Numer. Math. **13** (1993) 345–355.

[13] N.J. Higham, *Accuracy and Stability of Numerical Algorithms*, Second Edition, SIAM, Philadelphia, 2002.

[14] S. Karlin, *Total Positivity*, Stanford University Press, 1968.

[15] J.M. Peña, *Shape Preserving Representations in Computer-Aided Geometric Design*, Nova Science Publishers, 1999.

[16] A. Pinkus, *Total positive matrices*, Cambridge University Press, 2009.

[17] I.J. Schoenberg, *Uber Variationsverminderende lineare Transformationen*, Math. Z. **32** (1930) 321–328.

[18] D.C. Sorensen, *Analysis of pairwise pivoting in Gaussian elimination*, IEEE Trans. Comput. **C-34** (1985) 274–278.

[19] L.N. Trefethen, R.S. Schreiber, *Average case stability of Gaussian elimination*, SIAM J. Matrix Anal. Appl. **11** (1990) 335–360.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

Looking for efficiency when avoiding order reduction in nonlinear problems with Strang splitting

I. Alonso-Mallo[1], B. Cano[1] and N. Reguera[2]

[1] *IMUVA, Departmento de Matemática Aplicada, Universidad de Valladolid*

[2] *IMUVA, Departamento de Matemáticas y Computación, Universidad de Burgos*

emails: isaias@mac.uva.es, bego@mac.uva.es, nreguera@ubu.es

Abstract

In this paper, we offer a comparison in terms of computational efficiency between two techniques to avoid order reduction when using Strang method to integrate nonlinear initial boundary value problems with time-dependent boundary conditions. Considering different implementations for each of the techniques, we show that the technique suggested by Alonso et al. is more efficient than the one suggested by Einkemmer et al.

Key words: Strang splitting, avoiding order reduction, computational comparison,
MSC 2000: 65M12 65M20

1 Introduction

There are several papers in the literature concerning the important fact of avoiding the order reduction in time which turns up when integrating with splitting methods nonlinear problems of the form

$$\begin{aligned} u'(t) &= Au(t) + f(t, u(t)), \quad 0 \le t \le T, \\ \partial u(t) &= g(t), \\ u(0) &= u_0, \end{aligned} \tag{1}$$

where A is an elliptic differential operator, f is a smooth real function which acts as a reaction term, ∂ is a boundary operator, g is the boundary condition which in principle does not vanish and is time-dependent and u_0 is a smooth initial condition which makes that the solution of (1) is regular enough.

 ISBN: 978-84-617-8694-7

More particularly, in [6, 7] a technique is suggested to do it, in which each part of the splitting is assumed to be solved in an exact way for the analysis and, in the numerical experiments, standard subroutines are used to integrate each part in space and time. Although, from the point of view of the analysis, the technique in both papers is equivalent, the difference is that, in [6], the solution of this elliptic problem is required at each time $t \in [0, T]$,

$$Az(t) = 0, \quad \partial z(t) = g(t),$$

and a suggestion for $z_t(t)$ must be given. This is very simple analytically in one dimension but much more complicated and expensive in several dimensions. Nevertheless, that is avoided in [7] by considering just a function q which coincides with $f(g)$ at the boundary. That can also be done analytically in one dimension and simple domains in two dimensions, and numerically in more complicated domains, according to a remark made in [7] although it is not in fact applied to such a problem there. In this paper, we will concentrate on the technique in [7] for 1-dimensional and 2-dimensional simple domains.

On the other hand, in [3], another different technique is suggested in which appropriate boundary conditions are suggested for each part of the splitting. The analysis there considers both the space and time discretization. The linear and stiff part is integrated 'exactly' in time through exponential-type functions while the nonlinear but smooth part is assumed to be numerically integrated by a classical integrator just with the order of accuracy that the user wants to achieve with the whole method. Although the latter seems to be the more natural, in order to be more similar in the comparison with the technique in [7], we will use standard subroutines which use variable stepsizes with given small tolerances for the nonlinear and smooth problems of both techniques.

We will concentrate on the extensively used second-order Strang splitting and the aim of the paper is to compare both techniques in terms of computational efficiency, considering different space discretizations, different tolerances for the standard subroutines which integrate in time some of the split problems, and different (although standard) ways to tackle the calculation of terms which contain exponential-type functions of matrices. For that, both one-dimensional and bidimensional problems will be considered. There is already another comparison in the literature between both techniques [8] but there they just compare in terms of error against the time stepsize without entering into the details of implementation and its computational cost, which we believe that is the interesting comparison. Moreover, they just consider time-independent boundary conditions and 1-dimensional problems, for which many simplifications can be made.

The paper is structured as follows. Section 2 gives some preliminaries on the description of the different techniques and suggest different implementations for each of them. Section 3 presents results for all the methods suggested in terms of error against cpu time when using accurate spectral collocation methods in space and very small tolerances for the standard subroutines in time. Section 4 offers also a numerical comparison, but now using less

 ISBN: 978-84-617-8694-7

accurate finite differences in space and less small tolerances for the standard subroutines in time. Moreover, numerical differentiation is also considered in order to achieve local order 3 instead of just 2, and again the computational comparison is performed.

2 Preliminaries and suggestion of different implementations

The technique suggested in [7] consists of the following: A function $q(t)$ is constructed which satisfies $\partial q(t) = \partial f(t, u(t))$. Then, given the numerical approximation at the previous step u_n, the numerical approximation at the next step u_{n+1} is given by the following procedure:

$$\begin{aligned}
&\left\{\begin{array}{rcl} v'_{n,1}(t) &=& Av_{n,1}(t) + q(t), \\ v_{n,1}(t_n) &=& u_n, \\ \partial v_{n,1}(t) &=& g(t), \end{array}\right. \\
&\left\{\begin{array}{rcl} w'_n(t) &=& f(t, w_n(t)) - q(t), \\ w_n(t_n) &=& v_{n,1}(t_n + \frac{k}{2}), \end{array}\right. \\
&\left\{\begin{array}{rcl} v'_{n,2}(t) &=& Av_{n,2}(t) + q(t), \\ v_{n,2}(t_n + \frac{k}{2}) &=& w_n(t_n + k), \\ \partial v_{n,2}(t) &=& g(t), \end{array}\right. \\
&u_{n+1} = v_{n,2}(t_n + k).
\end{aligned} \tag{2}$$

However, we notice that two of three problems which turn up here are stiff and therefore solving them will be more expensive than solving the unique nonlinear but smooth problem. In order to reverse that, the decomposition of the splitting method can be done in another order and then the following procedure would turn up, for which with similar arguments, no order reduction would either turn up:

$$\begin{aligned}
&\left\{\begin{array}{rcl} w'_{n,1}(t) &=& f(t, w_{n,1}(t)) - q(t), \\ w_{n,1}(t_n) &=& u_n, \end{array}\right. \\
&\left\{\begin{array}{rcl} v'_n(t) &=& Av_n(t) + q(t), \\ v_n(t_n) &=& w_{n,1}(t_n + \frac{k}{2}), \\ \partial v_n(t) &=& g(t) \end{array}\right. \\
&\left\{\begin{array}{rcl} w'_{n,2}(t) &=& f(t, w_{n,2}(t)) - q(t), \\ w_{n,2}(t_n + \frac{k}{2}) &=& v_n(k), \end{array}\right. \\
&u_{n+1} = w_{n,2}(t_n + k).
\end{aligned} \tag{3}$$

Then, two of the problems are cheap and just one is more expensive.

On the other hand, in [3], the main idea is to consider, from u_n,

$$\left\{\begin{array}{rcl} w'_n(s) &=& Aw_n(s), \\ w_n(0) &=& \Psi^{f,t_n}_{\frac{k}{2}}(u_n), \\ \partial w_n(s) &=& \partial[u(t_n) + \frac{k}{2} f(t_n, u(t_n)) + sAu(t_n)], \end{array}\right.$$

 ISBN: 978-84-617-8694-7

$$u_{n+1} = \Psi_{\frac{k}{2}}^{f,t_n+\frac{k}{2}}(w_n(k)).$$

where $\Psi_{\frac{k}{2}}^{f,t_n}$ and $\Psi_{\frac{k}{2}}^{f,t_n+\frac{k}{2}}$ integrate respectively with order 2 the following problems from $s = 0$ to $s = k/2$:

$$v_n'(s) = f(t_n + s, v_n(s)), \quad z_n'(s) = f(t_n + \frac{k}{2} + s, z_n(s)).$$

Moreover, the procedure to integrate this is more explicitly stated. Firstly, in [3] (see also [1, 2, 5]), a general space discretization is introduced which discretizes the elliptic problem

$$Au = F, \quad \partial u = g,$$

through the 'elliptic projection' $R_h u$ which satisfies

$$A_{h,0}R_h u + C_h g = P_h F,$$

for a certain matrix $A_{h,0}$, an associated boundary operator C_h and a projection operator P_h. Then, given the numerical approximation at the previous step U_h^n, the procedure in [3] to obtain U_h^{n+1} reads as follows:

$$\begin{aligned} V_h^n &= \Psi_{\frac{k}{2}}^{f,t_n}(U_h^n), \\ W_{h,n}(k) &= e^{kA_{h,0}}V_h^n + k\varphi_1(kA_{h,0})C_h[g(t_n) + \frac{k}{2}\partial f(t_n, u(t_n))] \\ &\quad + k^2\varphi_2(kA_{h,0})C_h[g'(t_n) - \partial f(t_n, u(t_n)] \\ U_h^{n+1} &= \Psi_{\frac{k}{2}}^{f,t_n+\frac{k}{2}}(W_{h,n}(k)), \end{aligned} \tag{4}$$

where φ_1 and φ_2 are the standard functions which are used in exponential methods [3]. The original suggestion used this order for the decomposition thinking that Ψ_k is just an explicit method which is applied with a single stepsize k, and therefore it would be cheaper than the equation in $W_{h,n}(k)$. We still believe that would be the best. However, as in this paper, in order to do it more similarly to [7], we will solve that part with a standard variable stepsize subroutine for non-stiff problems until a given small tolerance, the first and last problem may be more expensive than the middle one. Therefore, we will also consider this other implementation which comes from reversing the order of the problems in the decomposition:

$$\begin{aligned} W_{h,n}(\frac{k}{2}) &= e^{\frac{k}{2}A_{h,0}}U_h^n + \frac{k}{2}\varphi_1(\frac{k}{2}A_{h,0})C_h g(t_n) + \frac{k^2}{4}\varphi_2(\frac{k}{2}A_{h,0})C_h\partial Au(t_n) \\ V_h^n &= \Psi_k^{f,t_n}(W_{h,n}(\frac{k}{2})), \\ U_h^{n+1} &= e^{\frac{k}{2}A_{h,0}}V_h^n + \frac{k}{2}\varphi_1(\frac{k}{2}A_{h,0})C_h\partial[u(t_n) + \frac{k}{2}Au(t_n) + kf(t_n, u(t_n)] \\ &\quad + \frac{k^2}{4}\varphi_2(\frac{k}{2}A_{h,0})C_h\partial Au(t_n). \end{aligned} \tag{5}$$

 ISBN: 978-84-617-8694-7

In the following, we will denote by **EO1** to (3), by **EO2** to (2), by **ACR1** to (4) and by **ACR2** to (5).

3 Numerical comparison with spectral collocation methods in space and high accuracy in time

In this section a comparison in terms of computational efficiency among the different techniques is given when solving all the problems at hand with high accuracy. In such a way, we will be seeing the error which comes from the splitting itself for a large range of values of the timestepsize k. More precisely, spectral collocation methods [4] are used for the space discretization with the two implementations of both techniques. Moreover, the nonlinear and non-stiff problems of both techniques are integrated with MATLAB subroutine ode45 with relative tolerance 10^{-12} and absolute tolerance 10^{-15}. As for the linear and stiff part of **EO1** and **EO2** ((3) and (2) respectively), we have considered subroutine ode15s with the same tolerances. On the other hand, in this section, for the implementation of the equation on $W_{h,n}$ in **ACR1** and the first and last equation in **ACR2** ((4) and (5) respectively), as $A_{h,0}$ is a full but small matrix and we consider k as fixed during the whole integration, we have calculated once and for all the matrices $e^{kA_{h,0}}$, $\varphi_1(kA_{h,0})C_h$ and $\varphi_2(kA_{h,0})C_h$. The calculation of these matrices is not included in the measured computational time since this initial cost is negligible when integrating till large times but may be very big when T is small. (For another implementation of those terms with no initial cost, look at Section 4.)

In a first place, we have considered the following one-dimensional Dirichlet boundary value problem whose exact solution is $u(x,t) = e^{t+x^3}$:

$$\begin{aligned} &u_t(x,t) = u_{xx}(x,t) + u^2 - e^{t+x^3}(9x^4 + 6x + e^{t+x^3} - 1), \quad 0 \le x \le 1, \quad t \in [0, 0.2], \\ &u(x,0) = e^{x^3}, \\ &u(0,t) = e^t, \quad u(1,t) = e^{t+1}. \end{aligned} \tag{6}$$

For the spectral space discretization, 16 Gauss-Lobatto interior nodes have been used so that the error in space is negligible. We also notice that the matrix $e^{kA_{h,0}}$ has dimension 16×16 while $\varphi_1(kA_{h,0})C_h$ and $\varphi_2(kA_{h,0})C_h$ have dimension 16×2 since they are just multiplied by the values at the boundary. Therefore, the calculation of the terms in $e^{kA_{h,0}}$ is more expensive than the calculation of those in $\varphi_i(kA_{h,0})C_h$ $(i = 1, 2)$. On the other hand, as it is a one-dimensional problem, the function q in **EO1** and **EO2** is calculated directly for every value of t as the straight line which joins the corresponding values $f(t, 0, e^t)$ and $f(t, 1, e^{t+1})$ at $x = 0$ and $x = 1$ respectively. The values of the time stepsize which have been displayed have been $k = 10^{-3}, 5 \times 10^{-4}, 2.5 \times 10^{-4}, 1.25 \times 10^{-4}, 6.25 \times 10^{-5}$ for **EO** and $k = 10^{-3}, 5 \times 10^{-4}, 2.5 \times 10^{-4}, 1.25 \times 10^{-4}, 6.25 \times 10^{-5}, 3.125 \times 10^{-5}$ for **ACR**. The results in terms of maximum error against computational cost are in Figure 1 and we can

 ISBN: 978-84-617-8694-7

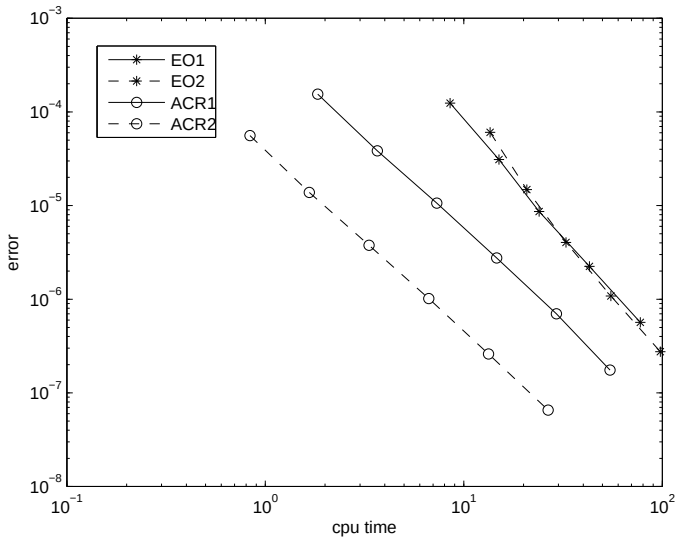


Figure 1: Numerical comparison with spectral collocation methods in space and high accuracy in time for the 1-dimensional problem (6)

see that techniques **ACR1 and ACR2 are more efficient than EO1 and EO2**, that **ACR2** is three times more efficient than **ACR1** and that **EO1** and **EO2** are very similar in efficiency. In any case, we can check that, as already remarked in the previous section, for a fixed value of k, **ACR2** is cheaper than **ACR1** and **EO1** cheaper than **EO2**. Moreover we can observe that, at least for this particular problem, for fixed k, the error is smaller with the second implementation of both techniques than with the first.

In a second place, we have considered the two-dimensional problem

$$\begin{aligned}
&u_t(x,y,t) = u_{xx}(x,y,t) + u_{yy}(x,y,t) + f(t,x,y,u(x,y,t)), \quad 0 \le x,y \le 1, \quad t \in [0,0.2],\\
&u(x,y,0) = e^{x^3+y^3},\\
&u(0,y,t) = e^{t+y^3}, \quad u(1,y,t) = e^{t+1+y^3}, \quad u(x,0,t) = e^{t+x^3}, \quad u(x,1,t) = e^{t+1+x^3}. \qquad (7)
\end{aligned}$$

where $f(t,x,y,u) = u^2 - e^{t+x^3+y^3}(9(x^4+y^4)+6(x+y)+e^{t+x^3+y^3}-1)$, so that the exact solution is $u(x,y,t) = e^{t+x^3+y^3}$. Now, 16 interior Gauss-Lobatto nodes have been taken in each direction of the square for the space discretization and the implementation has been performed with similar remarks to those of the one-dimensional case. The only remarkable difference is that the function $q(t,x,y)$ in **EO1** and **EO2** must be chosen in a different way. We consider a function of the form $q(t,x,y) = r(t,x)f(t,1,y,e^{t+1+y^3}) + s(t,x)f(t,0,y,e^{t+y^3})$ which satisfies the corresponding conditions at the boundary and that is achieved if $r(t,x)$

 ISBN: 978-84-617-8694-7

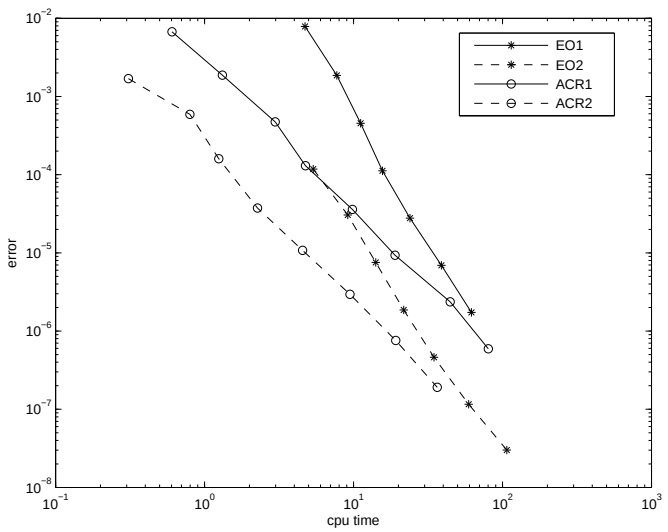


Figure 2: Numerical comparison with spectral collocation methods in space and high accuracy in time for the 2-dimensional problem (7)

and $s(t,x)$ satisfy

$$\begin{pmatrix} f(t,1,0,e^{t+1}) & f(t,0,0,e^{t}) \\ f(t,1,1,e^{t+2}) & f(t,0,1,e^{t+1}) \end{pmatrix} \begin{pmatrix} r(t,x) \\ s(t,x) \end{pmatrix} = \begin{pmatrix} f(t,x,0,e^{t+x^3}) \\ f(t,x,1,e^{t+1+x^3}) \end{pmatrix}.$$

(Notice that this technique to calculate q analytically can be applied in a rectangular domain but not in more complicated domains in two dimensions.) In Figure 2, which corresponds to the following values of the timestepsize $k = 2\times10^{-2}, 10^{-2}, 5\times10^{-3}, 2.5\times10^{-3}, 1.25\times10^{-3}, 6.25\times10^{-4}, 3.125\times10^{-4}$ for **EO** and $k = 2.5\times10^{-3}, 1.25\times10^{-3}, 6.25\times10^{-4}, 3.125\times10^{-4}, 1.5625\times10^{-4}, 7.8125\times10^{-5}, 3.9063\times10^{-5}, 1.9531\times10^{-5}$ for **ACR**, we can see that the second implementation of both techniques is cheaper than the first and that **the best of all implementations is ACR2**, at least for a range of errors $\geq 10^{-7}$.

4 Numerical comparison with finite difference methods in space and middle accuracy in time

In this section we have been a bit less demanding when solving each part of the splitting. We have just considered 10^{-7} and 10^{-8} as relative and absolute tolerances respectively

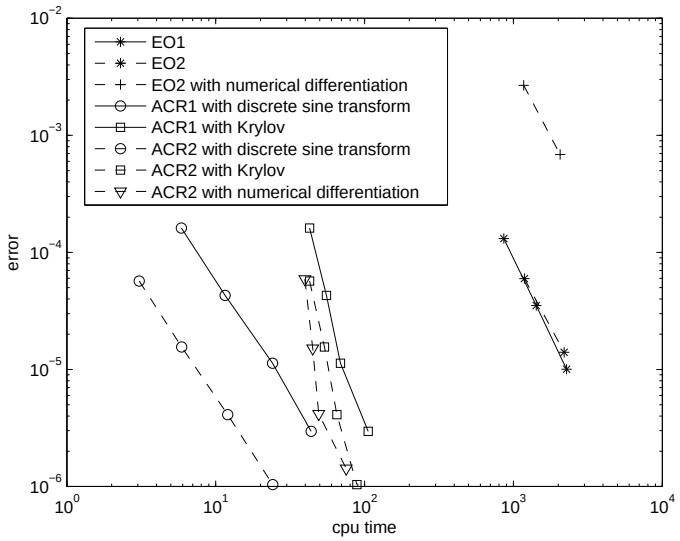


Figure 3: Numerical comparison with finite difference methods in space and middle accuracy in time for the 1-dimensional problem (6)

for the standard subroutines ode45 and ode15s. As Strang method just has second-order accuracy, it is usually used for problems in which a very high precision is not required. Moreover, in space we have considered finite differences of just second order accuracy in the space grid. More particularly, as in the problems above the operator A is the Laplacian, we have taken the standard symmetric second-order difference scheme in 1 dimension and the five-point formula in 2 dimensions [11]. We have considered as space grid $h = 5 \times 10^{-4}$ for the 1-dimensional case and $h = 2 \times 10^{-2}$ for the 2-dimensional case. With this type of implementation, the matrix $A_{h,0}$ is sparse and, in this particular case, their eigenvalues and eigenvectors are well-known [9]. Because of the former, it is natural to use standard Krylov subroutines [10] in **ACR1** and **ACR2** to calculate the application of exponential-type functions over vectors. Due to the latter, which is more specific of this particular example and space discretization, in order to calculate the same terms, it seems advantageous to use the discrete sine transform in the same way that FFT is used in Poisson solvers [9]. When using Krylov subroutines [10], we have considered the default tolerance 10^{-7}. The comparison is performed in Figure 3 for the 1-dimensional problem (6) with $k = 10^{-3}, 5 \times 10^{-4}, 2.5 \times 10^{-4}$ for **EO1**, $k = 10^{-3}, 5 \times 10^{-4}$ for **EO2** and $k = 10^{-3}, 5 \times 10^{-4}, 2.5 \times 10^{-4}, 1.25 \times 10^{-4}$ for **ACR**. **We again see that ACR1 and ACR2 are more competitive than EO1 and EO2**. Although, for a fixed value of k, **EO2** takes more computational time than

 ISBN: 978-84-617-8694-7

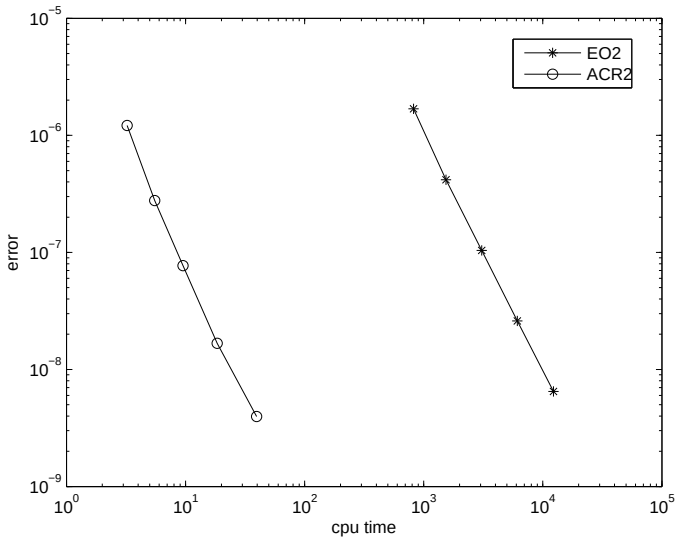


Figure 4: Numerical comparison with finite difference methods in space and middle accuracy in time for the 2-dimensional problem (8)

EO1, in the end they are very similar in efficiency since, at least in this case, the error is also quite smaller. As for **ACR1** and **ACR2**, **ACR2** is more competitive since not only the computational time is smaller for a fixed value of k but also the error is smaller. In this particular case, considering discrete sine transforms is much cheaper than using Krylov techniques. However, for a general operator A, that may not be possible and that is why it is also interesting to see the comparison when using these techniques. In any case, **the worst of ACR implementations is about 20 times cheaper than the best of EO**.

Moreover, following [8], we have also considered numerical differentiation in order to get local order 3 with **EO1** and **EO2** in (6). More precisely, a function q has been taken for which $\partial q(t) = \partial f(t, u(t))$ and $\partial Aq(t) = \partial Af(t, u(t))$. Although the order for the global error does not improve, it is interesting to see whether the fact that the local errors maybe smaller implies a better overall behaviour. Notice that, in (6),

$$\frac{d}{dx^2} f = f_{xx} + 2f_{x,u}u_x + f_{uu}u_x^2 + f_u u_{xx}.$$

As $\partial u_{xx} = g'(t) - \partial f(t, u)$, numerical differentiation is just required to calculate ∂u_x. For that, we have considered the second-order scheme

$$u_x(0,t) \approx \frac{-\frac{3}{2}u(0,t) + 2u(h,t) - \frac{1}{2}u(2h,t)}{h}, u_x(1,t) \approx \frac{\frac{3}{2}u(1,t) - 2u(1-h,t) + \frac{1}{2}u(1-2h,t)}{h}.$$

 ISBN: 978-84-617-8694-7

Although numerical differentiation is a badly-posed problem, its effect is still not visible with the considered value of h for the first derivative and the range of errors which we are considering. What is true is that the error constants may grow significantly because the size of q may be very big, as it happens in this problem and Figure 3 shows. The latter is the explanation that **numerical differentiation is not worth when using EO2**, as it was already remarked in [8] for some other particular problems and fixed k.

As for **ACR2**, considering also terms of second order in s for the boundaries of the problems in which the operator A appears, the following full scheme turns up:

$$\begin{aligned}
W_{h,n}(\frac{k}{2}) &= e^{\frac{k}{2}A_{h,0}}U_h^n + \frac{k}{2}\varphi_1(\frac{k}{2}A_{h,0})C_h g(t_n) + \frac{k^2}{4}\varphi_2(\frac{k}{2}A_{h,0})C_h\partial A u(t_n) \\
&\quad + \frac{k^3}{8}\varphi_3(\frac{k}{2}A_{h,0})C_h\partial A^2 u(t_n) \\
V_h^n &= \Psi_k^{f,t_n}(W_{h,n}(\frac{k}{2})), \\
U_h^{n+1} &= e^{\frac{k}{2}A_{h,0}}V_h^n \\
&\quad + \frac{k}{2}\varphi_1(\frac{k}{2}A_{h,0})C_h\partial[u(t_n) + k(\frac{1}{2}Au(t_n) + f(t_n,u(t_n)) \\
&\qquad + k^2(\frac{1}{8}A^2u(t_n) + \frac{1}{2}f_u(t_n,u(t_n))Au(t_n) + \frac{1}{2}(f_t(t_n,u(t_n)) + f_u(t_n,u(t_n))f(t_n,u(t_n)))] \\
&\quad + \frac{k^2}{4}\varphi_2(\frac{k}{2}A_{h,0})C_h\partial[Au(t_n) + \frac{k}{2}A^2u(t_n) + kAf(t_n,u(t_n))] \\
&\quad + \frac{k^3}{8}\varphi_3(\frac{k}{2}A_{h,0})C_h\partial A^2 u(t_n).
\end{aligned}$$

As $\partial A^2 u = \partial A\dot{u} - \partial Af = \ddot{g} - \partial(f_t + f_u\dot{u}) - \partial Af$, what is again necessary is to approximate u_x with numerical differentiation and we have done it in the same way as before. As it is observed, **there is a small ganancy in efficiency when using numerical differentiation with ACR2 although it is not extremely significant**.

Notice that, surprisingly, for both **EO2** and **ACR2**, for a fixed value of k, the computational cost does not increase but is slightly smaller when using numerical differentiation. This must be due to the fact that the standard subroutines which are used converge more quickly when numerical differentiation is applied. A full explanation for that is out of the scope of this paper although it might be a subject of future research.

In this section, in order to assure that the errors in space are negligible without having to decrease too much the space grid, we have considered as bidimensional problem

$$\begin{aligned}
&u_t(x,y,t) = u_{xx}(x,y,t) + u_{yy}(x,y,t) + f(t,x,y,u(x,y,t)), \quad 0 \le x,y \le 1, \quad t \in [0,0.2], \\
&u(x,y,0) = x^2 + y^2, \\
&u(0,y,t) = e^t y^2, \quad u(1,y,t) = e^t(1+y^2), \quad u(x,0,t) = e^t x^2, \quad u(x,1,t) = e^t(1+x^2). \qquad (8)
\end{aligned}$$

 ISBN: 978-84-617-8694-7

where $f(t,x,y,u) = u^2 - e^{2t}(x^2+y^2)^2 + e^t(x^2+y^2-4)$, so that the exact solution is $u(x,y,t) = e^t(x^2+y^2)$. For the sake of brevity, we have concentrated on **ACR2** and **EO2** because they were the best implementations in previous problems. We have implemented **EO2** calculating q in a similar way as in the bidimensional problem of the previous section and **ACR2** again with Krylov subroutines [10]. In Figure 4 we can see that **ACR2 is 500 times more efficient than EO2** where $k = 10^{-2}, 5\times10^{-3}, 2.5\times10^{-3}, 1.25\times10^{-3}, 6.25\times10^{-4}$ has been displayed for **EO2** and $k = 1.25\times10^{-3}, 6.25\times10^{-4}, 3.125\times10^{-4}, 1.5625\times10^{-4}, 7.8125\times10^{-5}$ for **ACR2**.

Acknowledgements

This work has been supported by project MTM 2015-66837-P.

References

[1] I. Alonso–Mallo, B. Cano and N. Reguera, *Avoiding order reduction when integrating linear initial boundary value problems with exponential splitting methods*, submitted for publication.

[2] I. Alonso–Mallo, B. Cano and N. Reguera, *Avoiding order reduction when integrating linear initial boundary value problems with Lawson methods*, accepted in IMA J. Numer. Anal., doi: 10.1093/imanum/drw052.

[3] I. Alonso-Mallo, B. Cano and N. Reguera, *Avoiding order reduction when integrating reaction-diffusion boundary value problems with exponential splitting methods*, arXiv:1705.01857, submitted for publication.

[4] C. Bernardy and Y. Maday, *Approximations spectrales de problemes aux limites elliptiques*, Springer-Verlag France, Paris, 1992. MR 94f:65112.

[5] B. Cano and N. Reguera, *Avoiding order reduction when integrating reaction-diffusion boundary value problems with exponential splitting methods*, J. Comp. Appl. Math. **316** (2017) 86–99.

[6] L. Einkemmer and A. Ostermann, *Overcoming order reduction in diffusion-reaction splitting. Part 1: Dirichlet boundary conditions*, SIAM J. Sci. Comput. **37 (3)** (2015), A1577–A1592.

[7] L. Einkemmer and A. Ostermann, *Overcoming order reduction in diffusion-reaction splitting. Part 2: Oblique boundary conditions*, SIAM J. Sci. Comput. **38** (2016) A3741-A3757.

[8] L. Einkemmer and A. Ostermann, *A comparison of boundary correction methods for Strang splitting*, arXiv:1609.05505v1.

[9] A. Iserles, *A First Course in the Numerical Analysis of Differential Equations*, Cambridge University Press, Cambridge, 2008.

[10] J. Niesen, and W. M. Wright, *Algorithm 919: a Krylov subspace algorithm for evaluating the φ-functions appearing in exponential integrators*, ACM Trans. Math. Software **38**, no. 3, Art. 22 (2012).

[11] J. C. Strikwerda, *Finite Difference Schemes and Partial Differential Equations*, Wadsworth & Brooks, United States of America, 1989.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

Fast Iterative Block QR Updating

Fran J. Alventosa[1], Pedro Alonso[1], Antonio M. Vidal[1] and Gema Piñero[2]

[1] *Depto. de Sistemas Informaticos y Computación, Universitat Politècnica de València*

[2] *Instituto de Telecomunicaciones y Aplicaciones Multimedia (iTEAM), Universitat Politècnica de València*

emails: fraalrue@upv.es, palonso@upv.es, avidal@dsic.upv.es, gpinyero@iteam.upv.es

Abstract

The processing of some digital sound signals can involve the QR factorization of a rectangular system matrix. But, sometimes, only a given (and probably small) part of the system matrix varies from the current sample to the next one. We exploit this fact to reuse some computations of the former samples processing to save execution time in the processing of the actual sample. This saving in execution time can be critical for real time application addressed to low power consumption devices with high mobility.

Key words: QR factorization, QR Updating, real-time, block QR.

1 Introduction

The QR factorization of a matrix is a very known method to solve systems of linear equations or to approximate the solution to an overdetermined system (more equations than unknowns). Let $A \in R^{m \times n}$, with $m \geq n$, be a matrix that represents the linear system, its QR factorization is $A = QR$, being $Q \in \Re^{m \times m}$ orthogonal and $R \in \Re^{m \times n}$ upper triangular.

This factorization can be computed in $O(mn^2)$ flops either through Householder reflections or through Givens rotations [3]. Also, there exist block algorithms that improve performance on processors with a hierarchical set of cache memories [4, 2]. In this work we start form a sequential block algorithm that performs this factorization [5], and that we modify to incorporate our proposal. We use OpenBLAS [1] in the basis of our algorithm.

We use the QR factorization with the aim of extracting the most approximate solution of a least squares problem. The least squares problem arises as the main computational kernel

 ISBN: 978-84-617-8694-7

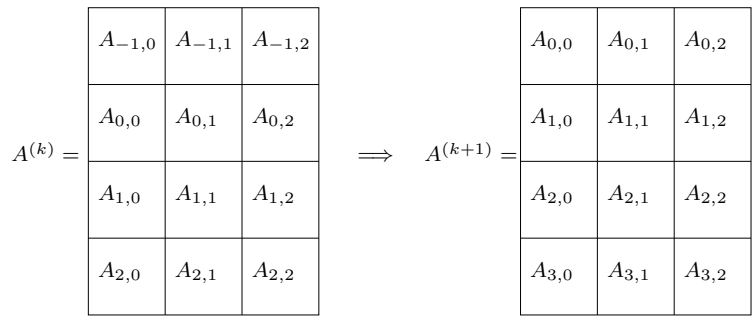


Figure 1: Updating of system matrix A from iteration k to $k+1$.

of a more complex process used to evaluate a sound signal sample. The digital sound signal is sampled real time, what means that there exists a very short slot of time to perform all the computations associated to the sample processing. Furthermore, the problem is critical since we are interested in the use of devices with low computational capabilities (mobile phones, tablets, etc.). Hence, what we propose is to speed up the process of performing the QR factorization of each matrix at each sample by reusing some computations performed in the processing of the former sample since there exist some common data to matrices consecutive samples that can be reused.

2 The QR updating

The process we tackle here is iterative (a new sample is processed at each iteration). At iteration k, $k = 0, \ldots$, a QR factorization is carried on the system matrix $A^{(k)}$. Matrix $A^{(k)}$ is "updated" with new data to form the system matrix of the following iteration $A^{(k+1)}$. This "update" consist of 1) deleting the most ancient rows, which are the first t_s rows; and 2) appending a new set of t_s rows to the bottom of the system matrix, which are coming from the sampled signal. The updating carried out between iterations k and $k+1$ can be observed in Figure 1, represented by matrices $A^{(k)}$ and $A^{(k+1)}$. The figure represents the matrix partitioned in tiles of size $t_s \times t_s$.

As stated before, the most time consuming process at iteration k consists of the QR factorization, i.e. $A^{(k)} = Q_k R_k$. In order to save flops in each factorization we propose to work on matrices that we denote as *jagged*, and that are represented by J in Figure 2. Each block row of J^k is the "R" factor of each corresponding row in $A^{(k)}$. Hence, to form J^{k+1} there exists an intermediate step that consists of performing the QR factorization of the new t_s rows $\begin{pmatrix} A_{3,0} & A_{3,1} & A_{3,2} \end{pmatrix}$, and append the "R" factor i.e. $\begin{pmatrix} J_{3,0} & J_{3,1} & J_{3,2} \end{pmatrix}$ to the bottom of the matrix. The "Q" is not explicitly formed since it is not needed. It is

 ISBN: 978-84-617-8694-7

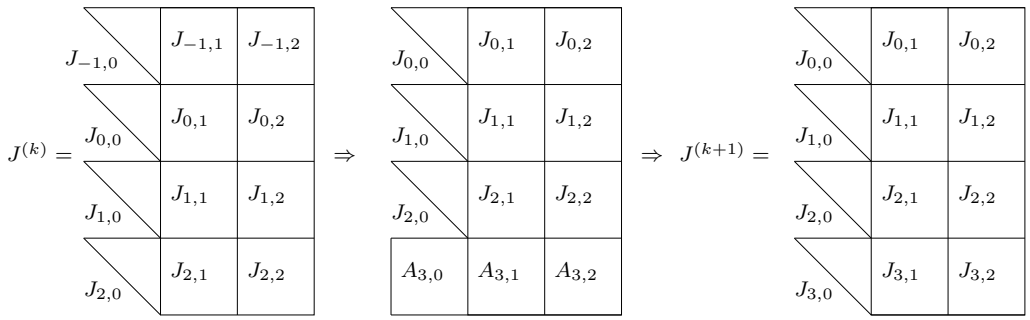


Figure 2: Updating of system matrix J from iteration k to $k+1$ with a *jagged* matrix.

clear that computing the QR factorization of the *jagged* matrix is cheaper than to compute the QR of the original one.

3 Experimental Results

The experimental results have been obtained in a single ARM® Cortex-A15 core of a NVIDIA Jetson TK1 development kit. Table 1 shows the time in seconds to perform the QR factorization of each matrix $A^{(k)}$ (column QR($A^{(k)}$)), and the factorization of the *jagged* matrix $J^{(k)}$ (column QR($J^{(k)}$)). We show two problem sizes and two different tile sizes for each problem. Also, we vary the block size b_s, which is a parameter that must be tuned to exploit the cache memory when working on a tile. Both sizes influence the performance of the algorithm, but while the tile size t_s is problem dependent (number of rows to be updated on the system matrix) there exist freedom to select the block size (b_s).

The best performance of the QR factorization of the two matrices, i.e. QR($A^{(k)}$) and QR($J^{(k)}$), is obtained always for the same (t_s, b_s) combination. In these cases, the speed up in percentage of QR($J^{(k)}$) with regard to QR($A^{(k)}$) is in the range $\approx [41, 45]$.

4 Conclusion

Real time applications are characterized by performing the same computation repeatedly over new coming data. Sometimes, data, which frequently is represented as a system matrix, changes from one iteration to the next one only in a small part. If the computation on this matrix is expensive, as it is the QR factorization, we can use this fact to save processing time. We propose to work on a modified matrix, called *jagged*, instead of on the original system matrix. With this simple idea, it is possible to increase performance by a factor close to 1.45 in the particular case we tackled in this paper, i.e. when data is represented

 ISBN: 978-84-617-8694-7

Table 1: Time in seconds of QR factorization of matrix A vs. matrix J for problem sizes 1280×960 and 2560×1920 varying the tile size and the block size.

$m \times n$	t_s	b_s	QR($A^{(k)}$)	QR($J^{(k)}$)
1280×960	160	20	0.630 s.	0.434 s.
		32	0.647 s.	0.445 s.
		40	0.652 s.	0.452 s.
		80	0.724 s.	0.515 s.
	320	20	0.601 s.	0.417 s.
		32	0.615 s.	0.426 s.
		40	0.624 s.	0.432 s.
		80	0.667 s.	0.479 s.

$m \times n$	t_s	b_s	QR($A^{(k)}$)	QR($J^{(k)}$)
2560×1920	320	20	4.583 s.	3.273 s.
		32	4.572 s.	3.249 s.
		40	4.554 s.	3.227 s.
		80	4.721 s.	3.360 s.
	640	20	4.542 s.	3.227 s.
		32	4.484 s.	3.160 s.
		40	4.436 s.	3.137 s.
		80	4.504 s.	3.224 s.

by a 4×3 tiles matrix and 50% of data (25% discarded rows and 25% new rows) changes from one iteration to the next one. In the future works we plan to perform this in parallel using the several cores of the NVIDIA Jetson.

Acknowledgements

This work has been supported by projects TEC2015-67387-C4-1-R of the Spanish Ministerio de Economía y Competitividad and PROMETEOII/2014/003 of the Generalitat Valenciana.

References

[1] Openblas. http://www.openblas.net. Accessed on May 2017.

[2] A. Buttari, J. Langou, J. Kurzak, and J. Dongarra. A class of parallel tiled linear algebra algorithms for multicore architectures. *Parallel Computing*, 35(1):38 – 53, 2009.

[3] G.H. Golub and C.F. Van Loan. *Matrix Computations*. Johns Hopkins Studies in the Mathematical Sciences. Johns Hopkins University Press, 2013.

[4] Brian C. Gunter and Robert A. van de Geijn. Parallel out-of-core computation and updating the QR factorization. *ACM Trans. on Math. Soft.*, 31(1):60–78, March 2005.

[5] Gregorio Quintana-Ortí, Enrique S. Quintana-Ortí, Robert A. Van De Geijn, Field G. Van Zee, and Ernie Chan. Programming matrix algorithms-by-blocks for thread-level parallelism. *ACM Trans. Math. Softw.*, 36(3):14:1–14:26, July 2009.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

Computer Aided Ship Analysis using Subdivision Schemes

S. Amat[1], M. J. Legaz[2] and J. Ruiz[3]

[1] *Departamento de Matemática Aplicada y Estadística, Universidad Politécnica de Cartagena (Spain)*

[2] *Departamento de Ciencias y Técnicas de la Navegación y Construcciones Navales, Universidad de Cádiz (Spain)*

[3] *Departmento de Física y Matemáticas, Universidad de Alcalá (Spain)*

emails: sergio.amat@upct.es, mariajose.legaz@uca.es, juan.ruiza@uah.es

Abstract

This document tries to give an overview of some applications related to ship analysis using subdivision schemes. It presents a little introduction of some subdivision schemes and its properties. It reviews the modeling, simulation and design of the ship structures.

Key Words. Computer-Aided Ship Analysis, simulation, modeling, structural analysis, design.

1 Introduction

The importance of subdivision to applications in computer-aided geometric design is clear. The surfaces generated by their subdivision schemes can generate bivariate functions, but also they can easily represent surfaces of arbitrary topology (Doo and Sabin (1978) and by Catmull and Clark (1978)).

The first work on a subdivision scheme was by de Rahm (1956), but in the pioneering work of Chaikin (1974) was introduced the first subdivision as a practical algorithm for curve design.

Subdivision algorithms are well suited to computer applications: they are simple to compute, highly flexible, and well adapted for dynamical problems.

The subdivision tools have been applied in several design applications, such as in the 3D animation industry, but is not the case in other more demanding application related to geometric modelling in the industry.

 ISBN: 978-84-617-8694-7

For overviews and tutorials on subdivision schemes and their applications, the reader may turn to Cavaretta, Dahmen and Micchelli (1991), Schröder (2001), Zorin and Schröder (2000) and Warren (1995) or Dyn and Levin (2002).

In this paper, we are interesting in the use of subdivision schemes in the analysis and desing of ship applications. After a little review of some mathematical tools for Ship applications, we mention the advantages of the use of subdivision schemes.

2 Ship design and analysis: A brief introduction

We start with a brief review of some mathematical tools used in the Ship design and analysis.

2.1 Ship design

The modern engineering designs are becoming increasingly more complex because of the pressing demands of economies of scale, higher speed, lower motions, operational ability in adverse weather conditions, energy efficiency, clean environment, and advancements in material science and technologies. Today, modern engineering design and analysis around the world can no longer be imagined without intensive use of computers and computational methods.

The discipline of Computer-Aided Ship Design (CASD) was certainly well established by 1960, although it has several roots and historical precursors we can recognize three main roots:

- The need for digital media in the numerical control of manufacturing automata.
- The desire for digital representation of the ship geometric product model replacing the tedious and error prone graphical process of ship lines definition.
- The application of computers for computationally intensive, time-consuming tasks of ship design calculations as in ship stability, hydrodynamic and structural analysis.

The digital modeling of ship hull geometries was a primary prerequisite in capturing the complex, curved, free form shape of ship lines and surfaces for all purposes of hull form analysis.

Ship design is a decision-making process that leads from given requirements to a product definition with all relevant information for the performance assessment and production of the ship.

The definition of the ship hull form is one of the earliest steps in ship design because initially many necessary assessments depend on at least a provisional description of hull shape. Later this tentative description is refined into a definitive and fair ship hull form definition for production purposes.

 ISBN: 978-84-617-8694-7

The predominant curve representations in hull form definition as in other branches of CAD (Computer Aided Design) applications soon became parametric polynomial forms, in particular:

- Bézier curves [5, 12, 1].
- B-splines [5, 12, 13].
- Non-uniform rational B-Splines [5, 12, 11].

These representations have the advantage of being manipulable interactively by control points, hence also on computer monitors. They offer any desired orders of piecewise continuity and can be elevated to high polynomial degrees. The rational variant of Bézier and B-spline curves also includes conic sections, sometimes used for special features of hull form. For an overview of these geometric modeling capabilities see [5, 12]. For a nice history introduction see [9].

A frequent approach is the generation of hull surfaces from given curves. Initially a set of piecewise continuous characteristic curves is generated from provided offset data and end constraints. As an example of offset data in Maxsurf. They may come from another ship with features similar to the one which is modeled. Together they form a regular or irregular mesh, not unlike a lines plan, connected at the mesh knots.

There exist several software, some of them that can be use for the ship design and that they use Bézier and/or NURBS curves and/or surfaces: Autoship (Autoship Systems Corporation), DefCar (DefCar Engineering), Fastship (Proteus Engineering), HullCAO (HullCAO), Hull Form (Blue Peter Marine Systems), Maxsurf (Formation Design Systems), MultiSurf (Aerohydro), Prolines (Vacanti Yacht Design), ProSurf (New Wave Systems), Rhino (Robert McNeel and Assoc.), Naval Designer (US Sales by Forum Marine), SeaSolution (SeaSolution), TouchCAD (Lundström Design).

2.2 Ship structural analysis

The application of ship structural analysis is a key element in ensuring the safety and economy of the ship. The ship must meet unique safety standards to operate safely in the hostile maritime scenario without jeopardizing human lives, cargo and hull, or the marine environment. The design solutions must realize low light ship weights and production costs in order to be economically competitive. These basic objectives are not new, but the approach to achieving viable and attractive solutions has changed significantly during the recent decades of Computer-Aided Ship Design.

Traditionally before the computer era the structural design of merchant ships was generally based on classification rules and international ship safety conventions which were derived in part on analytical grounds, but in large measure also on empirical observations

 ISBN: 978-84-617-8694-7

and accident statistics. Load assumptions and operating risk scenarios were often hidden behind declarative statements in the rules. Most of the design was founded on deterministic load cases combined with protective safety factors to cope with uncertainties. This methodology was applied responsibly and without undue risks to conventional designs.

The picture has grossly changed during the following decades driven by advances in structural analysis, load assumptions, probabilistic modeling, reliability analysis, nonlinear optimization, and many innovative ideas in design solutions. Ship structural analysis has benefited from general progress in this field, such as by Finite Element Method, Finite Volume Method, and Boundary Element Method developments. But much of the innovation was also due to the maritime field with its unique requirements. The common denominator of many small steps of innovation seems to be a long-term trend toward more rational design assumptions, based on probabilistic models of many actual load cases and operating scenarios, treated by systematic optimization strategies.

Aside of many advances in computational capabilities, mainly by discrete element methods, which were also shared with many other engineering branches, I want to single out the important influence of probabilistic models in ship structural design. Ships operate in uncertain conditions, mainly because of their environment in irregular seaways of varying character and because of their vulnerability to random accidents such as collisions, groundings etc. A safe design must take these risk factors into account. The response of the ship to such random influences could be modeled probabilistically as soon as analytical methods for describing random processes had become available. This entry into the area of ship theory is marked by two milestone papers:

Ship structural design today has become a rationally based, optimization oriented, mature discipline of ship design. It uses accurate, statistically described load assumptions, multiple load cases and failure modes, reliability-based constraints, and modern analysis and optimization techniques. A modern approach based on these principles is well described by Hughes [8].

2.3 Isogeometric analysis

The finite element method is an useful technique in solving partial differential equations, which has been widely applied to the solve problems in engineering. Starting from the variational model, piecewise low order polynomials on the subdivision meshes of the domain are computed to approximate the solution of the differential model.

The main motivation in isogeometric analysis is to reduce the gap between the the finite element community and the computer-aided design community (Hughes et al., 2005). This gap has become increasingly with the past of the years. It is not only a shortcoming of the current technology but of the entire engineering process. The idea in isogeometric analysis has been to connect in a natural way the design and analysis camps. In particular, isogeometric analysis seeks to unify the fields of CAD and FEA.

 ISBN: 978-84-617-8694-7

There are many computational geometry technologies that could serve as a basis for isogeometric analysis. Maybe, the most studied is Non-Uniform Rational B-Spline (NURBS) basis functions. There are several reasons for selecting NURBS as the initial basis, but the most important is that these basis are widely used computational geometry technology in engineering design.

3 Why Subdivision schemes?

There are a main problem with NURBS is that they are not good adapted for dynamics applications. Some artifacts appear on the generated surfaces that can be corrected by hand. For these reason, subdivision schemes have been considered as an alternative in the animation industry.

Subdivision schemes are a tool that has become widely used in the computer aided geometric design (CAGD) community in the past years. They are usually used to generate smooth curves and surfaces. If the subdivision scheme is convergent, a discrete set of data points can be used for generating a smooth limit curve or surface using low complexity algorithms. In general, the limit function is obtained applying recursively the subdivision scheme.

In the CAGD field, the criteria used to decide the quality of the algorithms are not only based on complexity, convergence or regularity of the schemes. It is also important to have into account the order of approximation, in order to determine the precision of the scheme. Numerical artifacts introduced by the scheme are also another point to have into account: Gibbs oscillations usually occur in the proximity of data with strong gradients, possibly obtained from the sampling of discontinuous data. These oscillations that appear in the limit function are undesirable.

In the past years, several attempts have been done in order to improve the results of linear subdivision algorithms. Some of the mentioned attempts are related to the modification of linear schemes in order to introduce data dependent non linearities. In these cases, the subdivision rule also becomes data dependent and the stability is not anymore a consequence of the convergence of the scheme. Thus, it is a must to introduce a stability property in order to assure that the nonlinear scheme is linearly affected by perturbations of the data.

On the other hand, mixed finite element methods based on surface subdivision technology have been used to construct high-order smooth surfaces with specified boundary conditions [4, 7, 10, 11, 12]. Moreover, subdivision surfaces are compatible with NURBS. The geometry models can be refined to arrive at a satisfactory accuracy of the numerical simulation where the subdivision schemes are simple, efficient and can be applied to meshes with arbitrary topology.

Isogeometric analysis based on Catmull-Clark solids was first investigated in [5]. In

 ISBN: 978-84-617-8694-7

[10], the authors present a make-up approach for isogeometric analysis where they utilize the finite element function space induced from the limit form of the Catmull-Clark surface subdivision to uniformly describe the geometric domain and perform numerical simulation on it. It is compatible with NURBS as the standard in CAD system where the boundary of geometry is modelled as piecewise cubic B-spline curves. The advantage of this strategy admits quadrilateral meshes of arbitrary topology and any-shaped boundary. The computational domains considered in this paper are planar geometries. They develop the approximation character of Catmull-Clark surface subdivision function, which provides the mathematical support for the isogeometric analysis based on Extended Catmull-Clark surface subdivision.

Acknowledgements

Research supported in part by Programa de Apoyo a la investigación de la fundación Séneca-Agencia de Ciencia y Tecnología de la Región de Murcia 19374/PI/14 and MTM2015-64382-P (MINECO/FEDER).

References

[1] Bézier P., Définition numérique des courbes et surfaces, parts I et II. Automatisme 1967; 12:17-21.

[2] Birk L., Introduction to nonlinear programming. In: Course notes, 39th WEGEMT (Western European Graduate Education in Marine Technology) summer school. 2003., p. 53-82.

[3] BirkL. and Harries S. (Eds.), OPTIMISTIC-optimization in marine design. In: Course notes, 39th WEGEMT(Western European Graduate Education in Marine Technology) summer school. 2003.

[4] Coons S.A., Surfaces for computer-aided design of space figures. Unpublished notes, M.I.T.(Massachusett Institute of Technology), Mechanical Eng. Dept., Cambridge,Massachusett, January 1964.

[5] Hoschek J. and Lasser D., Fundamentals of computer aided geometric design. Wellesley (Massachusett): A.K. Peters Publ; 1996.

[6] http://www.aerohydro.com/products/marine/multisurf.htm

[7] http://www.formsys.com/maxsurf

 ISBN: 978-84-617-8694-7

[8] Hughes O., Ship structural design: A rationally-based, computer-aided, optimization approach. Jersey City: SNAME (Society of Naval Architects and Marine Engineers) ; 1988.

[9] Laurent P.J., Sablonnière P., Pierre Bézier, an engineer and a mathematician. Computer Aided Geometric Design 18 (2001), 609-617.

[10] Pan Q., Xub G., Xuc G. and Zhangd Y. Isogeometric Analysis Based on Extended Catmull-Clark Subdivision, Report No. ICMSEC-15-01 March 2015.

[11] Piegl L. and Tiller W., Curve and surface constructions using rational B-splines. Comput. Aided Design 1987; 19:485-98.

[12] Piegl L. and Tiller W., NURBS, Springer Verlag, 1997.

[13] Riesenfeld R., Applications of B-spline approximation to geometric problems of computer-aided design. Ph.D. thesis. Dept. of Computer Science, Syracuse University; 1973.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

A variational approach for chemical kinetics: A case study

S. Amat[1], M. J. Legaz[2] and J. Ruiz[3]

[1] *Departamento de Matemática Aplicada y Estadística., Universidad Politécnica de Cartagena (Spain)*

[2] *Departmento de Construcciones Navales, Universidad de Cádiz (Spain)*

[3] *Department of Physics and Mathematics, Universidad de Alcalá (Spain)*

emails: sergio.amat@upct.es, mjlegazalmansa@gmail.com, juan.ruiza@uah.es

Abstract

The solution of chemical kinetics is one of the most computationally intensive tasks in atmospheric chemical transport simulations. Due to the stiff nature of the system, implicit time stepping algorithms which repeatedly solve linear systems of equations are necessary. In some recent works, we have proposed a variational approximation of stiff systems of differential equations based on an analysis of a certain error functional associated. By using standard descent schemes, the procedure can never get stuck in local minima, but will always and steadily decrease the error until getting to the original solution. We show the efficiency of this approach in some standard chemical kinetics problems found in the literature.

Key words: Variational methods, stiff problems, chemical kinetics, implicit Runge-Kutta, variable step implementation.

1 Introduction

A typical property of chemical calculations is that different reactions occur at dramatically different time scales, with some species achieving quasi-steady-state solutions rather quickly, while others can evolve very slowly. For this reason, the solution of chemical kinetics is one of the most computationally intensive tasks in atmospheric chemical transport simulations. Due to the stiff nature of the system, implicit time stepping algorithms which repeatedly solve linear systems of equations are necessary.

 ISBN: 978-84-617-8694-7

Implicit Runge-Kutta schemes are widely used in mathematics and engineering to numerically solve stiff differential equations. These methods require the solution to a (possibly large) nonlinear system of equations at every time step, which in the past was deemed prohibitively costly. The increasing ability for scientists to solve large-scale nonlinear systems with relative efficiency, thanks to advances in numerical methods and the growth in computing power, has increased the use of these methods.

Every integration method is associated with a step-size h for the integration. If h is too large or too small, the efficiency of the scheme is relatively low. For this reason, variable step implementations are usually considered.

On the other hand, in some previous works [1, 2, 3, 4, 5], we have proposed a new variational approach for the approximation of different types of differential equations. The approach is based on an analysis of a certain error functional associated, in a natural way, with the original problem. The performance is very good due to the fact that we only need to approximate linear problems for which one can use very robust methods. In particular, the problem of having good initial guesses in the application of Newton-type methods, in order to approximate the associated nonlinear Runge-Kutta equations, is avoided.

In this paper, we show the efficiency of this approach in some standard chemical kinetics problems found in the literature.

2 Our basic variational approach: an overview

Our variational approach for the treatment of ODEs is based on the analysis of a certain error functional of the form

$$E(\mathbf{x}) = \frac{1}{2}\int_0^T |\mathbf{x}'(t) - \mathbf{f}(\mathbf{x}(t))|^2\, dt,$$

to be minimized among the absolutely continuous paths $\mathbf{x} : (0,T) \to \mathbf{R}^N$ with $\mathbf{x}(0) = \mathbf{x}_0$. This error functional is associated, in a natural way, with the Cauchy problem

$$\mathbf{x}'(t) = \mathbf{f}(\mathbf{x}(t)) \text{ in } (0,T), \quad \mathbf{x}(0) = \mathbf{x}_0. \tag{1}$$

The following claim is easy to check. It is one of the most appealing features of our viewpoint.

Proposition 1 ([2]) *Let $\overline{\mathbf{x}}$ be a critical point for the error E. Then $\overline{\mathbf{x}}$ is the solution of the Cauchy problem (1).*

On the other hand, suppose we start with an initial crude approximation $\mathbf{x}_{(0)}(\equiv \mathbf{x})$ to the solution of our basic problem (1). We would like to improve this approximation in such a way that the error is significantly decreased.

 ISBN: 978-84-617-8694-7

It is straightforward to find the Gâteaux derivative of E at a given feasible $\mathbf{x}$ in the direction $\mathbf{y}$ with $\mathbf{y}(0) = \mathbf{0}$. Namely

$$\langle E'(\mathbf{x}), \mathbf{y} \rangle = \int_0^T (\mathbf{x}'(t) - \mathbf{f}(\mathbf{x}(t))) \cdot (\mathbf{y}'(t) - \nabla \mathbf{f}(\mathbf{x}(t))\mathbf{y}(t))\, dt.$$

This expression suggests to select $\mathbf{y}$ as a solution of the linear problem

$$\mathbf{y}'(t) - \nabla \mathbf{f}(\mathbf{x}(t))\mathbf{y}(t) = \mathbf{f}(\mathbf{x}(t)) - \mathbf{x}'(t) \text{ in } (0,T), \quad \mathbf{y}(0) = \mathbf{0}.$$

In this way, it is easy to calculate that $\langle E'(\mathbf{x}), \mathbf{y} \rangle = -2E(\mathbf{x})$, and so the (local) decrease of the error is of the size $E(\mathbf{x})$. Finding $\mathbf{y}$ requires solving the above linear problem iteratively.

Theorem 1 *If the mapping* $\mathbf{f}$ *is smooth enough, and the horizon* T *is sufficiently small, the iterative procedure* $\mathbf{x}^{(j)}(t) = \mathbf{x}^{(j-1)}(t) + \mathbf{y}^{(j)}(t)$ *where*

$$(\mathbf{y}^{(j)})'(t) - \nabla \mathbf{f}(\mathbf{x}^{(j)}(t))\mathbf{y}^{(j)}(t) = \mathbf{f}(\mathbf{x}^{(j)}(t)) - (\mathbf{x}^{(j)})'(t) \text{ in } (0,T), \quad \mathbf{y}(0) = \mathbf{0},$$

converges strongly in $H^1(0,T;\mathbf{R}^N)$*, and in* $L^\infty(0,T;\mathbf{R}^N)$ *to the unique solution of problem (1).*

Note that this result can be used successively in an arbitrary time interval $[0,T]$ by dividing it in a sufficient big number of subintervals.

This strategy can be used, and extended for more general situations: differential-algebraic equations, delay differential-equations or Hamiltonian systems [1, 2, 3, 4, 5].

An implicit fixed-step solver computes the state at the next time step as an implicit function of the state at the current time step, and the state derivative at the next time step. The variable-step solvers dynamically vary the step size during the simulation. These solvers increase or reduce the step size using its local error control to achieve the tolerances that you specify. Computing the step size at each time step adds to the computational overhead, but can reduce the total number of steps, and the simulation time required to maintain a specified level of accuracy.

For a stiff problem, solutions can change on a time scale that is very small as compared to the interval of integration, while the solution of interest changes on a much longer time scale. Methods that are not designed for stiff problems are ineffective on intervals where the solution changes slowly because these methods use time steps small enough to resolve the fastest possible change.

If we denote by $\mathbf{y}(t_n, t_{n-1}, \mathbf{y}_{n-1})$ the solution of a given differential equation (system) $\mathbf{F}(t, \mathbf{y}, \mathbf{y}') = 0$, then a method of order p will verify that locally its error has the form

$$||e_n|| = ||\mathbf{y}(t_n, t_{n-1}, \mathbf{y}_{n-1}) - \mathbf{y}_n|| = C\ h^{p+1} + O(h^{p+2}).$$

 ISBN: 978-84-617-8694-7

In order to estimate C, we can use extrapolation techniques. Another possibility is to consider another scheme $\tilde{\mathbf{y}}_n$ with bigger order q starting from $\mathbf{y}_{n-1}$. In this case we have

$$||\tilde{e}_n|| = ||\mathbf{y}(t_n, t_{n-1}, \mathbf{y}_{n-1}) - \tilde{\mathbf{y}}_n|| = \tilde{C}\ h^{q+1} + O(h^{q+2}),$$

then

$$||e_n|| = ||\mathbf{y}_n - \tilde{\mathbf{y}}_n|| + O(h^{p+2}).$$

With an estimation of the error, the variable step codes select the new step such that the error is smaller than a prescribed tolerance TOL.

The desired error associated to a h_* will be

$$||e_n(h^{'})|| = Ch_*^{p+1} = TOL$$

and the real error

$$||e_n(h)|| = Ch^{p+1} = ||\mathbf{y}_n - \tilde{\mathbf{y}}_n||,$$

dividing both equations we obtain

$$h_* = h\ \sqrt[p+1]{\frac{TOL}{||\mathbf{y}_n - \tilde{\mathbf{y}}_n||}}.$$

Remark 1 *As every implicit integration scheme has a global error inherent to the scheme, in [6] the authors choose the total number of computations (in order to solve the inner nonlinear Runge-Kutta equations by Newton-type methods) in order to achieve a prescribed global error as a measure of efficiency of the integration scheme. This efficiency function is the critical component in making these methods variable step-size methods. This approach does not sense in our case since we use Runge-Kutta methods only in linear problems. In any case, we recommended to see [6] and its references.*

In our case, we can use the norm of the direction $\mathbf{y}^{(j)}$. These numbers are related to the local error (see the following corollaries). That is, without any extra computation (in fact this norm is used in the stop criterion, we stop in $j_{\max}$ when $|\mathbf{y}^{(j_{\max})}| \leq TOL$), we will be able to implement a variable step strategy.

The estimation of the error and the new h_* has been introduced from a mathematical point of view. However, in order to be effective in practice we need some control strategies.

Let $h_{n+1} = t_{n+1} - t_n$ be the discretization parameters. We introduce the following equation

$$h_{n+1} = \sigma h_n\ \sqrt[p+1]{\frac{TOL}{|\mathbf{y}^{(j_{\max})}(t_n)|}},$$

where σ is a security factor smaller than 1.

 ISBN: 978-84-617-8694-7

If in a step $|\mathbf{y}^{(j_{\max})}(t_n)| > TOL$ we reject $\mathbf{y}_n$, and compute a new iteration with

$$h_n = \sigma h_n \sqrt[p+1]{\frac{TOL}{|\mathbf{y}^{(j_{\max})}(t_n)|}}.$$

Finally, it is important to add some computational restrictions. Namely

$$h_{\min} \leq h_n \leq h_{\max}$$

and

$$\omega \leq \frac{h_{n+1}}{h_n} \leq \Omega$$

for some given positive constants $h_{\min}$, $h_{\max}$, ω and Ω.

Some classical examples are:

$$\begin{aligned} h_{\min} &= 10^{-6}, \\ h_{\max} &= 1, \\ \omega &= \frac{1}{5}, \\ \Omega &= 5. \end{aligned}$$

3 A stiff problem: Chapman atmosphere

This model represents the Chapman mechanism for the generation of the ozone and the oxygen singlet. In this example, the concentration of the oxygen $y_3 = [O_2]$ will be held constant. It is a severe test for a stiff ODE package governed by the following equations:

$$\begin{aligned} y_1'(t) &= 2k_3(t)y_3 + k_4(t)y_2(t) - (k_1 y_3 + k_2 y_2(t))y_1(t), \\ y_2'(t) &= k_1 y_1(t) y_3 - (k_2 y_1(t) + k_4(t))y_2(t), \end{aligned}$$

with $y_3 = 3.7 \times 10^{16}$, $k_1 = 1.63 \times 10^{-16}$, $k_2 = 4.66 \times 10^{-16}$,

$$k_i(t) = \begin{cases} \exp(\frac{a_i}{\sin(\omega t)}), & \text{if } \sin(\omega t) > 0 \\ 0, & \text{otherwise} \end{cases}$$

for $i = 3, 4$, with $a_3 = 22.62$, $a_4 = 7.601$ and $\omega = \frac{\pi}{43200}$. The constant 43200 is 12 h measured in seconds. The initial conditions are $y_1(0) = 10^6$ and $y_2(0) = 10^{12}$.

This problem has important features like:

- The Jacobian matrix is not a constant.

 ISBN: 978-84-617-8694-7

- The diurnal effect is present.
- The oscillations are fast.
- The time interval used is fairly long, $0 \leq t \leq 8.64\ 10^5$, or 10 days.

We obtain a good approximation, see Figure 1. Note that $y_2 = [0_3]$ looks like a staircase with a rise at midday every day and $y_1 = [O]$ looks like a spike with its amplitude increases each day.

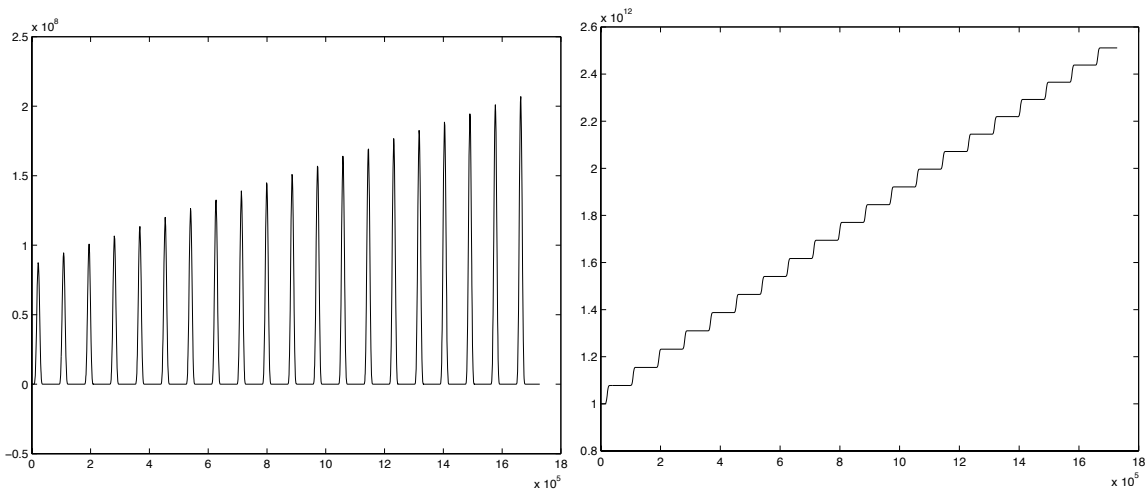


Figure 1: Chapman atmosphere approximated via our variational approximation. Left: first component. Right: second component.

Acknowledgements

Research supported in part by Programa de Apoyo a la investigación de la fundación Séneca-Agencia de Ciencia y Tecnología de la Región de Murcia 19374/PI/14 and MTM2015-64382-P (MINECO/FEDER).

References

[1] Amat, S., Pedregal, P., A variational approach to implicit ODEs and differential inclusions, ESAIM-COCV, **15**(1), (2009), 139-148.

 ISBN: 978-84-617-8694-7

[2] Amat, Sergio; Pedregal, Pablo On a variational approach for the analysis and numerical simulation of ODEs. Discrete Contin. Dyn. Syst. 33 (2013), no. 4, 12751291.

[3] Amat, S., López D.J.,Pedregal, P., Numerical approximation to ODEs using a variational approach I: The basic framework, to appear in Optimization.

[4] Amat, S.; Legaz, M. J.; Pedregal, P. Linearizing stiff delay differential equations. Appl. Math. Inf. Sci. 7 (2013), no. 1, 229232.

[5] Amat, S.; Legaz, M. J.; Pedregal, P. On a Newton-type method for differential-algebraic equations. J. Appl. Math. 2012, Art. ID 718608, 15 pp

[6] Holsapple R., Iyer R., Doman D., Variable step- size selection methods for implicit integration schemes for ODEs, International Journal of Numerical Analysis and Modeling, **4** 2, (2007), 210-240.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

Approximation of polynomial Hamiltonian systems using an alternative variational technique

S. Amat[1], M. J. Legaz[2] and J. Ruiz[3]

[1] *Departamento de Matemática Aplicada y Estadística., Universidad Politécnica de Cartagena (Spain)*

[2] *Departmento de Construcciones Navales, Universidad de Cádiz (Spain)*

[3] *Department of Physics and Mathematics, Universidad de Alcalá (Spain)*

emails: sergio.amat@upct.es, mjlegazalmansa@gmail.com, juan.ruiza@uah.es

Abstract

Hamiltonian systems appear in numerous areas and have many applications. Many interesting Hamiltonian systems arising from different fields of study are defined by polynomial Hamiltonian functions. Methods preserving the qualitative properties of the exact solution are essential. In general, the integration of these systems requires the use of geometric integrators. In this paper, we propose the use of a variational approach for models which are formulated naturally as Hamiltonian systems. Our variational method for polynomial Hamiltonian systems does not require the use of any integrator and moreover it penalize the not preservation of the energy. We indicate its most basic properties, and test its numerical performance in some examples.

Key words: Polynomial Hamiltonian systems, symplecticity, energy preserving, variational approach.

1 Introduction

It is well-known that numerical methods, such as the ordinary Runge-Kutta schemes, are not particularly efficient in integrating Hamiltonian systems, because Hamiltonian systems are not generic in the set of all dynamic systems. They are not structurally stable against non-Hamiltonian perturbations. Numerical solution of Hamiltonian systems is frequently carried out by symplectic integrators due to their good performance in moderate and long-time integration, see [10, 12, 14, 15, 19]. Symplectic numerical methods belong to the family

 ISBN: 978-84-617-8694-7

of Geometric Numerical Integrators, which preserve important qualitative and geometric properties of the underlying differential system, and are arguably the most popular methods in this class. Certain qualitative properties of the evolution, like symplecticity, are preserved and, in general they exhibit smaller error growth along the numerical trajectory.

Some pioneering works on symplectic integrations are due to Vogelaere [21], Ruth [17], and Feng Kang [9]. The derivation of higher-order methods is covered by several approaches such as composition methods, classical Runge-Kutta methods (RK) as well as partitioned Runge-Kutta (PRK) methods, and methods based on generating functions. The systematic study of symplectic Runge-Kutta methods started around 1988, and a complete characterization has been found independently by Lasagni [13] (using the approach of generating functions), and by Sanz-Serna [18] and Suris[20] (using the ideas of the classical papers of Burrage and Butcher [7] and Crouzeix [8] on algebraic stability).

Nowadays, it is well-known that certain implicit RK methods of Radau type (generalizing the implicit Euler method) are useful in the context of systems with strong dissipation, like electronic circuits or chemical reaction dynamics. Partitioned Runge- Kutta (PRK) methods are another way to approximating the solution trajectory which it is based on using different approximation formulas for different components of the solution. (use different sets of quadrature rules for each subset of the variables). The starting point of generating function (GF) theory was the discovery of Hamilton that the motion of the system is completely described by a characteristic function S, and that S is the solution of a partial differential equation, now called the Hamilton-Jacobi differential equation. It was noticed later, especially by Siegel (Siegel and Moser 1971), that such a function S is directly connected to any symplectic map. It was called generating function. See [10, 14].

Another important point should be taken into account regarding Hamiltonian systems, even with symplectic maps, and that is the lack of energy conservation in the map. It would seem to be an obvious goal for Hamiltonian integration methods both to preserve the symplectic structure and to conserve the energy, but it was shown that this was in general impossible. Thus a symplectic map which only approximates a Hamiltonian cannot conserve energy [22]. Recently, some research has been carried out about energy-preserving symplectic methods based on the key tool line integral associated with conservative vector fields, as well as its discrete version. See,[5, 6].

Polynomial Hamiltonian systems appear frequently in physics and it is important to build good integrators for numerical simulations. In general, it is not possible to find analytical expressions for, so numerical methods are required. Of course, methods preserving the qualitative properties of the exact solution are essential in order to have a good picture of the stability regions. Basically, we have to consider symplectic integrators.

In [1], we introduce a new variational approach for models which are formulated naturally as conservative systems of ODEs, most importantly Hamiltonian systems. The variational method for Hamiltonian systems, which is proposed here, is in some sense symplectic

 ISBN: 978-84-617-8694-7

and energy preserving, and is based on a natural modification of the schemes introduced in [2, 3, 4].

We would like to highlight the advantages of our approach.

- A symplectic map which only approximates a Hamiltonian cannot conserve energy [22]. In our case, by definition the functional penalizes the nonpreservation of the energy, and our linearization can be approximated by well known sympletic quadrature rules. Moreover, for polynomial systems, the formula for the direction $\mathbf{Y}$ can be obtained exactly.
- In the non-stiff situation, the long-time behaviour of symplectic methods is well understood, and can be explained with the help of a backward error analysis (modified equations). In the highly oscillatory (stiff) case, this theory breaks down. However, our linearization is well understood in both cases.

2 A variational approach for Hamiltonian systems

Consider the hamiltonian dynamical system

$$\mathbf{x}'(t) = \frac{\partial \mathbf{H}}{\partial \mathbf{p}}(\mathbf{x}(t), \mathbf{p}(t)), \quad \mathbf{p}'(t) = -\frac{\partial \mathbf{H}}{\partial \mathbf{x}}(\mathbf{x}(t), \mathbf{p}(t)), \tag{1}$$

to hold in a certain time interval $(0, T)$, subject to initial boundary conditions $(\mathbf{x}(0), \mathbf{p}(0)) = (\mathbf{x}_0, \mathbf{p}_0)$. Both $\mathbf{x}$ and $\mathbf{p}$ take on values in $\mathbf{R}^N$. We are interested in setting up a method to understand, and approximate, the trajectories of such a system. In particular, we would like to focus on how the a priori knowledge of conserved quantities may help in improving our ability to approximate such system. It is well-known that the hamiltonian itself $\mathbf{H}(\mathbf{x}, \mathbf{p}) : \mathbf{R}^N \times \mathbf{R}^N \to \mathbf{R}$ is one such conserved quantity so that $\mathbf{H}(\mathbf{x}(t), \mathbf{p}(t)) = \mathbf{H}(\mathbf{x}_0, \mathbf{p}_0) \equiv \mathbf{H}_0$ for all times t in $(0, T)$.

Recently, an alternative to the analysis and numerical approximation of dynamical systems has been introduced ([2, 3, 4]), based on the minimization of the error functional

$$E(\mathbf{X}) = \int_0^T \frac{1}{2} |\mathbf{X}'(t) - \mathbf{F}(\mathbf{X}(t))|^2 \, dt$$

regarded as a measure of how far a given absolutely-continuous path $\mathbf{X}$ complying with $\mathbf{X}(0) = \mathbf{X}_0$ is from being a solution of the underlying dynamical system

$$\mathbf{X}'(t) = \mathbf{F}(\mathbf{X}(t)) \text{ in } (0, T), \quad \mathbf{X}(0) = \mathbf{X}_0. \tag{2}$$

It is elementary to realize that solutions of the system are precisely those $\mathbf{X}$ for which $E(\mathbf{X}) = 0$, and preserve the initial condition. We would like to explore one possibility of

 ISBN: 978-84-617-8694-7

taking advantage of the hamiltonian structure of the system for this variational approach. Namely, if the dimension is even $2N$, and (1) holds for some hamiltonian $\mathbf{H}$, then, as remarked above, $\mathbf{H}$ must be constant on integral curves, and so we could modify the error functional to take into account this extra information to write

$$E(\mathbf{x},\mathbf{p}) = \int_0^T \left[\frac{1}{2}|\mathbf{x}' - \frac{\partial \mathbf{H}}{\partial \mathbf{p}}(\mathbf{x},\mathbf{p})|^2 + \frac{1}{2}|\mathbf{p}' + \frac{\partial \mathbf{H}}{\partial \mathbf{x}}(\mathbf{x},\mathbf{p})|^2 + \frac{1}{2}|\mathbf{H}(\mathbf{x},\mathbf{p}) - \mathbf{H}_0|^2\right] dt. \tag{3}$$

The basis for our proposal is the following.

Proposition 2.1 *Suppose the path* $\mathbf{x}$ *is a critical point for the functional in (3), under the initial condition* $(\mathbf{x}(0),\mathbf{p}(0)) = (\mathbf{x}_0,\mathbf{p}_0)$*. Then* $\mathbf{x}$ *is the unique solution of (1).*

What Proposition 2.1 ensures is that the only critical points of the error functional (3) are the solutions of the hamiltonian system (1) itself, and so an approximation procedure based on minimizing the error functional can never get stuck in local minima, but proceed to steadily approximate the true solution of the dynamical system. Moreover, by construction, the functional penalizes the non conservation of the energy.

We focus on finding the steepest descent direction with respect to the norm

$$\int_0^T |\mathbf{Y}'(t)|^2 dt,$$

in this case, the steepest descent direction can be found as the solution of a variational problem of the following form.

Minimize $\mathbf{Y}$:

$$\frac{1}{2}\int_0^T [|\mathbf{Y}'|^2 + (\mathbf{X}' - \Omega\nabla\mathbf{H}(\mathbf{X})(\mathbf{Y}' + \Omega^T\nabla^2\mathbf{H}(\mathbf{X})\mathbf{Y}) + (\mathbf{H}(\mathbf{X}) - \mathbf{H}(\mathbf{X}(0)))\nabla\mathbf{H}(\mathbf{X})\mathbf{Y}]\, dt,$$

under $\mathbf{Y}(0) = 0$, where

$$\Omega = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}. \tag{4}$$

This is a standard quadratic variational problem whose optimal solution can be found through the system

$$-\frac{d}{dt}\left[\mathbf{Y}' + \mathbf{X}' + \Omega^T\nabla\mathbf{H}(\mathbf{X})\right] + \nabla^2\mathbf{H}(\mathbf{X})\Omega\left(\mathbf{X}' + \Omega^T\nabla\mathbf{H}(\mathbf{X})\right) + \left(\mathbf{H}(\mathbf{X}) - \mathbf{H}(\mathbf{X}_0)\right)\nabla\mathbf{H}(\mathbf{X}) = 0$$

 ISBN: 978-84-617-8694-7

in $(0,T)$, together with the end-point conditions

$$\mathbf{Y}(0) = 0, \mathbf{Y}'(T) + \mathbf{X}'(T) + \Omega^T \nabla \mathbf{H}(\mathbf{X}(T)) = 0.$$

It is remarkable that the solution of this system can be given in an fully explicit form as

$$\mathbf{Y}(t) = -\int_0^t [s\mathbf{G}(s) + \mathbf{F}(s)]\, ds - t\int_t^T \mathbf{G}(s)\, ds, \tag{5}$$

where

$$\begin{aligned} \mathbf{F} &= \mathbf{X}' + \Omega^T \nabla \mathbf{H}(\mathbf{X}) \\ \mathbf{G} &= \nabla^2 \mathbf{H}(\mathbf{X})\Omega(\mathbf{X}' + \Omega^T \nabla \mathbf{H}(\mathbf{X})) + (\mathbf{H}(\mathbf{X}) - \mathbf{H}(\mathbf{X}(0)))\nabla \mathbf{H}(\mathbf{X}). \end{aligned}$$

We can therefore establish a steepest descent strategy to decrease the error in each iteration. Note that the derivative of the error at $\mathbf{X}$ in the direction $\mathbf{Y}$ is

$$-\int_0^T |\mathbf{Y}'(t)|^2 dt$$

Then we can choose $\eta \in (0,1)$ such that

$$E(\mathbf{X} + \eta \mathbf{Y}) \leq E(\mathbf{X}) - \frac{1}{2}\int_0^T |\mathbf{Y}'(t)|^2 dt.$$

Moreover, it is possible to check uniform convergence of this procedure for smooth Hamiltonian systems (see [16]). The following statement is a direct adaptation of similar results in [16].

Theorem 2.2 *The iterative procedure $\mathbf{X}^{(j)} = \mathbf{X}^{(j-1)} + \eta_j \mathbf{Y}^{(j)}$, starting from arbitrary feasible $\mathbf{X}^{(0)}$, converges strongly in $H^1(0,T)$ to the unique solution of any sufficiently smooth Hamiltonian system.*

3 Numerical results

The iterative numerical procedure is easily implementable.

1. Start with an initial approximation $\mathbf{X}^0(t)$ compatible with the initial conditions, for instance $\mathbf{X}^{(0)}(t) = \mathbf{X}_0$.
2. Assume we have approximation $(\mathbf{X}^{(j)})(t)$ in $[0,T]$.
3. Compute its derivative $(\mathbf{X}^{(j)})'(t)$.
4. Define $\mathbf{F}$ and $\mathbf{G}$ through the formulas

$$\begin{aligned} \mathbf{F}^{(j)} &= (\mathbf{X}^{(j)})' + \Omega^T \nabla \mathbf{H}(\mathbf{X}^{(j)}), \\ \mathbf{G}^{(j)} &= \nabla^2 \mathbf{H}(\mathbf{X}^{(j)})\Omega((\mathbf{X}^{(j)})' + \Omega^T \nabla \mathbf{H}(\mathbf{X}^{(j)})) \\ &\quad + (\mathbf{H}(\mathbf{X}^{(j)}) - \mathbf{H}(\mathbf{X}(0)))\nabla \mathbf{H}(\mathbf{X}^{(j)}). \end{aligned}$$

 ISBN: 978-84-617-8694-7

5. Approximate the mapping

$$\mathbf{Y}^{(j)}(t) = -\int_0^t \left[s\mathbf{G}^{(j)}(s) + \mathbf{F}^{(j)}(s)\right] ds - t\int_t^T \mathbf{G}^{(j)}(s)\, ds, \tag{6}$$

using (symplectic) quadrature formulas. In the case of polynomial Hamiltonians perform an exact integration-

6. Update $\mathbf{X}^{(j)}$ to $\mathbf{X}^{(j+1)}$ by using the formula

$$\mathbf{X}^{(j+1)}(t) = \mathbf{X}^{(j)}(t) + \eta_j \mathbf{Y}^{(j)}(t).$$

7. Iterate (3), (4), (5) and (6) until numerical convergence.

Many interesting Hamiltonian systems arising from different fields of study are defined by polynomial Hamiltonian functions. For example the Fermi-Pasta-Ulam problem and the polynomial pendulum oscillator of degree k, obtained from the Hamiltonian function of the nonlinear pendulum equation $H(p,q) = 1/2p^2 + 1 - cosq$, by retaining a finite number of terms in the Taylor expansion of the cosine.

It is well known that symplectic RK-methods only conserve quadratic Hamiltonian functions but, in general, they fail to yield conservation for higher degree, and so do symmetric methods.

We consider the polynomial problem:

$$H(p,q) = \frac{1}{3}p^3 - \frac{1}{2}p - \frac{1}{3}q^3 + \frac{1}{2}q^2 + \frac{1}{6} \tag{7}$$

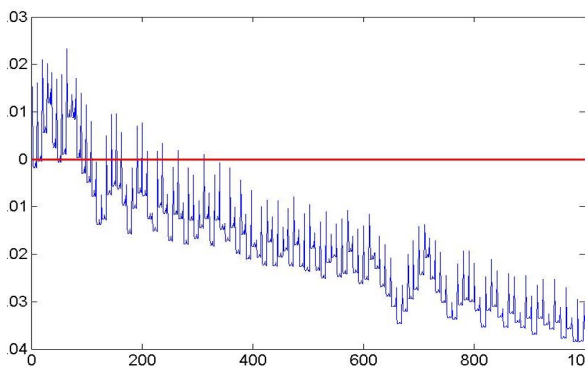


Figure 1: Energy function associated to the integration by Lobatto IIIB.

 ISBN: 978-84-617-8694-7

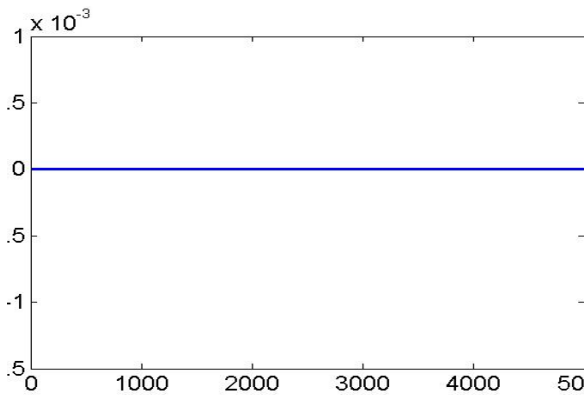


Figure 2: Energy function associated to the integration by our variational approach.

Energy function H evaluated over the numerical solution obtained by the Lobatto IIIB method is given in Figure 1 and by our variational approach in Figure 2 applied to problem (7).

It is possible to modified the original RK method in order to preserve energy but in a most sophisticated way [11]. Moreover, as we said before, in the highly oscillatory (stiff) case, the classical theory for RK breaks down. However, our linearization is well understood in all the cases.

Acknowledgements

Research supported in part by Programa de Apoyo a la investigación de la fundación Séneca-Agencia de Ciencia y Tecnología de la Región de Murcia 19374/PI/14 and MTM2015-64382-P (MINECO/FEDER).

References

[1] S. Amat, M.J. Legaz, P. Pedregal, Approximation of Hamiltonian Systems using an Alternative Variational Technique, Appl. Math. Inf. Sci. 9, No. 5, 2389-2394 (2015).

[2] S. Amat, P. Pedregal, A variational approach to implicit ODEs and differential inclusions, ESAIM-COCV **15**, 139-148 (2009).

[3] S. Amat, P. Pedregal, On a variational approach for the analysis and numerical simulation of ODEs. Discrete Contin. Dyn. Syst. **33**, 1275-1291 (2013).

 ISBN: 978-84-617-8694-7

[4] S. Amat, D. J. López , P. Pedregal, Numerical approximation to ODEs using a variational approach I: The basic framework, Optimization **63**, 337-358 (2014).

[5] L. Brugnano, F. Iavernaro, D. Trigiante, On the existence of energy preserving symplectic integrators based upon Gauss collocation formulae, Math. N.A., 2010.

[6] L. Brugnano, F. Iavernaro, D. Trigiante, A two step, fourth order, nearly-linear method with energy preserving properties. Math. N. A., 2011.

[7] K. Burrage, J. C. Butcher, Stability criteria for implicit Runge-Kutta methods, SIAM J. Numer. Anal. **16**, 46-57 (1979).

[8] M. Crouzeix, Sur la B-stabilité des methods de Runge-Kutta, Numer. Math. **32**, 75-82 (1979).

[9] K. Feng, On difference schemes and symplectic geometry, Proceedings of the 5-th Intern. Symposium on differential geometry and differential equations, Beijing, 42-58, (1985).

[10] E. Hairer, C. Lubich, G. Wanner, Geometric Numerical Integration: Structure-Preserving Algorithms for Ordinary Differential Equations, Springer Series in Computational Mathematics, Springer, 2006.

[11] F. Iavernaro and B. Pace, s-stage trapezoidal methods for the conservation of Hamiltonian functions of polynomial type, in: American Institute of Phisycs, T. Simos et al. eds., 603-606 (2007).

[12] A. Iserles, A First Course in the Numerical Analysis of Differential Equations. 2nd edn. Cambridge University Press, Cambridge, 2008.

[13] F. M. Lasagni, Canonical Runge-Kutta methods, ZAMP **39**, 952-953 (1988).

[14] B. Leimkuhler, S. Reich, Simulating Hamiltonian Dynamics. Cambridge University Press, Cambridge, 2004.

[15] R. I. McLachlan, R. G. W. Quispel, Splitting methods. Acta Numerica **11**, 341-434 (2002).

[16] P. Pedregal, A variational approach to dynamical systems, and its numerical simulation, Numer. Funct. Anal. Opt., **31**, 1532-2467 (2010).

[17] R. D. Ruth, A canonical integration technique, IEEE Trans. Nuclear Science **30**, 2669-2671 (1983).

[18] J. M. Sanz-Serna, Runge-Kutta schemes for Hamiltonian systems, BIT **28**, 877-883 (1988).

 ISBN: 978-84-617-8694-7

[19] J. M. Sanz-Serna, M. P. Calvo, Numerical Hamiltonian Problems. Chapman and Hall, London, 1994.

[20] Y. B. Suris, On the conservation of the symplectic structure in the numerical solution of Hamiltonian systems (in Russian), In: Numerical Solution of Ordinary Differential Equations, ed. S.S. Filippov, Keldysh Institute of Applied Mathematics, USSR Academy of Sciences, Moscow, 148-160 (1988).

[21] R. Vogelaere, Methods of integration which preserve the contact transformation property of the Hamiltonian equations, Report No. 4, Dept. Math., Univ. Of Notre Dame, Notre Dame, Ind., 1956.

[22] G. Zhong, J. Marsden, Lie-Poisson Hamilton-Jacobi theory and Lie-Poisson integrators, Phys. Lett. **133**, 134 (1988).

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

Stability analysis of a parametric family of seventh-order iterative methods for solving nonlinear problems

Abdolreza Amiri[1], Alicia Cordero[1], M. Taghi Darvishi[2] and Juan R. Torregrosa[2]

[1] *Department of Mathematics, Razi University, Iran*

[2] *Instituto de Matemáticas Multidisciplinar, Universitat Politècnica de València, Spain*

emails: amiriabdolreza@ymail.com, acordero@mat.upv.es, darvishimt@yahoo.com, jrtorre@mat.upv.es

Abstract

In this paper, a parametric family of seventh-order of iterative method to solve systems of nonlinear equations is presented. Its local convergence is studied and quadratic polynomials are used to investigate its dynamical behavior. The study the fixed and critical points of the rational function associated to this class allows us to obtain regions of the complex plane where the method is stable. By depicting parameter planes and dynamical planes we obtain complementary information of the analytical results.

Key words: Nonlinear system of equations, iterative method, basin of attraction, dynamical plane, stability

1 Introduction

The problem of finding roots of nonlinear system $F(x) = 0$, where $F : D \subseteq \mathbb{R}^n \longrightarrow \mathbb{R}^n$ arises in different areas of scientific computing and engineering. The solution of this problem is usually obtained by a fixed point function $\overline{G} : D \subseteq \mathbb{R}^n \longrightarrow \mathbb{R}^n$ associated to a fixed point iteration scheme. There are many iterative schemes to solve this problem. The most common root-finding method for solving systems of nonlinear equations is the second order Newton's scheme. But in recent years, some researchers have used new iterative methods as an alternative to classical one that have higher order and better efficiency (see, for example [5, 8]). However, the analysis of the stability of these iterative methods when $n > 1$ is so complicated and most of the times is impossible, but the study of the behavior of the

 ISBN: 978-84-617-8694-7

iterative method in scalar case that is, for $n = 1$, in the complex plane is an interesting field of study(see, for instance [1, 7]). Specially the advent of computers in last decades, made it practically possible to study the structure of the dynamical and parameter planes of iterative methods closely, since large amount of computational power is need to obtain their precise shape, that can be easily performed in computers.

The main aim of this analysis is finding the regions in the complex plane where our function shows better stability behavior when converges to the zeros of the function. But even in the scalar case, finding stable regions for a high order iterative method is not easy. High order iterative methods, even for simple scalar nonlinear function $f(z)$, usually results in a high degree fixed point operator, since the key of stability analysis is the study of the fixed point operator.

In this paper we propose the following family of seventh-order iterative methods to solve systems of nonlinear equations, whose iterative expression is

$$\begin{aligned} y^{(k)} &= x^{(k)} - \left[F'(x^{(k)})\right]^{-1} F(x^{(k)}), \\ z^{(k)} &= y^{(k)} - \frac{1}{\beta}\left[F'(x^{(k)})\right]^{-1} F(y^{(k)}), \\ w^{(k)} &= z^{(k)} - \left[F'(x^{(k)})\right]^{-1} \left((2 - 1/\beta - \beta)F(y^{(k)}) + \beta F(z^k)\right), \\ x^{(k+1)} &= w^{(k)} - G(t^{(k)}) \left[F'(x^{(k)})\right]^{-1} F(w^{(k)}), \end{aligned} \tag{1}$$

where $t^{(k)} = I - \frac{1}{\beta}\left[F'(x^{(k)})\right]^{-1}[y^{(k)}, z^{(k)}; F]$ and $G : \mathbb{R}^{n\times n} \longrightarrow \mathbb{R}^{n\times n}$ is a matrix weight function that is chosen in order to obtain the seventh-order of convergence. Let us remark that the first three steps of 1 correspond to family M4 of parametric methods. It was shown in [3] to have fourth-order of convergence for all non-zero values of β.

In this paper, we are going to analyze the dynamical behavior of class (1) on scalar functions. This analysis will be made on quadratic polynomials, as they are the simplest nonlinear functions. This analysis will give us important information about the stability of the family in terms of the value of the parameter.

The rational function associated with a subclass of (1) on the quadratic polynomial $p(z) = (z - a)(z - b)$ is used in the following and denoted by $O_p(z)$. The obtained results can be extrapolated to more complicated nonlinear function.

The rest of paper is organized as follows: in Section 2 the local convergence of the iterative class (1) is presented. In Section 3, we obtain the operator $O_p(z)$ of a particular subfamily of (1) on $p(z)$ and then the stability of fixed points of the operator is studied. Critical points of $O_p(z)$ are investigated in Section 4. Finally, Section 5 is devoted to study the parameter and dynamic planes for some members of the subclass.

 ISBN: 978-84-617-8694-7

2 Convergence analysis

In this section, we present a local convergence theorem for family (1). In order to get this result, we introduce the following notation.

Let us denote by $X = \mathbb{R}^{n\times n}$ the space of all $n \times n$ real matrices; the weight function in this context is $G : X \to X$ such that

(i) $G'(u)(v) = G_1uv$, being G' the first derivative of G, $G' : X \to \mathcal{L}(X)$, $G_1 \in \mathbb{R}$ and $\mathcal{L}(X)$ denotes the space of linear mappings from X to itself.

(ii) $G''(u, v)(w) = G_2uvw$, being G'' the second derivative of G, $G'' : X \times X \to \mathcal{L}(X)$ and $G_2 \in \mathbb{R}$.

Then, the Taylor expansion of G around the identity matrix gives

$$G(\eta^{(k)}) \approx G\left(\frac{\beta-1}{\beta}I\right) + \left(t^{(k)} - \frac{\beta-1}{\beta}I\right)G_1 + \frac{1}{2}\left(t^{(k)} - \frac{\beta-1}{\beta}I\right)^2 G_2.$$

The following result establishes that this class have order of convergence at least seven.

Theorem 1 *Let $F : D \subseteq \mathbb{R}^n \longrightarrow \mathbb{R}^n$ be a sufficiently differentiable function in an open convex set D and let $\overline{x} \in D$ be a solution of the system of nonlinear equations $F(x) = 0$. We suppose that $F'(x)$ is continuous and nonsingular at $\overline{x}$. Then, the sequence $\{x^{(k)}\}_{k\geq 0}$, obtained by expression (1), converges to $\overline{x}$ with order of convergence at least seven when $G_0 = G\left(\frac{\beta-1}{\beta}I\right) = I$, $G_1 = \beta$ and $G_2 = \frac{1}{2}(-\beta + 6\beta^2)$. Moreover, the special case of $\beta = \frac{1}{5}$ has order of convergence eight.*

One could use different weight functions $G(t)$, satisfying the conditions of previous result, for designing specific families of iterative schemes. For example,

$$G(t^{(k)}) = I + \beta\left(t^{(k)} - \frac{\beta-1}{\beta}I\right) + \frac{1}{4}(-\beta + 6\beta^2)\left(t^{(k)} - \frac{\beta-1}{\beta}I\right)^2 \tag{2}$$

or the rational function,

$$G(t^{(k)}) = I + \frac{10\beta-1}{(1-6\beta)\beta}[t^{(k)}]^{-1}\left[\frac{1-6\beta}{10\beta-1}I - \frac{(1-2\beta)\beta}{10\beta-1}t^{(k)}\right]$$

The subclass corresponding to the first function $G(t^{(k)})$ is denoted by M7.

 ISBN: 978-84-617-8694-7

3 Fixed points of operator $O_p(z)$

In this section, we study the fixed points of operator $O_p(z)$ as a function of β and in the next section, we investigate its critical points. To get this aim, we are going to recall some dynamical concepts of complex dynamics (see [2]) that we use in this work.

Given a rational function $R : \hat{\mathbb{C}} \to \hat{\mathbb{C}}$, where $\hat{\mathbb{C}}$ is the Riemann sphere, the *orbit of a point* $z_0 \in \hat{\mathbb{C}}$ is defined as:

$$\{z_0,\, R(z_0),\, R^2(z_0), ..., R^n(z_0), ...\}.$$

We analyze the phase plane of the map R by classifying the starting points from the asymptotic behavior of their orbits. A $z_0 \in \hat{\mathbb{C}}$ is called a *fixed point* if $R(z_0) = z_0$. A *periodic point* z_0 of period $p > 1$ is a point such that $R^p(z_0) = z_0$ and $R^k(z_0) \neq z_0$, for $k < p$. A *pre-periodic point* is a point z_0 that is not periodic but there exists a $k > 0$ such that $R^k(z_0)$ is periodic. A *critical point* z_0 is a point where the derivative of the rational function vanishes, $R'(z_0) = 0$. Moreover, a fixed point z_0 is called *attractor* if $|R'(z_0)| < 1$, *superattractor* if $|R'(z_0)| = 0$, *repulsor* if $|R'(z_0)| > 1$ and *parabolic* if $|R'(z_0)| = 1$. The fixed points that do not correspond to the roots of the polynomial $p(z)$ are called *strange fixed points*.

The basin of attraction of an attractor α is defined as:

$$\mathcal{A}(\alpha) = \{z_0 \in \hat{\mathbb{C}} \ :\ R^n(z_0) \to \alpha,\ n \to \infty\}.$$

The *Fatou set* of the rational function R, $\mathcal{F}(R)$, is the set of points $z \in \hat{\mathbb{C}}$ whose orbits tend to an attractor (fixed point, periodic orbit or infinity). Its complement in $\hat{\mathbb{C}}$ is the *Julia set,* $\mathcal{J}(R)$. That means that the basin of attraction of any fixed point belongs to the Fatou set and the boundaries of these basins of attraction belong to the Julia set.

The fixed and critical points as well as their asymptotic behavior depend on the values of parameter β. Since even for quadratic polynomials, $O_p(z)$ is so complicated and difficult to work with, instead we use conjugacy maps to get a simpler operator. By using the following conjugacy map

$$M(z) = \frac{z-a}{z-b}, \qquad M^{-1} = \frac{zb-a}{z-1},$$

with the properties:

$$M(\infty) = 1, \qquad M(a) = 0, \qquad M(b) = \infty,$$

P. Blanchard in [2] proved that, for quadratic polynomials, Newton's operator is conjugated to rational function z^2,

In a similar way, it is easy to show that operator $O_p(z)$ is conjugated to the rational function

$$M_p(z,\beta) = z^7 \frac{(-1+5\beta+14\beta z+14\beta z^2+6\beta z^3+\beta z^4)r(z)}{(\beta+6\beta z+14\beta z^2+14\beta z^3-z^4+5\beta z^4)s(z)},$$

 ISBN: 978-84-617-8694-7

where $r(z) = 4\beta^2 - 24\beta^3 + 80\beta^4 + (\beta + 26\beta^2 - 188\beta^3 + 900\beta^4)z + (6\beta + 72\beta^2 - 608\beta^3 + 4592\beta^4)z^2 + (14\beta + 108\beta^2 - 968\beta^3 + 13952\beta^4)z^3 + (10\beta + 128\beta^2 - 688\beta^3 + 28080\beta^4)z^4 + (-1 + 3\beta + 114\beta^2 + 8\beta^3 + 39580\beta^4)z^5 + (-4\beta + 44\beta^2 + 320\beta^3 + 40400\beta^4)z^6 + (184\beta^3 + 30496\beta^4)z^7 + (40\beta^3 + 17216\beta^4)z^8 + (4\beta^3 + 7260\beta^4)z^9 + 2240\beta^4 z^{10} + 480\beta^4 z^{11} + 64\beta^4 z^{12} + 4\beta^4 z^{13}$ and $s(z) = 4\beta^4 + 64\beta^4 z + 480\beta^4 z^2 + 2240\beta^4 z^3 + (4\beta^3 + 7260\beta^4)z^4 + (40\beta^3 + 17216\beta^4)z^5 + (184\beta^3 + 30496\beta^4)z^6 + (-4\beta + 44\beta^2 + 320\beta^3 + 40400\beta^4)z^7 + (-1 + 3\beta + 114\beta^2 + 8\beta^3 + 39580\beta^4)z^8 + (10\beta + 128\beta^2 - 688\beta^3 + 28080\beta^4)z^9 + (14\beta + 108\beta^2 - 968\beta^3 + 13952\beta^4)z^{10} + (6\beta + 72\beta^2 - 608\beta^3 + 4592\beta^4)z^{11} + (\beta + 26\beta^2 - 88\beta^3 + 900\beta^4)z^{12} + (4\beta^2 - 24\beta^3 + 80\beta^4)z^{13}$.

The fixed points are the roots of equation $M_p(z, \beta) = z$, or $z(z-1)q(z) = 0$. In this case, $q(z)$ is a polynomial of degree 22, whose roots are different from $z = 0$ and $z = 1$. That is, fixed points of operator $M_p(z, \beta)$ are, $z = 0$, $z = 1$, $z = \infty$ and the 22 roots of $q(z)$, denoted by $s_i(\beta)$, $i = 1, 2, \cdots, 22$.

In order to study the stability of the fixed points, we calculate the first derivative of $M_p(z, \beta)$ and evaluate it at every fixed point. The resulting absolute value gives us information about the asymptotic behavior of the point. We conclude that $z = 0$ and $z = \infty$ are always superattracting fixed points, but the stability of other fixed points depend on the value of the parameter β.

On the other hand, changes in the multiplicity of the fixed points imply also alterations in their dynamical behavior. For different values of β we have:

- If $\beta = 0$, we have only one simple fixed point.
- There are many different values of β in the complex plane that for these values, strange fixed points $s_i(\beta)$, $i = 1, 2, \cdots, 22$ are equal to 1, so strange fixed point 1 can have different multiplicities.
- However, the behavior of the strange fixed points are different in the complex plane especially around 0, as in a small region near 0 some of them satisfy $|M_p'(s_i(\beta), \beta)| < 1$. Figure 1, depicts $|M_p'(s_i(\beta), \beta)| < 1$, for $i = 18, 19, \cdots, 22$ and strange fixed point 1 simultaneously near 0. These six strange fixed points among all strange fixed points have the bigger regions in which $|M_p'(s_i(\beta), \beta)| < 1$, in this figure S_{one} denoted strange fixed point $z = 1$. One can see more details of these stability functions in Figure 2, where boundaries of regions around the basin of attraction of strange fixed point $z = 1$ are shown. These four regions are denoted by $B1$, $B2$, $B3$ and $B4$.

4 Critical points of operator $M_p(z, \beta)$

Let us recall that critical points of $M_p(z, \beta)$ are the roots of $M_p'(z, \beta) = 0$, since we have

$$M_p'(z, \beta) = (z + 1)^{12} z^6 u(z),$$

 ISBN: 978-84-617-8694-7

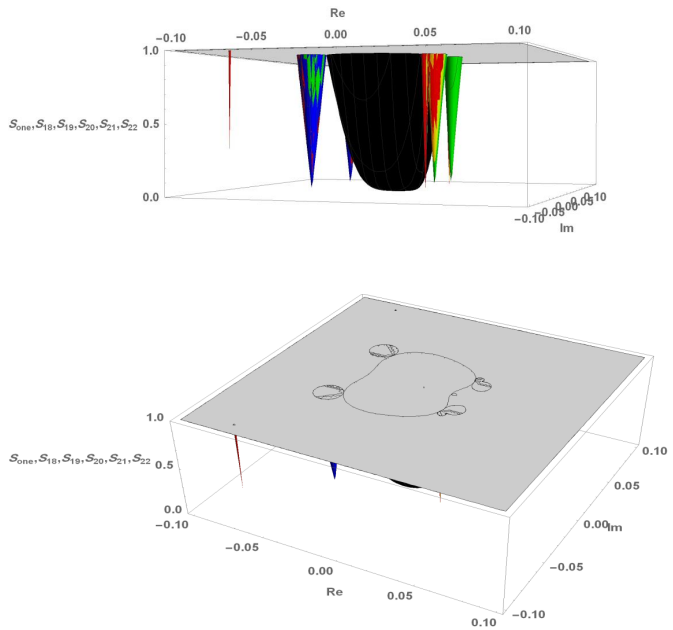
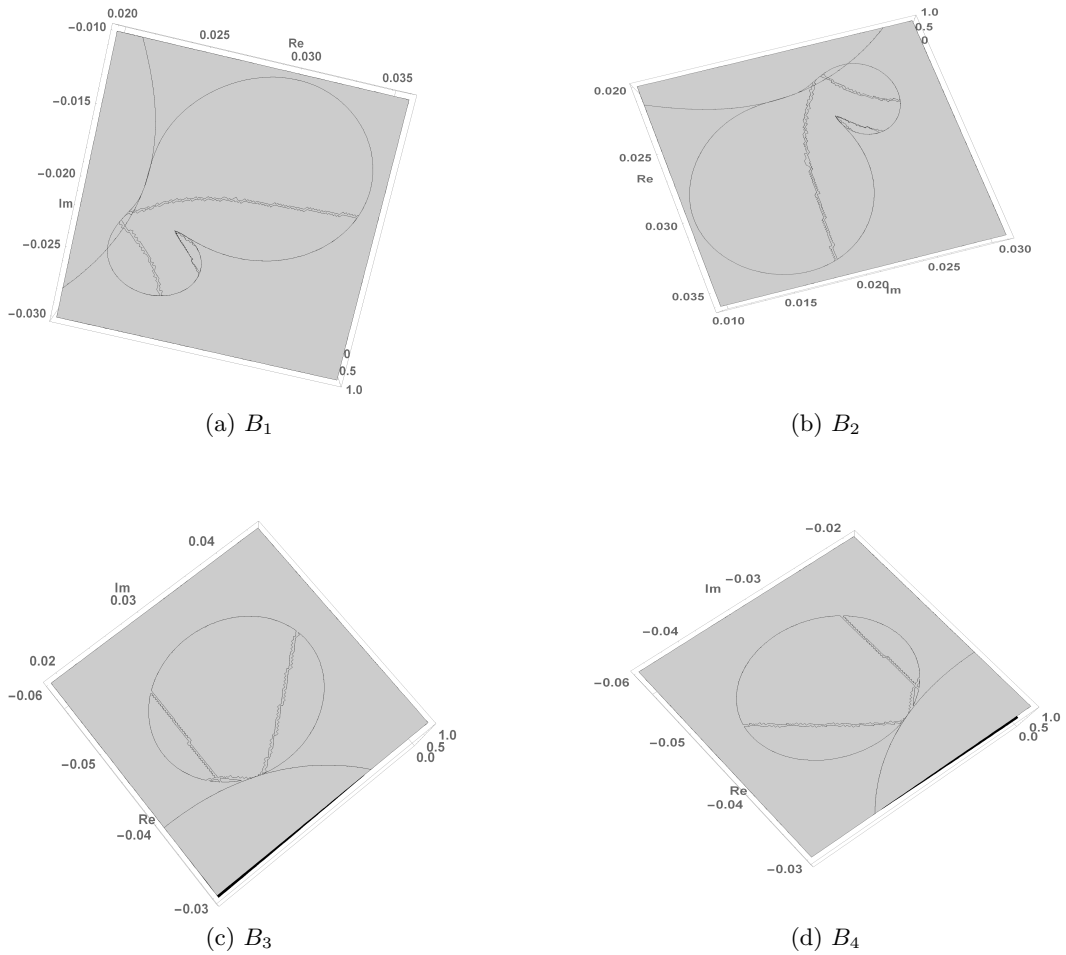


Figure 1: Stability functions of strange fixed points S_{one} and $S_{18}, S_{19}, \cdots, S_{22}$.

where $u(z) = 1400\beta^5 + (-4\beta - 40\beta^2 + 1012\beta^3 - 6504\beta^4 + 18080\beta^5)z + (-16\beta + 71\beta^2 + 2506\beta^3 - 25464\beta^4 + 106640\beta^5)z^2 + (-12\beta + 440\beta^2 + 2244\beta^3 - 51208\beta^4 + 377120\beta^5)z^3 + (54\beta + 274\beta^2 + 510\beta^3 - 49668\beta^4 + 886920\beta^5)z^4 + (6 - 36\beta - 224\beta^2 + 2184\beta^3 - 7824\beta^4 + 1458240\beta^5)z^5 + (-11 - 42\beta - 310\beta^2 + 4540\beta^3 + 20592\beta^4 + 1716960\beta^5)z^6 + (6 - 36\beta - 224\beta^2 + 2184\beta^3 - 7824\beta^4 + 1458240\beta^5)z^7 + (54\beta + 274\beta^2 + 510\beta^3 - 49668\beta^4 + 886920\beta^5)z^8 + (-12\beta + 440\beta^2 + 2244\beta^3 - 51208\beta^4 + 377120\beta^5)z^9 + (-16\beta + 71\beta^2 + 2506\beta^3 - 25464\beta^4 + 106640\beta^5)z^{10} + (-4\beta - 40\beta^2 + 1012\beta^3 - 6504\beta^4 + 18080\beta^5)z^{11} + (-14\beta^2 + 154\beta^3 - 700\beta^4 + 1400\beta^5)z^{12}$.

So the critical points of $M_p(z, \beta)$ are $z = 0$, $z = \infty$ and $z = -1$ and the 12 roots of polynomial $u(z)$. We denote these 12 roots of $p(z)$ as $c_i(\beta)$, $i = 1, 2, \cdots 12$. To obtain some properties of $c_i(\beta)$ we discretize the square of $[-2, 2] \times [-2, 2]$ in the complex plane with 400×400 mesh points. We denote this mesh points by (i, j), where $i, j = 1, 2, \cdots, 400$, then

 ISBN: 978-84-617-8694-7



(a) B_1 (b) B_2

(c) B_3 (d) B_4

Figure 2: Details of the stability functions of strange fixed points S_{one} and $S_{18}, S_{19}, \cdots, S_{22}$.

 ISBN: 978-84-617-8694-7

by obtaining the roots of $p(z)$ for each $\beta = (i, j)$, a numerical value of critical points c_i in the square $[-2, 2] \times [-2, 2]$, for $i = 1, 2, \cdots, 12$ is obtained.

A classical result establishes that there is at least one critical point associated with each invariant Fatou component. Points $z = \infty$ and $z = 0$ are both superattractive fixed points of $M_p(z, \beta)$, so they also are critical points and give rise to their respective Fatou components. So, the way to calculate the parameter planes associate to $M_p(z, \beta)$ is to study the orbits of each free independent critical point for all the complex value of the parameter β in the defined mesh. In the next section we investigate parameter planes of $M_p(z, \beta)$.

5 The parameter space

As mentioned in the previous sections, the dynamical behavior of operator $M_p(z, \beta)$ depends on the values of the parameter β. In order to get the parameter plane we study the orbits of the free critical points for each β in the complex square $[-2, 2] \times [-2, 2]$. Strange fixed points are attracting near zero, so we only focus our attention in this region. To find the most stable members of the family M7, we are looking for regions in which the orbits of free critical points converge to zero or infinity. Since 12 critical points of $M_p(z, \beta)$ are two-

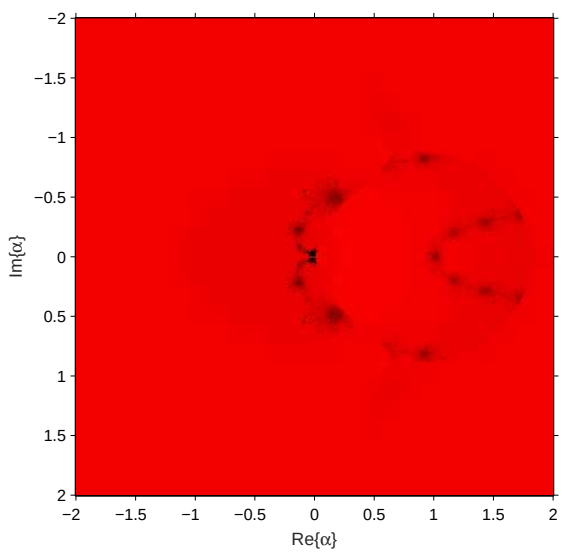
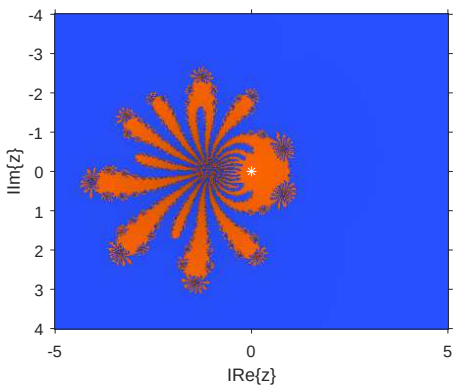


Figure 3: Parameter plane for critical point c_1

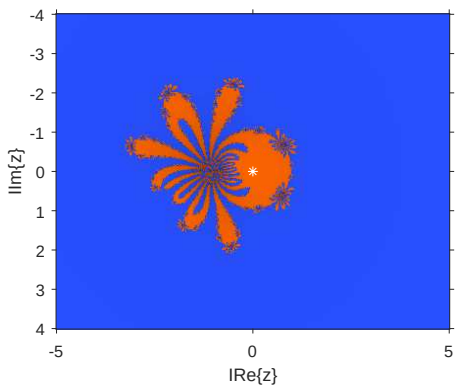
by-two dependent and we here study the parameter plane of one of these 6 critical points, shown in Figure 3, calculated by using the routines presented in [4]. We have painted in red the values of the parameter that makes the critical point converge to $z =$ or $z = \infty$

 ISBN: 978-84-617-8694-7

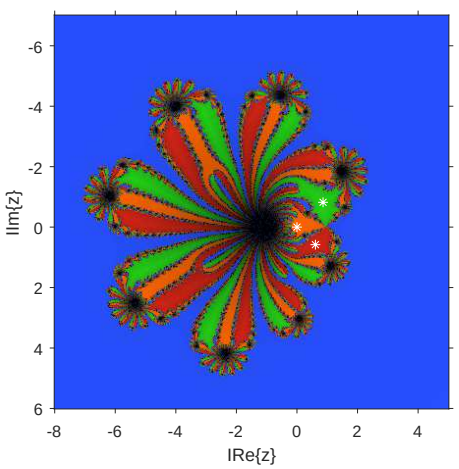
(brighter as lower is the number of iterations needed), and in black if the critical point has not converged after 600 iterations, or has converged to another element (attracting strange fixed point or periodic orbit).



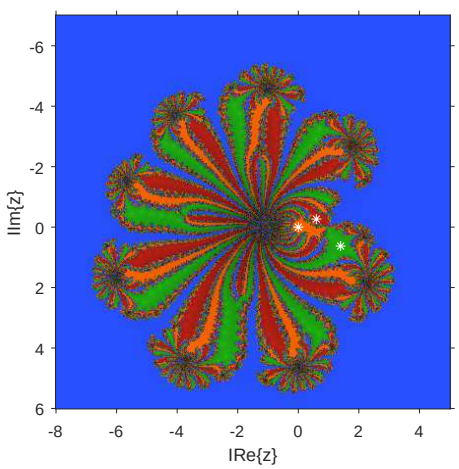
(a) $\beta = 0.1 - 0.1i$



(b) $\beta = 0.5 + 0.5i$



(c) $\beta = -0.05 + 0.03i$



(d) $\beta = 0.032 - 0.021i$

Figure 4: Some dynamical planes of family M7

By analyzing the conclusions obtained from the parameter planes of the family M7, we select some elements of this family to depict their dynamical planes. These planes show different behaviors of the members of the family (see Figure 4). These dynamical planes have generated by iterating an element of the family M7 that is, for a constant β and using

each point of the complex plane as an initial estimation the dynamical plane associated to each member obtained. These dynamical planes are built (again with the routines appearing in [4]) by using an a mesh of 400×400 points, a maximum number of 100 iterations and a tolerance of 10^{-3}. Figures 4a and 4b show the dynamical planes of members of $\beta = 0.1-0.1i$ and $0.5 + 0.5i$, respectively, corresponding to the red area of the parameter plane. In both figures there are only two basins of attraction that correspond to superattracting $z = 0$ and $z = \infty$, respectively shown in orange and blue colors. On the other hand, Figure 4c and 4d show the members $\beta = -0.05+0.03i$ and $\beta = 0.032-0.021i$, corresponding to values of β in the stability regions appearing in Figure (2a) and (2c) of some strange fixed points. For both of them there exists four basins of attraction. Value $\beta = -0.05 + 0.03i$ results in a member of the family whose dynamical plane shows the basins of attraction of $z = 0$, $z = \infty$ and, in red and green colors, the basins of attraction of two strange fixed points which numerical values for $\beta = -0.05 + 0.03i$ are equal to $0.608121 + 0.58028i$ and $0.860704 - 0.821302i$. Also $\beta = 0.032-0.021i$ results in a member of the family that has four basins of attractions corresponding to $z = 0$, $z = \infty$ and two strange fixed points $z = 1.39925 + 0.626732i$ and $z = 0.59525 - 0.266617i$ whose basins of attraction are represented in the dynamical planes in green and red color, respectively.

Acknowledgements

First and third authors have been partially supported by Ministry of Science of Islamic Republic of Iran. Second and fourth authors have been partially supported by Ministerio de Economía y Competitividad, MTM2014-52016-C2-2-P, and Generalitat Valenciana PROMETEO/2016/089.

References

[1] S. Amat, S. Busquier, S. Plaza, *A construction of attracting periodic orbits for some classical third-order iterative methods*, J. of Computational and Applied Mathematics **189** (2006) 22–33.

[2] P. Blanchard, *The dinamics of Newton's method*, Proc. Symp. Appl. Math. **49** (1994) 139–154.

[3] A. Cordero, J.M. Gutiérrez, Á. A. Magreñán, J.R. Torregrosa, *Stability analysis of a parametric family of iterative methods for solving nonlinear models*, Applied Mathematics and Computation **285** (2016) 26–40.

 ISBN: 978-84-617-8694-7

[4] F. Chicharro, A. Cordero, J.R. Torregrosa, *Drawing dynamical and parameters planes of iterative families and methods*, The Scientific World Journal **2013** (2013) Article ID 780153.

[5] J.L. Hueso, E. Martínez, C. Teruel,*Convergence, effiency and dinamimics of new fourth and sixth order families of iterative methods for nonlinear systems*, Comput. Appl. Math. **275** (2015) 412–420.

[6] J.M. Ortega, W.C. Rheinboldt,*Iterative Solution of Nonlinear Equations in Several Variables*, Academic Press, 1970.

[7] M. Scott, B. Neta, C. Chun, *Basin attractors for various methods*, Applied Mathematics and Computation **218** (2011) 2584–2599.

[8] J.R. Sharma, H. Arora, *Efficient Jarratt-like methods for solving systems of nonlinear equations*,Calcolo **51** (2014) 193–210.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

Square cylinder with passive flow control

Bo AN[1], Josep M Bergada[1] and Aditya Mushyam[1]

[1] *Fluid Mechanics Department, University Polithecnic of Catalunya, Terrassa, Barcelona, Spain*

emails: bo_alan_an@163.com, josep.m.bergada@upc.edu, mushyam.aditya@gmail.com

Abstract

In the present paper it is presented the flow around a 2D square cylinder which is located downstream of a splitter plate and at a certain distance of it. The fluid velocity below and above the splitter plate is different, several velocity ratios are considered, it is interesting to see that the downstream vortex shedding frequency and amplitude highly depends on the velocity ratio defined upstream. So far, the Reynolds numbers evaluated fall into the laminar unsteady regime, yet the interaction between the upstream mixing layer and the wake generate fully different downstream vortex shedding for different upstream velocity ratios, lift, drag and Strouhal numbers are as well highly dependent on the velocity ratios. In the present paper, the comparison between the resuls obtained via CFD finite volumes and Lattice Boltzmann Method are being presented. For these initial cases studied the agreement is very good.

Key words: Laminar flow, CFD, Finite volumes, Lattice Boltzmann Methods, Passive flow control.

1 Introduction

It has been shown by past investigations on circular cylinders, e.g., [1–6], that the flow approaching with linear shear greatly altered the vortex dynamics in the wake when compared to the uniform flow case. They attributed this phenomena to the constant vorticity embedded in the free-stream. For square-sectional cylindrical bodies, shear effects have been reported by, e.g., [7–14]. Saha et al. [9] studied numerically the same problem for a wide range of Reynolds numbers. They showed that due to influence of shear, Karman Vortex Street mainly consisted of clockwise vortices, whose decay was very slow when compared to that of uniform flow. Cheng et al. [12, 13] reported that vortex shedding disappeared

for large shear parameters. The mean lift and drag coefficients tended to decrease with increasing the shear parameter. They also observed that the vortex shedding frequency tended to decrease with the increase of the shear parameter, although they highlighted that this observation was opposite to the one obtained by Kiya et al. [2] when studding shear flow past circular cylinders.

Lankadasu and Vengadesan [14] also reported a decrease of mean lift and drag coefficients with the increasing shear, for a given Reynolds number. The same phenomenon was observed for a given shear and when Reynolds number was increased. It was also found that the critical Reynolds number at which the flow becomes unsteady periodic, was reduced as shear increased. Bhattacharyya and Maiti [10] studied laminar shear flow past a square cylinder placed nearby a wall, Reynolds up to 1400. For a gap height of 0.25 times the square cylinder height, vortex shedding suppression and steadiness of the wake was observed up to a Reynolds 250. For a Reynolds number equal and above 500, only negative vortices behind the cylinder at closed proximity to the wall were found. Bhattacharyya and Maiti [11] investigated as well laminar flow over a square cylinder placed nearby a wall, Reynolds <1500. They found that the critical gap height for vortex shedding suppression was dependent on the Reynolds number.

There are typical situations, where the non-uniformity of the approaching flow is due to the effect of a body located in front or behind the one to be studied. For example, when a small cylinder is placed in the separated shear layer of a large main cylinder to alter the vortex shedding phenomena behind the main cylinder, the small cylinder is inevitably subjected to the effect of shear induced from the large cylinder. Onset of periodic flow from the small cylinder induces oscillatory forces on the body, which may trigger flow-induced vibrations, and in turn alter the vortex shedding on the main cylinder.

Some of the latest most relevant numerical simulations of flow past a square cylinder with the incorporation of a control plate to alter the wake, which is having similarities to the present study, were carried out among others by, Lesage and Gartshore [15], Sakamoto et al [16, 17], Zhou et al [18], Doolan [19], Ali et al [20, 21], Malekzadeh and Sohankar [22] and Salinas et al [23]. It must be highlithed that none of these previous studies resembles the one presented in this paper, in fact according to the authors knowledge, just a single previous paper undertaken by the authors, see [24] matches with the work presented here.

2 Problem definition

The computational domain under study is defined in figure 1a, notice that it consists of a 2D square cylinder located downstream of a splitter plate, the plate thickness is negligible, being the distance between the splitter plate and the square three times the square cylinder lateral side. This distance, which will remain constant in this research, plays an important role, since it controls the mixing flow upstream. The boundary conditions employed, already

 ISBN: 978-84-617-8694-7

presented in figure 1, consist of, newmann boundary conditions for pressure and Dirichlet boundary conditions for velocity at the inlet, no slip boundary conditions were applied to all solid boundaries, at the upper and lower boundaries Newmann boundary conditions for velocty and pressure vere defined, finally, at the outlet, Newmann boundary conditions for pressure and velocity were considered. A grid of 200x150 was used for all the finite volume simulations carried out in the present study, with 50 cells allocated on the square cylinder side in both x and y directions. The non-dimensional time step employed was of dt=0.001. Three diferent Reynolds numbers Re=100; 150 and 200, defined as a function of the fluid velocity below the plate and the square cylinder side, were evaluated. For each Reynolds six different velocity ratios were considered r= 1.5; 2; 2.5; 3; 3.5 and 4. Via using the conventional LBM, just few cases were considered, the idea behind the employment of this particular method was to compare the results with the previous ones obtained via finite volumes. With LBM, just a single Reynolds number was studied Re=100, two velocity ratios were considered r=1.5 and 2. The advantage of using LBM is that it is computationally less expensive and therefore an extreme fine mesh could be used, the mesh employed was having 6200x3200 cells, 200 cells were used along each square cylinder side. The non-dimensional time step used for the case of LBM was of dt=0.0005.

In the present paper, two different self-made codes for laminar flow were used, the first one is based on finite volumes implementation of the Navier Stokes equations (NS), the second one employs the Lattice Boltzmann Method (LBM) to perform the CFD siulations. The code based on NS equations was already validated in [24], the validation of the LBM code is implementd in the next section. In figure 1b, the mesh used for the finite volumes code is presented, figures 1c and d respectively introduce the mesh used for the LBM code and a zoom view of it.

3 Code validation

The Lattice boltzmann code was validated via comparing the results with the previous researchers ones. It was observed that the flow around a square cylinder, was steady for Reynolds numbers 50 and 52, and unsteady periodical at Re=53. According to Sohankar et al [25], Von Karman vortex street appeared and therefore vortex shedding was happening in the wake of the square cylinder at $Re = 51.2 \pm 1$. Kelkar and Patankar [26] through stability analysis of the flow, found that the critical Reynolds number for the onset of vortex shedding was 54. From the present LBM code, at Reynolds 52, the steady downstream bubble lenght obtained when using LBM differed in 2.9% versus the steady bubble length obtained in [24], based on these results we considered the LBM code presented in this paper, fully validated for the present application at laminar Reynolds numbers.

 ISBN: 978-84-617-8694-7

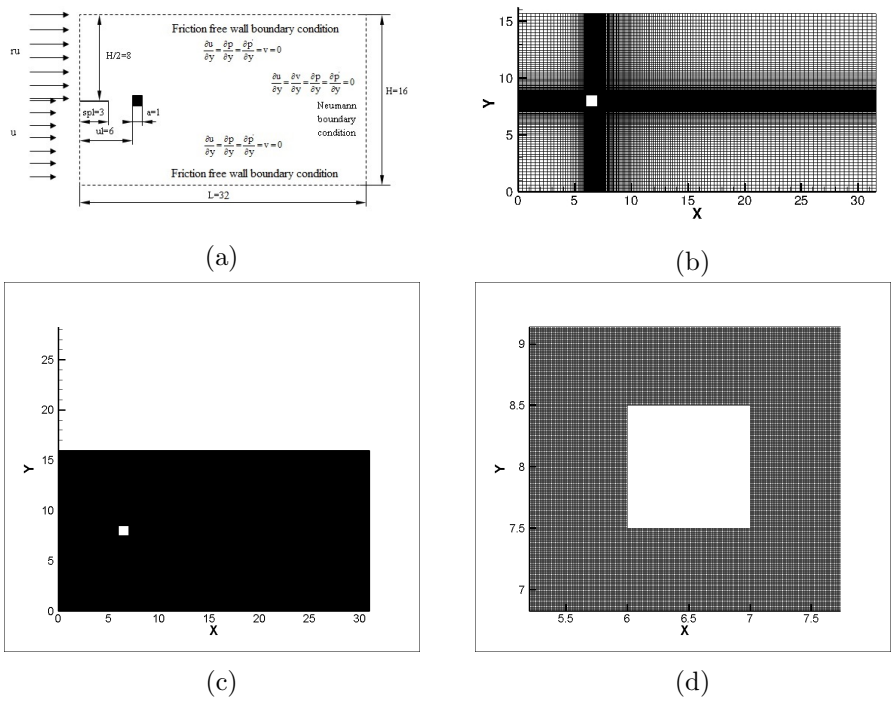


Figure 1: (a) Problem definition and boundary conditions used for the simulation, (b) overview of the mesh used for FV, (c) mesh used for LBM, (d) zoom of the mesh used for LBM.

4 Results

The vast majority of the results presented in this section are the ones obtained with the NS equations finite volumes code, the intention of the authors is comparing these results with the ones obtained using LBM. For the cases studied, this comparisons are presented in figures 2 to 4.

Figure 2 presents the stagnation point location at the square cylinder upstream vertical face. It is important to realize that the stagnation point location fluctuates regardless of the upstream conditions established, the fluctuation amplitude initially decreases and then tends to increase with the velocity ratio increase. What is even more interesting, is that the stagnation point location suffers a drastic jump for velocity ratios bigger than 2, clearly indicating that vortex shedding will suffer a drastic change. The results obained via using LBM are represented by rounded dots, and it is noticed that at small velocity ratios, r=1.5, the stagnation point location fluctuation, is much bigger than the one obtained when using finite volumes. Whenever the velocity ratio was 2, the results were in good agreement with

the ones obtained using finite volumes. At this point, the origin of the disagreement at low velocity ratios is so far unknown, further cases need to be computed to find out which of the two codes produces more accurate results. The wide amplitude associated to the stagnation point location at low velocity ratios is due to the wide fluctuation observed on the upstream mixing layer.

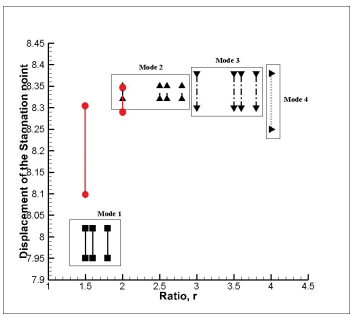


Figure 2: Displacement of the stagnation point vs. the Ratio, r for Reynolds number, Re=100 and upstream length, ul=6a. The rounded dots characterize the results obtained via using LBM.

Figure 3 presents, the average lift and drag coefficients as a fucntion of the velocity ratio and the upstream reynolds number, obtained when employing the velocity below the plate. This figure clearly indicates that the most important parameter is the velocity ratio, Reynolds number plays a secondary role. It also clarifies that four vortex shedding modes are obtained when increassing the velocity ratio from 1.5 to 4. A further explanation of the different modes and its associated downstream vortex shedding is to be found in [24]. When comparing the previous results obtained via finite volumes with the ones obtained using LBM, it can be seen that the agreement is very good. Clearly more results need to be obtained via LBM, but so far the expectatives are promissing.

Figure 4 presents the dowstream vortex sheddding non dimensional frequency as a function of the velocity ratio and the Reynols number, again a very similar trend to the one presented in figure 3 is to be spotted. Yet, although velocity ratio plays a very relevant role, it appears that the vortex shedding frequency is more deeply affected by the Reynolds number than the lift and drag coefficients. The increase in Strouhal number between two consecutive modes is larger for higher velocity ratios and higher Reynolds numbers, because as the velocity ratio increases, it leads to an increase in flow kinetic energy causing higher oscillations in the mixing layer making it unstable and enhancing the vortex shedding frequency downstream. It is observed that in modes 1 and 2 the predominant effect when considering the Strouhal number change is due to velocity ratio r, whereas in modes 3 and 4, the Reynolds number increase generates a higher variation in Strouhal number. This is due to the higher difference in of the kinetic energy, which is 9 and 16 times bigger respec-

 ISBN: 978-84-617-8694-7

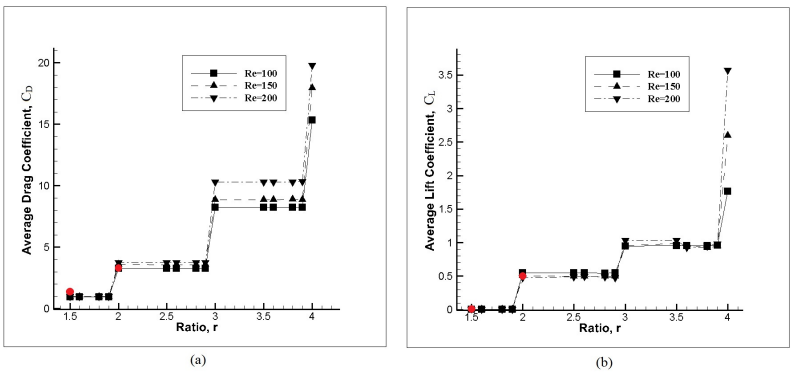


Figure 3: (a) Average drag coefficient, $C_{D_{avg}}$ versus ratio, r. (b) Average Lift coefficient, $C_{L_{avg}}$ versus ratio, r. Upstream length ul = 6a. The rounded dots characterize the results obtained via using LBM.

tively for modes 3 and 4 on the top of the splitter plate when compared to the kinetic energy below the plate, this kinetic energy difference is much higher than the one at lower Reynolds numbers, promoting the vortex dissipation in the wake of the square cylinder. When comparing the results obtained via using LBM with the ones gathered via finite volumes, the agreement is very good, which certifies that at low Reynolds numbers, even when the flow has relatively high shear stresses, the conventional LBM is a reliable and fast methodology to be employed.

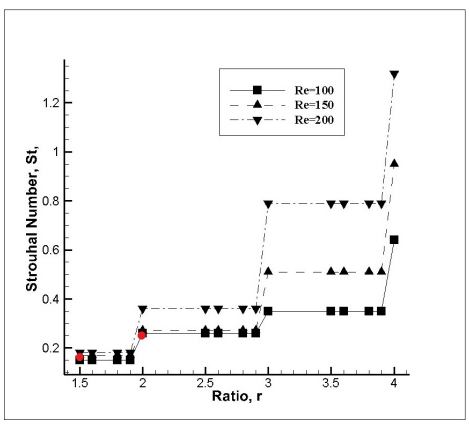


Figure 4: Strouhal Number, St versus Ratio, r for (a) ul=6a. The rounded dots characterize the results obtained via using LBM.

 ISBN: 978-84-617-8694-7

5 Conclusions

This paper is comparing the results obtained when evaluating the flow around a square cylinder with a splitter plate located in front of it, via employing two self made codes, one based on finite volumes and the other based on LBM.

A drastic change in lift and drag coefficients was observed as a function of the upstream velocity ratio. The effect of Reynolds number plays a secondary role, being this role more relevant at high velocity ratios due to the high differences in the splitter plate above/below kinetic energy, which drastically modifies the mixing and shear layers.

For the few cases evaluated using LBM, the results were pretty accurate, althoug the entire set of Reynolds and velocity ratios need to be evaluated using this methodology, the resuls obtained and presented in this paper give confidence to the authors regarding the final outcome of the reserach.

Acknowledgements

The present paper presents part of the results obtained thanks to a competitive research project number FIS0016-77849-R founded by Spanish economy ministry.

References

[1] Stanley K Jordan and Jacob E Fromm. Laminar flow past a circle in a shear flow. *The Physics of Fluids*, 15(6):972–976, 1972.

[2] Masaru Kiya, Hisataka Tamura, and Mikio Arie. Vortex shedding from a circular cylinder in moderate-reynolds-number shear flow. *Journal of Fluid Mechanics*, 101(04):721–735, 1980.

[3] Tae Soon Kwon, Hyung Jin Sung, and Jae Min Hyun. Experimental investigation of uniform-shear flow past a circular cylinder. *ASME J. Fluids Eng*, 114:457–460, 1992.

[4] A Mukhopadhyay, P Venugopal, and SP Vanka. Numerical study of vortex shedding from a circular cylinder in linear shear flow. *Journal of fluids engineering*, 121(2):460–468, 1999.

[5] Y Xu and C Dalton. Computation of force on a cylinder in a shear flow. *Journal of fluids and structures*, 15(7):941–954, 2001.

[6] D Sumner and OO Akosile. On uniform planar shear flow around a circular cylinder at subcritical reynolds number. *Journal of Fluids and Structures*, 18(3):441–454, 2003.

[7] K Ayukawa, J Ochi, G Kawahara, and T Hirao. Effects of shear rate on the flow around a square cylinder in a uniform shear flow. *Journal of Wind Engineering and Industrial Aerodynamics*, 50:97–106, 1993.

[8] Robert R Hwang, YC Sue, et al. Numerical simulation for shear effect on vortex shedding behind a square cylinder. In *The Eighth International Offshore and Polar Engineering Conference*. International Society of Offshore and Polar Engineers, 1998.

[9] AK Saha, G Biswas, and K Muralidhar. Two-dimensional study of the turbulent wake behind a square cylinder subject to uniform shear. *TRANSACTIONS-AMERICAN SOCIETY OF MECHANICAL ENGINEERS JOURNAL OF FLUIDS ENGINEERING*, 123(3):595–603, 2001.

[10] S Bhattacharyya and DK Maiti. Shear flow past a square cylinder near a wall. *International Journal of Engineering Science*, 42(19):2119–2134, 2004.

[11] S Bhattacharyya and DK Maiti. Vortex shedding suppression for laminar flow past a square cylinder near a plane wall: a two-dimensional analysis. *Acta mechanica*, 184(1):15–31, 2006.

[12] M Cheng, SHN Tan, and KC Hung. Linear shear flow over a square cylinder at low reynolds number. *Physics of Fluids*, 17(7):078103, 2005.

[13] M Cheng, DS Whyte, and J Lou. Numerical simulation of flow around a square cylinder in uniform-shear flow. *Journal of fluids and structures*, 23(2):207–226, 2007.

[14] A Lankadasu and S Vengadesan. Onset of vortex shedding in planar shear flow past a square cylinder. *International Journal of Heat and Fluid Flow*, 29(4):1054–1059, 2008.

[15] F Lesage and IS Gartshore. A method of reducing drag and fluctuating side force on bluff bodies. *Journal of Wind Engineering and Industrial Aerodynamics*, 25(2):229–245, 1987.

[16] H Sakamoto, K Tan, and H Haniu. An optimum suppression of fluid forces by controlling a shear layer separated from a square prism. *Journal of Fluids Engineering*, 113(2):183–189, 1991.

[17] H Sakamoto, K Tan, N Takeuchi, and H Haniu. Suppression of fluid forces acting on a square prism by passive control. *Journal of Fluids Engineering*, 119(3):506–511, 1997.

[18] L Zhou, M Cheng, and KC Hung. Suppression of fluid force on a square cylinder by flow control. *Journal of Fluids and Structures*, 21(2):151–167, 2005.

 ISBN: 978-84-617-8694-7

[19] Con J Doolan. Flat-plate interaction with the near wake of a square cylinder. *AIAA journal*, 47(2):475–479, 2009.

[20] Mohamed Sukri Mat Ali, Con J Doolan, and Vincent Wheatley. Low reynolds number flow over a square cylinder with a splitter plate. *Physics of Fluids*, 23(3):033602, 2011.

[21] Mohamed Sukri Mat Ali, Con J Doolan, and Vincent Wheatley. Low reynolds number flow over a square cylinder with a detached flat plate. *International Journal of Heat and Fluid Flow*, 36:133–141, 2012.

[22] S Malekzadeh and A Sohankar. Reduction of fluid forces and heat transfer on a square cylinder in a laminar flow regime using a control plate. *International Journal of Heat and Fluid Flow*, 34:15–27, 2012.

[23] M Salinas-Vazquez, W Vicente, E Barrera, and E Martinez. Numerical analysis of the drag force of the flow in a square cylinder with a flat plate in front. *Revista mexicana de física*, 60(2):102–108, 2014.

[24] Aditya Mushyam and Josep M Bergada. A numerical investigation of wake and mixing layer interactions of flow past a square cylinder. *Meccanica*, 52(1-2):107–123, 2017.

[25] A Sohankar, C Norberg, and L Davidson. Low-reynolds-number flow around a square cylinder at incidence: study of blockage, onset of vortex shedding and outlet boundary condition. *International journal for numerical methods in fluids*, 26(1):39–56, 1998.

[26] Kanchan M Kelkar and Suhas V Patankar. Numerical prediction of vortex shedding behind a square cylinder. *International Journal for Numerical Methods in Fluids*, 14(3):327–341, 1992.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

Nondominated solutions in a fully fuzzy linear programming problem

M. Arana-Jiménez[1]

[1] *Department of Statistics and Operations Research, University of Cádiz*

emails: manuel.arana@uca.es

Abstract

In this work, it is presented a new method for solving the fuzzy optimal (nondominated) solutions of a fully fuzzy linear programming problem with inequality constraints and triangular fuzzy numbers, not necessarily symmetric, by the means of solving a multiobjective linear problem. It is proved an equivalence between the set of nondominated solutions of the fully fuzzy linear programming problem and the set of weakly efficient solutions of the considered and related mulitobjective linear problem.

Key words: Fully fuzzy linear programming problem, fuzzy numbers, multiobjective optimization

1 Introduction

Nowadays, the concept of decision making in fuzzy environment introduced by Bellman and Zadeh [1] is well-known and adopted by researchers in fields close to fuzzy linear programming [2, 3, 4, 5, 6]. It was usual that not all parts of the fuzzy linear problem were assumed to be fuzzy. In fact, an interesting problem in the recent literature, at the same time that a challenge, in to solve a fuzzy linear programming problems in which all the parameters as well as the variables are represented by fuzzy. In this regard, Lofti et al. [7] pointed out that there was no method in literature for finding the fuzzy optimal solution of fully fuzzy linear programming (FFLP) problems and proposed a new method to find the fuzzy optimal solution of (FFLP) problems with equality constraints with symmetric fuzzy numbers. Kumar et al. [8] claim that there was no method in the literature to obtain the exact solution of (FFLP) problems with equality constraints, and that in [7] the solutions are approximate not exact and also it is very difficult to apply the existing method. In this regard, they propose a new method for finding the fuzzy optimal solution of (FFLP) problems with equality constraints, with triangular fuzzy numbers involved, although they use ranking function (see [9] and

 ISBN: 978-84-617-8694-7

the bibliography there in) to compare the objective function values. In this way, Khan et al. [10] deal with (FFLP) with inequalities, and they also compare the objective function values via ranking functions (see also [11, 12]).

In this work, it is presented a new method to find the fuzzy optimal (nondominated) solutions of (FFLP) problems with inequality constraints with triangular fuzzy numbers and not necessarily symmetric, via solving a multiobjective linear problem with crips numbers. It is proved that there exists an equivalence between the set of fuzzy optimal (nondominated) solutions of (FFLP) and the set of weakly efficient solutions of its related mulitobjective linear problem. No ranking functions are needed in this method. Due to lenght requirements on this paper, proofs are omited.

2 Notation on fuzzy numbers

We denote by $\mathcal{K}_C$ the family of all bounded closed intervals in $\mathbb{R}$, i.e.,

$$\mathcal{K}_C = \left\{ \left[\underline{a}, \overline{a}\right] \mid \underline{a}, \overline{a} \in \mathbb{R} \text{ and } \underline{a} \leq \overline{a} \right\},$$

A fuzzy set on $\mathbb{R}^n$ is a mapping $u : \mathbb{R}^n \to [0, 1]$. For each fuzzy set u, we denote its α-level set as $[u]^\alpha = \{x \in \mathbb{R}^n \mid u(x) \geq \alpha\}$ for any $\alpha \in (0, 1]$. The support of u we denote by $supp(u)$ where $supp(u) = \{x \in \mathbb{R}^n \mid u(x) > 0\}$. The closure of $supp(u)$ defines the 0-level of u, .i.e. $[u]^0 = cl(supp(u))$ where $cl(M)$ means the closure of the subset $M \subset \mathbb{R}^n$.

Definition 1 *A fuzzy set u on $\mathbb{R}$ is said to be a fuzzy interval if:*

1. *u is normal, i.e. there exists $x_0 \in \mathbb{R}$ such that $u(x_0) = 1$;*
2. *u is an upper semi-continuous function;*
3. *$u(\lambda x + (1 - \lambda)y) \geq min\{u(x), u(y)\}$, $x, y \in \mathbb{R}$, $\lambda \in [0, 1]$;*
4. *$[u]^0$ is compact.*

Let $\mathcal{F}_C$ denote the family of all fuzzy intervals. So, for any $u \in \mathcal{F}_C$ we have that $[u]^\alpha \in \mathcal{K}_C$ for all $\alpha \in [0, 1]$ and thus the α-levels of a fuzzy interval are given by $[u]^\alpha = \left[\underline{u}_\alpha, \overline{u}_\alpha\right]$, $\underline{u}_\alpha, \overline{u}_\alpha \in \mathbb{R}$ for all $\alpha \in [0, 1]$. If $[u]^1$ is a singleton then we say that u is a fuzzy number. In this regard, the representation of fuzzy numbers has been deeply discussed by Stefanini et al. [13]. Triangular fuzzy numbers are a special type of fuzzy numbers, well-known in the literature (see, for instance, [14, 13, 7, 10]) which are well determined by three real numbers $a \leq b \leq c$. Its α-levels are formulated as

$$[u]^\alpha = [a + (b - a)\alpha, c - (c - b)\alpha],$$

 ISBN: 978-84-617-8694-7

for all $\alpha \in [0,1]$. Also we can denote a triangular fuzzy number $u = (a,b,c)$ by $\tilde{b}$. This formulation of α-levels characterizes a unique trianggular fuzzy number, what can be established by the following result, which makes the connection between a fuzzy interval and their endpoint functions (Goestschel and Voxman [15]).

Theorem 1 *Let u be a fuzzy interval. Then the functions* $\underline{u},\ \overline{u} : [0,1] \to \mathbb{R}$*, defining the endpoints of the* α*-level sets of u (*$\underline{u}(\alpha) = \underline{u}_\alpha$ *and* $\overline{u}(\alpha) = \overline{u}_\alpha$*), satisfy the following conditions:*

(i) $\underline{u}$ *is a bounded, non-decreasing, left-continuous function in* $(0,1]$ *and it is right-continuous at* 0*.*

(ii) $\overline{u}$ *is a bounded, non-increasing, left-continuous function in* $(0,1]$ *and it is right-continuous at* 0*.*

(iii) $\underline{u}(1) \leq \overline{u}(1)$*.*

Reciprocally, given two functions that satisfy the above conditions they uniquely determine a fuzzy interval.

The nonnegative conditions in some optimizations problems makes useful the following special consideration of triangular fuzzy numbers.

Definition 2 *Let* $u = (\underline{u}, \overline{u})$ *be a fuzzy interval. We say that u is a nonnegative fuzzy interval (nonpositive fuzzy interval, respectively) if* $\underline{u}(0) \geq 0$ *(*$\overline{u}(0) \leq 0$*, repectively).*

So, from the previus definition, a nonnegative triangular fuzzy number $\tilde{b} = (a,b,c)$ is characterized by $a \geq 0$.

Following, we consider some classical arithmetic operations on interval and fuzzy numbers. Given $A = [\underline{a}, \overline{a}],\ B = [\underline{b}, \overline{b}] \in \mathcal{K}_C$ and $\tau \in \mathbb{R}$:

(i) $A + B = [\underline{a} + \underline{b}, \overline{a} + \overline{b}]$,

(ii) $\tau A = \{\tau a : a \in A\} = \begin{cases} [\tau\underline{a}, \tau\overline{a}], & \text{if } \tau \geq 0, \\ [\tau\overline{a}, \tau\underline{a}], & \text{if } \tau \leq 0 \end{cases}$

We refer to Moore [16, 17] and Alefeld and Herzberger [18] for further details on the topic of interval analysis. As a natural extension of the previous operations, it is well known that if we consider the fuzzy intervals $u, v \in \mathcal{F}_C$ represented by $\left[\underline{u}_\alpha, \overline{u}_\alpha\right]$ and $\left[\underline{v}_\alpha, \overline{v}_\alpha\right]$, respectively, and a real number λ, then the addition $u + v$ and scalar multiplication λu are as follows:

$$(u+v)(x) = \sup_{y+z=x} \min\{u(y), v(z)\}$$

$$(\lambda u)(x) = \begin{cases} u\left(\frac{x}{\lambda}\right), & \text{if } \lambda \neq 0, \\ 0, & \text{if } \lambda = 0. \end{cases}$$

This is equivalent to say that, for every $\alpha \in [0,1]$,

$$[u+v]^\alpha = \left[(\underline{u+v})_\alpha, (\overline{u+v})_\alpha\right] = \left[\underline{u}_\alpha + \underline{v}_\alpha\ ,\ \overline{u}_\alpha + \overline{v}_\alpha\right] \tag{1}$$

 ISBN: 978-84-617-8694-7

and

$$[\lambda u]^{\alpha} = \left[(\underline{\lambda u})_{\alpha}, (\overline{\lambda u})_{\alpha}\right] = \left[\min\{\lambda \underline{u}_{\alpha}, \lambda \overline{u}_{\alpha}\}, \max\{\lambda \underline{u}_{\alpha}, \lambda \overline{u}_{\alpha}\}\right]. \tag{2}$$

Taking into account the previous operations, we have the following arithmetic operations on the set of triangular fuzzy numbers (see, for instance, [14, 8]). Given $\tilde{b} = (a, b, c)$ and $\tilde{e} = (d, e, f)$:

(i) $\tilde{b} + \tilde{e} = (a, b, c) + (d, e, f) = (a + d, b + e, c + f)$.

(ii) $\lambda\tilde{b} = (\lambda a, \lambda b, \lambda c)$, if $\lambda \geq 0$; and, $\lambda\tilde{b} = (\lambda c, \lambda b, \lambda a)$, if $\lambda < 0$.

(iii) If $\tilde{e}$ is a nonnegative triangular fuzzy number, then

$$\tilde{b}\tilde{e} = \begin{cases} (ad, be, cf), & \text{if } a \geq 0, \\ (af, be, cf), & \text{if } a < 0,\ c \geq 0, \\ (af, be, cd), & \text{if } c < 0. \end{cases} \tag{3}$$

With respect to this last multiplication operation (iii), we can find another proposal by Khan et al. [10], as follows: $\tilde{b}\tilde{e} = (ad, be, cf)$. Note that this definition coincides with a particular case in the previous operation (iii). We could think of the multiplication operation given by Khan et al. [10] is applied in the cases when $\tilde{b}$ and $\tilde{e}$ are not nonnegative, since they introduce this operation for two triangular fuzzy numbers. To this respect, if we consider $\tilde{b} = (-5, -3, -1)$ and $\tilde{e} = (-4, -2, 0)$, then $\tilde{b}\tilde{e} = (20, 6, 0)$, which does not represent a triangular fuzzy number. In fact, in the numerical example given by them [10], the variables and objetive coefficients, which appear multiplying, are nonnegative triangular fuzzy numbers. Therefore, we keep (3) as the multiplication operation through out this paper, although coincides with that given in [10] in the case of nonnegative triangular fuzzy numbers.

In order to compare two fuzzy numbers, there exist some definitions as generalization of relationship on intervals (see [19]), in the recent literature. In this regard, given $u, v \in \mathcal{F}_C$, we write their α-levels as $u_{\alpha} = [\underline{u}_{\alpha}, \overline{u}_{\alpha}] \in \mathcal{K}_C$ and $v_{\alpha} = [\underline{v}_{\alpha}, \overline{v}_{\alpha}] \in \mathcal{K}_C$, respectively, for all $\alpha \in [0, 1]$.

Definition 3 *Given $u, v \in \mathcal{F}_C$, we say that*

(i) $u \leqq v$ if and only if $\underline{u}_{\alpha} \leq \underline{v}_{\alpha}$ and $\overline{u}_{\alpha} \leq \overline{v}_{\alpha}$, for all $\alpha \in [0, 1]$,

(ii) $u \prec v$ if and only if $\underline{u}_{\alpha} < \underline{v}_{\alpha}$ and $\overline{u}_{\alpha} < \overline{v}_{\alpha}$, for all $\alpha \in [0, 1]$.

In the previous definition, the relationships given by $\leqq$ and $\prec$ are usually presented under the notation $\leqq_{LU}$ and $\prec_{LU}$ (see [19]). In a similar way, the relations $\succ$ and $\geqq$ are considered. These relationships establish partial orders in $\mathcal{F}_C$.

Remark 1 *Note that, as an inmediate consequence of Theorem 1 and Definition 3, we have that to say $u \leqq v$ and $v \leqq u$ is equivalent to say $u = v$.*

 ISBN: 978-84-617-8694-7

For convenience, we denote $\tilde{0} = (0, 0, 0)$. Observe that a triangular fuzzy number $\tilde{v}$ is nonnegative if and only if $\tilde{v} \geqq \tilde{0}$.

Proposition 1 *Given two triangular fuzzy numbers $\tilde{u} = (u^-, \hat{u}, u^+)$ and $\tilde{v} = (v^-, \hat{v}, v^+)$, it follows that*

(i) $\tilde{u} \prec \tilde{v}$ if and only if $u^- < v^-$, $\hat{u} < \hat{v}$ and $u^+ < v^+$.

(ii) $\tilde{u} \leqq \tilde{v}$ if and only if $u^- \leq v^-$, $\hat{u} \leq \hat{v}$ and $u^+ \leq v^+$.

3 Fully fuzzy linear programming problem

We consider the following formulation of a Fully Fuzzy Linear Programming Problem:

$$\begin{aligned} \text{(FFLP)} \quad \text{Minimize} \quad & \tilde{z} = \sum_{j=1}^{n} \tilde{c}_j \tilde{x}_j \\ \text{subject to} \quad & \sum_{j=1}^{n} \tilde{a}_{ij} \tilde{x}_j \geqq \tilde{b}_i, \quad i = 1, \ldots, m, \\ & \tilde{x}_j \text{ is a nonnegative fuzzy triangular number}, \quad j = 1, \ldots, n, \end{aligned}$$

where $\tilde{z}$ is the fuzzy objective function, $\tilde{c} = (\tilde{c}_1, \ldots, \tilde{c_n})$ is the fuzzy vector with the fuzzy objective function coefficients, $\tilde{x} = (\tilde{x}_1, \ldots, \tilde{x}_n)$ is the vector with the fuzzy decision variables, and $\tilde{a}_{ij}$ and $\tilde{b}_i$ are the thechnical coeficients.

We deal with (FFLP) without any kind of ranking function. And, in this regard, we define the following nondominated solution.

Definition 4 *Let $\tilde{\bar{x}}$ be a feasible solution for (FFLP). $\tilde{\bar{x}}$ is said to be a nondominated solution of (FFLP) if there does not exist a feasible solution $\tilde{x}$ for (FFLP) such that $\sum_{j=1}^{n} \tilde{c}_j \tilde{x}_j \prec \sum_{j=1}^{n} \tilde{c}_j \tilde{\bar{x}}_j$.*

Following the notation for triangular fuzzy number, we have:

$$\tilde{z} = (z^-, \hat{z}, z^+),$$

$$\tilde{x}_j = (x_j^-, \hat{x}_j, x_j^+), \quad j = 1, \ldots, n,$$

$$\tilde{c}_j = (c_j^-, \hat{c}_j, c_j^+), \quad j = 1, \ldots, n,$$

$$\tilde{a}_{ij} = (a_{ij}^-, \hat{a}_{ij}, a_{ij}^+), \quad i = 1, \ldots, m, \; j = 1, \ldots, n,$$

$$\tilde{b}_i = (b_i^-, \hat{b}_i, b_i^+), \quad i = 1, \ldots, m.$$

 ISBN: 978-84-617-8694-7

Since every $\tilde{x}_j$ is a nonnegative triangular fuzzy number, then the the multiplication role given in (3) is considered. We have three possible expressions for $\tilde{c}_j\tilde{x}_j$ depending on $\tilde{c}_j$. So, given $j \in \{1,\dots,n\}$ and $\tilde{c}_j$, then the expression for $\tilde{c}_j\tilde{x}_j$ is unique, and it is the same for all nonnegative fuzzy triangular numbers $\tilde{x}_j$. A similar remark for the multiplication $\tilde{a}_{ij}\tilde{x}_j$, that is, given $i \in \{1,\dots,m\}$, $j \in \{1,\dots,n\}$ and $\tilde{a}_{ij}$, then the expression for $\tilde{a}_{ij}\tilde{x}_j$ is unique, and it is the same for all nonnegative fuzzy triangular numbers $\tilde{x}_j$.

4 An approach as multiobjective linear programming problem

Taking into account the previous arithmetic opertations, the problem (FFLP) can be approach under the following formualation:

$$
\begin{aligned}
\text{(MLP)}\quad \text{Minimize}\quad & f(x) = (f_1(x), f_2(x), f_3(x)) = \Big(\sum_{j=1}^{n}(\tilde{c}_j\tilde{x}_j)^-, \sum_{j=1}^{n}\widehat{(\tilde{c}_j\tilde{x}_j)}, \sum_{j=1}^{n}(\tilde{c}_j\tilde{x}_j)^+\Big)\\
\text{subject to}\quad & \sum_{j=1}^{n}(\tilde{a}_{ij}\tilde{x}_j)^- \leqq b_i^-, \quad i = 1,\dots,m,\\
& \sum_{j=1}^{n}\widehat{(\tilde{a}_{ij}\tilde{x}_j)} \leqq \hat{b}_i, \quad i = 1,\dots,m,\\
& \sum_{j=1}^{n}(\tilde{a}_{ij}\tilde{x}_j)^+ \leqq b_i^+, \quad i = 1,\dots,m,\\
& x_j^- - \hat{x}_j \leqq 0, \quad j = 1,\dots,n,\\
& \hat{x}_j - x_j^+ \leqq 0, \quad j = 1,\dots,n,\\
& x_j^- \geqq 0, \hat{x}_j \geqq 0, x_j^+ \geqq 0, \quad j = 1,\dots,n,
\end{aligned}
$$

$f : \mathbb{R}^{3n} \to \mathbb{R}^3$ is a vector function, with the variable $x = (x_1^-, \hat{x}_1, x_1^+, \dots, x_n^-, \hat{x}_n, x_n^+) \in \mathbb{R}^{3n}$, with f_i linear functions, $i = 1, 2, 3$. All constraints are represented as linear inequalities on the variable x. Then, (MLP) is a multiobjective linear programming problem (for further details, see [20]).

The relationship between (FFLP) and (MLP) is as follows.

Theorem 2 *$\tilde{x} = (\tilde{x}_1,\dots,\tilde{x}_n)$ with $\tilde{x}_j = (x_j^-, \hat{x}_j, x_j^+) \in \mathcal{F}_C$, $j = 1,\dots,n$, is a nondominated solution of (FFLP) if and only if $x = (x_1^-, \hat{x}_1, x_1^+, \dots, x_n^-, \hat{x}_n, x_n^+) \in \mathbb{R}^{3n}$ is a weakly efficient solution of (MLP).*

5 Conclusions

It has been shown that the fully fuzzy linear programming problem (FFLP) can be considered equivalent to the multiobjective lineal programming problem (MLP). In this regard, a fuzzy problem has

been linked to a multiobjetive crisp linear problem without information loss, what is not usually avoided when this transformation is made via ranking functions.

References

[1] R.E. Bellman, L.A. Zadeh, *Decision making in a fuzzy environment*, Manage. Sci. **17** (1970) 141–164.

[2] L. Campos, J.L. Verdegay, em Linear programming problems and ranking of fuzzy numbers, Fuzzy Set. Syst. **32** (1989) 1–11.

[3] K. Ganesan, P. Veeramani, *Fuzzy linear programs with trapezoidal fuzzy numbers*, Ann. Oper. Res. **143** (2006) 305–315.

[4] H.R. Maleki, M. Tata, M. Mashinchi, *Linear programming with fuzzy variables*, Fuzzy Set. Syst. **109** (2000) 21–33.

[5] H.R. Maleki, *Ranking functions and their applications to fuzzy linear programming*, Far East J. Math. Sci. **4** (2002) 283–301.

[6] A. Ebrahimnejad, S.H. Nasseri, F.H. Lotfi, M. Soltanifar, *A primal-dual method for linear programming problems with fuzzy variables*, Eur. J. Ind. Eng. **4** (2010) 189–209

[7] F.H. Lotfi, T. Allahviranloo, M.A. Jondabeha, L. Alizadeh, *Solving a fully fuzzy linear programming using lexicography method and fuzzy approximate solution*, Appl. Math. Modell. **33** (2009) 3151–3156.

[8] A. Kumar, J. Kaur, P. Singh, A new method for solving fully fuzzy linear programming problems, Applied Mathematical Modelling 35 (2011) 817-823.

[9] M. Arana-Jiménez , A. Rufián-Lizana, Y. Chalco-Cano, H. Román-Flores, *Generalized convexity in fuzzy vector optimization through a linear ordering*, Information Sciences **312** (2015) 13–24.

[10] I.U. Khan, T. Ahmad, N. Maan, *A simplified novel technique for solving fully fuzzy linear programming problems*, J. Optim. Theory Appl. **159** (2013) 536-546.

[11] B. Bhardwaj, A. Kumar, *A note on the paper "A simplified novel technique for solving fully fuzzy linear programming problems*, J. Optim. Theory Appl. **163** (2014) 685-696.

[12] I.U. Khan, T. Ahmad, N. Maan, *A Reply to a Note on the Paper "A simplified Novel Technique for Solving Fully Fuzzy Linear Programming Problems"*, J. Optimiz. Theory Appl. **173** (2017) 353-356.

 ISBN: 978-84-617-8694-7

[13] L. STEFANINI, L. SORINI, M.L. GUERRA, *Parametric representation of fuzzy numbers and application to fuzzy calculus*, FUZZY SETS AND SYSTEMS **157** (18) (2006) 2423-2455

[14] A. KAUFMANN, M.M. GUPTA, *Introduction to Fuzzy Arithmetic Theory and Applications, Van Nostrand Reinhold*, NEW YORK, 1985.

[15] R. GOESTSCHEL, W. VOXMAN, *Elementary fuzzy calculus*, FUZZY SETS AND SYSTEMS **18** (1986) 31-43.

[16] R.E. MOORE, *Interval Analysis Prentice-Hall*, ENGLEWOOD CLIFFS, NJ, 1966.

[17] R.E. MOORE, *Method and Applications of Interval Analysis*, SIAM, PHILADELPHIA, 1979.

[18] G. ALEFELD, J. HERZBERGER, *Introduction to Interval Computations*, ACADEMIC PRESS, NEW YORK, 1983.

[19] M.L. GUERRA, L. STEFANINI, *A comparison index for interval based on generalized Hukuhara difference*, SOFT. COMPUT. **16** (2012) 1931-1943.

[20] M. ARANA-JIMÉNEZ (ED.), *Optimiality conditions in vector optimization*, BENTHAM SCIENCE PUBLISHERS, LTD., BUSSUM, 2010.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

Numerical methods for nonlinear option pricing models with variable transaction costs

Iñigo Arregui[1], Daniel Ševčovič[2] and Carlos Vázquez[1]

[1] *Dept. of Mathematics, University of A Coruña, Spain*

[2] *Dept. of Applied Mathematics and Statistics, Comenius University, Slovakia*

emails: arregui@udc.es, sevcovic@fmph.uniba.sk, carlosv@udc.es

Abstract

The classical Black-Scholes equation for options pricing exhibits several limitations when applied to real markets in certain conditions. In many settings, the consideration of a constant volatility is no more realistic. In the present paper, we consider the case where the volatility is assumed to depend on the product of the asset price and the second derivative of the option with respect to the asset price (an option Greek which is known as Gamma). This hypothesis has been made in models that incorporate transaction costs, market feedback effects related to stocks trading strategies or illiquid markets, risks related to unprotected portfolios, etc. In these settings, the corresponding nonlinear Black-Scholes equation can be transformed into a quasilinear equation (Gamma equation) in a new unknown variable related to the Gamma of the option.

Once this semilinear Gamma equation has been obtained, we propose a duality method, combined with a characteristics scheme and finite elements methods. The duality method is applied to the maximal monotone operator that governs the nonlinear term in the Gamma equation. By a suitable numerical integration technique the value of the European option can be recovered. Finally, we present some examples of European options to show the good performance of the new numerical global strategy.

Key words: option pricing, option gamma, nonlinear Black–Scholes, duality methods, finite elements

1 Introduction

The classical linear Black–Scholes presented in 1973 establishes that the price V of an option can be obtained as the solution of the parabolic equation:

$$\partial_t V + \frac{\sigma^2}{2} S^2 \, \partial_S^2 V + rS\partial_S V - rV = 0\,, \tag{1}$$

 ISBN: 978-84-617-8694-7

where $r > 0$ denotes the risk–free interest rate and σ is the (constant) volatility of the underlying asset, the price of which is assumed to be a stochastic process that follows the stochastic differential equation

$$dS_t = \mu S_t\, dt + \sigma S_t\, dW_t\,,$$

the constant μ being the drift and the process W_t a geometric Brownian motion. Note that the option price, V_t, is a process that is obtained from the solution V of (1), by the expression $V_t = V(t, S_t)$. Equation (1) has been derived under several restrictive assumptions.

However, from the analysis of market data, the need of more realistic models arises. For example, in several setting, different models assume nonconstant volatility expressions that depend on the gamma of the option in the form:

$$\sigma = \hat{\sigma}(S\,\partial_S^2 V)\,,$$

so that the following nonlinear Black–Scholes equation is posed:

$$\partial_t V + \frac{1}{2}\hat{\sigma}(S\partial_S^2 V)^2 S^2\,\partial_S^2 V + rS\partial_S V - rV = 0. \tag{2}$$

For example, this kind of dependency arises in option pricing models that take into account non-trivial variable transaction costs related to assets buying and selling [1, 2, 10], market feedback effects due to large traders choosing given stock–trading strategies [6, 7], risk from volatile and unprotected portfolios [8], or investor preferences [4], among others.

2 Mathematical model

In this section, we just remind a result by Ševčovič and Žitňanská [11] that establishes the equivalence between the nonlinear Black–Scholes equation (2) and a quasilinear parabolic equation. For this purpose, we introduce the function

$$\beta(H) = \frac{\hat{\sigma}(H)^2}{2} H\,.$$

Proposition (Ševčovič–Žitňanská, [11]) *Assume the function $V = V(S,t)$ is a solution to the nonlinear Black–Scholes equation*

$$\partial_t V + S\,\beta(S\partial_S^2 V) + rS\partial_S V - rV = 0\,, \qquad S > 0\,, \quad t \in (0,T)\,. \tag{3}$$

Then the transformed function $H = H(x,\tau) = S\partial_S^2 V(S,t)$, where $x = \ln(S/E)$, $\tau = T - t$, is a solution to the quasilinear parabolic (Gamma) equation:

$$\partial_\tau H = \partial_x^2 \beta(H) + \partial_x \beta(H) + r\partial_x H\,. \tag{4}$$

 ISBN: 978-84-617-8694-7

On the other hand, if H is a solution to (4) such that $H(-\infty,\tau) = \partial_x H(-\infty,\tau) = 0$ and $\beta'(0)$ is finite, then the function

$$V(S,t) = aS + b\, exp(-r(T-t)) + \int_{-\infty}^{+\infty} (S - E\, exp(\xi))^+ H(\xi, T-t)\, d\xi \tag{5}$$

is a solution to the nonlinear Black–Scholes equation (3) for any $a, b \in \mathbb{R}$.

Moreover, if we consider the initial condition

$$H(x,0) = \delta(x)\,, \tag{6}$$

where $\delta(x)$ denotes the Dirac delta function with basis point x, then we can recover the payoffs of the European vanilla options with the choices:

- $a = b = 0$, for the call option (i.e. $V(S,T) = (S-E)^+$),
- $a = -1$, $b = E$, for the put option (i.e. $V(S,T) = (E-S)^+$),

the constant E being the strike price. As the analytical expression for the solution of (4) is not available, in next section we propose a set of numerical techniques for its approximation (see e.g. [8, 9, 11] for other numerical strategies).

3 Numerical solution of the quasilinear parabolic equation

In order to solve numerically the nonlinear equation (4) jointly with the initial condition (6), we note the main difficulties: the possibility of advection term dominating diffusion one, the nonlinear diffusion term, the presence of an unbounded domain and the Dirac delta function in the initial condition. First, as in other problems in which advection can dominate over diffusion, we propose the characteristics method for the time discretization. Secondly, as the nonlinear term can be related to maximal monotone operators [3], we make use of a duality method based on a result in [5]. In order to deal with the unbounded domain, as usually in financial problems, we propose a domain truncation in the asset variable by taking $S_\infty = 4E$, which corresponds to x_∞ as the upper finite boundary of the computational domain in x. Also we consider x_0 as the lower boundary of this domain $\Omega = (x_0, x_\infty)$. Concerning to the Dirac delta function, we approximate it by a Gaussian density. Finally, a finite element method is proposed for the discretization in the spatial-like variable x at each time step.

So, first following [3, 5], we introduce the parameter $\omega > 0$ and a new variable θ given in terms of the function β^ω by:

$$\theta = \beta^\omega(H) = \beta(H) - \omega H\,. \tag{7}$$

 ISBN: 978-84-617-8694-7

As $\beta(H) = \theta + \omega H$, equation (4) can be equivalently written as:

$$H_\tau - (\omega + r)H_x - \omega H_{xx} = \theta_x + \theta_{xx}\,. \tag{8}$$

Next, in order to apply the method of characteristics, we introduce the material derivative of function H:

$$\frac{DH}{D\tau} = H_\tau - (\omega + r)\,H_x\,, \tag{9}$$

which represents the derivative associated to the constant scalar velocity field $-(\omega + r)$, so that (8) turns into:

$$\frac{DH}{D\tau} - \omega H_{xx} = \theta_x + \theta_{xx}\,. \tag{10}$$

Note that (10) is still a nonlinear problem, as θ and H are related by (7).

In order to discretize (10) in time by the characteristics (also known as semilagrangian) method, we introduce the time stepsize $\Delta\tau > 0$ and mesh points in time $\tau^n = n\Delta\tau$ for $n = 0, 1, 2, \ldots$, so that we consider the following final value problem:

$$\begin{cases} \dfrac{d\chi}{d\tau} = -(\omega + r)\,\chi(\tau) \\ \chi(\tau^{n+1}) = x\,, \end{cases}$$

that provides the characteristics curve (associated to the scalar velocity field) passing through the point x at time τ^{n+1}. Its analytical solution provides the position at time τ^n to be used in the characteristics method:

$$\chi^n(x) = \chi(x, \tau^{n+1}; \tau^n) = x\exp((\omega + r)\Delta\tau)\,.$$

We can now approximate the material derivative in (10) by a first order upwinded quotient. If we denote $H^n(\cdot) = H(\cdot, \tau^n)$, then (10) is approximated by:

$$\frac{H^{n+1} - H^n \circ \chi^n}{\Delta\tau} - \omega H_{xx}^{n+1} = \theta_x^{n+1} + \theta_{xx}^{n+1}\,. \tag{11}$$

We will consider homogeneous Dirichlet boundary conditions on $\partial\Omega$, i.e. $H(x_0) = H(x_\infty) = 0$. Thus, the variational formulation of (11) consists in finding $H^{n+1} \in W_0^{1,2}(\Omega)$, such that:

$$\int_\Omega H^{n+1}\varphi - \Delta\tau\omega\int_\Omega H_{xx}^{n+1}\varphi = \int_\Omega (H^n \circ \chi^n)\varphi + \Delta\tau\int_\Omega \theta_x^{n+1}\varphi + \Delta\tau\int_\Omega \theta_{xx}^{n+1}\varphi\,, \quad \forall\varphi \in W_0^{1,2}(\Omega)$$

where $W_0^{1,2}(\Omega)$ stands for the classical notation of Sobolev spaces. Next, using Green's theorem, we get:

$$\begin{aligned}\int_\Omega H^{n+1}\varphi + \Delta\tau\omega\int_\Omega H_x^{n+1}\varphi_x = \int_\Omega (H^n \circ \chi^n)\varphi + \Delta\tau\int_\Omega \theta_x^{n+1}\varphi - \Delta\tau\int_\Omega \theta_x^{n+1}\varphi_x \\ + \Delta\tau\omega\int_{\partial\Omega} H_x^{n+1}\varphi + \Delta\tau\int_{\partial\Omega}\theta_x^{n+1}\varphi\,.\end{aligned}$$

 ISBN: 978-84-617-8694-7

Taking into account the homogeneous boundary conditions, the two integrals on $\partial\Omega$ vanish and we get:

$$\int_\Omega H^{n+1}\varphi + \Delta\tau\omega \int_\Omega H_x^{n+1}\varphi_x = \int_\Omega (H^n \circ \chi^n)\varphi + \Delta\tau \int_\Omega \theta_x^{n+1}\varphi - \Delta\tau \int_\Omega \theta_x^{n+1}\varphi_x \,, \qquad (12)$$

jointly with the relation:

$$\theta^{n+1} = \beta^\omega(H^{n+1}) \,. \qquad (13)$$

We propose the following fixed point algorithm to solve (12)-(13) at each time instant τ^{n+1}. Assume $(H^{n+1,0}, \theta^{n+1,0})$ is given. Then, for $k = 0, 1, \ldots$

- For given $(H^{n+1,k}, \theta^{n+1,k})$, we search $H^{n+1,k+1}$ as the solution of the linear problem

$$\int_\Omega H^{n+1,k+1}\varphi + \Delta\tau\omega \int_\Omega H_x^{n+1,k+1}\varphi_x = \int_\Omega (H^n \circ \chi^n)\varphi + \Delta\tau \int_\Omega \theta_x^{n+1,k}\varphi - \Delta\tau \int_\Omega \theta_x^{n+1,k}\varphi_x \qquad (14)$$

  for all $\varphi \in W_0^{1,2}(\Omega)$.

- We update $\theta^{n+1,k+1}$ by solving the nonlinear equation (13). As the exact solution is not available in most cases, we make use the theory of maximal monotone operators as in [5] and propose the updating:

$$\theta^{n+1,k+1} = \beta_\lambda^\omega(H^{n+1,k+1} + \lambda\theta^{n+1,k}) \,, \qquad (15)$$

  where β_λ^ω denotes the Yosida regularization of function β^ω with parameter λ:

$$\beta_\lambda^\omega(H) = \inf_G \left(\beta^\omega(G) + \frac{(G-H)^2}{2\lambda} \right) .$$

  Moreover, for convergence reasons, we choose $\lambda = 1/(2\omega)$.

We note that Yosida regularization is strongly dependent on the function β and requires the computation of the inverse of an operator. Therefore, it is not always possible to get its analytical expression. This is the reason why we replace (15) by first order Taylor expansion:

$$\begin{aligned}
\theta^{n+1,k+1} &= \beta_\lambda^\omega(H^{n+1,k+1} + \lambda\theta^{n+1,k}) \\
&= \beta^\omega(H^{n+1,k+1} + \lambda\theta^{n+1,k} - \lambda\theta^{n+1,k+1}) \\
&= \beta^\omega\left(H^{n+1,k+1} + \lambda(\theta^{n+1,k} - \theta^{n+1,k+1})\right) \\
&= \beta^\omega\left(H^{n+1,k+1}\right) + (\beta^\omega)'\left(H^{n+1,k+1}\right)\lambda(\theta^{n+1,k} - \theta^{n+1,k+1}) \\
&\quad + o\left(\lambda^2(\theta^{n+1,k} - \theta^{n+1,k+1})^2\right) \,, \qquad (16)
\end{aligned}$$

 ISBN: 978-84-617-8694-7

which does not require the computation of the Yosida regularization and is accurate enough if λ is small. From (16), we deduce:

$$\theta^{n+1,k+1}\left[1+(\beta^\omega)'(H^{n+1,k+1})\lambda\right]=\beta^\omega(H^{n+1,k+1})+(\beta^\omega)'(H^{n+1,k+1})\lambda\theta^{n+1,k}$$

so that:

$$\theta^{n+1,k+1}=\frac{\beta^\omega(H^{n+1,k+1})+(\beta^\omega)'(H^{n+1,k+1})\lambda\theta^{n+1,k}}{1+\lambda(\beta^\omega)'(H^{n+1,k+1})}\,.$$

Finally, taking into account that $\beta^\omega(H)=\beta(H)-\omega H$ we obtain:

$$\begin{aligned}\theta^{n+1,k+1}&=\frac{\beta(H^{n+1,k+1})-\omega H^{n+1,k+1}+\left[\beta'(H^{n+1,k+1})-\omega\right]\lambda\theta^{n+1,k}}{1+\lambda\left[\beta'(H^{n+1,k+1})-\omega\right]}\\&=\frac{\beta(H^{n+1,k+1})+\beta'(H^{n+1,k+1})\lambda\theta^{n+1,k}-\omega\left[H^{n+1,k+1}+\lambda\theta^{n+1,k}\right]}{1-\omega\lambda+\lambda\beta'(H^{n+1,k+1})}\,.\end{aligned}\tag{17}$$

The last expression is used instead of (15) to update θ^{n+1}. Let us remark that the first derivative of β is used in (14). In practice, it is approximated by a second order central differences formula. If the function β is not differentiable, it can be replaced by a regularized function $\hat{\beta}$.

For solving (14), we implement a finite element method. Thus, for a fixed natural number $M>0$, we consider a uniform mesh of the computational domain $\Omega=[x_0,x_\infty]$, the nodes of which are $x_j=x_0+j\Delta x$, $j=0,\ldots,M+1$, where $\Delta x=(x_\infty-x_0)/(M+1)$ denotes the constant mesh step. Associated to this uniform mesh a piecewise linear Lagrange finite elements discretization is considered.

More precisely, we search $H_h^{n+1,k+1}\in W_{0,h}$ such that:

$$\int_\Omega H_h^{n+1,k+1}\varphi+\Delta\tau\omega\int_\Omega H_{h,x}^{n+1,k+1}\varphi_x=\int_\Omega(H_h^n\circ\chi^n)\varphi+\Delta\tau\int_\Omega\theta_x^{n+1,k}\varphi-\Delta\tau\int_\Omega\theta_x^{n+1,k}\varphi_x\,,$$

for all $\varphi\in W_{0,h}$, where the space of finite elements is:

$$W_{0,h}=\left\{v_h:\Omega\to\mathbb{R}\,/\;\;v_h|_{[x_k,x_{k+1}]}\in\mathcal{P}_1\text{ for }k=0,1,\ldots,M,\;\;v_h=0\text{ on }\partial\Omega\right\}\,,$$

$\mathcal{P}_1$ being the space of polynomials of degree less or equal than one. The coefficients of the matrix and right hand side vector defining the linear system associated to the fully discretized problem are approximated by adequate quadrature formulae, when necessary. In particular, a five nodes Gaussian formula has been used. Finally, the system of linear equations is solved by a conjugate gradient method.

Once the function H is approximated at each time instant, we can recover the value of the derivative by means of (5), where $a=b=0$ for a call option and $a=-1, b=E$ for a put option.

 ISBN: 978-84-617-8694-7

4 Numerical results

In this section we present a numerical result concerning Amster *et al* model [1, 2], in which the nonlinear function β is given by

$$\beta(H) = \frac{\sigma^2}{2}(H - \mathrm{Le}|H| + \kappa H^2)\,,$$

with $\sigma = 0.95$, $\kappa = 0.10$ and $\mathrm{Le} = 0.30$. As the function β is not differentiable due to the presence of the absolute value, we introduce the regularized function β_ϵ:

$$\beta_\epsilon(H) = \frac{\sigma^2}{2}(H - \mathrm{Le}\, f_\epsilon(H) + \kappa H^2)\,.$$

The function f_ϵ is a smooth approximation of the absolute value function and its first derivative is given by:

$$f'_\epsilon(H) = \begin{cases} -1\,, & \text{if } H < -\epsilon \\ s(H)\,, & \text{if } -\epsilon \le H \le \epsilon \\ 1\,, & \text{if } H > \epsilon \end{cases}$$

s being a cubic spline and $\epsilon = 10^{-3}$.

We have considered the case of European call and put options, the payoff of which is given in terms of the strike price $E = 100$. Moreover, we have taken the risk-free interest rate $r = 0.05$ and the maturity $T = 4$.

For the numerical solution, the time domain has been discretized in 800 steps, thus $\Delta\tau = 0.005$ and the spatial variable x is in $[-4, 1.4]$, for which we have considered a uniform mesh consisting of 1601 nodes. Figure 1 shows the payoff of the call option as well as the solution at time $t = 0$ (or $\tau = 4$), while Figure 2 shows analogous results for the put option.

5 Conclusions

A nonlinear model for derivatives pricing is solved by a numerical strategy including duality methods based on maximal monotone operators, characteristics methods for time discretization and finite elements. The method is independent of the nonlinear function β.

Acknowledgements

First and third authors have been partially supported by Spanish Government (Ministerio de Economía y Competitividad, project MTM2016-76497-R) and Xunta de Galicia (Grupos de Referencia Competitiva 2014). The second author has been supported by grant VEGA 1/0780/15.

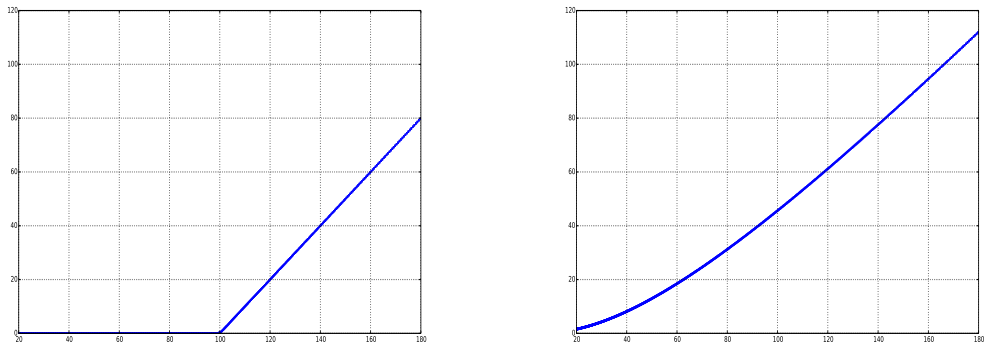


Figure 1: Call option. The terminal condition (payoff) and numerical solution

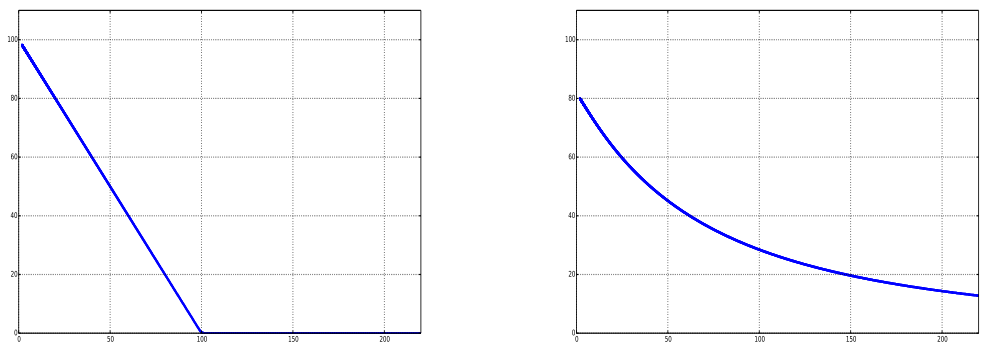


Figure 2: Put option. The terminal condition (payoff) and numerical solution

References

[1] P. Amster, C. G. Averbuj, M. C. Mariani, D. Rial, *A Black–Scholes option pricing model with transaction costs*, J. Math. Anal. Appl. **303** (2005) 688-695.

[2] P. Amster, A. P. Mogni, *On a pricing problem for a multi-asset option with general transaction costs*, arXiv: 1704.02036 [q-fin.CP].

[3] I. Arregui, J. J. Cendán, C. Vázquez, *A duality method for the compressible Reynolds equation. Application to simulation of read/write processes in magnetic stor-*

age devices, J. Comput. Appl. Math. **175** (2005) 1, 31-40.

[4] G. Barles, H. M. Soner, *Option pricing with transactions costs and a nonlinear Black–Scholes equation*, Finance Stochast. **2** (1998) 369-397.

[5] A. Bermúdez, C. Moreno, *Duality methods for solving variational inequalities*, Comput. Math. Appl. **7** (1981) 43-58.

[6] R. Frey, P. Patie, *Risk management for derivatives in illiquid markets: A simulation study*, In: Advances in Finance and Stochastics, Springer, Berlin, 2002.

[7] R. Frey, A. Stremme, *Market volatility and feedback effects from dynamic hedging*, Mathematical Finance **4** (1997) 351-374.

[8] M. Jandačka, D. Ševčovič, *On the risk adjusted pricing methodology based valuation of vanilla options and explanation of the volatility smile*, Journal of Applied Mathematics **3** (2005) 235-258.

[9] M. N. Koleva, L. G. Vulkov, *A second–order positivity preserving numerical method for Gamma equation*, Appl. Math. Comput. **220** (2013) 722-734.

[10] H. E. Leland, *Option pricing and replication with transaction costs*, Journal of Finance **40** (1985) 1283-1301.

[11] D. Ševčovič, M. Žitňanská, *Analysis of the nonlinear option pricing model under variable transaction costs*, Asia-Pacific Financial Markets **23** (2016) 2, 153-174.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

A Fast and Stable Square Root Free Unitary QR Algorithm

Jared L. Aurentz[1], Thomas Mach[2], Raf Vandebril[3] and David S. Watkins[4]

[1] *Instituto de Ciencias Matemáticas, Universidad Autónoma de Madrid*

[2] *Department of Mathematics, Nazarbayev University*

[3] *Department of Computer Science, KU Leuven*

[4] *Department of Mathematics, Washington State University*

emails: jared.aurentz@icmat.es, thomas.mach@nu.edu.kz,
raf.vandebril@cs.kuleuven.be, watkins@math.wsu.edu

Abstract

In 1968 Pal, Walker and Kahan presented a stable implementation of the QR algorithm for symmetric tridiagonal matrices that avoided computing square roots. Removing the square roots significantly lowers the computation time making it superior to the traditional QR algorithm when only the eigenvalues are needed. Here we present a square root free version of the QR algorithm for unitary upper-hessenberg matrices. This method is faster than the traditional version and just as stable making it competitive when only the eigenvalues are needed.

1 Introduction

For an $n \times n$ symmetric tridiagonal matrix there exists a number of stable methods for computing all n eigenvalues in $O(n^2)$ flops. An important example is the symmetric tridiagonal QR algorithm. While this algorithm has the same $O(n^2)$ complexity it requires noticeably more arithmetic than competing LR and divide-and-conquer methods. In 1968 Pal, Walker and Kahan [3, p. 178] showed that one can reorder the arithmetic in the symmetric tridiagonal QR algorithm to avoid the most costly arithmetic operation: square roots. Casting out the square roots significantly reduced the total computation time making their algorithm one of the most competitive methods for solving this class of problems.[1] It was also shown that removing the square roots does not affect stability.

[1] This algorithm is implemented as `xSTERF` in LAPACK.

 ISBN: 978-84-617-8694-7

In 1986 Gragg [2] presented a variant of Francis' implicitly-shifted QR algorithm [1] for computing all the eigenvalues of a unitary upper-hessenberg matrix in $O(n^2)$ flops. Just as in the symmetric case, Gragg's method requires the computation of costly square roots. In this paper we illustrate how one can rearrange the arithmetic to completely remove all square roots from the algorithm while preserving the stability.

2 Unitary QR algorithm

Given an $n \times n$ unitary upper-hessenberg matrix U with real subdiagonal entries, Gragg's algorithm begins by factoring U into a product of n essentially 2×2 matrices, $U = U_1 U_2 \cdots U_n$. In the case $n = 3$ we have,

$$\begin{bmatrix} u_1 & -v_1 u_2 & v_1 v_2 u_3 \\ v_1 & \overline{u}_1 u_2 & -\overline{u}_1 v_2 u_3 \\ & v_2 & \overline{u}_2 u_3 \end{bmatrix} = \begin{bmatrix} u_1 & -v_1 & \\ v_1 & \overline{u}_1 & \\ & & 1 \end{bmatrix} \begin{bmatrix} 1 & & \\ & u_2 & -v_2 \\ & v_2 & \overline{u}_2 \end{bmatrix} \begin{bmatrix} 1 & & \\ & 1 & \\ & & u_3 \end{bmatrix},$$

where $|u_1|^2 + v_1^2 = 1$, $|u_2|^2 + v_2^2 = 1$ and $|u_3|^2 = 1$. Such a factorization is always possible.

Given a complex shift ρ, $|\rho| = 1$, a single iteration of Gragg's algorithm begins by constructing a unitary matrix Q_1 such that $Q_1 e_1 = \alpha(U - \rho I)e_1$, where $e_1 = [1, 0, \ldots, 0]^T$ and α is a nonzero complex number, and applying the similarity transformation $U^{(1/2)} = Q_1^* U Q_1$. To construct such a matrix Q_1 when $n = 3$ we set $q_1 = u_1 - \rho$, $p_1 = v_1$, $m_1 = \sqrt{|q_1|^2 + p_1^2}$ and take Q_1 as

$$Q_1 = \begin{bmatrix} q_1/m_1 & -p_1/m_1 & \\ p_1/m_1 & \overline{q}_1/m_1 & \\ & & 1 \end{bmatrix}.$$

From the definitions of q_1 and p_1 we can also express Q_1 as follows,

$$Q_1 = \left(\begin{bmatrix} u_1 & -v_1 & \\ v_1 & \overline{u}_1 & \\ & & 1 \end{bmatrix} + \begin{bmatrix} -\rho & & \\ & -\overline{\rho} & \\ & & 1 \end{bmatrix} + \begin{bmatrix} 0 & & \\ & 0 & \\ & & -1 \end{bmatrix} \right) \begin{bmatrix} m_1^{-1} & & \\ & m_1^{-1} & \\ & & 1 \end{bmatrix}.$$

Taking

$$R_1 = \begin{bmatrix} -\rho & & \\ & -\overline{\rho} & \\ & & 1 \end{bmatrix}, \quad J_1 = \begin{bmatrix} 0 & & \\ & 0 & \\ & & -1 \end{bmatrix} \quad \text{and} \quad M_1 = \begin{bmatrix} m_1^{-1} & & \\ & m_1^{-1} & \\ & & 1 \end{bmatrix}$$

we have that $Q_1 = (U_1 + R_1 + J_1)M_1 = M_1(U_1 + R_1 + J_1)$, where the second equality comes from the fact that M_1 commutes with R_1, J_1 and U_1.

Continuing with our 3×3 example we see that Q_1 commutes with U_3 which gives $U^{(1/2)} = Q_1^* U Q_1 = Q_1^* U_1 U_2 Q_1 U_3$. Using the definition of Q_1 one can show that $Q_1^* U_1 =$

 ISBN: 978-84-617-8694-7

$R_1^*Q_1$. This gives $U^{(1/2)} = R_1^*Q_1U_2Q_1U_3$. To finish the initialization we move R_1^* to the other side by applying the similarity transform $U^{(1)} = R_1U^{(1/2)}R_1^* = Q_1U_2Q_1R_1^*U_3$. The remainder of the iteration involves chasing the matrices Q_1 and R_1^* down the matrix by performing $n-1$ turnovers.

2.1 The turnover

To proceed with Gragg's algorithm we must perform a turnover on the product $Q_1U_2Q_1R_1^*$. A *turnover* is a refactorization of the product $Q_1U_2Q_1R_1^*$. The following lemma shows how this can be done.

Lemma 1. *Let $q_1, \gamma_1, u_2, \rho \in \mathbb{C}$ and $p_1, m_1, v_2 \in \mathbb{R}$ satisfy,*

$$|q_1|^2 + p_1^2 = m_1^2, \quad |\gamma_1| = 1, \quad |u_2|^2 + v_2^2 = 1, \quad |\rho| = 1,$$

and let U_2, Q_1 and R_1 be as above and let $\tilde{Q}_1$ be as follows,

$$\tilde{Q}_1 = \begin{bmatrix} \gamma_1 q_1/m_1 & -p_1/m_1 & \\ p_1/m_1 & \overline{\gamma}_1\overline{q}_1/m_1 & \\ & & 1 \end{bmatrix}.$$

Then the product $\tilde{Q}_1U_2Q_1R_1^*$ can be refactored as the product $Q_2R_2^*U_1\tilde{Q}_2$ where

$$\hat{U}_1 = \begin{bmatrix} \hat{u}_1 & -\hat{v}_1 & \\ \hat{v}_1 & \overline{\hat{u}}_1 & \\ & & 1 \end{bmatrix}, \quad R_2 = \begin{bmatrix} 1 & & \\ & -\rho & \\ & & -\overline{\rho} \end{bmatrix}, \quad Q_2 = \begin{bmatrix} 1 & & \\ & q_2/m_2 & -p_2/m_2 \\ & p_2/m_2 & \overline{q}_2/m_2 \end{bmatrix},$$

and

$$\tilde{Q}_2 = \begin{bmatrix} 1 & & \\ & \gamma_2 q_2/m_2 & -p_2/m_2 \\ & p_2/m_2 & \overline{\gamma}_2\overline{q}_2/m_2 \end{bmatrix},$$

with

$$\begin{aligned} q_2 &= q_1 + u_2\overline{\gamma}_1\overline{q}_1, \\ p_2 &= m_1v_2, \\ m_2 &= \sqrt{|q_2|^2 + p_2^2}, \\ \gamma_2 &= -\rho\gamma_1, \\ \hat{u}_1 &= \overline{\rho}(p_1^2u_2 - \gamma_1q_1^2)/m_1^2, \\ \hat{v}_1 &= m_2v_1/m_1. \end{aligned}$$

Proof. The result follows by direct comparison of the products $\tilde{Q}_1U_2Q_1R_1^*$ and $Q_2R_2^*U_1\tilde{Q}_2$. □

 ISBN: 978-84-617-8694-7

We can compute the turnover for $Q_1U_2Q_1R_1^*$ by setting $\gamma_1 = 1$ and applying the previous lemma. The algorithm continues by applying the similarity transform $Q_2R_2^*$ to $U^{(1)}$ to get $U^{(2)} = \hat{U}_1\tilde{Q}_2U_3Q_2R_2^*$. Since $\tilde{Q}_2U_3Q_2R_2^*$ has the same form as the above lemma, only one row lower and one column to the right, we can perform another turnover. This continues until we reach the last row and column of the matrix.

2.2 Casting out the square roots

To cast out the square roots we simply need to store and update the quantities v_i^2, $\hat{v}_i^2$, p_i^2 and m_i^2 for $i = 1, 2, \ldots, n$. The turnover equations now look as follows

$$\begin{aligned} q_2 &= q_1 + u_2\overline{\gamma}_1\overline{q}_1, \\ p_2^2 &= m_1^2v_2^2, \\ m_2^2 &= |q_2|^2 + p_2^2, \\ \gamma_2 &= -\rho\gamma_1, \\ \hat{u}_1 &= \overline{\rho}(p_1^2u_2 - \gamma_1q_1^2)/m_1^2, \\ \hat{v}_1^2 &= m_2^2v_1^2/m_1^2. \end{aligned}$$

2.3 Stability

With a change variables one can show that the square root free turnover is the exact turnover of a nearby matrix, which means the square root free turnover is normwise backward stable. The stability of the turnover implies that a square root free iteration of Gragg's algorithm is also backward stable and thus the computed eigenvalues are the exact eigenvalues of a nearby matrix.

Acknowledgements

Jared L. Aurentz acknowledges financial support from the Spanish Ministry of Economy and Competitiveness, through the Severo Ochoa Programme for Centres of Excellence in R&D (SEV-2015-0554).

References

[1] J. G. F. Francis, *The QR transformation: a unitary analogue to the LR transformation. I.*, Comput. J. **4** (1961) 265–271.

[2] W. B. Gragg, *The QR algorithm for unitary Hessenberg matrices*, J. Comput. Appl. Math. **16** (1986) 1–8.

[3] B. Parlett, *The Symmetric Eigenvalue Problem*, SIAM, Philadelphia, 1998.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference on Computational and Mathematical Methods in Science and Engineering, CMMSE 2017 4–8 July, 2017.

A stochastic mathematical model of pre-diagnostic glioma growth based on blood glucose levels

L.E. Ayala-Hernández[1], A. Gallegos[1], J. E. Macías-Díaz[2], M.L. Miranda-Beltrán[1] and H. Vargas-Rodríguez[1]

[1] *Centro Universitario de los Lagos, Universidad de Guadalajara*

[2] *Departamento de Matemáticas y Física, Universidad Autónoma de Aguascalientes*

emails: luis.ayala@alumno.udg.mx, gallegos@culagos.udg.mx, jemacias@correo.uaa.mx, lmiranda@culagos.udg.mx, hvargas@culagos.udg.mx

Abstract

In this paper, we propose a stochastic mathematical model in which the values of the factors involved in the development of a glioma, vary randomly in a biologically congruent range. Stability analysis revealed three fixed points, which allude to a growing glioma, an advanced glioma and a patient without glioma. The graphics of the solutions are presented, we also show the order of influence of the parameters. The results obtained disclose a decay in serum glucose levels when there is glioma, moreover they indicate that glucose consumption by glioma is an important element for its growth.
Key words: Pre-diagnostic glioma, glucose decay, nonlinear differential equations, stochastic noise, dynamical system.

1 Introduction

The most common primary brain tumors are gliomas, believed to originate in glial cells, which are the supporting structure of neurons [1]. The aggressive conduct of these tumors is reflected in their mortality rate of almost 100 % within six to twelve months after being diagnosed [2, 3]. Surgery is the main treatment option and is usually followed by chemotherapy and radiotherapy, but these treatments usually only prolong the patient's life for a short time [4]. Cell culture studies and recent data from glioma patients have revealed that patients with low serum glucose levels have a higher risk of developing gliomas [5]. In this work, a stochastic mathematical model is used to analyze the behavior of glioma and blood

 ISBN: 978-84-617-8694-7

glucose levels when the parameters involved take values with random noise, which shows which parameters have a greater influence on glioma growth and decreased blood glucose levels.

2 The mathematical model

Our start point is the mathematical model proposed in [4] which consider a simplified process of a small glioma growing in the brain which elicits a response from the host immune system. Both the host immune system and the glioma require energy to sustain their functions. Therefore, we also keep track of an energy source, specifically in the form of glucose which can exist in the brain or blood. The model consists of four variables denoted T, σ_{brain}, I, σ_{serum}, which represent the concentration of glioma cells, the concentration of glucose in the brain, the concentration of immune system cells and the concentration of serum glucose levels respectively. We describe this assumptions in the following system of differential equations (the set of values of parameters is presented in table 1 along with descriptions and a reference):

$$\frac{dT}{dt} = \underbrace{\alpha_T \sigma_{brain} T(1 - \frac{T}{K_T})}_{Production} - \underbrace{d_T T}_{Apoptosis} - \underbrace{d_{TI} TI}_{Immune\ response} \tag{1}$$

$$\frac{d\sigma_{brain}}{dt} = \underbrace{\alpha_\sigma(\sigma_{serum} - \sigma_{brain})}_{Glucose\ exchange} - \underbrace{d_{T\sigma} T \sigma_{brain}}_{Glioma\ consumption} - \underbrace{(d_{\sigma 1} + \alpha_s(\nu + I))\sigma_{brain}}_{Natural\ consumption} \tag{2}$$

$$\frac{dI}{dt} = \underbrace{\alpha_s(\nu + I)\sigma_{brain} + \alpha_{TI} TI}_{Production} - \underbrace{d_I I}_{Natural\ decay} - \underbrace{d_{TT} TI}_{Glioma\ response} \tag{3}$$

$$\frac{d\sigma_{serum}}{dt} = \underbrace{\alpha_\sigma(\sigma_{brain} - \sigma_{serum})}_{Glucose\ exchange} + \underbrace{F(t)}_{Glucose\ intake} - \underbrace{d_{\sigma 2}\sigma_{serum}}_{Natural\ consumption} \tag{4}$$

Equation 1 governs the temporal evolution of glioma growth , equation 2 describe the glucose concentration in the brain, equation 3 models immune system activity in the brain and equation 4 explain the behavior of glucose concentration in serum.

 ISBN: 978-84-617-8694-7

Parameter	Description	Value	Source
α_T	Growth rate of glioma	1.575 ($ml^2g^{-1}day^{-1}$)	[4]
K_T	Carrying capacity of glioma	2 (g/ml)	[4]
d_{TI}	Decay rate of glioma due to immune response	0.072 (day^{-1})	[4]
α_{TI}	Recruitment rate of immune systems cells due to glioma	0.0003 (day^{-1})	[4]
d_T	Natural decay rate of glioma	0.0001 (day^{-1})	[4]
d_I	Natural decay rate of immune system cells	0.01 (day^{-1})	[4]
α_s	Immune system cell recruitment rate	0.7 (day^{-1})	[4]
ν	Baseline immune system cell production rate	0.7 (day^{-1})	[4]
$d_{T\sigma}$	Glucose consumption rate by glioma	1 (day^{-1})	[4]
α_σ	Transfer rate of glucose from serum to brain	20 (day^{-1})	[4]
σ_{min}	Minimum glucose intake rate to serum	0.0008 (g/ml)	[4]
σ_0	Maximum variation in glucose intake rate	0.0016 (g/ml)	[4]
$d_{\sigma 1}$	Glucose consumption in brain by healthy cells	0.01 (day^{-1})	[4]
$d_{\sigma 2}$	Metabolic glucose consumption in serum	0.00625 (day^{-1})	Estimate
d_{TT}	Rate of glioma cells killing immune cells	0.072 (day^{-1})	[4]

Table 1: Model parameter values and description

3 Stability analysis

The stability analysis of the model yielded six fixed points. Three of them were biologically unrealistic due to the negative sign of some of its variables, so they were discarded. The other three were used as initial conditions in the numerical simulations, which refer to a patient with a growing glioma (initial conditions 1), a patient with an advanced stage glioma (initial conditions 2) and a patient without glioma (initial conditions 3). The Lyapunov method is applied to know the stability of each of the biologically real fixed points, resulting in that the fixed points 1 and 2 are unstable states, while the fixed points 3 is a stable state.

4 Numerical simulations

We solved the system of equations for a time corresponding to 9 years, adding stochastic noise according to the following expression:

$$\varepsilon_i + \beta * \varepsilon_i * \Delta B \tag{5}$$

where ε_i corresponds to each parameter presented in table 1, β is the noise intensity, ΔB is an independent random variable with normal distribution of standard deviation 1 and centered in 0. The system of equations (1-4) was solved numerically for each parameter with

noise leaving the rest constant using a noise intensity $\beta = 0.01(1\%)$. The average deviation (S_{ε_i}) between the noisy solutions with respect to the no-noise solution is also calculated. The average deviations obtained for a patient with growing glioma (initial conditions 1) can be seen in Table 2, for a patient with advanced stage glioma (initial conditions 2) in Table 3 and for a patient without glioma (initial conditions 3) in Table 4, where the order of influence of the parameters is observed.

Evolution of glioma (T)	**Glucose levels in the brain** ($\sigma_{\mathbf{brain}}$)	**Immune system activity in the brain** (I)	**Serum glucose levels** ($\sigma_{\mathbf{serum}}$)
$S\alpha_T = 4.72327\times10^{-4}$	$S\alpha_\sigma = 1.68637\times10^{-5}$	$S\alpha_s = 2.86956\times10^{-5}$	$S\alpha_\sigma = 1.64504\times10^{-5}$
$S\nu = 3.5827\times10^{-4}$	$Sd_{T\sigma} = 1.35894\times10^{-5}$	$S\nu = 2.79093\times10^{-5}$	$Sd_{T\sigma} = 1.31053\times10^{-5}$
$Sd_{T\sigma} = 3.04307\times10^{-4}$	$S\alpha_s = 9.53708\times10^{-6}$	$Sd_{TT} = 2.05259\times10^{-5}$	$S\alpha_s = 9.28816\times10^{-6}$
$S\alpha_s = 1.81806\times10^{-4}$	$S\nu = 9.21831\times10^{-6}$	$Sd_{T\sigma} = 8.01755\times10^{-6}$	$S\nu = 8.96285\times10^{-6}$
$Sd_T = 6.50825\times10^{-5}$	$S\sigma_0 = 7.2453\times10^{-7}$	$S\alpha_T = 1.00573\times10^{-6}$	$S\sigma_0 = 7.87681\times10^{-7}$
$SK_T = 5.85242\times10^{-5}$	$S\sigma_{min} = 5.7075\times10^{-7}$	$Sd_I = 6.74882\times10^{-7}$	$S\sigma_{min} = 6.20798\times10^{-7}$
$Sd_{TT} = 5.66263\times10^{-5}$	$S\alpha_T = 2.42869\times10^{-7}$	$S\sigma_0 = 5.32647\times10^{-7}$	$S\alpha_T = 2.42792\times10^{-7}$
$Sd_{TI} = 5.24222\times10^{-5}$	$Sd_{\sigma1} = 1.84758\times10^{-7}$	$S\alpha_\sigma = 4.81307\times10^{-7}$	$Sd_{\sigma1} = 1.79384\times10^{-7}$
$S\sigma_0 = 1.62426\times10^{-5}$	$Sd_{\sigma2} = 1.16689\times10^{-7}$	$S\sigma_{min} = 4.20722\times10^{-7}$	$Sd_{\sigma2} = 1.25781\times10^{-7}$
$Sd_{\sigma1} = 5.94224\times10^{-6}$	$Sd_{TT} = 3.84346\times10^{-8}$	$Sd_{\sigma1} = 1.60575\times10^{-7}$	$Sd_{TT} = 3.84224\times10^{-8}$
$Sd_{\sigma2} = 3.64688\times10^{-6}$	$Sd_T = 2.53313\times10^{-8}$	$Sd_{TI} = 1.165\times10^{-7}$	$Sd_T = 2.53231\times10^{-8}$
$Sd_I = 2.6147\times10^{-6}$	$Sd_{TI} = 2.36531\times10^{-8}$	$Sd_{\sigma2} = 1.01543\times10^{-7}$	$Sd_{TI} = 2.36457\times10^{-8}$
$S\sigma_{min} = 2.41258\times10^{-6}$	$SK_T = 1.96256\times10^{-8}$	$Sd_T = 8.34023\times10^{-8}$	$SK_T = 1.96192\times10^{-8}$
$S\alpha_\sigma = 8.8973\times10^{-7}$	$Sd_I = 2.08661\times10^{-9}$	$SK_T = 3.37889\times10^{-8}$	$Sd_I = 2.08596\times10^{-9}$
$S\alpha_{TI} = 3.32817\times10^{-8}$	$S\alpha_{TI} = 2.3631\times10^{-11}$	$S\alpha_{TI} = 1.08784\times10^{-8}$	$S\alpha_{TI} = 2.36235\times10^{-11}$

Table 2: Average deviation values (ordered from highest to lowest) obtained for the initial conditions 1 using a noise intensity equal to 1% ($\beta = 0.01$) with respect to each parameter.

The parameters influenced very similarly for a patient with an advanced stage glioma (initial conditions 2) Table 3 and for a patient with a growing glioma (initial conditions 1). In the case of a patient without glioma (initial conditions 3) the parameters related to the tumor have no influence and affect in a different way with respect to the other initial conditions (see Table 4). In accordance with Table 2 one of the parameters that most influences a patient with growing glioma (initial conditions 1) is the baseline immune system cell production rate (ν), the result of which is shown in the Figure 1. It is of interest to observe that the glucose levels and the activity of the immune system diminish as the glioma grows. For the case of a patient with advanced stage glioma (initial conditions 2), Table 3 shows that the glucose consumption rate by glioma ($d_{T\sigma}$) is an influential parameter, the results for noise included in this parameter are shown in Figure 2. In this Figure it can be

 ISBN: 978-84-617-8694-7

seen that the glucose levels and the activity of the immune system are quite decayed because the glioma is already big. Finally in Table 4 it can be observed that the immune system cell recruitment rate (α_s) is an important parameter for the case of a patient without glioma (initial conditions 3) whose results are shown in the Figure 3. It is seen that glucose levels and activity of the immune system remained constant.

Evolution of glioma (T)	**Glucose levels in the brain** ($\sigma_{\mathbf{brain}}$)	**Immune system activity in the brain** (I)	**Serum glucose levels** ($\sigma_{\mathbf{serum}}$)
$S\alpha_T = 3.14535\times10^{-4}$	$S\alpha_\sigma = 1.72783\times10^{-5}$	$S\nu = 9.16772\times10^{-6}$	$S\alpha_\sigma = 1.66865\times10^{-5}$
$Sd_{T\sigma} = 2.67975\times10^{-4}$	$Sd_{T\sigma} = 1.40994\times10^{-5}$	$S\alpha_s = 9.12024\times10^{-6}$	$Sd_{T\sigma} = 1.34064\times10^{-6}$
$Sd_T = 9.71563\times10^{-5}$	$S\nu = 4.65053\times10^{-6}$	$Sd_{TT} = 7.90895\times10^{-6}$	$S\nu = 4.42455\times10^{-6}$
$S\nu = 9.00448\times10^{-5}$	$S\alpha_s = 4.61761\times10^{-6}$	$Sd_{T\sigma} = 4.15178\times10^{-6}$	$S\alpha_s = 4.39445\times10^{-6}$
$S\alpha_s = 5.85406\times10^{-5}$	$S\sigma_0 = 5.97225\times10^{-7}$	$S\alpha_\sigma = 3.96968\times10^{-7}$	$S\sigma_0 = 6.7143\times10^{-7}$
$SK_T = 3.8292\times10^{-5}$	$S\sigma_{min} = 4.72989\times10^{-7}$	$S\sigma_0 = 1.814\times10^{-7}$	$S\sigma_{min} = 5.31797\times10^{-7}$
$Sd_{TI} = 1.83604\times10^{-5}$	$Sd_{\sigma1} = 9.39898\times10^{-8}$	$S\sigma_{min} = 1.44595\times10^{-7}$	$Sd_{\sigma1} = 8.94567\times10^{-8}$
$Sd_{TT} = 1.74568\times10^{-5}$	$S\alpha_T = 7.78463\times10^{-8}$	$S\alpha_T = 8.33845\times10^{-8}$	$S\alpha_T = 7.78217\times10^{-8}$
$S\sigma_0 = 1.63468\times10^{-5}$	$Sd_{\sigma2} = 6.05967\times10^{-8}$	$Sd_I = 7.818\times10^{-8}$	$Sd_{\sigma2} = 6.81077\times10^{-8}$
$S\sigma_{min} = 6.27594\times10^{-6}$	$Sd_T = 2.40913\times10^{-8}$	$Sd_{\sigma1} = 2.80896\times10^{-8}$	$Sd_T = 2.40837\times10^{-8}$
$Sd_{\sigma1} = =2.61524\times10^{-6}$	$SK_T = 9.24783\times10^{-9}$	$Sd_T = 2.57339\times10^{-8}$	$SK_T = 9.24485\times10^{-9}$
$Sd_{\sigma2} = 1.3272\times10^{-6}$	$Sd_{TI} = 4.55963\times10^{-9}$	$Sd_{\sigma2} = 1.86688\times10^{-8}$	$Sd_{TI} = 4.5582\times10^{-9}$
$S\alpha_\sigma = 6.416\times10^{-7}$	$Sd_{TT} = 4.10278\times10^{-9}$	$SK_T = 9.75197\times10^{-9}$	$Sd_{TT} = 4.10143\times10^{-9}$
$Sd_I = 2.54664\times10^{-7}$	$Sd_I = 5.87371\times10^{-11}$	$Sd_{TI} = 4.89012\times10^{-9}$	$Sd_I = 5.87164\times10^{-11}$
$S\alpha_{TI} = 1.80112\times10^{-8}$	$S\alpha_{TI} = 4.32786\times10^{-12}$	$S\alpha_{TI} = 3.37733\times10^{-9}$	$S\alpha_{TI} = 4.32649\times10^{-12}$

Table 3: Average deviation values (ordered from highest to lowest) obtained for the initial conditions 2 using a noise intensity equal to 1% ($\beta = 0.01$) with respect to each parameter.

 ISBN: 978-84-617-8694-7

Evolution of glioma (T)	**Glucose levels in the brain**($\sigma_{\mathbf{brain}}$)	**Immune system activity in the brain** (I)	**Serum glucose levels** ($\sigma_{\mathbf{serum}}$)
$S\alpha_T = 0$	$S\alpha_s = 3.1726\times10^{-5}$	$S\nu = 5.96052\times10^{-4}$	$S\alpha_s = 3.12607\times10^{-5}$
$Sd_{T\sigma} = 0$	$S\nu = 2.83901\times10^{-5}$	$S\alpha_s = 5.19598\times10^{-4}$	$S\nu = 2.7994\times10^{-5}$
$Sd_T = 0$	$S\alpha_\sigma = 1.62421\times10^{-5}$	$Sd_I = 3.74108\times10^{-4}$	$S\alpha_\sigma = 1.60875\times10^{-5}$
$S\nu = 0$	$S\sigma_0 = 1.07689\times10^{-6}$	$S\sigma_0 = 9.72883\times10^{-6}$	$S\sigma_0 = 1.12092\times10^{-6}$
$S\alpha_s = 0$	$S\sigma_{min} = 8.83677\times10^{-7}$	$S\sigma_{min} = 8.8455\times10^{-6}$	$S\sigma_{min} = 9.19126\times10^{-7}$
$SK_T = 0$	$Sd_I = 7.73228\times10^{-7}$	$Sd_{\sigma1} = 6.29084\times10^{-6}$	$Sd_I = 7.72986\times10^{-7}$
$Sd_{TI} = 0$	$Sd_{\sigma1} = 5.74781\times10^{-7}$	$Sd_{\sigma2} = 4.56442\times10^{-6}$	$Sd_{\sigma1} = 5.66526\times10^{-7}$
$Sd_{TT} = 0$	$Sd_{\sigma2} = 3.71462\times10^{-7}$	$S\alpha_\sigma = 1.05216\times10^{-6}$	$Sd_{\sigma2} = 3.86275\times10^{-7}$
$S\sigma_0 = 0$	$S\alpha_T = 0$	$S\alpha_T = 0$	$S\alpha_T = 0$
$S\sigma_{min} = 0$	$SK_T = 0$	$SK_T = 0$	$SK_T = 0$
$Sd_{\sigma1} = 0$	$Sd_{TI} = 0$	$Sd_{TI} = 0$	$Sd_{TI} = 0$
$Sd_{\sigma2} = 0$	$S\alpha_{TI} = 0$	$S\alpha_{TI} = 0$	$S\alpha_{TI} = 0$
$S\alpha_\sigma = 0$	$Sd_T = 0$	$Sd_T = 0$	$Sd_T = 0$
$Sd_I = 0$	$Sd_{T\sigma} = 0$	$Sd_{T\sigma} = 0$	$Sd_{T\sigma} = 0$
$S\alpha_{TI} = 0$	$Sd_{TT} = 0$	$Sd_{TT} = 0$	$Sd_{TT} = 0$

Table 4: Average deviation values (ordered from highest to lowest) obtained for the initial conditions 3 using a noise intensity equal to 1% ($\beta = 0.01$) with respect to each parameter.

 ISBN: 978-84-617-8694-7

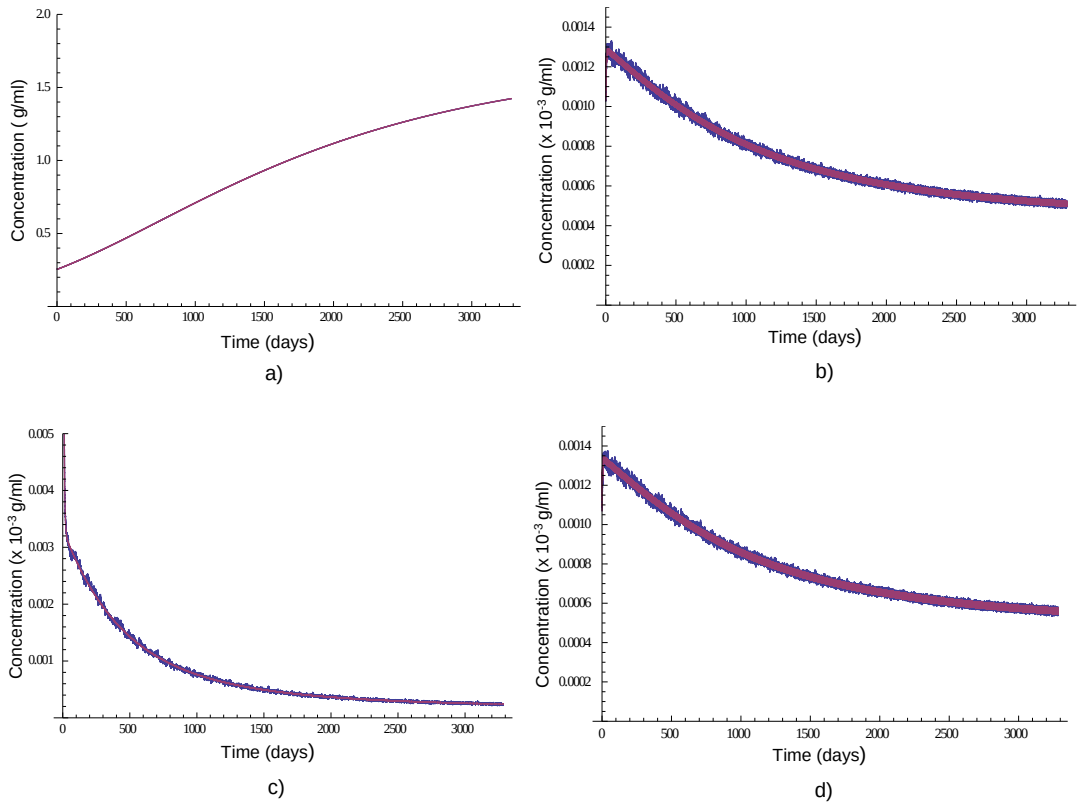


Figure 1: The system solution is shown for a patient with a small growing glioma (initial conditions 1) over a period of 3285 days (9 years) without noise (purple) and with noise (blue) of 1% ($\beta = 0.01$) in the baseline immune system cell production rate (ν). (a) Evolution of the glioma T, (b) glucose levels in the brain σ_{brain}, (c) immune system activity in the brain I and (d) serum glucose levels σ_{serum}.

 ISBN: 978-84-617-8694-7

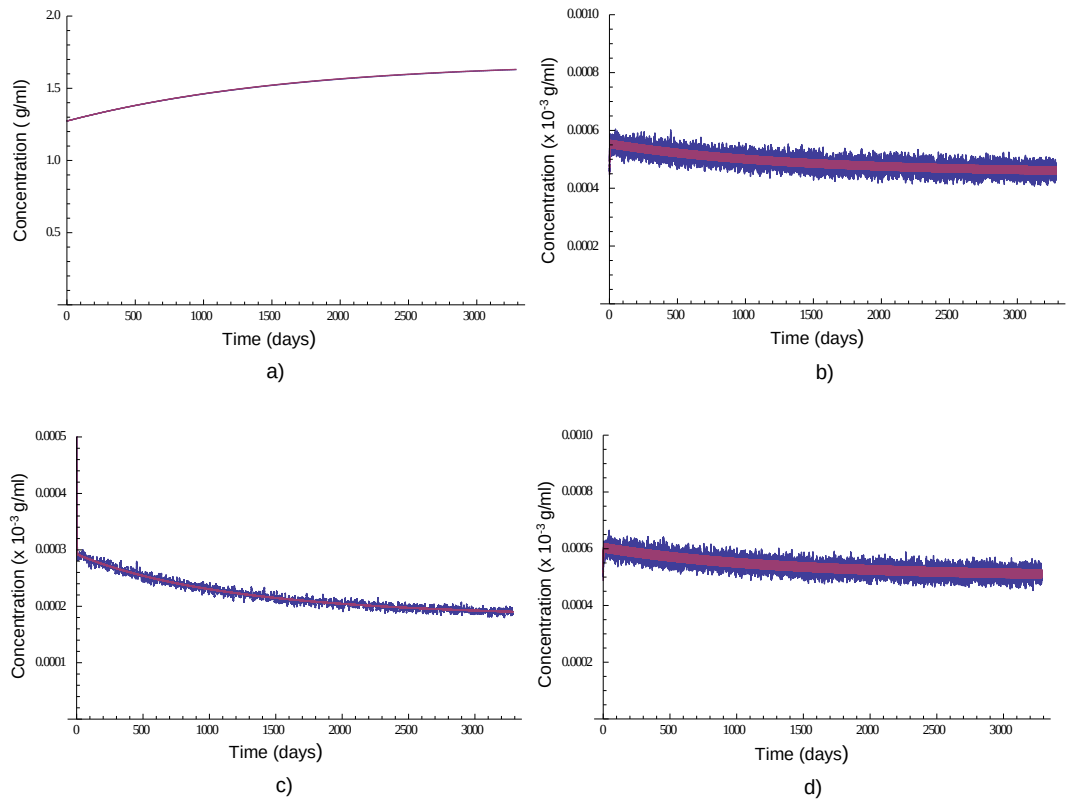


Figure 2: The system solution is shown for a patient with an advanced stage glioma (initial conditions 2) over a period of 3285 days (9 years) without noise (purple) and with noise (blue) of 1% ($\beta = 0.01$) in the glucose consumption rate by glioma ($d_{T\sigma}$). (a) Evolution of the glioma T, (b) glucose levels in the brain σ_{brain}, (c) immune system activity in the brain I and (d) serum glucose levels σ_{serum}.

 ISBN: 978-84-617-8694-7

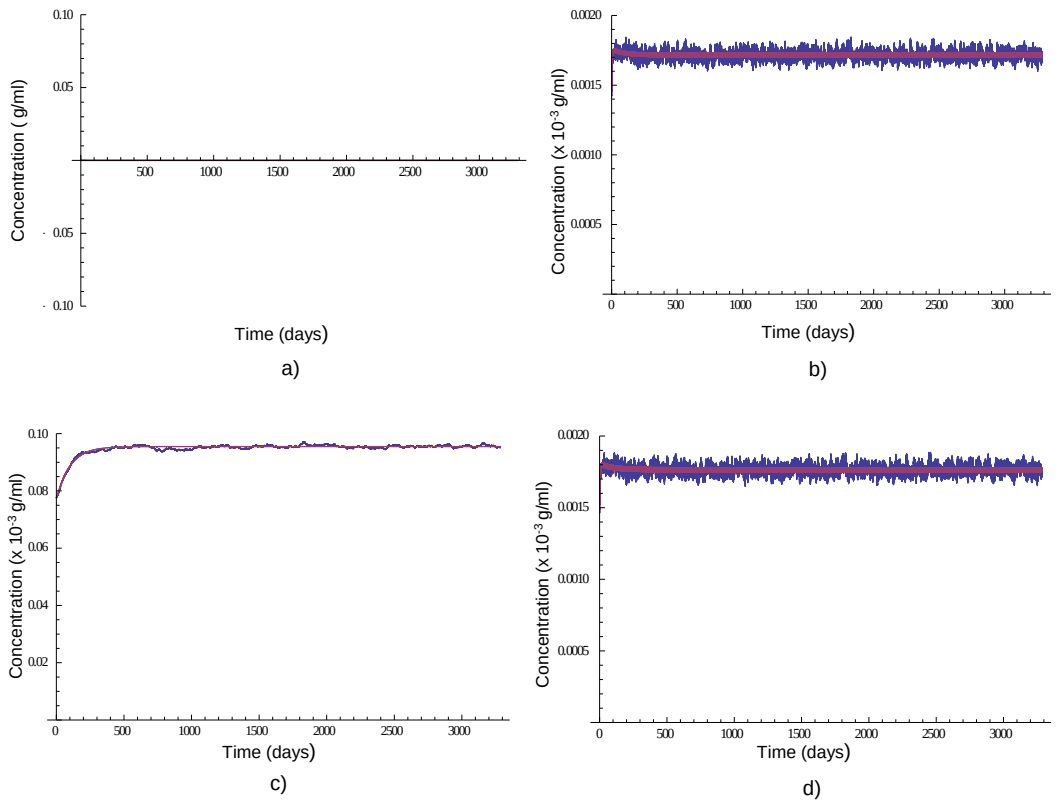


Figure 3: The system solution is shown for a patient without glioma (initial conditions 3) over a period of 3285 days (9 years) without noise (purple) and with noise (blue) of 1% ($\beta = 0.01$) in the immune system cell recruitment rate (α_s). (a) Evolution of the glioma T, (b) glucose levels in the brain σ_{brain}, (c) immune system activity in the brain I and (d) serum glucose levels σ_{serum}.

5 Conclusion

The model shows that when glioma exists (initial conditions 1 and 2), serum glucose levels fall below healthy levels, which represents a close dependence between tumor growth and available glucose, being consistent with the results obtained in the order of influence of the parameters, given that for the evolution of the glioma one of the parameters that more influences is the glucose consumption rate by glioma ($d_{T\sigma}$) and in general the model is

very sensitive to this parameter. The immune system plays a key role in tumor growth, because the immune system cell recruitment rate (α_s) and the baseline immune system cell production rate (ν) have a considerable influence. Also for the case of the patient without glioma the immune system is highly linked to serum glucose concentration, which suggests that it consumes a considered quantity of glucose for its activity.
The stability analysis shows that the model is stable when there is no glioma and is not stable when there is. In general, it can be concluded that this model serves to determine the level of influence of the parameters in this system. The results obtained can be used to search for blood-biomarkers for early detection of gliomas.

References

[1] N. Sanai, A. Alvarez-Buylla, M. S. Berger, *Neural stem cells and the origin of gliomas*, N. Engl. J. Med. **353** (2005) 811–822.

[2] S. S. Stylli, A. H. Kaye, L. MacGregor, M. Howes, P. Rajendra, *Photodynamic therapy of high grade glioma-long term survival*, J. Clin. Neurosci. **12** (2005) 389–398.

[3] S. Lonardi, A. Tosoni, A. A. Brandes, *Adjuvant chemotherapy in the treatment of high grade gliomas*, Cancer Treat. Rev. **31** (2005) 79–89.

[4] M. Sturrock, W. Hao, J. Schwartzbaum, G. Rempala, *A mathematical model of pre-diagnostic glioma growth*, J. Theor. Biol. **380** (2015) 299–308.

[5] W. Flavahan, Q. Wu, M. Hitomi, N. Rahim, Y. Kim, A. Sloan, R. Weil, I. Nakano, J. Nakano, B. Stringer, B. Day, L. Li, D. Justin, J. Rich, A. Hjelmeland, *Brain tumor initiating cells adapt to restricted nutrition through preferential glucose uptake*, Nat. Neurosci. **16** (2013) 1373–1382.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

DRBEM Solutions of the Direct and Inverse Formulations of Cauchy Problem for the Magnetohydrodynamic Duct Flow

Cemre Aydın[1] **and Münevver Tezer-Sezgin**[1]

[1] *Department of Mathematics, Middle East Technical University*

emails: acemre@metu.edu.tr, munt@metu.edu.tr

Abstract

In this paper, the MHD flow in a rectangular duct with a slipping wall portion is solved as a Cauchy problem when underspecified boundary conditions are present for the velocity on the slip wall. The aim is to regain the slipping velocity on this part of the duct walls. The dual reciprocity boundary element method (DRBEM) is used to discretize the Cauchy problem which results in an ill-conditioned system of linear algebraic equations, hence a regularization technique is necessary to solve the system. In this study, three regularization techniques, namely Tikhonov regularization, the direct-inverse iteration method, and the well-posed iteration method are used and compared in terms of velocity and induced magnetic field behaviors. Slip velocity behavior of the fluid is investigated for several values of the Hartmann number (Ha) with insulated and conducting vertical duct walls. It is found that all the three methods reconstruct the slip on the wall on which the direct solution has a slip assumption, especially when Ha increases.

Key words: DRBEM, Cauchy problem, MHD duct flow, regularization

1 Introduction

The electrodynamics of magnetic fields in the proximity of electrically conducting fluids, such as liquid metals and blood plasmas is concerned by magnetohydrodynamics (MHD). Conducting fluid movements under the external magnetic field are decribed with the combination of the Navier-Stokes equations with the Maxwell's equations through Ohm's law. The magnetohydrodynamic flow in ducts has many industrial and biological applications as MHD generators and MHD pipes, cooling of nuclear fusion apparatus and measuring the

blood flow pressure [1].

The distance from the fluid to the duct walls within the solid stage where the flow velocity vanishes is defined as the slip length. As current experimental data shows the slip in the MHD flow will likely occur in fusion reactors with liquid metal flows in contact with ceramics. The presence of the Dirichlet or mixed boundary conditions for the velocity on the complete duct walls results in direct problems. These conditions usually refer to no-slip velocity or the slip velocity on the whole duct walls, respectively. The MHD equations have exact solutions only for some special duct geometries used for no-slip velocity or slip velocity and insulated or perfectly conducting duct walls. For the general case of wall conditions, these equations are mostly solved by numerical techniques [2]. In some engineering applications one part of the boundary may allow both the velocity slip and conductivity change depending on the material it is made of. In this case, the boundary conditions are incomplete either in the form of underspecified or overspecified on different parts of the boundary. These are inverse problems and it is well-known that they are generally ill-posed [3]. Thus, a regularization method must be used.

The DRBEM transforms the differential equations defined in the region to integral equations defined on the boundary, approximating also the inhomogeneities of the equations using radial basis functions which are related to differential operator with particular solutions. This way, a system of discretized equations for the boundary nodes and at some selected interior points is solved [4]. There are quite a number of DRBEM solutions of MHD duct flow problems with no-slip velocity condition and various combinations of wall conductivities [5, 6, 7, 8].

In this paper, we present the DRBEM solution of both the direct and inverse MHD flow problems when one part of the duct walls contain the velocity slip. The slip length is not known and thus, both the velocity and its normal derivative are going to be determined which are underspecified boundary conditions. Since the other parts of the boundary have both of these values, they are overspecified which form the Cauchy MHD problem. The Tikhonov regularization method, the direct-inverse iteration method, and the well-posed iteration method are used to regularize the ill-posed discretized problem. The results are obtained for Hartmann number values $\leqslant 50$ and these three methods are compared in terms of the convergence to the direct problem solutions in the sense that direct problem is solved with an approximation added to the underspecified velocity wall conditions. The DRBEM has the advantage of discretizing only the boundary and providing both the velocity and its normal derivative values on the boundary. Thus, it gives the solution of Cauchy MHD flow problem at a small computational expense.

 ISBN: 978-84-617-8694-7

2 Mathematical Formulation of the Problem

The two-dimensional, steady and fully developed MHD flow is considered in a rectangular duct under an external magnetic field applied horizontally. The vertical walls have certain electrical conductivity and the left wall allows also the slip of the fluid. The non-dimensional governing equations in terms of the velocity $V(x,y)$ of the fluid and the induced magnetic field $B(x,y)$ are given as [1]

$$\begin{aligned} &\nabla^2 V + Ha\frac{\partial B}{\partial x} = -1 \\ &\qquad\qquad\qquad\qquad \text{in} \quad -1 \le x, y \le 1 \\ &\nabla^2 B + Ha\frac{\partial V}{\partial x} = 0 \end{aligned} \tag{1}$$

where $Ha = LB_0\sqrt{\sigma/\nu\rho}$ is the Hartmann number resulted from the nondimensionalization of the equations, and L, B_0, σ, ν and ρ are the characteristic length, the external magnetic field intensity, electrical conductivity, kinematic viscosity, and the density of the fluid, respectively.

The physical configuration of the duct walls results in the boundary conditions of the velocity and the induced magnetic field as ($V = 0$, $B = 0$, no-slip and insulated wall)

$$\begin{aligned} &V = 0, && B = 0 && \text{on} \quad y = \mp 1, \quad -1 < x < 1 \\ &V = 0, && B = k && \text{on} \quad x = 1, \quad -1 < y < 1 \\ &V + \alpha\frac{\partial V}{\partial n} = 0, && B = -k && \text{on} \quad x = -1, \quad -1 < y < 1 \end{aligned} \tag{2}$$

where α is the dimensionless slip length and $\mp k$ are the conductivity values of the vertical walls. By taking $U_1 = V + B$ and $U_2 = V - B$ equations (1) are decoupled first as

$$\begin{aligned} &\nabla^2 U_1 = -1 - Ha\frac{\partial U_1}{\partial x} \\ &\qquad\qquad\qquad\qquad \text{in} \quad -1 \le x, y \le 1 \\ &\nabla^2 U_2 = -1 + Ha\frac{\partial U_2}{\partial x} \end{aligned} \tag{3}$$

in which both of the equations are of diffusion-convection type. The aim of the study is to obtain the slip velocity on the left wall $x = -1, -1 \leqslant y \leqslant 1$. Meantime, there will be a conductivity change on this left slipping wall due to the newly computed velocity values. Thus, the MHD problem is reconstructed as a Cauchy problem first in terms of U_1 and U_2 as shown in Figure 1.

 ISBN: 978-84-617-8694-7

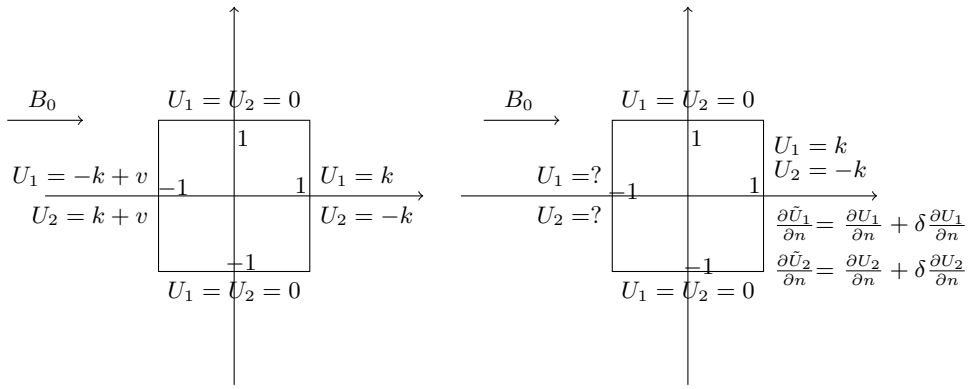


Figure 1: Boundary conditions for the direct (left) and the Cauchy (right) problems.

Here, v is the approximation added to U_1 and U_2 in the direct problem to differ from the no-slip condition $V = 0$ to a slip condition $V + \alpha \dfrac{\partial V}{\partial n}$ on the left wall where both V and $\dfrac{\partial V}{\partial n}$ are not known. Then, the solution of direct problem using DRBEM gives $\dfrac{\partial U_1}{\partial n}$, $\dfrac{\partial U_2}{\partial n}$ everywhere on the boundary. The normal derivative conditions with noisy data on the $x = 1$ wall are obtained as $\frac{\partial \tilde{U}_1}{\partial n} = \frac{\partial U_1}{\partial n} + \delta \frac{\partial U_1}{\partial n}$ and $\frac{\partial \tilde{U}_2}{\partial n} = \frac{\partial U_2}{\partial n} + \delta \frac{\partial U_2}{\partial n}$, where $\delta \frac{\partial U_1}{\partial n}$ and $\delta \frac{\partial U_2}{\partial n}$ are Gaussian random variables with mean zero and standard deviation $\sigma = max_{x=1} |\frac{\partial U_1}{\partial n}, \frac{\partial U_2}{\partial n}| \times (\dfrac{p_u}{100})$ and p_u is the percentage of additive noise for $\frac{\partial U_1}{\partial n}$ and $\frac{\partial U_2}{\partial n}$. Then, the Cauchy problem is solved by DRBEM to determine both U_1, U_2, $\dfrac{\partial U_1}{\partial n}$, $\dfrac{\partial U_2}{\partial n}$ on $x = -1$. Finally, V, B, $\dfrac{\partial V}{\partial n}$, $\dfrac{\partial B}{\partial n}$ are obtained through the relationships $U_1 = V + B$, $U_2 = V - B$.

3 The DRBEM Application

The DRBEM is applied to the decoupled equations (3) of the MHD problem by using the fundamental solution of the Laplace equation which is $u^* = ln(\frac{1}{r})/2\pi$, [4]. Thus, all the terms other than Laplacian are considered as inhomogeneity. That is, by weighting the equations (3) by u^* and applying Green's second identity, we obtain the following equations

$$c_i U_{1i} + \int_\Gamma q^* U_1 d\Gamma - \int_\Gamma u^* \frac{\partial U_1}{\partial n} d\Gamma = - \int_\Omega (-1 - Ha \frac{\partial U_1}{\partial x}) u^* d\Omega \tag{4}$$

$$c_i U_{2i} + \int_\Gamma q^* U_2 d\Gamma - \int_\Gamma u^* \frac{\partial U_2}{\partial n} d\Gamma = - \int_\Omega (-1 + Ha \frac{\partial U_2}{\partial x}) u^* d\Omega \tag{5}$$

 ISBN: 978-84-617-8694-7

where $q^* = \dfrac{\partial u^*}{\partial n}$, Γ is the boundary $x = \mp 1, y = \mp 1$, and the index i denotes the source point. The constant c_i is 1/2 and 1 when the source point is on the boundary and in the interior of the domain, respectively.

The right hand side domain integrals are treated as inhomogeneity and they can be approximated by radial basis functions $f(r) = 1 + r$ which are connected to the particular solutions $\hat{u}_j$'s with the equation $\nabla^2 \hat{u}_j = f_j$. The approximations of the domain integrals are given by $\sum_{j=1}^{N+L} \alpha_j f_j$ and $\sum_{j=1}^{N+L} \beta_j f_j$ for the equations (4) and (5), respectively, where α_j's and β_j's are undetermined coefficients, $f_{ij} = 1 + r_{ij}$, and r_{ij} is the distance between the nodes i and j, N and L denote the number of boundary and interior nodes, respectively, when the bpundary Γ is dicretized using N constant boundary elements.

Then, the right hand sides of the equations (4)-(5) are rewritten as

$$c_i U_{1i} + \int_\Gamma q^* U_1 d\Gamma - \int_\Gamma u^* \frac{\partial U_1}{\partial n} d\Gamma = \sum_{j=1}^{N+L} \alpha_j (c_i \hat{u}_{ij} + \int_\Gamma q^* \hat{u}_j d\Gamma - \int_\Gamma u^* \frac{\partial \hat{u}_j}{\partial n} d\Gamma) \tag{6}$$

$$c_i U_{2i} + \int_\Gamma q^* U_1 d\Gamma - \int_\Gamma u^* \frac{\partial U_2}{\partial n} d\Gamma = \sum_{j=1}^{N+L} \beta_j (c_i \hat{u}_{ij} + \int_\Gamma q^* \hat{u}_j d\Gamma - \int_\Gamma u^* \frac{\partial \hat{u}_j}{\partial n} d\Gamma) \tag{7}$$

by applying the BEM also to the inhomogeneities connected to the same Laplace operator. The discretization of the boundary results in system of matrix vector equations

$$HU_1 - G\frac{\partial U_1}{\partial n} = (H\hat{U} - G\hat{Q})F^{-1}\{-1 - Ha\frac{\partial U_1}{\partial x}\} \tag{8}$$

$$HU_2 - G\frac{\partial U_2}{\partial n} = (H\hat{U} - G\hat{Q})F^{-1}\{-1 + Ha\frac{\partial U_2}{\partial x}\}. \tag{9}$$

The matrices $\hat{U}$, $\hat{Q}$ and F are constructed by taking each of the vectors $\hat{u}_j$, $\hat{q}_j$ and f_{ij} as columns, respectively. The components of the H and G matrices are given as

$$H_{ij} = c_i \delta_{ij} + \frac{1}{2\pi} \int_{\Gamma_j} \frac{\partial}{\partial n}(ln(\frac{1}{r}))d\Gamma_j, \quad H_{ii} = -\sum_{j=1, j\neq i}^{N} H_{ij}$$

$$G_{ij} = \frac{1}{2\pi} \int_{\Gamma_j} ln(\frac{1}{r})d\Gamma_j, \quad G_{ii} = \frac{l}{2\pi}(ln(\frac{2}{l}) + 1)$$

 ISBN: 978-84-617-8694-7

where l is the length of the elements and δ_{ij} is the Kronecker delta function.

The space derivatives for U_1 and U_2 are computed by using the coordinate matrix as

$$\frac{\partial U_1}{\partial x} = \frac{\partial F}{\partial x} F^{-1} U_1 \quad \text{and} \quad \frac{\partial U_2}{\partial x} = \frac{\partial F}{\partial x} F^{-1} U_2.$$

The solution of the system (8)-(9) gives the unknown vectors U_1, U_2, $\frac{\partial U_1}{\partial n}$ and $\frac{\partial U_2}{\partial n}$ which are $(N+L)\times 1$ vectors everywhere on the boundary.

4 Tikhonov Regularization

The unknown vectors U_1, U_2, $\frac{\partial U_1}{\partial n}$ and $\frac{\partial U_2}{\partial n}$ which are $(N+L)\times 1$ vectors everywhere on the boundary are obtained from the system of equations (8)-(9). Since the aim is to find the slip velocity on the left wall ($x=-1$), first the direct problem is solved by adding approximations to U_1 and U_2 on this wall. Then, with the obtained $\frac{\partial U_1}{\partial n}$ and $\frac{\partial U_2}{\partial n}$ values on $x=1$, the Cauchy problem will be solved as an inverse formulation using Tikhonov regularization with the overspecified boundary conditions $(U_1, U_2, \frac{\partial U_1}{\partial n}, \frac{\partial U_2}{\partial n})$ on $x=1$, and underspecified boundary conditions (no boundary condition) on $x=-1$. The resulting ill-conditioned system of equations in the form $Ax=b$ obtained from the DRBEM discretized system for the Cauchy problem configured in Figure 1 is solved by using Tikhonov regularization as defined

$$(A^T A + \lambda^2 I)x = A^T b.$$

λ is the regularization parameter obtained from the L-curve method and reduces the system to least-squares equations for $\lambda = 0$. In the case of λ being large, the residual norm $||Ax-b||_2$ will be large. On the other hand, when λ is small, then the solution will be dominated by the contributions from the data errors and so, $||x||_2$ will be very large. Hence, the choice of the regularization parameter is very crucial. We choose the regularization parameter λ by using the L-curve method. The curve $(||Ax_\lambda - b||_2, ||x_\lambda||_2)$, x_λ is the regularized solution, which is obtained from plotting the norms of the residual and the solution provides controlling these two quantities. The corner of this curve gives the optimal regularization parameter λ, [9].

5 The Direct-Inverse Iterations for the Cauchy Problem

Instead of using one-step Tikhonov regularization, an iterative method including the direct and inverse solutions for the governing equations (3) can be developed. In the inverse

 ISBN: 978-84-617-8694-7

formulations, the Tikhonov regularization is used. In this method, while the direct problem is solved with the Dirichlet type boundary conditions for U_1 and U_2 on each part of the boundary, but containing approximations on $x = -1$, the inverse problem is solved with overspecified boundary conditions $U_1, U_2, \frac{\partial U_1}{\partial n}, \frac{\partial U_2}{\partial n}$ on $x = 1$. The iterations follow as;

1. $U_1 = -k+v$, $U_2 = k+v$ are taken with approximation v to achieve a slip on $x = -1$.

2. Solve the discretized equations (8)-(9) for the direct problem configured in Figure 1.

3. Use $\frac{\partial U_1}{\partial n}, \frac{\partial U_2}{\partial n}$ obtained from step 2 in addition to the Dirichlet boundary conditions for U_1, U_2 on $x = 1$ for solving the inverse problem using Tikhonov regularization with overspecified and underspecified boundary conditions for $x = 1$ and $x = -1$, respectively.

4. Update the values U_1, U_2 with the obtained ones from step 3.

5. Repeat steps 2-4 until the velocity $V = \frac{U_1 + U_2}{2}$ on $x = -1$ converges to the velocity of the direct solution containing the approximation.

6 The Well-posed Iterations for the Cauchy Problem

This method deduces the ill-posed problem to a successive well-posed problems by alternating the given data on the overspecified part of the boundary. This iterative algorithm advanced by Kozlov et. al. [10] for Cauchy problems includes the following steps:

1. $U_1 = -k + v$, $U_2 = k + v$ are taken with an approximation v to achieve slip on $x = -1$ wall.

2. Solve the discretized equations (8)-(9) for the direct problem with the Dirichlet type boundary conditions for U_1 and U_2 given as in Figure 1.

3. Use the normal derivative conditions for U_1 and U_2 on $x = -1$ and $x = 1$ obtained from step 2 for solving the direct problem.

4. Update the values U_1 and U_2 with the values obtained from step 3 on $x = -1$.

5. Repeat steps 2-4 until the velocity $V = \frac{U_1 + U_2}{2}$ on $x = -1$ converges to the perturbed velocity $V = v$.

 ISBN: 978-84-617-8694-7

7 Numerical Results

The MHD duct flow equations (3) are solved numerically in terms of the velocity V of the fluid and the induced magnetic field B with the boundary conditions imposed for the direct and inverse problems as given in Figure 1. The aim is to regain the slipping velocity and compute slip length on the left wall. In the application of DRBEM, $N = 100, 160, 240$ boundary elements and $L = 625, 1600, 3600$ interior nodes are taken for $Ha = 1, 10, 50$, respectively. Both the insulated ($k = 0$) and conducting ($k = 1$) vertical walls are considered.

Figures 2, 3 show the velocity and induced magnetic field profiles for the direct solution with a slipping left wall where the slip length is taken $\alpha = 0.2$, [11], respectively for $k = 0$ (insulated) and $k = \mp 1$ (conducting) vertical walls. This means that the velocity V is perturbed with v corresponding to slip length $\alpha = 0.2$. In Figures 4, 5, we present velocity behaviors obtained from the Tikhonov regularization, the direct-inverse iterations, and the well-posed iterations employed to solve the Cauchy MHD flow problem, respectively for $k = 0$ and $k = 1$. They demonstrate that the slip velocity is reconstructed and converged to the direct solution which is perturbed with a slip length $\alpha = 0.2$. When $Ha = 1$ is taken, all the three methods converge to the direct solution. On the other hand, when Hartmann number increases to $Ha = 10$, the well-posed iterations seem to converge to the direct solution very well while the others show much sharper slip on the left wall. However, all the three methods capture the same velocity behavior of the direct solution. When the Hartmann number is further increased to Ha=50, the velocity slip is detected close to the corners of the left wall and there is no slip on the middle portion of the right wall. In all the methods, an increase in Hartmann number causes boundary layers (Hartmann layers of order $1/Ha$ on the walls perpendicular to the applied magnetic field and side layers of order $1/\sqrt{Ha}$ on the parallel walls [1]). However, Hartmann layers are weakened on the left wall due to the slip effect of the velocity as Ha increases. As Ha increases further to Ha=50, the core region enlarges, and fluid becomes stagnant and the flow is flattened. The slip velocity is regained on the left wall by using DRBEM with all the three methods for the Cauchy MHD duct flow problem.

In Figure 6, we depict also induced magnetic field behaviors obtained from the three methods. The well-posed iterations give exactly the same induced magnetic field of the direct solution since there is no change involved in B. On the other hand, the Tikhonov and direct-inverse methods result in conductivity changes on $x = -1$ due to the perturbations imposed on the direct formulation. However, all the three methods show the same behavior of B especially when Ha increases.

 ISBN: 978-84-617-8694-7

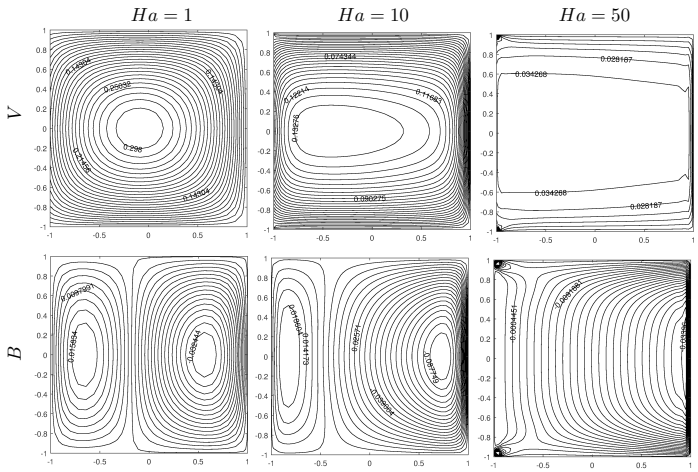


Figure 2: Velocity and induce magnetic field profiles from direct solution with slip walls, $k = 0$, $\alpha = 0.2$

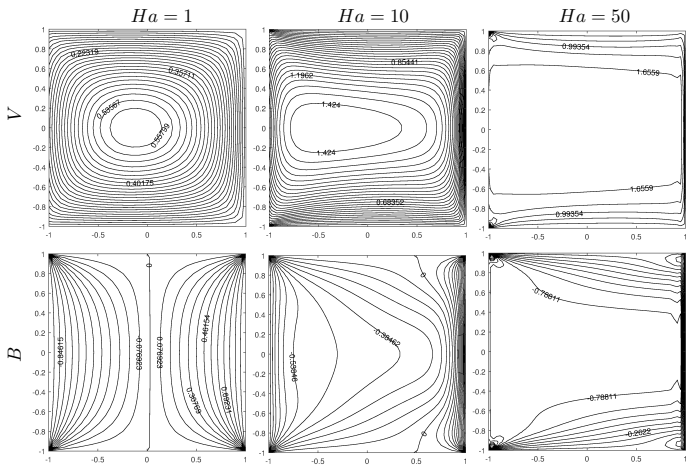


Figure 3: Velocity and induce magnetic field profiles from direct solution with slip walls, $k = 1$, $\alpha = 0.2$

 ISBN: 978-84-617-8694-7

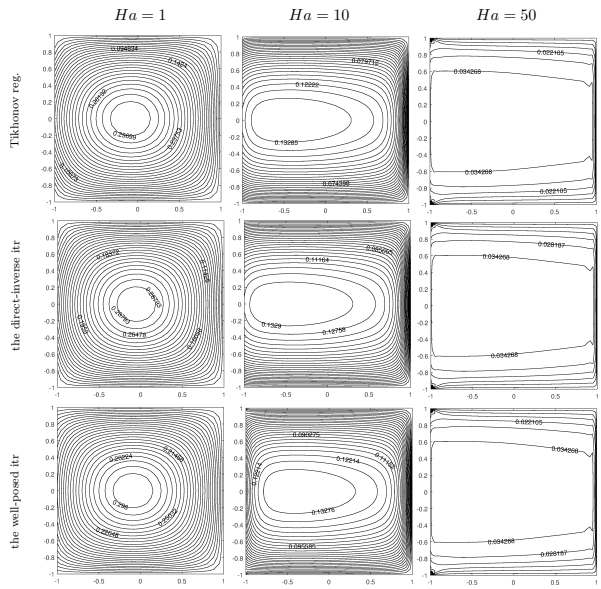


Figure 4: Velocity profiles from inverse solution, $k = 0$

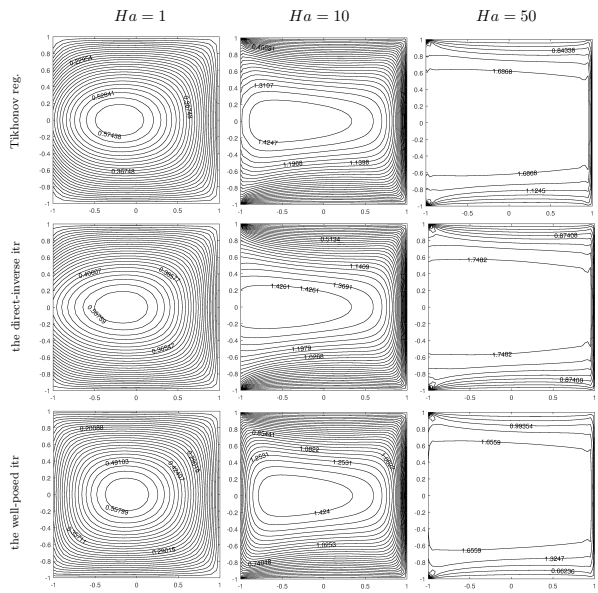


Figure 5: Velocity profiles from inverse solution, $k = 1$

 ISBN: 978-84-617-8694-7

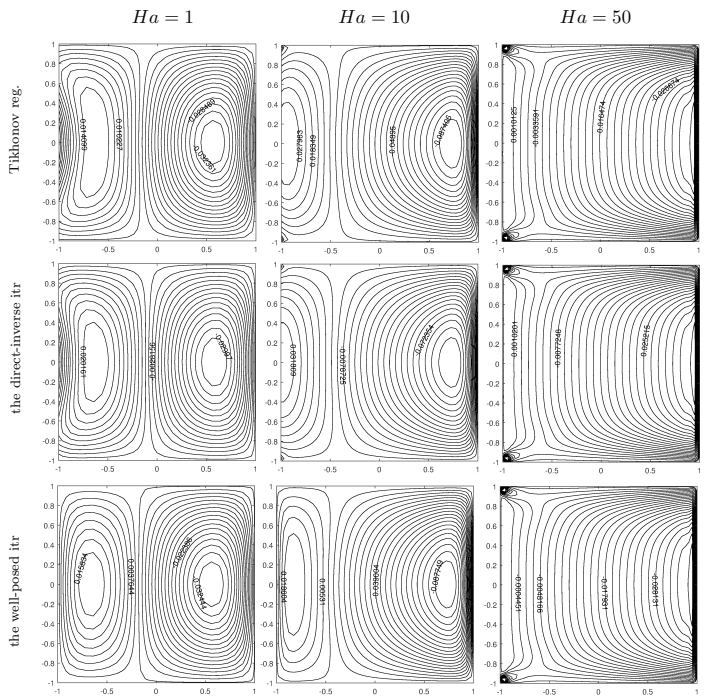


Figure 6: Induced magnetic field profiles from inverse solution, $k = 0$

8 Conclusion

The MHD rectangular duct flow is formulated as a direct and a Cauchy problem in terms of the velocity slip on the left wall and is solved by using the DRBEM. The Cauchy problem is solved with the normal derivative conditions obtained from the direct solution to determine the underspecified velocity information on the left wall by using Tikhonov regularization, the direct-inverse iterations, and the well-posed iterations. The Cauchy problem reconstructs the slip velocity on the left wall in accordance with the solution of perturbed direct problem for Ha=1, 10, 50 in all the methods mentioned. The well-posed iterations show very well convergence to the solution of direct problem on the slipping wall while Tikhonov regularization and direct-inverse iterations capture both the slip and conductivity change on this wall. When Hartmann number increases, all the three methods reconstruct the slip velocity on the left wall. The DRBEM is the most suitable numerical method for the Cauchy MHD duct problem since it provides both the velocity and its normal derivative values on the underspecified wall extracting the slip length α between them.

 ISBN: 978-84-617-8694-7

References

[1] L. Dragoş, *Magnetofluid Dynamics*, Abacus Press, England, 1975.

[2] E. Ligere, I. Dzenite, A. Matvejevs, *Analytical solution of a problem on MHD flow in a rectangular duct*, Recent Advances in Mathematics

[3] L. Marin, et al., *Dual reciprocity boundary element method solution of the Cauchy problem for Helmholtz-type equations with variable coefficients*, Jour. of Sound and Vibration **297** (2006) 89-105.

[4] P.W. Partridge, C.A Brebbia, L.C Wrobel, *The Dual Reciprocity Boundary Element Method*, Computational Mechanics Publications, Southampton Boston, 1992.

[5] A. I. Nesliturk and M. Tezer-Sezgin, *Finite Element method solution of electrically driven Magnetohydrodynamic flow*, Jour. of Comput. and Appl. Math. **192** (2006) 339-352.

[6] M. Tezer-Sezgin and S. Han Aydin, *Solution of MHD Flow Problems using the Boundary Element Method*, Engrg. Analy. with Bound. Elem. **30** (2006) 411-418.

[7] C. Bozkaya and M. Tezer-Sezgin, *Fundamental Solution for Coupled Magnetohydrodynamic Flow Equations*, Jour. of Comput. and Appl. Math. **203** (2007) 125-144.

[8] S. Han Aydin, M. Tezer-Sezgin, *DRBEM Solution of MHD Pipe Flow in a Conducting Medium*, Jour. of Comput. and Appl. Math. **259** (2014) 720-729.

[9] P.C. Hansen, *The L-curve and its use in the Numerical Treatment of Inverse Problems*, Computational Inverse Problems in Electrocardiology, WIT Press, 2001.

[10] V.A. Kozlov, V.G. Maz'ya, A.F. Fomin, *An iterative method for solving the Cauchy problem for elliptic equations*, Comput. Maths. Math. Phys **31** (1991) 45-52.

[11] P. Şenel and M.Tezer-Sezgin, *MHD Duct Flow with Slipping Velocity on the Hartmann and Side Walls*, EMI Int. Conf. presentation (19-22 2017).

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

Fluidic actuator performance variation via internal dimensions modifications

Masoud Baghaei[1], Josep M Bergada[2] and David Del Campo[1]

[1] *Physics Department, UPC-ESEIAAT Colon 7-11, Terrassa, Spain*

[2] *Fluid Mechanics Department, UPC-ESEIAAT Colon 7-11, Terrassa, Spain*

emails: masoudbaghaie@gmail.com, josep.m.bergada@upc.edu, david.del.campo@upc.edu

Abstract

When aimed to modify the downstream vortex shedding of a given bluff body, whether any road vehicle or wing profile, the use of Active Flow Control (AFC) appears to be an efficient technology. Among the different (AFC) methodologies the use of periodic forcing is ment to have better efficiency since it requires less energy to activate the shear layer, the reason behind this efficiency lies on the fact that periodic forcing interacts with the shear layer natural instabilites. In the present paper, one of the devices widely emloyed to generate pulsating flow, is carefully studied via 3D-CFD and using OpenFOAM. Initially the base flow is being determined and compared with previous experimental results, in a second step several internal dimensions of the fluidic actuator are being modified to characterize the output frequency and amplitude variations, among the conclusions obtained it is found that a given fluidic actuator is capable of generating several output frequencies and amplitudes when modifiying some internal dimensions while maintaining a constant incoming flow Reynolds number.

Key words: Fluidic actuators, Active flow control, Computational fluid dynamics.

1 Introduction

One of the newest technology to modify the lift and drag of a given body is via injecting or sucking flow to or from the boundary layer nearby the separation point. The technique which seems to be more effective, mostly because it involves the use of the smallest amount of energy, is the use of periodic forcing, the main advantage of using periodic forcing is that

the injected flow interacts with the shear layer natural instabilities, and threrefore deeply activates the flow. To generate pulsating flow, two main types of fluidic oscillators are being used, the Zero Net Mass Flow (ZNMF)fluidic oscillators and the Fluidic actuators (FA). The former consist of a membrane located inside of an open chamber, the membrane moves back and forward and so a pulsating flow is being generated, with positive and negative velocities alternating every half cycle, as a result, the net mass flow at every cycle is null. The later generates a sinusoidal outgoing flow and has the advantage of having no moving parts, this particular advantage is very handy when designing long lasting and reliable systems.

Original (FA) designs goes back to the 60s and 70s, left nearly unchanged for over 45 years. Their possible output frequency ranges from several Hz to KHz and the flow rate is usually of a few dm^3/min. Among their applications in flow control, it is worth to mention their use in combustion control [1, 2, 3], mixing enhancement [4], flow separation in aerofoils [5], boundary layer control on hump diffusers used in turbomachinery [6], flow separation control on stator vanes of compressors [7], drag reduction on trucks [8] and cavity noise reduction [9]. It appears that fluidic actuators have the potential of being much widely used in the near future, and according to the authors there is the need of better understanding their behaviour in order to further improve their performance. Regarding the fluidic actuators design two main groups exist, the one based on Coanda effect [10], and the one based on a jet mixing chamber, also called vortex oscillators [11]. The former group had an early application as pressure, temperature and flow measuring devices [12, 13, 14], the latter group has recently been applied as a flow control device [15]. To push forward (FA) boundaries several new designs have been recently created. Uzol and Camci [16], studied experimentally and via (CFD) a fluidic oscillator based on two elliptical cross-sections placed transversally and an after-body located in front of them. Such configuration was in fact proposed by Bauers patent [17, 18]. The device operates at frequencies of around 30 Hz and under laminar flow. The relation frequency versus Reynolds number was found to be perfectly linear. Huang and Chang [19] performed a deep experimental study on a V-shaped fluidic oscillator. Playing with the dimensions and the internal oscillator circular cavity, they defined the regimes under which oscillation was generated and they proved that frequencies from few Hz to several KHz could be obtained by modifying oscillator parameters. Additionally, an analysis of the streamline patterns behind the oscillator was also presented. Khelfaoui et al [20], presented an experimental and numerical analysis of non-symmetrical mini and micro oscillators. They found a linear relationship between the actuator frequency and the feedback channel volume, and noticed that above a certain input pressure choked flow appeared. Gebhard et al [21] studied a micro-oscillator operated with water, finding a linear relationship between the output frequency and the input volumetric flow. Raman and Raghu [9] evaluated the decrease of a cavity tone by using fluidic oscillators. The main acoustic frequency was reduced by over 10 dB, concluding that fluidic

 ISBN: 978-84-617-8694-7

excitation is a candidate in noise control applications. A numerical simulation of a two dimensional fluidic oscillator by using Navier-Stokes equations in laminar and incompressible flow, was performed by Nakayama et al [22]. They were able to visualize the periodical flow movement and measured the temporal axial and tangential fluid velocities, oscillation frequency being of 40Hz. Gregory and Raghu [23], created a fluidic oscillator based on Coanda effect but driven by piezoelectric devices. One of the main interesting performances of such device is that the oscillating frequency can be decoupled from the input flow and pressure differential. Frequency just depends on input electrical signal, being the oscillator able to work at a range of velocities which goes up to sonic conditions.
The present paper will introduce a numerical evaluation of a fluidic actuator previously studied in [24, 25, 26, 27]. In these previous studies, and extensive CFD model including the analysis of several turbulent models in order to find out which one was the most appropriate, was undertaken. Besides, they performed an experimental study obtaining a good agreement between experimental and CFD results. In the present paper, experimental results obtained in [24, 25], will be compared with the new CFD calculations. A discussion regarding how different fluidic actuators internal parts and dimensions may affect its performance will be carried on. The authors main aim is to give to the reader some hints to be able to modify a given oscillator to fulfil a particular application.

2 Numerical problem definition and boundary conditions

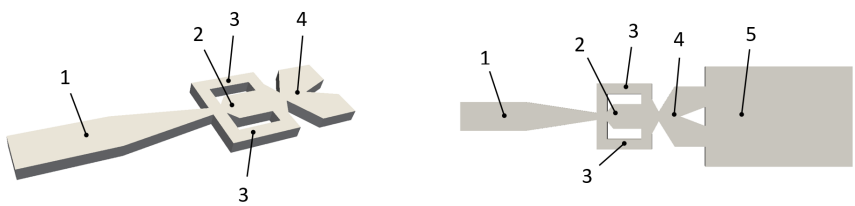
The three dimensional fluidic actuator considered in the present paper is depicted in figure 1, its thickness was of 3.25mm. The incoming flow enters the actuator mixing chamber (2) through the flattered pipe located on the left hand side of the figure (1), on both sides of the mixing chamber there are the feedback channels (3), their function is to allow transporting fluid from the downstream mixing chamber site to the upstream one and vice-versa, the fluid leaves the actuator alternatively through one of the two exit surfaces located on both sides of the external chamber (4). Notice that a second fluidic actuator with a buffer zone (5) is also presented, the idea behind this second configuration is evaluating the efect of the outlet boundary conditions onto the pulsating flow. The mesh employed for the present simulation had a total of 2242000 cells, the grid used was structured and care was taken to obtain a very small y+ in all directions, in fact the maximum y+ respectively obtained in x, y, and z directions was, 1.8, 4.7 and 1.2 for a Reynolds number of 16034. Boundary conditions employed were, fluid velocity at the entrance and absolute pressure $1.01978 * 10^5 Pa$ at the output, Dirichlet boundary conditions were set to all walls. A range of different input velocities from 0.758 to 1.23 m/s were studied, its minimum and maximum Reynolds number associated was 8711 and 16034. The fluid employed was water and it was considered as incompressible. Fluid dynamic viscosity was chosen as 0.001003 $Kg/(m * s)$ and fluid density was 998.2 Kg/m^3. The characteristic length was chosen to be the inlet

 ISBN: 978-84-617-8694-7

width, which value was $2.55 * 10^{-3}m$. The turbulence model used was the SpalartAlmaras DDES, which is a hybrid LES model. OpenFOAM version 3.0, was employed for all 3D simulations, finite volumes approach was employed. Inlet turbulence intensity was set to 0.05% in all cases; PISO was used as a solution method, being the time step of $10^{-6}s$, spatial discretization was set to second order. The CFD model designed had a probe covering one of the two actuator exits, at this section the frequency and amplitude associated to the temporal mass flow were measured. The frequencies obtained were compared with the ones experimentally obtained in [24, 25], table 1 compares both results. Notice that the difference is minimum giving confidence to the CFD simulations undertaken.

Table 1: Comparison experimental and CFD results.

Reynolds number	8711	11152	13593	16034
Frequency [Hz], ref. [24, 25]	12.9	15.5	18.7	21.8
Frequency [Hz], present paper	12.98	15.87	19.41	22.7
Difference %	0.62%	2.3%	3.7%	4.1%



(a) (b)

Figure 1: Fluid actuator general view and its different parts, (a) original fluidic actuator (b) fluidic actuator with buffer zone.

As previously stated, and in order to characterize the possible effect of the boundary conditions on the flow performance, a fluidic actuator with a buffer zone was generated, for this particular case the outlet boundary conditions were maintained the same as in the original case but the outlet was located at the end of the buffer zone, the total number of cells used in this new model was of 2854500. A single Reynolds number of 16034 was studied, the frequency obtained from this particular buffer zone model increased by 2.6% versus the one obtained with the original case. The authors have considered that the effect of buffer zone is pretty negligible and the rest of the cases will be studied without it. Finally and in order to compare the effect of the mesh on the results obtained, the original fluidic actuator was modelled via using 4.4 million cells, almost twice as much as the ones

employed initially. The simulation was done at the maximum Reynolds number Re=16034. The values of maximum x+, y+ and z+ obtained for this new mesh were of 1.8; 4.7 and 0.6 respectively. The frequency obtained when using this extremely dense mesh was of 22.83Hz which involves an increase of 0.57% versus the original case. Understanding that at lower Reynolds numbers the differences will be even smaller and considering that the time required to simulate the (FA) with 4.4 million cells is 74% higher than the one needed to perform the simulation with 2242000 cells, it can be concluded that using 2.2 million cells is precise enought for the cases under study.

3 Internal dimensions modifications

Figure 2 introduces the three modifications considered in this paper, the mixing chamber inlet width (a), was increased and decreased respectively by 64% and 114%, nine different positions were considered. The outlet width (b), increase and decrease was of 82%, again nine different positions were evaluatd. Regarding the outlet angle reduction (c), it was of 36.5%, the angle increase was of 100%, as in the previous two cases, nine different angles were simulated. In what follows the results obtained when performing these modifications are presented for a single Reynolds number, Re=16034.

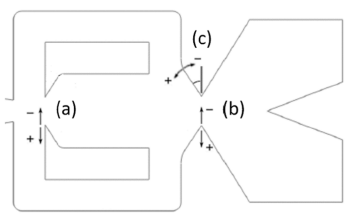


Figure 2: Fluidic actuator mixing chamber internal dimensions modifications.

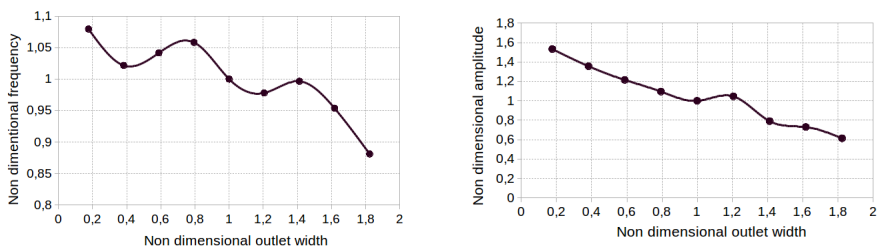
4 Results

In the present section, the main performance characteristics of a fluidic actuator having some internal dimensions modified will be presented. For all cases, the frequency from the pulsating mass flow measured at one of the two outlets and the amplitude of such pulsating flow will be presented. Some pictures will be introduced for each case to better understand the flow behaviour inside the actuator.

 ISBN: 978-84-617-8694-7

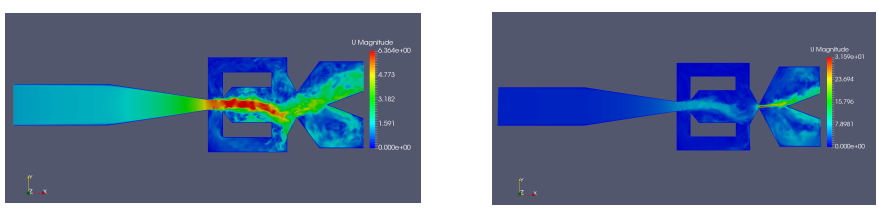
4.1 Outlet width modification

Fiure 3 presents the variation of fluidic actuator non dimensional output frequency and amplitude, whenever the outlet width is being modified. It is interesting to realize that as the width decreases the pulsating frequency increases, increasing as well the output amplitude, and vice-versa. This increase in amplitude is explained when observing that the maximum fluid velocity increases with the width decrease. The temporal mass flow amplitude and the fluid velocity amplitude go hand by hand. Figure 4 shows the velocity magnitudes inside the (FA) for the highest and lowest outlet widths evaluated in this study, notice the difference in velocity at the (FA) output. From figure 3 it must be realized that the change in frequency versus the original actuator one, for the cases evaluated, is about $\pm 10\%$ while the amplitude is suffering an increase of nearly 60% and a decrease of about 40%.



(a) (b)

Figure 3: Fluid actuator performance when modifying the outlet width. (a) Frequency variation (b) Amplitude variation.



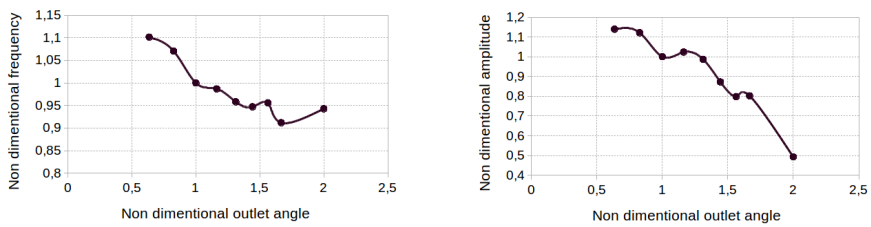
(a) (b)

Figure 4: Fluid actuator internal field velocity magnitude, (a) Maximum outlet width (b) Minimum outlet width.

 ISBN: 978-84-617-8694-7

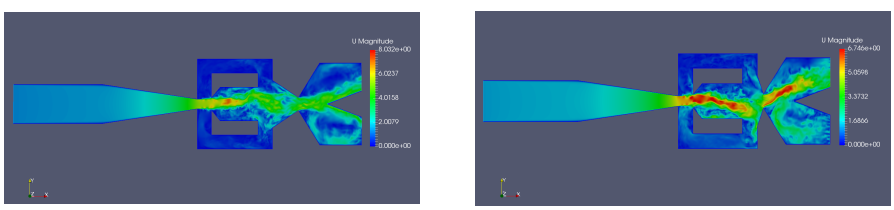
4.2 Outlet angle modification

The variation of non dimensional output frequency and amplitude when the mixing chamber outlet angle is modified, is presented in figure 5. Notice that when the angle decreases, see figures 2, 5 and 6, the output frequency tends to increase, notice as well from figure 6b), that these small angles tend to direct the flow alternatively through the feedback channels, therefore explaining why frequency increases. On the other hand, as the angle increases, the flow is being directed towards the (FA) outlet, jeopardizing the degrees of freedom of the fluid, and so minimizing the output frequency and amplitudes. Notice that the angles variation studied, affected the output frequencies by approximately ±10%, the amplitude was affected by +11% and minus 50%. Clearly, these two modifications already presented, affect much deeply the output amplitudes than the frequencies.



(a) (b)

Figure 5: Fluid actuator performance when modifying the mixing chamber outlet angle. (a) Frequency variation (b) Amplitude variation.



(a) (b)

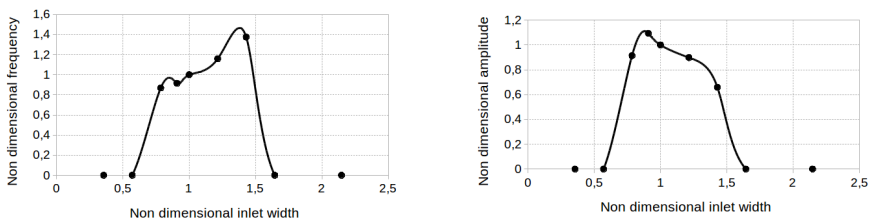
Figure 6: Fluid actuator internal field velocity magnitude, (a) Maximum outlet angle (b) Minimum outlet angle.

4.3 Mixing chamber inlet width variation

The last modification to be introduced consist of changing the mixing chamber inlet width. The first thing to be realized when seeing figures 7 and 8, is that the inlet width modifi-

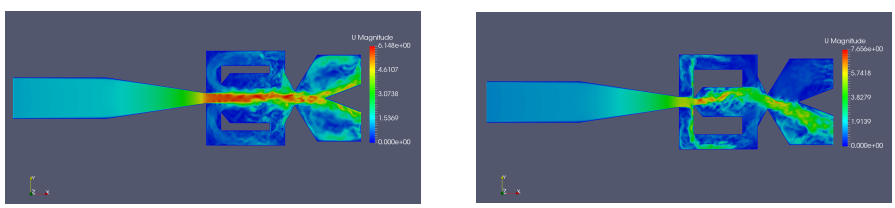
 ISBN: 978-84-617-8694-7

cation, generates a completely different pattern onto the outgoing flow. If the inlet width overcomes a certain minimum or maximum values, there is no pulsating flow at the outlet, the flow simply goes straight fom inlet to outlet. In fact, for small values of inlet width, the incoming jet borders impinge onto the walls and create a flow stream which goes from left to right,upstream to downstream, along both feedback channels at the same time, preventing any feedback from downstream to upstream. Notice that this feedback is imprescindible to generate pulsating flow inside the mixing chamber. It is interesting to ralize that an increase of the inlet width is capable of producing a frequency increase of about 40%, while making the output amplitude to decrease nearly a 30%. This tendency is completely different than the one obtained with the previous two modifications, and it has mostly do to with the fact that for the present modification, when high widths are considered, the mixing chamber incoming jet, just suffers a slight wavering inside the chamber, causing at the (FA) exit a small variation of amplitude.



(a) (b)

Figure 7: Fluid actuator performance when modifying the mixing chamber inlet width, (a) Frequency variation (b) Amplitude variation.



(a) (b)

Figure 8: Fluid actuator internal field velocity magnitude, (a) Maximum inlet width (b) Minimum inlet width.

 ISBN: 978-84-617-8694-7

5 Conclusions

A given Fluidic actuator has been studied via 3D-CFD and when modifiying its internal parameters, such as mixing chamber inlet and outlet widths, and outlet angle. When outlet width or outlet angle are being decreased, flow output frequency and amplitude increases and vice-versa. The modification of the inlet width produces quite an oposite effect, as inlet width increses the freqeuncy increases generating a decrease of the output amplitude.

Acknowledgements

The present paper presents part of the results obtained thanks to a competitive research project number FIS0016-77849-R founded by Spanish economy ministry.

References

[1] Daniel Guyot, Bernhard Bobusch, Christian Oliver Paschereit, and Surya Raghu. Active combustion control using a fluidic oscillator for asymmetric fuel flow modulation. In *44th AIAA/ASME/SAE/ASEE Joint Propulsion Conference & Exhibit*, page 4956, 2008.

[2] Arnaud Lacarelle. *Modeling, control, and optimization of fuel/air mixing in a lean premixed swirl combustor using fuel staging to reduce pressure pulsations and NOx emissions*. Universitätsverlag der TU Berlin, 2011.

[3] Ahmed Abdelrazek Emara. Interactions of flow field and combustion characteristics in a swirl stabilized burner. 2011.

[4] G Raman, S Packiarajan, G Papadopoulos, C Weissman, and S Raghu. Jet thrust vectoring using a miniature fluidic oscillator. *The Aeronautical Journal*, 109(1093):129–138, 2005.

[5] Roman Seele, Philipp Tewes, René Woszidlo, Michael A McVeigh, Nathaniel J Lucas, and Israel J Wygnanski. Discrete sweeping jets as tools for improving the performance of the v-22. *Journal of Aircraft*, 46(6):2098–2106, 2009.

[6] Ciro Cerretelli and Kevin Kirtley. Boundary layer separation control with fluidic oscillators. *Journal of Turbomachinery*, 131(4):041001, 2009.

[7] Dennis E Culley, Michelle M Bright, Patricia S Prahst, and Anthony J Strazisar. Active flow separation control of a stator vane using surface injection in a multistage

compressor experiment. In *ASME Turbo Expo 2003, collocated with the 2003 International Joint Power Generation Conference*, pages 1039–1050. American Society of Mechanical Engineers, 2003.

[8] A Seifert, O Stalnov, D Sperber, G Arwatz, V Palei, S David, I Dayan, and I Fono. Large trucks drag reduction using active flow control. *The Aerodynamics of Heavy Vehicles II: Trucks, Buses, and Trains*, pages 115–133, 2009.

[9] Ganesh Raman and Surya Raghu. Cavity resonance suppression using miniature fluidic oscillators. *AIAA journal*, 42(12):2608–2612, 2004.

[10] Brian D Giles. Fluidics, the coanda effect, and some orographic winds. *Archiv für Meteorologie, Geophysik und Bioklimatologie, Serie A*, 25(3):273–279, 1977.

[11] V Tesař, S Zhong, and F Rasheed. New fluidic-oscillator concept for flow-separation control. *AIAA journal*, 51(2):397–405, 2012.

[12] Toshihiko Shakouchi. A new fluidic oscillator, flowmeter, without control port and feedback loop. *ASME J. Dyn. Syst., Meas., Control*, 111(3):535–539, 1989.

[13] and others. Pressure and temperature insensitive flueric oscillator, October 28 1969. US Patent 3,474,805.

[14] Earl Don Webb, Roger L Schultz, Robert G Howard, James Craig Tucker, et al. Next generation fluidic oscillator. In *SPE/ICoTA Coiled Tubing Conference & Exhibition*. Society of Petroleum Engineers, 2006.

[15] James Gregory, John Sullivan, Ganesh Raman, and Surya Raghu. Characterization of a micro fluidic oscillator for flow control. In *2nd AIAA Flow Control Conference*, page 2692, 2004.

[16] Oguz Uzol and Cengiz Camci. Experimental and computational visualization and frequency measurements of the jet oscillation inside a fluidic oscillator. *Journal of Visualization*, 5(3):263–272, 2002.

[17] Peter Bauer. Fluidic oscillator flowmeter, January 13 1981. US Patent 4,244,230.

[18] Peter Bauer. Fluidic oscillator and spray-forming output chamber, November 20 1990. US Patent RE33,448.

[19] Rong Fung Huang and Kuo Tong Chang. Fluidic oscillation influences on v-shaped bluffbody flow. *AIAA journal*, 43(11):2319–2328, 2005.

 ISBN: 978-84-617-8694-7

[20] Rachid Khelfaoui, Stephane Colin, Stephane Orieux, Robert Caen, and Lucien Baldas. Numerical and experimental analysis of monostable mini-and micro-oscillators. *Heat Transfer Engineering*, 30(1-2):121–129, 2009.

[21] Ute Gebhard, H Hein, and U Schmidt. Numerical investigation of fluidic micro-oscillators. *Journal of Micromechanics and Microengineering*, 6(1):115, 1996.

[22] A Nakayama, F Kuwahara, and Y Kamiya. A two-dimensional numerical procedure for a three dimensional internal flow through a complex passage with a small depth (its application to numerical analysis of fluidic oscillators). *International Journal of Numerical Methods for Heat & Fluid Flow*, 15(8):863–871, 2005.

[23] James W Gregory, Ebenezer P Gnanamanickam, John P Sullivan, and Surya Raghu. Variable-frequency fluidic oscillator driven by a piezoelectric bender. *AIAA journal*, 47(11):2717–2725, 2009.

[24] BC Bobusch. *Experimentelle und Numerische Bestimmumg der Innerdurschtrömung eines Fluidisched Oszillators*. PhD thesis, Diplomarbeit. TU Berlin, 2010.

[25] Bernhard C Bobusch, René Woszidlo, JM Bergada, CN Nayeri, and CO Paschereit. Experimental study of the internal flow structures inside a fluidic oscillator. *Experiments in fluids*, 54(6):1559, 2013.

[26] David del Campo Sud, Bergadà Granyó, Josep Maria, and Vanessa del Campo Gatell. Preliminary study on fluidic actuators. design modifications. In *Recent advances on mechanics, materials, mechanical engineering and chemical engineering: proceedings of the International Conference on Mechanics, Materials, Mechanical Engineering and Chemical Engineering (MMMCE 2015); Barcelona, Spain, April 7-9, 2015*, pages 53–61, 2015.

[27] Masoud Baghaei, Bergadà Granyó, Josep Maria, David del Campo Sud, and Vanessa del Campo Gatell. Research on fluidic amplifiers dimensional modifications via computer simulation (cfd). In *9th International Conference on Computational Fluid Dynamics, ICCFD9, Istanbul, Turkey, July 11-15, 2016: proceedings*, pages 1–10. Istanbul Tecnical University, 2016.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

Analysis of OpenACC Performance Using Different Block Geometries

Daniel Barba[1], Arturo Gonzalez-Escribano[1] and Diego R. Llanos[1]

[1] *Departamento de Informatica, Universidad de Valladolid, Spain*

emails: daniel@infor.uva.es, arturo@infor.uva.es, diego@infor.uva.es

Abstract

OpenACC is a parallel programming model for automatic parallelization of sequential code using compiler directives or pragmas. OpenACC is intended to be used with accelerators such as GPUs and Xeon Phi. The different implementations of the standard, although still in early development, are primarily focused on GPU execution. In this study, we analyze how the different OpenACC compilers available under certain premises behave when the clauses affecting the underlying block geometry implementation are modified. These clauses are the Gang number, Worker number, and Vector Size defined by the standard.

Key words: OpenACC, GPU, block geometry, thread geometry

1 Introduction

OpenACC is an open standard intended to automatically parallelize sequential code and manage its execution in accelerators like GPUs or Xeon Phi coprocessors. It defines a number of compiler directives, also called pragmas. The main goal of OpenACC is to reduce both learning and coding time in a portable way [1]. The version of the OpenACC standard at the time of writing is the 2.5 [2].

The OpenACC standard was founded by Nvidia, CRAY, CAPS and PGI. The number of members now is larger, including both academic institutions and companies like the Oak Ridge National Laboratory, the University of Houston, AMD, and the Edinburgh Parallel Computing Centre (EPCC), among others.

There are several compilers that implement the OpenACC standard. The PGI compiler, developed by the Portland Group (subsidiary of Nvidia) is being distributed as part of

 ISBN: 978-84-617-8694-7

the Nvidia OpenACC Toolkit under a free 90-day license. Cray Inc. has its own OpenACC compiler, only available for use with their supercomputers. Pathscale Inc., a software developer for compilers and multicore software, also has an OpenACC implementation, the ENZO compiler.

Among the many academic or open-source alternatives there are the OpenUH compiler [3] by the University of Houston and accULL [4] from Universidad de La Laguna (Spain).

This work presents a study on the impact of different values for the clauses that affect the underlying block geometry of OpenACC-generated code. GPUs are very sensitive to the geometry of the thread-block chosen [5], and OpenACC makes use of the terms "gang", "worker" and "vector" in order to define different levels of parallelism. According to [6], the specification is ambiguous and this functionality depends directly on how each compiler is implemented. In this work, we measure the impact of the choice of an appropriate *thread-block geometry* when running a representative benchmark. By default, the geometry is decided by the compiler unless the specific clause is used inside the OpenACC directive. Our aim is to compare the resulting behaviour among different compilers and options. For thread-block geometry testing, we will modify the most representative of the benchmarks, testing several combinations of values for the clauses to specify gang, workers and vectors, and analyzing the differences in execution time for each compiler. This will offer some insight on the implementation of these clauses on each compiler.

Our contribution shows that the decisions made by each compiler is not always optimal, but manual tuning of the different values is not always possible for every compiler.

The rest of this paper is organized as follows. Section 2 describes the selected compilers. Section 3 shows our selected microbenchmark for testing the behaviour of the generated code when modifying the block geometry. Section 4 contains the result of our analysis about the impact on performance when changing the underlying block geometry. Finally, Section 5 concludes our paper.

2 Available Compilers

We mentioned several compilers in the Introduction. In this section we describe with more detail the compilers we were able to use for this study.

2.1 PGI Compiler

The PGI Compiler [7] is being developed by The Portland Group, being owned by Nvidia. This compiler is frequently presented in webinars, workshops, and conferences.

At the time of writing this paper, the PGI compiler is available for download as part of the OpenACC Toolkit from Nvidia. This toolkit includes a 90-day free trial, the possibility of acquiring an academic license for a whole year, or buying a commercial license.

 ISBN: 978-84-617-8694-7

2.2 accULL

The accULL [4] compiler developed by Universidad of La Laguna (Spain) is an open-source initiative. accULL consists on a structure of two layers containing YaCF [8] (Yet another Compiler Framework) and Frangollo [9], a runtime library. YaCF acts as a source-to-source translatori, while Frangollo works as an interface that provides the most common operations found in accelerators.

2.3 OpenUH

The OpenUH [3] compiler, developed by the University of Houston (USA) is another open-source initiative. It makes use of Open64, a discontinued open-source optimizing compiler.

3 Microbenchmark Description

In OpenACC, block size is defined by gangs, workers and vectors.Their choices affect the performance on memory-bound applications. To study this issue, we are going to use a very simple matrix addition implemented both in CUDA and OpenACC. Our decision is made by the fact that the problem is embarrassingly parallel, memory acceses are perfectly coalesced, and the computational load per global memory access is low (memory-bound application).

The standard only defines the different levels of parallelism, but it is up to each implementation to decide how are these levels exploited in the actual architecture. In order to obtain comparable results, some details should be taken into account. The CUDA version needs to be implemented using elastic kernels, using a fixed number of blocks, which equals the gang number in the OpenACC code. Also, since the OpenACC standard establishes that grid dimensions depend on the use of the collapse clause with nested loops, we have decided to make a one-dimensional grid. We evaluate a number of blocks in the grid ranging from one to 2048 (using only powers of two). The sequential code can be seen in Fig. 1 and the CUDA kernel in Fig. 2.

In OpenACC the X dimension of the CUDA block translates to vector length, whereas the Y dimension equals the worker number. We have also decided to use 512 threads per block, trying each possible combination of X and Y dimension values using powers of two.

4 Evaluation

In this section we analyze the impact of different choices for the geometry of the underlying thread-blocks in the OpenACC generated code.

 ISBN: 978-84-617-8694-7

```
#pragma acc data copyin(p_A[0:Size*Size],p_B[0:Size*Size]), copyout(p_C[0:Size*Size])
{
    int j;
    #pragma acc kernels
    #pragma acc loop independent gang(GANG), worker(WORKER), vector(VECTOR)
    for (j = 0; j < Size*Size; ++j)
    {
         p_C[j] = ALPHA*p_A[j] + p_B[j];
    }
} //End data region
```

Figure 1: Sequential Code for the Matrix Addition

```
__global__ void matrixKernel(float* p_A, float* p_B, float* p_C)
{
    int iterations = ((SIZE/blockDim.x)*(SIZE/blockDim.y)) / GANG;

    int iter;
    for (iter = 0; iter < iterations; ++iter)
    {
        int tx = (blockIdx.x + iter*GANG)*blockDim.x + threadIdx.x;
        int ty = threadIdx.y;
        int i = (tx/SIZE)*blockDim.y + ty;
        int j = (tx%SIZE);
        int offset = i*SIZE + j;
        if (offset < SIZE*SIZE) p_C[offset] = ALPHA*p_A[offset] + p_B[offset];
    }
}
```

Figure 2: CUDA Kernel for the Matrix Addition

 ISBN: 978-84-617-8694-7

4.1 Experimental Setup

We used a Nvidia GTX Titan Black to run the experiments. This GPU contains 2 880 CUDA cores with a clock rate of 980MHz and 15 SMs. It has 6GB of RAM, and Compute Capability 3.5. The host is a Xeon E5-2690v3 with 12 cores at a clock rate of 1.9GHz, and 64GB in four 12GB modules.

The PGI compiler is the one contained in the Nvidia OpenACC Toolkit, version 15.7-0, published in Jul 13, 2015. We used OpenUH version 3.1.0 (published in November 4, 2015), based on Open64 version 5.0 and using GCC 4.2.0, prebuilt, downloaded from the High Performance Computing Tools group website [10]. accULL is version 0.4alpha (published in November 28, 2013), downloaded from Universidad de La Laguna's research group "Computación de Altas Prestaciones" [11].

4.2 Block Geometry Sensibility of Generated Code

Observing the results obtained using the CUDA code in Fig. 3, we can see that the best results are obtained when the block number is high enough to make the GPU reach proper levels of occupation. The X dimension plays a huge role, but we expected to see an improvement in performance when the X dimension was at least 16. We suspect this is due to several factors, being the most important the behaviour of the cache when elastic kernels are used.

The results of the OpenACC code generated by PGI (Fig. 4) show that the only factor affecting the performance is the gangs number, whereas the variation of workers number and vector length does not play a significant role except when a vector length of one is used. In this case, performance is severely affected, which indicates a poor use of the cache. Overall, PGI's behaviour is the closest one to CUDA for this example.

When using OpenUH as the compiler for the OpenACC code (Fig. 5), all three parameters affect the performance of the generated code, which means that the parameters are actually being used to map the computation to the architecture resources. There is an exception when any of the three parameters is set to one. In this case, OpenUH seems to assume direct control and choose what it considers an adequate set of parameters for gangs number, workers number and vector length.

Finally, the results obtained by using accULL (Fig. 6) as our OpenACC compiler show that none of the parameters have any effect on the performance of the generated code, and it is the compiler itself who decides the value to be used.

5 Conclusions

During this work, we have realized that the OpenACC standard is very unspecific about how the different compilers should implement the three levels of parallelism. This allows

 ISBN: 978-84-617-8694-7

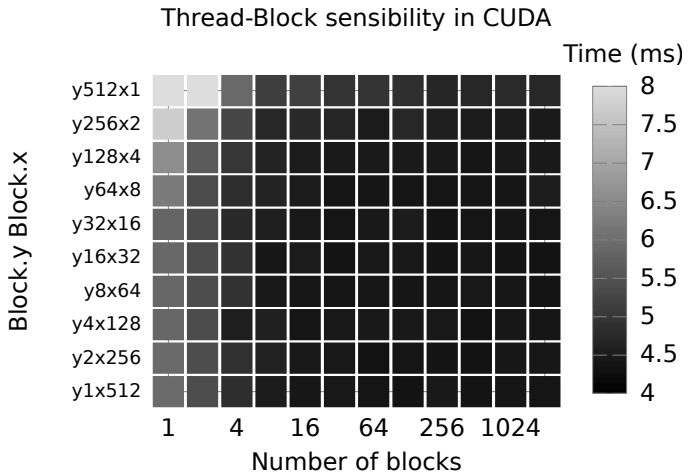


Figure 3: Effects of the Gang Number, Worker and Vector Length in the measured execution time using CUDA. Execution time is in milliseconds (lower is better). Darker is lower. Brighter is higher.

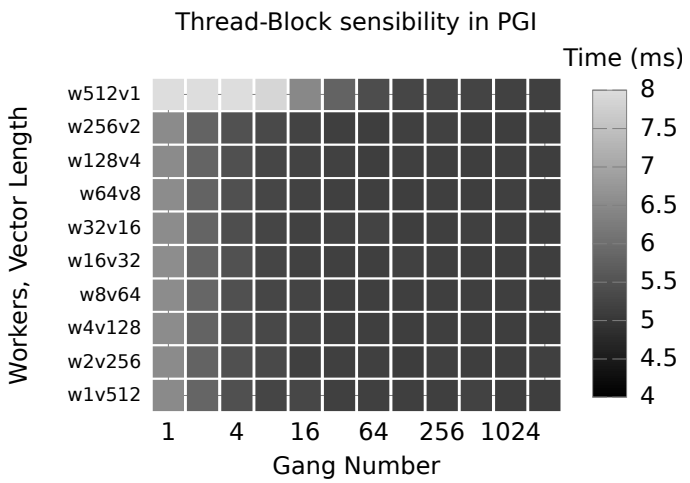


Figure 4: Effects of the Gang Number, Worker and Vector Length in the measured execution time using PGI compiler. Execution time is in milliseconds (lower is better). Darker is lower. Brighter is higher.

 ISBN: 978-84-617-8694-7

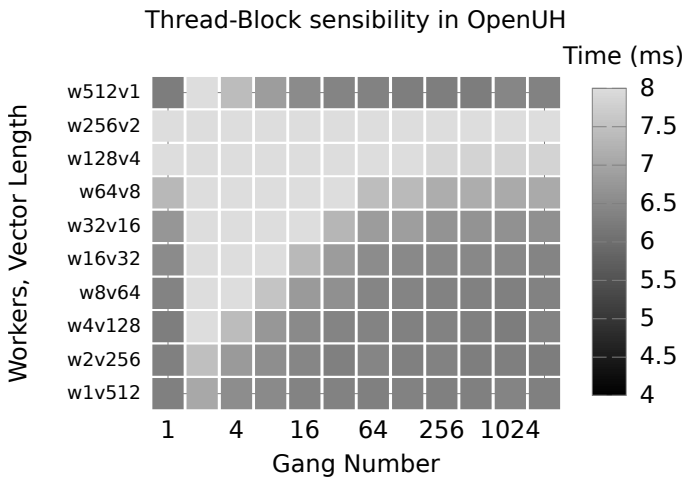


Figure 5: Effects of the Gang Number, Worker and Vector Length in the measured execution time using OpenUH compiler. Execution time is in milliseconds (lower is better). Darker is lower. Brighter is higher.

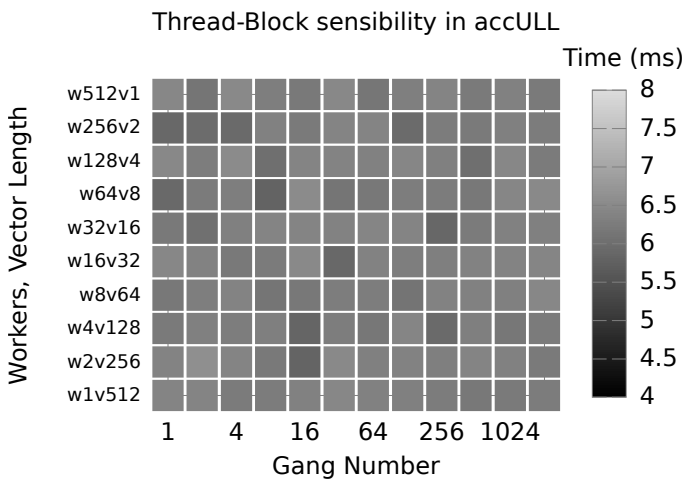


Figure 6: Effects of the Gang Number, Worker and Vector Length in the measured execution time using accULL compiler. Execution time is in milliseconds (lower is better). Darker is lower. Brighter is higher.

 ISBN: 978-84-617-8694-7

very different behaviours while unifying the basic concepts of automatic parallelization for both GPUs and Xeon Phi coprocessors.

Due to the standard leaving freedom for the implementation of the different levels of parallelism to the different compilers, the maturity of the latter directly affects the performance of the OpenACC-generated code. Thus we find that the PGI compiler, being the more mature of the analyzed compilers, generates code that resembles an optimized CUDA implementation. OpenUH shows an implementation that takes the defined levels of parallelism into consideration, but with room for a performance improvement. On the other hand, accULL seems to avoid hand-made changes to the levels of parallelism, choosing always the same configuration.

Although compiler implementations are not very mature yet, the simplicity of our microbenchmark allows us to see the effects of the variations in the different clauses: Gang number, Worker number, Vector length. Our results remark that the performance boost obtained by the tested OpenACC compilers in GPUs is dependant on the implementation of these clauses. However, since the mission of these clauses is to unify different concepts among GPUs and Xeon Phi accelerators, we argue this is complex task for the compiler implementations.

Acknowledgements

This research has been partially supported by MICINN (Spain) and ERDF program of the European Union: HomProg-HetSys project (TIN2014-58876-P), CAPAP-H5 network (TIN2014-53522-REDT), and COST Program Action IC1305: Network for Sustainable Ultrascale Computing (NESUS).

References

[1] OpenACC-standard.org, "About OpenACC."

[2] OpenACC-Standard.org, "The OpenACC application programming interface version 2.5," oct 2015.

[3] X. Tian, R. Xu, Y. Yan, Z. Yun, S. Chandrasekaran, and B. Chapman, "Compiling a high-level directive-based programming model for GPGPUs," in *Languages and Compilers for Parallel Computing*, pp. 105–120, Springer, 2014.

[4] R. Reyes, I. López-Rodríguez, J. J. Fumero, and F. de Sande, "accULL: an OpenACC implementation with CUDA and OpenCL support," in *Euro-Par 2012 Parallel Processing*, pp. 871–882, Springer, 2012.

[5] H. Ortega-Arranz, Y. Torres, A. Gonzalez-Escribano, and D. R. Llanos, "Optimizing an apsp implementation for nvidia gpus using kernel characterization criteria," *The Journal of Supercomputing*, vol. 70, no. 2, pp. 786–798, 2014.

[6] C. Wang, R. Xu, S. Chandrasekaran, B. Chapman, and O. Hernandez, "A validation testsuite for OpenACC 1.0," in *Parallel Distributed Processing Symposium Workshops (IPDPSW), 2014 IEEE International*, pp. 1407–1416, May 2014.

[7] PGI, "Pgi accelerator compilers with OpenACC directives." `https://www.pgroup.com/resources/accel.htm`, nov 2015.

[8] U. de La Laguna, "YaCF." `https://bitbucket.org/ruyman/llcomp`, nov 2015.

[9] U. de La Laguna, "Frangollo." `https://bitbucket.org/ruyman/frangollo`, nov 2015.

[10] U. of Houston, "Open-source UH compiler." `http://web.cs.uh.edu/~openuh/download/`, nov 2015.

[11] U. de La Laguna, "accULL." `http://cap.pcg.ull.es/es/accULL`, nov 2015.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

Some statistical approaches to deal with change and confusion matrices obtained from spatial data

Inmaculada Barranco-Chamorro[1]

[1] *Department of Statistics and Operations Research, University of Sevilla*

emails: chamorro@us.es

Abstract

In this paper several statistical methods are considered to deal with change and confusion matrices obtained from spatial data. The proposed methodologies focus on the study of the off-diagonal elements of these matrices. So tests for the marginal homogeneity are considered and correspondence analysis methods tailored for square matrices.

Key words: confusion matrix, correspondence analysis, marginal homogeneity

1 Introduction

Confusion matrices are a quite common tool used to asses the quality of spatial data. Usually, these matrices are obtained as result from classification of images and fotointerpretations. In practice, a variety of situations can be found in which proper statistical techniques are necessary to carry out the statistical inference associated to these matrices. For instance, we can cite: matrix size (it can range from 3×3 to more than 40×40), number of total data (from only a few to thousands), diversity of classes, dates of studies, etc.

To assess the overall accuracy of the classification depicted in a confusion matrix or to compare matrices, the kappa index is commonly used. One of the main drawbacks of this index is that it is an overall measure of agreement in a matrix. The aim of this paper is not to contribute to the debate about the salience of kappa and similar statistics for describing change or accuracy. On the other hand, our aim is to introduce statistical measures that can help us to describe and understand issues of interest in a confusion matrix and/or compare main features in two (or more) confusion matrices. Our results can be applied in two different contexts where these kind of matrices can be obtained related to spatial data:

 ISBN: 978-84-617-8694-7

1. To quantify spatial differences in land cover and land use change over two time periods. In this case, we have a square contingency table, which summarizes the coincident areas or spatial intersection of land classified over the two time periods of time. In this way, the matrix reports us about the change in the area from time 1 to time 2 in which the surveys were carried out. They are usually referred as transition or change matrices.

2. To asses the suitability of a geographical classification method or to compare several ones. In this case, we have a square contingency table showing the agreement between the mapped classes and the reference classes. So the diagonal gives us the number of elements properly classified, and the off-diagonal elements the wrong ones. These tables are usually referred as confusion matrices.

From the statistical point of view, we have, in both cases, a $k \times k$ square asymmetric contingency table where the rows and columns refer to the same set of objects, and their entries are non negative integer numbers. Quite often, standard statistical methods designed for contingency tables are unsuccessful to deal with this kind of matrices because of the strong effect of the diagonal values on the statistical summaries. So, we focus on statistical methods to deal with the off-diagonal elements of these matrices. Two approaches are considered. They are described in the next two sections.

2 Plots and statistical tests to deal with off-diagonal elements in change/confusion matrices

Agreement plots The agreement plots provide a graphical representation for the diagonal and off-diagonal elements in a change/confusion matrix. Large off-diagonal values in the matrix are depicted by the areas around the diagonal. The size, orientation and colour of rectangles indicate the direction of change (gain/loss) in a change matrix, and the existence of possible biases for a classifier in a confusion matrix. These plots are implemented in *vcd* package of R, and they can be the basis to formulate hypothesis tests about the direction of changes and the presence/absence of bias in the previously mentioned matrices.

Marginal homogeneity tests Change/confusion matrices correspond to before/after experiments, so statistical methods for matched-pairs data can be applied. In particular, marginal homogeneity tests are of interest. In these tests, the null hypothesis is the marginal homogeneity, that is, the equality (or lack of significant difference) between the row marginal proportions and the corresponding column proportions, i.e., for a $k \times k$ matrix, it would be

$$H_0 : \pi_{i.} = \pi_{.i}, \quad \text{for} \quad i = 1, \ldots, k-1$$

 ISBN: 978-84-617-8694-7

versus the alternative hypothesis, H_1 : not H_0. In H_0, $\pi_{i.}$ and $\pi_{.i}$ denote the row and the column marginal proportions, respectively.

In the present context, the marginal homogeneity would mean there was no significant change in the change matrix, and there was no significant bias in the confusion matrix. As marginality homogeneity tests, we consider the McNemar test for 2×2 tables, Generalized McNemar / Stuart- Maxwell test and Bhapkar test for $k \times k$ tables, with $k > 2$. The statistical properties of these tests will be studied from a theoretical and computational point of view.

3 Statistical methods based on correspondence analysis

Correspondence analysis (CA) is a multivariate statistical technique designed for visualizing the underlying relationships in a contingency table. There are several ways of defining CA, see for instance Benzécri [2] or Greenacre [4]. We highlight that CA is a least square method of data analysis based on the singular value decomposition (SVD) of certain matrix associated to this technique. For square asymmetric tables, such as those we obtain dealing with spatial data, standard CA usually fails due to the strong effect of the diagonal elements. However, following Greenacre [4], the table can be splitted into symmetric and skew symmetric components. The separate analysis of these parts allows us to gain in interpretability and representation of information. So two separate two-dimensional maps can be obtained which involve one set of points each one. Both maps must be constructed on the same scale in order to show the relative sizes of the variation in each component. In this way, the magnitude of off-diagonal values in change/confusion matrices can be interpreted. The use of these methods is illustrated in change/confusion matrices obtained from real spatial data.

Acknowledgements

This work has been partially supported by grant CTM2015-68276-R.

References

[1] A. Agresti, *Categorical Data Analysis*, Wiley, 2013.

[2] J.P. Benzécri et al. , *L'Analyse des Données, vol.2, l'Analyse des Correspondences*, Dunod, Paris, 1973.

[3] A. Comber et al., *Methods to quantify regional differences in land cover change*, Remote Sens. **8** 176 (2016) 1–19.

[4] M. Greenacre, *Correspondence analysis of square asymmetric matrices*, Appl. Statist. **49** Part 3 (2000) 297–310.

[5] R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/. (2017)

[6] Z. Yang, X. Sun and J.W. Hardin, *Testing marginal homogeneity in clustered matched-pair data*, J. Stat. Plan. Infer. **141** (2011) 1313–1318.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

Hermite interpolation by many-knot cubic splines: error analysis

D. Barrera[1]

[1] *Department of Applied Mathematics, University of Granada, 18071-Granada, Spain*

emails: dbarrera@ugr.es

Abstract

The interpolation error is analyzed for three different schemes for Hermite interpolation by many-knot cubic splines. Two of them are local and the third one is global. The first method is of order two, and the two other are of order one. The error estimates are given for C^4 functions.

Key words: Spline functions, B-splines, Hermite interpolation, Bernstein basis
MSC 2000: 41A05, 41A15, 65D07

1 Introduction

Cubic spline interpolation on $\mathbb{R}$ at a strictly increasing sequence $X := (x_i)_{i\in\mathbb{Z}}$ of primary knots can be made more flexible by adding secondary knots, i.e. by using a sequence of knots $T := (t_i)_{i\in\mathbb{Z}}$ containing X.

A very detailed description of different situations in Lagrange interpolation can be found in Dahmen, Goodmand and Micchelli (1988), following a preceding study by Qi (1981), whereas Hermite interpolation is analyzed in Qi and Zhou (1982).

On the other had, a Jackson-type estimate is established for the interpolation error in the completely local Lagrangian scheme in de Villiers (1993).

We analyze the interpolation error associated with three different schemes for Hermite interpolation by cubic splines obtained by adding two knots in each interval $[x_i, x_{i+1}]$ symmetrically distributed around the midpoint. This last restriction is reasonable since the minimization of the uniform error in any scheme designed to recover all functions of a certain smoothness from their values at a fixed finite set of points leads to consider canonical knots of the type indicated above, as shown in Micchelli, Rivlin and Winograd (1976).

The first problem concerns the error analysis of the Hermite interpolation scheme of order two shose data are $(f(x_i), f'(x_i), f''(x_i))_{i\in\mathbb{Z}}$. When the data $f''(x_i)$ are not available, they can be replace by $p''(x_i)$, where p is the cubic polynomial interpolating $f(x_j)$, $j = i-1, j, j+1$, and $f'(x_i)$. The error for the resulting Hermite interpolation scheme of order one is then analyzed. Finally, a third Hermite scheme of order one is proposed and its interpolation error analyzed. The second derivative of the interpolant at the knots is now the solution of a tridiagonal linear system arising from C^3 continuity at x_i.

2 The Hermite interpolation problem of order two

Let $X := (x_i)_{i\in\mathbb{Z}}$ be some arbitrary strictly increasing sequence of real numbers and $Y := (y_{i,j})_{i\in\mathbb{Z},j=1,2}$ be a sequence satisfying the conditions

$$x_i < y_{i,1} < y_{i,2} < x_{i+1},\ i \in \mathbb{Z},$$

where

$$y_{i,1} := x_i + \alpha_i h_i \quad \text{and} \quad y_{i,2} := x_i + (1-\alpha_i) h_i,\ i \in \mathbb{Z},$$

with

$$h_i := x_{i+1} - x_i,\ 0 < \alpha_i < \frac{1}{2}.$$

Define $T := (t_i)_{i\in\mathbb{Z}} = X \cup Y$, where

$$t_{3i} = x_i,\ t_{3i+1} = y_{i,1},\ t_{3i+2} = y_{i,2},\ i \in \mathbb{Z},$$

and consider the space

$$S_4(T) := \left\{ f \in C^2(\mathbb{R}) : f_{|[t_i,t_{i+1}]} \in \mathbb{P}_3 \right\},$$

where $\mathbb{P}_3$ denotes the space of all polynomials of degree at most three.

The Schoenberg-Whitney theorem (see [3]) allows us to state that for any any $i \in \mathbb{Z}$ there exist functions φ_i, ψ_i and χ_i in $S_4(T)$ such that

$$\operatorname{supp}\varphi_i = \operatorname{supp}\psi_i = \operatorname{supp}\chi_i = [x_{i-1}, x_{i+1}]$$

and

$$\begin{aligned}
&\varphi_i(x_j) = \delta_{i,j},\ \varphi_i'(x_j) = 0,\ \varphi_i''(x_j) = 0,\\
&\psi_i'(x_j) = \delta_{i,j},\ \psi_i(x_j) = 0,\ \psi_i''(x_j) = 0,\\
&\chi_i''(x_j) = \delta_{i,j},\ \chi_i(x_j) = 0,\ \chi_i'(x_j) = 0.
\end{aligned}$$

 ISBN: 978-84-617-8694-7

The functions φ_i, ψ_i and χ_i can be expressed in terms of normalized B-splines associated with the knots in T (see [1, 4]) or by means of Bernstein polynomials (see [2]) in each subinterval of their support.

These functions allow us to define an Hermite interpolant. For a C^2 function f, to interpolate $\left(f\left(x_i\right), f'\left(x_i\right), f''\left(x_i\right)\right)_{i\in\mathbb{Z}}$ we define the operator

$$H^{(2)} : C^2\left(\mathbb{R}\right) \longrightarrow S_4\left(T\right)$$

by means of

$$H^{(2)} f := \sum_{i\in\mathbb{Z}} \left(f\left(x_i\right)\varphi_i\left(x\right) + f'\left(x_i\right)\psi_i\left(x\right) + f''\left(x_i\right)\chi_i\left(x\right)\right).$$

This operator is exact on $\mathbb{P}_3$.

The following result holds.

Proposition 1 *For every $i \in \mathbb{Z}$ and any $x \in [x_i, x_{i+1}]$, define, for a function $f \in C^4\left(\mathbb{R}\right)$,*

$$\mathcal{L}f = f\left(x\right) - H^{(2)} f\left(x\right).$$

Then

$$\left|\mathcal{L}f\right| \leq \frac{3 - 15\alpha_i + 16\alpha_i^2 + 48\alpha_i^3 - 96\alpha_i^4 + 32\alpha_i^5}{1152\left(1-\alpha_i\right)} h_i^4 \left\|f^{(4)}\right\|_i,$$

where $\|\cdot\|_i$ denotes the uniform norm in $[x_i, x_{i+1}]$.

3 The Hermite interpolation problem of order one

When the the second derivative is not available at the knots, it is possible to estimate it from the values at some points in a neighbourhood. Let p the cubic polynomial interpolating the data $f\left(x_j\right)$, $j = i-1, j, j+1$, and $f'\left(x_i\right)$. Then $f''\left(x_i\right)$ is estimated from

$$s_i := p''\left(x_i\right) = a_{i-1} f\left(x_{i-1}\right) + b_i f\left(x_i\right) + b_i' f'\left(x_i\right) + c_{i+1} f\left(x_{i+1}\right),$$

where

$$\begin{aligned} a_{i-1} &= \frac{2h_i}{h_{i-1}^2\left(h_{i+1}+h_i\right)}, \\ b_i &= -2\frac{h_{i-1}^3 + h_i^3}{h_{i-1}^2\left(h_{i-1}+h_i\right)h_i^2}, \quad b_i' = 2\frac{h_{i-1}-h_i}{h_{i-1}h_i}, \\ c_{i+1} &= 2\frac{h_{i-1}}{\left(h_{i-1}+h_i\right)h_i^2}. \end{aligned}$$

 ISBN: 978-84-617-8694-7

We define the operator

$$H^{(1)} : C^1(\mathbb{R}) \longrightarrow S_4(T)$$

by means of

$$H^{(1)} f := \sum_{i\in\mathbb{Z}} \left(f(x_i)\,\varphi_i(x) + f'(x_i)\,\psi_i(x) + s_i\chi_i(x) \right).$$

Equivalently,

$$H^{(1)} f = \sum_{i\in\mathbb{Z}} \left(f(x_i)\,\widetilde{\varphi}_i(x) + f'(x_i)\,\widetilde{\psi}_i(x) \right),$$

with

$$\begin{aligned} \widetilde{\varphi}_i &= \varphi_i + c_i\chi_{i-1} + b_i\chi_i + a_i\chi_{i+1}, \\ \widetilde{\psi}_i &= \psi_i + b_i'\chi_i. \end{aligned}$$

We have the following result.

Proposition 2 *Consider $f \in C^4(\mathbb{R})$. For every $i \in Z$ and $x \in [x_i, x_{i+1}]$ the following error estimate holds:*

$$\left| H^{(1)} f(x) - f(x) \right| \le C(\alpha_i)\, h_i^3 \left\| f^{(4)} \right\|_{i-1,3\ j\in\{i-1,i,i+1\}} \max \{h_j\},$$

where

$$\|g\|_{i-1,3} := \max_{x_{i-1}\le x\le x_{i+2}} |g(x)|$$

and

$$C(\alpha_i) := \frac{32\alpha_i^5 - 96\alpha_i^4 + 48\alpha_i^3 - 3\alpha_i + 3}{1152\,(1-\alpha_i)}.$$

4 Another Hermite interpolation scheme of order one

Let $I = [a, b]$ be an interval of $\mathbb{R}$ al let us consider two sets $X_n := (x_i)_{0\le i\le n}$ and $Y_n := (y_{i,j})_{0\le i\le n, j=1,2}$ such that $x_0 = a$, $x_n = b$, and for $0 \le i \le n-1$

$$\begin{aligned} x_i < y_{i,1} < y_{i,2} < x_{i+1}, \\ y_{i,1} = x_i + \alpha_i h_i,\ y_{i,2} = x_i + (1-\alpha_i)\,h_i, \end{aligned}$$

with $h_i := x_{i+1} - x_i$ and $0 < \alpha_i < \frac{1}{2}$.

We define $T_n := X_n \cup Y_n = (t_i)_{0\le i\le 3n}$, where

$$\begin{aligned} t_{3i} &= x_i, \\ t_{3i+1} = y_{i,1}, \quad t_{3i+2} &= y_{i,2}, \quad 0 \le i \le n-1, \\ t_{3n} &= x_n, \end{aligned}$$

 ISBN: 978-84-617-8694-7

and consider the space $S_4(I,T_n)$ of C^2 cubic splines on the partition of I induced by T_n. There exist functions $\varphi_{i,n}$, $\psi_{i,n}$ and $\chi_{i,n}$ in $S_4(I,T_n)$ such that

$$\begin{aligned}\operatorname{supp}\varphi_{0,0}=\operatorname{supp}\psi_{0,0}=\operatorname{supp}\chi_{0,0}&=[x_0,x_1],\\ \operatorname{supp}\varphi_{i,n}=\operatorname{supp}\psi_{i,n}=\operatorname{supp}\chi_{i,n}&=[x_{i-1},x_{i+1}],\ 1\leq i\leq n-1,\\ \operatorname{supp}\varphi_{n,n}=\operatorname{supp}\psi_{n,n}=\operatorname{supp}\chi_{n,n}&=[x_{n-1},x_n],\end{aligned}$$

and

$$\begin{aligned}\varphi_{i,n}(x_j)&=\delta_{i,j}, \quad \varphi'_{i,n}(x_j)=\varphi''_{i,n}(x_j)=0,\\ \psi'_{i,n}(x_j)&=\delta_{i,j}, \quad \psi_{i,n}(x_j)=\psi''_{i,n}(x_j)=0,\\ \chi''_{i,n}(x_j)&=\delta_{i,j}, \quad \chi_{i,n}(x_j)=\chi'_{i,n}(x_j)=0\end{aligned}$$

for $0\leq j\leq n$.

Proposition 3 *We define the operator*

$$H_n^{(2)}:C^2(I)\longrightarrow S_4(I,T_n)$$

by means of

$$H_n^{(2)}f=\sum_{i=0}^{n}\left(f(x_i)\,\varphi_{i,n}(x)+f'(x_i)\,\psi_{i,n}(x)+f''(x_i)\,\chi_{i,n}(x)\right).$$

For each $\lambda=(\lambda_0,\lambda_1,\ldots,\lambda_n)$, define the operator

$$S_{n,\lambda}:C^1(I)\longrightarrow S_4(I,T_n)$$

as follows:

$$S_{n,\lambda}f=\sum_{i=0}^{n}f(x_i)\,\varphi_{i,n}(x)+f'(x_i)\,\psi_{i,n}(x)+\lambda_i\chi_{i,n}(x).$$

Proposition 4 *Consider $f\in C^2(I)$ and define $\lambda_0=f''(x_0)$ and $\lambda_n=f''(x_n)$. Then there exist unique values λ_i, $1\leq i\leq n-1$, such that for $\lambda=(\lambda_0,\lambda_1,\ldots,\lambda_n)$ the function $S_{n,\lambda}f$ is of class C^3 at the knots x_i, $1\leq i\leq n-1$.*

The following result holds.

Proposition 5 *Consider $f\in C^4(I)$. Then, for each $0\leq i\leq n-1$ and $x\in[x_i,x_{i+1}]$ the following inequality is satisfied:*

$$|S_{n,\lambda}f(x)-f(x)|\leq\frac{h_i^2}{1152\,(1-\alpha_i)}D^{(1)}_{\underline{\alpha},\underline{h}}\left\|f^{(4)}\right\|_{\infty,I},$$

 ISBN: 978-84-617-8694-7

where

$$D^{(1)}_{\underline{\alpha},\underline{h}} := 4\alpha_i\,(3-4\alpha_i)\,D^{(2)}_{\underline{\alpha},\underline{h}} + p\,(\alpha_i)\,h_i^2,$$

with

$$\begin{aligned}\underline{\alpha} &:= (\alpha_0,\alpha_1,\ldots,\alpha_{n-1})\,,\\ \underline{h} &:= (h_0,h_1,\ldots,h_{n-1})\,,\\ p\,(\alpha_i) &:= 3-15\alpha_i+16\alpha_i^2+48\alpha_i^3-96\alpha_i^4+32\alpha_i^5\end{aligned}$$

and

$$D^{(2)}_{\underline{\alpha},\underline{h}} := \min_{1\le j\le n-2}\left\{h_j\alpha_j\,(1-\alpha_j)\right\}\max_{0\le j\le j\le n-1}\left\{\frac{1-6\alpha_j+18\alpha_j^2-16\alpha_j^3}{\alpha_j\,(1-\alpha_j)}h_j\right\}.$$

References

[1] C. de Boor. A practical guide to splines. Springer, New York, 1978.

[2] W. Dahmen, T.N.T. Goodman and C.A. Micchelli. Compactly supported fundamental functions for spline interpolation. Numer. Math. 52 (1988) 639–644.

[3] W. Dahmen, T.N.T. Goodman and C.A. Micchelli. Local spline interpolation schemes in one and several variables. In: A. Gómez, F. Guerra, M.A. Jiménez and G. López (eds.), Approximation & Optimisation, Lecture Notes in Mathematics 1354. Springer, 1987, pp. 11–24.

[4] G. Farin. Curves and surfaces for Computer Aided Geometric Design. Academic Press, San Diego, CA, 2nd ed., 1990.

[5] S. Karlin. Total Positivity. Stanford University Press, Stanford, CA, 1968.

[6] C.A. Micchelli, T.J. Rivlin, S. Winograd. The optimal recovery of smooth functions. Numer. Math. 26 (1976) 191–200.

[7] D.X. Qi. A class of local explicit many-knots spline interpolation schemes. MRC Technical Summary Report, No. 2238, Madison, WI, 1981.

[8] D.X. Qi, S.Z. Zhou. Local explicit many-knot spline Hermite approximation schemes. Technical Summary Report, No. 2359, Madison, WI, 1982.

[9] L.L. Schumaker. Spline functions. Basic theory. Wiley, New York, 1981.

[10] J.M. de Villiers. A convergence result in nodal spline interpolation. Journal of Approximation Theory 74 (1993) 266–279.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

Non uniform quasi-interpolation for solving nonlinear Fredholm integral equations of the second kind

Domingo Barrera[1], Fadila El Mokhtari[1,2] and Driss Sbibih[2]

[1] *Department of Applied Mathematics, University of Granada, Granada, Spain*

[2] *Department of Mathematics, University Mohammed I, FSO, Oujda, Morocco*

emails: dbarrera@ugr.es, elmokhtari.fadila@gmail.com, sbibih@yahoo.fr

Abstract

In this paper, we use the Nyström method and a quadrature formula based on non-uniform quasi-intepolation for solving nonlinear Fredholm integral equations of the second kind. This method results in a system of nonlinear equations for determining approximate values of the true solution of the integral equations. Numerical examples illustrate the performance of the method.

Key words: Quasi-interpolation, Nonlinear Fredholm integral equations of the second kind, Nyström method Quadrature formula

1 Introduction

We consider the nonlinear Fredholm integral equation

$$u(x) - \int_a^b k(x,t)\, g(u(t))\, dt = f(x), \ x \in I := [a,b], \tag{1}$$

where f is a known continuous function defined on I, and g is a nonlinear function defined on I. The nonlinear integral operator $\mathcal{K}$ is defined as

$$\mathcal{K}u(x) := \int_a^b k(x,t)\, g(u(t))\, dt, \ x \in I.$$

For a continuous kernel, $\mathcal{K}$ is a compact operator on $C(I)$ into $C(I)$.

 ISBN: 978-84-617-8694-7

Many different methods are presented in the literature to solve (1) (see, [4, 5, 6, 7] and references therein). In this work, we construct a quadrature formula based on quadratic spline quasi-interpolation on non-uniform partition, then we approximate the integral operator with this formula, i.e.

$$\mathcal{K}u\left(x\right) \approx \sum_{j} \omega_j k\left(x, \theta_j\right) g\left(u\left(\theta_j\right)\right).$$

where ω_j and θ_j denote the weights and the knots of the formula, respectively. Then we obtain the approximate solution by using the Nyström method.

2 Quadratic spline quasi-interpolation on non-uniform partition

Let $P := \{a = x_0 < x_1 < \cdots < x_{N-1} < x_N = b\}$ be a non-uniform partition of $I = [a, b]$ into N sub-intervals $I_i := [x_{i-1}, x_i]$ of length $h_i := x_i - x_{i-1},\ 1 \leq i \leq N$. The space $S_3(P)$ of splines of order 3 on P admits a basis composed by $N+1$ B-splines $\{B_j,\ 0 \leq j \leq N+1\}$ constructed from an extended partition associated with P. We choose the extended partition with multiple knots at the endpoints, i.e. $x_{-2} = x_{-1} = x_0 = a$ and $x_{N+2} = x_{N+1} = x_N = b$. Then, the support of B_j is the interval $[x_{j-2}, x_{j+1}]$ (see [1]).

We consider the quadratic spline quasi-interpolant to a function f given by

$$\mathcal{Q}f := \sum_{j=0}^{N+1} \mu_j(f) B_j,$$

where the functional coefficients $\mu_j(f)$ are specific linear combinations of the values of f on the set $\Theta := \{\theta_i,\ 0 \leq j \leq (N+1)\}$, where $\theta_0 := a$, $\theta_j := \frac{1}{2}\left(x_{j-1} + x_j\right),\ 1 \leq j \leq N$, and $\theta_{N+1} := b$.

More precisely,

$$\mu_0(f) = f_0,\ \mu_{N+1}(f) = f_{N+1}$$

$$\mu_j(f) = \alpha_j f_{j-1} + \beta_j f_j + \gamma_j f_{j+1},\ 1 \leq j \leq N,$$

where $f_j := f(\theta_j)$, $0 \leq j \leq N+1$,

$$\alpha_j = -\frac{\tau_j^2 \tau'_{j+1}}{\tau_j + \tau'_{j+1}},\ \beta_j = 1 + \tau_j \tau'_{j+1},\ \gamma_j = -\frac{(\tau'_{j+1})^2 \tau_j}{\tau_j + \tau'_{j+1}}$$

and

$$\tau_j = \frac{h_j}{h_{j-1} + h_j},\ \tau'_j = \frac{h_{j-1}}{h_{j-1} + h_j}.$$

 ISBN: 978-84-617-8694-7

We can write the spline quasi-interpolant Q under the following quasi-Lagrange form (see [3])

$$\mathcal{Q}f = \sum_{j=0}^{N+1} f(\theta_j) L_j,$$

with

$$\begin{aligned} L_0 &:= B_0 + \alpha_1 B_1,\ L_1 := \beta_1 B_1 + \alpha_2 B_2, \\ L_j &:= \gamma_{j-1} B_{j-1} + \beta_j B_j + \alpha_{j+1} B_{j+1},\ 2 \le j \le N-1, \\ L_N &:= \gamma_{N-1} B_{N-1} + \beta_N B_N,\ L_{N+1} := \alpha_N B_N + B_{N+1} \end{aligned}$$

Theorem 1 *[3] For f bounded on I and for any partition P of I, the infinity norm of the quadratic quasi-interpolant $\mathcal{Q}_P$ is uniformly bounded by* 2.5.

As a consequence of Theorem 1, the exactness of $\mathcal{Q}_P$ on the space $\mathbb{P}_2$ of quadratic polynomials and a classical result on approximation [1], the following result holds:

Theorem 2 *There exists a constant C independent of h. such that for all $f \in C^3(I)$ and for all partition P of I, with $h := \max h_i$,*

$$\|f - \mathcal{Q}_P f\|_\infty \le C h^3 \left\| f^{(3)} \right\|_\infty .$$

3 Quadrature formula

For any continuous function f, the quadrature formula associated with the quadratic spline quasi-interpolant on a non-uniform partition is obtained by integrating $\mathcal{Q}f$ in the above quasi-Lagrange form:

$$\mathcal{I}_{\mathcal{Q}}(f) = \int_a^b \mathcal{Q}f(t)\, dt = \sum_{j=0}^{N+1} \omega_j f(\theta_j),$$

where $\omega_j := \int_a^b L_j(t)\, dt$.
Explicitly, the weights ω_j are

$$\begin{aligned} \omega_0 &:= \vartheta_0 + \alpha_1 \vartheta_1,\ \omega_1 := \beta_1 \vartheta_1 + \alpha_2 \vartheta_2, \\ \omega_j &:= \gamma_{j-1} \vartheta_{j-1} + \beta_j \vartheta_j + \alpha_{j+1} \vartheta_{j+1},\ 2 \le j \le N-1, \\ \omega_N &:= \gamma_{N-1} \vartheta_{N-1} + \beta_N \vartheta_N,\ \omega_{N+1} := \alpha_N \vartheta_N + \vartheta_{N+1}, \end{aligned}$$

where

$$\vartheta_0 := \frac{1}{3} h_1,\ \vartheta_1 := \frac{1}{3}(h_1 + h_2),\ \vartheta_N := \frac{1}{3}(h_{N-1} + h_N),\ \vartheta_{N+1} := \frac{1}{3} h_N$$
$$\vartheta_j := \frac{1}{3}(h_{j-1} + h_j + h_{j+1}),\ 2 \le j \le N-1.$$

 ISBN: 978-84-617-8694-7

Theorem 3 *Let assume that the points of evaluation* $\Theta := \{\theta_i,\ 0 \leq j \leq (N+1)\}$ *are symmetric with respect to the midpoint of* I*. Then for* $f \in C^4(I)$*, we have*

$$\int_a^b f(t)\,dt - \mathcal{I}_{\mathcal{Q}}(f) = \mathcal{O}\left(h^4\right).$$

4 Nyström method

For $g \in C(I)$, we set

$$\mathcal{K}u(x) \approx \mathcal{K}_N u(x) = \sum_{j=0}^{N+1} \omega_j k(x,\theta_j)\, g(u(\theta_j)).$$

Thus, we approximate the integral equation (1) by

$$u_N(x) - \sum_{j=0}^{N+1} \omega_j k(x,\theta_j)\, g(u_N(\theta_j)) = f(x),\ x \in I. \tag{2}$$

This is equivalent to first solve the system of nonlinear equations

$$u_N(\theta_i) - \sum_{j=0}^{N+1} \omega_j k(\theta_i,\theta_j)\, g(u_N(\theta_j)) = f(\theta_i),\ 0 \leq i \leq N+1$$

where the unknows are $\{u_N(\theta_i),\ 0 \leq i \leq N+1\}$. Then, after using the Nyström method formula and solving the nonlinear system, we obtain the following approximate solution

$$u_N(x) = f(x) + \sum_{j=0}^{N+1} \omega_j k(x,\theta_j)\, g(u_N(\theta_j)),\ x \in I$$

Theorem 4 *Assume* $k \in C^{0,4}\left(I^2\right)$ *and* $g \in C^4(I)$*. Let* u *be the solution of the integral equation (1) and* u_N *be the solution of the approximate integral equation (2). Then, it holds*

$$\|u - u_N\| = \mathcal{O}\left(h^4\right).$$

5 Numerical results and comparisons

In this section, five examples are given to illustrate the results established in the previous sections

$$u_i(x) - \int_0^1 k_i(x,t)\, g_i(u_i(t))\,dt = f_i(x),\ x \in [0,1].$$

 ISBN: 978-84-617-8694-7

i	$k_i(x,t)$	$u_i(x)$	$g_i(x)$
1	xt	$2-x^2$	$\sqrt{x}$
2	$\frac{1}{5}\cos(\pi x)\sin(\pi t)$	$\sin(\pi x)+\frac{1}{3}\left(20-\sqrt{391}\right)\cos(\pi x)$	x^3
3	e^{x-t}	x	$\cos(x)$
4	$-x$	x	e^x
5	$-e^{x-2t}$	e^x	x^3

Table 1: Data for the tested integral equations.

N	$\|u_i-u_{N,i}\|_{\infty,I}$				
	$i=1$	$i=2$	$i=3$	$i=4$	$i=5$
8	$2.79\,(-6)$	$2.43\,(-5)$	$1.64\,(-6)$	$3.41\,(-7)$	$3.01\,(-7)$
16	$1.82\,(-7)$	$1.43\,(-6)$	$1.05\,(-7)$	$2.18\,(-8)$	$1.92\,(-8)$
32	$1.14\,(-8)$	$8.88\,(-8)$	$6.53\,(-9)$	$1.36\,(-9)$	$1.20\,(-9)$
64	$7.09\,(-10)$	$5.54\,(-9)$	$4.07\,(-10)$	$8.46\,(-11)$	$7.48\,(-11)$
128	$4.43\,(-11)$	$3.46\,(-10)$	$2.55\,(-11)$	$5.29\,(-12)$	$4.67\,(-12)$
256	$2.77\,(-12)$	$2.16\,(-11)$	$1.59\,(-12)$	$3.30\,(-13)$	$2.91\,(-13)$

Table 2: Estimated infinity norms for the numerical approximations to the solutions of the tested integral equations.

The corresponding data are given in Table 1. The function f_i is chosen so that u_i is the solution of the integral equation.

Let us consider the partition of I given by the extrema of Chebyshev polynomials of the first kind of degree N on the interval I, namely

$$P=\left\{x_i=a+\frac{b-a}{2}\left(\cos\left(\frac{(N-i)\,\pi}{N}\right)+1\right),\ 0\leq i\leq N\right\}.$$

For different values of N, we present in Table 2 the maximum errors of the approximate solution obtained by using our method. We also give in Table 3 the numerical convergence orders, denoted by $\mathcal{NCO}$.
In order to give a comparison with other methods, we present some numerical results obtained by these methods in Table 4 and Table 5.

 ISBN: 978-84-617-8694-7

	$\mathcal{NCO}$				
	$i=1$	$i=2$	$i=3$	$i=4$	$i=5$
8/16	3.93739	4.0861	3.9703	3.97032	3.97031
16/32	4.00176	4.01039	4.00438	4.00438	4.00438
32/64	4.00168	4.00249	4.0019	4.00189	4.00191
64/128	4.00051	4.00062	4.0005	4.00045	4.00076
128/256	4.0009	4.00016	3.99951	4.0023	4.00481

Table 3: Numerical convergence orders from the results in Table 2.

N	Haar wavelet method (2013)	Our method
8	$1.0\,(-3)$	$2.79\,(-6)$
16	$2.6\,(-4)$	$1.82\,(-7)$
32	$6.6\,(-5)$	$1.14\,(-8)$
64	$1.7\,(-5)$	$7.09\,(-10)$
128	$4.2\,(-6)$	$4.43\,(-11)$

Table 4: Comparison of the errors with the Haar wavelet method [6] for equation 1.

N	Triangular factorization method (2010)	Our method
8	$9.9\,(-3)$	$2.43\,(-5)$
16	$2.5\,(-3)$	$1.43\,(-6)$
32	$7.9\,(-4)$	$8.88\,(-8)$

Table 5: Comparison of the errors with the triangular function method [5] for equation 2.

 ISBN: 978-84-617-8694-7

References

[1] R.A DeVore and G.G. Lorentz, *Constructive Approximation*. Springer, Berlin (1993).

[2] P. Sablonnière, *A quadrature formula associated with a univariate quadratic spline quasi-interpolant*. BIT **47** (2007) 825–837.

[3] P. Sablonnière, *Quadratic spline quasi-interpolants on bounded domains of* $\mathbb{R}^d$, $d = 1, 2, 3$.Rend. Sem. Mat. Univ. Pol. Torino **61** (2003) 61–78, Spline and radial functions.

[4] C. Allouch, D. Sbibih and M. Tahrichi, *Superconvergent Nystrm and degenerate kernel methods for Hammerstein integral equations*, Comput. Appl. Math. **258** (2014) 30–41.

[5] K. Maleknejad, H. Almasieh and M. Roodaki, *Triangular functions (TF) method for the solution of nonlinear VolterraFredholm integral equations*, Commun Nonlinear Sci. Numer. Simulat. **15** (2010) 3293–3298.

[6] I. Aziz and S. U. Islam, *New algorithms for the numerical solution of nonlinear Fredholm and Volterra integral equations using Haar wavelets*, Comput. Appl. Math. **239** (2013) 333–345.

[7] E. Babolian and A. Shahsavaran, *Numerical solution of nonlinear Fredholm integral equations of the second kind using Haar wavelets*, Comput. Appl. Math. **225** (2009) 87–95.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

A spline quasi-interpolation based method to obtain the reset voltage in Resistive RAMs in the Charge-Flux domain

D. Barrera[1], M. J. Ibáñez[1], F. Jiménez-Molinos[2], A. M. Roldán[2] and J. B. Roldán[2]

[1] *Department of Applied Mathematics, University of Granada, 18071-Granada, Spain*

[2] *Department of Electronics, University of Granada, 18071-Granada, Spain*

emails: dbarrera@ugr.es, mibanez@ugr.es, jmolinos@ugr.es, amroldan@ugr.es, jroldan@ugr.es

Abstract

Resistive RAMs (RRAMs) are the most promising devices for near future in terms of non-volatile applications. The devices present amazing features that make both Academia and Industry deep in the research of these devices. Among the characteristics, we can list the very good endurance, short writing and reading times in comparison to current Flash technology, low voltage operation, CMOS compatibility, etc. This new technology needs advances in all the fronts that need to be addressed prior to industrialization. One of them is connected with compact modeling, i.e, the development of analytical expressions to account for the most important physical effects that are needed to calculate the current, capacitance, transient response, etc. The models are essential for circuit simulation and design. The device models should be accurate and this issue is achieved by implementing the correct physics in a flexible and robust mathematical architecture. We will focus on this latter problem in this work. In particular, we will deal with a good numerical approximation of experimental data based on B-splines construction to perform the integrals of the current and voltage as function of time. We do so to transform the usual modeling domain, consisting of a current-voltage representation, to a charge-flux domain; i.e., the time integral of the current and voltage measured variables. In this new domain an interesting modeling approach can be developed. In particular, once the new representation is obtained, we have introduced a new method to obtain the reset voltage of RRAMs.

The new numerical procedure we have implemented allows a correct integration and help with the treatment of the usual measurement noise that is found in these kind of devices, since their operation is based on the stochastic processes that form and

 ISBN: 978-84-617-8694-7

rupture internal conductive filaments. The main features of the mathematical technique we propose along with a practical example built upon real experimental data will be explained.

Key words: Resistive RAM, parameter extraction, B-splines, quasi-interpolation
MSC 2000: 41A15, 65D07, 65D15

1 Introduction

Since the breakthrough that represented the announcement of HP [14] of having successfully implemented the first memristor, these elements have drawn a great deal of interest both in the Academia and in the Industry. Particularly, among the great variety of devices that work as memristors, those based on resistive switching (RS) mechanisms, usually known as Resistive RAMs (RRAMs), can store information without the need of a power source when they are switched off. These RRAMs present interesting features concerning their capability for low power applications, switching speed, endurance, etc., [17, 14]. Thus, RRAMs are a promising emerging new technology for non-volatile memories [9].

Many papers devoted to the fabrication and characterization of RRAMs are being published nowadays [17, 9]; in addition, modeling and simulation tools are being developed to analyze the physics of these devices [15, 7, 3, 6]. The availability of compact models to simulate circuits based on these new devices is essential for the development of this type of technology [9]. Several models have been presented to describe these devices, although new models accounting for the main characteristics of RRAMs are needed because of the variety of physical mechanisms that been detected in these devices [9, 8].

The important issue connected to the intrinsic variability in RRAMs is essential. The cycle to cycle variability is mainly due to the random processes that lead to the formation and destruction of the conductive filaments. In fact, it is so unpredictable that it has even been used to implement random number generators. As an example of this variability, see below, where a set of experimentally measured reset transitions are shown [5]. We would like to highlight that all these curves correspond to different RS cycles of the same device, so the big spread from cycle to cycle is apparent.

The modeling of RRAM can be performed in different domains; i.e., in a I-V domain or in a charge (Q) and flux (ϕ), which in some cases can be simpler as reported by L. Chua [2] and others [13] in the past. This Q-flux domain can be an advantage for modeling since the typical parameters that characterize a technology could be more easily calculated.

The flux ϕ can be calculated as the time integral of the voltage while the charge Q is obtained as the current time integral [2]. Notice that a more formally accurate name for these magnitudes would be voltage momentum and current momentum, as explained in [2]. However for simplicity we will keep the names flux and charge:

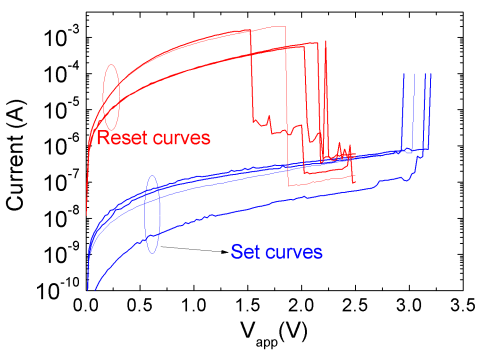
 ISBN: 978-84-617-8694-7

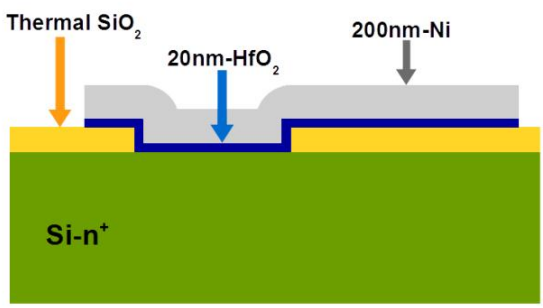
$$\phi(t) = \int_{t_0}^{t} V(\tau)\, d\tau \qquad \text{and} \qquad Q(t) = \int_{t_0}^{t} I(\tau)\, d\tau.$$

The calculation of these magnitudes and the determination of parameters such as reset voltage and set voltage need of an accurate representation of the experimental measurements at the mathematical level. We do so in this work, were advanced numerical techniques based on B-Splines are employed to build an approximant that allow the accurate calculation of the integrals to determine the charge and flux previously reported.

In order to apply the new numerical technique to real devices and experimental measurements we have employed structures fabricated at the Institut of Microelectronics of Barcelona IMB-CNM (CSIC), they are based on a Ni/HfO_2/$Si - n^+$ stack (see figure below). The ALD fabricated dielectric layer was 20nm thick; further details of the fabrication process and measurement setup can be found in [5]. The conduction is filamentary in these devices, i.e., it takes place trough conductive filaments (CFs) that are formed and destroyed within the RS device operation [5, 16].

A few experimental current-voltage curves corresponding to several RS (set-reset) cycles are shown below (left). Reset curves for devices based on a Ni/HfO_2/$Si - n^+$ stack are shown in red while set curves are plotted in blue. The curves correspond to different cycles from a set-reset series of almost three thousands cycles. Although the Ni electrode had a negative voltage applied while the substrate was grounded [5], we have considered absolute values for the applied voltage for the sake of simplicity.The V_{Reset} and V_{Set} points are highlighted. The curves are different in all the cases due to the stochastic nature of the processes behind the conductive filament formation that fix the device resistance, i.e., the ratio between the device voltage and the corresponding current [9, 5]. Also the physical estructure of the devices fabricated is shown [5].





 ISBN: 978-84-617-8694-7

2 Mathematical background

Spline quasi-interpolation is an easy method to define approximants to a given function f defined on a interval $I := [a, b]$. The class of the quasi-interpolant can be fixed in advance by using the appropriate spline space. Here we use C^1 quadratic quasi-interpolation from the values of f at the points in $X := \{x_0, x_1, \ldots, x_n\}$.

Let us suppose that $a = x_0 < x_1 < \cdots < x_n = b$. Let $\mathcal{S}_2(X)$ be the space of C^1 quadratic splines on this partition. It is a $n + 2$-dimensional linear space. To compute a good basis $\mathcal{B}_2(X) := \{B_j, 1 \leq j \leq n + 2\}$ for $\mathcal{S}_2(X)$ some auxiliary knots $x_{-2} \leq x_{-1} \leq a$ and $b \leq x_{n+1} \leq x_{n+2}$ are needed. We choose $x_{-2} = x_{-1} = a$ and $x_{n+1} = x_{n+2} = b$. From the extended partition $X_* := \{x_{-2}, x_{-1}, x_0, x_1, \ldots, x_n, x_{n+1}, x_{n+2}\}$, the B-splines B_j can be defined in terms of divided differences as follows:

$$B_j(x) := (x_j - x_{j-3})\,[x_{j-3}, x_{j-2}, x_{j-1}, x_j]\,(\cdot - x)_+^2,$$

where $(\cdot)_+^2$ stands for the quadratic truncated power. The B-spline B_j is supported on the interval $[x_{j-3}, x_j]$ and positive on (x_{j-3}, x_j). It holds (see [12, Theorem 4.21 and Remark 4.1]) that

$$\sum_{j=1}^{n+2} B_j(x) = 1,\ \sum_{j=1}^{n+2} \theta_j^{(1)} B_j(x) = x,\ \ \sum_{j=1}^{n+2} \theta_j^{(2)} B_j(x) = x^2,\ x \in I,$$

where

$$\theta_j^{(1)} := \theta_j := \frac{1}{2}(x_{j-2} + x_{j-1}) \quad \text{and} \quad \theta_j^{(2)} := x_{j-2}x_{j-1}.$$

Taking into account the main goal of this contribution, among all the possibilities at our disposal we associate with f the quasi-interpolant $Q_2 f \in S_2(X)$ defined as

$$Q_2 f := \sum_{j=1}^{n+2} \mu_j(f)\, B_j,$$

where every coefficient $\mu_j(f)$ is a linear combination of values of f at some points lying in a neighbourhood of the support of B_j. More precisely, we impose the following structure for $\mu_j(f)$:

$$\begin{aligned}
\mu_1(f) &:= \alpha_1 f(x_0) + \beta_1 f(x_1) + \gamma_1 f(x_2),\\
\mu_j(f) &:= \alpha_j f(x_{j-2}) + \beta_j f(x_{j-1}) + \gamma_j f(x_j),\ 2 \leq j \leq n+1,\\
\mu_{n+2}(f) &:= \alpha_{n+2} f(x_{n-2}) + \beta_{n+2} f(x_{n-1}) + \gamma_{n+2} f(x_n).
\end{aligned}$$

It can be proved that there are unique coefficients α_j, β_j and γ_j such that the the associated quasi-interpolation operator Q_2 is exact on $\mathbb{P}_2$, i.e. such that $Q_2 p = p$ for all quadratic

 ISBN: 978-84-617-8694-7

polynomial p (see [1]). Since Q is a uniformly bounded operator (see [10, 11]) and exact on $\mathbb{P}_2$, there exists a constant C independent of f and X such that

$$\|f - Q_2 f\|_{\infty,I} \leq Ch^3,$$

where $h := \max_{1\leq i\leq n} h_i$ and $h_i := x_i - x_{i-1}$.

3 Algorithm

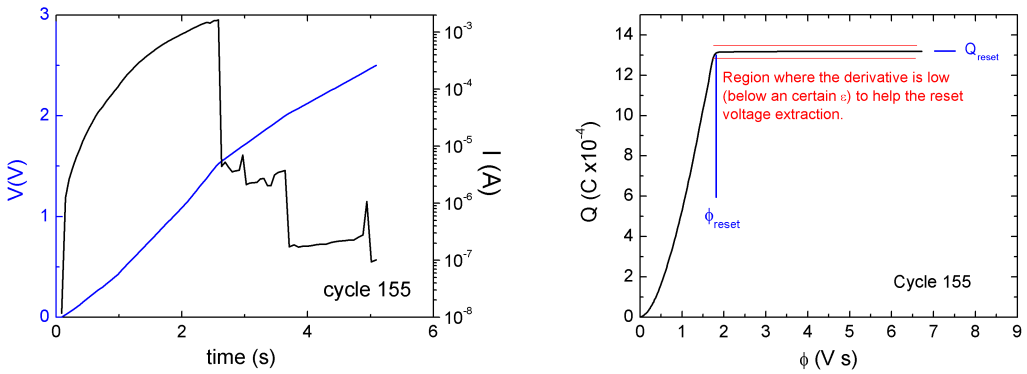
Once we know how to construct a spline quasi-interpolant to a given function, we have to determine approximations to the voltage V and the current I from experimental measurements V_i and I_i made at times x_i.To do so,

1. Define in the space $S_2(X)$ with knots x_i the C^1 quadratic quasi-interpolants V_{app} and I_{app} of V and I, respectively.
2. Define approximations ϕ_{app} and Q_{app} of ϕ and Q, respectively, as follows:
$$\phi_{app}(t) = \int_{x_0}^{t} V_{app}(\tau)\, d\tau \qquad \text{and} \qquad Q_{app}(t) = \int_{x_0}^{t} I_{app}(\tau)\, d\tau.$$
3. Consider a uniform partition of the interval $[x_0, x_n]$ and compute the numerical derivative of Q_{app} with respect to ϕ_{app} at the resulting points, $\overline{x}_i$.
4. Let $J_\epsilon := \min\left\{j : \left|\frac{dQ_{app}}{d\phi_{app}}(\overline{x}_j)\right| < \epsilon\right\}$ for a given threshold ϵ.
5. Let the reset charge Q_{rst} be equal to $Q_{app}(\overline{x}_{J_\epsilon})$.
6. Fit a model $Q = Q(\phi)$ to data $Q_{app}(\overline{x}_i)$, $i = 0, \ldots, k$, for a enough large k.
7. Let the estimated reset flux be the solution of the equation $Q(\phi_{rst}) = Q_{rst}$.

4 An example

The proposed method has been applied to real measurements in order to calculate the charge versus flux relationship. The original data measured are shown in figure bottom left (curve 155 was selected among a long series of resistive switching cycles). The corresponding charge-flux representation employed to estimate the reset flux and finally the reset voltage has been determined using quadratic quasi-interpolation from the values at the knots of V and I.

 ISBN: 978-84-617-8694-7



It shows voltage applied to the device and current versus measurement time (left) and the charge versus flux calculated making use of the data plotted. Some of the parameters employed in the proposed algorithm are plotted for the sake of clarity.

Notice that the region with low derivative (below a certain threshold that can be fitted, corresponding to step 5 in our algorithm) is shown to be contained among the red lines. Once this region is isolated, the reset charge can be estimated, as suggested in the algorithm (Q_{reset}, step 5). Once the determination of Q_{reset} is done, the ϕ_{reset} is obtained as the value needed for the final reset voltage determination. The application of the algorithm to the current curve under study led us to obtain an estimated reset flux equal to 1.72 and an estimated reset voltage of 1.52 V.

Acknowledgements

We thank the Spanish Ministry of Economy and Competitiveness for Project TEC2014-52152-C3-2-R (also supported by the FEDER program).

References

[1] G. Chen, C.K. Chui, M.J. Lai. Construction of real-time spline quasi-interpolation schemes. Approx. Theory Appl. 4 (1988) 61–75.

[2] L.O. Chua. Resistance switching memories are memristors. Applied Physics A 102 (2011) 765–783.

[3] R. Degraeve et al. Hourglass concept for RRAM: A dynamic and statistical device model. Proceedings of the 21th International Symposium on the Physical and Failure Analysis of Integrated Circuits (IPFA) (2014).

 ISBN: 978-84-617-8694-7

[4] R.A. DeVore, G.G. Lorentz. Constructive Approximation. Springer Verlag, 1993.

[5] M.B. González, J. Rafi, O. Beldarrain, M. Zabala, F. Campabadal. Analysis of the switching variability in Ni/HfO2-based RRAM devices. IEEE Transactions on Device and Materials Reliability 14 (2014) 769771.

[6] F. Jiménez-Molinos, M.A. Villena, J.B. Roldán, A.M. Roldán. A SPICE compact model for unipolar RRAM reset process analysis. IEEE Transactions on Electron Devices 62 (2015) 955–962.

[7] S. Larentis, F. Nardi, S. Balatti, D.C. Gilmer, D. Ielmini. Resistive switching by voltage-driven ion migration in bipolar RRAM. part II: Modeling. IEEE Transactions on Electron Devices 59 (2012) 24682475.

[8] J. Lee, S. Lee, T. Noh. Resistive switching phenomena: A review of statistical physics approaches. Applied Physics Reviews 2 (2015) 031303.

[9] F. Pan, S. Gao, C. Chen, C. Song, F. Zeng. Recent progress in resistive random access memories: materials, switching mechanisms and performance. Materials Science and Engineering 24 (2014) 421.

[10] P. Sablonnière. Spline quasi-interpolants and applications to numerical analysis. Rend. Sem. Univ. Pol. Torino 63 No 2 (2005) 107–118.

[11] P. Sablonnière. Quasi-interpolants splines : Exemples et applications. ESAIM: Proceedings 20 (2007) 195–207.

[12] L.L. Schumaker. Spline Functions: Basic Theory. Wiley, New York (1973).

[13] S. Shin, K. Kim, S.M. Kang. Compact models for memristors based on charge-flux constitutive relationships. Computer-Aided Design of Integrated Circuits and Systems 29 (2010) 590–598.

[14] D.B. Strukov, G.S. Snider, D.R. Stewart, R.S. Williams. The missing memristor found. Nature 453 (2008) 80–83.

[15] M.A. Villena, F. Jiménez-Molinos, J.B. Roldán, J. Sune, S. Long, X. Lian, F. Gámiz, M. Liu. An in-depth simulation study of thermal reset transitions in resistive switching memories. Journal of Applied Physics 114 (2013) 144505.

[16] M.A. Villena, M.B. González, F. Jiménez-Molinos, F. Campabadal, J.B Roldán, J. Sune, E. Romera, E. Miranda. Simulation of thermal reset transitions in resistive switching memories including quantum effects. Journal of Applied Physics 115 (2014) 214504.

 ISBN: 978-84-617-8694-7

[17] R. Waser, M. Aono. Nanoionics-based resistive switching memories. Nature Materials 6 (2007) 833–840.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

BIGEOMETRIC COMPLEX CALCULUS

Agamirza E. Bashirov[1,2] and Sajedeh Norozpour[1]

[1] *Department of Mathematics, Eastern Mediterranean University, Famagusta*

[2] *Institute of Control Systems, ANAS, Baku*

emails: agamirza.bashirov@emu.edu.tr, sajedeh.norozpour@cc.emu.edu.tr

Abstract

In most (if not all) textbooks on complex calculus the differentiation and integration of complex functions are presented by employing the algebraic form of complex variables because the respective formulae in terms of polar form are inappropriate. In this paper we demonstrate that by transferring the field structure of the system of complex numbers to the Riemann surface of complex logarithm and changing the sense of derivative and integral, complex calculus can be delivered in terms of polar form of complex variable identically to the presentation in terms of algebraic form.

Key words: complex calculus, complex logarithm, complex exponent, complex differentiation, complex integration, bigeometric calculus
MSC 2000: Primary: 30E20; Secondary: 30E99

1 Introduction

Complex analysis is a continuation of real analysis to complex variables, nicely explicating many eccentric results of the real case. In most (if not all) textbooks on complex calculus [6, 9] the differentiation and integration of complex functions are presented by employing the algebraic form of complex variables. The polar form is used only in a few situations when the capacity of the algebraic form is insufficient.

We investigate the presentation of complex calculus essentially on the basis of the polar form of complex variables and demonstrate that if the field structure of $\mathbb{C}$ is transferred to the Riemann surface of the complex logarithm, which will be denoted by $\mathbb{B}$, and the sense of the complex differentiation and integration is changed to bigeometric, then the polar presentation of complex calculus over $\mathbb{B}$ is identical to its algebraic presentation over $\mathbb{C}$.

2 Elements of bigeometric real calculus

In the 60th decade Michael Grossman and Robert Katz [7] gave an underlying idea for creating different presentations of Newtonian calculus. Bigeometric calculus is one of them. Briefly, on the basis of the exponential function e^x and its inverse the field structure of the real number system $\mathbb{R}$ can be transferred to the interval $(0,\infty)$ by setting exp-operations

(i) $a \oplus_{\exp} b = e^{\ln a + \ln b} = ab$,

(ii) $a \otimes_{\exp} b = e^{\ln a \ln b} = a^{\ln b} = b^{\ln a}$,

(iii) $a \ominus_{\exp} b = e^{\ln a - \ln b} = a/b$,

(iv) $a \oslash_{\exp} b = e^{\ln a / \ln b} = a^{1/\ln b}$.

The neutral elements of exp-addition and exp-multiplication are 1 and e, respectively.

The exp-operations create exp-derivative and exp-integral. Both of them can be expressed in terms of the derivative and integral of Newtonian calculus. An informal derivation of these expressions are as follows:

$$\begin{aligned}(f(y) \ominus_{\exp} f(x)) \oslash_{\exp} (y \ominus_{\exp} x) &= (f(y)/f(x))^{\frac{1}{\ln(y/x)}} \\ &= (f(y)/f(x))^{\frac{1}{y-x}\cdot\frac{y-x}{\ln y-\ln x}} \\ &= \left(e^{\frac{\ln f(y)-\ln f(x)}{y-x}}\right)^{\frac{y-x}{\ln y-\ln x}} \\ &\to e^{\frac{(\ln f(x))'}{(\ln x)'}} = e^{x(\ln f(x))'}\end{aligned}$$

and

$$\begin{aligned}\bigoplus_{i=1}^{n}{}_{\exp} f(c_i) \otimes_{\exp} (x_i \ominus_{\exp} x_{i-1}) &= \prod_{i=1}^{n} f(c_i)^{\ln(x_i/x_{i-1})} \\ &= e^{\sum_{i=1}^{n} \ln f(c_i)(x_i-x_{i-1})\cdot\frac{\ln x_i-\ln x_{i-1}}{x_i-x_{i-1}}} \\ &\to e^{\int_a^b \ln f(x)(\ln x)'dx} = e^{\int_a^b \frac{\ln f(x)}{x}dx}.\end{aligned}$$

These formulae gave rise to two pairs of exp-derivative and exp-integral. Bigeometric derivative and integral or, briefly, $^{\pi}$derivative and $^{\pi}$integral are defined by

$$f^{\pi}(x) = e^{x(\ln f(x))'} \quad \text{and} \quad \int_a^b f(x)^{\mathbf{d}x} = e^{\int_a^b \frac{\ln f(x)}{x}dx}.$$

They are initiated by Grossman [8]. It is seen that if they are applied repeatedly, then the factors x and $1/x$ eliminate each other making $^{\pi}$form the fundamental theorem of calculus.

 ISBN: 978-84-617-8694-7

Deleting these factors simplifies the formulae for $^{\pi}$derivative and $^{\pi}$integral. Consequently, multiplicative derivative and integral or, briefly, *derivative and *integral, which are widely investigated in [1, 2], are defined by

$$f^*(x) = e^{(\ln f(x))'} \quad \text{and} \quad \int_a^b f(x)^{dx} = e^{\int_a^b \ln f(x)\, dx}.$$

Bigeometric calculus or, briefly, $^{\pi}$calculus is based on the $^{\pi}$derivative and $^{\pi}$integral.

3 The field $\mathbb{B}$

Previously, complex calculus was investigated by use of *operations in [3, 4, 5] over the field $\mathbb{C}$ of complex numbers. Deficiencies of the obtained results can be removed if we transfer the field structure of $\mathbb{C}$ to the Riemann surface of complex logarithm and use $^{\pi}$operations.

The Reimann surface of the complex logarithm can be defined in the algebraic form as

$$\mathbb{B} = \{(r, \theta) : r > 0,\ -\infty < \theta < \infty\}.$$

The subset

$$\mathbb{B}_\alpha = \{(r, \theta) : r > 0,\ \alpha - \pi < \theta < \alpha + \pi\}$$

of $\mathbb{B}$ will be called an α-branch of $\mathbb{B}$. We will identify $\mathbb{B}_0$ and $\mathbb{C} \setminus \{0\}$ by

$$\mathbb{B}_0 \ni (r, \theta) = re^{i\theta} \in \mathbb{C} \setminus \{0\},$$

where θ is the principal argument of $z = re^{i\theta} \in \mathbb{C}$. Therefore, $\mathbb{C} \setminus \{0\} \subset \mathbb{B}$. Similarly, $\mathbb{R} \setminus \{0\} \subset \mathbb{B}$ with $x = (x, 0)$ if $x > 0$ and $x = (-x, -\pi)$ if $x < 0$, where $\mathbb{R}$ is the system of real numbers.

We will denote the elements of $\mathbb{B}$ by boldface letters like $\mathbf{z}$. The boldface letters like $\mathbf{f}$ will be used for functions with domain or range essentially in $\mathbb{B}$ as well. The symbols like f will be used for functions with domain and range in $\mathbb{C}$. We will refer to an element from $\mathbb{R}$ as r-number, from $\mathbb{C}$ as c-number, and from $\mathbb{B}$ as b-number for brevity.

The addition and subtraction of c-numbers lose the sense for b-numbers whereas the multiplication and division operations can be modified to $\mathbb{B}$ as follows

$$\mathbf{z}_1\mathbf{z}_2 = (r_1, \theta_1)(r_2, \theta_2) = (r_1 r_2, \theta_1 + \theta_2)$$

and

$$\mathbf{z}_1/\mathbf{z}_2 = (r_1, \theta_1)/(r_2, \theta_2) = (r_1/r_2, \theta_1 - \theta_2),$$

assuming that $\mathbf{z}_1 = (r_1, \theta_1)$ and $\mathbf{z}_2 = (r_2, \theta_2)$. In fact, these are **exp**-addition and **exp**-subtraction on $\mathbb{B}$ if we introduce **exp**- and **log**-functions as follows.

 ISBN: 978-84-617-8694-7

Proposition 1. *The function*

$$\mathbf{E}^z = (e^x, y),\ z = x + iy \in \mathbb{C}, \tag{1}$$

where e^x is a natural exponent of the r-number x, is a bijection from $\mathbb{C}$ onto $\mathbb{B}$. Its inverse is a function from $\mathbb{B}$ to $\mathbb{C}$ defined by

$$\mathbf{Log}\,\mathbf{z} = \ln r + i\theta,\ \mathbf{z} = (r, \theta) \in \mathbb{B}, \tag{2}$$

where $\ln r$ is the natural logarithm of the r-number r. Moreover,

(a) For $\mathbf{z}_1, \mathbf{z}_2 \in \mathbb{B}$, $\mathbf{Log}\,(\mathbf{z}_1\mathbf{z}_2) = \mathbf{Log}\,\mathbf{z}_1 + \mathbf{Log}\,\mathbf{z}_2$.

(b) For $\mathbf{z}_1, \mathbf{z}_2 \in \mathbb{B}$, $\mathbf{Log}\,(\mathbf{z}_1/\mathbf{z}_2) = \mathbf{Log}\,\mathbf{z}_1 - \mathbf{Log}\,\mathbf{z}_2$.

(c) For $z_1, z_2 \in \mathbb{C}$, $\mathbf{E}^{z_1+z_2} = \mathbf{E}^{z_1}\mathbf{E}^{z_2}$.

(d) For $z_1, z_2 \in \mathbb{C}$, $\mathbf{E}^{z_1-z_2} = \mathbf{E}^{z_1}/\mathbf{E}^{z_2}$.

The proof of this proposition is elementary. Note that our motivation to use the symbols $\mathbf{E}^z$ and $\mathbf{Log}\,\mathbf{z}$ in (1) and (2) is as follows. If

$$\bar{\mathbb{B}} = \{(r, \theta) : r \geq 0,\ -\infty < \theta < \infty\},$$

then the periodic extension of $\mathbf{E}^z$ to $\bar{\mathbb{B}}$ has the form

$$\mathbf{e}^{\mathbf{z}} = \mathbf{e}^{(r,\theta)} = (e^{r\cos\theta}, r\sin\theta),\ \mathbf{z} = (r, \theta) \in \bar{\mathbb{B}}.$$

Since the range of $\mathbf{e}^{\mathbf{z}}$ equals to $\mathbb{B}$, the inverse of $\mathbf{e}^{\mathbf{z}}$ is the multivalued function

$$\mathbf{log}\,\mathbf{z} = \left(\sqrt{\ln^2 r + \theta^2}, \mathrm{atan2}\,(\theta, \ln r) + 2\pi n\right),\ n = 0, \pm 1, \ldots,\ \mathbf{z} = (r, \theta) \in \mathbb{B},$$

where $\mathrm{atan2}\,(y, x)$ is the arctan-function of two variables. Therefore, $\mathbf{Log}\,\mathbf{z}$ becomes the principal branch of $\mathbf{log}\,\mathbf{z}$ for $n = 0$. This is very similar to the exponential and logarithmic functions of complex calculus.

By Proposition 1, $\mathbf{z}_1 \oplus_{\mathbf{exp}} \mathbf{z}_2 = \mathbf{e}^{\mathbf{Log}\,\mathbf{z}_1 + \mathbf{Log}\,\mathbf{z}_2} = \mathbf{z}_1\mathbf{z}_2$ and $\mathbf{z}_1 \ominus_{\mathbf{exp}} \mathbf{z}_2 = \mathbf{E}^{\mathbf{Log}\,\mathbf{z}_1 - \mathbf{Log}\,\mathbf{z}_2} = \mathbf{z}_1/\mathbf{z}_2$ which are defined previously. Thus the product and ratio of $\mathbf{z}_1$ and $\mathbf{z}_2$, defined previously, are indeed **exp**-additon and **exp**-subtraction on $\mathbb{B}$, respectively. This motivates to define **exp**-multiplication and **exp**-division on $\mathbb{B}$ by

$$\mathbf{z}_1 \otimes_{\mathbf{exp}} \mathbf{z}_2 = \mathbf{E}^{\mathbf{Log}\,\mathbf{z}_1 \mathbf{Log}\,\mathbf{z}_2} = (e^{\ln r_1 \ln r_2 - \theta_1\theta_2}, \theta_2 \ln r_1 + \theta_1 \ln r_2)$$

and

$$\mathbf{z}_1 \oslash_{\mathbf{exp}} \mathbf{z}_2 = \mathbf{E}^{\mathbf{Log}\,\mathbf{z}_1 / \mathbf{Log}\,\mathbf{z}_2} = \left(e^{\frac{\ln r_1 \ln r_2 + \theta_1\theta_2}{\ln^2 r_2 + \theta_2^2}}, \frac{\theta_1 \ln r_2 - \theta_2 \ln r_1}{\ln^2 r_2 + \theta_2^2}\right),$$

assuming that $\mathbf{z}_1 = (r_1, \theta_1)$ and $\mathbf{z}_2 = (r_2, \theta_2)$. These operations make $\mathbb{B}$ a field with the neutral elements of **exp**-additon and **exp**-multiplication to be $\mathbf{0} = (1, 0)$ and $\mathbf{1} = (e, 0)$, respectively. For breviety, the subscript in the symbols of these operations will be dropped. For example, instead of $\oplus_{\mathbf{exp}}$ we will write $\oplus$.

 ISBN: 978-84-617-8694-7

4 Elementary functions on $\mathbb{B}$

Complex functions have analogs in complex $^{\pi}$calculus. The $^{\pi}$analog of a complex function f will be denoted by $\mathbf{f}$.

The $^{\pi}$analogs of e^z and $\log z$ are the periodic $\mathbf{e^z}$ and multivalued $\mathbf{\log z}$. One can verify the equality

$$\mathbf{E}^z = \mathbf{E}^{e^{\mathbf{Log\, z}}} \quad \text{for} \quad \mathbf{z} = z \in \mathbb{B}_0 \subset \mathbb{B}.$$

This motivates to define the $^{\pi}$analog of a single-valued complex function f by $\mathbf{f}(\mathbf{z}) = \mathbf{E}^{f(\mathbf{Log\, z})}$, $\mathbf{z} \in B \subseteq \mathbb{B}$. In such a way, it can be derived the analogs of the complex trigonometric and hyperbolic functions in the form

$$\begin{aligned}
\mathbf{cos\, z} &= (e^{\cosh\theta \cos\ln r}, -\sinh\theta \sin\ln r),\\
\mathbf{sin\, z} &= (e^{\cosh\theta \sin\ln r}, \sinh\theta \cos\ln r),\\
\mathbf{cosh\, z} &= (e^{\cos\theta \cosh\ln r}, \sin\theta \sin\ln r),\\
\mathbf{sinh\, z} &= (e^{\cos\theta \sinh\ln r}, \sin\theta \cosh\ln r).
\end{aligned}$$

Almost similar rule can be applied to multivalued functions of complex calculus as well. For example, the raising to c-power of b-number is defined by $\mathbf{z}^w = \mathbf{E}^{w\log(\mathbf{Log\, z})}$, which suggests the definition

$$\mathbf{z^w} = \mathbf{e^{w \otimes log\,(log\, z)}}$$

for the raising to b-power. This is a multivalued function.

5 $^{\pi}$Derivative

The following lemma is crucial for setting $^{\pi}$derivative.

Lemma 1. *Let f be a non-vanishing function from some nonempty open connected subset C of $\mathbb{C}$ to $\mathbb{C}$. Assume that f has the algebraic and polar representations*

$$f(z) = u(r,\theta) + iv(r,\theta) = R(r,\theta)e^{i\Theta(r,\theta)} \quad \textit{for} \quad z = re^{i\theta}.$$

If $f'(z)$ exists and, consequently, the Cauchy–Riemann conditions in polar form

$$ru'_r = v'_\theta \quad \textit{and} \quad rv'_r = -u'_\theta$$

hold, then

$$r(\ln R)'_r = \Theta'_\theta \quad \text{and} \quad r\Theta'_r = -(\ln R)'_\theta \tag{3}$$

and

$$z(\log f(z))' = r((\ln R)'_r + i\Theta'_r), \tag{4}$$

assuming that f transfers C into a branch of log-function.

 ISBN: 978-84-617-8694-7

Note that, (3) expresses a version of the Cauchy–Riemann conditions when both argument and value of a complex function are represented in the polar form.

Motivated from the definition of $^{\pi}$derivative in the real case and Lemma 1, the $^{\pi}$derivative of $\mathbf{f}(r,\theta) = (R(r,\theta), \Theta(r,\theta))$ can be defined in the following way.

Definition 1. A function $\mathbf{f}(r,\theta) = (R(r,\theta), \Theta(r,\theta))$ is said to be $^{\pi}$differentiable at $\mathbf{z} = (r,\theta)$ if R and Θ have continuous partial derivatives and the Cauchy–Riemann conditions from (3) hold at (r,θ). If $\mathbf{f}$ is $^{\pi}$differentiable at $\mathbf{z}$, then its $^{\pi}$derivative at $\mathbf{z}$(in accordance to (4)) will be defined by

$$\mathbf{f}^{\pi}(\mathbf{z}) = (e^{r(\ln R)'_r}, r\Theta'_r).$$

$\mathbf{f}$ is said to be $^{\pi}$analytic on $B \subseteq \mathbb{B}$ if $\mathbf{f}$ is $^{\pi}$differentiable at every $\mathbf{z} \in B$.

Theorem 1. *Let* $\mathbf{f}$ *be the* $^{\pi}$*analog of the single-valued function* f*. Then its* $\mathbf{f}^{\pi}$ *and* f' *are related as*

$$\mathbf{f}^{\pi}(\mathbf{z}) = \mathbf{E}^{f'(\mathbf{Log}\,\mathbf{z})},$$

assuming that at least one of them exists.

This theorem establishes a method for calculation of $^{\pi}$derivatives. For example,

- $(\mathbf{a} \otimes \mathbf{z} \oplus \mathbf{b})^{\pi} = \mathbf{a}$.
- $(\mathbf{E}^{z})^{\pi} = \mathbf{E}^{z}$ and also $(\mathbf{e}^{\mathbf{z}})^{\pi} = \mathbf{e}^{\mathbf{z}}$.
- $(\mathbf{sin}\,\mathbf{z})^{\pi} = \mathbf{cos}\,\mathbf{z}$.
- $(\mathbf{cos}\,\mathbf{z})^{\pi} = \mathbf{0} \ominus \mathbf{sin}\,\mathbf{z}$.
- $(\mathbf{sinh}\,\mathbf{z})^{\pi} = \mathbf{cosh}\,\mathbf{z}$.
- $(\mathbf{cosh}\,\mathbf{z})^{\pi} = \mathbf{sinh}\,\mathbf{z}$.

Complex logarithm is multivaliued. But still we have

- $(\mathbf{Log}\,\mathbf{z})^{\pi} = \mathbf{1} \oslash \mathbf{z}$ and also $(\mathbf{log}\,\mathbf{z})^{\pi} = \mathbf{1} \oslash \mathbf{z}$.

Theorem 2. *Assume that* $\mathbf{f}^{\pi}(\mathbf{z})$ *and* $\mathbf{g}^{\pi}(\mathbf{z})$ *exist. Then*

(i) $(\mathbf{z}_0\mathbf{f})^{\pi}(\mathbf{z}) = \mathbf{f}^{\pi}(\mathbf{z})$ *for constant* $\mathbf{z}_0 \in \mathbb{B}$.

(ii) $(\mathbf{fg})^{\pi}(\mathbf{z}) = \mathbf{f}^{\pi}(\mathbf{z})\mathbf{g}^{\pi}(\mathbf{z})$.

(iii) $(\mathbf{f}/\mathbf{g})^{\pi}(\mathbf{z}) = \mathbf{f}^{\pi}(\mathbf{z})/\mathbf{g}^{\pi}(\mathbf{z})$.

(iv) $\mathbf{f}(\mathbf{g}(\mathbf{z}))^{\pi} = \mathbf{f}^{\pi}(\mathbf{g}(\mathbf{z})) \otimes \mathbf{g}^{\pi}(\mathbf{z})$, *assuming additionally that* $\mathbf{f}^{\pi}(\mathbf{g}(\mathbf{z}))$ *exists.*

Moreover, remaining theorems of complex differentiation can be stated and proved in terms of $^{\pi}$derivative.

 ISBN: 978-84-617-8694-7

6 $^{\pi}$Integral

Definition 2. Given $\mathbf{f} : B \subseteq \mathbb{B} \to \mathbb{B}$ by $\mathbf{f}(\mathbf{z}) = (R(r,\theta), \Theta(r,\theta))$ for $\mathbf{z} = (r,\theta)$, we assume that C is a contour in B with a parameterization $\mathbf{z}(t) = (r(t), \theta(t))$, $a \le t \le b$. Then we define the $^{\pi}$integral of $\mathbf{f}$ along the curve C by

$$\int_C \mathbf{f}(\mathbf{z})^{\mathbf{dz}} = \mathbf{E}^{\int_C \left(\frac{\ln R}{r}\, dr - \Theta\, d\theta\right) + i \int_C \left(\frac{\Theta}{r}\, dr + \ln R\, d\theta\right)},$$

assuming that the line integrals in the right side exist.

We present the following properties of $^{\pi}$integration.

Theorem 3. *Assume that $\int_C \mathbf{f}(\mathbf{z})^{\mathbf{dz}}$ and $\int_C \mathbf{g}(\mathbf{z})^{\mathbf{dz}}$ exist, where C is a contour in the domains of the functions $\mathbf{f}$ and $\mathbf{g}$. Then*

(i) $\int_C \mathbf{f}(\mathbf{z})^{\mathbf{dz}} = \int_{C_1} \mathbf{f}(\mathbf{z})^{dz} \int_{C_2} \mathbf{f}(\mathbf{z})^{\mathbf{dz}}$, where $C = C_1 + C_2$.

(ii) $\int_C (\mathbf{f}(\mathbf{z})\mathbf{g}(\mathbf{z}))^{\mathbf{dz}} = \int_C \mathbf{f}(\mathbf{z})^{\mathbf{dz}} \int_C \mathbf{g}(\mathbf{z})^{\mathbf{dz}}$.

(iii) $\int_C (\mathbf{f}(\mathbf{z})/\mathbf{g}(\mathbf{z})^{\mathbf{dz}} = \int_C \mathbf{f}(\mathbf{z})^{\mathbf{dz}} / \int_C \mathbf{g}(\mathbf{z})^{\mathbf{dz}}$.

Theorem 4. *Let $\mathbf{f} : B \subseteq \mathbb{B} \to \mathbb{B}$ be $^{\pi}$analytic and let C be a contour in a connected set B with the initial and end points $\mathbf{z}_1$ and $\mathbf{z}_2$, respectively. Then*

$$\int_C \mathbf{f}^{\pi}(\mathbf{z})^{\mathbf{dz}} = \frac{\mathbf{f}(\mathbf{z}_2)}{\mathbf{f}(\mathbf{z}_1)}.$$

Note that if C_1 and C_2 are two contours in the simply connected set $B \subseteq \mathbb{B}$ with the same initial and end points $\mathbf{z}_1$ and $\mathbf{z}_2$, respectively, and $\mathbf{f}$ is $^{\pi}$analytic on B, then

$$\int_{C_1} \mathbf{f}(\mathbf{z})^{\mathbf{dz}} = \int_{C_2} \mathbf{f}(\mathbf{z})^{\mathbf{dz}},$$

that means the $^{\pi}$integral is independent on the shape of the contours. Therefore, this integral can be denoted as

$$\int_{\mathbf{z}_1}^{\mathbf{z}_2} \mathbf{f}(\mathbf{z})^{\mathbf{dz}}.$$

Theorem 5. *If $\mathbf{f}$ is $^{\pi}$analytic on a connected set $B \subseteq \mathbb{B}$ and C is a closed contour in B, then*

$$\oint_C \mathbf{f}^{\pi}(\mathbf{z})^{\mathbf{dz}} = (1,0) = \mathbf{0}.$$

Theorem 6. *Let $\mathbf{f} : B \subseteq \mathbb{B} \to \mathbb{B}$ be the $^{\pi}$analog of the complex function f and let C be a contour in B. Denote $\mathbf{Log}\, C = \{\mathbf{Log}\, \mathbf{z} : \mathbf{z} \in C\}$. If $\int_{\mathbf{Log}\, C} f(z)\, dz$ exists, then $\int_C \mathbf{f}(\mathbf{z})^{\mathbf{dz}}$ exists*

$$\int_C \mathbf{f}(\mathbf{z})^{\mathbf{dz}} = \mathbf{E}^{\int_{\mathbf{Log}\, C} f(z)\, dz}. \tag{5}$$

 ISBN: 978-84-617-8694-7

Example 1. Since $(\mathbf{a}\otimes\mathbf{z})^{\pi}=\mathbf{a}$, By Theorem 4,

$$\int_{\mathbf{z}_1}^{\mathbf{z}_2}\mathbf{a}^{\mathbf{dz}}=\frac{\mathbf{a}\otimes\mathbf{z}_2}{\mathbf{a}\otimes\mathbf{z}_1}=\frac{\mathbf{E}^{\mathbf{Log\,a\,Log\,z}_2}}{\mathbf{E}^{\mathbf{Log\,a\,Log\,z}_1}}=\mathbf{E}^{\mathbf{Log\,a\,(Log\,z}_2-\mathbf{Log\,z}_1)}=\mathbf{a}\otimes(\mathbf{z}_2\ominus\mathbf{z}_2).$$

Example 2. The equality $(\mathbf{log}\,\mathbf{z})^{\pi}=\mathbf{1}\oslash\mathbf{z}$ is a complex $^{\pi}$version of $(\ln x)'=1/x$. Based on this we would like to find $^{\pi}$version of the important formula

$$\oint_{|z|=a}\frac{dz}{z}=2\pi i\ \text{ for }\ a>0,$$

where the orientation on the circle $|z|=a$ is counterclockwise. This circle has the parameterisation $z(t)=a\cos t+ia\sin t$, $-\pi\le t\le\pi$. Therefore, we will look to the closed curve C in $\mathbb{B}$ with parameterisation $\mathbf{z}(t)=(e^{a\cos t},a\sin t)$, $-\pi\le t<\pi$. Then the $^{\pi}$analog of the preceding integral is

$$\oint_C(\mathbf{1}\oslash\mathbf{z})^{\mathbf{dz}}.$$

Theorem 5 is not applicable to this integral since the function $\mathbf{f}(\mathbf{z})=\mathbf{1}\oslash\mathbf{z}$ has a singularity at $\mathbf{0}=(1,0)$ located inside C. Therefore, this $^{\pi}$integral will be calculated in accordance to Theorem 6 as follows:

$$\oint_C(\mathbf{1}\oslash\mathbf{z})^{\mathbf{dz}}=\mathbf{E}^{\oint_{\mathbf{Log}\,C}\frac{1}{z}dz}=\mathbf{E}^{\oint_{|z|=a}\frac{1}{z}dz}=\mathbf{E}^{2\pi i}=(e^0,2\pi)=(1,2\pi).$$

This example demonstrates that the $^{\pi}$integrals respond to residues.

7 Conclusion

Resuming, we see that the formulae of complex $^{\pi}$calculus are identical to those of complex calculus. Although this is verified only for its starting elements, there is no doubts that this expands to the whole. The identity is achieved by implementing the following three actions:

- Transferring the field structure of $\mathbb{C}$ to the larger set $\mathbb{B}$.
- Changing the sense of derivative and integral $^{\pi}$sense.
- Considering $^{\pi}$analogs of functions of complex calculus.

One can think that complex $^{\pi}$calculus is a small part of complex calculus on Riemann surfaces since $\mathbb{B}$ is the Riemann surface of the complex logarithm. This thought is misleading because complex $^{\pi}$calculus is based on rather easy concepts of $^{\pi}$derivative and $^{\pi}$integral and does not use the heavy machinery of integration and differentiation on manifolds. Additionally, complex $^{\pi}$calculus still considers multivalued logarithmic and power functions while they are single-valued in complex calculus on the Riemann surface of complex logarithm.

 ISBN: 978-84-617-8694-7

Since every α-calculus, including $^{\pi}$calculus as well, transfers a structure over $\mathbb{R}$ to an identical structure, it should not be surprising in the complex case as well. But unlike to the real case, complex $^{\pi}$calculus is a nontrivial transformation of complex calculus because it requires setting the new exponential function $\mathbf{E}^z$ with the range $\mathbb{B}$ being a supset of $\mathbb{C} \setminus \{0\}$ while in the real case the range of e^x is the subset $(0, \infty)$ of $\mathbb{R}$.

We think that most interesting element of complex $^{\pi}$calculus is the clear meaning of the variable (r, θ), which is the polar components if a complex variable and in the use in complex calculus as well, for example, to prove Moivre-Laplace formula establishing n different nth roots a nonzero complex number. In terms of cosmology, there are two universes—complex calculus and $^{\pi}$calculus. They behave identically, but have interaction. This can be used to obtain polar form of different formulae of complex calculus. Just replace the algebraic components x and y by polar components r and θ and transform the formula back from $^{\pi}$calculus to complex calculus. In this regard it is interesting to investigate Navier–Stokes equations. There is a believe (while not yet proved) that Navier–Stokes equations include the turbulence property of fluid flow. Roughly speaking, turbulence appears when the angular velocity of liquid is essentially nonzero. Since the derivative in θ presents the angular velocity (while in r the radial velocity) it is interesting to obtain Navier–Stokes equations in terms of r and θ.

Also, complex $^{\pi}$calculus can be included to textbooks on complex calculus in the form of exercises or research projects. It could serve as an introductory element to Riemann surfaces.

References

[1] A. E. Bashirov, E. Kurpinar and A. Özyapici, *Multiplicative calculus and its applications*, J. Math. Anal. Appl. **337** (2008) 36-48.

[2] A. E. Bashirov, E. Misirli, Y. Tandoğdu and A. Özyapici, *On modelling with multiplicative differential equations*, Appl. Math. J. Chinese Univ. Ser. B **26** (2011) 425-438.

[3] A. E. Bashirov, *On line and double multiplicative integrals*, TWMS J. Appl. Eng. Math. **3** (2013) 103-107.

[4] A. E. Bashirov and M. Riza, *On complex multiplicative differentiation*m TWMS J. Appl. Eng. Math. **1** (2011) 75–85.

[5] A. E. Bashirov and S. Norozpour, *On complex multiplicative integration* TWMS J. Appl. Eng. Math. (accepted).

[6] J. W. Brown and R. V. Churchill, *Complex variables and applications*, McGraw-Hill, New York, 2009.

 ISBN: 978-84-617-8694-7

[7] M. Grossman and R. Katz, *Non-Newtonian Calculi*, Lee Press, Pigeon Cove, 1972.

[8] M. Grossman, *Bigeometric Calculus: A System with a Scale-Free Derivative*, Archimedes Foundation, Rockport, 1983.

[9] D. Sarason, *Complex Functions Theory*, AMS, Providence, 2007.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

Probabilistic Evolution Theoretical Formulation of Anharmonic Symmetric Quantum Oscillator by Using Quantum Evolver Dynamics

Semra BAYAT ÖZDEMİR[1] and Metin DEMİRALP[1]

[1] *Computational Science and Engineering Department, Informatics Institute, Istanbul Technical University*

emails: bayat@itu.edu.tr, metin.demiralp@gmail.com

Abstract

This work focuses on the construction of evolver dynamics equations for an exponential quantum anharmonic oscillator. We also focus on the point that how squarification is used to solve the obtained second degree ordinary differential equation system. Space extension concept is used for creating the PREVTH specific system vector which is composed of a set conically closed under Poisson bracketing. A super ODE system on operators is obtained but third degree products obtained at the first attempt on the right hand sides force us to apply an improved version of space extension is again to get second degree multinomiality. A method called squarification, which is developed by our group members, helps us to get a recursive equation to solve the obtained equations for expectation values of the operators.

Key words: PREVTH, Evolver Dynamics, Space Extension, CASE, Squarification, Anharmonic Symmetric Quantum Oscillator

1 Introduction

Solution of Schrödinger equation is one of the most challenging problem in Quantum Mechanics. Several methodologies are studied for the general solution and for the special systems. But a method to meet all the needs could not have been developed. General expression for the Schrödinger equation [1] is given below.

$$i\hbar \frac{\partial \psi(\mathbf{x},t)}{\partial t} = \widehat{H}\psi(\mathbf{x},t), \quad \psi(\mathbf{x},0) = \psi_0(\mathbf{x}) \tag{1}$$

 ISBN: 978-84-617-8694-7

where $\psi(\mathbf{x},t)$ denotes the wave function and ψ_0 symbolizes the initial form of it. $\hbar$ is the reduced Planck constant. The Hamiton operator $\widehat{H}$ defines the behaviour of the quantum systems. The system studied in this work is anharmonic symmetric quantum operator which is given with the following Hamiltonian

$$\widehat{H} \equiv \frac{1}{2\mu}\widehat{p}^2 + \alpha\left(\mathrm{e}^{\frac{\kappa}{2}\widehat{q}^{\,2}} - \widehat{I}\right) \qquad \alpha,\kappa > 0 \tag{2}$$

$$\widehat{p}g(x) \equiv -i\hbar g'(x), \quad \widehat{q}g(x) \equiv xg(x), \quad \widehat{I}g(x) \equiv g(x) \qquad x \in (-\infty,\infty) \tag{3}$$

where $\widehat{I}$, $\widehat{p}$, $\widehat{q}$ denote the identity, momentum, position operators and μ, α, κ stand for the mass, potential amplitude and anharmonicity constant respectively.

Evaluation of the wave equation is painfull when working on that partial differential equation (1). But construction of ordinary differential equations (ODEs) over expectation values of the operators may facilitates the analysis [2–4]. This approach is called "Quantum Expectation Value Dynamics". Our group have numerous studies on this subject. Recently developed methodology "Probabilistic Evolution Theory (PREVTH)" [5–13] is based on the idea of constructing an infinite set of linear homogeneous ODE or ODE set over the expectation values of the basis operators. First of all, an appropriate system vector is defined for the basis operators and then its Kronecker powers are obtained for use in the solution equation.

2 System Vector of Operators for PREVTH

To define a compact equation for poisson bracket of Hamiltonian with the operators, first of all, we have to define a system vector ($\widehat{\mathbf{s}}$) composed of the necessary operators. The exponential part seen in the potential term of Hamiltonian forces us to define a new independent variable s_3 which ease to handle the singularity at infinity.

$$\widehat{\mathbf{s}} \equiv \begin{bmatrix} \widehat{s}_1 \\ \widehat{s}_2 \\ \widehat{s}_3 \end{bmatrix} \equiv \begin{bmatrix} \widehat{p} \\ \widehat{q} \\ \mathrm{e}^{\frac{\kappa}{2}\widehat{q}^2} \end{bmatrix} \tag{4}$$

Poisson brackets of Hamiltonian and system vector elements can be written as below.

$$\left\{\widehat{H},\widehat{s}_1\right\} = \left\{\widehat{H},\widehat{p}\right\} = \frac{1}{2\mu}\left\{\widehat{p}^2,\widehat{p}\right\} + \alpha\left\{\left(\mathrm{e}^{\frac{\kappa}{2}\widehat{q}^2} - 1\right),\widehat{p}\right\} = -\alpha\kappa\widehat{q}\mathrm{e}^{\frac{\kappa}{2}\widehat{q}^2} = -\alpha\kappa\widehat{s}_2\widehat{s}_3 \tag{5}$$

$$\left\{\widehat{H},\widehat{s}_2\right\} = \left\{\widehat{H},\widehat{q}\right\} = \frac{1}{2\mu}\left\{\widehat{p}^2,\widehat{q}\right\} + \alpha\left\{\left(\mathrm{e}^{\frac{\kappa}{2}\widehat{q}^2} - 1\right),\widehat{q}\right\} = \frac{1}{\mu}\widehat{p} = \frac{1}{\mu}\widehat{s}_1 \tag{6}$$

 ISBN: 978-84-617-8694-7

$$\begin{aligned}\left\{\widehat{H},\widehat{s}_3\right\} &= \left\{\widehat{H},\mathrm{e}^{\frac{\kappa}{2}\widehat{q}^2}\right\}=\frac{1}{2\mu}\left\{\widehat{p}^2,\mathrm{e}^{\frac{\kappa}{2}\widehat{q}^2}\right\}+\alpha\left\{\left(\mathrm{e}^{\frac{\kappa}{2}\widehat{q}^2}-1\right),\mathrm{e}^{\frac{\kappa}{2}\widehat{q}^2}\right\}\\ &= \frac{1}{2\mu}\left(\widehat{p}\left\{\widehat{p},\mathrm{e}^{\frac{\kappa}{2}\widehat{q}^2}\right\}+\left\{\widehat{p},\mathrm{e}^{\frac{\kappa}{2}\widehat{q}^2}\right\}\widehat{p}\right)\\ &= \frac{\kappa}{2\mu}\left(\widehat{p}\widehat{q}\mathrm{e}^{\frac{\kappa}{2}\widehat{q}^2}+\widehat{q}\mathrm{e}^{\frac{\kappa}{2}\widehat{q}^2}\widehat{p}\right)\\ &= \frac{\kappa}{2\mu}\left(\widehat{s}_1\widehat{s}_2\widehat{s}_3+\widehat{s}_2\widehat{s}_3\widehat{s}_1\right)\end{aligned} \tag{7}$$

Triple multiplications of the operators are present in (7). The degree of the right hand side polynomials can be reduced to two by space extension [14, 15]. To write a compact second degree multinomial equation from above equations, we have to define new independent variables s_4 and s_5 which are multiplications of two operators.

$$\widehat{s}_4 \equiv \widehat{s}_2\widehat{s}_3 = \widehat{s}_3\widehat{s}_2 \tag{8}$$

$$\widehat{s}_5 \equiv \widehat{s}_2\widehat{s}_1 \tag{9}$$

New augmented 1×5 system vector $\widehat{\mathbf{s}}_{aug}$ is created by adding new variables to the original system vector $\widehat{\mathbf{s}}$ (4). Its Kronecker square is 1×5^2 type vector which is labeled as $\widehat{\mathbf{s}}^{\otimes 2}_{aug}$.

$$\begin{aligned}\widehat{\mathbf{s}}_{aug} &\equiv \left[\begin{array}{ccccc}\widehat{s}_1 & \widehat{s}_2 & \widehat{s}_3 & \widehat{s}_4 & \widehat{s}_5\end{array}\right]^T\\ \widehat{\mathbf{s}}^{\otimes 2}_{aug} &\equiv [\widehat{s}_1^2 \;\; \widehat{s}_1\widehat{s}_2 \;\; \widehat{s}_1\widehat{s}_3 \;\; \widehat{s}_1\widehat{s}_4 \;\; \widehat{s}_1\widehat{s}_5 ...\\ &\quad \widehat{s}_2\widehat{s}_1 \;\; \widehat{s}_2^2 \;\; \widehat{s}_2\widehat{s}_3 \;\; \widehat{s}_2\widehat{s}_4 \;\; \widehat{s}_2\widehat{s}_5 ...\\ &\quad \widehat{s}_3\widehat{s}_1 \;\; \widehat{s}_3\widehat{s}_2 \;\; \widehat{s}_3^2 \;\; \widehat{s}_3\widehat{s}_4 \;\; \widehat{s}_3\widehat{s}_5 ...\\ &\quad \widehat{s}_4\widehat{s}_1 \;\; \widehat{s}_4\widehat{s}_2 \;\; \widehat{s}_4\widehat{s}_3 \;\; \widehat{s}_4^2 \;\; \widehat{s}_4\widehat{s}_5 ...\\ &\quad \widehat{s}_5\widehat{s}_1 \;\; \widehat{s}_5\widehat{s}_2 \;\; \widehat{s}_5\widehat{s}_3 \;\; \widehat{s}_5\widehat{s}_4 \;\; \widehat{s}_5^2]^T\end{aligned} \tag{10}$$

Poisson brackets of newly defined operators and Hamitonian can be written as

$$\begin{aligned}\left\{\widehat{H},\widehat{s}_4\right\} &= \left\{\widehat{H},\widehat{s}_2\widehat{s}_3\right\}=\left\{\widehat{H},\widehat{q}\mathrm{e}^{\frac{\kappa}{2}\widehat{q}^2}\right\}=\left\{\widehat{H},\widehat{q}\right\}\mathrm{e}^{\frac{\kappa}{2}\widehat{q}^2}+\widehat{q}\left\{\widehat{H},\mathrm{e}^{\frac{\kappa}{2}\widehat{q}^2}\right\}\\ &= \frac{1}{\mu}\widehat{p}\mathrm{e}^{\frac{\kappa}{2}\widehat{q}^2}+\widehat{q}\frac{\kappa}{2\mu}\left(\widehat{p}\widehat{q}\mathrm{e}^{\frac{\kappa}{2}\widehat{q}^2}+\widehat{q}\mathrm{e}^{\frac{\kappa}{2}\widehat{q}^2}\widehat{p}\right)\\ &= \frac{1}{\mu}\widehat{s}_1\widehat{s}_3+\frac{\kappa}{2\mu}\left(\widehat{s}_2\widehat{s}_1\widehat{s}_2\widehat{s}_3+\widehat{s}_2^2\widehat{s}_3\widehat{s}_1\right)\\ &= \frac{1}{\mu}\widehat{s}_1\widehat{s}_3+\frac{\kappa}{2\mu}\left(\widehat{s}_2\widehat{s}_1\widehat{s}_2\widehat{s}_3+\widehat{s}_2\widehat{s}_3\widehat{s}_2\widehat{s}_1\right)\end{aligned} \tag{11}$$

$$\begin{aligned}\left\{\widehat{H},\widehat{s}_5\right\} &= \left\{\widehat{H},\widehat{s}_2\widehat{s}_1\right\}=\left\{\widehat{H},\widehat{q}\widehat{p}\right\}=\left\{\widehat{H},\widehat{q}\right\}\widehat{p}+\widehat{q}\left\{\widehat{H},\widehat{p}\right\}\\ &= \frac{1}{\mu}\widehat{p}^2-\alpha\kappa\widehat{q}\widehat{q}\mathrm{e}^{\frac{\kappa}{2}\widehat{q}^2}\\ &= \frac{1}{\mu}\widehat{s}_1^2-\alpha\kappa\widehat{s}_2\widehat{s}_2\widehat{s}_3\end{aligned} \tag{12}$$

 ISBN: 978-84-617-8694-7

For convenience, all the poisson bracket results can be seen in the system form.

$$\begin{aligned}
\left\{\widehat{H},\widehat{s}_1\right\} &= -\alpha\kappa\widehat{s}_4 \\
\left\{\widehat{H},\widehat{s}_2\right\} &= \frac{1}{\mu}\widehat{s}_1 \\
\left\{\widehat{H},\widehat{s}_3\right\} &= \frac{\kappa}{2\mu}\left(\widehat{s}_1\widehat{s}_4+\widehat{s}_4\widehat{s}_1\right) \\
\left\{\widehat{H},\widehat{s}_4\right\} &= \frac{1}{\mu}\widehat{s}_1\widehat{s}_3+\frac{\kappa}{2\mu}\left(\widehat{s}_5\widehat{s}_4+\widehat{s}_4\widehat{s}_5\right) \\
\left\{\widehat{H},\widehat{s}_5\right\} &= \frac{1}{\mu}\widehat{s}_1^2-\alpha\kappa\widehat{s}_2\widehat{s}_4
\end{aligned} \tag{13}$$

To gather all the coefficients of the equations (13) in matrices, we have two define $\mathbf{H}_1$ and $\mathbf{H}_2$ with the help of unit vectors of cartesian space.

$$\begin{aligned}
\mathbf{H}_1 &\equiv \mathbf{e}_2\mathbf{e}_1^T-\alpha\kappa\mathbf{e}_1\mathbf{e}_4^T \\
\mathbf{H}_2 &\equiv \frac{\kappa}{2\mu}\mathbf{e}_3\left(\mathbf{e}_1\otimes\mathbf{e}_4+\mathbf{e}_4\otimes\mathbf{e}_1\right)^T+\frac{1}{\mu}\mathbf{e}_4\left(\mathbf{e}_1\otimes\mathbf{e}_3+\frac{\kappa}{2}\left(\mathbf{e}_4\otimes\mathbf{e}_5+\mathbf{e}_5\otimes\mathbf{e}_4\right)\right)^T \\
&\quad +\frac{1}{\mu}\mathbf{e}_5\left(\mathbf{e}_1\otimes\mathbf{e}_1-\alpha\kappa\mathbf{e}_2\otimes\mathbf{e}_4\right)^T
\end{aligned} \tag{14}$$

A compact equation can be written as below to express the equations in (13).

$$\left\{\widehat{H},\widehat{\mathbf{s}}_{aug}\right\}=\mathbf{H}_1\widehat{\mathbf{s}}_{aug}\left(t\right)+\mathbf{H}_2\widehat{\mathbf{s}}_{aug}\left(t\right)^{\otimes 2} \tag{15}$$

3 Constancy Adding Space Extension (CASE)

The expression given in the (15) is not unique. By defining a constant as a new vector element for $\widehat{\mathbf{s}}_{aug}$, the equations in (13) can be rewritten in a way that will not affect the final result. This method is called "Constancy Adding Space Extension (CASE)"which is developed in our group [16–19]. With appropriate choices, the coefficient of first term of (15) can be obtained as identity matrix multiplied by a constant. Thus, the obtained dynamic equation can be solved easily.

Let's start by defining $\widehat{s}_6$ as unit operator multiplied by nonzero constant a.

$$\widehat{s}_6\equiv a\widehat{I} \tag{16}$$

$$\left\{\widehat{H},\widehat{s}_6\right\}=0 \tag{17}$$

$$\widehat{I}=\frac{1}{a}\widehat{s}_6 \qquad \Longrightarrow \qquad \frac{1}{a^2}\widehat{s}_6^2-\frac{1}{a}\widehat{s}_6=0 \qquad \Longrightarrow \qquad \beta\widehat{s}_6-\frac{\beta}{a}\widehat{s}_6^2=0 \tag{18}$$

 ISBN: 978-84-617-8694-7

Multiplying both sides of the equation with a nonzero constant β does not change the result as it is seen in the third equation of (18).

$$\widehat{\mathbf{s}}_{ca} \equiv \left[\begin{array}{cccccc} \widehat{s}_1 & \widehat{s}_2 & \widehat{s}_3 & \widehat{s}_4 & \widehat{s}_5 & \widehat{s}_6 \end{array} \right]^T \tag{19}$$

$\widehat{\mathbf{s}}_{ca}$ is an augmented system vector which is created by CASE. We can write the Kronecker square of $\widehat{\mathbf{s}}_{ca}$ as

$$\begin{array}{rl} \widehat{\mathbf{s}}_{ca}^{\otimes 2} = & [\widehat{s}_1^2 \;\; \widehat{s}_1\widehat{s}_2 \;\; \widehat{s}_1\widehat{s}_3 \;\; \widehat{s}_1\widehat{s}_4 \;\; \widehat{s}_1\widehat{s}_5 ... \\ & \widehat{s}_2\widehat{s}_1 \;\; \widehat{s}_2^2 \;\; \widehat{s}_2\widehat{s}_3 \;\; \widehat{s}_2\widehat{s}_4 \;\; \widehat{s}_2\widehat{s}_5 ... \\ & \widehat{s}_3\widehat{s}_1 \;\; \widehat{s}_3\widehat{s}_2 \;\; \widehat{s}_3^2 \;\; \widehat{s}_3\widehat{s}_4 \;\; \widehat{s}_3\widehat{s}_5 ... \\ & \widehat{s}_4\widehat{s}_1 \;\; \widehat{s}_4\widehat{s}_2 \;\; \widehat{s}_4\widehat{s}_3 \;\; \widehat{s}_4^2 \;\; \widehat{s}_4\widehat{s}_5 ... \\ & \widehat{s}_5\widehat{s}_1 \;\; \widehat{s}_5\widehat{s}_2 \;\; \widehat{s}_5\widehat{s}_3 \;\; \widehat{s}_5\widehat{s}_4 \;\; \widehat{s}_5^2 ... \\ & a\widehat{s}_1 \;\; a\widehat{s}_2 \;\; a\widehat{s}_3 \;\; a\widehat{s}_4 \;\; a\widehat{s}_5]^T \end{array} \tag{20}$$

The Poisson bracket equations can be written as below.

$$\begin{aligned} \left\{\widehat{H}, \widehat{s}_1\right\} &= -\alpha\kappa\widehat{s}_4\frac{\widehat{s}_6}{a} + \beta\widehat{s}_1 - \frac{\beta}{a}\widehat{s}_1\widehat{s}_6 \\ \left\{\widehat{H}, \widehat{s}_2\right\} &= \frac{1}{\mu}\widehat{s}_1\frac{\widehat{s}_6}{a} + \beta\widehat{s}_2 - \frac{\beta}{a}\widehat{s}_2\widehat{s}_6 \\ \left\{\widehat{H}, \widehat{s}_3\right\} &= \frac{\kappa}{2\mu}\left(\widehat{s}_1\widehat{s}_4 + \widehat{s}_4\widehat{s}_1\right) + \beta\widehat{s}_3 - \frac{\beta}{a}\widehat{s}_3\widehat{s}_6 \\ \left\{\widehat{H}, \widehat{s}_4\right\} &= \frac{1}{\mu}\widehat{s}_1\widehat{s}_3 + \frac{\kappa}{2\mu}\left(\widehat{s}_5\widehat{s}_4 + \widehat{s}_4\widehat{s}_5\right) + \beta\widehat{s}_4 - \frac{\beta}{a}\widehat{s}_4\widehat{s}_6 \\ \left\{\widehat{H}, \widehat{s}_5\right\} &= \frac{1}{\mu}\widehat{s}_1^2 - \alpha\kappa\widehat{s}_2\widehat{s}_4 + \beta\widehat{s}_5 - \frac{\beta}{a}\widehat{s}_5\widehat{s}_6 \\ \left\{\widehat{H}, \widehat{s}_6\right\} &= \beta\widehat{s}_6 - \frac{\beta}{a}\widehat{s}_6^2 \end{aligned} \tag{21}$$

All the coefficients at the right hand side can be written in the following matrix forms.

$$\begin{aligned} \mathbf{F}_1 &\equiv \beta\mathbf{I}_6 \\ \mathbf{F}_2 &\equiv -\frac{\alpha\kappa}{a}\mathbf{e}_1\left(\mathbf{e}_4 \otimes \mathbf{e}_6\right) + \frac{1}{a\mu}\mathbf{e}_2\left(\mathbf{e}_1 \otimes \mathbf{e}_6\right) + \frac{\kappa}{2\mu}\mathbf{e}_3\left(\mathbf{e}_1 \otimes \mathbf{e}_4 + \mathbf{e}_4 \otimes \mathbf{e}_1\right)^T \\ &\quad + \frac{1}{\mu}\mathbf{e}_4\left(\mathbf{e}_1 \otimes \mathbf{e}_3 + \frac{\kappa}{2}\left(\mathbf{e}_4 \otimes \mathbf{e}_5 + \mathbf{e}_5 \otimes \mathbf{e}_4\right)\right)^T + \frac{1}{\mu}\mathbf{e}_5\left(\mathbf{e}_1 \otimes \mathbf{e}_1 - \alpha\kappa\mathbf{e}_2 \otimes \mathbf{e}_4\right)^T \\ &\quad - \frac{\beta}{a}\left(\mathbf{e}_1(\mathbf{e}_1 \otimes \mathbf{e}_6 + \mathbf{e}_4 \otimes \mathbf{e}_6) + \mathbf{e}_2(\mathbf{e}_2 \otimes \mathbf{e}_6 + \mathbf{e}_1 \otimes \mathbf{e}_6)\right. \\ &\qquad \left. + \mathbf{e}_3(\mathbf{e}_3 \otimes \mathbf{e}_6) + \mathbf{e}_4(\mathbf{e}_4 \otimes \mathbf{e}_6) + \mathbf{e}_5(\mathbf{e}_5 \otimes \mathbf{e}_6) + \mathbf{e}_6(\mathbf{e}_6 \otimes \mathbf{e}_6)\right) \end{aligned} \tag{22}$$

 ISBN: 978-84-617-8694-7

which shows that an appropriate choice of $\mathbf{F}_2$ enables us to get $\mathbf{F}_1$ as the product of 6×6 identity matrix, $(\mathbf{I}_6)$, with the scalar β.

These coefficient matrices $\mathbf{F}_1$ and $\mathbf{F}_2$ allow the equations in (21) to be written in the following compact form.

$$\left\{\widehat{H}, \widehat{\mathbf{s}}_{ca}\right\} \;=\; \mathbf{F}_1\,\widehat{\mathbf{s}}_{ca} + \mathbf{F}_2\,\widehat{\mathbf{s}}_{ca}^{\otimes 2} = \beta\,\widehat{\mathbf{s}}_{ca} + \mathbf{F}\,\widehat{\mathbf{s}}_{ca}^{\otimes 2} \tag{23}$$

Since $\mathbf{F}_1$ does not appear at the rightmost part of these equations we have also dropped the index 2 from $\mathbf{F}_2$.

4 Evolver Dynamics

(23) is not a set of ODE which is an expected issue for PREVTH. Hence we need to produce a set of ODE on which PREVTH can be applied. To this end, we can start by defining the transition operators as follows

$$\psi(\mathbf{x}, t) = \widehat{T}\,(t_0, t)\,\psi\,(\mathbf{x}, t_0) \tag{24}$$

$$\frac{d\widehat{T}}{dt}\,(t_0, t) = -\frac{i}{\hbar}\widehat{H}(t)\widehat{T}\,(t_0, t)\,, \qquad \widehat{T}\,(t_0, t_0) = \widehat{I} \tag{25}$$

$\widehat{T}\,(t_0, t)$ denotes the transition operator of a quantum system under consideration without any specification about its interaction with its environment.

We can also define an Evolver superoperator as follows.

$$\widehat{\mathbb{E}}(\widehat{o})(t) \equiv \widehat{T}\,(t, t_0)^{\dagger}\,\widehat{o}(t)\widehat{T}\,(t, t_0)\,. \tag{26}$$

$\widehat{\mathbb{E}}(\widehat{o})(t)$ maps from the operator space where $\widehat{o}(t)$ resides to the same space. Therefore it is a superoperator whose operand is denoted as its first argument. It evolves in time because of not only its temporal dependence coming from interaction operator but also from the possible dependence of the operand $\widehat{o}(t)$. All these dependencies are shown in the second but separate argument.

The equation needed for PREVTH solution can be constructed as

$$\frac{d\widehat{\mathbb{E}}(\,\widehat{\mathbf{s}}_{ca})(t)}{dt} = \beta\widehat{\mathbb{E}}(\,\widehat{\mathbf{s}}_{ca})(t) + \mathbf{F}\widehat{\mathbb{E}}(\,\widehat{\mathbf{s}}_{ca})\,(t)^{\otimes 2} \qquad\qquad \widehat{\mathbb{E}}(\,\widehat{\mathbf{s}}_{ca})\,(t_0) = \widehat{\mathbf{s}}_{ca}. \tag{27}$$

This is the desired super ODE set on operator unknowns and in PREVTH format. Hence the rest is just to get the PREVTH solution.

 ISBN: 978-84-617-8694-7

5 Constructing the PREVTH Solution

The PREVTH solution of (27) is constructed by using an integral recursion between positive integer Kronecker powers of the unknown evolver. Details are given in our relevant publications [20–23]. PREVTH solution can be explicitly given below

$$\mathbb{E}\left(\widehat{\mathbf{s}}_{ca}\right)(t) = \mathrm{e}^{\beta(t-t_0)} \sum_{j=0}^{\infty} \frac{1}{j!} \left(\frac{\mathrm{e}^{\beta(t-t_0)} - 1}{\beta} \right)^{j} \mathbf{T}_j \, \widehat{\mathbf{s}}_{ca}^{\otimes (j+1)} \tag{28}$$

where $\widehat{\mathbf{s}}_{ca}$ is the augmented initial system vector formed through CASE and telescope matrix, $\mathbf{T}_j$, is defined as

$$\mathbf{T}_j \equiv \mathbf{M}_1 ... \mathbf{M}_j, \quad \mathbf{T}_0 \equiv \mathbf{I}_n, \quad j = 0, 1, 2, ... \tag{29}$$

where the monocular matrices, $\mathbf{M}_k$s, whose cascaded forms form the telescope matrices, are also defined as

$$\mathbf{M}_k \equiv \sum_{\ell=0}^{k-1} \mathbf{I}_n^{\otimes \ell} \otimes \mathbf{F} \otimes \mathbf{I}_n^{\otimes (k-1-\ell)}, \qquad k = 1, 2, ... \tag{30}$$

Here $\mathbf{F}$ is the $\mathbf{F}_2$ matrix obtained in the previous section. It is in the form of $n \times n^2$ type matrices. We can rewrite it as the sum of Kronecker product of Cartesian unit vectors with $n \times n$ submatrices ($\mathbf{F}^{(i)}$) as below.

$$\mathbf{F} = \sum_{i=1}^{n} \mathbf{e}_i^T \otimes \mathbf{F}^{(i)} \tag{31}$$

Let us represent the squarified telescope matrix (SquTelMat) $\widehat{\mathbf{S}}_j\left(\widehat{\mathbf{s}}_{ca}\right) \widehat{\mathbf{s}}_{ca}$ through the following formula

$$\mathbf{T}_j \, \widehat{\mathbf{s}}_{ca}^{\otimes (j+1)} = \mathbf{S}_j(\widehat{\mathbf{s}}_{ca}) \, \widehat{\mathbf{s}}_{ca} \qquad j = 0, 1, 2, ... \tag{32}$$

where each SquTelMat is a square matrix of $n \times n$. For a given $\mathbf{F}$ rectangular matrix of $n \times n^2$ and an n-element vector $\mathbf{a}$, the squarification is defined as follows

$$\lfloor \mathbf{F}, \mathbf{a} \rceil \equiv \sum_{i=1}^{n} a_i \mathbf{F}^{(i)}, \tag{33}$$

where a_is are the elements of the vector $\mathbf{a}$.

As we could have been able to show quite recently, the SquTelMats satisfy the following recursion

$$\widehat{\mathbf{S}}_j\left(\widehat{\mathbf{s}}_{ca}\right) = \sum_{i=1}^{j-1} \binom{j-1}{k} \lfloor \mathbf{F}, \widehat{\mathbf{S}}_k\left(\widehat{\mathbf{s}}_{ca}\right) \widehat{\mathbf{s}}_{ca} \rceil \widehat{\mathbf{S}}_{j-1-k}\left(\widehat{\mathbf{s}}_{ca}\right), \quad \widehat{\mathbf{S}}_0\left(\widehat{\mathbf{s}}_{ca}\right) = \widehat{\mathbf{I}}, \quad j = 1, 2, ... \tag{34}$$

 ISBN: 978-84-617-8694-7

where $\widehat{\mathbf{I}}$ is the identity matrix operator whose diagonal elements are identity operator, $\widehat{I}$.

(34) is a matrix recursion and therefore its computational complexity may be desired to be further suppressed. To this end, we can define

$$\widehat{\mathbf{v}}_j\left(\widehat{\mathbf{s}}_{ca}\right) \equiv \mathbf{S}_j\left(\widehat{\mathbf{s}}_{ca}\right) \widehat{\mathbf{s}}_{ca}, \qquad j = 0, 1, 2, ... \tag{35}$$

and rewrite (34) as follows in terms of $\widehat{\mathbf{v}}_j$s instead of $\widehat{\mathbf{S}}_j$s.

$$\widehat{\mathbf{v}}_j\left(\widehat{\mathbf{s}}_{ca}\right) = \sum_{i=1}^{j-1} \binom{j-1}{k} \lfloor \mathbf{F}, \widehat{\mathbf{v}}_k\left(\widehat{\mathbf{s}}_{ca}\right) \rceil \widehat{\mathbf{v}}_{j-1-k}\left(\widehat{\mathbf{s}}_{ca}\right), \quad \widehat{\mathbf{v}}_0\left(\widehat{\mathbf{s}}_{ca}\right) = \widehat{\mathbf{s}}_{ca}, \quad j = 1, 2, ... \tag{36}$$

Now the use of (35) in (28) produces

$$\mathbb{E}\left(\widehat{\mathbf{s}}_{ca}\right)(t) = \mathrm{e}^{\beta(t-t_0)} \sum_{j=0}^{\infty} \frac{1}{j!} \left(\frac{\mathrm{e}^{\beta(t-t_0)} - 1}{\beta}\right)^j \widehat{\mathbf{v}}_j\left(\widehat{\mathbf{s}}_{ca}\right) \tag{37}$$

6 System Expectation Values at No Fluctuation Limit

The expectation value of the system vector with operator elements can be evaluated by taking the expectation values of both sides in (37) under the initial wave function. This requires the evaluation of $\widehat{\mathbf{v}}_j$'s expectation values, which can be accomplished by using the both-sides'-expectation-value in the recursion (36). However, this expectation valued recursion form is not so practical for applications. Hence we can use the no fluctuation approximation which is based on the fluctuationlessness theorem which was conjectured and proven in our group studies. Fluctuationlessness Theorem dictates us that "The expectation value of a multivariate function depending on various linear operators is equal to the image of those operators' expectation values under the same function at the no fluctuation limit" [24–30] This theorem implies that the equality obtained by taking expectation values of both sides in an equality can be approximated by replacing each argument operator with its expectation values under the same initial wave function. Thus we can write the following approximate equality by taking both sides' expectation value and then using no fluctuation approximation from (36)

$$\widehat{\mathbf{v}}_j\left(\langle \widehat{\mathbf{s}}_{ca} \rangle\right) = \sum_{i=1}^{j-1} \binom{j-1}{k} \lfloor \mathbf{F}, \widehat{\mathbf{v}}_k\left(\langle \widehat{\mathbf{s}}_{ca} \rangle\right) \rceil \widehat{\mathbf{v}}_{j-1-k}\left(\langle \widehat{\mathbf{s}}_{ca} \rangle\right), \ \widehat{\mathbf{v}}_0\left(\langle \widehat{\mathbf{s}}_{ca} \rangle\right) = \langle \widehat{\mathbf{s}}_{ca} \rangle, \ j = 1, 2, .. \tag{38}$$

On the other hand, we can write the following equality from (37) by taking expectation value of its both sides and then using no fluctuation approximation

$$\mathbb{E}\left(\langle \widehat{\mathbf{s}}_{ca} \rangle\right)(t) = \mathrm{e}^{\beta(t-t_0)} \sum_{j=0}^{\infty} \frac{1}{j!} \left(\frac{\mathrm{e}^{\beta(t-t_0)} - 1}{\beta}\right)^j \widehat{\mathbf{v}}_j\left(\langle \widehat{\mathbf{s}}_{ca} \rangle\right) \tag{39}$$

 ISBN: 978-84-617-8694-7

All these equalities convert all operator elemented system vectors to standard linear algebraic vectors depending on the ordinary linear algebraic system vector which is the expectation value of the system vector with operator elements.

7 Fluctuation Expansion and Initial Wave Function Selection

No fluctuation approximation brings the question: How can we construct a scheme which brings corrections to the non fluctuation expansion systematically? The answer can be based on the following fluctuation identities for our present case where there are six system operators:

$$\widehat{s}_i \equiv \langle\, \widehat{s}_i \,\rangle\, \widehat{I} + \widehat{\phi}_i, \qquad i = 1, 2, ..., 6 \tag{40}$$

which are fluctuation operator defining equations at the same time. These definition urge us to define the following system fluctuation operator:

$$\widehat{\boldsymbol{\phi}} \equiv \left[\begin{array}{cccc} \widehat{\phi}_1 & ... & \widehat{\phi}_6 \end{array} \right]^T \tag{41}$$

Now Mathematical Fluctuation Theory dictates us that the expectation values encountered in PREVTH can be expanded to expectation values of Kronecker power series (nonnegative Kronecker powers) of the fluctuation. The initial wave function structure to be used in the fluctuations gains great importance for the convergence of relevant Kronecker power series. To get uniform convergence the basis function for representing the initial wave function must have an appropriate weight function such that all terms' expectation values should be finite and preferably analytically calculable. This issue may need to be elaborately investigated, even though we are not going to attempt to do so here and then report the results since it is somehow out of the conceptuality which is basic goal of this presentation.

8 Concluding Remarks

In this work we have developed evolver dynamics for a quantum system of exponential anharmonic oscillator. This work is dominated by conceptuality because of its importance. Hence we do not present any illustrative implementation in this proceeding paper even though we are going to realize a much more comprehensive presentation in the conference and we are planning to write an extended paper for some journals with more satisfactory and detailed content in close future.

References

[1] P. A. M. Dirac, *The fundamental equations of quantum mechanics*, Proc. R. Soc. Lond. A **109** (1925) 642–653.

 ISBN: 978-84-617-8694-7

[2] M. Demiralp, *Determination of the quantum motion of the one dimensional harmonic oscillator via expectation value evolutions*, Bull. Tech. Univ. Istanbul **47**(4) (1994) 357.

[3] M. Demiralp, *Quantum mechanical matrix ordinary differential equations and their solutions by characteristic evolutions*, Proceedings of MMACTEE'09, Vouliagmeni, Athens, (2009) 657–662.

[4] E. Meral and M. Demiralp, *Determination of the external field amplitude and deviation parameter through expectation value based quantum optimal control of multi-harmonic oscillators under linear control agents*, J. Math. Chem. **46** (2009) 834–852.

[5] M. Demiralp, E. Demiralp and L. Hernandez-Garcia, *A probabilistic foundation for dynamical systems: theoretical background and mathematical formulation* J. Math. Chem. **50** (2012) 850–869.

[6] E. Demiralp, M. Demiralp and L. Hernandez-Garcia, *A probabilistic foundation for dynamical systems: phenomenological reasoning and principal characteristics of probabilistic evolution* J. Math. Chem. **50** (2012) 870–880.

[7] M. Demiralp, *A probabilistic evolution approach trilogy, part 1: quantum expectation value evolutions, block triangularity and conicality, truncation approximants and their convergence*, J. Math. Chem. **51**(4) (2012) 1170–1186.

[8] M. Demiralp and N. Baykara, *A probabilistic evolution approach trilogy, part 2: spectral issues for block triangular evolution matrix, singularities, space extension*, J. Math. Chem. **51**(4) (2012) 1187–1197.

[9] M. Demiralp and B. Tunga, *A probabilistic evolution approach trilogy, part 3: temporal variation of state variable expectation values from Liouville equation perspective*, J. Math. Chem. **51**(4) (2012) 1198–1210.

[10] M. Demiralp and E. Demiralp, *A contemporary linear representation theory for ordinary differential equations: probabilistic evolutions and related approximants for unidimensional autonomous systems*, J. Math. Chem. **51**(1) (2013) 58–72.

[11] M. Demiralp and E. Demiralp, *A contemporary linear representation theory for ordinary differential equations: multilinear algebra in folded arrays (folarrs) perspective and its use in multidimensional case*, J. Math. Chem. **51**(1) (2013) 38–57.

[12] M. Ayvaz and M. Demiralp, *Probabilistic evolution approach to the expectation value dynamics of quantum mechanical operators, part I: integral representation of Kronecker power series and multivariate Hausdorff moment problems*, J. Math. Chem. **52**(8) (2014) 2161–2182.

 ISBN: 978-84-617-8694-7

[13] M. Ayvaz and M. Demiralp, *Probabilistic evolution approach to the expectation value dynamics of quantum mechanical operators, part II: the use of mathematical fluctuation theory*, J. Math. Chem. **52**(8) (2014) 2294–2315.

[14] M. Demiralp and H. Rabitz, *Lie algebraic factorization of multivariable evolution operators: Definition and the solution of the canonical problem*, Int. J. Eng. Sci **31**(2) (1993) 307–331.

[15] M. Demiralp and H. Rabitz, *Lie algebraic factorization of multivariable evolution operators: Convergence theorems for the canonical case*, Int. J. Eng. Sci **31**(2) (1993) 333–346.

[16] M. Ayvaz and M. Demiralp, *Space extension strategies for probabilistic evolution approach: classical symmetric quartic anharmonic oscillator*, *Proceedings of ISTASC'13*, (2013) 81–86.

[17] B. Kalay and M. Demiralp, *Constancy added space extension for the fluctuation free expectation value dynamics of hydrogen-like quantum systems*, Proceedings of ISTASC'13, (2013) 96–100.

[18] S. Bayat and M. Demiralp, *Space extensions including constancy addition for exponentially anharmonic symmetric quantum oscillator in fluctuation free expectation value dynamics*, Proceedings of ISTASC'13, (2013) 111–116.

[19] C. Gözükirmizi and M. Demiralp, *Constancy adding space extension for ODE sets with second degree multinomial right hand side functions*, AIP Conference Proceedings **1618** (2014) 875–878.

[20] M. Demiralp, *Squarificating the telescope matrix images of initial value vector in probabilistic evolution theory (PET)*, Proceedings of AMATH'14 (2014) 99-104.

[21] C. Gözükirmizi and M. Demiralp, *Probabilistic evolution approach for the solution of explicit autonomous ordinary differential equations. Part 1: Arbitrariness and equipartition theorem in Kronecker power series*, J. Math. Chem. **52**(3) (2014) 866–880.

[22] C. Gözükirmizi and M. Demiralp, *Probabilistic evolution approach for the solution of explicit autonomous ordinary differential equations. Part 2: Kernel separability, space extension, and, series solution via telescopic matrices*, J. Math. Chem. **52**(3) (2014) 881–898.

[23] C. Gözükirmizi, M. E. Kirkin and M. Demiralp, *Probabilistic evolution theory for the solution of explicit autonomous ordinary differential equations: squarified telescope matrices*, J. Math. Chem. **55** (2017) 175–194.

 ISBN: 978-84-617-8694-7

[24] M. Demiralp, *Determination of quantum expectation values via fluctuation expansion*, Lecture Series on Computer and Computational Sciences, Selected Papers from ICCMSE 2005, Loutraki, Greece, **4A** (2005) 146–149.

[25] M. Demiralp, *A fluctuation expansion method for the evaluation of a function's expectation value*, Proceedinds of ICNAAM 2005, Rhodes, Greece, (2005) 711–714.

[26] M. Demiralp, *Fluctuationlessness theorem to approximate univariate functions' matrix representations*, WSEAS Trans. on Math., **8**(6) (2009) 258–267.

[27] N. Altay and M. Demiralp, *Numerical solution of ordinary differential equations by fluctuationlessness theorem*, J. Math. Chem. **47**(4) (2010) 1323–1344.

[28] M. Ayvaz and M. Demiralp, *Utilization of the fluctuationlessness theorem in the evaluation of certain operator matrix representations for optimally controlled simple quantum harmonic oscillator*, Proceedings of MAASE'08 (2008) 216-220.

[29] M. Ayvaz and M. Demiralp, *A fluctuation analysis at the classical limit for the dxpectation dynamics of a single quartic quantum anharmonic oscillator*, AIP Conf. Proc. **1281** (2010) 1950–1953.

[30] M. Demiralp, *No fluctuation approximation in any desired precision for univariate matrix representations*, J. Math. Chem., **47** (2010) 99-110.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

Function Approximation via Contour Integration and Tridiagonal Kernel Enhanced Multivariance Products Representation (TKEMPR), both Applied to the Remainder Term of Taylor Expansion, Expressed in Integral Form

N.A. Baykara[1] and Ercan Gürvit[2]

[1] *"Interinstitutional Group for Science and Methods of Computing" located at Informatics Institute, İstanbul Technical University*

[2] *Faculty of Sciences and Letters, Department of Mathematics, Marmara University*

emails: nabaykara@gmail.com, ercangurvit@gmail.com

Abstract

This work takes into consideration a single variable function to be approximated. To start with, Taylor expansion of the considered function is taken. The remainder term is explicitly expressed in integral form. Then making a series of transformations on the integrand, it is represented as a contour integral where the new kernel fuction is expressed in terms of two variables. This new representation allows us to apply Tridiagonal Kernel Enhanced Multivariance Products Representation (TKEMPR) method onto it and provides us with a good approximation of the kernel function which then is integrated to obtain an approximation for the initially transformed integrand. Then a final step allows us to calculate the approximation to the univariate function under consideration. Because of the overall conceptual approach to the problem through this article, applications are intended to be given during the oral representation.

Key words: Univariate Approximation, multivariate approximation, EMPR, TKEMPR, Contour integrals, Taylor Expansion
MSC 2000: 00A69, 15A12, 65D15

1 Introduction

This work consists of using the TKEMPR method which decomposes a linear integral operator on univariate functions by using high dimensional modelling with the basic idea

to use Enhanced Multivariance Products Representation (EMPR) [1-18] technique created and conjectured by Demiralp. The representation used here is not based on the general EMPR and is a specific EMPR construction for bivariate function decomposition We call this decomposition Tridiagonal Kernel Enhanced Multivariance Products Representation (TKEMPR) [19-24]. It uses EMPR bivariate function decomposition consecutively such that in each step the remainder term is expanded to again a bivariate EMPR but with different support functions. To make a function approximation using TKEMPR, first we represent our function in terms of a Taylor expansion whose remainder term is expressed in integral form. Then after making necessary manipulations the kernel of the integral in question is written in terms of a contour integration, then following a procedure which allows us to obtain the kernel to be used in TKEMPR, we obtain a two variable form and finally use it to approximate our function.

2 Basics of EMPR

Enhanced Multivariance Products Representation (EMPR) is a decomposition method recently developed by M. Demiralp. For a given multivariate function $f(x_1, ..., x_N)$ it can be written as

$$\begin{aligned} f(x_1, ..., x_N) \;=\; & f_0 \prod_{i=1}^{N} s_i(x_i) + \sum_{j=1}^{N} f_j(x_j) \prod_{i=1, i\neq j}^{N} s_i(x_i) \\ & + \sum_{\substack{j_1, j_2 = 1 \\ j_1 < j_2}}^{N} f_{j_1, j_2}(x_{j_1}, x_{j_2}) \prod_{i=1, i\neq j_1, j_2}^{N} s_i(x_i) \\ & + \cdots + f_{1,2,...,N}(x_1, x_2, ..., x_N) \end{aligned} \tag{1}$$

where f_js stand for the EMPR components ordered in ascending multivariance. The univariate functions denoted by s_i are called support functions and x_js are elements of the corresponding interval $[\, a_j, b_j \,]$

There are 2^N unknown components given in a single equation. Some constraints apply to uniquely determine EMPR components. These constraints are constructed via vanishing integrals over the EMPR components except f_0. To define these integrals univariate weight functions the product of which defines a single multivariate weight function, are used.

$$W(x_1, ..., x_N) \equiv \prod_{i=1}^{N} W_i(x_i)\,. \tag{2}$$

These univariate weight function integrals over the relevant intervals given above are set equal to 1 to facilitate further analysis; these constraints are not necessary but provide the

averaging property to the weight function.

$$\int_{a_i}^{b_i} dx_i W_i\left(x_i\right) = 1, \qquad i = 1, ..., N \tag{3}$$

In order to be able to calculate the constant term f_0, the univariate terms $f_i(x_i)$, bivariate terms $f_{i_1 i_2}\left(x_{i_1}, x_{i_2}\right)$ and other multivariate terms the normalization condition defined by

$$\int_{a_i}^{b_i} dx_i W_i(x_i) s_i(x_i)^2 = 1; \quad i = 1, ..., N \tag{4}$$

and the vanishing condition defined by

$$\int_{a_i}^{b_i} dx_i W_i\left(x_i\right) s_i\left(x_i\right) f_{i_1 ... i_k}\left(x_{i_1}, ..., x_{i_k}\right) = 0, \;\; x_i \in \left(x_{i_1}, ..., x_{i_k}\right), \; 1 \leq i \leq k \leq N \tag{5}$$

are to be used.

3 EMPR for Bivariate Functions

A new matrix decomposition method called "Tridiagonal Matrix Enhanced Multivariance Products Representation ", has been developed. The elements of a matrix being indicated by two indices which take positive integer values, independently from the other index values, this allows us to construct, so-called "Discrete Bivariate EMPR, over the matrices. Following this step we have to search for a method to apply this construction to corresponding bivariate functions. To this end we can write the following equality for a given bivariate function $f\left(x, y\right)$

$$f\left(x, y\right) = f_0 u(x) v(y) + f_1(x) v(y) + f_2(y) u(x) + f_{1,2}(x, y) \tag{6}$$

where u and v are the support functions. In this construction we will use the unit interval $\left[\, 0, 1 \,\right]$ for both independent variables, x and y without any loss of generality Beyond that the weight factors are defined as follows

$$W_1\left(x\right) \equiv 1, \;\; W_2\left(y\right) \equiv 1, \qquad x, y \in \left[\, 0, 1 \,\right]. \tag{7}$$

The support functions are assumed to have unit norms over the above intervals and under the above weights

$$\int_0^1 dx u\left(x\right)^2 = 1, \qquad \int_0^1 dy v\left(y\right)^2 = 1 \tag{8}$$

 ISBN: 978-84-617-8694-7

For this case the vanishing conditions should be taken as constraints imposed on $f_1(x)$, $f_2(y)$ and, $f_{1,2}(x,y)$. They can be written as follows

$$\int_0^1 dx f_1(x)u(x) = 0, \qquad \int_0^1 dy f_2(y)v(y) = 0 \tag{9}$$

$$\int_0^1 dx f_{1,2}(x,y)u(x) = 0, \qquad \int_0^1 dy f_{1,2}(x,y)v(y) = 0 \tag{10}$$

These equations allow us to evaluate the EMPR terms uniquely as

$$f_0 = \int_0^1 dx \int_0^1 dy f(x,y)\, u(x)\, v(y) \tag{11}$$

$$f_1(x) = \int_0^1 dy f(x,y)\, v(y) - f_0 u(x) \tag{12}$$

$$f_2(y) = \int_0^1 dx f(x,y)\, u(x) - f_0 v(y) \tag{13}$$

$$f_{1,2}(x,y) = f(x,y) - f_0 u(x)\, v(y) - f_1(x)\, v(y) - f_2(y)\, u(x). \tag{14}$$

These four equations can be used to obtain more concise expressions for EMPR terms and EMPR's itself.

EMPR is an orthogonal decomposition of the target function into two components in the space spanned by $u(x)$ and in its complementary space and another two components in the space spanned by $v(y)$ and in its complementary space. Therefore "EMPR is a decomposition projecting onto the spaces spanned by support functions and their complementary spaces.

Finally to reach (TKEMPR) construct which consists of using EMPR bivariate function decomposition consecutively such that in each step the remainder term is expanded to again a bivariate EMPR but with different support functions.

4 Contour Integration

n-th order derivative of a function can be expressed in terms of a contour integral as follows

$$f^{(n)}(x) = \frac{n!}{2\pi i} \oint_C d\xi \frac{f(\xi)}{(\xi - x)^{n+1}}, \tag{15}$$

 ISBN: 978-84-617-8694-7

where we can make a change of variable

$$\xi = x + re^{i\theta}, \qquad d\xi = ire^{i\theta}d\theta \tag{16}$$

Thus we obtain a new expression of the nth derivative of the function followed by separation of the integration interval into two

$$\begin{aligned} f^{(n)}(x) &= \frac{in!r}{2\pi i}\int_0^{2\pi} d\theta \frac{e^{i\theta} f\left(x + re^{i\theta}\right)}{r^{n+1}e^{i(n+1)\theta}} \\ &= \frac{n!}{2\pi r^n}\int_0^{2\pi} d\theta e^{-in\theta} f\left(x + re^{i\theta}\right) \\ &= \frac{n!}{2\pi r^n}\int_0^{\pi} d\theta e^{-in\theta} f\left(x + re^{i\theta}\right) + \frac{n!r}{2\pi r^n}\int_{\pi}^{2\pi} d\theta e^{-in\theta} f\left(x + re^{i\theta}\right) \end{aligned} \tag{17}$$

Making another change of variable for the second of the integrals

$$\theta \to 2\pi - \theta, \qquad d\theta \to -d\theta \tag{18}$$

and factoring out the constants and rearranging all the rest, we obtain the below expression

$$\begin{aligned} f^{(n)}(x) &= \frac{n!}{2\pi r^n}\left[\int_0^{\pi} d\theta e^{in\theta} f\left(x + re^{i\theta}\right) + \int_0^{\pi} d\theta e^{-in\theta} f\left(x + re^{-i\theta}\right)\right] \\ &= \frac{n!}{\pi r^n}\int_0^{\pi} d\theta \frac{1}{2}\left[e^{in\theta} f\left(x + re^{i\theta}\right) + e^{-in\theta} f\left(x + re^{-i\theta}\right)\right] \end{aligned} \tag{19}$$

Now, setting $\theta = \pi\theta$ gives

$$f^{(n)}(x) = \frac{n!}{r^n}\int_0^{1} d\theta \frac{1}{2}\left[e^{in\pi\theta} f\left(x + re^{i\pi\theta}\right) + e^{-in\pi\theta} f\left(x + re^{-i\pi\theta}\right)\right] \tag{20}$$

To proceed let us define the following function

$$1_f \equiv \begin{cases} 1 & 0 \le \theta \le 1 \\ 0 & other \end{cases} \tag{21}$$

Therefore same expression can be written as

$$f^{(n)}(x) = \frac{n!}{r^n}\int_0^{1} d\theta K(\theta, x) 1_f \tag{22}$$

where $K(\theta, x)$ is the bivariate kernel function necessary for us to apply TKEMPR

$$K(\theta, x) \equiv \frac{1}{2}\left[e^{in\pi\theta} f\left(x + re^{i\pi\theta}\right) + e^{-in\pi\theta} f\left(x + re^{-i\pi\theta}\right)\right] \tag{23}$$

 ISBN: 978-84-617-8694-7

5 Method

Let us consider the following identity

$$f(x) = f(a) + \int_a^x dt f'(t) \tag{24}$$

which can be rewritten as follows

$$f(x) = f(a) - \int_a^x d(x-t) f'(t) \tag{25}$$

and via integration by parts

$$f(x) = f(a) + f'(a)(x-a) + \int_a^x dt(x-t) f''(t) \tag{26}$$

Finally via repeated use of integration by parts the following expression is obtained

$$f(x) = f(a) + f'(a)(x-a) + \ldots + \frac{1}{k!} f^{(k)}(a)(x-a)^k + \frac{1}{k!} \int_a^x dt(x-t)^k f^{(k+1)}(t) \tag{27}$$

The integral at the end of the above expression can also be written as

$$\int_a^x dt(x-t)^k f^{(k+1)}(t) = \frac{(x-a)^{k+1}}{k+1} \int_0^1 dt(k+1)(1-t)^k f^{(k+1)}\left((x-a)t + a\right) \tag{28}$$

Now after constructing a bivariate form for the integrand we have to adapt it to the proper form mentioned in bivariate (EMPR)

At this stage, this is again a bivariate function which needs our attention in terms of (EMPR). This process can be repeated as many times as we need and each time we add a new step a better approximation will be obtained.

A final step after making sufficient number of iterations to approximate the bivariate function $K(\theta, x)$ is to integrate the latter over the interval $[0, 1]$ to obtain back the intended approximation of our original integrand $f^{(k+1)}(x)$

6 Conclusion

The decomposition obtained by applying (TKEMPR) to a univariate function expanded to Taylor series, after having obtained a bivariate counterpart of the integrand it via contour integration provides us with a good approximation to the bivariate kernel. Finally if we integrate this kernel over the unit interval we obtain an approximation to the original function.

 ISBN: 978-84-617-8694-7

Acknowledgements

Both authors are grateful to Prof. Metin Demiralp who had an invaluable scientific support in this present work

References

[1] Ö. F Alış and H. Rabitz: *General Foundations of High Dimensional Model Representation*, Journal of Mathematical Chemistry, 25, pp.197-233, **(1999)**.

[2] M. Demiralp : *High Dimensional Model Representation and Its Application Varieties*, The Fourth International Conference on Tools for Mathematical Modelling, St. Petersburg, Russia, June 23-28, **(2003)**.

[3] M. Demiralp and E. Demiralp : *An Orthonormal Decomposition Method for Multidimensional Matrices* in AIP Proceedings for the International Conference of Numerical Analysis and Applied Mathematics (ICNAAM 2009), vol. 1168, Rethymno, Crete, Greece, 18-22 September 2009, pp. 424427, doi:http://dx.doi.org/10.1063/1.3241487, **(2009)**.

[4] M. Demiralp and E. Demiralp : *Dimensionality Reduction and Approximation via Space Extension and Multilinear Array Decomposition* in AIP Proceedings for the International Conference of Computational Methods in Science and Engineering (ICCMSE 2009), Mini Symposium on Recent Developments in Numerical Schemes for Hilbert Space Related Issues in Science and Engineering, Rhodes, Greece, 29 September-4 October 2009, pp. 837 840, doi:http://dx.doi.org/10.1063/1.4771824, **(2009)**.

[5] M. Demiralp and E. Demiralp : *A New Straightforward Decomposition Method without Iteration to Approximate Matrices via Dominant Basis Matrices* in The Interna- tional Conference on Scientific Computing - WorldComp09 (CSC09), Las Vegas, Nevada, USA, 13-16 July 2009, pp. 7983, **(2009)**.

[6] B. Tunga and M. Demiralp : *An Iterative Scheme for Enhanced Multivariance Product Representation Method* in Proceedings for the 1st IEEEAM Conference on Applied Computer Science (ACS), Malta, pp. 247255, **(2010)**.

[7] M. Demiralp : *New Generation HDMR Based Multiway Array Decomposers: Enhanced Multivariance Products Representation (EMPR)* in Proceedings for the 1st IEEEAM Conference on Applied Computer Science (ACS), ser. ISBN: 978-960-474-225-7, Malta, pp. 1616, keynote Speech, **(2010)**.

[8] E. Demiralp and M. Demiralp : *Reductive Multilinear Array Decomposition Based Support Functions in Enhanced Multivariance Products Representation (EMPR)* in Pro-

ceedings for the 1st IEEEAM Conference on Applied Computer Science (ACS), Malta, pp. 448454, **(2010)**.

[9] C. Gözükırmızı and M. Demiralp : *Numerical Studies on the Use of Enhanced Multivariance Product Representation as a Multiway Array Decomposer* in AIP Proceedings for the International Conference of Numerical Analysis and Applied Mathematics (ICNAAM 2010), Symposium 112, Recent Developments in Hilbert Space Tools and Methodology for Scientific Computing, vol. 1281, Rhodes, Greece, pp. 19221925, doi:http://dx.doi.org/10.1063/1.3498300, **(2010)**.

[10] L. Divanyan and M. Demiralp : *Weighted Reductive Multilinear Array Decomposition* in AIP Proceedings for the 9th International Conference on Numerical Analysis and Applied Mathematics (ICNAAM2011), vol. 1389, Halkidiki, Greece, pp. 11561159, doi:http://dx.doi.org/10.1063/1.3637820, **(2011)**.

[11] S. Tuna, N. A. Baykara, and M. Demiralp : *Weighted Singular Value Decomposition for Folded Matrices* in Proceedings of the 2nd Inter- national Conference on Applied Informatics and Computing Theory (AICT11), IEEEAM, ser. ISBN: 978-1-61804-034-3, N. Mastorakis, M. Demiralp, and N. A. Baykara, Eds., Prague, Czech Republic, pp. 7075, **(2011)**.

[12] M. Demiralp : *Decomposing Functions, Arrays, Function Arrays* in Lecture Talk based on the symposium 48s preface, AIP Proceedings for the 9th International Conference on Numerical Analysis and Applied Math- ematics (ICNAAM2011), vol. 1389, Halkidiki, Greece, pp. 11381138, doi:http://dx.doi.org/10.1063/1.3637815, **(2011)**.

[13] M. Ayvaz and M. Demiralp : *Towards a New Multiway Array Decomposition Algorithm: Elementwise Multiway Array High Dimensional Model Representation (EMAHDMR)* in Proceedings of the 2nd International Conference on Applied Informatics and Computing Theory (AICT11), IEEEAM, ser. ISBN: 978-1-61804-034-3, N. Mastorakis, M. Demiralp, and N. A. Baykara, Eds., Prague, Czech Republic, pp. 7681, **(2011)**.

[14] E. K. Özay : *A New Multi-way Array Decomposition via Enhanced Multivariance Product Representation*, AIP Proceedings for the 10th International Conference of Numerical Analysis and Applied Mathematics (ICNAAM), Kos, Greece, pp. 2015-2018,Volume 1479, **(2012)**

[15] E. K. Özay and M. Demiralp : *A New Multi-way Array Decomposition*, 2012 SIAM Conference on Applied Linear Algebra, Valencia, Spain, pp. 78–78, **(2012)**

[16] E. K. Özay and M. Demiralp : *Tridiagonal Matrix Enhanced Multivariance Products Representation (TMEMPR) Studies: Decomposing the Planarly Unfolded Three-way Ar-*

rays, Proceedings of the 14th International Conference on Computational and Mathematical Methods in Science and Engineering (CMMSE), Cadiz, Spain, **(2014)**.

[17] B. Tunga and M. Demiralp : *The Influence of the Support Functions on the Quality of Enhanced Multivariance Product Representation* Journal of Mathematical Chemistry, Volume 48, Issue 3, pp 827-840, 2010.

[18] A. Okan, N.A. Baykara and M. Demiralp : *Weight Optimization in Enhanced Multivariance Product Representation (EMPR) Method* Int. Conf. on Numer. Anal. and Appl. Math., AIP Conference Proceedings, Volume 1281, pp. 1935-1938,Rhodes, Greece, 2010.

[19] A. Okan and M. Demiralp, *Tridiagonal Kernel Enhanced Multivariance Products Representation (TKEMPR) for Univariate Integral Operator Kernels*,The 2014 International Conference Mathematics and Computers in Sciences and Industry (MCSI 2014), 2014 International Conference,13-15 Sept., Varna, Bulgaria, pp. 195-200, doi: 10.1109/MCSI.2014.26, print ISBN: 978-1-4799-4744-7

[20] A. Okan and M. Demiralp, *Tridiagonal Kernel Enhanced Multivariance Products Representation (TKEMPR) for Outer Product Sums: Arrowheading EMPR for Kernel (AEMPRK)*, The 12th International Conference of Numerical Analysis and Applied Mathematics (ICNAAM 2014), 22-28 September 2014, Rhodes, Greece.

[21] A. Okan and M. Demiralp, *Arrowheading Enhanced Multivariance Products Representation for a Kernel (AEMPRK) in a Taylor Series Expansion*, 11th International Conference of Computational Methods in Sciences and Engineering, ICCMSE 2015, 20-23 March 2015, Athens, Greece, doi: 10.1063/1.4912452

[22] A. Okan and M. Demiralp, *Numerical Implementations for Tridiagonal Kernel Enhanced Multivariance Products Representation (TKEMPR) Method: Bivariate Case*, in The Proceedings of International Journal of Signal Processing, ISSN: 2367-8984, pages 102 - 107, Volume 1, 2016

[23] M. Demiralp and E. Demiralp : *A New Straightforward Decomposition Method without Iteration to Approximate Matrices via Dominant Basis Matrices* in The International Conference on Scientific Computing - WorldComp09 (CSC09), Las Vegas, Nevada, USA, 13-16 July 2009, pp. 7983, 2009.

[24] S. Tuna, and M. Demiralp : *Zero Interval Limit Perturbation Expansion for the Spectral Entities of Hilbert-Schmidt Operators Combined with Most Dominant Spectral Component Extraction: Convergence and Confirmative Implementations*, Journal of Mathematical Chemistry: 1-23., doi:10.1007/s10910-017-0740-1, 2017.

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

General Analytical Laws for Metabolic Pathways

L. Bayón[1], P. Fortuny Ayuso[1], J. M. Grau[1], M.M. Ruiz[1] and P.M. Suárez[1]

[1] *Department of Mathematics, University of Oviedo*

emails: bayon@uniovi.es, fortunypedro@uniovi.es, grau@uniovi.es, mruiz@uniovi.es, pedrosr@uniovi.es

Abstract

In this paper a general formulation for the kinetics of multi-step enzymatic reactions is presented. The optimal enzyme and metabolite concentrations are studied for the problem of minimizing the operation time in which the substrate is converted into the product. We give an analytic solution for three different kinetic models for both the unbranched and branched cases. Sufficient conditions for the optimality of the solution are studied. Several examples are presented.

Key words: Optimal Control, Kinetic models
MSC 2000: 49J30, 49M05, 80A30.

1 Introduction

The kinetics of multi-step enzymatic reactions is an ongoing research topic and, in it, the minimization of the operation time in which the substrate is transformed into the product is one of the classical problems which is being currently studied. In it, one measures the optimal profiles of both the enzyme and the metabolite. Our aim in this work is to obtain a general analytic solution for this problem. This way, we avoid the unwieldy numerical solutions, which are always tainted by the specific traits of each particular problem.

In the last years, several results have been presented on the topic. In many of them, an unbranched reaction chain of n irreversible reaction steps is studied (e.g. [1], where an explicit solution for the simplest case, $n = 2$ with linear kinetic model, is given). A mathematical model of an unbranched reaction chain with $n = 3$ and obeying the Michaelis-Menten (MM) kinetic model is used in [2]. In [3], a quasi-analytic solution is found for $n = 3$

and linear kinetics. In [4], the general case of n steps with MM kinetic model is analyzed and quantitative properties are presented, although the authors do not give an explicit analytical solution. In [5], we use a linear kinetic model for the solution of the general case of n steps. Later, in [6], we improve our results with a quasi-analytical solution for the n-steps MM model using the Lambert W-function.

Branched pathways have also been studied, certainly, but only specific cases. In [7], a network inspired in the glycolysis, with the MM model, is considered. The same example is revisited in [8] and a new pathway with two outputs is presented. Other objective functionals may also be considered, like maximizing the productivity of the metabolite [9] or the flux of a particular metabolite [10]. However, in this work we shall minimize the operation time to obtain a specified concentration of the final product.

The numerical methods for the solution of our dynamic optimization problem are usually classified into two groups: direct and indirect methods. Direct methods include complete parametrization [11], multiple shooting [12] or control vector parametrization [13]. In all of them, the basic idea is to transform the original problem into a non-linear programming problem by discretizing and approximating the control and the state variables.

On the other hand, the indirect methods solve the optimization problem using Pontryagin's Minimum Principle (PMP) taking into account the necessary optimality conditions. In this paper, the problem is stated as an Optimal Control Problem (OCP) and using PMP [14] we obtain the solution. Even more (and this is unusual in the literature), we shall also study the sufficient conditions to obtain an optimum. We also remark that we allow the possibility of using three different kinetic models in the same example. Finally, we consider not only an unbranched metabolic pathway but also a branched scheme. We obtain general, model-independent laws, for the first time.

The paper is organized as follows. In Section 2 we present the statement of the problem. The general laws of the optimal solution are obtained in Section 3, and a new kinetic model, the power law, is also presented. Section 4 contains a study on the verification of the sufficient conditions. Then we generalize the problem to the branched case with a statement valid for any graph satisfying some specific conditions. In Section 6 we present a model based on glycolysis which we aim to study numerically.

2 Statement of the problem

For the sake of simplicity we start with the simplest case of an unbranched metabolic pathway composed of n irreversible reaction steps converting substrate x_1 into product p. The value $x_1(t)$ is the substrate concentration at time t, $p(t)$ the concentration of the final product, $x_i(t), (i = 2, \ldots, n)$ the concentration of the intermediate compounds, and $u_i(t)$ $(i = 1, \ldots, n)$ the concentration of the enzyme catalyzing the i-th reaction (see Fig. 1).

 ISBN: 978-84-617-8694-7

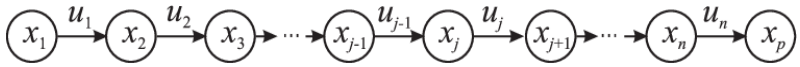


Fig. 1. Unbranched scheme.

Once we have fully studied this case, we shall perform the generalization to the branched one. The rate of the i-th reaction, $v_i(x_i(t), u_i(t))$ is linear in the enzyme concentrations, u_i:

$$v_i(x_i(t), u_i(t)) = w_i(x_i(t)) \cdot u_i(t) \tag{1}$$

The following are frequently used kinetic models:

$$\begin{aligned} w_i(x_i) &= k_i x_i && \text{(Mass action)} \\ w_i(x_i) &= \frac{k_i x_i}{K_i + x_i} && \text{(Michaelis-Menten)} \\ w_i(x_i) &= k_i x_i^c && \text{(Power law)} \end{aligned} \tag{2}$$

The dynamical model for the pathway is given by the law of conservation of mass:

$$\dot{x}_i(t) = v_{i-1}(x_{i-1}(t), u_{i-1}(t)) - v_i(x_i(t), u_i(t)); \ (i = 1, \ldots, n). \tag{3}$$

The objective is to transform x_1 into p as fast as possible; we denote t_f the final time. We assume an exhaustible initial substrate, x_1, and imposing $p(t_f) = C_f$ $(0 < C_f < 1)$, we obtain:

$$x_1(t_f) + x_2(t_f) + \ldots + x_n(t_f) = 1 - C_f \tag{4}$$

so that the optimization problem may thus be stated as the control problem (Pr):

$$\begin{aligned} &\text{(Pr):} && \tau_{C_f} = \min_{u_1, \ldots u_n} \int_0^{t_f} dt = \min_{u_1, \ldots u_n} t_f \\ &\text{subject to:} && (3), (4) \\ & && u_1 \geq 0, \ldots u_n \geq 0; \ u_1 + \ldots + u_n \leq 1 \end{aligned} \tag{5}$$

3 Optimal Solution

In two previous papers, we used PMP to obtain the solution to (Pr) for the mass action model [5] and for the MM model [6]. When the control appears linearly, as is the case for the problem under consideration, the control switches between its upper and lower bounds at discrete instants: the optimal control is said to be a *bang-bang type control* and those instants are called the *switching times*. The general form of the solution can be described

as follows: there exist n switching times, as many as enzymes, so that the optimal i-enzyme profile is proved to be of bang-band type and satisfies:

$$u_i(t) = \begin{cases} 1 & \text{for} \quad t \in [t_{i-1}, t_i) \\ 0 & \text{for} \quad t \notin [t_{i-1}, t_i) \end{cases} \; ; \; i = 1, \ldots, n \tag{6}$$

where $\{t_0, t_1, t_2, ..., t_n\}$ are the switching times, with $t_0 = 0$ and $t_n = t_f$.

We shall denote by $x_{ji}(t)$ (for $i, j = 1, \ldots, n$) the optimal j-th metabolite concentration, in the i-th interval $[t_{i-1}, t_i]$. The optimal solution of the complete system can be described on each interval, knowing that on the i-th interval, $[t_{i-1}, t_i]$ (for $i = 2, \ldots, n-1$), there are 4 laws governing the metabolite concentrations:

(a) Metabolites before the i-th remain at a constant value given by:

$$x_{ji}(t) = x_{jj}(t_j) \text{ for } j = 1, \ldots, i-1 \tag{7}$$

(b) The i-th metabolite follows a law given by a function depending on: the parameters of the model, the previous switching time, and the value of the i-th metabolite on the previous interval:

$$x_{ii}(t) = f(x_{ii-1}(t_{i-1}), t_{i-1}, t) \tag{8}$$

(c) The $i+1$-th metabolite follows a law obtained from the one of the previous metabolite as follows:

$$x_{i+1i}(t) = x_{ii-1}(t_{i-1}) - x_{ii}(t) \tag{9}$$

(d) Metabolites from the $i+2$-th on have not been activated yet, so that their value is zero:

$$x_{ji}(t) = 0 \text{ for } j = i+2, \ldots, n \tag{10}$$

A schematic idea is shown in Fig. 2.

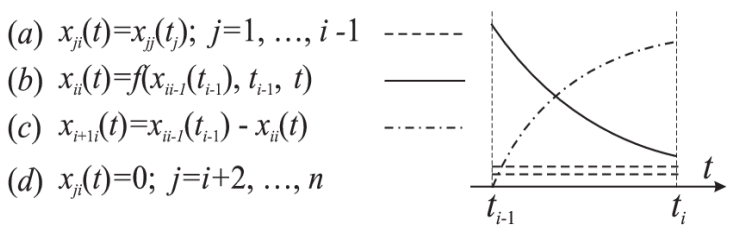


Fig. 2. Optimal concentration laws for the metabolites.

On the first interval ($i = 1$), letting $x_{10}(t_0) = 1$, only Laws (b), (c) and (d) apply; on the last-but-one ($i = n-1$), only Laws (a), (b) and (c) apply; whereas on the last one ($i = n$), only Laws (a) and (b).

 ISBN: 978-84-617-8694-7

Notice that the formulas above are general and they only depend on the kinetic model (Law (b)). In [5] we obtained the law for the mass action model (i.e. the linear one), getting:

$$x_{ii}(t) = f(x_{ii-1}(t_{i-1}), t_{i-1}, t) = x_{ii-1}(t_{i-1}) \exp(-k_i(t - t_{i-1})) \tag{11}$$

Elsewhere, in [6], we obtained the law for the Michaelis-Menten model, which gives:

$$\begin{aligned} x_{ii}(t) &= f(x_{ii-1}(t_{i-1}), t_{i-1}, t) & (12)\\ &= K_{mi} W\left(\frac{x_{ii-1}(t_{i-1})}{K_{mi}} \exp\left(\frac{x_{ii-1}(t_{i-1})}{K_{mi}}\right) \exp\left(-\frac{k_i}{K_{mi}}(t - t_{i-1})\right)\right) & (13)\end{aligned}$$

where W is the Lambert W-function.

We present, for the first time, in this paper, the expression of Law (b) for the kinetic model given by the power law. The expression is now:

$$\begin{aligned} x_{ii}(t) &= f(x_{ii-1}(t_{i-1}), t_{i-1}, t) & (14)\\ &= \left[(x_{ii-1}(t_{i-1}))^{1-c} - (1-c)k_i(t - t_{i-1})\right]^{(1-c)^{-1}} & (15)\end{aligned}$$

The method for computing this optimal solution is analogous to the one we shown in our previous papers, and we refer the reader to them (mainly Appendix 1 in [6], where it is given in detail).

The idea is to define the Hamiltonian $H(x_1, \ldots, x_n, u_1, \ldots, u_n, \lambda_1 \ldots, \lambda_n, t)$ associated to the problem (Pr):

$$\begin{aligned} H &= 1 + \sum_{i=1}^{n} \lambda_i \dot{x}_i(t) & (16)\\ &= 1 + \sum_{i=1}^{n} \lambda_i \left[w_{i-1}(x_{i-1}(t)) \cdot u_{i-1}(t) - w_i(x_i(t)) \cdot u_i(t)\right] & (17)\end{aligned}$$

and compute the optimum values for x_i and u_i applying the necessary conditions in PMP on that Hamiltonian. In this case:

$$\begin{aligned} &\text{(i)} \quad \dot{\lambda}_i(t) = -\frac{\partial H}{\partial x_i}; \lambda_i(t_f) = 0,\ i = 1, ..., n\\ &\text{(ii)} \quad \min_{u_1, \ldots u_n} H\\ &\text{(iii)} \quad \dot{x}_i(t) = w_{i-1}(x_{i-1}(t)) \cdot u_{i-1}(t) - w_i(x_i(t)) \cdot u_i(t);\ x_i(0) = x_{i0},\ i = 1, ..., n\\ &\text{(iv)} \quad H(x_1, \ldots, x_n, u_1, \ldots, u_n, , \lambda_1 \ldots, \lambda_n, t_f) = 0\end{aligned} \tag{18}$$

Our system is autonomous, so that $H_t \equiv 0 \Rightarrow H(t) = cte$. This condition together with (iv) implies that $H(t) = 0$.

 ISBN: 978-84-617-8694-7

We obtain the optimal solution constructively by intervals, starting at $t = 0$ and concatenating the results. Once these values are computed, one still needs to calculate the switching times $t_1, t_2, \ldots, t_{n-1}$ and the operation time t_f. To this end, we use the restriction (4) and define the augmented functional:

$$L(t_1, t_2, \ldots, t_{n-1}, t_f, \beta) = t_f + \beta(x_{1n}(t_f) + x_{2n}(t_f) + \cdots + x_{nn}(t_f) - C_f) \tag{19}$$

where the values of the concentrations $x_{1n}(t_f) = x_{1n}(t_1)$, $x_{2n}(t_f) = x_{2n}(t_2)$, $\ldots, x_{nn}(t_f)$ are given, and where the unknowns $t_1, t_2, \ldots, t_{n-1}$ and t_f appear. Then we solve the non-linear system:

$$\frac{\partial L}{\partial t_1} = 0; \frac{\partial L}{\partial t_2} = 0; \ldots; \frac{\partial L}{\partial t_{n-1}} = 0; \frac{\partial L}{\partial t_f} = 0; \frac{\partial L}{\partial \beta} = 0 \tag{20}$$

Once the optimal values of the switching times $t_1, t_2, \ldots, t_{n-1}$ and t_f are obtained numerically, the remaining values of the solution are immediately obtained analytically using the closed-form formulas (a) to (d), and the problem is completely solved.

Notice that, remarkably, the laws given above allow the simultaneous consideration of several models when studying the pathway: one just needs to use the appropriate function $f(x_{ii-1}(t_{i-1}), t_{i-1}, t)$ in Law (b).

4 Sufficient Conditions

Considering the optimal control problem, with $x(t)$ and $u(t)$ denoting n-dimensional vectors:

$$\min_{u(t)} J = \int_0^T F(x(t), u(t), t)dt + B[T, x(T)] \tag{21}$$

$$\dot{x}(t) = f(x(t), u(t), t);\ x(0) = x_0 \tag{22}$$

$$u(t) \in U(t),\ 0 \leq t \leq T \tag{23}$$

we can guarantee the sufficient conditions for the existence of an optimal solution using Arrow's Theorem [14] (this sufficiency is not considered in the previous works [4], [5], [6]).

In our case:

$$\min_{u(t)} J = \min_{u_1, \ldots u_n} \int_0^{t_f} dt \tag{24}$$

$$\dot{x}_i(t) = w_{i-1}(x_{i-1}(t)) \cdot u_{i-1}(t) - w_i(x_i(t)) \cdot u_i(t), \quad i = 1, \ldots, n \tag{25}$$

$$u_1 \geq 0, \ldots, u_n \geq 0;\ u_1 + \ldots + u_n \leq 1 \tag{26}$$

Upon minimizing $H(x, u, \lambda, t)$ in $u \in U(t)$ one obtains a function $u = u^0(x, \lambda, t)$ from which $H^0(x, u^0(x, \lambda, t), \lambda, t)$ can be computed. In each case, the equations of the model (Mass action, Michaelis-Menten and Power law) provide the sufficient conditions in one way or another.

 ISBN: 978-84-617-8694-7

5 Generalization

We present a generalization for branched pathways. The cases we cover are pathways whose graph of *temporal dependencies* satisfies a specific property (which essentially says that the i-th metabolite follows temporally the j-th one if and only if vertex j is joined with vertex i by a directed path (see [15] for the elementary notions).

These graphs represent the fact that the synthesis of one enzyme requires the degradation of *all the previous ones*; this is the case, for example, of the glycolysis, see [8]. Following that paper, we assume that the optimal profile follows a pattern matching the topology of the pathway, reflecting the fact that the enzymes are activated sequentially. Notice that the type of graphs we deal with have a strictly upper-triangular adjacency matrix (and more conditions, but this one is easy to verify).

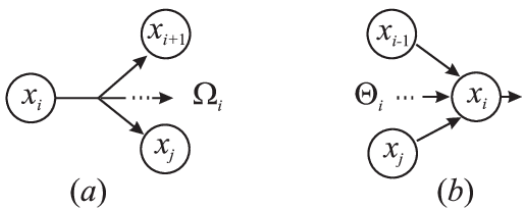


Fig. 3. Branched pathways.

With the same notation as above, the optimal solution of these branched systems can be described on the i-th interval $[t_{i-1}, t_i]$ ($i = 2, \ldots, n-1$) with these 4 laws:

- (a') For the metabolites before the $i-$th one:

$$x_{ji}(t) = x_{jj}(t_j) \text{ for } j = 1, \ldots, i-1 \tag{27}$$

- (b') For the i-th metabolite:

$$x_{ii}(t) = f(x_{ii-1}(t_{i-1}), t_{i-1}, t) \tag{28}$$

- (c') All the metabolites $j \in \Omega_i$ follow the same law:

$$x_{ji}(t) = x_{ji-1}(t_{i-1}) + x_{ii-1}(t_{i-1}) - x_{ii}(t) \tag{29}$$

- (d') The metabolites j-th such that $i < \min \Theta_j$, have not been activated yet:

$$x_{ji}(t) = 0 \tag{30}$$

The only differences with the unbranched system appear in (c') and (d').

 ISBN: 978-84-617-8694-7

6 Numerical Examples

We provide several simulations which illustrate the general formulations above. We shall use a test example already studied by several authors and inspired by the upper part of glycolysis. The original problem was stated by Bartl et al. in [7] and was also considered in [8] with a new formulation incorporating the enzyme dynamics.

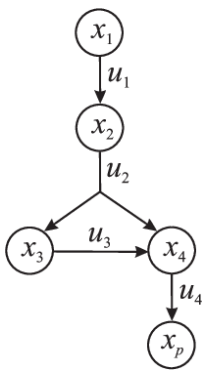


Fig. 4. Glycolysis inspired network.

The pathway (Fig. 4) consists of four enzymatic reactions with one branch. Recall that x_1 corresponds to the substrate, x_2, x_3 and x_4 are the intermediate metabolites and x_p represents the product. The alternative route of the glycolysis is represented by u_2 (corresponding to the enzyme aldolase) metabolizing the intermediate x_2 to x_3 and x_4.

The single aim is to minimize the time needed to transform the substrate x_1 into a fixed amount (90%) of product x_p, (i.e. $C_f = 0.9$). We assume unbuffered or exhaustible substrate x_1 (i.e. the substrate is consumed during the process) and enzymes are assumed to become activated instantaneously (just-in-time activation). The following initial conditions at $t = 0$: $x_0 = [1, 0, 0, 0, 0]^T$; $u_0 = [0, 0, 0, 0]^T$ are imposed. Metabolites and enzymes are expressed in concentration units and time in seconds.

Additionally, restrictions on enzyme concentrations and their total amount are introduced in a normalized form as:

$$u_1 \geq 0, \ldots u_n \geq 0; \; u_1 + \ldots + u_n \leq 1 \tag{31}$$

This is in agreement with the assumption that the cell can only allocate a certain amount of protein to a pathway, and with experimental observations [16] in Escherichia coli.

The classical theoretical studies ([7], [8]) are based on the Michaelis–Menten kinetics:

$$w_i(x_i) = \frac{k_i x_i}{K_i + x_i} \tag{32}$$

with unity rate constants for k_i and K_i: $k_i = 1(s^{-1})$, $K_i = 1(mM)$. However, we also give results for the other two laws we have studied: mass action and the power law.

 ISBN: 978-84-617-8694-7

References

[1] E. Klipp, R. Heinrich, H.G. Holzhutter, *Prediction of temporal gene expression. Metabolic optimization by re-distribution of enzyme activities*, Eur. J. Biochem. **269 (22)** (2002) 5406-5413.

[2] A. Zaslaver, A. Mayo, R. Rosenberg, P. Bashkin, H. Sberro, M. Tsalyuk, M. Surette, U. Alon, *Just-in-time transcription program in metabolic pathways*, Nat. Genet. **36(5)** (2004) 486-491.

[3] M. Bartl, P. Li, S. Schuster, *Modelling the optimal timing in metabolic pathway activation-Use of Pontryagin's Maximum Principle and role of the Golden section*, BioSystems **101** (2010) 67-77.

[4] D. Oyarzun, B. Ingalls, R. Middleton, D. Kalamatianos, *Sequential activation of metabolic pathways: a dynamic optimization approach*, Bull. Math. Biol. **71(8)** (2009) 1851-1872.

[5] L. Bayon, J.A. Otero, M.M. Ruiz, P.M. Suarez, C. Tasis, *Sensitivity analysis of a linear unbranched chemical process with n steps*, J. Math. Chem. **53**(**3**) (2015) 925-940.

[6] L. Bayon, J.A. Otero, P.M. Suarez, C. Tasis, *Solving linear unbranched pathways with Michaelis-Menten kinetics using the Lambert W-Function*, J. Math. Chem. **54(7)** (2016) 1351-1369 .

[7] M. Bartl, M. Kotzing, C. Kaleta, S. Schuster, P. Li, *Just-in-time activation of a glycolysis inspired metabolic network - solution with a dynamic optimization approach*, Proc. 55nd Int. Sci. Colloq. (2010) 217-222.

[8] G. de Hijas-Liste, E. Klipp, E. Balsa-Canto, J. Banga, *Global dynamic optimization approach to predict activation in metabolic pathways*, BMC Syst. Biol. **8(1)** (2014) 1-15.

[9] A.F. Villaverde, S. Bongard, K. Mauch, E. Balsa-Canto, J.R. Banga, *Metabolic engineering with multi-objective optimization of kinetic models*, J. Biotech. **222** (2016) 1-8.

[10] G. Xu, L. Wang, *An improved geometric programming approach for optimization of biochemical systems*, J.Appl. Math. (2014) 1-10.

[11] L.T. Biegler, A.M. Cervantes, A. Watcher, *Advances in simultaneous strategies for dynamic process optimization*, Chem. Eng. Sci. **57(4)** (2002) 575-593.

 ISBN: 978-84-617-8694-7

[12] H.G. Bock, K.J. Plitt, *A multiple shooting algorithm for direct solution of optimal control problems*, In Proceedings 9th IFAC World Congress, New York: Pergamon Press (1984) 242-247.

[13] V.S. Vassiliadis, R.W.H. Sargent, C.C. Pantelides, *Solution of a class of multistage dynamic optimization problems. 1. Problems without path constraints*, Ind. Eng. Chem. Res. **33(9)** (1994) 2111-2122.

[14] A. Chiang, *Elements of Dynamic Optimization*, Waveland Press, 2000.

[15] R. Gould, *Graph Theory*, Dover Publications, 2012.

[16] K.M. Slade, R. Baker, M. Chua, N.L. Thompson, G.J. Pielak, *Effects of recombinant protein expression on green fluorescent protein diffusion in Escherichia coli*, Biochemistry, **48(23)** (2009) 5083-5089.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

A Leslie-Gower type predation model considering double Allee effect on prey and a sigmoid functional response

Ruth Becerra-Klix[1] **and Eduardo González-Olivares**[2]

[1] *Facultad de Ingeniera y Ciencias, Universidad Adolfo Ibañez, Chile.*

[2] *Pontificia Universidad Católica de Valparaíso, Chile,*

emails: ruth.becerra@uai.cl, ejgonzal@ucv.cl

Abstract

This work deals with a modified continuous time Leslie-Gower type predator-prey model, assuming the functional response is of sigmoid type and prey is affected by a double Allee effect. The main characteristic of this type of models is that the growth predator equation is a modification of the equation logistic, considering the environmental carrying capacity of predators proportional to the amount of available prey.

It may be observed the modified system has different dynamics which appear according to the parameter values; the existence of one, two or three equilibrium points can be proven; also, the existence of a limit cycle surrounding a unique positive equilibrium point may exist.

The existence of separatrix curves on the phase plane dividing the behavior of the trajectories is also demonstrated, which imply that two solutions very closed can have different $\omega - limit$; as consequence, solutions near of these separatrix curves, are highly sensitive to initial conditions.

Key words: Predator–prey model, functional response, bifurcation, limit cycle, separatrix curve, stability.

MSC 2000: AMS codes 92D25; 34C23; 58F14; 58F21

1 Introduction

In this work a modified Leslie-Gower predator-prey model is analyzed, which is described by an autonomous differential equation system, considering the following aspects:

i) the prey population is affected by a double Allee effect [11, 30],

ii) the functional response or predator consumption rate is sigmoid [27, 33], and

 ISBN: 978-84-617-8694-7

iii) the equation for predator growth is the logistic type. [1, 28, 33].

In this type of model, the conventional environmental carrying capacity for predators K_y is a function of the available prey quantity [2, 16, 25], in which K_y is proportional to prey abundance $x = x(t)$, that is, $K_y = K(x) = nx$. Denoting by $y = y(t)$ the predator population size in the logistic predator model, the quotient $\frac{y}{nx}$ is called the Leslie-Gower term. It measures the loss in the predator population due to scarce availability (per capita $\frac{y}{x}$) of its favorite food.

The formulation of the Leslie-Gower type predator-prey model is based on the assumption that reduction in a predator population has a reciprocal relationship with per capita availability of its preferred food. The importance of this model is highlighted by Collings [10] assuring that it provides a way to avoid the biological control paradox.

The Leslie-Gower type model may present anomalies in its predictions, as it can deduce that even in very low prey population density, when the consumption rate per predator is almost zero, predator population might increase, if the predator/prey ratio is very small [33]. This type of model has been used in different recent articles as [1, 17, 22, 23].

A important factor in the predation model is the predator functional response or consumption function, which refers to the change in attacked prey density per unit of time per predator when the prey population size changes [27]. In many predator-prey models it is assumed that the functional response grows monotonic, being the inherent assumption the more prey in the environment, the better for the predator [33].

We will consider that the predator consumption function is expressed by the sigmoid function $h(x) = \frac{q\ x^2}{x^2 + a^2}$ [5, 32], corresponding to the Holling type III [27]. The parameter a is a measure of the abruptness on the functional response [15]. If $a \to 0$, the curve grows quickly, while if $a \to K$, the curve grows slowly, i.e., a bigger amount of prey is required to attain $\frac{q}{2}$ [19].

On the other hand, Allee effects occur whenever fitness of an individual in a small or sparse population decreases as the population size or density also declines.

Various mechanisms generate Allee effects, for example: mate finding, reproductive facilitation, cooperative antipredator behaviour, predator dilution, etc. It can also be represented by many mathematical expressions [7, 16].

In [6, 12] it is suggested two or more Allee effects act simultaneously in the same population; the combined influence of this phenomenon has been named *multiple Allee effect* [6]. In [7] is proposed the equation

$$\frac{dx}{dt} = r\left(1 - \frac{x}{K}\right)\left(1 - \frac{m+b}{x+b}\right)x,$$

Here, if $m > 0$ it has a *strong Allee effect* and if $m = 0$ we have a special case of *weak Allee effect* [8, 34]. Algebraically, this equation can be rewritten as:

$$\frac{dx}{dt} = \frac{rx}{x+b}\left(1 - \frac{x}{K}\right)(x - m),$$

which represents a double Allee effect in prey [3, 20]. since the factor $r(x) = \frac{rx}{x+b}$ indicates the impact of a second Allee effect exerted by the non-fertile population, represented

 ISBN: 978-84-617-8694-7

by the parameter b.

In order to establish the local stability of the equilibrium points, the well-known results on the phase portraits of the planar systems near equilibria are utilized, determining the signs of the determinants and trace of the Jacobian matrix at the equilibria [5, 9, 24].

2 THE MODEL

The model to be analyzed is described by the following autonomous bidimensional differential equation system of Kolmogorov type [14]

$$X_\mu : \left\{ \begin{array}{rcl} \frac{dx}{dt} & = & \left(r\left(1-\frac{x}{K}\right)\left(1-\frac{m+b}{x+b}\right) - \frac{qxy}{x^2+a^2} \right) x \\ \frac{dy}{dt} & = & s\left(1-\frac{y}{nx}\right) y \end{array} \right. , \tag{1}$$

where $x = x(t)$ and $y = y(t)$ indicate the prey and predator population sizes respectively for $t \geq 0$, measure as the number of individuals, density or biomass. Clearly, system (1) can be rewritten as

$$X_\mu : \left\{ \begin{array}{rcl} \frac{dx}{dt} & = & \left(\frac{r}{x+b}\left(1-\frac{x}{K}\right)(x-m) - \frac{qxy}{x^2+a^2} \right) x \\ \frac{dy}{dt} & = & s\left(1-\frac{y}{nx}\right) y \end{array} \right. , \tag{2}$$

The parameters are all positive, i.e., $\mu = (r, K, q, a, s, n, b, m) \in \mathbb{R}^7_+ \times]-K, K[$ and for biological reasons $a, b < K$, having the following meanings:

r is the intrinsic prey growth rate or biotic potential,

K is the prey environmental carrying capacity,

b is the fraction of sterile population members,

m is the strong Allee effect threshold or minimum of viable prey population,

q is the maximum number of prey that can be eaten by a predator at each time unit (satiation rate),

a is the amount of prey to achieve one-half of the maximum rate q (the half saturation rate),

s is the intrinsic predator growth rate,

n is a measure of the food quality indicating how the predators turn eaten prey into new predator births.

The analysis must be made separately for $m > 0$, $m = 0$ and $m < 0$, because the properties of the system can change with respect to this parameter; however, due to the algebraic difficulties, we analyzed the model when $m = 0$ and partially when $m > 0$.

Clearly in system (1), the predator carrying capacity is $K(x) = nx$. It is interesting to note that the system is not defined at $x = 0$, but later will show that the point $(0, 0)$ has a strong influence on the dynamics of the system.

System (1) or vector field X_μ is defined at

$\Omega = \left\{(x,y) \in \mathbb{R}^2 / \ x > 0, \ y \geq 0\right\} = \mathbb{R}^+ \times \mathbb{R}_0^+$.

The equilibrium points are $(0,0)$, $(m,0)$ $(K,0)$ and (x_e, y_e) satisfying the equation of the isoclines $y = nx$ and $y = \frac{r}{qx(x+b)}\left(1-\frac{x}{K}\right)(x-m)\left(x^2+a^2\right)$. The point (x_e, y_e) can be a positive equilibrium point (equilibrium at interior of the first quadrant) or cannot exist there. So, the abscise x_e satisfies the equation

$$P(x) = x^4 - (K+m-Knq)x^3 + (a^2+Km+Kbnq)x^2 - a^2(K+m)x + Ka^2m = 0$$

In order to simplify the calculus we follow the methodology used in [18, 21], making a change of variable and a time rescaling; we have

Proposition 1 *System* (1) *is topologically equivalent to*

$$U_\eta : \left\{ \begin{array}{rcl} \frac{du}{d\tau} = & \left((1-u)(u-M)\left(u^2 + A^2\right) - Quv(u+B)\right)u^2 \\ \frac{dv}{d\tau} = & S(u-v)(u+B)\left(u^2 + A^2\right)v \end{array} \right. , \tag{3}$$

defined in $\bar{\Omega} = \{(u,v) \in \mathbb{R}^2 / \ u \geq 0, \ v \geq 0\}$, *where* $\eta = (A,B,Q,S,M) \in \Delta =]0,1[^2 \times \mathbb{R}_+^2 \times]-1,1[$, *with* $A = \frac{a}{K}$, $B = \frac{b}{K}$, $Q = \frac{nq}{r}$, $S = \frac{s}{r}$ *and* $M = \frac{m}{K}$.

Proof. Using the change of variables given by $x = Ku$ and $y = nKv$, replacing in (1), we have

$$V_\mu : \left\{ \begin{array}{rcl} \frac{du}{dt} & = & \left(r(1-u)\left(1-\frac{m+b}{Ku+b}\right) - \frac{q(Ku)(nKv)}{(Ku)^2+a^2}\right)u \\ \frac{dv}{dt} & = & s\left(1-\frac{nKv}{nKu}\right)v \, . \end{array} \right.$$

After simplification and factoring, the new vector field is

$$V_\mu : \left\{ \begin{array}{rcl} \frac{du}{dt} & = & r\left((1-u)\left(1-\frac{\frac{m}{K}+\frac{b}{K}}{u+\frac{b}{K}}\right) - \frac{qn}{r}\frac{uv}{u^2+\left(\frac{a}{K}\right)^2}\right)u \\ \frac{dv}{dt} & = & s\left(1-\frac{v}{u}\right)v \, . \end{array} \right.$$

By means of the time rescaling given by $t = \frac{1}{r}\left(u+\frac{b}{K}\right)u\left(u^2+\left(\frac{a}{K}\right)^2\right)\tau$ and by using the chain rule, it follows

$$Y_\mu : \left\{ \begin{array}{rcl} \frac{du}{dt} & = & \left((1-u)\left(u-\frac{m}{K}\right)\left(u^2+\left(\frac{a}{K}\right)^2\right) - \frac{qn}{r}uv\left(u+\frac{b}{K}\right)\right)u \\ \frac{dv}{dt} & = & \frac{s}{r}(u-v)\left(u+\frac{b}{K}\right)\left(u^2+\left(\frac{a}{K}\right)^2\right)v \end{array} \right.$$

Finally, replacing the new parameters A, B, Q, S, M, we obtain system (2).

Therefore, we have constructed the diffeomorphism [9] $\varphi : \bar{\Omega} \times \mathbb{R} \longrightarrow \Omega \times \mathbb{R}$,

so that

$\varphi(u,v,\tau) = \left(Ku, nKv, \frac{1}{r}\left(u+\frac{b}{K}\right)u\left(u^2+\left(\frac{a}{K}\right)^2\right)\tau\right) = (x,y,t)$

and we have that

$\det D\varphi(u,v,\tau) = \frac{Kn}{r}\left(u+\frac{b}{K}\right)u\left(u^2+\left(\frac{a}{K}\right)^2\right) > 0.$

Thus, φ is a diffeomorphism preserving the time orientation; thus, the vector field X_μ, is topologically equivalent to the vector field $U_\eta = \varphi \circ X_\mu$ with $U_\eta = P(u,v)\frac{\partial}{\partial u} +$

 ISBN: 978-84-617-8694-7

$Q(u,v)\ \frac{\partial}{\partial v}$and the associated differential equation system is given by the polynomial system of fourth degree. ■

The equilibrium points of system (3) or singularities of vector field U_η are $(0,0)$, $(M,0)$, $(1,0)$, and the points lie in the intersection of the isoclinic curves

$v = \frac{(1-u)(u-M)\left(u^2\ +\ A^2\right)}{Qu(u+B)}$ and $v = u$.

Then, the abscissa u is solution of the fourth degree equation:

$$p(u) = u^4 - (1+M-Q)u^3 + (A^2+M+BQ)u^2 - A^2(M+1)u + A^2M = 0 \qquad (4)$$

This equation can have up to four positive roots and different cases must be studied. Due to this algebraic complexities and because the dynamics of model depend on the parameter M, we will focused our study for the case $M = 0$.

The local stability of equilibrium points is determined by the Jacobian or community matrix [4] given by

$$DU_\eta\,(u,v) = \begin{pmatrix} DU_\eta\,(u,v)_{11} & -Q\,(u+B)\,u^3 \\ DU_\eta\,(u,v)_{21} & S\,(u-2v)\left(A^2+u^2\right)(u+B) \end{pmatrix}$$

where

$DU_\eta\,(u,v)_{11} = -u(6u^4 - 5\,(M+1)\,u^3 + 4\left(A^2+M+Qv\right)u^2 + 3\left(BQv - A^2 - A^2M\right)u + 2A^2M)$,

$DU_\eta\,(u,v)_{21} = Sv\left(4u^3 + 3\,(B-v)\,u^2 + 2\left(A^2 - Bv\right)u + A^2\,(B-v)\right)$.

3 MAIN RESULTS

For system (3) we have the following general properties:

Lemma 2 *1, The set* $\bar{\Gamma} = \left\{(u,v) \in \tilde{\Omega}/\ 0 \le u \le 1,\ v \ge 0\right\}$ *is an invariant region*

2. The solutions are bounded

Proof. 1. Clearly the $u-axis$ and the $v-axis$ are invariant sets because the system is of Kolmogorov type. If $u = 1$, we have

$\frac{du}{d\tau} = -Q\,(1+C)\,v < 0$

and whatever it is the sign of

$\frac{dv}{d\tau} = S\,(1+A)\,(\ 1 + C -\ v)\ v$

the trajectories enter and remain in the region $\tilde{\Gamma}$.

2. We use the Poincaré compactification [9, 13] given by the change of variables $u = \frac{w}{z}$ and $v = \frac{1}{z}$, to obtain a new system with a zero Jacobian matrix. To desingularize the origin we use the blowing-up method; then, we have that $(0,0)$ is a non-hyperbolic saddle point; hence, the point $(0,\infty)$ in the compactified system (3) is a non-hyperbolic saddle point. ■

 ISBN: 978-84-617-8694-7

3.1 Nature of equilibrium points over axis

Lemma 3 *For all* $\eta = (A, S, C, Q) \in]0, 1[\times\mathbb{R}^3_+$

1. *The singularity* $(1, 0)$ *is a hyperbolic saddle point.*
2. *The equilibrium* $(M, 0)$ *is a hyperbolic repellor.*

Proof. Evaluating the Jacobian matrix in each point is immediate that
1) $\det DY_\eta\,(1,0) = -S\,(A+1)^2\,(1+B)\,(1-M) < 0$.
Therefore, the equilibrium $(1, 0)$ is saddle point.
2) $\det DY_\eta\,(M,0) = SM^3\,(1-M)\left(A^2+M^2\right) > 0$ and
$\mathrm{tr}DY_\eta\,(M,0) = M(A^2+M^2)(M(1-M)+S(B+M)) > 0$.
Then, the equilibrium $(M, 0)$ is a repellor point. ■ $\bar{\Sigma}$

Lemma 4 *The point* $(0, 0)$ *of the vector field* U_η *has a hyperbolic and a parabolic sector [29], determined for the line* $v = \frac{M+BS}{BS}u$*, i.e., there exists a separatrix curve* $\bar{\Sigma}$*, inthephaseplanethatdividesthebel is then an attractor point for certain trajectories and a saddle point for others.*

Proof. As $DY_\eta\,(0,0)$ is the zero matrix, we use the horizontal blowing-up given by the function $W\,(p,q) = (p, pq) = (u, v)$.

In the new vector field U_η the singularities are $(0, 0)$ and $\left(0, \frac{M+BS}{BS}\right)$, which are a hyperbolic saddle point and an attractor point, respectively. Then, a separatrix straight line exists in the phase plane pq, determined by $v = \frac{M+BS}{BS}u$.

Then, using the blowing down, the point $(0, 0)$ is a saddle-node in the vector field U_η; thus, there exist a curve dividing the behavior of trajectories on the phase plane. ■

3.2 A Special Case of a Weak Allee Effect

When $M = 0$, a particular case of the weak Allee effect is described by the system:

$$U_\zeta : \begin{cases} \frac{du}{d\tau} &= ((1-u)(u^2+A^2) - Q(u+B)v)u^3 \\ \frac{dv}{d\tau} &= S(u-v)(u+B)(u^2+A^2)v, \end{cases} \tag{5}$$

where $\zeta = (A, B, Q, S) \in (]0, 1[)^2 \times \mathbb{R}^2_+$. In this case, there are only two equilibrium points on the axes, $(0, 0)$ and $(1, 0)$, since $(M, 0)$ coincides with the origin. Now, the polynomial (4) can be factored as $up^*\,(u)$, where

$$p^*\,(u) = u^3 - (1-Q)\,u^2(\left(A^2+BQ\right)u - A^2. \tag{6}$$

Using Descartes's rule of signs on the polynomial $p^*\,(u)$ we have that:

1. There exists a unique positive root, if and only if, $1 - Q \leq 0$.

2. There are three, two (one of multiplicity two) or one positive real roots, if and only if, $1 > Q$. In any case, we can assure that there is at least one positive real root denoted by H. Making synthetic division between $p^*\,(u)$ and $u - H$ we obtain:

 ISBN: 978-84-617-8694-7

$$p^*(u) = (u-H)\left(u^2 - (1-Q-H)u + BQ + A^2 - H(1-Q-H)\right),$$

and the condition

$$Q = \frac{(1-H)\left(A^2+H^2\right)}{H(B+H)}.$$

As $Q > 0$ then $H < 1$. So, the quadratic factor of p^* can be written as

$$q(u) = u^2 - \frac{(1-H)\left(BH-A^2\right)}{H(B+H)}u + \frac{A^2}{H}.$$

Again, using Descartes's rule of signs for the polynomial $q(u)$ we have that:

1. No positive roots exist, if and only if, $BH - A^2 \leq 0$.
2. Assuming, $BH - A^2 > 0$, we analyze the discriminant:

$$D_q = \left(\frac{(1-H)\left(BH-A^2\right)}{H(B+H)}\right)^2 - 4\frac{A^2}{H}.$$

Then,

2.1 There no exist positive roots, if and only if, $D_q < 0$.

2.2 There exists a unique positive root of multiplicity two, if and only if, $D_q = 0$. This root is given by

$H_0 = \frac{(1-H)\left(BH-A^2\right)}{2H(B+H)}$.

2.3 There exist two positive roots, if and only if, $D_q > 0$, which are given by

$H_{1,2} = \frac{1}{2}\left(\frac{(1-H)\left(BH-A^2\right)}{2H(B+H)} \pm \sqrt{D_q}\right)$.

Lemma 5 *For system* (5) *inside the region* $\bar{\Gamma}$ *we have:*

a. If $BH \leq A^2$ *or else* $A^2 < BH$ *and* $D_q < 0$*, then the system has a unique positive equilibrium point:* $P_H = (H, H)$.

b. If $A^2 < BH$ *and* $D_q = 0$*, then the system has two singularities:* (H, H) *and* (H_0, H_0).

c. If $A^2 < BH$ *and* $D_q > 0$*, then the system has three singularities:* (H, H), (H_1, H_1) *and* (H_2, H_2).

In this case, the Jacobian matrix of system (5) has the same components that the Jacobian matrix of system (3), except the first entry given by

$DU_\zeta(u,v)_{11} = -u^2\left(6u^3 - 5u^2 + 4\left(A^2+Qv\right)u + 3\left(BQv - A^2\right)\right)$.

We recall that, for all parameter values the equilibrium point $(1, 0)$ is a hyperbolic saddle point for all parameter values.

 ISBN: 978-84-617-8694-7

Lemma 6 *The equilibrium point* $(0,0)$ *of the vector field* U_ζ *has two hyperbolic sectors and one parabolic sector determined by the straight line* $v = u$*, i.e., there exists a separatrix curve* $\bar{\Sigma}_0$ *in the phase plane that divides the behavior of trajectories; the point* $(0,0)$ *is a point attractor for certain trajectories, a saddle point for others and a repellor point for others.*

Proof. The proof is similar to that of Lemma 4. The Jacobian matrix of the vector field U_ζ evaluated at $(0,0)$ is the zero matrix, then we use the horizontal blowing-up. The demonstration is in two parts with the understanding that in this case the origin represents the collapse of three points: two obtained in Lemma 4 and the other point $(M,0)$.

Based on the properties of the points $(0,0)$, $\left(0, \frac{M+BS}{BS}\right)$ and $(M,0)$ of system (3), we can see that there exists the separatrix curve $\bar{\Sigma}_0$ and the origin is an attractor point for certain trajectories, a saddle point for some and a repellor point for others. ■

The determinant of the Jacobian matrix evaluated in the equilibrium point (u,u) is
$\det DU_\zeta(u,u) = Su^3\left(A^2+u^2\right)(B+u)\left((1-Q)u^2 - 2\left(A^2+BQ\right)u + 3A^2\right)$;
then, the sign of $\det DU_\zeta(u,u)$ depends on the factor
$\mathrm{d}[u] = (1-Q)u^2 - 2\left(A^2+BQ\right)u + 3A^2$.
The trace of the Jacobian matrix is
$TrDU_\zeta(u,u) = -u\,(t\,[u])$
where the factor $t\,[u]$ is
$\mathrm{t}[u] = (1-2Q+S)u^3 - \left(2A^2+3BQ-BS\right)u^2 + A^2(S+3)u + A^2BS$.

Now, suppose $BH \leq A^2$ or $A^2 < BH$ and $D_q < 0$; then, the system has a unique positive equilibrium point (H,H), and we have that:

Theorem 7 *The singularity* (H,H) *is:*

1. *a stable equilibrium point, if and only if,*
$S > \frac{H^2\left(H^2-2H^3-A^2B-3BH^2+2BH-A^2\right)}{\left(A^2+H^2\right)(B+H)^2}$,
2. *an unstable singularity surrounding a stable limit cycle, if and only if,*
$S < \frac{H^2\left(H^2-2H^3-A^2B-3BH^2+2BH-A^2\right)}{\left(A^2+H^2\right)(B+H)^2}$.

Proof. At the point (H,H) we have that
$d[H] = \frac{1}{B+H}\left(\left(A^2+H^3\right)(B+H) - H(1-H)\left(BH-A^2\right)\right)$
In both cases when there is a unique positive equilibrium point we have that
$d\,[u] > 0$, then, $\det DU_\zeta(u,u) > 0$
and the nature of singularity (H,H) depends on the trace, that is, it depends of
$t\,[H] = \left(A^2+H^2\right)(B+H)S - \frac{H^2\left(-2H^3+(1-3B)H^2+2BH-A^2(B+1)\right)}{B+H}$
Clearly,

1. If $t\,[H] > 0$, then $TrDU_\zeta(u,u) < 0$, and the point (H,H) is a stable equilibrium point.

 ISBN: 978-84-617-8694-7

2. If $t[H] < 0$, then $TrDU_\zeta(u,u) > 0$, and the point (H,H) is an unstable equilibrium point. Moreover, verifying the transversality condition [9], we have

$\frac{\partial\left(traDU_\zeta(H,H)\right)}{\partial S} = \left(A^2 + H^2\right)(B+H) > 0.$

By Hopf Bifurcation Theorem a stable limit cycle is generated at this point. ■

Remark. The existence of more than a limit cycle in the system (9) must be demonstrated, since by simulations it is possible to show the existence of two limit cycles .

4 CONCLUSIONS

In this work we have studied the dynamics of a modified Leslie-Gower type predator-prey model, considering a Holling type III functional response and double Allee effect on prey. The particular case of a weak Allee effect was mainly analyzed where $M = 0$, due to the algebraic complexities to obtain equilibrium points in the general case.

The expressions for the coordinates of these equilibrium points when $M > 0$, correspond to the roots of a polynomial of the fourth degree, which are quite complex because the polynomial coefficients depend on the system parameters.

Another difficulty for the analysis is that the original system (1) describing our model it is not defined in the origin $(0,0)$. To solve this problem and to reduce the number of parameters, we made a reparameterization and a time rescaling, by using a diffeomorphism, obtaining a polynomial topologically equivalent system described for system (3).

For this new system (3), we determined the nature of singularities located over the axis, when $M > 0$, emphasizing the importance of the point $(0,0)$ on the global dynamics of the model. It has been shown that this point has a parabolic and a hyperbolic sectors [29].

Using the directional blowing-up method, we show the existence of a curve determined by the stable manifold of the origin, which divides the behavior of the trajectories, implying that trajectories near the curve are highly sensitive to initial conditions. This means that the paths above this separatrix curve have the point $(0,0)$ as its $\omega-limit$, while those trajectories under the separatrix may have different $\omega - limits$, could be an attractor equilibrium point or a stable limit cycle.

For $M = 0$ it is possible to obtain three possible dynamics of system according the amount of positive equilibrium points, (one, two or three). In this work, the local stability was analyzed with a unique positive equilibrium point, obtaining highly interesting results, in the sense that populations can coexist for a wide set of parameter values.

For example, when there is a single interior point, this can be stable or unstable, but is unstable when the appearance of the stable limit cycle, via Hopf bifurcation, prevents the global stability of the origin, which means that there is no extinction of both species. However, for certain parameter values, this limit cycle is broken and the origin $(0,0)$ becomes a global asymptotically stable equilibrium point, implying the extinction of both species.

 ISBN: 978-84-617-8694-7

Acknowledgements

This work has been partially supported by by DIEA-PUCV 124.730/2012 project

References

[1] P. Aguirre, E. González-Olivares, E. Sáez, Two limit cycles in a Leslie-Gower predator-prey model with additive Allee effect, Nonlinear Analysis: Real World Applications 10 (2009) 1401-1416.

[2] P. Aguirre, E. González-Olivares, E. Sáez, Three limit cycles in a Leslie-Gower predator-prey model with additive Allee effect, SIAM Journal on Applied Mathematics 69(5) (2009) 1244-1269.

[3] E. Angulo, G. W. Roemer, L. Berec, J. Gascoigne and F. Courchamp, Double Allee effects and extinction in the island fox, Conservation Biology 21 (2007) 1082-1091.

[4] D. K. Arrowsmith and C. M. Place, *Dynamical Systems. Differential equations, maps and chaotic behaviour*, Chapman and Hall (1992).

[5] A. D. Bazykin, *Nonlinear Dynamics of interacting populations*, World Scientific Publishing Co. Pte. Ltd., 1998.

[6] L. Berec, E. Angulo and F. Courchamp, Multiple Allee effects and population management,Trends in Ecology and Evolution 22 (2007) 185-191.

[7] D. S. Boukal and L. Berec, Single-species models and the Allee effect: Extinction boundaries, sex ratios and mate encounters, Journal of Theoretical Biology 218 (2002) 375-394.

[8] D. S. Boukal,M. W. Sabelis and L. Berec, How predator functional responses and Allee effects in prey affect the paradox of enrichment and population collapses, Theoretical Population Biology 72 (2007) 136-147.

[9] C. Chicone, *Ordinary differential equations with applications* (2nd edition), Texts in Applied Mathematics 34, Springer (2006).

[10] J. B. Collings, The effect of the functional response on the bifurcation behavior of a mite predator-prey interaction model, Journal of Mathematical Biology 36 (1997) 149-168.

[11] F. Courchamp, T. Clutton-Brock and B. Grenfell, Inverse dependence and the Allee effect, Trends in Ecology and Evolution 14 (1999) 405-410.

 ISBN: 978-84-617-8694-7

[12] F. Courchamp, L. Berec and J. Gascoigne, *Allee effects in Ecology and Conservation*, Oxford University Press 2008.

[13] F. Dumortier, J. Llibre and J. C. Artés, *Qualitative theory of planar differential systems*, Springer (2006).

[14] H. I. Freedman, *Deterministic mathematical models in Population Ecology*, Marcel Dekker (1980).

[15] W. M. Getz, A hypothesis regarding the abruptness of density dependence and the growth rate populations, Ecology 77(7) (1996) 2014-2026.

[16] E. González-Olivares, B. González-Yañez, J. Mena-Lorca, R. Ramos-Jiliberto, Modelling the Allee effect: Are the different mathematical forms proposed equivalents? In: R. Mondaini (Ed.) *Proceedings of the International Symposium on Mathematical and Computational Biology BIOMAT 2006*, E-papers Serviços Editoriais Ltda. Rio de Janeiro (2007) 53-71.

[17] E. González-Olivares, J. Mena-Lorca, A. Rojas-Palma, J. D. Flores, Dynamical complexities in the Leslie-Gower predator-prey model as consequences of the Allee effect on prey, Applied Mathematical Modelling 35 (2011) 366-381.

[18] E. González-Olivares and A. Rojas-Palma, Multiple limit cycles in a Gause type predator-prey model with Holling type III functional response and Allee effect on prey, Bulletin of Mathematical Biology 73 (2011) 1378-1397.

[19] E. González-Olivares, P. Tintinago-Ruiz and A. Rojas-Palma, A Leslie-Gower type predator-prey model with sigmoid funcional response, International Journal of Computer Mathematics 93(9) (2015) 1895-1909.

[20] E. Gonzalez-Olivares B. González-Yañez and A. Rojas-Palma, Multiple limit cycles in a Leslie-Gower type predator-prey model considering weak Allee effect on prey, Nonlinear Analysis: Modelling and Control 22(3) (2017) 347-365.

[21] B. González-Yañez, E. González-Olivares and J. Mena-Lorca, Multistability on a Leslie-Gower type predator-prey model with nonmonotonic functional response, In R. Mondaini and R. Dilao (eds.), *BIOMAT 2006 - International Symposium on Mathematical and Computational Biology*, World Scientific Co. Pte. Ltd. (2007) 359-384.

[22] R.P. Gupta and P. Chandra, Bifurcation analysis of modified Leslie-Gower predator-prey model with Michaelis-Menten type prey harvesting, Journal of Mathematical Analysis and Applications 398 (2013) 278-398, 278-295.

 ISBN: 978-84-617-8694-7

[23] R.P. Gupta, M. Banerjee and P. Chandra, Bifurcation analysis and control of Leslie-Gower predator-prey model with Michaelis-Menten type prey-harvesting, Differential Equations and Dynamical Systems 20 (2012) 339-366.

[24] Y. A. Kuznetsov, *Elements of Applied Bifurcation Theory, (2nd Edition)*, Springer-Verlag 1998.

[25] P. H. Leslie, Some further notes on the use of matrices in Population Mathematics, Biometrika 35 (1948) 213-245.

[26] P. H. Leslie and J. C. Gower, The properties of a stochastic model for the predator-prey type of interaction between two species, Biometrica 47 (1960) 219-234.

[27] R. M. May, *Stability and complexity in model ecosystems* (2nd edition), Princeton University Press (2001).

[28] J. Mena-Lorca, E. González-Olivares, B. González-Yañez, The Leslie-Gower predator-prey model with Allee effect on prey: A simple model with a rich and interesting dynamics, in: R. Mondaini (Ed.), *Proceedings of the 2006 International Symposium on Mathematical and Computational Biology BIOMAT 2006*, E-papers Serviços Editoriais Ltda., Rio de Janeiro, 2007 105-132.

[29] L. Perko, *Differential Equations and Dynamical Systems,* (3rd ed), Texts in Applied Mathematics 7, Springer-Verlag 2001.

[30] P. A. Stephens and W. J. Sutherland, Consequences of the Allee effect for behaviour, ecology and conservation. Trends in Ecology and Evolution 14 (1999) 401-405.

[31] P. A. Stephens, W. J. Sutherland and R. P. Freckleton, What is the Allee effect?, Oikos 87 (1999) 185-190.

[32] R. J. Taylor, *Predation*, Chapman and Hall, 1984.

[33] P. Turchin, *Complex population dynamics. A theoretical/empirical synthesis*, Monographs in Population Biology 35, Princeton University Press (2003).

[34] G. A. K. van Voorn, L. Hemerik, M. P. Boer and B. W. Kooi, Heteroclinic orbits indicate overexploitation in predator-prey systems with a strong Allee, Mathematical Biosciences 209 (2007) 451-469.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

An optimal scheme for multiple roots of nonlinear equations with eighth-order convergence

Ramandeep Behl[1], Ali Saleh Alshomrani[2] and Sandile S. Motsa[3,1]

[1] *School of Mathematics, Statistics and Computer Sciences, University of KwaZulu-Natal, Private Bag X01, Scottsville 3209, Pietermaritzburg, South Africa.*

[2] *Department of Mathematics, King Abdulaziz University, Jeddah 21577, Saudi Arabia.*

[3] *Mathematics Department, University of Swaziland, Private Bag 4, Kwaluseni, M201, Swaziland.*

emails: ramanbehl87@yahoo.in, aszalshomrani@kau.edu.sa, sandilemotsa@gmail.com

Abstract

We have a good number of eighth-order iterative methods for simple zeros of nonlinear equations in the available literature. But, unfortunately, we don't have a single iterative method of eighth-order for multiple zeros with known or unknown multiplicity. Some scholars from the worldwide have tried to present optimal or non-optimal multipoint eighth-order iteration functions. But, unfortunately, none of them get success in this direction and attained maximum sixth-order convergence in the case of multiple zeros with known multiplicity m. Motivated and inspired by this fact, we propose an optimal scheme with eighth-order convergence based on weight function approach. In addition, an extensive convergence study is discussed in order to demonstrate the eighth-order convergence of the proposed scheme. The proposed scheme is optimal in the sense of Kung-Traub conjecture. Moreover, we also show the applicability of the proposed scheme on some real life as well as academic problems. These problems illustrate that our proposed iterative functions are more efficient among the available multiple root finding techniques.

Key words: Multiple zeros, iterative functions, Kung-Traub conjecture, Nonlinear equations, efficiency index

MSC 2000: AMS codes (optional)

1 Introduction

Finding new higher-order multipoint iteration functions for the multiple zeros with known multiplicity m of univariate function of the form

$$f(x) = 0, \tag{1.1}$$

(where $f : \mathbb{C} \to \mathbb{C}$ is analytic function in a neighborhood of the required zero) is one of the most important and challenging task in the field of numerical analysis. The advantages of multipoint iterative methods over the one-point iterative methods can be seen in well-known standard text books Ostrowski [1], Traub [2] and Petković et al. [3].

In the past and recent years, several scholars given optimal fourth-order iteration functions for multiple zeros with known multiplicity m e.g. Li et al. [4] in (2009), Sharma and Sharma [5] and Li et al. [6] in (2010), Zhou et al. [7] in (2011), Sharifi et al. [8] in (2012), Soleymani et al. [9], Soleymani and Babajee [10], Liu and Zhou [11] and Zhou et al. [12] in (2013), Thukral [13] in (2014), Behl et al. [14] and Hueso et al. [15] in (2015) and Behl et al. [16] in (2016) . In addition, Li et al. [6] (expect two of them are optimal) and Neta [17] given non-optimal iterative functions of order four. Most of the above mentioned iteration functions are either the extension or modification of modified Newton's method (also known as Rall's method [2]) or Newton like method at the expense of additional functional evaluations or increase the substep of the original methods.

In 2013, Thukral [18] presented a multi-point iterative method with sixth-order convergence, which is given by

$$\begin{aligned} y_n =& x_n - m\frac{f(x_n)}{f'(x_n)}, \\ z_n =& x_n - m\frac{f(x_n)}{f'(x_n)}\sum_{i=1}^{3} i\left(\frac{f(y_n)}{f(x_n)}\right)^{\frac{i}{m}}, \\ x_{n+1} =& z_n - m\frac{f(x_n)}{f'(x_n)}\left(\frac{f(z_n)}{f(x_n)}\right)^{\frac{1}{m}}\left[\sum_{i=1}^{3} i\left(\frac{f(y_n)}{f(x_n)}\right)^{\frac{i}{m}}\right]^2. \end{aligned} \tag{1.2}$$

Recently, Geum et al. in [19], presented a non-optimal family of two-point sixth-order methods to find multiple zeros, given as follows:

$$\begin{aligned} y_n &= x_n - m\frac{f(x_n)}{f'(x_n)}, \quad m > 1, \\ x_{n+1} &= y_n - Q(u_n, s_n)\frac{f(y_n)}{f'(y_n)}, \end{aligned} \tag{1.3}$$

where, $u_n = \sqrt[m]{\frac{f(y_n)}{f(x_n)}}$, $s_n = \sqrt[m-1]{\frac{f'(y_n)}{f'(x_n)}}$ and Q is an analytic function in a neighborhood of $(0, 0)$. The problem with this scheme is that it does not work for simple zeros ($m = 1$).

 ISBN: 978-84-617-8694-7

Very recently, another non-optimal family of three-point sixth-order methods for multiple zeros was proposed by Geum et al. [20], and it is given by

$$\begin{aligned} y_n &= x_n - m\frac{f(x_n)}{f'(x_n)}, \quad m \geq 1, \\ w_n &= y_n - mG(u_n)\frac{f(x_n)}{f'(x_n)}, \\ x_{n+1} &= w_n - mK(u_n, v_n)\frac{f(x_n)}{f'(x_n)}, \end{aligned} \tag{1.4}$$

where $u_n = \sqrt[m]{\frac{f(y_n)}{f(x_n)}}$ and $v_n = \sqrt[m]{\frac{f(w_n)}{f(x_n)}}$. The weight functions $G : \mathbb{C} \to \mathbb{C}$ and $K : \mathbb{C}^2 \to \mathbb{C}$ are analytic in a neighborhood of 0 and $(0, 0)$, respectively.

The above discussion illustrate that many scholars have tried to construct iteration functions for multiple zeros with optimal/non-optimal eighth-order convergence. But, none of them get success in this direction till date and the highest attained order is six. So, this means there is not a single research article where any scholar claims optimal eighth-order convergence is attained to this date, according to our knowledge. In addition, the optimal multipoint iterative functions have more importance than the non-optimal ones because of the higher efficiency index, timing and faster convergence.

Keeping all these points in our mind, we try to solve this problem which has been pending for several years. So, we propose an optimal scheme for multiple zeros with eighth-order convergence. Weight function approach is used in the construction of this scheme. In addition, our proposed methods not only give the faster convergence but also have smaller residual error and asymptotic error constants. Each member of the proposed scheme also satisfies the classical Kung-Traub conjecture [21]. So, all the members are optimal in the sense of Kung-Traub conjecture. We have demonstrated the usefulness of the presented iteration functions by performing several applied science problems for numerical tests. Finally, we concluded from the numerical results that our methods have far better numerical results, than the existing robust iterative schemes (in the regards of smaller asymptotic error constants, stable computational order of convergence and smaller residual errors).

2 Construction of the proposed scheme

In this section, we present the following new scheme

 ISBN: 978-84-617-8694-7

$$\begin{aligned} y_n &= x_n - m\frac{f(x_n)}{f'(x_n)}, \\ z_n &= y_n - \frac{f(x_n)}{f'(x_n)} u_n G_f(u_n), \\ x_{n+1} &= z_n + \frac{f(x_n)}{f'(x_n)} \frac{w_n u_n}{1-w_n}\Big(H_f(u_n) + K_f(v_n)\Big), \end{aligned} \tag{2.1}$$

where the weight functions $G_f, H_f, K_f : \mathbb{C} \to \mathbb{C}$ are analytic functions in a neighborhood of (0) with $u_n = \left(\frac{f(y_n)}{f(x_n)}\right)^{\frac{1}{m}}$, $v_n = \left(\frac{f(z_n)}{f(x_n)}\right)^{\frac{1}{m}}$ $w_n = \left(\frac{f(z_n)}{f(y_n)}\right)^{\frac{1}{m}}$.

In the next Theorem 2.1, we demonstrate that the order of convergence of the proposed scheme will reach at eight without using any additional function evaluations. In addition, we will also discuss the conditions on weight functions in the following Theorem 2.1.

Theorem 2.1 *Let $x = \xi$ be a multiple zero with multiplicity m of an analytic function $f : \mathbb{C} \to \mathbb{C}$ in the region enclosing the multiple zero ξ. Then, the proposed scheme* (2.1) *attain eighth-order convergence when the weight functions G_f, H_f, K_f will satisfy the following conditions*

$$\begin{cases} G(0) = m, \quad G'(0) = 2m, \quad H(0) = -m - K(0), \quad H'(0) = -2m, \quad H''(0) = -G''(0) - 2m, \\ K'(0) = -2m, \quad H'''(0) = -\big(6G''(0) + G'''(0) - 24m\big). \end{cases} \tag{2.2}$$

Proof Let us assume that $e_n = x_n - \alpha$ be the error at nth step. We will obtain the following expressions by expanding $f(x_n)$ and $f'(x_n)$ about $x = \alpha$ with the help of Taylor's series

$$f(x_n) = \frac{f^{(m)}(\alpha)}{m!} e_n^m \Big(1 + c_1 e_n + c_2 e_n^2 + c_3 e_n^3 + c_4 e_n^4 + c_5 e_n^5 + c_6 e_n^6 + c_7 e_n^7 + c_8 e_n^8 + O(e_n^9)\Big), \tag{2.3}$$

and

$$\begin{aligned} f'(x_n) = &\frac{f^m(\alpha)}{(m)!} e_n^{m-1} \Big(m + c_1(m+1)e_n + c_2(m+2)e_n^2 + c_3(m+3)e_n^3 + c_4(m+4)e_n^4 \\ &+ c_5(m+5)e_n^5 + c_6(m+6)e_n^6 + c_7(m+7)e_n^7 + c_8(m+8)e_n^8 + O(e_n^9)\Big), \end{aligned} \tag{2.4}$$

respectively.
With the help of above expressions (2.3) and (2.4), we have

$$y_n - \alpha = \frac{c_1 e_n^2}{m} + \frac{2c_2 m - c_1^2(m+1)}{m^2} e_n^3 + \sum_{k=0}^{4} \xi_k e_n^{k+4} + O(e_n^9), \tag{2.5}$$

 ISBN: 978-84-617-8694-7

where $\xi_k = \xi_k(m, c_1, c_2, \ldots, c_8)$ are given in terms of $m, c_2, c_3, \ldots, c_8$ with explicitly written two coefficients $\xi_0 = \frac{1}{m^3}\{3c_3m^2 + c_1^3(m+1)^2 - c_2c_1m(3m+4)\}$ and $\xi_1 = \frac{1}{m^4}\{c_1^4(m+1)^3 - 2c_2c_1^2m(2m^2+5m+3) + 2c_3c_1m^2(2m+3) + 2m^2(c_2^2(m+2) - 2c_4m)\}$, etc.
Again by using Taylor Series expansion, we further yield

$$f(y_n) = f^{(m)}(\alpha)e_n^{2m}\Bigg[\frac{\left(\frac{c_1}{m}\right)^m}{m!} + \frac{(2c_2m - c_1^2(m+1))\left(\frac{c_1}{m}\right)^m e_n}{c_1m!} + \left(\frac{c_1}{m}\right)^{1+m}\frac{1}{2m!c_1^3}\Big\{(3+3m+3m^2 + m^3)c_1^4 - 2m(2+3m+2m^2)c_1^2c_2 + 4(-1+m)m^2c_2^2 + 6m^2c_1c_3\Big\}e_n^2 + \sum_{k=0}^{5}\bar{\xi}_k e_n^{k+3} + O(e_n^9)\Bigg] \tag{2.6}$$

and

$$u_n = \frac{c_1e_n}{m} + \frac{(2c_2m - c_1^2(m+2))e_n^2}{m^2} + \frac{c_1^3(2m^2+7m+7) + 6c_3m^2 - 2c_2c_1m(3m+7)}{2m^3}e_n^3 + \lambda_0e_n^4 + \lambda_1e_n^5 + O(e_n^6), \tag{2.7}$$

where $\lambda_0 = -\frac{1}{6m^4}\big[c_1^4(6m^3+29m^2+51m+34) - 6c_2c_1^2m(4m^2+16m+17) + 12c_3c_1m^2(2m+5) + 12m^2(c_2^2(m+3) - 2c_4m)\big]$ and $\lambda_1 = \frac{1}{24m^5}\big[-24m^3(c_2c_3(5m+17) - 5c_5m) + 12c_3c_1^2m^2(10m^2 + 43m + 49) + 12c_1m^2\{c_2^2(10m^2+47m+53) - 2c_4m(5m+13)\} - 4c_2c_1^3m(30m^3+163m^2 + 306m + 209) + c_1^5(24m^4+146m^3+355m^2+418m+209)\big]$.
It is straightforward to say from the expression (2.7) that u_n is of order e_n. So, we can expand weight function $G_f(u_n)$ in the neighborhood of origin by Taylor series expansion up to fourth-order term for the eighth order convergence as follow:

$$G_f(u_n) = G(0) + G'(0)u_n + \frac{1}{2!}G''(0)u_n^2 + \frac{1}{3!}G'''(0)u_n^3 + \frac{1}{4!}G''''(0)u_n^4. \tag{2.8}$$

By inserting the expressions (2.3) – (2.8) in the second substep of the proposed scheme (2.1), we have

$$z_n - \alpha = \frac{(m - G(0))c_1}{m^2}e_n^2 + \frac{\big(3G - G'(0) - m + Gm - m^2\big)c_1^2 + 2m(-G+m)c_2}{m^3}e_n^3 + \sum_{k=1}^{5}\Omega_k e_n^{k+3} + O(e_n^9). \tag{2.9}$$

where $\Omega_k = \Omega_k(G(0), G'(0), G''(0), G'''(0), G''''(0), m, c_1, c_2, \ldots, c_8)$.
Fourth-order convergence can be attained, when the coefficient of e_n^2 and e_n^3 simultaneously are equaled to zero. That is possible only for the following particular values of $G(0)$ and $G'(0)$

$$G(0) = m, \quad G'(0) = 2m, \tag{2.10}$$

which can be obtained from the expression (2.9).
Insert the above expression (2.10), we obtain

$$z_n - \alpha = \frac{\big(-G''(0)+m(9+m)\big)c_1^3 - 2m^2c_1c_2}{2m^4}e_n^4 + \sum_{k=1}^{4}\bar{\Omega}_k e_n^{k+4} + O(e_n^9). \tag{2.11}$$

Now, again by using the Taylor series expansion, we have

$$f(z_n) = f^{(m)}(\alpha)e_n^{4m}\left[\frac{2^{-m}}{m!}\left(\frac{(m(9+m)-G''(0))c_1^3 - 2m^2c_1c_2}{m^4a_1^2}\right)^m + \sum_{k=1}^{5}\bar{\bar{\Omega}}_k e_n^k + O(e_n^6)\right], \tag{2.12}$$

$$v_n = \frac{\big((-G''(0)+m(9+m))c_1^3 - 2m^2c_1c_2\big)}{2m^4}e_n^3 + \beta_0 e_n^4 + \beta_1 e_n^5 + O(e_n^6), \tag{2.13}$$

and

$$w_n = \frac{(m(9+m)-G''(0))c_1^2 - 2m^2c_2}{2m^4}e_n^2 + \gamma_0 e_n^3 + \gamma_1 e_n^4 + O(e_n^5), \tag{2.14}$$

where $\beta_0 = -\frac{1}{6m^5}\Big[\big(G'''(0) - 3G''(0)(8+3m) + m(152+87m+7m^2)\big)c_1^4 - 6m\big(m(29+4m) - 3G''(0)\big)c_1^2c_2 + 12m^3c_2^2 + 12m^3c_1c_3\Big]$, $\beta_1 = \frac{1}{24m^6}\Big[\big(-G''''(0) + 2061m + 2246m^2 + 711m^3 + 46m^4 + 8G'''(0)(5+2m) - 12G''(0)(37+30m+6m^2)\big)c_1^5 - 4m\big(8G'''(0) - 9G''(0)(20+7m) + m(1123 + 624m + 53m^2)\big)c_1^3c_2 + 12m^2(-9G''(0) + m(87+13m))c_1^2c_3 - 168m^4c_2c_3 - 12m^2c_1\big((12G''(0) - m(121+17m))c_2^2 + 6m^2c_4\big)\Big]$, $\gamma_0 = -\frac{1}{6m^4}\Big[\big(G'''(0) + 98m + 54m^2 + 4m^3 - 6G''(0)(3+m)\big)c_1^3 + 12m(G''(0) - m(9+m))c_1c_2 + 12m^3c_3\Big]$, $\gamma_1 = \frac{1}{24m^5}\Big[\big(899m + 1002m^2 + 313m^3 + 18m^4 + 4G'''(0)(8+3m) - 6G''(0)(43+33m+6m^2) - G''''(0)\big)c_1^4 - 12m\big(2G'''(0) - G''(0)(33+10m) + m(167+87m+6m^2)\big)c_1^2c_2 + 24m^2(-3G''(0) + m(26+3m))c_1c_3 + 12m^2\big((-4G''(0) + m(35+3m))c_2^2 - 6m^2c_4\big)\Big]$.

We observed from the expressions namely, (2.7) and (2.14) that u_n and v_n are of order e_n and e_n^3, respectively. Therefore, we can expand both weight functions $H_f(u_n)$ and $K_f(v_n)$ in the neighborhood of origin (0) by Taylor series expansion up to fourth-order and third-order terms, respectively as follow:

$$H_f(u_n) = H(0) + H'(0)u_n + \frac{1}{2!}H''(0)u_n^2 + \frac{1}{3!}H'''(0)u_n^3 + \frac{1}{4!}H''''(0)u_n^4 \tag{2.15}$$

and

$$K_f(v_n) = K(0) + K'(0)v_n + \frac{1}{2!}K''(0)v_n^2 + \frac{1}{3!}K'''(0)v_n^3. \tag{2.16}$$

By using the expression (2.3) – (2.16), we have

$$e_{n+1} = \frac{(K(0)+H(0)+m)\big(m(9+m)-G''(0)\big)c_1^3 - 2m^2c_1c_2\big)}{2m^5}e_n^4 + \sum_{k=0}^{3}\phi_k e_n^{k+5} + O(e_n^9), \tag{2.17}$$

 ISBN: 978-84-617-8694-7

where $\phi_k = \phi_k(m, G''(0), G'''(0), G''''(0), H(0), H'(0), H''(0), H'''(0), H''''(0), K(0), K'(0), K''(0), K'''(0), c_1, c_2, \ldots, c_8)$.
We can easily obtain at least fifth-order convergence, when we choose the following value of $H(0)$

$$H(0) = -K(0) - m, \tag{2.18}$$

which is obtained from the expression (2.17).
In addition, by using this value $H(0) = -K(0) - m$ in $\phi_0 = 0$. Then, we have

$$H'(0) = -2m. \tag{2.19}$$

By inserting the expressions namely, (2.18) and (2.19) in $\phi_1 = 0$, we get

$$(G''(0) + H''(0) + 2m) = 0 \tag{2.20}$$

which further gives

$$H''(0) = -(G''(0) + 2m). \tag{2.21}$$

Now, with the help of expressions (2.18), (2.19) and (2.21) in $\phi_2 = 0$, we have

$$K1+2m = 0, \;\; 3G''(0)K'(0) - m\left(G'''(0) + H'''(0) + 27K'(0) + 30m + 3K'(0)m + 6m^2\right) = 0, \tag{2.22}$$

the above two independent expressions which leads us

$$K'(0) = -2m \quad H'''(0) = -(6G''(0) + G'''(0) - 24m). \tag{2.23}$$

Finally, we will use the expressions (2.18), (2.19), (2.21) and (2.23) in the expression (2.17), in order to obtain final asymptotic error constant term. Then, we obtain

$$e_{n+1} = \frac{c_1^2\big((m(9+m) - G''(0))c_1^2 - 2m^2c_2\big)}{48m^9}\Big[\big(8G'''(0) - 36G''(0) + G''''(0) + H''''(0) + 268m + 84m^2 + 8m^3\big)c_1^3 - 24m^2(7+m)c_1c_2 + 24m^3c_3\Big]e_n^8 + O(e_n^9), \tag{2.24}$$

where $G''(0), G'''(0), G'''(0), H''(0), K(0), K''(0)$ and $K'''(0)$ are finite real/complex parameters.
The above asymptotic error constant term (2.24) reveals that the proposed scheme (2.1) attain eighth-order convergence by using only four functional evaluations (viz. $f(x_n)\, f'(x_n)\, f(y_n)$ and $f(z_n)$) per iteration. So, it is optimal in the sense of classical Kung-Traub conjecture. This completes the proof. □

Remark 2.2 *It is worthy to note that the weight functions G_f, H_f and K_f play an important role in the construction of the proposed scheme with desired convergence order. However, only terms namely, $G''(0), G'''(0), G''''(0)$ and $H''''(0)$ are involve in the asymptotic error constant term* (2.24)*. On the other hand, other terms namely $K''(0)$ and $K'''(0)$ do not effect the asymptotic error constant term* (2.24) *at all. So, we can say that these are dummy parameters. However, we can't leave them in the beginning.*

 ISBN: 978-84-617-8694-7

2.1 Special cases of the proposed family

In this section, we will discuss some special cases of our proposed class (2.1) by assigning different weight functions Q_f and G_f.

1. Let us consider the following weight functions which are chosen directly from the proposed Theorem 2.1. Then, we get a new optimal class of order eight as follows:

$$\begin{aligned} y_n &= x_n - m\frac{f(x_n)}{f'(x_n)}, \\ z_n &= y_n - \frac{f(x_n)}{f'(x_n)}u_n\left[m + 2mu_n + \frac{G''(0)u_n^2}{2} + \frac{G'''(0)u_n^3}{6} + \frac{G''''(0)u_n^4}{24}\right], \\ x_{n+1} &= z_n + \frac{f(x_n)}{f'(x_n)}\frac{w_n u_n}{1-w_n}\Big[-m - 2mu_n - \frac{(G''(0)+2m)}{2}u_n^2 + \frac{1}{6}(24m - 6G''(0) - G'''(0))u_n^3 \\ &\quad + \frac{H''''(0)u_n^4}{24} - 2mv + \frac{K''(0)v^2}{2} + \frac{K'''(0)v^3}{6}\Big], \end{aligned} \tag{2.25}$$

where $G''(0), G'''(0), G''''(0)$, $K''(0), K'''(0)$ and $H''''(0)$ are free disposable parameters.
Sub cases of the given scheme (2.25):

(a) Let us consider $G''(0) = -2m, K''(0) = K'''(0) = G'''(0) = G''''(0) = 0, H''''(0) = -4m(85 + 21m + 2m^2)$ in expression (2.25), we obtain

$$\begin{aligned} y_n &= x_n - m\frac{f(x_n)}{f'(x_n)}, \\ z_n &= y_n - m\frac{f(x_n)}{f'(x_n)}u_n(1 + 2u_n - u_n^2), \\ x_{n+1} &= z_n + m\frac{f(x_n)}{f'(x_n)}\frac{w_n u_n}{1-w_n}\Big[-1 - 2u_n + 6u_n^3 - \frac{1}{6}(85 + 21m + 2m^2)u_n^4 - 2v\Big]. \end{aligned} \tag{2.26}$$

(b) For $G''(0) = K''(0) = K'''(0) = G''(0) = G''''(0) = H''''(0) = 0$, in expression (2.25), we have

$$\begin{aligned} y_n &= x_n - m\frac{f(x_n)}{f'(x_n)}, \\ z_n &= y_n - m\frac{f(x_n)}{f'(x_n)}u_n(1 + 2u_n), \\ x_{n+1} &= z_n + m\frac{f(x_n)}{f'(x_n)}\frac{w_n u_n}{1-w_n}\Big[-1 - 2u_n - u_n^2 + 4u_n^3 - 2v_n\Big]. \end{aligned} \tag{2.27}$$

 ISBN: 978-84-617-8694-7

2. Another choice of weight functions gives the following family of eighth-order methods

$$\begin{aligned} y_n &= x_n - m\frac{f(x_n)}{f'(x_n)}, \\ z_n &= y_n - m\frac{f(x_n)}{f'(x_n)}u_n(1+2u_n), \\ x_{n+1} &= z_n - m\frac{f(x_n)}{f'(x_n)}\frac{w_n u_n}{1-w_n}\Big[\frac{1+9u_n^2+2v_n+u_n(6+8v_n)}{1+4u_n}\Big]. \end{aligned} \tag{2.28}$$

3. By considering one more weight functions of the following form, we will obtain one more optimal family of eighth-order iteration function

$$\begin{aligned} y_n &= x_n - m\frac{f(x_n)}{f'(x_n)}, \\ z_n &= y_n - m\frac{f(x_n)}{f'(x_n)}u_n\left[\frac{m}{1-2u_n}\right], \\ x_{n+1} &= z_n + m\frac{f(x_n)}{f'(x_n)}\frac{w_n u_n}{1-w_n}\Big[\frac{5+u_n^2+10v_n-2u_n(1+12v_n)}{-5+12u_n}\Big]. \end{aligned} \tag{2.29}$$

In the similar fashion by arbitrary choice of assumed weight functions $G_f(u_n)$, $H_f(u_n)$ and $K_f(v_n)$ (provided the conditions on them in the above Theorem 2.1 should be satisfied), we can obtain several new optimal methods of eighth-order for multiple zeros.

3 Numerical experiments

In this section, we will check the efficiency and convergence behavior of our proposed iteration functions namely, expression (2.26), expression (2.27) expression (2.28) and expression (2.29) denoted by $PM1, PM2, PM3$ and $PM4$, respectively. In this regards, we consider a total number of three test problems: first two are real life problems; third one is a standard test problem, which are mentioned in Examples 3.1 to 3.3.

Unfortunately, we don't have any optimal eighth-order iteration function for multiple zeros for the comparison. So, we consider the highest-order methods of order six (multiple zeros) for comparison which is available in the literature. Now, we compare our proposed methods with family of two-point sixth-order method which were given by Geum et al. in [19], out of them we choose the following expression:

$$\begin{aligned} y_n &= x_n - m\frac{f(x_n)}{f'(x_n)},\ m>1, \\ x_{n+1} &= y_n - m\left[1+2(m-1)(u_n-s_n)-4u_n s_n+s_n^2\right]\frac{f(y_n)}{f'(y_n)}, \end{aligned} \tag{3.1}$$

 ISBN: 978-84-617-8694-7

called $GM1$.

Finally, we compare them with the non optimal family of sixth-order methods based on weight function approach presented by the same authors Geum et al. [20], out of them we consider the following expression (11)

$$\begin{aligned} y_n &= x_n - m\frac{f(x_n)}{f'(x_n)},\ m \geq 1, \\ w_n &= x_n - m\left[1 + u_n + 2u_n^2\right]\frac{f(x_n)}{f'(x_n)}, \\ x_{n+1} &= x_n - m\left[1 + u_n + 2u_n^2 + (1+2u_n)v_n\right]\frac{f(x_n)}{f'(x_n)}, \end{aligned} \tag{3.2}$$

denoted by $GM2$.

In the numerical tests presented in Tables 1 to 3, we have compared our methods with the known ones on the basis of approximated zeros, residual error of the involved functions, difference between the two consecutive iterations, asymptotic error constants. In Tables 1 – 3, we display the number of iteration indices (n), approximated zeros (x_n), absolute residual error of the corresponding function ($|f(x_n)|$), error in the consecutive iterations $|x_{n+1} - x_n|$, computational order of convergence $\left(\rho = \frac{\log|(x_{n+1}-x_n)/(x_n-x_{n-1})|}{\log|(x_n-x_{n-1})/(x_{n-1}-x_{n-2})|},\ n = 2, 3\right)$ (the details of this formula can be seen in Cordero and Torregrosa [22]), $\left|\frac{x_{n+1}-x_n}{(x_n-x_{n-1})^p}\right|$ (where p is either 6 or 8 corresponding to the considered iteration function), the estimation of asymptotic error constant $\eta \approx \lim_{n\to\infty}\left|\frac{x_{n+1}-x_n}{(x_n-x_{n-1})^p}\right|$ at the last iteration. We make our calculations with several number of significant digits (minimum 3000 significant digits) to minimize the roundoff error.

As we mentioned in the above paragraph, we calculate the values of all the constants and functional residuals up to several number of significant digits but we display the value of x_n up to 25 significant digits. In addition, we also display $\left|\frac{x_{n+1}-x_n}{(x_n-x_{n-1})^p}\right|$ and η up to 10 significant digits. Moreover, absolute residual error in the function $|f(x_n)|$ and error in the consecutive iterations $|x_{n+1} - x_n|$ are displayed up to 2 significant digits with exponent power which are mentioned in Tables 1 – 3. Finally, computational order of convergence is up to 5 significant digits. Furthermore, the approximated zeros up to 25 significant digits are also displayed in the Examples 3.1– 3.3 although minimum 3000 significant digits are available with us.

For these numerical tests, all computations have been performed using the programming package $Mathematica$ 11 with multiple precision arithmetic. Further, the meaning of $a(\pm b)$ is $a \times 10^{(\pm b)}$ in Tables 1–3.

 ISBN: 978-84-617-8694-7

Example 3.1 Population growth problem:
Law of population growth is defined as follows:

$$\frac{dN(t)}{dt} = \gamma N(t) + \eta, \tag{3.3}$$

where N(t) = population at time t, η= fixed/constant immigration rate and γ = fixed/constant birth rate of population. We can easily obtain the following solution of the above differential equation (3.3)

$$N(t) = N_0 e^{\gamma t} + \frac{\eta}{\gamma}(e^{\gamma t - 1}), \tag{3.4}$$

where N_0 is initial population.
For a particular case study, the problem is given as: Suppose a certain population contains 1000000 individuals initially, that 300000 individuals immigrate into the community in the first year and that 1365000 individuals are present at the end of one year. Find birth rate (γ) of this population.
To determine the birth rate, we must solve the equation

$$f_1(x) = 1365 - 1000e^x - \frac{300}{x}(e^x - 1). \tag{3.5}$$

wherein $x = \gamma$ and our desired zero of the above function f_1 is 0.055046224513351778274834 21.

Example 3.2 Van der Waals equation of state

$$\left(P + \frac{a_1 n^2}{V^2}\right)(V - na_2) = nRT,$$

explains the behavior of a real gas by introducing in the ideal gas equations two parameters, α_1 and α_2, specific for each gas. The determination of the volume V of the gas in terms of the remaining parameters requires the solution of a nonlinear equation in V

$$PV^3 - (na_2P + nRT)V^2 + \alpha_1 n^2 V - \alpha_1\alpha_2 n^2 = 0.$$

Given the constants α_1 and α_2 of a particular gas, one can find values for n, P and T, such that this equation has a three simple roots. By using the particular values, we obtain the following nonlinear function

$$f_2(x) = x^3 - 5.22x^2 + 9.0825x - 5.2675.$$

have three zeros and out of them one is a multiple zero $\alpha = 1.75$ of multiplicity of order two and other one simple zero $\alpha = 1.72$. However, our desired root is $\alpha = 1.75$.

 ISBN: 978-84-617-8694-7

Example 3.3 Let us consider the following standard nonlinear test function from Behl et al. [16]

$$f_3(x) = \left(-\sqrt{1-x^2} + x + \cos\left(\frac{\pi x}{2}\right) + 1\right)^3 \tag{3.6}$$

The above function has a multiple zero at $\xi = -0.72858404644482671671233310 2423$ of multiplicity 3.

4 Conclusions

In this study, we proposed an optimal scheme for multiple zeros with known multiplicity $m \geq 1$ having eighth-order convergence, for the first time according to our best knowledge. A detail convergence analysis is presented which demonstrate that the proposed scheme has attained eighth-order convergence theoretically. In addition, the proposed scheme also satisfy the classical Kung-Traub conjecture. So, each member of the scheme is optimal. The main advantage of the proposed iteration functions is that they have not only minimum residual errors and smaller absolute errors difference between two consecutive iterations corresponding to the listed test functions. But, they also show the stable computational order of convergence as compared to the other listed methods. Further, the computational efficiency index of the proposed scheme is $E = \sqrt[4]{8} \approx 1.682$ which is better than the efficiency index of Newton's method $E = \sqrt[2]{2} \approx 1.414$ and also the schemes proposed by Thukral [18] and Guem et al. [19, 20], $E = \sqrt[4]{6} \approx 1.565$. Moreover, we can easily develop many new interesting and optimal iteration functions having eighth-order convergence with the different choices of weight functions. Finally, we can claim on the basis of obtained numerical results that our proposed iteration functions are highly efficient and perform better than the existing robust methods.

References

[1] A. M. Ostrowski, *Solution of equations and systems of equations*, Academic Press, New York 1960.

[2] J. F. Traub, *Iterative methods for the solution of equations*, Prentice-Hall, Englewood Cliffs, 1964.

[3] M. S. Petković, B. Neta, L. D. Petković, J. Džunić, *Multipoint methods for solving nonlinear equations*, Academic Press, 2013.

[4] S. Li, X. Liao, L. Cheng, *A new fourth-order iterative method for finding multiple roots of nonlinear equations*, Appl. Math. Comput. **215** (2009) 1288–1292.

 ISBN: 978-84-617-8694-7

Table 1: Convergence behavior of different iterative methods on the test function $f_1(x)$

Cases	n	x_n	$\lvert f(x_n)\rvert$	$\lvert x_{n+1}-x_n\rvert$	ρ	$\frac{x_{n+1}-x_n}{(x_n-x_{n-1})^p}$	η
$GM1$	0	0.5	*	*			
	1	*	*	*		*	*
	2	*	*	*		*	
	3	*	*	*	*	*	
$GM2$	0	0.5	6.7(+2)	4.4(−1)			
	1	0.056338848506252073641397777	1.6	1.3(−3)		1.694976965(−1)	5.781466312(−1)
	2	0.055046224513351780961365 85	3.3(−15)	2.7(−18)		5.759153879(−1)	
	3	0.055046224513351778274834 21	2.6(−103)	2.2(−106)	5.9999	5.781466312(−1)	
$PM1$	0	0.5	6.7(+2)	4.4(−1)			
	1	0.055064035851103727975441 88	2.2(−2)	1.8(−5)		4.544713334(−4)	9.825483869(−157)
	2	0.055046224513351778274834 21	1.9(−36)	1.6(−39)		1.573201848(−20)	
	3	0.055046224513351778274834 21	7.5(−309)	6.2(−312)	8.0000	9.825483869(−157)	
$PM2$	0	0.5	6.7(+2)	4.4(−1)			
	1	0.055121754923328292519265 10	9.2(−2)	7.6(−5)		1.928222590(−3)	1.151040615(−135)
	2	0.055046224513351778274834 21	3.2(−21)	2.6(−34)		8.029691000(−18)	
	3	0.055046224513351778274834 21	6.5(−267)	5.4(−270)	8.0000	1.151040615(−135)	
$PM3$	0	0.5	6.7(+2)	4.4(−1)			
	1	0.055082910761477288152310 29	4.4(−2)	3.7(−5)		9.362394780(−4)	9.344020705(−147)
	2	0.055046224513351778274834 21	6.0(−34)	5.0(−37)		2.749508569(−19)	
	3	0.055046224513351778274834 21	7.0(−289)	5.7(−292)	8.0000	9.344020705(−147)	
$PM4$	0	0.5	6.7(+2)	4.4(−1)			
	1	0.055048025037086791920845 18	2.2(−3)	1.8(−6)		4.593527235(−5)	5.201135168(−198)
	2	0.055046224513351778274834 21	2.7(−46)	2.2(−49)		2.139253203(−26)	
	3	0.055046224513351778274834 21	1.6(−389)	1.3(−392)	8.0000	5.201135168(−198)	

(∗ means corresponding iterative method fails.)

Table 2: Convergence behavior of different iterative methods on the test function $f_2(x)$

Cases	n	x_n	$\lvert f(x_n)\rvert$	$\lvert x_{n+1}-x_n\rvert$	ρ	$\frac{x_{n+1}-x_n}{(x_n-x_{n-1})^p}$	η
$GM1$	0	1.8	2.0(−4)	4.9(−2)			
	1	1.750807526326236235652918	2.0(−8)	8.1(−4)		5.698554630(+4)	1.929012344(+7)
	2	1.750000000004488220780712	6.0(−25)	4.5(−12)		1.618579697(+7)	
	3	1.750000000000000000000000	7.5(−124)	1.6(−61)	5.9908	1.929012344(+7)	
$GM2$	0	1.8	2.0(−4)	4.9(−2)			
	1	1.751050232397918097498091	3.4(−8)	1.1(−3)		7.634517351(+4)	4.597479367(+7)
	2	1.750000000047058565701337	6.6(−23)	4.7(−11)		3.506923840(+7)	
	3	1.750000000000000000000000	7.5(−111)	5.0(−55)	5.9840	4.597479367(+7)	
$PM1$	0	1.8	2.0(−4)	4.9(−2)			
	1	1.750252237254488174294623	1.9(−9)	2.5(−4)		4.118271865(+1)	9.320484120(−87)
	2	1.750000000000000000018034	9.8(−42)	1.8(−20)		4.455165313(−6)	
	3	1.750000000000000000000000	2.9(−332)	9.9(−166)	8.9970	9.320484120(−87)	
$PM2$	0	1.8	2.0(−4)	4.9(−2)			
	1	1.750272652422040846473896	2.3(−9)	2.7(−4)		4.458904599(+1)	4.894866302(−62)
	2	1.750000000000000001063165	3.4(−38)	1.1(−18)		1.923814224(−4)	
	3	1.750000000000000000000000	1.2(−268)	6.3(−134)	7.9971	4.894866302(−62)	
$PM3$	0	1.8	2.0(−4)	4.9(−2)			
	1	1.750256304692013573664840	2.0(−9)	2.6(−4)		4.186049682(+1)	2.649021470(−66)
	2	1.750000000000000000140097	5.9(−40)	1.4(−19)		3.246400883(−5)	
	3	1.750000000000000000000000	3.1(−284)	1.0(−141)	8.0026	2.649021470(−66)	
$PM4$	0	1.8	2.0(−4)	4.9(−2)			
	1	1.750153658734633762425367	7.1(−10)	1.5(−4)		2.488995472(+1)	4.945789972(−74)
	2	1.750000000000000000001714	8.8(−44)	1.7(−21)		3.074155534(−6)	
	3	1.750000000000000000000000	5.5(−315)	4.3(−157)	7.9990	4.945789972(−74)	

Table 3: Convergence behavior of different iterative methods on the test function $f_3(x)$.

Cases	n	x_n	$\lvert f(x_n)\rvert$	$\lvert x_{n+1}-x_n\rvert$	ρ	$\frac{x_{n+1}-x_n}{(x_n-x_{n-1})^p}$	η
	0	-0.5	$4.0(-2)$	$2.3(-1)$			
$GM1$	1	$-0.72857684585826792541127 39$	$9.5(-16)$	$7.2(-6)$		$5.048640676(-2)$	$1.291686374(+1)$
	2	$-0.72858404644482671671233 31$	$1.5(-89)$	$1.8(-30)$		$1.2.1444621(+1)$	
	3	$-0.72858404644482671671233 31$	$2.2(-532)$	$4.4(-178)$	6.0000	$1.291686374(+1)$	
	0	-0.5	$4.0(-2)$	$2.3(-1)$			
$GM2$	1	$-0.72856010289168166215402 63$	$3.5(-14)$	$2.4(-5)$		$1.679523409(-1)$	$2.241021157(+1)$
	2	$-0.72858404644482671671233 31$	$1.9(-79)$	$4.2(-27)$		$2.239528095(+1)$	
	3	$-0.72858404644482671671233 31$	$5.2(-471)$	$1.3(-157)$	6.0000	$2.241021157(+1)$	
	0	-0.5	$4.0(-2)$	$2.3(-1)$			
$PM1$	1	$-0.72858358235275882476135 16$	$2.6(-19)$	$4.6(-7)$		$1.699902478(-4)$	$3.659559514(-195)$
	2	$-0.72858404644482671671233 31$	$2.2(-147)$	$9.5(-50)$		$2.055108515(-24)$	
	3	$-0.72858404644482671671233 31$	$7.0(-1172)$	$3.0(-391)$	8.0000	$3.659559514(-195)$	
	0	-0.5	$4.0(-2)$	$2.3(-1)$			
$PM2$	1	$-0.72858358328243489974375 84$	$2.5(-19)$	$4.6(-7)$		$1.696497181(-4)$	$4.154136143(-194)$
	2	$-0.72858404644482671671233 31$	$9.4(-147)$	$1.5(-49)$		$3.356871797(-24)$	
	3	$-0.72858404644482671671233 31$	$3.4(-1166)$	$2.4(-389)$	8.0000	$4.154136143(-194)$	
	0	-0.5	$4.0(-2)$	$2.3(-1)$			
$PM3$	1	$-0.72858358520128916059336 90$	$2.5(-19)$	$4.6(-7)$		$1.689468637(-4)$	$6.330366939(-195)$
	2	$-0.72858404644482671671233 31$	$3.0(-147)$	$1.1(-49)$		$2.327297308(-24)$	
	3	$-0.72858404644482671671233 31$	$1.2(-1170)$	$7.8(-391)$	8.0000	$6.330366939(-195)$	
	0	-0.5	$4.0(-2)$	$2.3(-1)$			
$PM4$	1	$-0.72858360221533466225255 81$	$2.2(-19)$	$4.4(-7)$		$1.627148155(-4)$	$2.129227824(-196)$
	2	$-0.72858404644482671671233 31$	$3.3(-148)$	$5.0(-50)$		$1.292366015(-24)$	
	3	$-0.72858404644482671671233 31$	$6.5(-1179)$	$1.4(-393)$	8.0000	$2.129227824(-196)$	

[5] J. R. Sharma, R. Sharma, *Modified Jarratt method for computing multiple roots*, Appl. Math. Comput. **217** (2010) 878–881.

[6] S. G. Li, L.Z. Cheng, B. Neta, *Some fourth-order nonlinear solvers with closed formulae for multiple roots*, Comput. Math. Appl. **59** (2010) 126–135.

[7] X. Zhou, X. Chen, Y. Song, *Constructing higher-order methods for obtaining the muliplte roots of nonlinear equations*, J. Comput. Math. Appl. **235** (2011) 4199–4206.

[8] M. Sharifi, D. K. R. Babajee, F. Soleymani, *Finding the solution of nonlinear equations by a class of optimal methods*, Comput. Math. Appl. **63** (2012) 764–774.

[9] F. Soleymani, D. K. R. Babajee, T. Lofti, *On a numerical technique forfinding multiple zeros and its dynamic*, J. Egypt. Math. Soc. **21** (2013) 346–353.

[10] F. Soleymani, D. K. R. Babajee, *Computing multiple zeros using a class of quartically convergent methods*, Alex. Eng. J. **52** (2013) 531–541.

[11] B. Liu, X. Zhou, *A new family of fourth-order methods for multiple roots of nonlinear equations*, Non. Anal. Model. Cont. **18(2)** (2013) 143–152.

 ISBN: 978-84-617-8694-7

[12] X. Zhou, X. Chen, Y. Song, *Families of third and fourth order methods for multiple roots of nonlinear equations*, Appl. Math. Comput. **219** (2013) 6030–6038.

[13] R. Thukral, *A new family of fourth-order iterative methods for solving nonlinear equations with multiple roots*, J. Numer. Math. Stoch. **6 (1)** (2014) 37–44.

[14] R. Behl, A. Cordero, S.S. Motsa, J. R. Torregrosa, *On developing fourth-order optimal families of methods for multiple roots and their dynamics*, Appl. Math. Comput. **265(15)** (2015) 520–532.

[15] J. L. Hueso, E. Martínez, C. Teruel, *Determination of multiple roots of nonlinear equations and applications*, J. Math. Chem. **53** (2015) 880-892.

[16] R. Behl, A. Cordero, S. S. Motsa, J. R. Torregrosa, V. Kanwar, *An optimal fourth-order family of methods for multiple roots and its dynamics*, Numer. Algor. **71 (4)** (2016) 775–796.

[17] B. Neta, *Extension of Murakami's high-order non-linear solver to multiple roots*, Int. J. Comput. Math. **87(5)** (2010) 1023–1031.

[18] R. Thukral, *Introduction to higher-order iterative methods for finding multiple roots of nonlinear equations*, J. Math. **2013** (2013) Article ID 404635, 3 pages http://dx.doi.org/10.1155/2013/404635.

[19] Y. H. Geum, Y .I. Kim, B. Neta, A class of two-point sixth-order multiple-zero finders of modified double-Newton type and their dynamics, Appl. Math. Comput. **270** (2015) 387–400 .

[20] Y. H. Geum, Y. I. Kim, B. Neta, *A sixth-order family of three-point modified Newton-like multiple-root finders and the dynamics behind their extraneous fixed points*, Appl. Math. Comput. **283** (2016) 120–140.

[21] H. T. Kung, J. F. Traub, *Optimal order of one-point and multipoint iteration*, J. Assoc. Comput. Mach. **21** (1974) 643–651.

[22] A. Cordero, J. R.Torregrosa, *Variants of Newton's method using fifth-order quadrature formulas*, Appl. Math. Comput. **190 (1)** (2007) 686–698 .

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

Evaluating Sound Source Localization on Multi and Many-core Platforms

Jose A. Belloch[1], Jose M. Badia[1], Francisco D. Igual[2], Maximo Cobos[3] and Enrique S. Quintana-Ortí[1]

[1] *Depto. de Ingeniería y Ciencia de Computadores, Universitat Jaume I de Castelló, Spain*

[2] *Depto. de Arquitectura de Computadores y Automática, Universidad Complutense de Madrid, Spain*

[3] *Computer Science Department, Universitat de València, Universitat de València, Spain*

emails: jbelloch@uji.es, badia@uji.es, figual@ucm.es, maximo.cobos@uv.es, quintana@uji.es

Abstract

Sound source localization is an important topic in the field of audio signal processing, with multiple applications such as automatic camera steering systems, humanmachine interaction, video gaming and audio surveillance. The Steered Response Power with Phase Transform (SRP-PHAT) algorithm is a well-known method for sound source localization due to its robust performance in noisy and reverberant environments. SPR-PHAT implementations require to handle a high number of signals coming from a microphone array and a huge search grid that influences the localization accuracy of the system. In this context, high performance in the localization process can only be achieved using massively parallel computational resources. Different types of multi-core machines based on either multiple CPUs or GPUs are commonly employed in diverse fields of science for accelerating a number of applications. In this work, we evaluate and analyze OpenCL-based and OpenMP-based implementations on two different parallel platforms. Our results show that the proposed implementations achieve close-to-peak speedup on multi-CPU platforms. Furthermore, the OpenCL solution offers higher performance on GPU than on CPU.

Key words: Sound source localization, Audio processing, Microphone arrays, multi-core platforms, many-core platforms.

ISBN: 978-84-617-8694-7

1 Introduction

The localization of sound sources in acoustic environments with noise and reverberation is a challenging task for applications arising in a wide range of areas. Applications such as acoustic-based surveillance, gaming, spatial sound and virtual reality can be highly improved under a location-aware system [1].

In order to locate a sound source, it is necessary to process the input signals captured by a set of microphones in real time. The microphones can be arranged either in a distributed configuration or by following a specific geometry. The processing is usually based on the computation of the *Generalized Cross-Correlation* (GCC) functions [2, 3] of all the microphone pairs in the system. GCCs are usually obtained from the inverse Fourier transform of the cross-power spectral density of the microphone signals, multiplied by a proper spectral weighting function. A well-known algorithm that makes use of GCCs is the *Steered Response Power - Phase Transform* (SRP-PHAT). Indeed, this is the preferred choice due to its robustness in noisy and reverberant environments. Moreover, the SRP-PHAT method exhibits massive fine-grain parallelism, that is, the same operations are performed over different sets of data. Usually, these different data sets correspond to the audio samples of the different audio channels involved in the system. Basically, the SRP-PHAT algorithm evaluates a functional over a fine spatial grid and identifies its maximum value as the most likely source position.

The fact that current processors are composed of multiple cores and general-purpose GPUs has allowed to accelerate multiple applications in distinct fields, including audio signal processing [8]. In fact, there exists a high variety of hardware, CPUs and GPUs, and programming technologies, OpenMP [4] and OpenCL [5], that can be used for this purpose. However, deciding which is the most suitable option is not straightforward. To this end, in this work we propose an OpenCL-based implementation that runs on a multi-core CPU and a GPU, and an OpenMP-based implementation that is executed on the multi-core CPU.

In the next section, we describe briefly the SRP-PHAT algorithm. Section 3 is devoted to implementation issues and the performance analysis. Finally, Section 4 provides some conclusion remarks.

2 The SRP-PHAT Algorithm

Consider the output from a microphone l, $m_l(t)$, in a system composed of S microphones. Then, the SRP at the spatial point $\mathbf{x} = [x, y, z]^T$ for a time frame n of length T is defined as

$$P_n(\mathbf{x}) \equiv \int_{nT}^{(n+1)T} \left| \sum_{l=1}^{S} w_l m_l \left(t - \tau(\mathbf{x}, l)\right) \right|^2 dt, \tag{1}$$

 ISBN: 978-84-617-8694-7

where w_l is a weight and $\tau(\mathbf{x}, l)$ is the direct time of travel from location $\mathbf{x}$ to microphone l. DiBiase [6] showed that the SRP can be computed by summing up the GCCs for all possible pairs of the set of microphones in the system. In particular, the GCC for a microphone pair (k, l) is computed as

$$R_{m_k m_l}(\tau) = \int_{-\infty}^{\infty} \Phi_{kl}(\omega) M_k(\omega) M_l^*(\omega) e^{j\omega\tau} d\omega, \tag{2}$$

where τ is the time lag, * denotes complex conjugation, $M_l(\omega)$ is the Fourier transform of the microphone signal $m_l(t)$, and $\Phi_{kl}(\omega)$ is a combined weighting function in the frequency domain. The Phase Transform (PHAT) [2] has been demonstrated to be a suitable GCC weighting for time delay estimation in reverberant environments:

$$\Phi_{kl}(\omega) \equiv \frac{1}{|M_k(\omega) M_l^*(\omega)|}. \tag{3}$$

Taking into account the symmetries involved in the computation of Eq. (1), and removing some fixed energy terms [6], the part of $P_n(\mathbf{x})$ that changes with $\mathbf{x}$ is isolated as

$$P'_n(\mathbf{x}) = \sum_{k=1}^{S} \sum_{l=k+1}^{S} R_{m_k m_l}\left(\tau_{kl}(\mathbf{x})\right), \tag{4}$$

where $\tau_{kl}(\mathbf{x})$ is the *Inter-Microphone Time-Delay Function* (IMTDF). This function is very important, since it represents the theoretical direct path delay for the microphone pair (k, l) resulting from a point source located at $\mathbf{x}$. The IMTDF is mathematically expressed as [7]

$$\tau_{kl}(\mathbf{x}) = \frac{\|\mathbf{x} - \mathbf{x}_k\| - \|\mathbf{x} - \mathbf{x}_l\|}{c}, \tag{5}$$

where c is the speed of sound, and $\mathbf{x}_k$ and $\mathbf{x}_l$ are the microphone location vectors.

All in all, the SRP-PHAT algorithm consists in evaluating the functional $P'_n(\mathbf{x})$ on a fine grid G with the aim of finding the point-source location $\mathbf{x}_s$ that provides the maximum value:

$$\hat{\mathbf{x}}_s = \arg\max_{\mathbf{x} \in G} P'_n(\mathbf{x}). \tag{6}$$

Figure 1(a) shows schematically the intuition behind SRP-PHAT localization. In this figure, an anechoic environment is assumed so that the GCC for each microphone pair is a delta function located at the real TDOA (Time Difference of Arrival). Each TDOA defines a half-hyperboloid of potential source locations. The intersection resulting from all the half-hyperboloids matches the point of the grid having the largest accumulated value.

 ISBN: 978-84-617-8694-7

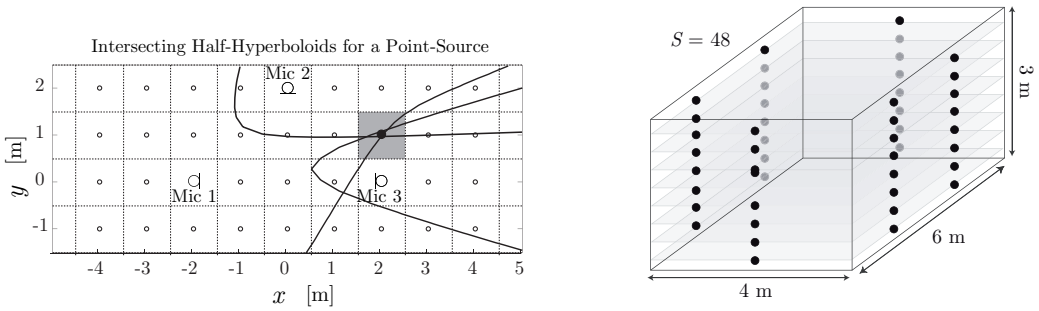


Figure 1: (a) Intersecting half-hyperboloids for $M = 3$ microphones. Each half-hyperboloid corresponds to a TDOA peak in the GCC. (b) Microphone set-up for 48 microphones.

2.1 Basic implementation of the SRP-PHAT

The SRP-PHAT algorithm is usually implemented on a grid by carrying out the following steps:

1. A spatial grid G is defined with a given spatial resolution r. The theoretical delays from each point of the grid to each microphone pair are pre-computed using Eq. (5).
2. For each analysis frame, the GCC of each microphone pair using Eq. (2).
3. For each position of the grid $\mathbf{x} \in G$, the contribution of the different cross-correlations are accumulated (using delays pre-computed in step 1), as in Eq. (4).
4. Finally, the position with the maximum score is selected as in Eq. (6).

Taking into account the above steps, we have implemented several parallel versions of the algorithm using two different parallel programming technologies, namely, OpenCL an OpenMP.

3 Performance on a GPU and a Multi-core CPU

All the CPU experiments have been performed on an Intel Xeon processor E5-2695 v3 at 2.3 GB with 28 cores and 64 GB of RAM. The GPU experiments have been executed on a NVIDIA GTX1080 GPU that implements Pascal architecture. This GPU has a clock frequency of 1.771 MHz and includes 8 GB of global memory. It provides 20 OpenCL compute units and allows a maximum work-group size of 1024 work-items.

Figure 2 shows the speedups obtained with the OpenMP version of the algorithm on the 28 cores of the CPU with respect to the sequential version of the algorithm executed on

one core. The left-hand side figure shows the results using 24 microphones and varying the grid resolution. We can observe that the performance improves as we increase the resolution and, therefore, the amount of parallel computations to perform. The right-hand side plot in the same figure, obtained with the maximum resolution ($r = 0.005$), shows that we obtain very good speedups for any number of microphones.

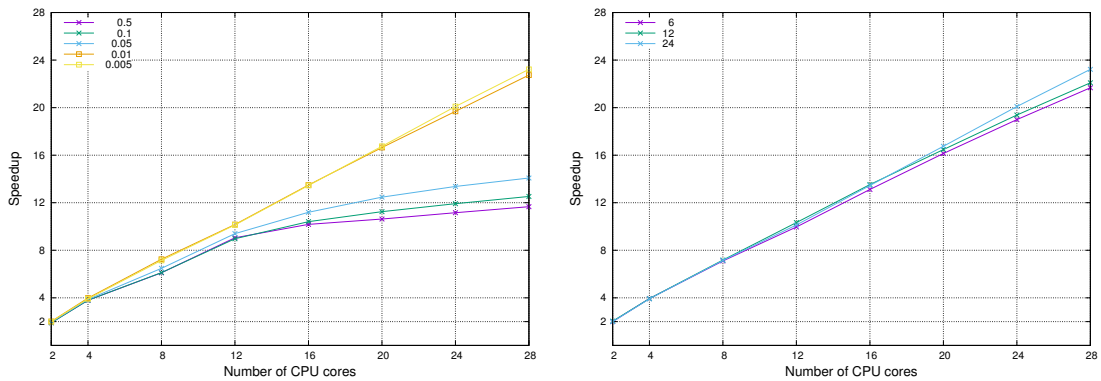


Figure 2: Speedups obtained with the OpenMP algorithm on the CPU.

We have evaluated the OpenCL implementation of the algorithm both on the CPU and the GPU. Figure 3 shows the runtime of this version of the algorithm using the 28 cores of the CPU and 24 microphones. We can observe that, for relatively small resolutions ($r < 0.005$), the cost of the algorithm only depends on the number of microphones. In this case most of the cost is due to the Fourier transforms and the computation of the GCC matrix, both of which depend only on the number of microphones, S, and the length of the audio frame, T. However, for larger resolutions, almost all the cost is due to the computation of the SRP matrix and so it grows very quickly with the resolution that defines the grid size. Figure 3 also shows that we can achieve real time performance even with 24 microphones and very fine resolutions (see black line in Figure 3).

Finally, we compare the speedups obtained with both versions of the algorithm (OpenMP and OpenCL) and using both parallel platforms (CPU and GPU) (in Figure 4). The speedups have been obtained with respect to the sequential algorithm executed on one of the cores of the CPU. In this experiment, we can observe that the OpenCL algorithm executed on the GTX1080 GPU obtains much higher speedups than the same version of the algorithm executed on the CPU. This clearly proves that an state-of-the art many-core GPU is better suited to run massive data-parallel algorithms even when compared with an up-to-date multi-core CPU with 28 cores.

On the other hand, if we compare the two versions of the algorithm on the 28 cores of the CPU, we can appreciate that the OpenCL version clearly outperforms the OpenMP

 ISBN: 978-84-617-8694-7

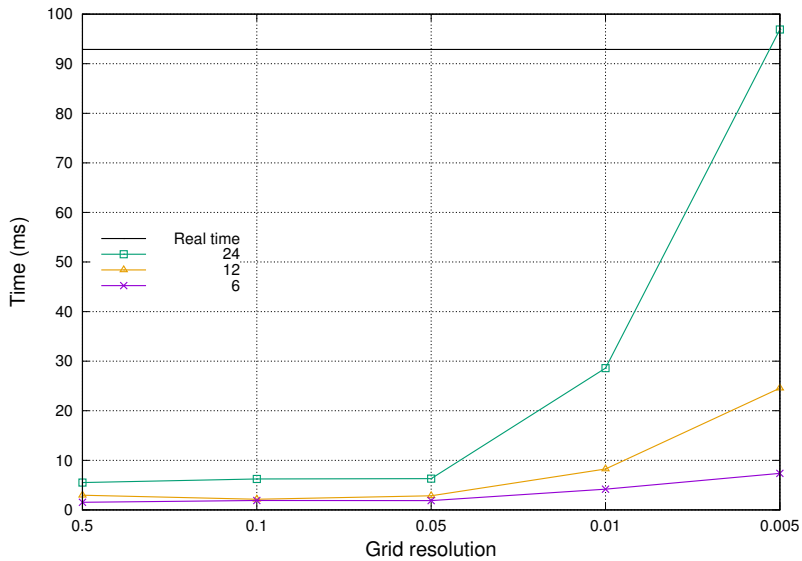


Figure 3: Time in milliseconds using the OpenCL algorithm on the CPU with a resolution r=0.005 and varying the number of microphones. The horizontal line represents the time threshold that allows to perform the localization process in real time.

version. The first version can obtain speedups much higher than the number of cores. For example, the OpenCL algorithm obtains an speedup of 45 with a grid resolution of 0.005. The OpenMP version of the algorithm delivers speedups close to the number of cores when running as many threads as cores, but is slower if we use more threads. However the OpenCL is able to run several work-items per core in parallel overlapping computations with accesses to the memory and so it reaches speedup rates that exceed the number of cores.

4 Conclusion

New emerging multi-core and many-core architectures help to overcome different computational problems in acoustic signal processing. This paper analyzed the specific case of sound source localization, where using very fine spatial resolutions or having a high number of microphones have a deep impact on the performance of real-time applications.

Currently, there are different techniques to exploit the parallel resources of the multi-core and many-core processors. In this work, we have compared performances on a 28-core CPU and a modern GPU with NVIDIA's Pascal architecture. Our results show that

 ISBN: 978-84-617-8694-7

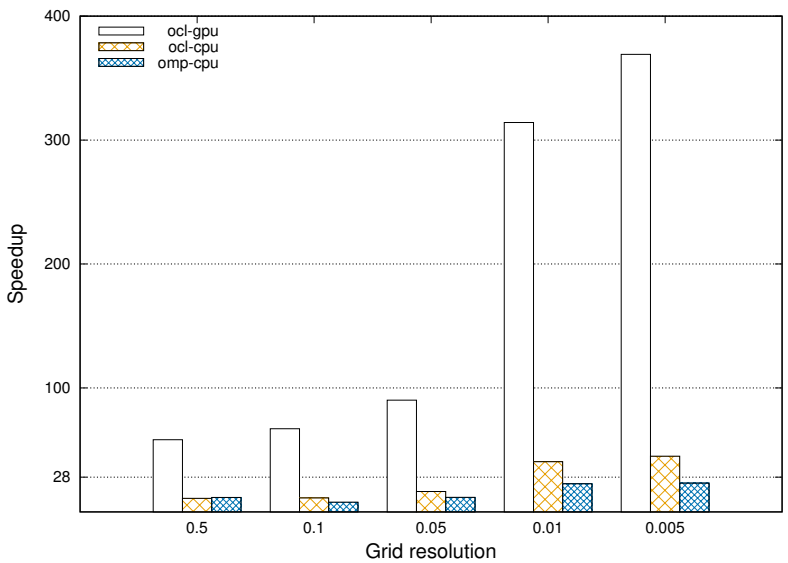


Figure 4: Comparison of the speedups of the different parallel versions of the algorithm using both platforms (GPU and GPU) and both programming technologies (OpenMP and OpenCL). We use 24 microphones and vary the grid resolution.

OpenCL on GPU clearly overcomes the implementation on CPU both using OpenCL and OpenMP. Regarding the CPU, the OpenMP-based implementation attains speedups close to the number of cores. However, this performance is clearly improved by the OpenCL version of the algorithm which is able to perform several tasks per core in parallel.

Acknowledgements

This work has been supported by the postdoctoral fellowship from Generalitat Valenciana APOSTD/2016/069, the Spanish Government through TIN2014-53495-R, TIN2015-65277-R and BIA2016-76957-C3-1-R, and the Universidad Jaume I project UJI-B2016-20.

References

[1] M. Brandstein, D. Ward, *Microphone arrays*, Springer (2001)

[2] C. H. Knapp, G. C. Carter, *The generalized correlation method for estimation of time delay*, IEEE Transactions on Acoustics, Speech and Signal Processing **27** (1976) 320–327.

[3] J. Chen, J. Benesty, Y. Huang, *Time delay estimation in room acoustic environments: An overview*, EURASIP Journal on Applied Signal Processing (2006) 1–19.

[4] B. Chapman, *Using OpenMP: Portable Shared Memory Parallel Programming*, The MIT Press (2007)

[5] M. Scarpino, *OpenCL in Action: How to Accelerate Graphics and Computation*, Manning (2012)

[6] J. H. DiBiase, *A high accuracy, low-latency technique for talker localization in reverberant environments using microphone arrays*, PhD. Thesis at Brown University Providence (2000)

[7] M. Cobos, A. Marti, J. J. Lopez, *A modified SRP-PHAT functional for robust real-time sound source localization with scalable spatial sampling*, IEEE Signal Processing Letters **18** (2011) 71–74.

[8] J. A. Belloch, M. Ferrer, A. Gonzalez, F. Martinez-Zaldivar, A. M. Vidal, *Headphone-based virtual spatialization of sound with a GPU accelerator*, Journal of the Audio Engineering Society **61** (2013) 546–561.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

Closed sets enumeration: a logical approach.

F. Benito-Picazo[1], P. Cordero[1], M. Enciso[1] and A. Mora[1]

[1] *Universidad de Málaga, Andalucía Tech, Málaga, Spain*

emails: fbenito@lcc.uma.es, pcordero@uma.es, enciso@lcc.uma.es, amora@ctima.uma.es

Abstract

Closed sets are the basis for the development of the concept lattice, a key issue in formal concept analysis. The enumeration of all the closed sets is a complex problem, having an exponential cost. In addition to the closed set, it is very useful for applications to add the information of all the minimal generators for each closed set. In this work we explain how to approach this problem from a complete set of implication by means of a sound and complete logic.

Key words: Formal concept analysis, closed sets, minimal generator, logic.

1 Introduction

Formal concept analysis (FCA) is a theoretical and practical framework to store information and manage them [GW99]. Data is stored in a table, representing a binary relation between a set of objects and attributes. The success of FCA relies on its solid theoretical framework and a wide set of methods and techniques to extract the knowledge from this data and manipulate it. One outstanding representation of the knowledge is the concept lattice, built over the closed sets, considering the subset relation as the order relation. Such representation depict a overall view of the information with a very strong formalism, opening the door to use the lattice theory as a metatheory to manage the information [BDVG17].

If-then rules have been introduced in several areas, dressed with different clothes. Thus, in relational databases [Cod71] they are named Functional Dependencies, in FCA they are named Implications and in Logic Programming (fuzzy logic) [BV06a] they are named if-then rules. All this notions captures a very intuitive idea: when the premise occurs, then the conclusion holds. Nevertheless, their semantics are very different and they further use are

 ISBN: 978-84-617-8694-7

also distinct. In this work we consider implications as elements to describe the information and we design a method to enumerate all closed sets and their minimal generators.

The proposed method is an evolution of [CEMO12], where the authors introduce a logic-based method based on $\mathbf{SL}_{FD}$, a sound and complete logic for implications. That method works by traversing the set of implications and applying a set of inference rules, following a tree paradigm in its execution. In that method, an exhaustive search was developed, producing the intended result but with an improvable performance. Here, we propose the design of several pruning strategies to improve such performance. These strategies are motivated by the idea of avoiding the opening of full branches in the tree or reducing the size of the information in their nodes.

The rest of the work is organized as follows: in the following section we present $\mathbf{SL}_{FD}$ and the axiomatic system which constitutes the basis of the MinGen algorithm. In Section 3 we present the algorithm to enumerate the closed sets and minimal generators and summarize the strategies to improve its execution in practice. The work ends with a brief conclusion.

2 Logic for implications

First, the main notions related formal concept analysis needed in this work are presented.

A **formal context** is a triple $\mathbf{K} := (G, M, I)$ where G is a set of objects, M is a set of attributes and $I \subseteq G \times M$ is a binary relation between G and M such that, for $o \in G$ and $a \in M$, $o\ I\ a$ means that the object o has the attribute a. Then, two mappings are defined $(\ )' \colon 2^G \to 2^M$ defined for all $A \subseteq G$ as $A' = \{m \in M \mid g\ I\ m \text{ for all } g \in A\}$, and $(\ )' \colon 2^M \to 2^G$ defined for all $B \subseteq M$ as $B' = \{g \in G \mid g\ I\ m \text{ for all } m \in B\}$. We use the same symbol since no confusion arises. This pair of mappings is a Galois connection.

The composition of the intent and the extent mappings, and vice versa, introduces two closure operators $(\ )'' \colon 2^G \to 2^G$ and $(\ )'' \colon 2^M \to 2^M$. The notion of closed set (as a fixpoint of a closure operator) is defined as follows:

Definition 1 *A **formal concept** is a pair (A, B) such that $A \subseteq G$, $B \subseteq M$, $A' = B$ and $B' = A$. Consequently, A and B are closed sets of objects and attributes, respectively called extent and intent.*

In this work we focus on the attributes closed sets. A key point in this work is the notion of the minimal generator (mingen) [GW99], which provides a minimal representation for each closed set, and is defined as follows:

Definition 2 *Let $\mathbf{K} = (G, M, I)$ be a formal context and $A \subseteq M$. The set of attributes A is said to be a minimal generator (**mingen**) if, for all set of attributes $X \subseteq A$ if $X'' = A''$ then $X = A$.*

 ISBN: 978-84-617-8694-7

Remark that the above definition allows to characterize each closed set by means of a minimal subset to provide a canonical representation of the closed sets. Moreover, we would like to remark that such representation is not unique, since a given closed sets can have several minimal generators.

The notion of minimal generator can also be defined from the point of view of implications. They are expressions $A \to B$ where A and B are attribute sets. A context satisfies the implication $A \to B$ if every object that has all the attributes from A also has all the attributes from B.

Definition 3 *An (attribute) implication of a formal context* $\mathbf{K} = (G, M, I)$ *is defined as a pair* (A, B)*, written* $A \to B$*, where* $A, B \subseteq M$ *and* $A \cap B = \varnothing$*. Implication* $A \to B$ *holds (is valid) in* $\mathbf{K}$ *if* $A' \subseteq B'$*.*

The set of all valid implications in a context satisfies the well-known Armstrong's axioms [Arm74], which constitutes the pioneer logic to manage implications. The author introduces a sound a complete axiomatic system to infer new implications holding in a context from a given set of implications. Moreover, this logic constitutes a proposal to solve the attribute closure, i.e. to find the maximal set of attributes A^+ such that the implication $A \to A^+$ holds. As we mentioned, this maximal set is a closed set as defined before and this closure operator $()^+$ allows us to guide the automatization the search for closed sets. Thus, we introduce a new logic suitable for this goal.

The introduction of the Simplification Logic [MECF12], named $\mathbf{SL}_{FD}$, opened the door to the development of automated reasoning methods directly based on its novel axiomatic system. $\mathbf{SL}_{FD}$ considers reflexivity as axiom scheme

$$\texttt{[Ref]} \quad \frac{}{A \to A}$$

together with the following inference rules called Fragmentation, Composition and Simplification respectively.

$$\texttt{[Frag]} \ \frac{A \to BC}{A \to B} \qquad \texttt{[Comp]} \ \frac{A \to B,\ C \to D}{AC \to BD} \qquad \texttt{[Simp]} \ \frac{A \to B,\ C \to D}{A(C \setminus B) \to D}$$

Similarly to the dual vision of closed set (in terms of Galois connection and implications), a dual definition of minimal generator can be done in terms of implications:

Definition 4 *Let* $\mathbf{K} = (G, M, I)$ *be a formal context and* $A \subseteq M$*. The set of attributes* A *is said to be a minimal generator (**mingen**) if, for all set of attributes* $X \subseteq A$ *if* $X \to A^+$ *then* $X = A$*.*

In the following section we introduce the algorithm to enumerate the minimal generator based on $\mathbf{SL}_{FD}$ and describe the strategies to improve its performance.

 ISBN: 978-84-617-8694-7

3 An algorithm to enumerate all closed sets and their minimal generators

Simplification logic has allowed us to design several executable methods to manage implications. Thus in [MECF12] we developed a novel method to compute attribute closure strongly based on $\mathbf{SL}_{FD}$ inference rules. This method has been showed to have a better performance than the classical methods based on indirect techniques. One outstanding characteristics of $\mathbf{SL}_{FD}$ closure is the output it renders: given a set of attributes $X \subseteq M$ and a set of implications Γ, it renders its closure X^+ and a new set of implications Γ' which describes the remaining knowledge in the set $M \setminus X^+$.

This logic-based closure method is the basis of another method, named MinGen, to compute the set of all minimal generators from a set of implicant set presented in [CEMO12]. The algorithm works by applying the $\mathbf{SL}_{FD}$Closure algorithm to each implication in the set, opening a new branch. This application provides a new candidate to be added to mingen and a smaller implications set which guides us in the search of new sets of attributes to be added to mingens, producing a tree-like execution.

In summary, the input of this algorithm is a set of attributes M and a set of implications Γ over the attributes in M. The output is the set of closed sets endowed with all the minimal generators, i.e. $\{\langle C, mg(C)\rangle : C \text{ is a closed set of attributes}\}$ where $mg(C) = \{D : D \text{ is a mingen and } D^+ = C\}$. In this work we only consider non-trivial minimal generators, i.e. pairs of closed set and minimal generator $\langle X, Y\rangle$ where $Y \subsetneq X$.

For example, if $M = \{a, b, c, d, e, f\}$ and $\Gamma = \{a{\to}b, bc{\to}d, de{\to}f, ace{\to}f\}$ the output is the set $\{\langle abcdef, \{ace\}\rangle, \langle abdef, \{ade\}\rangle, \langle abcde, \{ac\}\rangle, \langle bcdef, \{bce\}\rangle, \langle bcd, \{bc\}\rangle, \langle def, \{de\}\rangle, \langle ab, \{a\}\rangle, \langle c, \{c\}\rangle, \langle \varnothing, \{\varnothing\}\rangle\}$. The execution of the method is depicted in Figure 1. We refer the reader to [CEMO12] for a detailed description of the method and its theoretical results.

In this work, we propose two pruning strategies to improve MinGen method. We briefly describe them as follows:

- The first strategy characterizes the branches that can be considered a superfluous one because all their nodes explore closed sets and minimal generators already considered in another branches. To implement this strategy we will consider a subset test on the branches of the same level and in the same branch.

- The second strategy is to expedite the execution of the method by including the closure method in each node in two steps, so that in the first one the closure set is computed and, in the second one, the resulting set of implications is computed taking into account this closed set. In this way, the resulting set of implications will be a smaller one and the method will have a better performance.

 ISBN: 978-84-617-8694-7

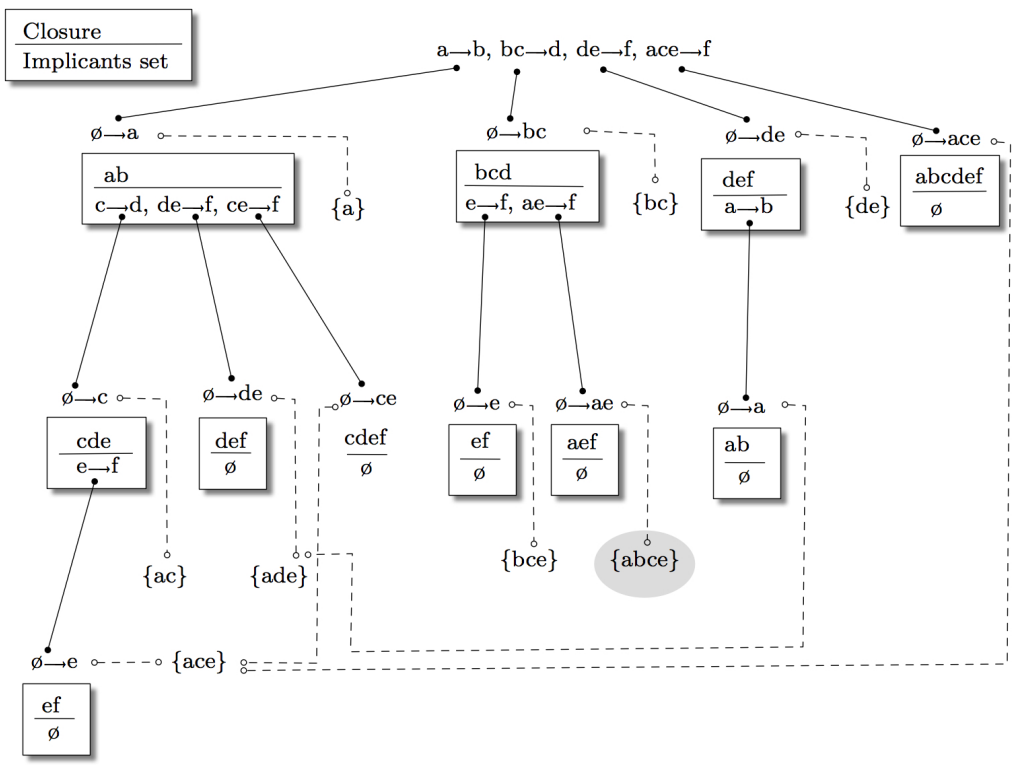


Figure 1: Exemplification of

 ISBN: 978-84-617-8694-7

4 Conclusion and future works

In this work we have studied the state of the art in the enumeration of closed sets and minimal generators based on logic. We have considered the MinGen method based on Simplification Logic as the target of our work and we propose to improve it by means of several prunes to improve its performance.

In a future work, we propose to establish the theoretical results to state these strategies and to develop an exhaustive practical experiment to show its benefits.

Acknowledgements

This work is partially supported by project TIN2014-59471-P of the Science and Innovation Ministry of Spain, co-funded by the European Regional Development Fund (ERDF).

References

[Arm74] W. Armstrong. Dependency structures of data base relationships. *Proceedings of the International Federation for Information Processing Congress*, pages 580–583, 1974.

[BV06a] R. Belohlávek and V. Vychodil. Computing non-redundant bases of if-then rules from data tables with graded attributes. In *2006 IEEE International Conference on Granular Computing*, pages 205–210, 2006.

[BDVG17] K. Bertet, C. Demko, J.-F. Viaud, and C. Guérin. Lattices, closures systems and implication bases: A survey of structural aspects and algorithms. *Theoretical Computer Science*, `http://dx.doi.org/10.1016/j.tcs.2016.11.021`, 2017.

[Cod71] E. F. Codd. Further normalization of the data base relational model. *IBM Research Report, San Jose, California*, RJ909, 1971.

[MECF12] A. Mora, M. Enciso, P. Cordero, and I. Fortes. Closure via functional dependence simplification. *International Journal of Computer Mathematics*, 89(4):510–526, 2012.

[CEMO12] P. Cordero, M. Enciso, A. Mora, and M. Ojeda-Aciego. Computing minimal generators from implications: a logic-guided approach. In *Proceedings of the Ninth International Conference on Concept Lattices and Their Applications*, pages 187–198, 2012.

[GW99] B. Ganter and R. Wille. *Formal Concept Analysis: Mathematical Foundations*. Springer-Verlag New York, Inc., 1999. Translator-C. Franzke.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

Numerical approximation for the mixed two-dimensional nonlinear Volterra-Fredholm integral equations

M. I. Berenguer[1] and D. Gámez[1]

[1] *Department of Applied Mathematics, University of Granada, Spain*

emails: maribel@ugr.es, domingo@ugr.es

Abstract

In this work we approximate the solution of the second order mixed nonlinear two-dimensional Volterra-Fredholm integral equation, by means of a new method based on the Fixed Point Theorem and making use of Schauder bases.

Key words: Two dimensional integral equations, Schauder bases, Banach spaces, Fixed Point Theorem, numerical methods.
MSC 2000: AMS 45A05, 45L05, 45N05,65R20.

1 Introduction

Several problems in engineering and physics can be modeled using nonlinear mixed two-dimensional Volterra-Fredholm integral equations of the second kind. On many occasions it is non possible to find an exact solution to these equations. Therefore, it is necessary use numerical methods for solving this type of equations. In recent years several numerical approaches have been proposed, deserving to be mentioned here, among others: the two-dimensional radial basis functions (see [1]), the Haar wavelet collocation method (see [2]), the discrete time collocation method of Brunner with a formulation of Kumar and Sloan (see [10]), the bivariate Chebyshev collocation method (see [11]), a meshless method using a radial basis function collocation scheme (see [12]), the piecewise constant two-dimensional block-pulse functions and their operational matrices (see [13]), the two-dimensional orthogonal triangular functions (see [14]), the homotopy perturbation method (see [15]), the He's variational iteration method (see [16])...

 ISBN: 978-84-617-8694-7

2 The problem and the method

In this work we develop an effective method for approximating the solution of the nonlinear mixed two–dimensional Volterra-Fredholm integral equation:

$$f(s,t) = g(s,t) + \int_{\gamma}^{t} \int_{\alpha}^{\alpha+\beta} K(s,t,x,y,f(x,y))\ dxdy \tag{1}$$

where $\alpha, \gamma \in \mathbb{R}$, $\beta, \delta \in \mathbb{R}^+$, $\Omega = [\alpha, \alpha+\beta] \times [\gamma, \gamma+\delta]$, $f \in C(\Omega)$ is the solution to be approximated, and g and K are given real-valued continuous functions defined, respectively, on Ω and $\Omega^2 \times \mathbb{R}$.

The numerical method to solve the equation (1) is based in two analytical techniques: the Banach fixed point theorem and Schauder bases in appropriate Banach spaces of continuous functions. Such tools have been used successfully in the study of certain types of one–dimensional integral and integro-differential equations (see [3], [4], [5]). We have also developed these analytical tools using the Geometric Series theorem with Schauder bases to solve numerically the linear two-dimensional Volterra integral equation (see [6]). The study of convergence and error are also described. Finally we illustrate the theoretical results with some numerical examples confirming the validity of the method employed.

Acknowledgements

Research partially supported by project MTM2016-80676-P (AEI/Feder, UE), by Junta de Andalucía Grant FQM359 and by E.T.S. Ingeniería de Edificación of the University of Granada (Spain).

References

[1] H. ALMASIED AND J. N. MELEH, *Numerical solution of a class of mixed two-dimensional nonlinear Volterra-Fredholm integral equations using multiquadric radial basis functions*, J. Comput. Appl. Math. **260** (2014) 10 pages.

[2] I. AZIZ, S. ISLAM AND F. KHAN, *A new method based on Haar wavelet for the numerical solution of two-dimensional nonlinear integral equations*, J. Comput Appl. Math. **272** (2014) 70–80.

[3] M. I. BERENGUER, AND D. GAMEZ , *Study on convergente and error of a numerical method for solving systems of nonlinear Fredholm-Volterra integral equations of Hammerstein type*, Appl. Anal. **96** (2017) 516-527.

 ISBN: 978-84-617-8694-7

[4] M. I. BERENGUER, M.V. FERNANDEZ MUÑOZ, A.I. GARRALDA GUILLEM AND M.RUIZ GALAN , *A sequential approach for solving the Fredholm integro-differential equation*, Appl. Numer. Math. **62** (2012) 297–304.

[5] M. I. BERENGUER, D. GAMEZ AND A. J. LOPEZ LINARES, *Solution of systems of integro–differential equations using numerical treatment of fixed point*, J. Comput. Appl. Math. **315** (2017), 343–353.

[6] M. I. BERENGUER, AND D. GAMEZ , *A computational method for solving a class of two dimensional Volterra integral equations*, J. Comput. Appl. Math. **318** (2017) 403-410.

[7] H. BREZIS, *Functional Analysis, Spaces and Sobolev Partial Differential Equations*, Universitext, Springer, New York, 2011.

[8] B. R. GELBAUM AND J. GIL DE LAMADRID, *Bases of tensor products of Banach spaces*, Pacific. J. Math. **11** (1961) 1281–1286.

[9] G. J. O. JAMESON, *Topology and Normed Spaces*, Chapman-Hall, London, 1974.

[10] M. HADIZADEH, *Posteriori error estimates for the nonlinear Volterra-Fredholm integral equations*, Comput. Math. Appl. **45** (2003) 677–687.

[11] M. HADIZADEH AND M. ASGARY, *An efficient numerical approximation for the linear class of mixed integral equations*, Appl. Math. Comput. **167** (2005) 1090–1100.

[12] H. LAELI, F.M. MAALEK AND M. HADIZADEH , *A meshless approximate solution of mixed Volterra-Fredholm integral equations*, Int. J. Comput. Math. **90** (2013) 527–538.

[13] K. MALEKNEJAD AND K. MAHDIANI, *Solving nonlinear mixed Volterra-Fredholm integral equations with two dimensional block-pulse functions using direct method*, Commun. Nonlinear Sci. **16** (2011) 3512–3519.

[14] K. MALEKNEJAD AND Z. JAFARI, *Applications of two-dimensional triangular functions for solving nonlinear class of mixed Volterra-Fredholm integral equations*, Math. Comput. Model. **55** (2012) 1833–1844.

[15] A. YILDIRIM, *Homotopy perturbation method for the mixed Volterra-Fredholm integral equations*, Chaos Soliton. Fract. **42** (2009) 2760–2764.

[16] S.A. YOUSEFI, A. LOTFI AND M. DEHGHAN, *He's variational iteration method for solving nonliner mexed Volterra-Fredholm integral equations*, Comput. Math. Appl. **58** (2009) 2172–2176.

 ISBN: 978-84-617-8694-7

[17] Z. SEMADENI, *Schauder Bases in Banach Spaces of Continuous Functions*, Springer–Verlag, Berlín, 1982.

[18] Z. SEMADENI, *Product Schauder bases and approximation with nodes in spaces of continuous functions*, Bull. Acad. Polon. Sci. **11** (1963) 387–391.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

Biorthogonal systems and their applications to nonlinear two–dimensional integral equations

M. I. Berenguer[1] and D. Gámez[1]

[1] *Department of Applied Mathematics, University of Granada, Spain*

emails: maribel@ugr.es, domingo@ugr.es

Abstract

Using a version of the Banach fixed-point theorem and biorthogonal systems in adequate Banach spaces, we propose an interesting method for approximating the solution of an important type of two-dimensional nonlinear integral equation.

Key words: Two dimensional integral equations, Schauder bases, Banach spaces, Fixed Point Theorem, numerical methods.
MSC 2000: AMS 45A05, 45L05, 45N05,65R20.

1 Introduction

The numerical solution of the nonlinear two-dimensional integral equations (NTDIEs) has been a subject of considerable interest. This equations arises in various physical, engineering and biological problems. In fact, few numerical methods have been known for approximating the solution of NTDIEs. For example, in [1] the authors present the definition and operation of both two-dimensional differential transformation method and their reduced form, for finding the numerical solution of two-dimensional Volterra integral equations. A method for approximating the solution of a two-dimensional second kind integral equation with a smooth kernel using a bivariate quadratic spline quasi-interpolant defined on a uniform criss-cross triangulation of a bounded rectangle is presented in [2]. In [12], the two-dimensional linear Fredholm integral equations of the second kind are solved by using two-dimensional modification of hat functions and operational matrix of integration.

In order to solve two-dimensional Volterra-Fredholm integral equations, the two-dimensional orthogonal triangular functions are used in [3], in [9] the authors presented a

 ISBN: 978-84-617-8694-7

method based on approximating unknown funcion with Berstein polynomials and in [13] the approximate solution is expressed as expansion of two-dimensional delta basis functions.

The aim of this work is to introduce a new numerical method in order to approximate the solution of the nonlinear two–dimensional integral equations of the form:

$$f(s,t) = g(s,t) + \int_{\gamma}^{t}\int_{\alpha}^{s} K(s,t,x,y,f(x,y))\ dxdy + \int_{\gamma}^{\gamma+\delta}\int_{\alpha}^{\alpha+\beta} H(s,t,x,y,f(x,y))\ dxdy \tag{1}$$

where $\alpha, \gamma \in \mathbb{R}$, $\beta, \delta \in \mathbb{R}^+$, $f(s,t)$ is an unknown continuous function to be approximated defined on $\Omega = [\alpha, \alpha+\beta]\times[\gamma, \gamma+\delta]$, $g(s,t)$ and $K(s,t,x,y,z)$ are given real-valued continuous functions defined, respectively, on Ω and $\Omega^2 \times \mathbb{R}$.

2 The numerical method and examples

Using fixed-point techniques and biorthogonal systems in adequate Banach spaces, we present a interesting numerical method for solving integral equation (1). Such tools have been used successfully in the study of certain types of one–dimensional integral and integro-differential equations (see [4], [5], [6]). Using the Geometric Series theorem and the Schauder bases we have also study the linear two-dimensional Volterra integral equation (see [7]). The study of the convergence and error will be done. Numerical examples of the performance of the method are provided, confirming its reliability and precision.

Acknowledgements

Research partially supported by project MTM2016-80676-P (AEI/Feder, UE), by Junta de Andalucía Grant FQM359 and by E.T.S. Ingeniería de Edificación of the University of Granada (Spain).

References

[1] R. ABAZARI AND A. KILIÇMAN, *Numerical study of two-dimensional Volterra integral equations by RDTM and comparison with DTM*, Abstr. Appl. Anal. **2013** (2013) 10 pages.

[2] C. ALLOUCH, P. SABLONNIRE AND D. SBIBIH, *A collocation method for the numerical solution of a two dimensional integral equation using a quadratic spline quasi-interpolant*, Numer. Algorithms **62** (2013) 445–468.

 ISBN: 978-84-617-8694-7

[3] E. BABOLIAN, K. MALEKNEJAD, M. ROODAKI AND H. ALMASIEH, *Two-dimensional triangular functions and their applications to nonlinear 2D Volterra-Fredholm integral equations*, Comput. Math. Appl. **60** (2010) 1711–1722.

[4] M. I. BERENGUER, AND D. GAMEZ , *Study on convergente and error of a numerical method for solving systems of nonlinear Fredholm-Volterra integral equations of Hammerstein type*, Appl. Anal. **96** (2017) 516-527.

[5] M. I. BERENGUER, D. GAMEZ AND A. J. LOPEZ LINARES, *Fixed–point iterative algorithm for the linear Fredholm–Volterra integro–differential equations*, J. Appl. Math. Vol. **2012** (2012) 12 pages.

[6] M. I. BERENGUER, D. GAMEZ AND A. J. LOPEZ LINARES, *Solution of systems of integro–differential equations using numerical treatment of fixed point*, J. Comput. Appl. Math. **315** (2017), 343–353.

[7] M. I. BERENGUER, AND D. GAMEZ , *A computational method for solving a class of two dimensional Volterra integral equations*, J. Comput. Appl. Math. **318** (2017) 403-410.

[8] H. BREZIS, *Functional Analysis, Spaces and Sobolev Partial Differential Equations*, Universitext, Springer, New York, 2011.

[9] M. SH. DAHAGHIN AND SH. ESKANDARI, *Solving two-dimensional Volterra-Fredholm integral equations of the second kind by using Bernstein polynomials*, Appl. Math. J. Chinese Univ. **32** (2017) 68–78.

[10] B. R. GELBAUM AND J. GIL DE LAMADRID, *Bases of tensor products of Banach spaces*, Pacific. J. Math. **11** (1961) 1281–1286.

[11] G. J. O. JAMESON, *Topology and Normed Spaces*, Chapman-Hall, London, 1974.

[12] F. MIRZAEE AND E. HADADIYAN, *Numerical solution of linear integral equations via two-dimensional modification of hat functions* , Appl. Math. Comput. **250** (2015) 805–816.

[13] F. MIRZAEE AND E. HADADIYAN, *A new numerical method for solving two-dimensional Volterra-Fredholm integral equations*, J. Appl. Math Comput. **52** (2016) 489–513.

[14] Z. SEMADENI, *Schauder Bases in Banach Spaces of Continuous Functions*, Springer–Verlag, Berlín, 1982.

[15] Z. SEMADENI, *Product Schauder bases and approximation with nodes in spaces of continuous functions*, Bull. Acad. Polon. Sci. **11** (1963) 387–391.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

A two-stage Jacobi-Davidson method with spectral preconditioners for the eigensolution of large SPD matrices

Luca Bergamaschi[1], Ángeles Martínez[2] and Filippo Zanetti[3]

[1] *Department of Civil Environmental and Architectural Engineering, University of Padua*

[2] *Department of Mathematics, University of Padua*

[3] *Department of Industrial Engineering, University of Padua*

emails: luca.bergamaschi@unipd.it, angeles.martinez@unipd.it, filippo.zanetti.4@studenti.unipd.it

Abstract

We propose a spectral preconditioner to accelerate the iterative solution of the correction equation i.e. the indefinite linear system to be solved at each step of the Jacobi-Davidson (JD) method fot the computation of the leftmost eigenpairs of large and sparse symmetric positive definite matrices. To construct the spectral preconditioner we use eigeninformation obtained by running the JD method itself at a low accuracy. The spectral preconditioner produces a clustering of the eigenvalues of the preconditioned correction equation thus speeding up the JD algorithm. Preliminary numerical tests accounts for the improving provided by the proposed approach.

Key words: eigenpairs, Jacobi-Davidson, spectral preconditioners

Extended abstract

The computation of $m \ll n$ eigenvalues of a symmetric positive definite (SPD) matrix A is a common task in many scientific applications. We will denote as $\lambda_1 < \lambda_2 < \cdots < \lambda_m < \ldots < \lambda_n$ the eigenvalues of A and $\boldsymbol{v}_1, \boldsymbol{v}_2, \cdots, \boldsymbol{v}_m, \ldots, \boldsymbol{v}_n$ the corresponding (normalized) eigenvectors. Recently in [2] a new preconditioning strategy is developed for accelerating the second stage of the DACG-Newton method [1]. The idea is to use eigenvector approximation to construct a spectral preconditioner for the efficient solution of the correction equation:

$$\begin{aligned} J_k \boldsymbol{u}_k &= -\boldsymbol{r}_k, \qquad \text{where} \\ J_k &= (I - \boldsymbol{u}_k \boldsymbol{u}_k^\top)(A - \theta_k I)(I - \boldsymbol{u}_k \boldsymbol{u}_k^\top), \quad r = -(A\boldsymbol{u}_k - \theta_k \boldsymbol{u}_k), \quad \theta_k = \frac{\boldsymbol{u}_k^\top A \boldsymbol{u}_k}{\boldsymbol{u}_k^\top \boldsymbol{u}_k} \end{aligned} \tag{1}$$

 ISBN: 978-84-617-8694-7

to be solved at each step of this projected Newton method. As the preconditioner for equation (1) a tuned preconditioner $\mathcal{P}$ is employed ([3]) i.e. a preconditioner satisfying $\mathcal{P}AV_m = V_m$, where V_m is a rectangular matrix containing an approximation of the m leftmost eigenvectors of A as columns.

In this paper we propose to use a spectral preconditioner to accelerate the correction equation (1) within the Jacobi-Davidson method. For the details of this method we refer to the original paper [5], as well as to successive works (see e.g. [4]) which analyze JD both theoretically and experimentally. Our approach is based on a twofold run of the JD solver. Namely, we first run JD to evaluate a number of (inaccurate) leftmost eigenpairs which will bw the columns of V_m. We then run a second time the JD method up to the final accuracy, using V_m both as the initial search subspace for JD and to construct the spectral preconditionerm as a low-rank update of a given inverse approximation of matrix A.

PCG solution of the correction equation

As a Krylov subspace solver for the correction equation we chose the Preconditioned Conjugate gradient (PCG) method since the Jacobian J_k has been shown to be SPD in the subspace orthogonal to $\boldsymbol{u}_k$. Regarding the implementation of PCG, we mainly refer to the work [4], where the author shows that it is possible to solve the linear system in the subspace orthogonal to $\boldsymbol{u}_k$ and hence the projection step needed in the application of J_k can be skipped. Moreover, we adopted the exit strategy for the linear system solution described in the above paper, which allows for stopping the PCG iteration, in addition to the classical exit test, whenever the l-th PCG iterate satisfies an eigenresidual test $\|\boldsymbol{r}_{k,l}\| < \tau$ or when the decrease of $\|\boldsymbol{r}_{k,l}\|$ is slower than the decrease of the residual of the linear system at step l because in this case further iterating does not improve the accuracy of the eigenvector.

Acceleration by spectral preconditioners

A class of spectral preconditioners is defined in [2] based on a very rough approximation of the sought eigenpairs. Denoted as P_0 an initial approximate inverse of A, we assume that the JD method has provided the m leftmost eigenpairs (to a low relative accuracy specified by parameter τ_0) satisfying

$$A\tilde{\boldsymbol{v}}_j = \lambda_j \tilde{\boldsymbol{v}}_j + \mathbf{res}_j, \qquad \|\mathbf{res}_j\| \leq \tau_0 \lambda_j, \ j = 1, \ldots, m, \tag{2}$$

Then, for a generic eigenvalue λ_j ($j < m$) we define the following tuned preconditioner, which will be kept constant throughout the second JD iterations to accurately compute the j-th eigenpair:

$$P_j = P_0 - W \left(W^\top A V_j \right)^{-1} W^\top, \qquad \text{with} \qquad W = P_0 A V_j - V_j \tag{3}$$

where

$$V_j = [\tilde{\boldsymbol{v}}_{j+1}, \ldots, \tilde{\boldsymbol{v}}_m] \qquad \text{and} \qquad \Lambda_j = \mathrm{diag}(\lambda_{j+1}, \ldots, \lambda_m). \tag{4}$$

 ISBN: 978-84-617-8694-7

A direct computation shows that P_j is a tuned preconditioner i.e. satisfies: $P_j A V_j = V_j$, irrespective of the error introduced by the computation of V_j. The previous relation implies that the preconditioned matrix $P_j A$ has the eigenvalue 1 with at least multiplicity $m - j$. When P_j is used to accelerate the Newton iteration, it must be projected in the space orthogonal to the previous computed eigenpairs as described in [5] (we will call $\widehat{P}_j$ the projected preconditioner).

Theorem 1 (whose proof is given in [2]) will characterize the eigenvalues of the preconditioned matrix $\widehat{P}_j J_k^{(j)}$, where index k identifies the JD iteration number in computing eigenpairs $(\lambda_j, \boldsymbol{v}_j)$.

Theorem 1 *Let matrix V_j be defined as in (4), P_j a tuned preconditioner, then each column of V_j i.e. $\tilde{\boldsymbol{v}}_s, s = j+1, \ldots, m$, is an approximate eigenvector of $P_j J_k^{(j)}$ corresponding to the approximate eigenvalue $1 - \dfrac{\theta}{\lambda_s} \approx 1 - \dfrac{\lambda_j}{\lambda_s}$. In particular the following relation holds:*

$$\widehat{P}_j J_k^{(j)} \tilde{\boldsymbol{v}}_s = \left(1 - \frac{\theta}{\lambda_s}\right) \tilde{\boldsymbol{v}}_s + \mathbf{err} \tag{5}$$

with $\|\mathbf{err}\| \leq \tau_0 \, C$, and $C \equiv C(\tau_0, \|P_j\|, \lambda_j, \lambda_{j+1})$.

Implementation

1. Limited memory implementation. We fix the maximum column size of matrix V_j, parameter $l_{\max}$.
2. Enlarging matrix V_m. A second variant consists in computing an additional number of approximated eigenpairs in the first JD stage to avoid matrix V_m to be empty when computing the $m-$th eigenpair. We introduce a further parameter, win, which counts these extra eigenpairs.

Taking into account these variants, in the computation of the j-th eigenpair we will use $V_j = [\tilde{\boldsymbol{v}}_{j+1}, \ldots, \ \tilde{\boldsymbol{v}}_{j_{end}}]$ with $j_{end} = \min\{m + \text{win}, l_{\max} + j\}$ in formula (3), namely

$$P_j = P_0 - W \left(W^\top A V_j\right)^{-1} W^\top, \qquad \text{with} \qquad W = P_0 A V_j - V_j \tag{6}$$

being $P_0 = (LL^T)^{-1}$ and $L = IC(\texttt{droptol}, A)$ an incomplete triangular Cholesky factor of A, with parameter `droptol` the threshold for dropping small elements in the factorization.

Preliminary numerical results

The preconditioned two-stage JD algorithm is tried for different values of the parameters on a model problem. We used the *MatLAB* JDQR package which can be found at the webpage: `http://www.staff.science.uu.nl/~sleij101/index.html`. This package has

 ISBN: 978-84-617-8694-7

been modified in order to allow the solution of indefinite shifted linear systems by the PCG method as described in [4] accelerated by the spectral preconditioner described so far.

We tested the proposed algorithm in the computation of the 20 smallest eigenpairs of the Laplacian matrix of order 300. The exit test is $\|A\boldsymbol{u} - q(\boldsymbol{u})\boldsymbol{u}\| \leq \tau$, with a tolerance $\tau = \tau_0 = 10^{-2}$ for the first and $\tau = \tau_1 = 10^{-12}$ for the second JD run.

Table 1: Computation of the 20 smallest eigenvalues of the Laplacian(300) with different win and l_{max} values. The initial IC preconditioner is computed using `droptol` $= 0.1$.

win	$l_{\max}$	PCG its.			JD its.		
		1st JD	2nd JD	total	1st JD	2nd JD	total
5	5	1740	3608	5348	79	155	234
5	10	1740	3535	5275	79	162	241
5	15	1740	3464	5204	79	153	232
5	20	1740	3269	5009	79	144	223
10	10	2360	3437	5797	102	166	268
10	15	2360	3544	5904	102	171	273
10	20	2360	3107	5467	102	141	243
IC precond.				8017			204

From these initial results we obtain in all cases an improvement in the total number of PCG iterations with respect to using the Incomplete Cholesky preconditioner alone. Moreover the low-rank update seems effective even when the number of eigenvectors used for updating the preconditioner is small (see Table 1 for win $= 5, l_{\max} = 5$).

References

[1] L. Bergamaschi and A. Martínez, *Efficiently preconditioned inexact Newton methods for large symmetric eigenvalue problems*, Optimization Methods & Software, 30 (2015), pp. 301–322.

[2] ——, *Two-stage spectral preconditioners for iterative eigensolvers*, Numer. Lin. Alg. Appl., 24 (2017), pp. 1–14.

[3] A. Martínez, *Tuned preconditioners for the eigensolution of large SPD matrices arising in engineering problems*, Numer. Lin. Alg. Appl., 23 (2016), pp. 427–443.

[4] Y. Notay, *Combination of Jacobi-Davidson and conjugate gradients for the partial symmetric eigenproblem*, Numer. Linear Algebra Appl., 9 (2002), pp. 21–44.

[5] G. L. G. Sleijpen and H. A. van der Vorst, *A Jacobi-Davidson method for linear eigenvalue problems*, SIAM J. Matrix Anal., 17 (1996), pp. 401–425.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

Efficient Parallel Stream Compaction on a Extremely Low-Cost SDC Cluster

Gregorio Bernabé[1] and Manuel E. Acacio[1]

[1] *Computer Engineering, University of Murcia*

emails: gbernabe@ditec.um.es, meacacio@ditec.um.es

Abstract

Many highly parallel algorithms usually generate large volumes of data containing both valid and invalid elements, and high-performance solutions to the stream compaction problem reveal extremely important in such scenarios. Although parallel stream compaction has been extensively studied in GPU-based platforms, and more recently, in the Intel Xeon Phi platform, no study has considered yet its parallelization using a low-cost computing cluster, even when general-purpose single-board computing devices are gaining popularity among the scientific community due to their high performance per $ and watt. In this work, we consider the case of an extremely low-cost cluster composed by four Odroid C2 single-board computers (SDCs), showing that stream compaction can also benefit–important speed-ups can be obtained–from this kind of platforms. To do so, we derive two parallel implementations for the stream compaction problem using MPI. Then, we evaluate them considering varying number of processes and/or SDCs, as well as different input sizes. In general, we see that unless the number of elements in the stream is too small, the best results are obtained when eight MPI processes are distributed among the four SDCs that conform the cluster.

Key words: General-purpose single-board computing (SDC); Odroid C2; Fast parallel stream compaction; Parallelization strategies; MPI; Speed-up

1 Introduction

Continuous improvements in the technologies used to build computers have recently made possible the fabrication of extremely low-cost general-purpose single-board computing devices. Nowadays, one can buy one of these *tiny* computers for a few dollars and make it run Windows 10 or Ubuntu-Linux operating systems [8]. Among the variety of vendors

providing these single-board computers (SBC), maybe the most renowned ones are Raspberry Pi and Odroid. Although the initial aim of these devices was to promote the teaching of basic computer science in schools [4] and developing countries [5], recent appearance of single-board computers with multicore CPU chips and several gigabytes of main memory have attracted interest of a multitude of projects trying to take advantage of their very low cost-performance ratio (i.e. for scientific computing [1]).

Whereas Raspberry Pi SBCs seem to have put the focus more on a "stand-alone" scenario, Odroid devices provide increased processor frequency, more main memory and higher bandwidth Ethernet capabilities. Particularly, the Raspberry Pi 3 model B that was launched in February 2016, features a 1.2 GHz, 4-core ARM Cortex-A53 CPU chip, 1 GB main memory and a 10/100 Ethernet port. Compared with its predecessor, the Raspberry Pi 2 model B released in February 2015, it adds wireless connectivity (2.4 GHz WiFi 802.11n and Bluetooth 4.1). On the contrary, the Odroid C2 sacrifices wireless connectivity in favor of higher clock frequencies (1.5 GHz, 4-core ARM Cortex-A53 CPU chip), larger main memory (2 GB) and Gigabit Ethernet connection. These characteristics make these particular devices more appropriate at building high-performance low-cost clusters able to meet the demands of some scientific applications.

On the other hand, a common characteristic found in many highly parallel algorithms is that they usually generate large volumes of data containing both valid and invalid elements. In these scenarios, high-performance solutions to the data reduction problem reveal extremely important. Stream compaction (also known as stream reduction) has been proposed to "compact" an input stream mixed with both valid and invalid elements to a subset with only the valid elements [12]. This way, stream compaction is found in many applications, that go from data mining and machine learning (in order to prune invalid nodes after each parallel breadth-first tree traversal step [11]) to deferred shading (to obtain the subset of pixels whose rays intersect, which allows for better workload balancing among the participating threads [6]).

Formally, given a list of elements $i_1, i_2, ..., i_n$ belonging to the set I and a predicate function $F : I \rightarrow \{true, false\}$, stream compaction divides I in valid and invalid elements (ones that satisfy the predicate F and others that do not), and keeps the relative order for all the valid elements in the output (O) [6]. As shown in Algorithm 1, the Serial stream compaction of I under the predicate function F is $O = \{i \epsilon I | F(i) = true\}$. Therefore, the output O simply contains all valid elements copied from the input I. An example of the execution of Algorithm 1 can be observed in Figure 1. The list of input elements is composed by numbers between 0 and 4. The Serial stream compaction selects all elements that are not zero (assuming that zero represents the invalid value), based on the predicate function F, as shown in the low part of Figure 1. Although Algorithm 1 is simple, the parallelization is not trivial because the output position of each valid element cannot be obtained until all its preceding elements have been discovered [10].

 ISBN: 978-84-617-8694-7

Algorithm 1 Serial stream compaction

```
Input: Vector I of length n
Input: Predicate function F
Output: Vector O of valid elements
Output: nvalid: the number of valid elements
1: nvalid = 0
2: for i = 0 to n − 1 do
3:     if F(I[i]) then
4:         O[nvalid + +] = I[i]
5:     end if
6: end for
```

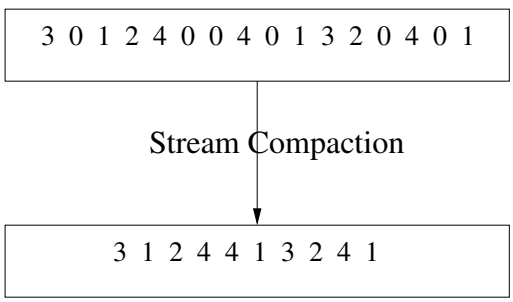


Figure 1: Example of serial stream compaction (Zero value is used to represent the invalid elements)

Parallel stream compaction has been extensively studied in GPU-based platforms [7, 12, 6, 10], and more recently parallel implementations for the Intel Xeon Phi processor have also been proposed [13]. In this work, we consider the case of an extremely low-cost cluster composed by four Odroid C2 single-board computers (SDCs), showing that stream compaction can also benefit–important speed-ups can be obtained–from this kind of platforms. To do so, we derive two parallel implementations for the stream compaction problem using MPI. Then, we evaluate them considering varying number of processes and/or SDCs, as well as different input sizes. In general, we see that unless the number of elements in the stream is too small, the best results are obtained when 8 MPI processes are distributed among the 4 SDCs that conform the cluster.

The rest of the paper is organized as follows. The parallelization strategies that we have implemented and evaluated in this work are explained in Section 2. In Section 3 we give the details of the cluster of Odroid C2 SDCs used for the evaluation, and then, we present the results. Finally, Section 4 concludes the paper and draws some lines of future work.

2 Parallelization on a cluster of Odroid C2s

In this Section we present the two parallelization strategies that we have considered in this work. In both cases, we have implemented them using MPI [3].

2.1 Parallel Stream Compaction

Algorithm 2 Parallel stream compaction

Input: Vector I of length n
Input: Predicate function F
Input: Number of processes p
Input: pid of process
Output: Vector O of valid elements
Output: nvalid: the number of valid elements
Output: pos: position to write

```
1:  nvalid = 0
2:  tamp = n/p
3:  scan[0 : (tamp − 1)] = 0
4:  for i = 0 to tamp − 1 in parallel do
5:      if F(I[i]) then
6:          temp[i] = 1
7:      end if
8:  end for
9:  for i = 0 to tamp − 1 in parallel do
10:     scan[i] = scan[i − 1] + temp[i − 1]
11: end for
12: nvalid = scan[tamp − 1] + temp[tamp − 1]
13: for i = pid to p − 1 in parallel do
14:     Send nvalid to process[i+1]
15: end for
16: if pid > 0 then
17:     for i = 0 to pid in parallel do
18:         Receive nvalid[i]
19:         pos = pos + nvalid[i]
20:     end for
21: end if
22: for i = 0 to tamp − 1 in parallel do
23:     if temp[i] then
24:         O[pos + scan[i]] = I[i]
25:     end if
26: end for
```

We have based on the implementation proposed in the Thrust library [9] to develop the Parallel stream compaction scheme shown in Algorithm 2. A vector of a particular length, the predicate function, the number of processes and the pid of each process are the inputs. We have divided Algorithm 2 in four phases namely: *Validation* phase (lines 4-8), *Scan* phase (lines 9-12), *Communication* phase (lines 13-21) and *Scatter* phase (lines 22-26). During the *Validation* phase, the input vector (I) is examined in parallel, and taking into consideration the predicate function, each process annotates the validity of each of its assigned elements in array *temp* (representing 1 a valid element and 0 an invalid one). The parallel *Scan* phase needs an additional array (*scan*) to compute the so called prefix-sum [2], where each element is the addition of all its preceding elements excluding itself. So, each process obtains in parallel the number of valid elements (*nvalid*) in its portion of the stream. Following this, in the *Communication* phase each process, identified by a pid, sends the number of valid elements that it has found to all the processes with higher pids. All the processes, except the first one, receive the number of valid elements and compute the position (*pos*) of the first of their valid elements. Finally, during the *Scatter* phase, based on the *scan* and *temp* arrays, all valid elements are transferred from the input array to the output one (I and O, respectively), preserving the order in which these elements appear in

 ISBN: 978-84-617-8694-7

the input array.

Figure 2 shows an example of an execution with four MPI processes for a list of input elements composed by numbers ranging between 0 and 4. In this case the predicate function F selects all elements that are not zero. Now, the input vector of length 16 positions is divided among the four MPI processes (*P0, P1, P2 and P3*). All the processes carry out the *Validation* and *Scan* phases in parallel. The position (*pos*) computed by each process is shown below the vector *scan*. Finally, the output O is built taking into account the *temp* and *scan* vectors, as well as the *pos*, previously computed.

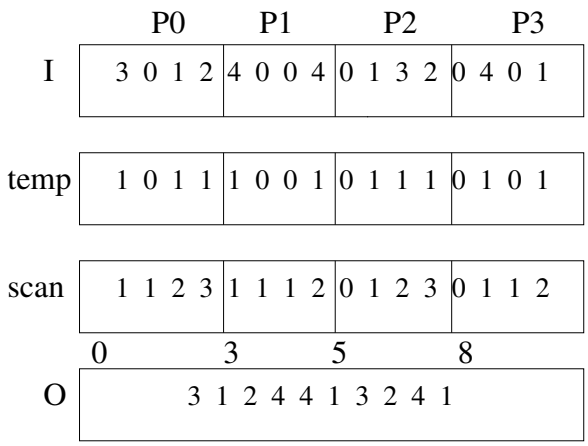


Figure 2: Example of Parallel stream compaction

2.2 Parallel Work-Efficient Stream Compaction

In [13], it is presented a work-efficient stream compaction algorithm aimed at improving the computing complexity of the Parallel stream compaction that was shown in Algorithm 2. Again, using MPI, we have developed the parallel version of this work-efficient stream compaction and we show it in Algorithm 3. Now, during the *Validation* phase (lines 5-10), each process saves the validity of each element on the array *scan* and stores the number of valid elements on the vector V. Therefore, the additional array of integers (*temp*) needed in Algorithm 2 is no longer necessary. In the *Communication* phase (lines 11-26), all processes except the first one send the number of valid elements to the first process (that with pid 0), which executes the inclusive prefix-sum on vector V [2], where each element is the addition of all its preceding elements including itself. Then, each position of the array V is sent back to the corresponding process. Following this, each process executes the *Scan* phase (lines 27-30) on its own segment independently, based on the shifting value received previously. Finally, in the *Scatter* phase (lines 31-35) the validity of each element is re-checked by

 ISBN: 978-84-617-8694-7

Algorithm 3 Parallel work-efficient stream compaction

Input: Vector I of length n
Input: Predicate function F
Input: Number of processes p
Input: pid of process
Output: Vector O of valid elements
Output: nvalid: the number of valid elements

```
1: nvalid = 0
2: tamp = n/p
3: scan[0:(tamp-1)] = 0
4: V[0:(t-1)] = 0
5: for i = 0 to tamp − 1 in parallel do
6:     if F(I[i]) then
7:         scan[i] = 1
8:         V[pid] = V[pid] + 1
9:     end if
10: end for
11: if pid > 0 then
12:     Send V[pid] to processpid0
13: end if
14: if pid == 0 then
15:     for i = 1 to npid do
16:         ReceiveV[i]
17:         V[i] = V[i] + V[i − 1]
18:     end for
19:     for i = 1 to npid do
20:         Send V[i − 1] to processpidi
21:     end for
22:     nvalid = V[p − 1]
23: end if
24: if pid > 0 then
25:     Receive V[pid − 1]
26: end if
27: scan[0] = temp[0] + V[pid − 1]
28: for i = 0 to tamp − 1 in parallel do
29:     scan[i] = scan[i − 1] + temp[i]
30: end for
31: for i = 0 to tamp − 1 in parallel do
32:     if scan[i]! = scan[i − 1] then
33:         O[scan[i − 1]] = I[i]
34:     end if
35: end for
```

 ISBN: 978-84-617-8694-7

evaluating two consecutive positions of the *scan* array, obtaining the output array (O) with the valid elements from the input array (I).

Figure 3 illustrates an example for a list of elements ranging between 0 and 4 and the predicate function F that selects all elements that are not zero for an execution of four MPI processes. As in the previous example, 4 input elements are assigned to each MPI process and the *Validation* phase is applied producing directly the validity of each element on vector *scan* together with the number of valid elements that each process finds out. The latter is stored on vector V. Then, the process $P0$ executes the inclusive prefix-sum on vector V and sends back the output to the rest of the processes as is indicated by the arrows in Figure 3. Finally, each process enters the *Scan* and *Scatter* phases taking into account the corresponding shifting value calculated by $P0$.

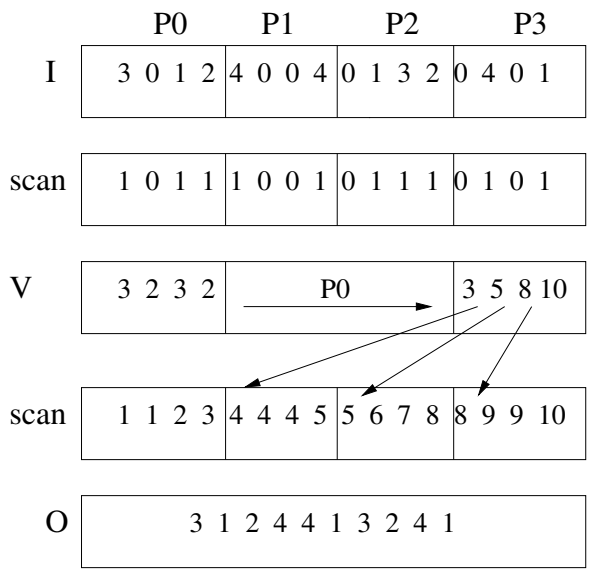


Figure 3: Example of Parallel work-efficient stream compaction

3 Experiments

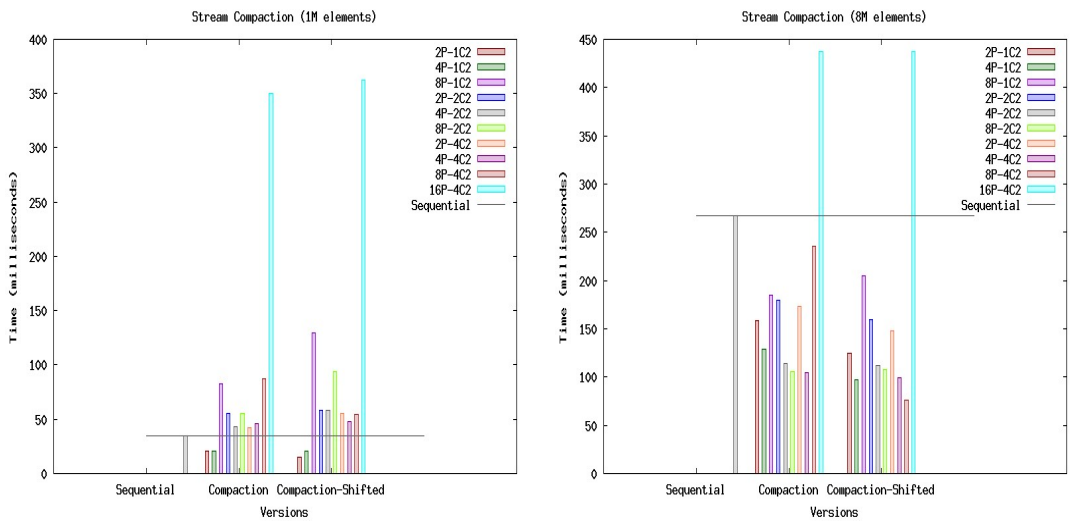
We have built a cluster which is composed by four Odroid C2 nodes. Each node contains a 1.5 GHz quad-core 64-bit ARM Cortex-A53 CPU and 2 GBytes of RAM memory. All the nodes are interconnected through a Gigabit Ethernet switch. The operating system installed on each node is Ubuntu 16.04.02 LTS. In this cluster we have installed MPICH (v3.2) as the MPI library implementation.

We have executed and measured the two parallelization strategies for stream compaction

 ISBN: 978-84-617-8694-7

presented in Section 2 on this cluster. The baseline for all the comparisons is the sequential version of Algorithm 2 without the *Communication* phase. Moreover, we have configured different parallel execution scenarios for the two parallel versions of the stream compaction problem explained before. We consider parallel executions with 2, 4, 8 and 16 MPI processes, running on the same Odroid C2 board or different boards (up to 4). We have chosen several input data sizes for our tests. In particular, we consider input arrays with $1M$, $8M$, $32M$ and $64M$ integer elements ranging between 0 and 4. The predicate function in all cases determines as valid all numbers that are not zero.

Figure 4a, Figure 4b, Figure 5a and Figure 5b show the execution times (in milliseconds) that are observed for input data sizes of $1M$, $8M$, $32M$ and $64M$ elements, respectively. For all these figures, from left to right, we first present the result obtained for the sequential version (`Sequential`), then we show the results for the Parallel stream compaction (`Compaction`) and Parallel work-efficient stream compaction (`Compaction-Shifted`) parallelization strategies respectively. For each one of them, we consider 2, 4 and 8 MPI processes running on one Odroid C2 board (`2P-1C2`, `4P-1C2` and `8P-1C2` respectively), and on 2 Odroid C2 boards, having 1, 2 and 4 MPI processes per board in each case (`2P-2C2`, `4P-2C2` and `8P-2C2` respectively), and finally 2, 4, 8 and 16 MPI processes running on 4 Odroid C2 boards, having 1, 2 or 4 processes per board as appropriate (`2P-4C2`, `4P-4C2` `8P-4C2` and `16P-4C2` respectively).



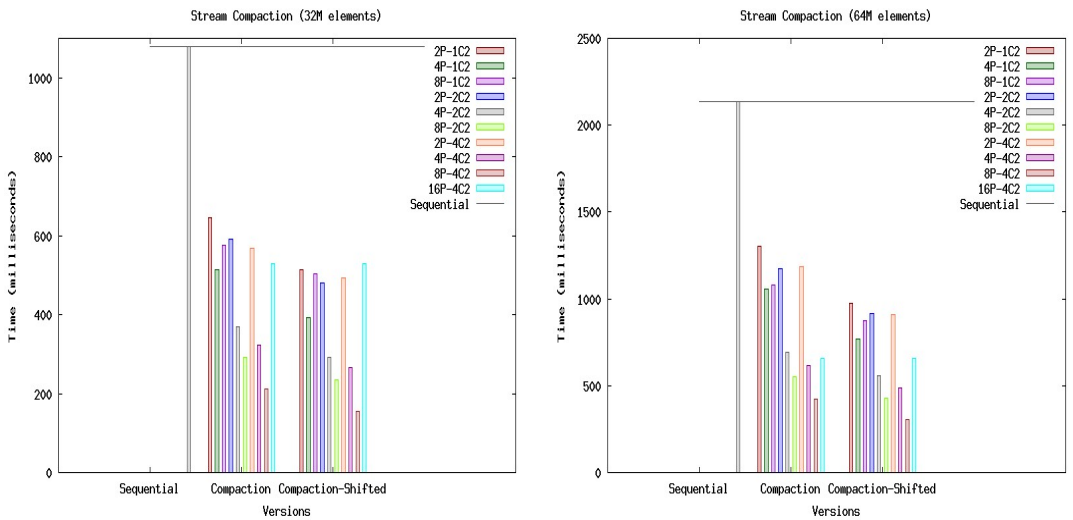
(a) (b)

Figure 4: Execution times (milliseconds) for stream compaction

From Figure 4a, we can see that the two proposed parallelization strategies for the

 ISBN: 978-84-617-8694-7

stream compaction problem obtain noticeable speed-ups when they are executed on a single Odroid C2 board with 2 or 4 MPI processes with regard to the sequential version. However, the executions on different Odroid C2 boards show negative outcomes from the performance point of view when the size of the input array is excessively small (1M elements). What makes the differences is that in the first case all communications take place on the same board, and therefore, can be performed with low latency. Contrarily what happens when communications involve several Odroid C2 boards. In this case, the time required for communication does not compensate the small processing time that is needed to obtain the stream compaction for such a small number of elements. Moreover, the executions on a single Odroid C2 SDC with 8 MPI processes (2 MPI processes per core) also show negative speed-ups revealing (as expected) that a configuration with more than one MPI process per core increments the communications and potentially slows computations.



(a) (b)

Figure 5: Execution times (milliseconds) for stream compaction

Taking a closer look at the results for one Odroid C2 board and 1M input size, we see that the speed-ups of the Parallel stream compaction strategy for 2 and 4 MPI processes are 1.68 and 1.64 respectively. Similarly, the work-efficient stream compaction parallelization strategy obtains speed-ups of 2.25 and 1.63 for 2 and 4 MPI processes respectively. Therefore, in both cases, fewer MPI processes, and therefore, less amount of communications among several processes brings the best results. These two parallel versions do not scale due to the small computation/communication ratio that they exhibit, which decreases as the number of processes grows.

In general, from Figures 4b, 5a and 5b, we can see that as the input data size increases, so it does the speed-ups obtained by the two parallelization strategies analyzed in this work when more cores are involved. The exception is the configuration with 16 processes running on 4 Odroid C2 boards (4 processes per board), which reaches lower speed-ups than that with 8 processes running on 4 Odroid C2 boards (2 processes per board).

More specifically, Figure 4b shows the results seen for 8M elements. In this case the two proposed parallelization strategies obtain significant speed-ups when executed on a single Odroid C2 board with 2, 4 or 8 MPI processes with regard to the sequential version. Additionally, the scalability is good for 2 and 4 MPI processes obtaining 1.69 and 2.06 for the Parallel stream compaction strategy and 2.14 and 2.74 for the Parallel work-efficient stream compaction approach. Therefore, for medium input data sizes the computation/communication ratio is appropriate. Although the two parallelization strategies also achieve gains for the configuration ($8P-1C2$) with 2 processes per core on a single Odroid C2 (speed-ups of 1.44 and 1.31, respectively), these speed-ups are (as expected) lower than those of the ($4P-1C2$) case. It is clear that the fact that there are twice the number of MPI processes than the total number of cores available introduces extra scheduling overhead and causes worse use of cores' resources (such as caches). On the other hand, the executions on different Odroid C2 SDCs (except for $8P-4C2$ and $16P-4C2$) present important speed-ups and good scalability for 2, 4 and 8 MPI processes for the two proposed parallelization strategies. Thus, the increment in the number of processes per Odroid C2 implies a suitable operation of the Odroid C2 cluster, where the communication latency among the different boards of the cluster does not ballast performance. In the $8P-4C2$ case is where the performance differences between the two parallelization strategies start appearing. Whereas the most efficient strategy (namely Compaction-Shifted) achieves the highest speed-up for this configuration, the other one cannot improve over the results reached by $4P-4C2$ demonstrating its more limited scalability for medium-sized workloads. Finally, the large number of processes involved in $16P-4C2$ results into excessively small computation/communication ratios, which is the reason for the negative outcomes observed in both cases.

As we can observe in Figures 5a and 5b, having higher input data sizes for the two parallel stream compaction strategies results into significant gains in all the configurations. For both input data sizes both Compaction and Compaction-Shifted obtain speed-ups that are close to that observed for the $8M$ elements case when executed on a single Odroid C2 SDC with 2, 4 or 8 MPI processes. However, the resulting speed-ups become even more important as the number of involved cores grows. Moreover, they scale nicely for 2, 4 and 8 MPI processes, achieving their highest values for 8 MPI processes running on 4 Odroid C2 SDCs (5.10 and 5.06 for the parallel stream compaction and input data sizes of $32M$ and $64M$, respectively; and 6.96 and 7.04 for the parallel work-efficient stream compaction and input data sizes of $32M$ and $64M$, respectively). It is also worth noting that even for these

 ISBN: 978-84-617-8694-7

large input sizes, the results reached for the $16P - 4C2$ configuration are worse than those of the $8P - 4C2$ in both cases. Now this the differences between them becomes narrower as input data sizes increase.

4 Conclusions

In this work, we have studied the parallelization of the stream compaction problem on a low-cost cluster of single-board computers. Particularly, we have configured the low-cost cluster from 4 Odroid C2 SDCs which are interconnected using a typical Gigabit Ethernet switch. We have implemented two parallel versions for the stream compaction problem using MPI. Then, we evaluate them considering varying number of processes and/or SDCs, as well as different input sizes. In general, we see that when the number of elements in the stream is too small, the most important benefits are observed when all participating processes are in the same Odroid board. In this case the low computation/communication ratio for small number of input elements cannot make up for the overhead entailed by the inter-SDC communications. As the number of elements in the input stream increases, so it does the number of processes that can participate in parallel executions, and important speed-ups are reached. Overall, the best results are reached when eight MPI processes are distributed among the four SDCs that conform the cluster. In this case, speed-ups of 6.96 and 7.04 are obtained for the Compaction and Compaction-Shifted strategies respectively, for the larger problem size considered in this work (input data size of $64M$).

Acknowledgements

This work was supported by the Spanish MINECO, as well as by European Commission FEDER funds, under grant TIN2015-66972-C5-3-R.

References

[1] David Abdurachmanov, Peter Elmer, Giulio Eulisse, and Shahzad Muzaffar. Initial explorations of ARM processors for scientific computing. *Journal of Physics: Conference Series*, 523(1):012009, 2014.

[2] Guy E. Blelloch. Prefix sums and their applications. Technical Report CMU-CS-90-190, School of Computer Science, Carnegie Mellon University, November 1990.

[3] W. Gropp, E. Lusk, and A. Skjellum. *Using MPI, Second Edition.* MIT Press, 1999.

[4] A. Hague, G. Hastings, M. Klling, B. Croston, A. Oldknow, B. Lockwood, and C. Beale. *The Raspberry Pi Education Manual Version 1.0. Computing at School.* Creative Commons License, 2012.

[5] R. Heeks and A. Robinson. Ultra-low-cost computing and developing countries. *Communications of the ACM*, 56(8):22–24, August 2013.

[6] Jared Hoberock, Victor Lu, Yuntao Jia, and John C. Hart. Stream compaction for deferred shading. In *Proceedings of the Conference on High Performance Graphics 2009*, HPG '09, pages 173–180, New York, NY, USA, 2009. ACM.

[7] D. Horn. Stream reduction operation for gpgpu applications. *GPU Gems 2*, pages 573–589, August 2005.

[8] P. Membrey and D. Hows. *Learn Raspberry Pi 2 with Linux and Windows 10 (2nd Edition).* Apress, 2015.

[9] Nvidia. Thrust. http://docs.nvidia.com/cuda/thrust, 2015.

[10] Alexandru Prjan. Solutions For Optimizing The Stream Compaction Algorithmic Function Using The Compute Unified Device Architecture. *Romanian Economic Business Review*, 6(1):216–231, May 2012.

[11] Bin Ren, Tomi Poutanen, Todd Mytkowicz, Wolfram Schulte, Gagan Agrawal, and James R. Larus. SIMD parallelization of applications that traverse irregular data structures. In *Proceedings of the 2013 IEEE/ACM International Symposium on Code Generation and Optimization (CGO)*, pages 1–10, 2013.

[12] David Roger, Ulf Assarsson, and Nicolas Holzschuch. Efficient stream reduction on the GPU. In *Workshop on General Purpose Processing on Graphics Processing Units*, oct 2007.

[13] Q. Sun, C. Yang, C. Wu, and L. F. Liu. Fast parallel stream compaction for IA-based multi/many-core processors. In *Proceedings of the 16th IEEE/ACM International Symposium on Cluster, Cloud, and Grid Computing*, May 2016.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

Quasi-monogenic functions

Swanhild Bernstein[1]

[1] *Department of Mathematics and Informatics, TU Bergakademie Freiberg*

emails: swanhild.bernstein@math.tu-freiberg.de

Abstract

We define quasi-monogenic functions based on a generalized Riesz-Hilbert operator. An application of this theory is the linear Riesz transform which commutes (up to a constant) with shears.

Key words: quasi-monogenic, monogenic functions, Riesz transforms, Hardy space decomposition

1 Introduction

Quaternionic and Clifford analysis are a refinement of classical harmonic analysis. Similar to the analytic signal that is based on classical function theory, the monogenic signal [2] is an important tool in higher dimensional signal analysis. The analytic signal is based on the Hilbert transform whereas the monogenic signal is based on the Riesz transforms. Using quaternions, the Riesz transforms build a higher dimensional analog of the Hilbert transform. More on applications of monogenic signals can be found in [1]. All Hilbert transform are related to Dirac and Cauchy-Riemann operators and Hardy spaces as spaces of boundary values of monogenic functions in upper/lower half spaces.
The Hilbert and Riesz transforms are the most important singular integrals and had been investigated in detail in [8]. Because they are also operators of convolution type, they can be studied in Fourier domain. That was done in [7] and extended to more general operators in [5]. The Riesz transforms can be used to construct monogenic wavelets [4] and can be applied to signal processing. Shearlets are similar to wavelets, but they are invariant under shears and not under rotations. Unfortunately, the Riesz transforms does not correspond to shear operations. But the in [3] introduced linearized Riesz transforms do. Therefore, quasi-monogenic shearlets are constructed by the aid of the linearized Riesz transforms.

The aim of this paper to construct Dirac operator D_L in $\mathbb{R}^2$ and Cauchy-Riemann operators in $\mathbb{R}^3$ which are generated by Riesz transforms $\mathcal{H}_L$. That will prove that the quasi-monogenic Riesz transforms generate an essentially equivalent function theory even though the operators can only be described by their Fourier multipliers.
Examples of quasi-monogenic Riesz transforms are the classical Riesz transforms, and the linearized Riesz transforms defined in [3].

2 Preliminaries

Let $\mathbb{H}$ the algebra of complex quaternions. The algebra consists of the basic elements e_0, e_1, e_2, e_3 which fulfill

$$e_0^2 = 1, \quad e_0 e_j = e_j e_0, \quad e_j^2 = -1, \quad e_i e_j = -e_j e_i, \ i, j = 1, 2, 3,$$

$$e_1 e_2 = e_3, \quad e_2 e_3 = e_1, \quad e_3 e_1 = e_2.$$

Because e_0 is the unit element of the algebra we identify e_0 with 1.

Definition 1 *Given u in the Schwartz space $\mathcal{S}(\mathbb{R}^2)$ we define the Fourier transform $\mathcal{F}(u) = \hat{u}$ of f as*

$$\widehat{u}(\underline{\omega}) := \frac{1}{2\pi} \int_{\mathbb{R}^2} u(\underline{x})\, e^{-i\underline{\omega}\cdot\underline{x}}\, d\underline{x}.$$

The inverse Fourier transform $\mathcal{F}^{-1}(u) = u^{\vee}$ is given by

$$u^{\vee}(\underline{x}) = \frac{1}{2\pi} \int_{\mathbb{R}^2} u(\underline{\omega})\, e^{i\underline{x}\cdot\underline{\omega}}\, d\underline{\omega}.$$

Theorem 2 (Hörmander-Mikhlin Mulitplier theorem) *Let $m(\underline{\omega})$ be a complex-valued bounded function on $\mathbb{R}^n\backslash\{0\}$ that satisfies either*

(a) Mikhlin's condition

$$|\partial_{\underline{\omega}}^{\alpha} m(\underline{\omega})| \leq A|\underline{\omega}|^{-|\alpha|}$$

for all multi-indices $|\alpha| \leq \left[\frac{n}{2}\right] + 1$,

(b) Hörmander's condition

$$\sup_{R>0} R^{-n+2|\alpha|} \int_{R<|\underline{\omega}|<2R} |\partial_{\underline{\omega}}^{\alpha} m(\underline{\omega})|\, d\underline{\omega} \leq A^2 < \infty$$

for all multi-indices $|\alpha| \leq \left[\frac{n}{2}\right] + 1$.

Then for all $1 < p < \infty$ m is an $L^p(\mathbb{R}^n)$ mulitplier and moreover, the operator $f \mapsto \mathcal{F}^{(-1)}(\mathcal{F}(f)m)$ maps $L^1(\mathbb{R}^n)$ to $L^{1,\infty}(\mathbb{R}^n)$, it suffices to prove that the distribution W coincides with a function K on $\mathbb{R}^n\backslash\{0\}$ that satisfies Hörmander's condition.

 ISBN: 978-84-617-8694-7

3 Quasi-monogenic functions

We consider a generalization of the theory of [6]. Let be

$$\mathcal{H}_L u = \mathcal{F}^{(-1)}(\mathcal{F}(u)h_L),$$

where h_L is an $L^p(\mathbb{R}^n)$ mulitplier and $h_L^2(\underline{\omega}) = 1$ and

$$P_L^{\pm} u = \frac{1}{2}\left(1 \pm \mathcal{H}_L\right) u.$$

The operator $|D|$, defined by $(|D|u)^\wedge(\underline{\omega}) = |\underline{\omega}|\widehat{u}(\underline{\omega})$, is a closed unbounded operator in $L^2(\mathbb{R}^2)$ with domain $\mathcal{D}(|D|) = \mathcal{D}(D) = H^1(\mathbb{R}^2)$, the Sobolev space $H^1(\mathbb{R}^2) = \{u \in L^2(\mathbb{R}^2 : \frac{\partial u}{\partial x_j} \in L^2(\mathbb{R}^2),\ 1 \le j \le 2\}$.

Definition 3 *The operator $D_L : H^1(\mathbb{R}^2) \to L^2(\mathbb{R}^2)$ is defined as*

$$D_L = -|D|\mathrm{sgn}(D_L) = -|D|\mathcal{H}_L.$$

Further,

$$|D| = |D_L| = -\mathrm{sgn}\{D_L\}D_L \quad \text{where} \quad \mathrm{sgn}\{D_L\} = P_L^+ - P_L^-,$$

is a bounded linear operator on $L^2(\mathbb{R}^2)$ satisfying $(\mathrm{sgn}\{D_L\})^2 = I$ and the operator D_L is self-adjoint and factorizes the Laplacian $D_L\overline{D_L} = -D_L D_L = \Delta$ on $\mathbb{R}^2$.

Given $u \in L^2(\mathbb{R}^2)$ and $t > 0$, define $u_+(t) \in L^2(\mathbb{R}^2)$ by

$$u_+(t) = e^{-t\{D_L\}}P_L^+ u = e^{-t|D|}P_L^+ u.$$

Definition 4 *A function u is left or right **quasi-monogenic** in $\mathbb{R}^3$ if*

$$(\partial_{x_0} + D_L)u = 0 \quad or \quad u(\partial_{x_0} + D_L) = 0.$$

Theorem 5 *Let $u \in L^2(\mathbb{R}^2)$, define functions U_+ on $\mathbb{R}^3_+$ and U_- on $\mathbb{R}^3_-$ by*

$$U_+(x_0 e_0 + \underline{x}) = u_+(x_0)(\underline{x}), \quad x_0 > 0, \underline{x} \in \mathbb{R}^2,$$
$$U_-(x_0 e_0 + \underline{x}) = u_-(x_0)(\underline{x}), \quad x_0 < 0, \underline{x} \in \mathbb{R}^2,$$

then

(1) $\dfrac{\partial U_\pm}{\partial x_0}(x_0 e_0 + \underline{x}) + D_L U_\pm(x_0 e_0 + \underline{x}) = 0, \quad x_0 e_0 + \underline{x} \in \mathbb{R}^3_\pm$, *or in other words, that the functions $U_\pm$ are left **quasi-monogenic** on their respective half-spaces.*

(2) $\lim\limits_{x_0 \to 0\pm} U_\pm(x_0 e_0 + \underline{x}) = P_L^\pm U(\underline{x})$ *for almost all* $\underline{x} \in \mathbb{R}^2$.
(Plemelj-Sochotzki formulae)

 ISBN: 978-84-617-8694-7

(3) $\lim_{x_0 \to \pm\infty} U_{\pm}(x_0 e_0 + \underline{x}) = 0$ *for all* $\underline{x} \in \mathbb{R}^2$.

Examples of quasi-monogenic Riesz transforms are the classical Riesz-Hilbert transform in Clifford analysis and the linearized Riesz transforms.

Definition 6 (Linearized Riesz transforms) *The linearized Riesz transforms* $\mathcal{R}_1$, $\mathcal{R}_2$, $L^2(\mathbb{R}^2) \to L^2(\mathbb{R}^2)$ *are defined by*

$$\mathcal{R}_1 u(\underline{x}) := \mathcal{F}^{-1}(-i\cos(\theta_L(\underline{\omega}))\widehat{u}(\underline{\omega})), \quad \mathcal{R}_2 u(\underline{x}) := \mathcal{F}^{-1}(-i\sin(\theta_L(\underline{\omega}))\widehat{u}(\underline{\omega})),$$

where $\theta(\underline{\omega}) = \arctan2(\underline{\omega}) = \arctan2\left(\frac{\omega_2}{\omega_1}\right)$ *and* $\mathcal{H} = e_1\mathcal{R}_1 + e_2\mathcal{R}_2$.

The linearized Riesz transforms are invariant under shears $S_s := \begin{pmatrix} 1 & s \\ 0 & 1 \end{pmatrix}, s \in \mathbb{R}$, ([3]).

References

[1] S. BERNSTEIN, J. L. BOUCHOT, M. REINHARDT, B. HEISE, *Generalized Analytic Signals in Image Processing: Comparison, Theory and Applications.* in S. Sangwine, E. Hitzer (eds) Quaternion and Clifford Fourier Transforms and Wavelets, Trends in Mathematics, 221–246, 2013.

[2] M. FELSBERG, G. SOMMER, *The monogenic signal,* IEEE Trans. Signal Proc., **49**(12), (2001), 3136–3144.

[3] S. HÄUSER, G. STEIDL, B. HEISE, *Linearized Riesz Transform and Quasi-Monogenic Shearlets,* in: Int.J.Wavelets. Multires.Informat. Proc., 12(3), 2014.

[4] S. HELD, M. STORATH, P. MASSOPUST, B. FORSTER, *Steerable wavelet frames based on the Riesz transform,* IEEE Trans. Image Proc., **19**(3) (2010) 653–667.

[5] L. HÖRMANDER, *The Analysis of Linear Partial Differential Operators* I, 2nd ed. Springer Verlag, Berlin-Heidelberg-New York, 1990.

[6] A. MCINTOSH, *Clifford Algebras, Fourier Theory, Singular Integrals, and Harmonic Functions on Lipschitz Domains,* in J. Ryan (ed.) Clifford Algebras in Analysis and Related Topics, Studies in Adv. Math., CRC Press, 1996, 33-88.

[7] S. G. MIKHLIN, *Multidimensional Singular Integrals and Integral Equations,* International Series of Monographs in Pure and Applied Mathematics, Vol. 83, Pergamon Press, 1965.

[8] E. STEIN, G. WEISS, *Singular Integrals and Differentiability Properties of Functions,* Princton University Press, Princton, New Jersey, 1970.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

Existence theorems and weak attractors for quasicrystal dynamics with non-linear gyroscopic effects

Luca Bisconti[1] **and Paolo Maria Mariano**[2]

[1] *DIMAI "U. Dini", Università di Firenze, V.le Morgagni 67/A, I-50134 Firenze,*

[2] *DICEA, Università di Firenze, V. Santa Marta 3, I-50139 Firenze,*

emails: luca.bisconti@unifi.it, paolo.mariano@unifi.it

Abstract

We review some existence results in the linear elastodynamics of quasicrystals with and without nonlinear phason inertia, and prove in the latter case the existence of a weak attractor.

Key words: Continuum mechanics, quasicrystals, gyroscopic phason inertia, existence theorems, weak attractors
MSC 2000: Primary 74A30; Secondary 74A60, 35A01

1 Introduction

The results of 1982 experiments by D. Shechtman, published in 1984 [20], showed atomic arrangements with icosahedral symmetry in aluminum-based synthetic alloys, and pentagonal symmetry in thin films of the same materials, not determined by twinned atomic structures. Natural quasi-periodic alloys have been found since then in meteorites. In a 1991 the International Union of Crystallography changed the definition of crystals including the possibility of quasi-periodic atomic arrangements [9].

In fact, a quasi-periodic atomic arrangement in a m-dimensional space can be viewed as the orthogonal projection of a portion of a periodic lattice, in a $2m$-dimensional space, onto an appropriate (incommensurate) subspace [5]. If we move orthogonally to itself the subspace in the lattice in an appropriate way, we recognize the need of atomic shifts to assure quasi-periodicity. They are commonly called **phason** defects. The hyperspace construction, however, is just an ideal geometric representation of what the nature is. Indeed, we can

 ISBN: 978-84-617-8694-7

consider phasons as *inner* degrees of freedom exploited locally by the atoms to assure quasi-periodicity, compatibly with the boundary conditions imposed to a quasi-crystalline body by the interaction with the external environment. As inner degrees of freedom, they are invariant with respect to rigid translations of the whole body.

In building up a continuum representation of quasicrystal mechanics, the geometric constructions above have suggested at a first glance of thinking just of a higher dimensional replica of traditional crystal elasticity or viscoelasticity (see, e.g., for the first proposals [11], [6] and also [10], [17], [8]). Researchers have then focused attention primarily on its linear version, having in this way the concrete advantage of a format where we can easily reproduce systematically, all the standard results of traditional linear elasticity. However, the circumstance that phasons are insensitive to global rigid translations of a body, as they are inner degrees of freedom, implies the existence of a self-action (see Theorem 2.1 below) rendering the description of the mechanics of quasicrystals more complicated than the mere reproduction of standard elasticity, in a space with dimension higher than the one of the physical ambient. A phason self-action has been already *assumed* to exist in [17] but just to have dissipative nature and to drive phason diffusion. A first proof of the existence of a possibly nonzero phason self-action with both conservative and dissipative nature appeared first in [13], and in [16].

Here, we consider the dynamics of quasicrystals with nonlinear rotational phason inertia and phason diffusion. We summarize the results obtained in [2] an a theorem in the same reference determining the appropriate balance equations. Then, we show new results concerning the existence of weak attractors.

2 Continuum mechanics of quasicrystals

2.1 Deformation and phason field

We write $\mathcal{B}$ for the macroscopic **reference** shape of a quasicrystalline body, assumed to be a bounded arcwise connected open region in the three-dimensional point space $\mathcal{E}^3$, coinciding with the interior of its closure and endowed with surface-like boundary uniquely oriented everywhere to within a finite number of corners and cusps. In an space isomorphic to $\mathcal{E}^3$, indicated by $\tilde{\mathcal{E}}^3$, we record shapes of the body that we consider **deformed** with respect to $\mathcal{B}$, reached by means of one-to-one, differentiable, orientation preserving maps $x \longmapsto y := \tilde{y}(x) \in \tilde{\mathcal{E}}^3$.

A field $\tilde{\nu}$ taking values in a three-dimensional real vector space $\mathcal{V}^3$, accounts point-by-point at the continuum scale for the atomic flips, which allow to match quasi-periodicity. This is the so-called **phason field** in Lagrangian representation, i.e., considered as a field over the reference place. We endow $\tilde{\mathcal{E}}^3$, $\mathcal{E}^3$ and $\mathcal{V}^3$ with Cartesian frames.

 ISBN: 978-84-617-8694-7

Motions are (in generalized sense) pairs

$$(x,t) \longmapsto y := \tilde{y}\,(x,t) \in \tilde{\mathcal{E}}^3, \quad (x,t) \longmapsto \nu := \tilde{\nu}\,(x,t) \in \mathcal{V}^3,$$

assumed to be sufficiently differentiable in time.

We will write F and N for the **deformation gradient** and the **phason field gradient**, evaluated at x and t. The assumption that the deformation preserves the local orientation of triples of linearly independent vectors implies $\det F > 0$. We define another vector field, the **displacement** as $(x,t) \longmapsto u := \tilde{u}\,(x,t) := \tilde{y}\,(x,t) - i\,(x)$, also called the **phonon field** in the jargon of the physics of quasicrystals. Consequently, we have $\nabla u := \nabla \tilde{u}\,(x,t) = F + I$, where I is the second-rank unit tensor. As a matter of notation, we will write u_t, u_{tt}, and ν_t for the values $\dot{u} := \frac{d\tilde{u}(x,t)}{dt}$, $\ddot{u} := \frac{d^2\tilde{u}(x,t)}{dt^2}$, and $\dot{\nu} := \frac{d\tilde{\nu}(x,t)}{dt}$, respectively, the latter chosen for the sake of notational uniformity. The velocity in the physical space and the phason time rate just listed are expressed in Lagrangian representation, i.e., as fields over the reference place and the time scale. We can have an Eulerian representation of such fields, i.e. we can consider them defined over the actual shape $\mathcal{B}_a = \tilde{y}(\mathcal{B},t)$. In this case we write $\mathrm{v}(y,t)$ and $\upsilon(y,t)$. Since at x and t, the vector $\dot{y}(x,t) = \dot{u}(x,t)$ is tangent to $\mathcal{B}_a$ at the point $y := \tilde{y}(x,t)$, we have the standard identity $\dot{y}(x,t) = \mathrm{v}(y,t)$. An analogous relation *does not hold* between $\upsilon(y,t)$ and $\dot{\nu}(x,t)$. The lack of identity depends on the circumstance that υ is the time rate of the Eulerian representation of the phason field, which is a map $\tilde{\nu}_a$ defined by $\tilde{\nu}_a := \tilde{\nu} \circ \tilde{y}^{-1}$, a definition possible for $\tilde{y}$ is one-to-one. The subscript a means *actual*, i.e. *referred to the deformed configuration*, here and in what follows. The condition $|\nabla u| \ll 1$ defines the **small strain regime**, in which we develop the analyses presented in Sections 3 and 4. In this setting we can 'confuse' $\mathcal{B}$ with $\mathcal{B}_a$ and ν with $\nu_a := \tilde{\nu_a}(y,t)$.

2.2 Changes in observers

According to the definition proposed explicitly in [13] and in [14], we define **observer** *a collection of frames of reference assigned on all spaces necessary to describe the shape of a body and its motion*. Here, an observer is then (1) a frame in $\tilde{\mathcal{E}}^3$, (2) one in $\mathcal{E}^3$, (3) a frame in $\mathcal{V}^3$ and (4) a time scale. We consider time-varying synchronous changes in observers leaving invariant the reference space and changing the frame(s) in $\tilde{\mathcal{E}}^3$ by a rigid body motion. Precisely, let us write $\mathcal{O}$ and $\mathcal{O}'$ for these two observers. Let v and v' the velocities recorded by $\mathcal{O}$ and $\mathcal{O}'$, respectively. Write v^* for the pull-back of v' into the frame of reference $\mathcal{O}$. For v^* we get

$$\mathrm{v}^* := c\,(t) + q\,(t) \times (y - y_0) + \mathrm{v},$$

where $c(t)$ and $q(t)$ are translational and rotational components of the rigid velocity of $\mathcal{O}'$ with respect to $\mathcal{O}$, as measured by $\mathcal{O}$. The distinction between the ambient space $\tilde{\mathcal{E}}^3$ and the phason one $\mathcal{V}^3$ is just matter of modeling. Atomic flips occur in the physical space. Also,

 ISBN: 978-84-617-8694-7

the notion of observer is just a formal representation of the concrete action of recording a phenomenon. Hence, when we rotate an observer in space we should perceive rotated the atomic flips. They are not affected by rigid translations in space for they are internal degrees of freedom. Consequently, in Eulerian representation of the phason rate we get

$$\upsilon^* = \upsilon + q \times \nu_a.$$

2.3 External power, invariance and balance

We derive balance equations from the invariance of power over a generic *part* of the body (see [12] for general complex materials as quasicrystals are). The word **part** indicates here a subset $\mathfrak{b}_a$ of $\mathcal{B}_a = \tilde{y}(\mathcal{B}, t)$ with non-null volume and the same geometric regularity of $\mathcal{B}_a$. Given a generic $\mathfrak{b}_a$, we divide as usual all actions exerted on $\mathfrak{b}_a$ by the environment and the rest of the body into bulk and contact families, the latter intended to be exerted through the boundary of $\mathfrak{b}_a$. Each family is also subdivided into standard and phason components, all defined by the expression of the power that the external actions must perform over $\mathfrak{b}_a$ to change its state of motion with velocity v in the physical space and phason rate υ. For this reason we call such a power **external**, indicating it by $\mathcal{P}^{ext}_{\mathfrak{b}}$ and defining it in Eulerian representation by

$$\mathcal{P}^{ext}_{\mathfrak{b}_a}(\mathrm{v}, \upsilon) := \int_{\mathfrak{b}_a} \left(b^{\ddagger}_a \cdot \mathrm{v} + \beta^{\ddagger}_a \cdot \upsilon \right) \, d\mu(y) + \int_{\partial \mathfrak{b}_a} (\mathfrak{t} \cdot \mathrm{v} + \tau \cdot \upsilon) \, d\mathcal{H}^2, \tag{1}$$

where $d\mathcal{H}^2$ is the surface measure along $\partial \mathfrak{b}_a$ and $d\mu(y)$ the volume measure in $\mathcal{B}_a$; $b^{\ddagger}_a$ and $\beta^{\ddagger}_a$ are standard and phason bulk actions respectively, while $\mathfrak{t}$ and τ are the contact ones.

At $y \in \partial \mathfrak{b}_a$, where $\partial \mathfrak{b}_a$ is oriented by the normal n, the standard traction $\mathfrak{t}$ is $\mathfrak{t} := \tilde{\mathfrak{t}}(y, n)$ (Cauchy's assumption). Here we impose also $\tau := \tilde{\tau}(y, n)$. We leave unexpressed the dependence on time for the sake of conciseness of some formulas below.

We subordinate $\mathcal{P}^{ext}_{\mathfrak{b}_a}(\mathrm{v}, \upsilon)$ to an axiom of invariance by following a view adopted in the general model-building framework of the mechanics of complex materials (see [12] for it).

Axiom 2.1. *$\mathcal{P}^{ext}_{\mathfrak{b}_a}(\mathrm{v}, \upsilon)$ is invariant under rigid-body-based changes in observers, i.e., $\mathcal{P}^{ext}_{\mathfrak{b}_a}(\mathrm{v}^*, \upsilon^*) = \mathcal{P}^{ext}_{\mathfrak{b}_a}(\mathrm{v}, \upsilon)$ for any choice of c, q, and $\mathfrak{b}_a$.*

Theorem 2.1. *The axiom of invariance implies the following list of assertions:*

*(a) If the fields $y \longmapsto b^{\ddagger}_a$, $y \longmapsto \nu \times \beta^{\ddagger}$, $y \longmapsto \mathfrak{t}$ and $y \longmapsto \tau$ are integrable over $\mathcal{B}_a$, the following integral balances hold for **any** part $\mathfrak{b}_a$ of $\mathcal{B}_a$ and for $\mathcal{B}_a$ itself:*

$$\int_{\mathfrak{b}_a} b^{\ddagger}_a \, d\mu(y) + \int_{\partial \mathfrak{b}_a} \mathfrak{t} \, d\mathcal{H}^2 = 0, \tag{2}$$

$$\int_{\mathfrak{b}_a} \left((y - y_0) \times b^{\ddagger}_a + \nu \times \beta^{\ddagger}_a \right) \, d\mu(y) + \int_{\partial \mathfrak{b}_a} ((y - y_0) \times \mathfrak{t} + \nu_a \times \tau) \, d\mathcal{H}^2 = 0. \tag{3}$$

 ISBN: 978-84-617-8694-7

(b) If the standard traction is continuous in y and the standard bulk action is bounded over $\mathcal{B}_a$ at every instant, $\mathfrak{t}$ satisfies the action-reaction principle $\mathfrak{t}(y,n) = -\mathfrak{t}(y,-n)$.

(c) In the same continuity conditions, a second-rank tensor σ independent of n exists and is such that $\mathfrak{t}(y,n) = \sigma(y)n(y)$.

(d) If the phason traction is continuous with respect to y and the field $y \longmapsto \nu \times \beta^{\ddagger}$ is bounded over $\mathcal{B}_a$ at every instant, τ satisfies a non-standard action-reaction principle, i.e., $\nu_a(y) \times (\tau(y,n) - \tau(y,-n)) = 0$.

(e) In the same regularity conditions above, a second-rank tensor $\mathcal{S}_a$ independent of n exists and is such that $\tau(y,n) = \mathcal{S}_a(y)n(y)$, a tensor that we call ***phason stress****.*

(f) If the field $y \longmapsto \sigma(y)$ is C^1 over $\mathcal{B}_a$ and just continuous over its boundary, equation (2) implies the validity of the standard pointwise balance of forces

$$b_a^{\ddagger} + \mathrm{div}\sigma = 0. \tag{4}$$

(g) If, in addition, the field $y \longmapsto \mathcal{S}_a(y)$ is C^1 over $\mathcal{B}_a$ and just continuous over its boundary, equation (3) implies the existence of a vector z_a such that

$$\mathrm{div}\mathcal{S}_a + \beta_a^{\ddagger} - z_a = 0, \tag{5}$$

and

$$\mathrm{Skw}\sigma = \mathrm{Skw}(\nu_a \otimes z_a + \nabla_y \nu_a)\mathcal{S}_a^T), \tag{6}$$

where the apex T means transposition, e *is Ricci's alternating symbol, and ∇_y is the gradient with respect to y.*

(h) The external power satisfies the following relation:

$$\mathcal{P}^{ext}_{\mathfrak{b}_a}(\mathrm{v}, \upsilon) = \int_{\mathfrak{b}_a} (\sigma \cdot \nabla_y \mathrm{v} + z_a \cdot \upsilon + \mathcal{S}_a \cdot \nabla_y \upsilon)\, d\mu(y), \tag{7}$$

for any choice of the rates involved. The right-hand side term takes the name of ***inner power****.*

Here, $b_a^{\ddagger}$ is the sum of inertial, (b_a^{in}), and non-inertial, (b_a), components. Since we do not know bulk forces acting directly over the phason field, we attribute to $\beta_a^{\ddagger}$ just inertial character. We identify b_a^{in} and $\beta_a^{\ddagger}$ by assuming that their power over $\mathfrak{b}_a$ equals the negative of the kinetic energy time rate, assumed to be the standard quadratic function of v, because experiments do no indicate peculiar phason inertia. By assumption the identity holds true for any velocity involved. Such arbitrariness and the assumed conservation of mass imply $b_a^{in} = -\rho a$, with $a := \dot{\mathrm{v}}$ the acceleration, and $\beta_a^{\ddagger} \cdot \upsilon = 0$. The latter identity may imply $\beta_a^{\ddagger} = -\ell\, \mathrm{curl}\, \mathrm{v} \times \upsilon$, with ℓ a positive constant, as showed in [16].

 ISBN: 978-84-617-8694-7

2.4 Constitutive structures

Clausius-Duhem inequality (combination of the first and the second law of thermodynamics) written in isothermal setting, imposes restrictions to the possible constitutive relations. It states that the rate of the free energy on any part of the body minus the external power developed by all actions on that part is lesser or equal to zero for any choice of the time rates involved in the inequality. When we presume that **free energy density** ψ, the stresses σ and $\mathcal{S}_a$, and the self-action z_a depend on F, ν, and N, we get

$$\sigma = (\det F)^{-1}\frac{\partial \psi}{\partial F}F^{\mathrm{T}},\ \ \mathcal{S}_a = (\det F)^{-1}\frac{\partial \psi}{\partial N}F^{\mathrm{T}},\ \ z_a = (\det F)^{-1}\frac{\partial \psi}{\partial \nu}.$$

These relations characterize the elastic setting of quasicrystals. By fixing ν and N, a standard argument shows that objectivity for ψ, i.e., invariance under the action of $SO(3)$ on the physical space, and convexity of ψ with respect to F are physically incompatible. Consequently, we commonly accept a polyconvex dependence of ψ on F. With respect to N, the free energy can be quadratic in the so-called *phason locked phase*, and ψ may admit a decomposed Ginzburg-Landau-type structure. In this case the existence of ground states (minimizers of the energy) has been found in [15] as a special case of a more general result presented there, further generalized in [7].

In small strain regime the dependence of the energy can be quadratic. With reference to the homogeneous and isotropic case, with $\varepsilon := \mathrm{Sym}\nabla u$ the small strain tensor and I the second-rank unit tensor, a rather general expression of the energy has been derived in [16]; it reads

$$\begin{aligned}\psi =& \frac{1}{2}\lambda\left(\varepsilon\cdot I\right)^2 + \mu\varepsilon\cdot\varepsilon + \frac{1}{2}k_1\left(N\cdot I\right)^2 + k_2\mathrm{Sym}N\cdot\mathrm{Sym}N + k_2'\mathrm{Skw}N\cdot\mathrm{Skw}N\\ &+ k_3\left(\varepsilon\cdot I\right)\left(N\cdot I\right) + k_3'\mathrm{Sym}N\cdot\varepsilon + \frac{1}{2}k_0\left|\nu\right|^2\end{aligned} \tag{8}$$

from which we get

$$\sigma = \lambda\left(\mathrm{tr}\varepsilon\right)I + 2\mu\varepsilon + k_3\left(\mathrm{tr}N\right)I + k_3'\mathrm{Sym}N, \tag{9}$$

$$z_a = k_0\nu, \tag{10}$$

$$\mathcal{S}_a = k_1\left(\mathrm{tr}N\right)I + 2k_2\mathrm{Sym}N + 2k_2'\mathrm{Skw}N + k_3\left(\mathrm{tr}\varepsilon\right)I + k_3'\varepsilon. \tag{11}$$

λ and μ are standard Lamé constants. k_0, k_1, k_2, k_2', k_3 are elastic constants related with the phason field.

Phason diffusion has dissipative character and may even determine viscous-like effects. To account for them, maintaining compatibility with the mechanical dissipative inequality, we presume additive decompositions of σ, $\mathcal{S}_a$ an z_a into energetic $(\sigma^e, \mathcal{S}_a^e, z_a^e)$ and dissipative $(\sigma^d, \mathcal{S}_a^d, z_a^d)$ components, the former depending on strain, phason filed and its gradient, the

 ISBN: 978-84-617-8694-7

latter being functions of the same variables and their first time rates. We declare the dissipative nature of σ^d, $\mathcal{S}_a^d$, and z_a^d, by imposing

$$\sigma^d \cdot \nabla\nu \geq 0, \quad \mathcal{S}_a^d \cdot \nabla\upsilon \geq 0, \quad z_a^d \cdot \upsilon \geq 0,$$

for any choice of $\nabla\nu$, $\nabla\upsilon$, and υ; the equality to zero occurs when the time rates vanish. These inequalities may imply that $\sigma^d = \epsilon\nabla\nu$, $\mathcal{S}_a^d = \delta\nabla\upsilon$, and $z_a^d = \varsigma\upsilon$, with ϵ, δ, and ς positive constants. In small strain regime $\nu \cong \dot{u} = u_t$, $\upsilon \cong \dot{\nu}_a = \dot{\nu} = \nu_t$, so that we write

$$\sigma^d = \epsilon\nabla u_t, \quad \mathcal{S}_a^d = \delta\nabla\nu_t, \quad z_a^d = \varsigma\nu_t.$$

Consequently, the constitutive equations (9), (10) and (11) become

$$\sigma = \lambda\,(\mathrm{tr}\varepsilon)\,I + 2\mu\varepsilon + k_3\,(\mathrm{tr}N)\,I + k_3'\mathrm{Sym}N + \epsilon\nabla u_t, \tag{12}$$

$$z_a = k_0\nu + \varsigma\nu_t, \tag{13}$$

$$\mathcal{S}_a = k_1\,(\mathrm{tr}N)\,I + 2k_2\mathrm{Sym}N + 2k_2'\mathrm{Skw}N + k_3\,(\mathrm{tr}\varepsilon)\,I + k_3'\varepsilon + \delta\nabla\nu_t. \tag{14}$$

3 Existence results

3.1 Dynamics with phason diffusion and absence of gyroscopic effects

In small strain regime and under the validity of the linear constitutive structures (9), (11) and (13), in absence of non-inertial body forces and gyroscopic-type phason inertia, by imposing u and ν along $\partial\mathcal{B}$ (Dirichlet boundary conditions) and their values together with those of the velocity u_t over $\mathcal{B}$ as initial conditions, the balance equations read

$$\begin{aligned}
&\rho u_{tt} = \mu\Delta u + \xi\nabla\,\mathrm{div}\,u + \kappa\Delta\nu + \bar{\xi}\nabla\,\mathrm{div}\,\nu && \text{in } (0,T)\times\mathcal{B},\\
&\varsigma\nu_t = \zeta\Delta\nu + \gamma\nabla\,\mathrm{div}\,\nu + \kappa\Delta u + \bar{\xi}\nabla\,\mathrm{div}\,u - \kappa_0\nu && \text{in } (0,T)\times\mathcal{B},\\
&u(t,x) = \bar{u}(x),\ \nu(t,x) = \bar{\nu}(x), && \text{on } (0,T)\times\partial\mathcal{B},\\
&u|_{t=0} = u_0,\ u_t|_{t=0} = \dot{u}_0,\ \nu|_{t=0} = \nu_0, && \text{on } \mathcal{B},
\end{aligned} \tag{15}$$

where u_0, $\dot{u}_0$ and ν_0 are the initial data, and the constitutive parameters are constants and satisfy the following relations: $\xi = \lambda + \mu$, $\bar{\xi} = k_3 + \frac{1}{2}k_3'$, $\zeta = k_2 + k_2'$, $\gamma = k_1 + k_2 - k_2'$, $\kappa = \frac{1}{2}k_3'$, and λ, μ, k_i, k_i', $i = 1,2,3$, and ρ is the mass density.

Definition 3.1. *We say that a pair (u,ν) is a **weak solution** of the system (15) if, for a given $T > 0$, the conditions listed below hold true.*

 ISBN: 978-84-617-8694-7

(1) Regularity:

$$\begin{aligned} &u \in L^\infty(0,T;\mathcal{H}^1) \cap C([0,T];L^2(\mathcal{B})) \cap C_{weak}([0,T];\mathcal{H}^1), \\ &\nu \in L^2(0,T;\mathcal{H}^1) \cap C([0,T];L^2(\mathcal{B})), \end{aligned} \tag{16}$$

$$\begin{aligned} &u_t \in L^\infty(0,T;L^2(\mathcal{B})) \cap C_{weak}([0,T];L^2(\mathcal{B})),\ u_{tt} \in L^2(0,T;\mathcal{H}^{-1}), \\ &\nu_t \in L^2(0,T;L^2(\mathcal{B})). \end{aligned} \tag{17}$$

(2) Weak formulation: For all $(w,h) \in C_0^\infty(0,T;\mathcal{H}^1) \times C_0^\infty(0,T;\mathcal{H}^1)$,

$$\begin{aligned} &\rho \int_0^T \int_{\mathcal{B}} u_{tt} \cdot w + \mu \int_0^T \int_{\mathcal{B}} \nabla u \cdot \nabla w + \kappa \int_0^T \int_{\mathcal{B}} \nabla \nu \cdot \nabla w \\ &= \int_0^T \int_{\partial\mathcal{B}} w \cdot \Big(\mu \frac{\partial u}{\partial n} + \kappa \frac{\partial \nu}{\partial n}\Big) + \xi \int_0^T \int_{\mathcal{B}} \nabla(\operatorname{div} u) \cdot w + \bar{\xi} \int_0^T \int_{\mathcal{B}} \nabla(\operatorname{div} \nu) \cdot w \end{aligned} \tag{18}$$

$$\begin{aligned} &\int_0^T \int_{\mathcal{B}} (\varsigma \nu_t + \kappa_0 \nu) \cdot h + \zeta \int_0^T \int_{\mathcal{B}} \nabla \nu \cdot \nabla h + \kappa \int_0^T \int_{\mathcal{B}} \nabla u \cdot \nabla h \\ &= \int_0^T \int_{\partial\mathcal{B}} h \cdot \Big(\kappa \frac{\partial u}{\partial n} + \zeta \frac{\partial \nu}{\partial n}\Big) + \gamma \int_0^T \int_{\mathcal{B}} \nabla(\operatorname{div} \nu) \cdot h + \bar{\xi} \int_0^T \int_{\mathcal{B}} \nabla(\operatorname{div} u) \cdot h, \end{aligned} \tag{19}$$

where, for the sake of conciseness, we have erased all volume, surface and time measures from the space-time integrals above, as we do below. An adapted Galerkin's approximation procedure, combined with a compactness argument (the Aubin-Lions lemma) and suitable a priori estimates, implies a first result.

Theorem 3.1. *[2] Assume* $\mu > -\lambda$, $\kappa > 0$, $\bar{\xi} > 0$, $\mu, \zeta > 2\kappa$, *and* $\xi, \gamma > 2\bar{\xi}$. *Assume also* $u_0, \nu_0 \in W^{1,2}(\mathcal{B})$ *so that* $\nabla u(0,x) = \nabla u_0$ *and* $\nabla \nu(0,x) = \nabla \nu_0(x)$ *on* $\mathcal{B}$ *and* $\bar{u}, \bar{\nu} \in L^2(\partial\mathcal{B})$. *Then, a unique regular weak solution to the problem* (15) *exists.*

3.2 Dynamics with phason diffusion and non-linear gyroscopic phason inertia

In the presence of gyroscopic-type phason inertia, previous problem becomes

$$\begin{aligned} &\rho u_{tt} = \mu \Delta u + \xi \nabla \operatorname{div} u + \kappa \Delta \nu + \bar{\xi} \nabla \operatorname{div} \nu && \text{in } \mathcal{B}_T, \\ &\varsigma \nu_t + \ell(\operatorname{curl} u_t) \times \nu_t = \zeta \Delta \nu + \gamma \nabla \operatorname{div} \nu + \kappa \Delta u + \bar{\xi} \nabla \operatorname{div} u - \kappa_0 \nu && \text{in } \mathcal{B}_T, \\ &u(t,x) = \bar{u}(x),\ \nu(t,x) = \bar{\nu}(x), && \text{on } \partial\mathcal{B}_T, \\ &u|_{t=0} = u_0,\ u_t|_{t=0} = \dot{u}_0,\ \nu|_{t=0} = \nu_0, && \text{on } \mathcal{B}. \end{aligned} \tag{20}$$

Definition 3.2 (Weak solution)**.** *We say that a pair* (u,ν) *is a* ***weak solution*** *to the system* (20) *if, for a given* $T > 0$, *the conditions defined below hold true.*

 ISBN: 978-84-617-8694-7

(1) Regularity:

$$u \in L^\infty(0,T;\mathcal{H}^1) \cap C([0,T];\mathcal{H}^1),\ \nu \in L^2(0,T;\mathcal{H}^1) \cap C([0,T];\mathcal{H}^1), \tag{21}$$

$$\begin{aligned} &u_t \in C([0,T];L^2(\mathcal{B})) \cap L^2(0,T;W^{1,2}(\mathcal{B})),\ u_{tt} \in L^2(0,T;\mathcal{H}^{-1}),\\ &\nu_t \in L^2(0,T;W^{1,2}(\mathcal{B})). \end{aligned} \tag{22}$$

(2) Weak formulation: For all $(w,h) \in C_0^\infty([0,T[\times\mathcal{B}) \times C_0^\infty([0,T[\times\mathcal{B})$,

$$\begin{aligned} &\rho\int_0^T\int_{\mathcal{B}} u_{tt}\cdot w + \mu\int_0^T\int_{\mathcal{B}} \nabla u\cdot\nabla w + \kappa\int_0^T\int_{\mathcal{B}} \nabla\nu\cdot\nabla w\\ &= \int_0^T\int_{\partial\mathcal{B}} w\cdot\Big(\mu\frac{\partial u}{\partial n} + \kappa\frac{\partial\nu}{\partial n}\Big) + \xi\int_0^T\int_{\mathcal{B}} \nabla(\operatorname{div} u)\cdot w + \bar\xi\int_0^T\int_{\mathcal{B}} \nabla(\operatorname{div}\nu)\cdot w, \end{aligned} \tag{23}$$

$$\begin{aligned} &\int_0^T\int_{\mathcal{B}} (\varsigma\nu_t + \kappa_0\nu)\cdot h + \ell\int_0^T\int_{\mathcal{B}} (\operatorname{curl} u_t)\times\nu_t\cdot h + \int_0^T\int_{\mathcal{B}} (\zeta\nabla\nu + \kappa\nabla u)\cdot\nabla h\\ &= \int_0^T\int_{\partial\mathcal{B}} h\cdot\Big(\kappa\frac{\partial u}{\partial n} + \zeta\frac{\partial\nu}{\partial n}\Big) + \gamma\int_0^T\int_{\mathcal{B}} \nabla(\operatorname{div}\nu)\cdot h + \bar\xi\int_0^T\int_{\mathcal{B}} \nabla(\operatorname{div} u)\cdot h. \end{aligned} \tag{24}$$

Boundary terms vanish because we have selected test functions in C_0^∞. We analyze the regularized counterpart of problem (20), obtained by introducing dissipative components of the stresses, at fixed the parameters $\epsilon > 0$ and $\delta > 0$:

$$\begin{aligned} &\rho u_{tt} - \epsilon\Delta u_t = \mu\Delta u + \xi\nabla\operatorname{div} u + \kappa\Delta\nu + \bar\xi\nabla\operatorname{div}\nu && \text{in } \mathcal{B}_T,\\ &\varsigma\nu_t - \delta\Delta\nu_t + \ell(\operatorname{curl} u_t)\times\nu_t = \zeta\Delta\nu + \gamma\nabla\operatorname{div}\nu + \kappa\Delta u + \bar\xi\nabla\operatorname{div} u - \kappa_0\nu && \text{in } \mathcal{B}_T,\\ &u(t,x) = \bar u(x),\ \nu(t,x) = \bar\nu(x), && \text{on } \partial\mathcal{B}_T,\\ &u|_{t=0} = u_0,\ u_t|_{t=0} = \dot u_0,\ \nu|_{t=0} = \nu_0, && \text{on } \mathcal{B}. \end{aligned} \tag{25}$$

Theorem 3.2 (see Theorem 4.1 in [2])**.** *Assume* $\mu > -\lambda$, $\kappa > 0$, $\bar\xi > 0$, $\mu,\zeta > 2\kappa$, *and* $\xi,\gamma > 2\bar\xi$. *Assume also* $u_0,\nu_0 \in W^{1,2}(\mathcal{B})$, $\dot u_0 \in W^{1,2}(\mathcal{B})$, *such that* $\ell\|\dot u_0\|_{1,2} < \varsigma/2$ *and that* $\bar u,\bar\nu \in L^2(\partial\mathcal{B})$. *Then, the system* (20) *admits a weak solution.*

4 Weak global attractor for the time-shift semiflow

Let (W,d) be a metric space. A *semigroup* on (W,d) is a family of operators $\big(S(t)\big)_{t\geq 0}$, $S(t)\colon W \to W$, that satisfies $S(0)w = w$ and $S(s)S(t)w = S(t+s)w$ for each $w \in W$ and for every $s,t \geq 0$. A *semiflow* on (W,d) is a mapping $\sigma\colon [0,+\infty)\times W \to W$ defined by $\sigma(t,w) = S(t)w$, where $\big(S(t)\big)_{t\geq 0}$ is a semigroup, and such that the restriction $\sigma\colon]0,+\infty)\times W \to W$ is continuous. A bounded subset $\mathfrak{B} \subset W$ is called an *absorbing set* if for any bounded set B of W, there exists $t_1 = t_1(B)$ such that $S(t)B \subseteq \mathfrak{B}$ for all $t \geq t_1$. A semiflow is said to

 ISBN: 978-84-617-8694-7

be *compact* if, for every bounded set $B \subset W$ and for every $t > 0$, $S(t)B$ lies in compact subset of W.

We have the following result [18, 19, 21] (see also [1]).

Theorem 4.1. *Let $S(t)$ define a compact semiflow admitting an absorbing set $\mathfrak{B}$ on a complete metric space W. Then $S(t)$ has a global attractor $\mathcal{A}$ in W and coincides with the omega-limit set of $\mathfrak{B}$:*

$$\mathcal{A} = \bigcap_{\tau \geq 0} \overline{\bigcup_{t \geq \tau} S(t)\mathfrak{B}}\,,$$

where the closure is taken in W.

Starting from problem (25), we introduce a bulk force $f(x,t)$ in the balance of standard force and rewrite the whole system as

$$\begin{aligned}
&u_t = v && \text{in } \mathcal{B}_T,\\
&\rho v_t - \epsilon\Delta v = \mu\Delta u + \xi\nabla \operatorname{div} u + \kappa\Delta\nu + \bar{\xi}\nabla \operatorname{div}\nu + f(t,x) && \text{in } \mathcal{B}_T,\\
&\varsigma\nu_t - \delta\Delta\nu_t + \ell(\operatorname{curl} v)\times\nu_t = \zeta\Delta\nu + \gamma\nabla\operatorname{div}\nu + \kappa\Delta u + \bar{\xi}\nabla\operatorname{div} u - \kappa_0\nu && \text{in } \mathcal{B}_T,\\
&u(t,x) = 0,\ \nu(t,x) = 0, && \text{on } \partial\mathcal{B}_T,\\
&u|_{t=0} = u_0,\ v|_{t=0} = \dot{u}_0,\ \nu|_{t=0} = \nu_0, && \text{on } \mathcal{B},
\end{aligned} \tag{26}$$

with homogeneous Dirichlet boundary conditions. Although the forcing term f does not appear in the existence result ([2, Thereom 4.1]), the proof can be adapted to the present case, with very minor changes, by assuming that $f \in L^2(0,\infty;L^2(\mathcal{B}))$.

Let us write $\boldsymbol{w}(t) = (u(t),\nu(t);u_t(t)) = (u(t),\nu(t);v(t))$ for brevity, and denote by

$$\begin{aligned}
\boldsymbol{\mathcal{W}} = \big\{\boldsymbol{w} = (u,\nu;v) \in L^2_{\mathrm{loc}}[0,+\infty;W^{1,2}(\mathcal{B})^2)\times L^2_{\mathrm{loc}}[0,+\infty;L^2(\mathcal{B}))\mid v(t) = u_t(t)\\
\text{and } \boldsymbol{w} \text{ satisfying (21)-(22) and the weak formulation of (26) as in Def. 3.2}\big\}
\end{aligned} \tag{27}$$

the set of the weak solutions to (26). In the sequel, the role of W will be played by the set $\boldsymbol{\mathcal{W}}$ of the weak solutions to (26) in $L^2_{\mathrm{loc}}[0,\infty;W^{1,2}(\mathcal{B})^2)\times L^2_{\mathrm{loc}}[0,\infty;L^2(\mathcal{B}))$, with metric d defined by

$$d(\boldsymbol{w}_1,\boldsymbol{w}_2) = \sum_{n=0}^{\infty} 2^{-n}\min\big\{1,\ |||\boldsymbol{w}_1 - \boldsymbol{w}_2|||_{L^2(0,n)}\big\}. \tag{28}$$

where, given $\boldsymbol{w} = (u,v;\nu)$, with L^2-norm $\|\cdot\|$, we have that

$$|||\boldsymbol{w}|||^2_{L^2(a,b)} = \int_a^b |||\boldsymbol{w}(s)|||^2 ds \ \text{ and } |||\boldsymbol{w}(t)|||^2 = (\|u\|^2 + \|\nabla u\|^2 + \|\nu\|^2 + \|\nabla\nu\|^2 + \|v\|^2)(t).$$

 ISBN: 978-84-617-8694-7

We also recall that a set B in a linear topological space $\mathcal{Z}$ is called *bounded* if for every neighborhood U of the origin in $\mathcal{Z}$ there exists an $r > 0$ such that $B \subset \{ru\colon u \in U\}$. In the case of the space $\mathcal{Z} = L^2_{\rm loc}[0,\infty; W^{1,2}(\mathcal{B})^2) \times L^2_{\rm loc}[0,\infty; L^2(\mathcal{B}))$, this reduces to

$$\sup\left\{|||\boldsymbol{w}|||_{L^2(0,n)} \,|\, \boldsymbol{w} \in B\right\} < +\infty, \ \ \forall\, n = 0,1,2,\dots$$

Actually our aim would be to prove a proposition as follows:

Proposition 4.1. *The time-shift operator $S(t)$, associated with* (26)*, admits a unique global attractor $\boldsymbol{\mathcal{A}}$ in $\boldsymbol{\mathcal{W}}$.*

Due to the possible lack of uniqueness for the weak solutions of (26), we find convenient to follow the approach developed by Sell in [18], which is suitable for non-well-posed problems, and to perform our dynamical analysis on the phase space given by $\boldsymbol{\mathcal{W}}$, in which each point is a weak solution.

In order to prove Proposition 4.1, we need to show that the time-shift operator $S(t)$ verifies the hypotheses of Theorem 4.1. That $\sigma(t,\boldsymbol{w}) = S(t)(\boldsymbol{w})$, $(t,\boldsymbol{w}) \in]0,+\infty) \times \boldsymbol{\mathcal{W}}$ is a semiflow can be proven by following a lemma by Sell [18, Lemma 7].

Proposition 4.2. *The mapping $\sigma\colon]0,+\infty) \times \boldsymbol{\mathcal{W}} \to \boldsymbol{\mathcal{W}}$ given by $S(t)\boldsymbol{w} = \boldsymbol{w}(\cdot + t)$ is a semiflow.*

Consider the following map

$$z(t) = \int_{\mathcal{B}} \Big(\rho|u_t+\alpha u|^2+\alpha^2|u|^2+(\mu-\epsilon\alpha)|\nabla u|^2+(\kappa_0+\alpha\varsigma)|\nu|^2+(\zeta+\alpha\delta)|\nabla\nu|^2\Big)(t)\,\mathrm{d}x \approx |||\boldsymbol{w}(t)|||$$

where α is a positive parameter which appears in the next result, and $\mu - \epsilon\alpha > 0$ (for a possible similar situation, see also [4, Ch. VI, §4])

Lemma 4.1. *Taking $\alpha > 0$ sufficiently small, then the following relation holds true*

$$z(t) \le z(\tau)\exp\{-\tilde{\beta}(t-\tau)\} + \tilde{c}, \ \ \textit{with } t \ge \tau \ge 0, \tag{29}$$

where $\tilde{\beta}$ and $\tilde{c}$ are positive constants.

The proof of such a lemma and the ones of all statements in this section are in [3].

As a consequence we have the existence of an absorbing set.

Proposition 4.3. *There exists an absorbing set $\mathfrak{B} \subset \boldsymbol{\mathcal{W}}$ that is bounded in $\boldsymbol{\mathcal{W}}$.*

Thus, Proposition 4.1 would be a theorem, indeed, after proving the following conjecture.

Conjecture 4.1. *The semiflow defined by $S(t)$ on $\boldsymbol{\mathcal{W}}$ is compact, i.e for each bounded set B in $\boldsymbol{\mathcal{W}}$ and for each $t > 0$, then $S(t)B$ lies in a compact subset of $\boldsymbol{\mathcal{W}}$.*

 ISBN: 978-84-617-8694-7

References

[1] L. Bisconti and D. Catania, *Remarks on global attractors for the 3D Navier–Stokes equations with horizontal filtering*, Discrete Contin. Dyn. Syst. Ser. B, **20 no. 1** (2015), 59–75.

[2] L. Bisconti and P. M. Mariano, *Existence results in the linear dynamics of quasicrystals with phason diffusion and nonlinear gyroscopic effects*, Multiscale Model. Simul. **15 no. 2** (2017), 745–767

[3] L. Bisconti and P. M. Mariano, *Weak attractors in the linear dynamics of quasicrystals with phason diffusion and nonlinear gyroscopic effects*, forthcoming.

[4] V. V. Chepyzhov, and M. I. Vishik, *Attractors for equations of mathematical physics*, American Mathematical Society Colloquium Publications, 49. American Mathematical Society, Providence, RI, 2002.

[5] M. Duneau and A. Kat, *Quasiperiodic patterns*, Phys. Rev. Lett., **54** (1985), 2688–2691.

[6] P. De and R. A. Pelcovits, *Linear elasticity theory of pentagonal quasicrystals*, Phys. Rev. B, **35** (1987), 8609–8620.

[7] M. Focardi, P. M. Mariano and E. Spadaro, *Multi-value microstructural descriptors for complex materials: analysis of ground states*, Arch. Rational Mech. Anal., **217** (2015), 899–933.

[8] C. Hu, R. Wang and D. H. Ding, *Symmetry groups, physical property tensors, elasticity and dislocations in quasicrystals*, Rep. Prog. Phys., **63** (2000), 1–39.

[9] International Union of Crystallography. Report of the Executive Committee for 1991, *Acta Cryst.* (1992), **A48**, 922–946.

[10] H.-C. Jeong and P. J. Steinhardt, *Finite-temperature elasticity phase transition in decagonal quasicrystals*, Phys. Rev. B, **48** (1993), 9394–9403.

[11] T. C. Lubensky, S. Ramaswamy and J. Toner, *Hydrodynamics of icosahedral quasicrystals*, *Phys. Rev. B*, **32** (1985), 7444–7452.

[12] P. M. Mariano, *Multifield theories in mechanics of solids*, Adv. Appl. Mech., **38** (2002), 1–93.

[13] P. M. Mariano, *Mechanics of quasi-periodic alloys*, J. Nonlinear Sci., **16** (2006), 45–77.

 ISBN: 978-84-617-8694-7

[14] P. M. Mariano, *Mechanics of material mutations*, Adv. Appl. Mech. **47** (2014), 1–91.

[15] P. M. Mariano and G. Modica, *Ground states in complex bodies*, ESAIM – Control, Optimization and Calculus of Variations, **15** (2009), 377–402.

[16] P. M. Mariano and J. Planas J., *Self-actions in quasicrystals*, *Physica D*, **249** (2013), 46–57.

[17] S. B. Rochal and V. L. Lorman, *Minimal model of the phonon-phason dynamics in icosahedral quasicrystals and its application to the problem of internal friction in the* i-*AlPbMn alloy*, Physical Review B, **66** (2002), pp. 1–9.

[18] G. R. Sell, *Global attractor for the three-dimensional Navier–Stokes equations*, J. Dynam. Differential Equations **8 n. 1** (1996), 1–33.

[19] G. R. Sell and Y. You, *Dynamics of evolutionary equations*, Applied Mathematical Sciences 143, Springer, 2002.

[20] D. Shechtman, I. Blech, D. Gratias and J. W. Cahn, *Metallic phase with long-range orientational order and no translational symmetry*, Phys. Rev. Letters, **53** (1984), 1951–1954.

[21] R. Temam, *Infinite dimensional dynamical systems in mechanics and physics*, Applied Mathematical Sciences 68, Springer, 1988.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

Triangular PN patches subject to surface-area constraints

Michal Bizzarri[1] **and Miroslav Lávička**[1,2]

[1] *New Technologies for the Information Society, Faculty of Applied Sciences, University of West Bohemia, Univerzitní 8, 306 14 Plzeň, Czech Republic*

[2] *Department of mathematics, Faculty of Applied Sciences, University of West Bohemia, Univerzitní 8, 306 14 Plzeň, Czech Republic*

emails: bizzarri@ntis.zcu.cz, lavicka@kma.zcu.cz

Abstract

This paper is devoted to the construction of polynomial surfaces with Pythagorean normals (PN surfaces) interpolating given data subject to prescribed constraints on the surface area of the patch. This is a problem analogous to the interpolation with Pythagorean hodograph (PH) curves satisfying the condition on the arc length. The special structure of PN surfaces allows the surface-area condition to be expressed as algebraic constraints on the surfaces coefficients. We employ these shapes for solving the G^1 Hermite interpolation problem by triangular PN patches with prescribed surface area. The presented technique is based on interpolating points on the unit sphere and consequently on solving a system of several linear and one quadratic equations. We show that for generic input data there exist at most two quartic PN patches depending on the particular value of the prescribed surface area.

Key words: G^1 Hermite interpolation; PN surfaces; polynomial area element

1 Introduction

Rational surfaces with Pythagorean normals (PN surfaces) were introduced in [9] as a surface analogy to Pythagorean hodograph (PH) curves defined before in [4]. For a survey of shapes with Pythagorean property see e.g. [3] and references therein. It holds that PH curves in plane and PN surfaces in space share some common properties, e.g. they both yield rational offsets. Nevertheless there also exist significant differences between these classes of rational varieties. For instance, the curves with Pythagorean hodographs were introduced as planar *polynomial* shapes and a compact formula for their description based on Pythagorean triples

 ISBN: 978-84-617-8694-7

of polynomials was presented. On the other hand, a description of *rational* Pythagorean normal vector surfaces reflecting their dual description was revealed first in [9] and it is still not known how to gain the polynomial subclass from this formula. Nonetheless, new attempts to the investigation of PN surfaces have emerged recently, see [6, 8, 1].

The area element $\mathrm{d}A$, and thus also the surface area $A(u,v) = \iint \sqrt{EG - F^2}\,\mathrm{d}u\mathrm{d}v$ of a polynomial PN surface $\mathbf{x}(u,v)$ is polynomial, in general (which is not the case for their rational counterparts). This special algebraic structure of polynomial PN surfaces offers many useful computational advantages over "standard" polynomial parametric surfaces. In addition, this also underlines a prominent role of polynomial PN surfaces and relates them suitably with polynomial PH curves that possess polynomial length element. Unfortunately, there is not known very much about them. We recall at least the investigation of a family of cubic polynomial PN surfaces in [7] and a novel approach to polynomial PN surfaces based on bivariate polynomials with quaternion coefficients, presented recently in [6].

A new technique to the shape reconstruction is presented herein, based on the polynomial PN surfaces. We will continue in the most recent approach from [1], where the interpolation by polynomial PN surfaces is transformed to solving a system of linear equations. As an extra added value, we consider the conditions on the surface-area. The special structure of PN surfaces allows the surface-area condition to be expressed as algebraic constraints on the surfaces coefficients. This is a problem analogous to the interpolation with Pythagorean hodograph (PH) curves satisfying the condition on the arc length, see e.g. [5]. In more detail, we show that three points and three tangent vector can be interpolated by 1-parameter family of polynomial PN surfaces of degree four. These surfaces will be consequently used for constructing G^1 interpolants with the prescribed surface area. Moreover, for a generic data, the existence of 0, 1 or 2 quartic PN patches with prescribed surface area L depending on the particular value of L will be shown. Solving this problem can be useful everywhere where the surface area is important such as in CNC painting.

The remainder of this paper is organized as follows. Section 2 recalls some basic facts concerning polynomial PN surfaces and presents shortly the method from [1] for their generation. In Section 3, these results are applied to a practical problem of G^1 Hermite interpolation by polynomial PN triangular patches which will be used in Section 4 to compute interpolants with prescribed surface area. Finally, we conclude the paper in Section 5.

2 Polynomial PN surfaces

A parametric surface $\mathbf{x}(u,v)$ in $\mathbb{R}^3$ is called a surface with *Pythagorean normal vector field* (*a PN surface*) if there exists a rational function $\sigma(u,v)$ such that it is satisfied

$$\|\mathbf{n}(u,v)\|^2 = \sigma(u,v)^2, \tag{1}$$

 ISBN: 978-84-617-8694-7

where $\mathbf{n}(u,v) = \mathbf{x}_u(u,v) \times \mathbf{x}_v(u,v)$ is a normal vector field of $\mathbf{x}(u,v)$.Clearly, the distinguishing property of PN surfaces is the rationality of their two-sided δ-offset surfaces

$$\mathbf{x}_\delta = \mathbf{x} \pm \delta \frac{\mathbf{n}}{\|\mathbf{n}\|} = \mathbf{x} \pm \delta \, \frac{\mathbf{x}_u \times \mathbf{x}_u}{\sigma}. \tag{2}$$

Next, we recall that the squared *area element* of the parametric surface $\mathbf{x}(u,v)$ has the form

$$\mathrm{d}A^2 = \begin{vmatrix} \mathbf{x}_u \cdot \mathbf{x}_u & \mathbf{x}_u \cdot \mathbf{x}_v \\ \mathbf{x}_u \cdot \mathbf{x}_v & \mathbf{x}_v \cdot \mathbf{x}_v \end{vmatrix} \mathrm{d}u^2 \mathrm{d}v^2 = (EG - F^2)\, \mathrm{d}u^2 \mathrm{d}v^2, \tag{3}$$

where $E = \mathbf{x}_u \cdot \mathbf{x}_u$, $F = \mathbf{x}_u \cdot \mathbf{x}_v$, $G = \mathbf{x}_v \cdot \mathbf{x}_v$ are the coefficients of the first fundamental form. Moreover, as it holds

$$\|\mathbf{x}_u \times \mathbf{x}_v\|^2 = (EG - F^2), \tag{4}$$

the family of rational PN surfaces (i.e., rational offsets surfaces) coincides with the family of surfaces with rational area element. In addition, all polynomial PN surfaces possess polynomial surface area $A(u,v) = \iint \sqrt{EG - F^2}\, \mathrm{d}u\mathrm{d}v$. This offers many useful computational advantages compared to other surfaces (including rational PN surfaces).

However until recently, the construction of PN surfaces was based only on their dual representation, i.e., any rational PN surface is represented as the envelope of its tangent planes

$$\mathbf{n}(u,v) \cdot \mathbf{x} = h(u,v), \tag{5}$$

where $\mathbf{n}(u,v)$ is a polynomial normal vector field satisfying (1) and $h(u,v)$ is a rational function. Differentiating (5) with respect to u and v gives 3 linear equations in variables x_i

$$\mathbf{M}\,\mathbf{x} = \mathbf{h}, \qquad \text{where} \quad \mathbf{M} = (\mathbf{n}, \mathbf{n}_u, \mathbf{n}_v)^\top \ \text{ and } \ \mathbf{h} = (h, h_u, h_v)^\top. \tag{6}$$

Solving (6) we arrive at $\mathbf{x}(u,v) = \mathbf{M}^{-1}\,\mathbf{h}$, a description of non-developable PN surfaces, cf. [9]. Unfortunately this method is not suitable for computing parameterizations of *polynomial* PN surfaces. Nevertheless a novel direct approach based on solving a suitable system of linear equations was recently introduced in [1].

In more detail, when looking for some polynomial PN surface (i.e., polynomial surface with polynomial area element) we prescribe first a suitable polynomial normal vector field $\mathbf{n}(u,v)$ of degree k satisfying (1). Its parameterization can be easily gained for instance from polynomial Pythagorean quadruples, cf. [2]. Then we determine an associated polynomial PN parameterization of degree ℓ in a direct way, i.e., we find a suitable polynomial patch

$$\mathbf{x}(u,v) = \left(\sum_{i+j\le\ell} x_{1ij} u^i v^j, \sum_{i+j\le\ell} x_{2ij} u^i v^j, \sum_{i+j\le\ell} x_{3ij} u^i v^j \right)^\top, \tag{7}$$

 ISBN: 978-84-617-8694-7

such that it satisfies the following conditions:

$$\mathbf{x}_u(u,v)\cdot\mathbf{n}(u,v)\equiv 0 \qquad \text{and} \qquad \mathbf{x}_v(u,v)\cdot\mathbf{n}(u,v)\equiv 0. \tag{8}$$

Hence the problem is now transformed to solving a system of $\binom{k+\ell+1}{2}$ homogeneous linear equations with $3\binom{\ell+2}{2}$ unknowns $x_{1ij}, x_{2ij}, x_{3ij}$. This system is solvable in general for ℓ large enough. Nonetheless, we must emphasize that the method does not guarantee a polynomial surface $\mathbf{x}(u,v)$ for which $\mathbf{x}_u\times\mathbf{x}_v=\mathbf{n}(u,v)$. We arrive at a polynomial PN parameterization such that

$$\mathbf{x}_u(u,v)\times\mathbf{x}_v(u,v)=f(u,v)\,\mathbf{n}(u,v), \tag{9}$$

where $f(u,v)$ is a factor relating suitably the degrees of $\mathbf{n}(u,v)$ and $\mathbf{x}(u,v)$, see [1] for more details. Of course, the existence of a non-constant factor $f(u,v)$ does not abolish the polynomiality of the corresponding surface area $A(u,v)$.

3 Hermite interpolation with triangular PN patches

In this section we present a *direct* method for interpolating given data by triangular *polynomial* PN surfaces (surfaces with polynomial area elemnts).

Consider three points $\mathbf{p}_{ij}$, $i,j=0,1$ and $i+j<2$, and three associated tangent planes τ_{ij} determined by the unit normal vectors $\mathbf{N}_{ij}$. Following the ideas presented in the previous section, we can formulate the whole algorithm consisting of two subparts:

(i) We construct a normal vector field $\mathbf{n}(u,v)$ interpolating data $\mathbf{n}_{ij}=\lambda_{ij}\mathbf{N}_{ij}$, $\lambda_{ij}\in\mathbb{R}^+$, and having the polynomial norm.

(ii) We compute a polynomial patch interpolating the points $\mathbf{p}_{ij}$ and possessing the normal vector field $\mathbf{n}(u,v)$ via solving (8).

As concerns (i), we apply a well-known method based on using the stereographic projection. We choose a suitable center $\mathbf{w}$ of the stereographic projection (such that it is not contained in the the Gauss image of the interpolating surface) and project data $\mathbf{N}_{ij}\in\mathcal{S}^2$ to the plane $\mathbb{R}^2$, i.e.,

$$\widehat{\mathbf{N}}_{ij}=\pi_{\mathbf{w}}(\mathbf{N}_{ij})=\mathbf{w}+\frac{(\mathbf{N}_{ij}-\mathbf{w})}{1-\mathbf{w}\cdot\mathbf{N}_{ij}}. \tag{10}$$

Then, we construct the linear patch in $\mathbb{R}^2$ interpolating $\pi_{\mathbf{w}}(\mathbf{N}_{ij})$, i.e.,

$$\widehat{\mathbf{N}}(u,v)=\widehat{\mathbf{N}}_{10}\,u+\widehat{\mathbf{N}}_{01}\,v+\widehat{\mathbf{N}}_{00}\,(1-u-v)\quad u\in[0,1],\ v\in[0,1-u]. \tag{11}$$

Finally, employing the inverse stereographic projection and omitting the least common denominator, we arrive at the quadratic polynomial normal vector field

$$\mathbf{n}(u,v)=2\widehat{\mathbf{N}}(u,v)+\left(\widehat{\mathbf{N}}(u,v)\cdot\widehat{\mathbf{N}}(u,v)-1\right)\mathbf{w} \tag{12}$$

 ISBN: 978-84-617-8694-7

with the PN property

$$\|\mathbf{n}(u,v)\|^2 = \left(\widehat{\mathbf{N}}(u,v)\cdot\widehat{\mathbf{N}}(u,v)+1\right)^2 = \sigma(u,v)^2, \tag{13}$$

and satisfying the prescribed interpolation conditions

$$\mathbf{n}(i,j) = \lambda_{ij}\,\mathbf{N}_{ij}, \qquad \lambda_{ij} = \widehat{\mathbf{N}}_{ij}\cdot\widehat{\mathbf{N}}_{ij}+1 \in \mathbb{R}^+. \tag{14}$$

Once we have a quadratic vector field $\mathbf{n}(u,v)$ we can continue with part (ii). Our goal is to find a polynomial patch (7) of prescribed degree ℓ possessing $\mathbf{n}(u,v)$ as its associated normal vector field and interpolating given position data, i.e., it must hold (8) and

$$\mathbf{x}(i,j) = \mathbf{p}_{ij}. \tag{15}$$

To conclude the method, expressions (8) and (15) depend linearly on the coefficients $x_{1ij}, x_{2ij}, x_{3ij}$ of $\mathbf{x}(u,v)$ and therefore can be rewritten as a system of linear equations yielding always a solution. Hence we can formulate

Theorem 3.1 *For generic points $\mathbf{p}_{ij}$ and associated tangent planes τ_{ij}, solving the equations from systems* (8) *and* (15) *yields a* 1*-parameter set of quartic polynomial PN patches interpolating the points $\mathbf{p}_{ij}$ and touching the planes τ_{ij} at these points.*

Proof: Using computer algebra system Mathematica we have verified symbolically that the system of linear equation (8) and (15) is solvable and for quartic surface (7) the rank of its matrix is equal to 44 whereas the number of coefficients of (7) is 45. Moreover from (15) and the construction of the normal vector field $\mathbf{n}(u,v)$, the constructed surface interpolates the points $\mathbf{p}_{ij}$ and is tangent to the planes τ_{ij} at these points. □

4 Triangular PN patches with prescribed surface area

Now, we are able to design a simple algorithm for the construction of triangular quartic PN patches with the prescribed surface area.

Consider three points $\mathbf{p}_{ij}$, $i,j = 0,1$ and $i+j<2$, three associated unit normal vectors $\mathbf{N}_{ij}$ and in addition a value L, the prescribed surface area of the patch. First, employing the method described in more detail in the previous section, we construct a 1-parametric family of quartic triangular patches interpolating $\mathbf{p}_{ij}$ and $\mathbf{N}_{ij}$, i.e., we have

$$\mathbf{x}(u,v,\alpha) = \mathbf{x}_1(u,v) + \alpha\,\mathbf{x}_2(u,v). \tag{16}$$

Then

$$F(u,v,\alpha) = \|\mathbf{x}_u(u,v,\alpha)\times\mathbf{x}_v(u,v,\alpha)\| = f(u,v,\alpha)\,\sigma(u,v), \tag{17}$$

 ISBN: 978-84-617-8694-7

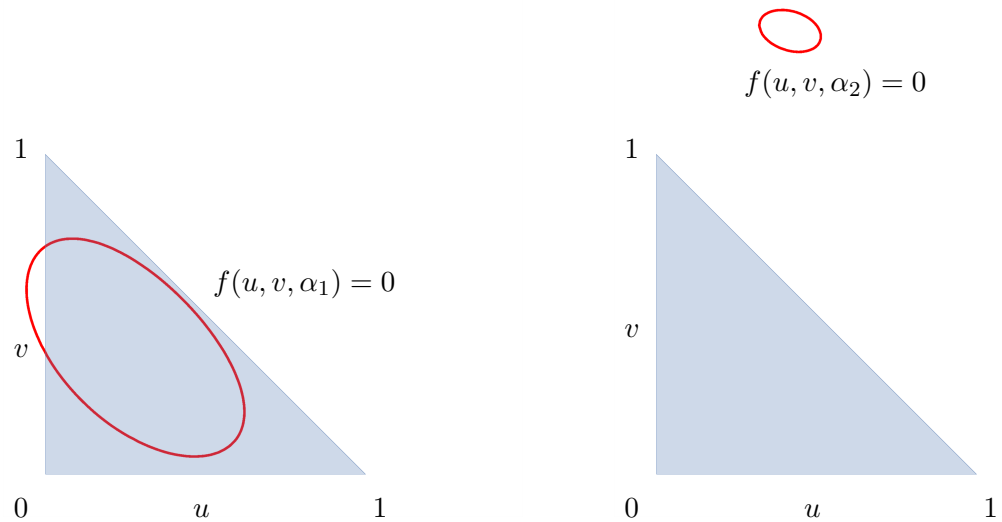


Figure 1: Curves in the parametr plane defined by the extra factors $f(u,v,\alpha_i)$ which correspond to singular curves on the resulting surfaces from Example 4.1.

where $\sigma(u,v)$ is given by (13) and $f(u,v,\alpha)$ is a quartic polynomial in u,v with coefficients quadratic in α. The corresponding surface area is then equal to

$$\int_0^1 \int_0^{1-u} F(u,v,\alpha)\,\mathrm{d}v\,\mathrm{d}u = A\alpha^2 + B\alpha + C = L, \quad A,B,C \in \mathbb{R}. \tag{18}$$

Hence for

$$L \geq \frac{4AC - B^2}{4A}, \tag{19}$$

we always obtain a solution, i.e, a quartic PN patch interpolating given data and subject to the surface-area constraint. In particular the interpolation problem has one solution in the case of equality and two solutions in the case of strong inequality in (19).

Example 4.1 Consider three points

$$\mathbf{p}_{00} = (0,0,0)^\top, \quad \mathbf{p}_{10} = (2,-1,1)^\top, \quad \mathbf{p}_{01} = (2,2,-1)^\top, \tag{20}$$

and the associated unit normal vectors

$$\mathbf{N}_{00} = (0,0,-1)^\top, \quad \mathbf{N}_{10} = \left(\frac{2}{3},-\frac{1}{3},-\frac{2}{3}\right)^\top, \quad \mathbf{N}_{01} = \left(-\frac{1}{3},-\frac{2}{3},-\frac{2}{3}\right)^\top. \tag{21}$$

Our goal is to construct a triangular PN patch interpolating the prescribed points and normals and possessing the surfaces area equal to $L = 4$.

 ISBN: 978-84-617-8694-7

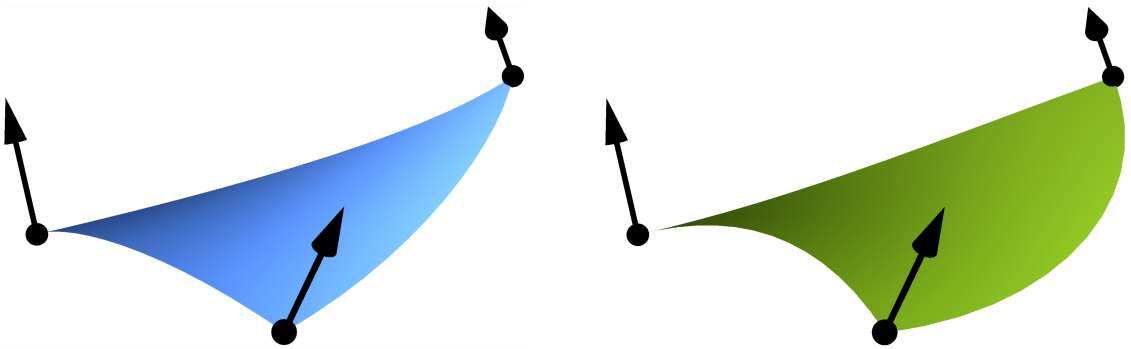


Figure 2: Quartic PN patch interpolating data (20) and (21) having areal element equal to 4 (left) and 4.5 (right) from Example 4.1.

We use the stereographic projection with $\mathbf{w} = (0, 0, 1)$, cf. (10), and construct the linear triangular planar patch (11), i.e.,

$$\widehat{\mathbf{N}}(u, v) = \left(\frac{1}{5}(2u - v), \frac{1}{5}(-u - 2v), 0\right)^{\top}. \tag{22}$$

Then lifting $\widehat{\mathbf{N}}$ back on the unit sphere $\mathcal{S}^2$ and omitting the denominator, cf. (12), yields the polynomial vector field

$$\mathbf{n}(u, v) = \left(\frac{2}{5}(2u - v), -\frac{2}{5}(u + 2v), \frac{1}{5}\left(u^2 + v^2 - 5\right)\right)^{\top} \tag{23}$$

fulfilling the PN condition

$$\|\mathbf{n}(u, v)\|^2 = \left[\frac{1}{5}\left(u^2 + v^2 + 5\right)\right]^2. \tag{24}$$

Then, we prescribe a polynomial parameterization (7) of degree four and solve the systems of linear equations (8) and (15) which yields 1-parametric solution (16). Finally computing the surface area (18) and solving

$$L = \frac{16001}{3686400}\alpha^2 + \frac{1219717}{6874560}\alpha + \frac{111456586}{26163225} = 4, \tag{25}$$

we arrive at the two values of the parameter $\alpha_1 \doteq -39.3537$ and $\alpha_2 \doteq -1.52238$ from which only α_2 corresponds to a patch without singularities – this can be verified by investigating the zeros of the extra factor $f(u, v, \alpha_i)$. In particular the curve defined by $f(u, v, \alpha) = 0$ in the parameter plane u, v corresponds to the singular curve on the corresponding patch,

hence our aim is to choose parameter α such that $f(u, v, \alpha)$ does not lie in the triangle $u \in [0, 1]$, $v \in [0, 1 - u]$. As mentioned before, this is satisfied only by α_2, see Fig. 1. In Fig. 2 two non-singular PN patches (with surface area equal to 4 and 4.5) interpolating points (20) and normals (21) are shown.

5 Conclusion

This paper was devoted to the construction of polynomial triangular PN patches interpolating given G^1 data and, in addition, possessing the prescribed surface area. We have shown that the first part (without considering the surface area constraint) is always possible by quartic surfaces. Moreover since the interpolation problem yields 1-parameter family of PN surfaces, the interpolation together with ensuring the prescribed area element gives 0, 1 or 2 solutions depending on the value of the particular are element. We also discussed that the resulting surfaces may contain the singular curve corresponding to the zeroes of the extra factor $f(u, v)$. Of course the good interpolants are those having $f(u, v) = 0$ outside the parameter domain. As far as we are aware of the literature, this paper is the first contribution solving for PN surfaces a problem analogous to the interpolation with PH curves satisfying the condition on the arc length. In our future work we would like to focus in more detail on ensuring the existence such interpolants and constructing piecewise PN surfaces with a prescribed continuity and subject to surface-area constraints.

Acknowledgements

The authors are supported by the project LO1506 of the Czech Ministry of Education, Youth and Sports.

References

[1] M. Bizzarri, M. Lávička, Z. Šír, and J. Vršek, *Hermite interpolation by piecewise polynomial surfaces with polynomial area element*, Computer Aided Geometric Design, 51 (2017), pp. 30 – 47.

[2] R. Dietz, J. Hoschek, and B. Jüttler, *An algebraic approach to curves and surfaces on the sphere and on other quadrics*, Computer Aided Geometric Design, 10 (1993), pp. 211–229.

[3] R. Farouki, *Pythagorean-Hodograph Curves: Algebra and Geometry Inseparable*, Springer, 2008.

 ISBN: 978-84-617-8694-7

[4] R. Farouki and T. Sakkalis, *Pythagorean hodographs*, IBM Journal of Research and Development, 34 (1990), pp. 736–752.

[5] M. Huard, R. T. Farouki, N. Sprynski, and L. Biard, C^2 *interpolation of spatial data subject to arc-length constraints using Pythagorean–hodograph quintic splines*, Graphical Models, 76 (2014), pp. 30–42.

[6] J. Kozak, M. Krajnc, and V. Vitrih, *A quaternion approach to polynomial PN surfaces*, Computer Aided Geometric Design, 47 (2016), pp. 172–188. SI: New Developments Geometry.

[7] M. Lávička and J. Vršek, *On a special class of polynomial surfaces with Pythagorean normal vector fields*, in Curves and Surfaces, J.-D. Boissonnat, P. Chenin, A. Cohen, C. Gout, T. Lyche, M.-L. Mazure, and L. Schumaker, eds., vol. 6920 of Lecture Notes in Computer Science, Springer Berlin Heidelberg, 2012, pp. 431–444.

[8] M. Lávička, Z. Šír, and J. Vršek, *Smooth surface interpolation using patches with rational offsets*, Computer Aided Geometric Design, 48 (2016), pp. 75–85.

[9] H. Pottmann, *Rational curves and surfaces with rational offsets*, Computer Aided Geometric Design, 12 (1995), pp. 175–192.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

On monogenic functions with line singularities

Sebastian Bock[1]

[1] *Institute of Mathematics/Physics, Bauhaus-Universität Weimar*

emails: sebastian.bock@uni-weimar.de

Abstract

In this article a class of monogenic functions with (logarithmic) line singularities is presented. These functions have special properties with respect to the hypercomplex derivative (Appell property) and can be generated by a two step recurrence formula.

Key words: monogenic functions, line singularities, recurrence formulae
MSC 2000: 30G35, 33B30

1 Introduction

Nowadays, Clifford analysis is commonly understood as the higher dimensional counterpart to the complex function theory and has applications in many fields of mathematical physics. In the complex one-dimensional analysis, series expansions of holomorphic functions take a key role in solving applied as well as theoretical problems. In recent years there has been a lot of progress in finding higher dimensional analogs to the complex series expansions. In [2, 3], generalized Taylor- and Fourier series expansions and in [4] a generalized Laurent series expansion for quaternion-valued functions taking values in $\mathbb{R}^3$ were constructed. This approach was then unified in [5] for dimensions 2, 3 and 4. The construction of these series expansions was based on the use of complete orthogonal systems of inner and outer solid spherical monogenics which generalize all the essential properties of the holomorphic $\boldsymbol{z}$-monomials, such as orthogonality, Appell property, two step recurrence relations, to $\mathbb{R}^3$ and $\mathbb{R}^4$. Furthermore, these series expansions were successfully applied in solving three dimensional problems in linear elasticity theory by using recently developed higher dimensional generalizations of the Kolosov-Muskhelishvilli formulae [6, 9]. Recent results in this context indicate that the function systems and series expansions developed so far are not sufficient for solving more specific boundary value problems with point and line singularities, as for

example in the spatial fracture mechanics. In [7], a new set of monogenic functions with (logarithmic) line singularities was studied. Based on these results, a hypercomplex version of the fundamental solution in linear elasticity (Kelvin solution) was constructed, which describes the elastic displacements of a concentrated force acting at the origin of an infinite body. The Kelvin solution is of fundamental importance as one can also construct solutions for force groups or distributed forces by using the principle of superposition.

In this article a class of monogenic functions with (logarithmic) line singularities is studied which extends the results of [7]. These functions have special properties with respect to the hypercomplex derivative (Appell property) and can be generated by an associated two step recurrence formula.

2 Preliminaries

Let $\mathbb{H}$ be the algebra of real quaternions with the standard basis $\{\mathbf{e}_0, \mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3\}$ subjected to the multiplication rules

$$\begin{aligned} \mathbf{e}_i\mathbf{e}_j + \mathbf{e}_j\mathbf{e}_i &= -2\delta_{ij}\,\mathbf{e}_0,\ i,j=1,2,3, \\ \mathbf{e}_1\mathbf{e}_2 &= \mathbf{e}_3,\ \mathbf{e}_0\mathbf{e}_i = \mathbf{e}_i\mathbf{e}_0 = \mathbf{e}_i,\ i=0,1,2,3. \end{aligned}$$

The real vector space $\mathbb{R}^4$ will be embedded in $\mathbb{H}$ by identifying $\boldsymbol{a} = [a_0, a_1, a_2, a_3]^{\mathrm{T}} \in \mathbb{R}^4$ with the quaternion $\boldsymbol{a} = a_0 + a_1\mathbf{e}_1 + a_2\mathbf{e}_2 + a_3\mathbf{e}_3$, $a_i \in \mathbb{R}$, $i = 0, 1, 2, 3$, where $\mathbf{e}_0 = [1, 0, 0, 0]^{\mathrm{T}}$ is the multiplicative unit element of the algebra $\mathbb{H}$. Further, we denote by $\mathbf{Sc}(\boldsymbol{a}) = a_0$ the scalar part and by $\mathbf{Vec}(\boldsymbol{a}) = \underline{\boldsymbol{a}} = \sum_{i=1}^{3} a_i\mathbf{e}_i$ the vector part of $\boldsymbol{a}$. In addition, $\overline{\boldsymbol{a}} = a_0 - \underline{\boldsymbol{a}}$ denotes the conjugate and $|\boldsymbol{a}| = \sqrt{\boldsymbol{a}\,\overline{\boldsymbol{a}}}$ the norm of $\boldsymbol{a}$.

Now, let us consider the subset $\mathcal{A} := \mathrm{span}_{\mathbb{R}}\{1, \mathbf{e}_1, \mathbf{e}_2\}$ which is only a real vector space but not a sub-algebra of $\mathbb{H}$. The real vector space $\mathbb{R}^3$ will be embedded in $\mathcal{A}$ by the identification of $\boldsymbol{x} = [x_0, x_1, x_2]^{\mathrm{T}} \in \mathbb{R}^3$ with the *reduced quaternion*

$$\boldsymbol{x} = x_0 + \mathbf{e}_1\boldsymbol{\zeta} \in \mathcal{A} \quad \text{with} \quad \boldsymbol{\zeta} = x_1 - \mathbf{e}_3 x_2.$$

Let now Ω be an open subset of $\mathbb{R}^3$ with a piecewise smooth boundary. An $\mathbb{H}$-valued function is a mapping $\boldsymbol{f} : \Omega \longrightarrow \mathbb{H}$ such that $\boldsymbol{f}(\boldsymbol{x}) = \sum_{i=0}^{3} f^i(\boldsymbol{x})\,\mathbf{e}_i\,,\ \boldsymbol{x} \in \Omega$. The coordinates $f^i(\boldsymbol{x})$ are real-valued functions defined in Ω, i.e., $f^i(\boldsymbol{x}) : \Omega \longrightarrow \mathbb{R}\,, i = 0, 1, 2, 3$. Continuity, differentiability or integrability of $\boldsymbol{f}$ are defined coordinate-wise.

Furthermore, the *generalized Cauchy-Riemann operator* and the corresponding *adjoint generalized Cauchy-Riemann operator* are defined by

$$\overline{\boldsymbol{\partial}} := \frac{\partial}{\partial x_0} + 2\mathbf{e}_1\frac{\partial}{\partial\overline{\boldsymbol{\zeta}}} \quad \text{and} \quad \boldsymbol{\partial} := \frac{\partial}{\partial x_0} - 2\mathbf{e}_1\frac{\partial}{\partial\overline{\boldsymbol{\zeta}}} \quad \text{with} \quad \frac{\partial}{\partial\overline{\boldsymbol{\zeta}}} = \frac{1}{2}\left(\frac{\partial}{\partial x_1} - \mathbf{e}_3\frac{\partial}{\partial x_2}\right),$$

which leads to the following definitions:

 ISBN: 978-84-617-8694-7

Definition 1 *A function* $\boldsymbol{f} \in C^1(\Omega;\mathbb{H})$ *is called monogenic in* $\Omega \subset \mathbb{R}^3$ *if* $\overline{\boldsymbol{\partial}}\boldsymbol{f} = 0$ *in* Ω.

Definition 2 (Hypercomplex derivative [8]) *Let* $\boldsymbol{f} \in C^1(\Omega;\mathbb{H})$ *be a continuous, real-differentiable function and monogenic in* Ω. *The expression* $\boldsymbol{\partial_x f} := \frac{1}{2}\boldsymbol{\partial f}$ *is called hypercomplex derivative of* $\boldsymbol{f}$ *in* Ω.

3 Constructing monogenic functions with line singularities

Let us consider the *Legendre differential equation*

$$(1-t^2)\,\frac{\mathrm{d}^2 y}{\mathrm{d}t^2} - 2t\,\frac{\mathrm{d}y}{\mathrm{d}t} + n(n+1)y = 0, \quad n > 0,\ |t| < 1\,.$$

It is well known that the general solution of the second-order ordinary differential equation is given by

$$y = A\,P_n(t) + B\,Q_n(t),$$

where $P_n(t)$ and $Q_n(t)$ are the Legendre functions of the 1^{st} and 2^{nd} kind, respectively. In previous works [3, 4, 6] it was shown that the Legendre functions $P_n(t)$ as well as the associated Legendre functions $P_n^m(t)$ of the 1^{st} kind play a crucial role in defining orthogonal Appell basis of inner and outer spherical monogenics. Here, we will work with the Legendre functions $Q_n(t)$ of the 2^{nd} kind which can be defined, see for example [1], by the recurrence relation

$$(n+1)\,Q_{n+1}(t) = (2n+1)\,t\,Q_n(t) - n\,Q_{n-1}(t), \quad n \geq 1$$

with

$$Q_0(t) = \frac{1}{2}\ln\frac{1+t}{1-t} \qquad \text{and} \qquad Q_1(t) = t\,Q_0(t) - 1\,.$$

Due to the logarithmic term in $Q_0(t)$ the functions $Q_n(t)$ have infinite discontinuities at $t = \pm 1$. Therefore, we consider in the following the functions $Q_n(t)$ on the interval $|t| < 1$. Using the $Q_0(t)$, the *associated Legendre functions of the* 2^{nd} are defined by

$$Q_n^l(t) \;=\; (1-t^2)^{\frac{l}{2}}\,\frac{\mathrm{d}^l}{\mathrm{d}t^l}Q_n(t), \quad n, l \in \mathbb{N}_0,$$

which are solutions of the *associated Legendre differential equation* (see [1]). The main result of this section is stated in the following theorem.

Theorem 3 *For each* $n, l \in \mathbb{N}_0$, *the* $\mathbb{H}$*-valued functions*

$$\boldsymbol{W}_n^l(\boldsymbol{x}) = 2\,\boldsymbol{H}_n^l(\boldsymbol{x}) + \mathbf{e}_1\boldsymbol{H}_n^{l+1}(\boldsymbol{x})$$

composed of the harmonic functions

$$\boldsymbol{H}_n^l(\boldsymbol{x}) = \frac{2^{l-1}n!}{(n+l)!}\,|\boldsymbol{x}|^n\,Q_n^l(t)\left(\frac{\boldsymbol{\zeta}}{|\boldsymbol{\zeta}|}\right)^l \quad \textit{where} \quad t = \frac{x_0}{|\boldsymbol{x}|} \quad \textit{and} \quad \boldsymbol{\zeta} = x_1 - \mathbf{e}_3 x_2$$

are defined in $\Omega = \left\{\boldsymbol{x} \in \mathbb{R}^3 \;\middle|\; \boldsymbol{x} \neq x_0,\, x_0 \in \mathbb{R}\right\}$ *and have the following properties:*

 ISBN: 978-84-617-8694-7

(i) Monogenicity: $\overline{\boldsymbol{\partial}}\, \boldsymbol{W}_n^l(\boldsymbol{x}) = 0 \ \ \forall\, n, l \in \mathbb{N}_0$,

(ii) Generalized Appell property: $\boldsymbol{\partial_x}\, \boldsymbol{W}_{n+1}^l(\boldsymbol{x}) = (n+1)\, \boldsymbol{W}_n^l(\boldsymbol{x}) \ \ \forall\, n \in \mathbb{N}, l \in \mathbb{N}_0$,

(iii) Recurrence relation: For all $n \in \mathbb{N}, l \in \mathbb{N}_0$ *and* $n+1 \neq l$ *it holds*

$$\frac{2(n-l+1)(n+l+2)}{n+1}\, \boldsymbol{W}_{n+1}^l(\boldsymbol{x}) = \Big((2n+3)\boldsymbol{x}+(2n+1)\overline{\boldsymbol{x}}\Big)\boldsymbol{W}_n^l(\boldsymbol{x}) - 2n|\boldsymbol{x}|^2\, \boldsymbol{W}_{n-1}^l(\boldsymbol{x}).$$

Note, that the functions studied in [7] are obtained for the case $l = 0$.

References

[1] Andrews, L.C.: *Special Functions of Mathematics for Engineers*. SPIE Optical Engineering Press, Bellingham, Oxford University Press, Oxford, (1998).

[2] Bock, S. and Gürlebeck, K.: *On a generalized Appell system and monogenic power series*, Mathematical Methods in the Applied Sciences, **33**, 394–411, (2010).

[3] Bock, S.: *On a three dimensional analogue to the holomorphic z-powers: Power series and recurrence formulae*, Complex Var. Elliptic Eqns, Vol.57(12), 1349–1370, (2012).

[4] Bock, S.: *On a three dimensional analogue to the holomorphic z-powers: Laurent series expansions*, Complex Var. Elliptic Eqns, Vol.57(12), 1271–1287, (2012).

[5] Bock, S.: *On orthogonal series expansions in dimensions 2, 3 and 4*, Proceedings of the 9th International Conference on Clifford Algebras and their Applications in Mathematical Physics (ICCA9), Weimar, 2011.

[6] Bock, S.: *On Monogenic Series Expansions with Applications to Linear Elasticity*, Advances in Applied Clifford Algebras, Vol. 24 (4), 931-943, (2014).

[7] Bock, S.: *On a hypercomplex version of the Kelvin solution in linear elasticity*, to appear in the book Modern Problems in Applied Analysis, Proceedings of the 3rd Conference Boundary Value Problems, Functional Equations and Applications, Rzeszow, (2016).

[8] Gürlebeck, K. and Malonek, H.R.: *A hypercomplex derivative of monogenic functions in* $\mathbb{R}^{n+1}$ *and its applications*, Complex Variables **39**, 199–228, (1999).

[9] Weisz-Patrault, D., Bock, S. and Gürlebeck, K.:*Three-dimensional elasticity based on quaternion-valued potentials*, International Journal of Solids and Structures, **51**(19), 3422-3430 (2014).

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

Separate Node Ascending Derivatives Expansion (SNADE) on a Sequence of Nodes Alternating Between Two Values

Derya BODUR[1] and Metin DEMİRALP[1]

[1] *Informatics Institute, Computational Science and Engineering Program, İstanbul Technical University*

emails: deryabodur@itu.edu.tr, metin.demiralp@gmail.com

Abstract

This work focuses on a method called as Separate Node Ascending Derivatives Expansion (SNADE) which can be considered as a new Taylor Series Expansion. This method is distinguished from Taylor Series Expansion by its different node points using features. This paper investigates certain details of this issue. However a specific importance is given to the case where the SNADE nodal sequence is composed of elements alternating between two nodal values.

Key words: SNADE, Taylor Expansion, Integral Operator, Generating Functions
MSC 2000: 41A10, 41A58, 47G10

1 Introduction

There are various function representing methods whose structurings change depending on their number of independent variables. The multivariate function representation is out of the scope of this work. Hence we focus on only univariate function decomposition here. One very widely used method is to expand the considered function to the nonnegative powers of a first degree polynomial (a monomial minus a common constant which may be called the expansion point). This is the backbone of very well known Taylor series [1–10] and the obtained series converges in the complex-plane analyticity domain of the target function. As long as that domain is not empty the target function can be exactly represented by that power series even though its finite truncations are preferred to be used for practical purposes. The convergence is pointwise in analyticity domain, that is, the series uniformly converges at every interior point of the domain while the boundary points of the analyticity domain may give not uniform but conditional convergences.

Power series are not the only series representations. They are in fact very specific form of series representations as an infinite linear combination of a complete basis function set with constant coefficients such that the basis functions can be expressed as the natural number powers of the same first degree polynomial composed of a monomial of the independent variable minus a common constant. The orthogonal polynomial or function series are amongst these other much more general series. The convergence-in-the-mean is mostly encountered situation in these cases. Fourier series are perhaps the best examples to these groups. On the other hand, Taylor, Maclaurin and Laurent series are basically well-known power series even though the Laurent series contain polarly singular terms.

Taylor or Maclaurin series can be constructed by using an integral-of-derivative identity and includes a remainder term such that the convergence of the infinite series as the limit of the polynomial part depends on the vanishment of that remainder term when the polynomial part's degree grows unboundedly.

SNADE can be considered as a construction of a new Taylor series expansion. The method contains denumerable infinitely many nodes in contrast to Taylor series expansion. SNADE is based on the derivative integration formula[11–16]. The core topic of this work, "Separate Node Ascending Derivatives Expansion (SNADE)" is also based on the same identity. However, it uses infinitely many expansion points in each step of its repeated utilization.

This paper is organised as follows. The formulation about SNADE is given in the following section. In the third section, a special case of SNADE is considered such that there are two different nodal values which are used infinitely many times repeatedly and consecutively. Details will be given therein. A convergence investigating section follows this section. The convergence, in other words, the fact that the validity of infinite representation by this specific SNADE is proven in the following section while the last section finalizes the paper as usual.

2 Recalling SNADE

SNADE is based on the Integral-of-Derivative Identity which can be expressed as follows

$$f(x) = f(x_1) + \int_{x_1}^{x} d\xi f'(\xi), \quad x, x_1 \in [a, b] \tag{1}$$

The following equality can be obtained when the functions $f(x)$ and $f'(\xi)$ appearing in this identity are replaced with $f'(\xi)$ and $f''(\xi_1)$ respectively, the following equality can be obtained.

$$f'(\xi) = f'(x_2) + \int_{x_2}^{\xi} d\xi_1 f''(\xi_1) \tag{2}$$

 ISBN: 978-84-617-8694-7

If (1) and (2) are combined, the following equality can be achieved.

$$f(x) = f(x_1) + f'(x_2)(x - x_1) + \int_{x_1}^{x} d\xi_1 \int_{x_2}^{\xi_1} d\xi_2 f''(\xi_2) \tag{3}$$

Taylor series expansion has a single node and calculations are done at this point. But here function value at x_1 and function's derivative value at x_2 are used until now. If one more step is taken, the following equality can be obtained.

$$\begin{aligned} f(x) &= f(x_1) + f'(x_2)(x - x_1) + \frac{1}{2}(x - x_1)(x + x_1 - 2x_2) f''(x_3) \\ &+ \int_{x_1}^{x} d\xi_1 \int_{x_2}^{\xi_1} d\xi_2 \int_{x_3}^{\xi_2} d\xi_3 f'''(\xi_3) \end{aligned} \tag{4}$$

This structure can be written in symbolic form as follows

$$\begin{aligned} f(x) &= f(x_1)\mathcal{I}_0 1_f + f'(x_2)\mathcal{I}_1(x_1)1_f + f''(x_3)\mathcal{I}_2(x_1, x_2)1_f \\ &+ \mathcal{R}_2(x; x_1, x_2, x_3) \end{aligned} \tag{5}$$

where 1_f stands for the unit constant function and $\mathcal{I}$ is an integral operator which can be defined as

$$\begin{aligned} \mathcal{I}_m(x_1, \cdots, x_m)\, g(x) \equiv \int_{x_1}^{x} d\xi_1 \cdots \int_{x_m}^{\xi_{m-1}} d\xi_m g(\xi_m), \\ m = 1, 2, \cdots, \quad \mathcal{I}_0 g(x) \equiv g(x) \end{aligned} \tag{6}$$

where $g(x)$ is an arbitrary integrable function and x takes the role of nonexisting dummy variable ξ_0. The m-th order remainder term is obtained as follows

$$\mathcal{R}_m(x; x_1, \cdots, x_{m+1}) \equiv \mathcal{I}_{m+1}(x_1, \cdots, x_{m+1}) f^{(m+1)}(x), \qquad m = 0, 1, 2, \cdots \tag{7}$$

So, the formula given in (5) can be generalized as follows

$$f(x) = \sum_{i=0}^{m} f^{(i)}(x_{i+1})\mathcal{I}_i(x_1, \cdots, x_i)1_f + \mathcal{R}_{m+1}(x; x_1, \cdots, x_{m+1}), \qquad m = 0, 1, \cdots \tag{8}$$

Remainder term in the above formula tends to vanish when m grows unboundedly. So, the following equation can be written

$$f(x) = \sum_{i=0}^{\infty} f^{(i)}(x_{i+1})\mathcal{I}_i(x_1, \cdots, x_i)1_f \tag{9}$$

and is named as "Infinite Order SNADE".

 ISBN: 978-84-617-8694-7

3 SNADE on a Sequence of Nodes Alternating Between Two Values

We start this section with recalling SNADE formula as follows

$$\begin{aligned} f(x) &= \sum_{j=0}^{\infty} f^{(j)}(x_j) \int_{x_1}^{x} d\xi_1 \int_{x_2}^{\xi_1} d\xi_2 \cdots \int_{x_j}^{\xi_{j-1}} d\xi_j 1_f \\ &= \sum_{j=0}^{\infty} f^{(j)}(x_j) \widehat{\mathcal{I}}_j(x; x_1, ..., x_j) 1_f, \qquad \widehat{\mathcal{I}}_0 \equiv 1 \end{aligned} \tag{10}$$

The title of this section may imply that there are only two nodes in the approach here although the truth is different. There are in fact a denumerable infinite number of nodes here as expected. However, the values of nodes are alternated between two nodal values represented by x_1 and x_2. This means that the nodes are given through the sequence x_1, x_2, x_1,... In other words there is a pattern defined by $x_{2j-1} \equiv x_1, x_{2j} \equiv x_2$ which leads us to the definition of two-fold integral operator as follows

$$\widehat{\mathcal{J}}(x_1, x_2) g(x) \equiv \int_{x_1}^{x} d\xi_1 \int_{x_2}^{\xi_1} d\xi_2 g(\xi_2) \tag{11}$$

which urges us to define the following polynomials

$$P_{2k}(x; x_1, x_2) \equiv \widehat{\mathcal{J}}(x_1, x_2)^k 1_f, \qquad k = 0, 1, 2, ... \tag{12}$$

$$P_{2k+1}(x; x_1, x_2) \equiv \int_{x_1}^{x} d\xi P_{2k}(\xi; x_2, x_1), \qquad k = 0, 1, 2, ... \tag{13}$$

It is possible to show that the following recursions between these polynomials hold.

$$P_{2k}(x; x_1, x_2) = \widehat{\mathcal{J}}(x_1, x_2) P_{2k-2}(x; x_1, x_2) \quad k = 1, 2, 3, ..., \quad P_0 \equiv 1 \tag{14}$$

$$v \equiv x_1 - x_2, \qquad u \equiv (x - x_2)/v, \tag{15}$$

The impressions from these polynomial expressions bring us to the following structure

$$P_{2k}(x; x_1, x_2) = \frac{v^{2k}}{(2k)!} \sum_{j=0}^{k} p_{k,j} u^{2j} \qquad k = 1, 2, 3, ..., \qquad p_{0,0} \equiv 1 \tag{16}$$

To determine $p_{k,j}$ coefficients, the action of $\widehat{J}$ on u^k is studied and the following equality is obtained

$$\widehat{J}(x_1, x_2) u^k = \frac{v^2}{(k+1)(k+2)} \left(u^{k+2} - 1 \right), \qquad k = 0, 1, 2, ... \tag{17}$$

 ISBN: 978-84-617-8694-7

By using the action of $\widehat{J}$ on u^k and the recursion between P_{2k} and P_{2k-2}, the following equality can be obtained without explicitly showing the intermediate operations.

$$P_{2k}\left(x;x_1,x_2\right)=\frac{v^{2k}}{(2k)!}\sum_{j=1}^{k}\frac{(2k-1)2k}{(2j-1)2j}p_{k-1,j-1}\left(u^{2j}-1\right),\qquad k=1,2,3,... \tag{18}$$

When the right hand sides of the polynomial expressions in (18) and (16) are set equal to each other, the following coefficients of the polynomials can be obtained.

$$\begin{aligned} p_{k,j} &= \frac{2k(2k-1)}{2j(2j-1)}p_{k-1,j-1},\quad k=1,2,3,...;\quad j=1,2,...,k; \\ p_{k,0} &= -\sum_{j=1}^{k}\frac{2k(2k-1)}{2j(2j-1)}p_{k-1,j-1}=-\sum_{j=1}^{k}p_{k,j},\qquad k=1,2,3,... \end{aligned} \tag{19}$$

As can be noticed immediately in these polynomials, the sum of the coefficients of each polynomial except the constant one is zero. On the other hand, there is a recursion between the coefficients of two consecutive terms. The determination of $p_{k,k}$ is given as follows.

$$\begin{aligned} p_{k,k} &= p_{k-1,k-1},\qquad k=1,2,3,...; \\ p_{0,0} &= 1,\quad p_{k,k}=1,\qquad k=0,1,2,... \end{aligned} \tag{20}$$

$p_{k,k-j}$ coefficients are determined by consecutive examinations as we have done above. This gives

$$p_{k,k-j}=\binom{2k}{2j}p_{j,0},\qquad j=0,1,2,...\qquad k=j,j+1,j+2,... \tag{21}$$

4 Constitution of Generating Function

Now we focus on (19) to determine $p_{j,0}$ coefficients. For this purpose, $p_{k,k-j}$ expressions given by (21) can be used in (19) to get the following equation.

$$\sum_{j=0}^{k}\binom{2k}{2j}p_{j,0}=\delta_{k,0} \tag{22}$$

If both sides of this equality is multiplied by $z^{2k}/(2k)!$ and the resulting equality's both sides are separately sum over k from 0 to infinity then the right hand side becomes just 1 while the left hand side becomes the result of a Cauchy product between two power series. This gives

$$\sum_{k=0}^{\infty}\frac{z^{2k}}{(2k)!}\sum_{j=0}^{\infty}\frac{z^{2j}}{(2j)!}p_{j,0}=1 \tag{23}$$

 ISBN: 978-84-617-8694-7

which enables us to get

$$\sum_{j=0}^{\infty} \frac{z^{2j}}{(2j)!} p_{j,0} = \frac{1}{\cosh z} \tag{24}$$

This allows us to determine $p_{j,0}$ as follows

$$p_{j,0} = \left\{ \frac{d^{2j}}{dz^{2j}} \left(\frac{1}{\cosh z} \right) \right\}_{z=0}, \qquad j = 0, 1, 2, ... \tag{25}$$

Now we can focus on the following generating function to determine the polynomials P_{2k}. We can write

$$\begin{aligned} G(z;u,v) &= \sum_{k=0}^{\infty} z^{2k} P_{2k}(x;x_1,x_2) = \sum_{k=0}^{\infty} \sum_{j=0}^{k} \frac{z^{2k} v^{2k} p_{k-j,0}}{(2j)!(2k-2j)!} u^{2j} \\ &= \sum_{j=0}^{\infty} \frac{(vuz)^{2j}}{(2j)!} \sum_{k=0}^{\infty} \frac{(vz)^{2k}}{(2k)!} p_{k,0} = \frac{\cosh vuz}{\cosh vz} \end{aligned} \tag{26}$$

$$P_{2k}(x;x_1,x_2) = \frac{1}{(2k)!} \left\{ \frac{d^{2k}}{dz^{2k}} \left(\frac{\cosh vuz}{\cosh vz} \right) \right\}_{z=0} \tag{27}$$

The above generating function depends on real valued v and uv, and complex valued z variables. Numerator and denumerator of this function that consist of known functions are entire functions. There are no singularities. But, if vz takes one of the values $\pm i(n+1/2)\pi$, the denumerator will be set equal to zero. So, it is faced with singularity in that case. For this reason, the Maclaurin series expansion of this generating function converges when the complex norm of vz is less than $\pi/2$, otherwise it diverges. Apart from that case, we have also been worked with divergent generating functions. Convergent integral representations were obtained by using Borel summation of these generating functions. On the other hand, we have also focused on circular sector contour integral to overcome this difficulty even though we do not intend to give the details here.

5 Validity of SNADE on a Sequence Alternating Between Two Values

In this section, we will prove that the SNADE on a sequence whose elements are alternating between two nodal values really represents the target function, f. For this purpose, the infinite sum is splitted into two parts as follows.

$$F_1(x;x_1,x_2) = \sum_{j=0}^{\infty} f^{(2j)}(x_1) \widehat{\mathcal{I}}_{2j}(x;x_1,x_2) 1_f, \qquad \widehat{\mathcal{I}}_0 \equiv 1,$$

 ISBN: 978-84-617-8694-7

$$F_2\left(x;x_1,x_2\right) = \sum_{j=0}^{\infty} f^{(2j+1)}\left(x_2\right)\widehat{\mathcal{I}}_{2j+1}\left(x;x_1,x_2\right)1_f, \tag{28}$$

These expressions can be transformed into other expressions composed of polynomials we have investigated above as follows.

$$\begin{aligned} F_1\left(x;x_1,x_2\right) &= \sum_{j=0}^{\infty} f^{(2j)}\left(x_1\right)P_{2j}\left(x;x_1,x_2\right), \\ F_2\left(x;x_1,x_2\right) &= \sum_{j=0}^{\infty} f^{(2j+1)}\left(x_2\right)P_{2j+1}\left(x;x_1,x_2\right) \end{aligned} \tag{29}$$

Now question "Does the following equality hold for any target function $f(x)$?" is the goal to be answered positively.

$$F_1\left(x;x_1,x_2\right) + F_2\left(x;x_1,x_2\right) = f(x) \tag{30}$$

To get answer, first, u and v are replaced with Υ and $-v$ respectively in (26). Then both sides of the produced equation are integrated over the interval $[\,0,1-u\,]$ with respect to Υ. At the last step, both sides of the resulting equation are multiplied by $-vz$ and the resulting equation can be written in the following simple form

$$P_{2k+1}\left(u,v\right) = -\frac{1}{(2k+1)!}\left\{\frac{d^{2k+1}}{dz^{2k+1}}\left(\frac{\sinh v(1-u)z}{\cosh vz}\right)\right\}_{z=0} \tag{31}$$

Now we focus on the determination of the function $F_1\left(x;x_1,x_2\right)$

$$F_1\left(x;x_1,x_2\right) = \sum_{k=0}^{\infty} f^{(2k)}\left(x_1\right)P_{2k}\left(u,v\right) \tag{32}$$

If the $(2k)$th derivative of the product between $1/\cosh(vz)$ and $\cosh(vuz)$ appearing at the right side of (27) where P_{2k} is given explicitly, is rewritten according to the Leibniz rule and vanished terms are not shown then the following equations can be written.

$$P_{2k}(u,v) = \sum_{j=0}^{k} \frac{(vu)^{2k-2j}}{(2j)!(2k-2j)!}\left\{\frac{d^{2j}}{dz^{2j}}\left(\frac{1}{\cosh(vz)}\right)\right\}_{z=0}, \qquad k=0,1,2,... \tag{33}$$

If this result is used in (32) then the following structure is obtained.

$$F_1\left(x;x_1,x_2\right) = \sum_{j=0}^{\infty} \frac{1}{(2j)!}\left\{\frac{d^{2j}}{dz^{2j}}\left(\frac{1}{\cosh(vz)}\right)\right\}_{z=0} \sum_{k=0}^{\infty} f^{(2k+2j)}\left(x_1\right)\frac{(vu)^{2k}}{(2k)!} \tag{34}$$

 ISBN: 978-84-617-8694-7

whose rightmost infinite sum can be simplifed in accordance with the following relation which can be established by using even function representation

$$\begin{aligned}\sum_{k=0}^{\infty} f^{(2k+2j)}(x_1)\frac{(vu)^{2k}}{(2k)!} &= \sum_{k=0}^{\infty} f^{(2k+2j)}(x_1)\frac{(x-x_2)^{2k}}{(2k)!} \\ &= \frac{1}{2}f^{(2j)}(x+x_1-x_2)+\frac{1}{2}f^{(2j)}(x_1+x_2-x) \end{aligned} \tag{35}$$

Using (35) in (34) allows us to write the following equality

$$\begin{aligned}F_1(x;x_1,x_2) &= \frac{1}{2}\sum_{j=0}^{\infty}\frac{1}{(2j)!}\left\{\frac{d^{2j}}{dz^{2j}}\left(\frac{1}{\cosh(vz)}\right)\right\}_{z=0} \\ &\times\left[f^{(2j)}(x+x_1-x_2)+f^{(2j)}(x_1+x_2-x)\right]\end{aligned} \tag{36}$$

Now it is time to focus on the function $F_2(x;x_1,x_2)$ whose explicit structure is given below.

$$F_2(x;x_1,x_2) = \sum_{k=0}^{\infty} f^{(2k+1)}(x_2)\,P_{2k+1}(u,v) \tag{37}$$

Similar steps we have taken above for the evaluation of F_1 can be traced for this equality and the following result is achieved without explicitly giving the intermediate operations.

$$\begin{aligned}F_2(x;x_1,x_2) &= \frac{1}{2}\sum_{j=0}^{\infty}\frac{1}{(2j)!}\left\{\frac{d^{2j}}{dz^{2j}}\left(\frac{1}{\cosh(vz)}\right)\right\}_{z=0} \\ &\times\left[f^{(2j)}(x-v)-f^{(2j)}(x_1+x_2-x)\right]\end{aligned} \tag{38}$$

(36) and (38) allow us to write the following equality

$$\begin{aligned}F_1(x)+F_2(x) &= \sum_{j=0}^{\infty}\frac{1}{(2j)!}\left\{\frac{d^{2j}}{dz^{2j}}\left(\frac{1}{\cosh(vz)}\right)\right\}_{z=0} \\ &\times\frac{1}{2}\left[f^{(2j)}(x+x_1-x_2)+f^{(2j)}(x-x_1+x_2)\right]\end{aligned} \tag{39}$$

To proceed, the expression between the brackets can be expanded to series in terms of natural number powers of (x_1-x_2) as follows

$$\begin{aligned}F_1(x;x_1,x_2)+F_2(x;x_1,x_2) &= \sum_{k=0}^{\infty}\frac{f^{(2k)}(x)}{(2k)!}v^{2k}\sum_{j}^{k}\binom{2k}{2j}\left\{\frac{d^{2j}}{dz^{2j}}\left(\frac{1}{\cosh(z)}\right)\right\}_{z=0} \\ &\times\left\{\frac{d^{2k-2j}}{dz^{2k-2j}}(\cosh(z))\right\}_{z=0}\end{aligned} \tag{40}$$

 ISBN: 978-84-617-8694-7

We can write the following equalities for the finite sum in (40)

$$\sum_{j}^{k} \binom{2k}{2j} \left\{ \frac{d^{2j}}{dz^{2j}} \left(\frac{1}{\cosh(z)} \right) \right\}_{z=0} \left\{ \frac{d^{2k-2j}}{dz^{2k-2j}} \left(\cosh(z) \right) \right\}_{z=0} =$$
$$= \left\{ \frac{d^{2k}}{dz^{2k}} \left(\frac{\cosh(z)}{\cosh z} \right) \right\}_{z=0} = \left\{ \frac{d^{2k}}{dz^{2k}} \left(1 \right) \right\}_{z=0} = \delta_{k,0} \tag{41}$$

which takes us to the following conclusion we want to reach in fact.

$$F_1 \left(x; x_1, x_2 \right) + F_2 \left(x; x_1, x_2 \right) = f(x) \tag{42}$$

6 Concluding Remarks

We have constructed Separate Node Ascending Derivatives Expansion (SNADE) on a nodal sequence with elements alternating on two different values in this paper. We enumerate important points below.

1. SNADE is based on the famous integral-of-derivative identity which can be used also in the Taylor series in format with remainder;

2. In contrast to Taylor expansion SNADE can use a denumerable infinite number of values as the nodes;

3. We have focused only on the sequences whose elements alternate between two nodal values here;

4. We could have been able to evaluate all terms of this very specific case;

5. The polynomials we have used in the construction are closely related to so-called "Swiss Knife" polynomials;

6. Despite very special specification we have obtained a lot of key points to focus on the cases of much more complicated sequences. Thus we can proceed to develop SNADE to make it a powerful theory;

7. The form of the specific SNADE we have focused here brings a lot of constraints on the nodal values and the independent variables. However it seems to generalize our findings to more complicated domains.

 ISBN: 978-84-617-8694-7

References

[1] G. Thomas and L.R. Finney, *Calculus and Analytic Geometry (9th ed.)*,Addison Wesley, ISBN 0-201-53174-7, 1996.

[2] M. Demiralp, *Convergence issues in the gaussian weighted multidimensional fluctuation expansion for the univariate numerical integration*, WSEAS Transactions on Mathematics, **4** (2005) 486-492.

[3] C. Gözükirmizi and M. Demiralp, *The application of the fluctuation expansion with extended basis set to numerical integration*, WSEAS Transactions on Mathematics, **8**(5) (2009) 205-212.

[4] N.A. Baykara, E. Gürvit and M. Demiralp, *The fluctuationlessness approach to the numerical integration of functions with a single variable by integrating taylor expansion with explicit remainder term*, J. Math. Chem., **49**(2) (2011) 393-406.

[5] N.A. Baykara, E. Gürvit and M. Demiralp, *Fluctuationless univariate integration through taylor expansion with remainder by using oscillatory function basis sets*, In AIP Proceedings for the International Conference of Numerical Analysis and Applied Mathematics (ICNAAM 2009), Rethymno, Crete, Greece, **1168** (2009) 428-431.

[6] E. Gürvit, N.A. Baykara and M. Demiralp, *Taylor series expansion with the fluctuation freely approximated remainder over highly oscillatory basis functions*, In AIP Proceedings for the International Conference of Numerical Analysis and Applied Mathematics (ICNAAM 2009), Rethymno, Crete, Greece, **1168** (2009) 432-435.

[7] S.Tuna, N.A. Baykara, and M. Demiralp, *Taylor series based integration with the fluctuation freely approximated remainder over gauss wave type basis functions*, In AIP Proceedings for the International Conference of Computational Methods in Science and Engineering (ICCMSE 2009), Mini Symposium on Recent Developments in Numerical Schemes for Hilbert Space Related Issues in Science and Engineering, page in print, Rhodes, Greece, 29 September-4 October 2009. Chaired by Metin Demiralp.

[8] N.A. Baykara, E. Gürvit and M. Demiralp, *Extended fluctuationlessness theorem and its application to numerical approximation via taylor series*, In Proceedings for the 1st IEEEAM Conference on Applied Computer Science (ACS), Malta, (2010) 317-323.

[9] M. Demiralp, *A new fluctuation expansion based method for the univariate numerical integration under gaussian weights*, In Proceedings of the 8th WSEAS International Conference on Applied Mathematics (MATH08), ISBN:960-8457-39-4, Tenerife, Canary Islands, Spain, (2005) 68-73.

[10] E. Gürvit, N.A. Baykara and M. Demiralp, *Numerical integration of bivariate functions over a non rectangular area by using fluctuationlessness theorem*, In Proceedings for the WSEAS Conference on the 2nd Multivariate Analysis and its Application in Science and Engineering (MAASE09), ISBN: 978-960-474-083-3, Istanbul, Turkey, **8**(5) 2009 81-86.

[11] M. Demiralp, *Separate Node Ascending Derivatives Expansion (SNADE) for Univariate Functions: Conceptuality and Formulation*, AIP Proceedings for the International Conference of Numerical Analysis and Applied Mathematics (ICNAAM 2014), 22-28 September 2014, Rhodes, Greece, **1648**(1) (2015) 160004.

[12] N.A. Baykara and M. Demiralp, *Separate Node Ascending Derivatives Expansion (SNADE) for Univariate Functions: Polynomial Recursions, Remainder Bounds and the Convergence*, AIP Proceedings for the International Conference of Numerical Analysis and Applied Mathematics (ICNAAM 2014), 22-28 September 2014, Rhodes, Greece, **1648**(1) (2015) 160005.

[13] N.A. Baykara and E. Gürvit, *Separate Node Ascending Derivatives Expansion (SNADE) for Univariate Functions: Univariate Numerical Integration*, AIP Proceedings for the International Conference of Numerical Analysis and Applied Mathematics (ICNAAM 2014), 22-28 September 2014, Rhodes, Greece, **1648**(1) (2015) 160006.

[14] E. Gürvit and N.A. Baykara, *Separate Node Ascending Derivatives Expansion (SNADE) for Univariate Functions: Node Optimization via Partial Fluctuation Suppression*, AIP Proceedings for the International Conference of Numerical Analysis and Applied Mathematics (ICNAAM 2014), 22-28 September 2014, Rhodes, Greece, **1648**(1) (2015) 160006.

[15] B. Tunga, *Separate multinode ascending derivatives expansion (Demiralps SMADE): Basis polynomial*, 11th International Conference of Computational Methods in Sciences and Engineering, ICCMSE 2015, 20-23 March 2015, Athens, Greece, **1702**(1) (2015) 170011.

[16] D. Bodur, and M. Demiralp, *Two Very Specific Cases for Separate Node Ascending Derivatives Expansion (SNADE)*, 11th International Conference of Computational Methods in Sciences and Engineering, ICCMSE 2015, 20-23 March 2015, Athens, Greece, **1702**(1) (2015) 170006.

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

Efficient Solution of Shifted Quasiseparable Systems and Applications

Paola Boito[1,2], Yuli Eidelman[3] and Luca Gemignani[4]

[1] *XLIM-MATHIS, Université de Limoges*

[2] *Laboratoire pour l'Informatique et le Parallélisme, ENS Lyon*

[3] *School of Mathematical Sciences, Raymond and Beverly Sackler Faculty of Exact Sciences, Tel-Aviv University*

[4] *Dipartimento di Informatica, Università di Pisa*

emails: paola.boito@unilim.fr, eideyu@post.tau.ac.il, l.gemignani@di.unipi.it

Abstract

We propose an efficient algorithm for the solution of shifted quasiseparable systems, which exploits the invariance of the quasiseparable structure under diagonal shifting and inversion. This algorithm is applied to compute various functions of matrices, and to solve a class of linear matrix equations. Numerical experiments show the effectiveness of our approach.

Key words: quasiseparable matrices; shifted linear system; QR factorization; matrix function; matrix equation.
MSC 2000: 65F05

1 Introduction

In this work we propose a novel method for computing the solution of shifted quasiseparable systems of the form

$$(A + \sigma_i I_N)\boldsymbol{x}_i = \boldsymbol{y}, \quad i = 1, \ldots, \ell, \tag{1}$$

and of more general parameter dependent linear matrix equations with quasiseparable representations. Our approach also has a noticeable potential for effectively solving some large-scale algebraic problems that reduce to evaluating the action of a quasiseparable matrix function to a vector.

Quasiseparable matrices find their application in several branches of applied mathematics and engineering. For instance, quasiseparable structure often arises in the discretization of continuous operators, due to the local properties of the discretization schemes and/or to the decay properties of the operator or of its finite approximations. As a consequence, in the last decade there has been considerable interest in the development of fast algorithms for working with quasiseparable matrices [3, 4, 9, 10].

2 Main Algorithm

It is well known that several operations with quasiseparable matrices can be performed in linear time with respect to matrix size. In particular, the QR factorization algorithm presented in [2] computes in linear time a QR decomposition of a quasiseparable matrix $A \in \mathbb{C}^{N\times N}$ of the form $A = V \cdot U \cdot R$, where R is upper triangular, whereas U and V are banded unitary matrices and – this is the crucial point – V only depends on the generators of the strictly lower triangular part of A. This implies that any shifted matrix $A + \sigma I_N$, $\sigma \in \mathbb{C}$, can also be factored as $A + \sigma I_N = V \cdot U_\sigma \cdot R_\sigma$ for suitable U_σ and R_σ.

Relying upon this fact, we design and implement an efficient algorithm [1] for solving a sequence of shifted quasiseparable linear systems. Its arithmetic complexity is linear with respect to N and to the number of shifts, and it is halved with respect to straightforward application of structured QR factorization to each shifted linear system. The Matlab code is available at `http://www.unilim.fr/pages_perso/paola.boito/software.html`.

Some applications are presented in the following sections; they include the computation of $f(A)\boldsymbol{v}$ via series expansion or contour integrals, and the solution of linear matrix equations.

3 A Model Problem for Boundary Value ODEs

Consider the non-local boundary value problem:

$$\frac{d\boldsymbol{v}}{dt} = A\boldsymbol{v}, \quad 0 < t < \tau,$$

$$\frac{1}{\tau}\int_0^\tau \boldsymbol{v}(t)\, dt = \boldsymbol{g},$$

where A is a linear operator in $\mathbb{R}^N$ and $\boldsymbol{g} \in \mathbb{R}^N$ is a given vector [11, 12, 5].

If all the numbers $\mu_k = 2\pi ik/\tau,\ k = \pm 1, \pm 2, \pm 3, \ldots$ are regular points of the operator A, the problem has a unique solution, given by

$$\boldsymbol{v}(t) = q_t(A)\boldsymbol{g}, \quad q_t(z) = \frac{\tau z e^{zt}}{e^{z\tau} - 1}.$$

 ISBN: 978-84-617-8694-7

Without loss of generality one can assume $\tau = 2\pi$.

We show that, for a suitable truncation index ℓ, $\boldsymbol{v}(t)$ can be approximated by the finite sum

$$\boldsymbol{v}_\ell(t) = \sum_{j=0}^{3} V_j(t) A^j g - 2 \sum_{k=1}^{\ell} \frac{1}{k^2} (A \cos kt + \frac{1}{k} A^2 \sin kt)(A^2 + k^2 I_N)^{-1} A^3 g,$$

with $0 \leq t \leq 2\pi$ and

$$V_0(t) = 1,\ V_1(t) = t - \pi,\ V_2(t) = \frac{\pi^2}{3} - \pi t + \frac{t^2}{2},\ V_3(t) = \frac{\pi^2}{3} t - \frac{\pi}{2} t^2 + \frac{t^3}{6}.$$

The computation of $\boldsymbol{v}(t_i)$, $0 \leq i \leq M+1$, requires the solution of a possibly large set of shifted systems and our algorithm proves to be effective for this task if A is quasiseparable.

More generally, a similar approach can be applied to the computation of a function of a quasiseparable matrix, multiplied by a vector, whenever the function can be represented as a series of partial fractions. The classes of meromorphic functions admitting such a representation were investigated for instance in [8]. Other partial fraction approximations of certain analytic functions can be found in [7].

4 Sylvester-type Matrix Equations

As a natural extension of the problem (1), the right-hand side $\boldsymbol{y}$ could also depend on the parameter, so that we have a different right-hand side for each linear system. This situation is common in many applications, such as control theory, structural dynamics and time-dependent PDEs [6]. The systems to be solved take the form of a linear matrix equation:

$$AX + XD = Y, \quad A \in \mathbb{R}^{N \times N},\ D = \operatorname{diag}[\sigma_1, \ldots, \sigma_\ell],\ Y = [\boldsymbol{y}_1, \ldots, \boldsymbol{y}_\ell].$$

The extension to the case where D is lower triangular can be carried out via backsubstitution, and a further generalization to the case where D is a general matrix is possible using the classical Bartels-Stewart approach based on Schur decomposition. This approach is especially interesting when ℓ is significantly smaller than N. If A is quasiseparable we can apply our structured approach as outlined above, with computational advantages with respect to the widespread method that relies on Kronecker products.

References

[1] P. Boito, Y. Eidelman, and L. Gemignani, *Efficient solution of parameter dependent quasiseparable systems and computation of meromorphic matrix functions*, arXiv 1611.09107 [math.NA], 2016.

[2] Y. Eidelman and I. Gohberg, *A modification of the Dewilde-van der Veen method for inversion of finite structured matrices*, Linear Algebra Appl. **343/344** (2002) 419–450.

[3] Y. Eidelman, I. Gohberg and I. Haimovici, *Separable type representations of matrices and fast algorithms. Vol. 1*, Operator Theory: Advances and Applications, 234. Birkhäuser/Springer, Basel, 2014.

[4] Y. Eidelman, I. Gohberg and I. Haimovici, *Separable type representations of matrices and fast algorithms. Vol. 2*, Operator Theory: Advances and Applications, 235. Birkhäuser/Springer, Basel, 2014.

[5] Y. S. Eidelman, V. B. Sherstyukov and I. V. Tikhonov, *The resolving formulas in the model nonlocal problem for the evolution equation*, In press.

[6] G. D. Gu and V. Simoncini, *Numerical solution of parameter-dependent linear systems*, Numer. Linear Algebra Appl. **12** (2005) 923–940.

[7] N. Hale, N. J. Higham and L. N. Trefethen, *Computing* $\mathbf{A}^{\alpha}$, $\log(\mathbf{A})$*, and related matrix functions by contour integrals*, SIAM J. Numer. Anal. **46** (2008) 2505–2523.

[8] V. B. Sherstyukov, *Expansion of the reciprocal of an entire function with zeros in a strip in the Kreĭn series*, Mat. Sb. **202** (2011) 137–156.

[9] R. Vandebril, M. Van Barel and N. Mastronardi, *Matrix computations and semiseparable matrices. Vol. 1*, Johns Hopkins University Press, Baltimore, 2008.

[10] R. Vandebril, M. Van Barel and N. Mastronardi, *Matrix computations and semiseparable matrices. Vol. 2*, Johns Hopkins University Press, Baltimore, 2008.

[11] I. V. Tikhonov, *On the solvability of a problem with a nonlocal integral condition for a differential equation in a Banach space*, Differential Equations **34** (1998) 841–844.

[12] I. V. Tikhonov, *Uniqueness theorems in linear nonlocal problems for abstract differential equations*, Izv. Math. **67** (2003) 333–363.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

Improved parallel simulations for fractional-order systems using HPC

Cosmin Bonchiş[1,2], Eva Kaslik[1,2] and Florin Roşu[1,2]

[1] *Institute e-Austria Timişoara, Romania*

[2] *Dept. of Mathematics and Computer Science, West University of Timişoara, Romania*

emails: cosmin.bonchis@e-uvt.ro, ekaslik@gmail.com, florin.rosu@e-uvt.ro

Abstract

A parallel numerical simulation algorithm is presented for fractional-order systems involving Caputo derivatives, based on the Adams-Bashforth-Moulton predictor-corrector scheme. The parallel algorithm is implemented using MPI that runs on an HPC cluster, and the results are compared to others recently reported in the literature. As an applied experiment, we numerically compute the solutions of a fractional-order version of a fractional-order system describing a forced series LCR circuit, depicting cascades of period-doubling bifurcations leading to the onset of chaotic behavior.

Key words: Fractional-order system, parallel numerical algorithm, HPC processing.

1 Introduction

Compared to their integer-order counterparts, over the past decades, fractional-order dynamical systems have proved to provide more accurate and realistic results in the modeling of real world processes arising from diverse applied fields [3, 6, 8, 9, 10].

Although many qualitative properties of fractional-order systems can be studied by analytical tools (such as local stability of equilibrium states), theoretical characterization of chaos in fractional-order dynamical systems is yet to be investigated. In order to assess chaotic behavior of fractional-order dynamical systems, accurate estimation of the solutions over large time intervals is of utmost importance. However, an essential observation is that the employed discretization should use a small step size, with the aim of providing an accurate estimation to the solution of the fractional-order system under investigation.

 ISBN: 978-84-617-8694-7

Several numerical methods are used for fractional-order systems, such as a generalization of the Adams-Bashforth-Moulton predictor-corrector method [5] or a class of p-fractional linear multistep methods [7]. The main drawback of these numerical schemes is that, in order to obtain a reliable estimation of the solution, at every iteration step, all previous iterations have to be taken into account, due to the hereditary nature of the problem. Therefore, this implies extreme computational costs whenever the solution is computed over a large time interval, with a small step size. These difficulties may be overcome using parallel computing algorithms implemented in a conventional way or using available high performance computing systems [1, 2].

In this paper, we will present an efficient parallel algorithm implemented using Message Passing Interface (MPI) and running on a high performance computing system BlueGene/P cluster that has 1024 processors and 4TB of RAM memory. The numerical method considered here for implementing the fractional-order system is the Adams-Bashforth-Moulton predictor-corrector scheme [5]. The main challenge for implementing this method is to parallelize the computation of the solution because, the computation of an iteration step requires to take into account all previous iterations.

2 Preliminaries

Consider an ordinary fractional differential equation of the form:

$$\begin{cases} \mathrm{D}_*^{\alpha} y(t) = f(t, y(t)), & t \in [0, T] \\ y^{(k)}(0) = y_0^k, & k \in \{0, \ldots, \lceil \alpha \rceil - 1\}, \end{cases} \tag{1}$$

where $\alpha > 0$ and $\lceil \cdot \rceil$ denotes the ceiling function that rounds up to the nearest integer. The fractional derivative of Caputo-type is defined as:

$$\mathrm{D}_*^{\alpha} y(t) = \frac{1}{\Gamma(\lceil \alpha \rceil - \alpha)} \int_0^T \frac{y^{(\lceil \alpha \rceil)}(t)}{(t-\tau)^{\alpha - \lceil \alpha \rceil + 1}} \, \mathrm{d}\tau.$$

The numerical method used to solve (1) is a fractional version of the Adams-Bashforth-Moulton predictor corrector scheme [5]. The domain $[0, T]$ is discretized into N intervals with a step size $h = \frac{T}{N}$ and the grid points $t_n = nh$, for $n \in \{0, \ldots, N\}$. We will also denote $y_n = y(t_n)$ and $f_n = f(t_n, y_n)$ with $y_0 = y_0^0$ as the initial condition.

The first step of the scheme is the **predictor**, which will give a first approximation y_{n+1}^P of our solution:

$$y_{n+1}^P = \sum_{k=0}^{\lceil \alpha \rceil - 1} \frac{t_{n+1}^k}{k!} y_0^{(k)} + h^{\alpha} \sum_{k=0}^{n} b_{n-k} f_k, \quad \text{where } b_n = \frac{(n+1)^{\alpha} + n^{\alpha}}{\Gamma(\alpha + 1)}. \tag{2}$$

 ISBN: 978-84-617-8694-7

The final approximation of the solution, called the **corrector**, is given by:

$$y_{n+1} = \sum_{k=0}^{\lceil\alpha\rceil-1} \frac{t_{n+1}^k}{k!} y_0^{(k)} + h^\alpha \left(c_n f_0 + \sum_{k=1}^{n} a_{n-k} f_k + \frac{f(t_{n+1}, y_{n+1}^P)}{\Gamma(\alpha+2)} \right),$$

where the weights a_n and c_n are defined as:

$$a_n = \frac{(n+2)^{\alpha+1} - 2(n+1)^{\alpha+1} + n^{\alpha+1}}{\Gamma(\alpha+2)} \quad \text{and} \quad c_n = \frac{n^{\alpha+1} - (n-\alpha)(n+1)^\alpha}{\Gamma(\alpha+2)}.$$

This numerical scheme can be generalized in a straight-forward way, when one has to deal with a system of fractional-order differential equations.

The main computational difficulty of this scheme arises from the fact that at each step, we require the complete history of the variable, i.e., when computing y_{n+1}, we need to know all previous values y_k that are used to compute f_k, for $k \leq n$. This makes numerical methods addressed at solving fractional differential equations (or systems) notoriously hard to parallelize.

3 Parallel numerical algorithm

The parallel implementation of Adams-Bashforth-Moulton algorithm as a method for solving an fractional-order system was first presented by Diethelm [4]. The solution presented there is not suitable to be running on a HPC cluster because there is an unbalanced workload and waiting (idle) times for the processes that can cause the performance to be very low, due to the HPC parallel implementation rules [12]. Also, the amount of messages passed between processes does not respect the HPC parallel implementation idea [12], and strongly influence the overall performance.

In the previous work [1], we explored how numerical calculations of Adams-Bashforth-Moulton method for fractional-order systems can be accelerated by using parallel computing techniques. We investigated the feasibility of parallel computing algorithms and their efficiency in reducing the computational costs over a large time interval. The results concerning the parallel implementation for the Adams-Bashforth-Moulton method on HPC and CUDA show that some execution times are quite high and thus, they limit the time frame needed for more accurate simulations.

In [1], the HPC implementation was made using a classical approach: a process is the master and all other are slaves to help compute the values for predictor and corrector. The classical execution flow is presented in Figure 1a and shows that the master process is working either by computing or communicating, while the slave processes have idle times. On a HPC architecture these idle times could causes huge delays in communication and drastically increase the overall simulation time.

 ISBN: 978-84-617-8694-7

Algorithm 1: Parallel Algorithm for the Adams-Bashforth-Moulton scheme.

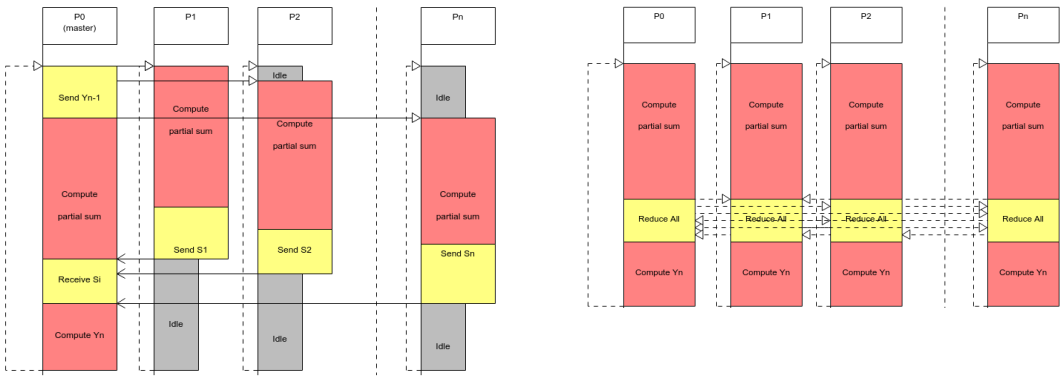
Data: T end of the time interval.
Data: N global number of points.
Data: P number of processes.
Data: p current process.
$N_P \leftarrow N/P$;
$y_0 \leftarrow$ initial condition;
for $n \in [1, N]$ **do**
  $Sp \leftarrow 0$;
  $Sc \leftarrow 0$;
  $n_{min} \leftarrow N_P p$;
  $n_{max} \leftarrow N_P(p+1)$;
  /* compute local sum for predictor and corrector */
  for $k \in [n_{min}, n_{max}]$ **do**
    $Sp \leftarrow Sp + b_{n-k} f_k$;
    $Sc \leftarrow Sc + c_{n-k} f_k$;
  end
  /* compute the global sum and sent to all processes */
  MPI_Allreduce(Sp, Sc);
  /* compute the predictor at time t_n */
  $y_n^P \leftarrow y_0 + a_n * Sp$;
  /* compute the corrector at time t_n */
  $y_n \leftarrow y_0 + c_n * (y_n^P + Sc)$;
end

In order to improve the overall simulation time, the workload needs to be balanced. First, the idle times of the processes need to be removed. Secondly, if possible, we need to decrease the number of messages, thus saving times in the communication part.

The parallel implementation method presented in Algorithm 1 reflects the core of our new implementation. The computation for the partial sum is done by each process and the final sum for the predictor and corrector is reduced to all processes (by MPI_Allreduce(Sp, Sc)). So instead of having the classical architecture where one process is the master and all the other processes are slaves just to compute the partial sum, in this approach all processes P compute the iteration function values (y_n), all acts as master processes.

In our new approach (see Figure 1b) we are able to avoid idle time for the processes, and we obtained a more balanced work load. Another advantage is that the overall communication between processes was reduced by removing the messages between the slaves

 ISBN: 978-84-617-8694-7



(a) The classical approach (b) The new improved approach

Figure 1: Workflow of execution processes

and master processes. There is only one message being exchanged when the global sum is reduced to all processes, and because the workload inside the process is balanced, the synchronized exchange time is very short. A similar method/solution was also presented and tested on PC using CPU core by [13] with very interesting results from the parallel computation point of view.

4 Simulation results

We implemented and tested the presented approach using the HPC cluster of the West University of Timişoara (Romania), namely, a BlueGene/P cluster that consists of a fully loaded single BlueGene/P rack that has more than 1000 CPUs and 4TB of RAM memory and can offer a performance up to 11.7 TFlops.

Table 1: Simulation results in seconds for different numbers of time steps

Case	No of time steps	HPC classic approach	CUDA run time	HPC run time
1	1000000	4621.25	2654.35	549.64
2	1500000	9162.33	4426.59	1158.13
3	2000000	14931.16	6616.77	2009.87
4	2500000	22697.66	9196.54	3066.92
5	3000000	31659.66	12206.16	4381.42

In Table 1 one can see the simulation run time results (in seconds) using different number of time steps (number of global points). We can see from the last column of this

 ISBN: 978-84-617-8694-7

table that the parallel algorithm proposed in this paper is up to 8 times more efficient than the classical approach of running the simulations on HPC, and also at least 3 times faster than CUDA simulations. The graphical representation of the running times (Figure 2) clearly presents the speed up of the new algorithm with all the processes as master processes.

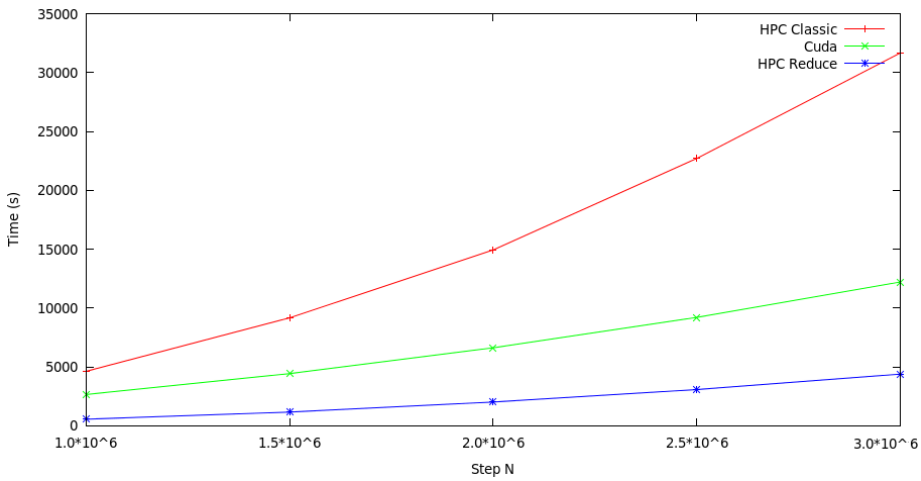


Figure 2: Running time comparison

Because our new implementation is much faster, we were able to implement simulations with even more than 5 million steps. This improvement allows us to compute the numerical solution of a fractional-order system over a large number of steps that provide us a better understanding about the system's behavior from the dynamic point of view. We will present more details about the numerical analysis of the test system in the next section.

In Figure 3 we present the average running time in seconds for the simulation on different number of steps. The time was computed between the simulation starting point until the results have been obtained. Those times include also the communication time between processes, but in our last proposed implementation the communication time is very low due to the fact that there are no idle times, respecting the HPC software development guidelines [12]. We observe from the simulations with the new implementation that the running time increases almost linearly with respect to the number of points even if more than 5 millions points are considered.

 ISBN: 978-84-617-8694-7

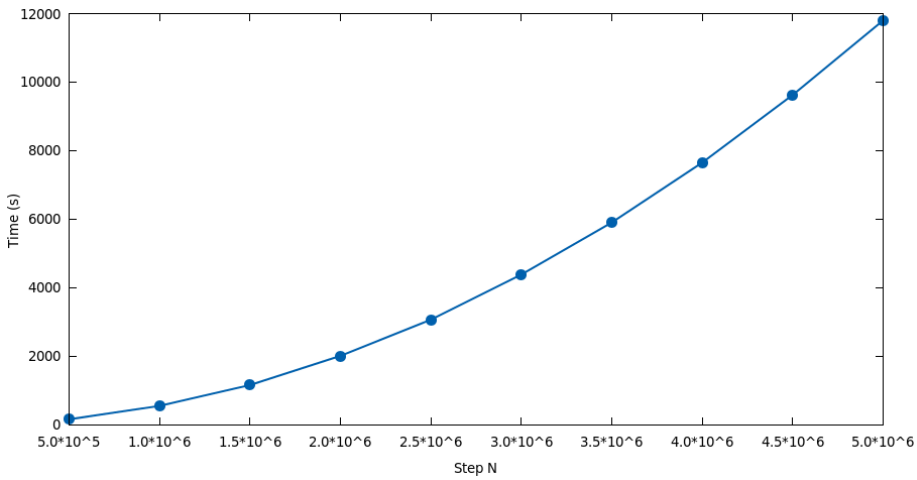


Figure 3: Running time speed HPC

5 Numerical experiment

Our test case is the fractional-order version of the normalized system describing a forced series LCR circuit [11]:

$$\begin{cases} D^{\alpha_1}x(t) = y - g(x) \\ D^{\alpha_2}y(t) = -\sigma y - x + f\sin(\omega t) \end{cases} \tag{3}$$

where $\alpha_1, \alpha_2 \in (0,1)$, $\sigma, f, \omega > 0$ and the function g is piecewise linear and is defined as:

$$g(x) = \begin{cases} bx - a + b, & \text{if } x \leq -1 \\ ax, & \text{if } |x| < 1 \\ bx + a - b, & \text{if } x \geq 1 \end{cases}$$

The parameter values considered for the numerical simulations are: $\sigma = 1.015$, $\omega = 0.55$, $a = -1.02$ and $b = -0.58$.

In the absence of the forcing term (i.e. $f = 0$), system (3) is autonomous and has three equilibrium states: $E_0 = (0,0)$ and $E_\pm = \left(\pm\frac{\sigma(a-b)}{1+\sigma b}, \pm\frac{b-a}{1+\sigma b}\right)$. However, when $f > 0$, the system (3) is non-autonomous and a series of period-doubling bifurcations leading to onset of chaotic behavior has been reported [11] when f is increased from 0 to 0.2, considering the fractional orders $\alpha_1 = \alpha_2 = 0.9$.

Using the HPC implementation of the parallel algorithm described in section 3, we are able to depict the dynamic behavior of system (3) with an improved precision compared to [11], using a small step size and computing the numerical solution over a large time

interval. Figure 4 shows the attractors of (3), for different values of the parameter f. For $f = 0.085$, the existence of two quasi-periodic attractors is observed and the period-doubling cascade actually involves both attractors, eventually leading to the appearance of two chaotic attractors (e.g. for $f = 0.117$). When the value of f is increased, these chaotic attractors collide and a double-scroll attractor takes their place (e.g. for $f = 0.125$). As we further increase f, a sequence of period-doubling bifurcations and reversed period-doubling bifurcations is observed, involving the single attractor of the system.

6 Conclusion and future work

Although the simulation execution times show good results, there is still room for improvement. A similar algorithm was implemented to run on PC using MPI or/and OpenMP which has similar results [13, 14]. Our next direction of research is having the software exploit the Blue Gene/P hardware in order to increase the performance.

This algorithm that implements the Adams-Bashforth-Moulton method is valid for solving any kind of fractional-order system with fractional derivatives of Caputo-type, hence, having the algorithm run with parameters and functions as input and execute the simulation as a black box would be nice to have.

As another direction for future research, the applicability of the Parareal algorithm (which has been initially developed to parallelize in time partial differential equations) to fractional-order differential equations, is also worth exploring.

Acknowledgements

This work was supported by a grant of the Romanian National Authority for Scientific Research and Innovation, CNCS-UEFISCDI, project no. PN-II-RU-TE-2014-4-0270.

References

[1] A. Baban, C. Bonchiş, A. Fikl, and F. Roşu, *Parallel simulations for fractional-order systems*, in Proceedings of SYNASC 2016, 141–144.

[2] D. Cafagna and G. Grassi, *Bifurcation and chaos in the fractional-order Chen system via a time-domain approach*, International Journal of Bifurcation and Chaos **18** (2008) 1845–1863.

[3] G. Cottone, M. D. Paola, and R. Santoro, *A novel exact representation of stationary colored gaussian processes (fractional differential approach)*, Journal of Physics A: Mathematical and Theoretical **43** (2010), p. 085002.

 ISBN: 978-84-617-8694-7

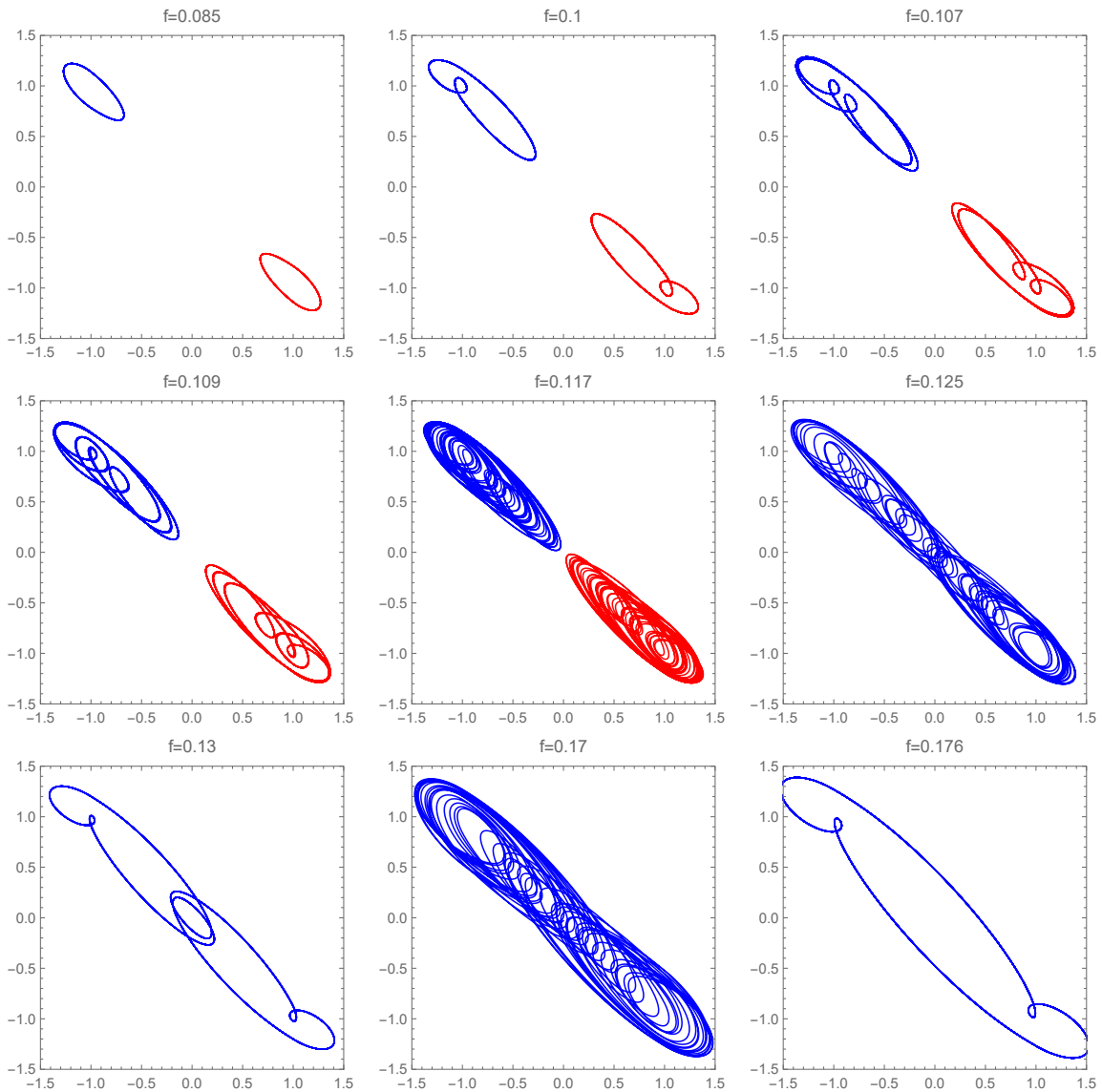


Figure 4: Rich dynamic behavior in system (3), for $\alpha_1 = \alpha_2 = 0.9$ and different values of the parameter f.

 ISBN: 978-84-617-8694-7

[4] K. Diethelm, *An efficient parallel algorithm for the numerical solution of fractional differential equations*, Fractional Calculus and Applied Analysis **14** (2011) 475–490.

[5] K. Diethelm, N. Ford, and A. Freed, *A predictor-corrector approach for the numerical solution of fractional differential equations*, Nonlinear Dynamics **29** (2002) 3–22.

[6] N. Engheia, *On the role of fractional calculus in electromagnetic theory*, IEEE Antennas and Propagation Magazine **39** (1997) 35–46.

[7] L. Galeone and R. Garrappa, *Explicit methods for fractional differential equations and their stability properties*, Journal of Computational and Applied Mathematics **228** (2009) 548–560.

[8] B. Henry and S. Wearne, *Existence of Turing instabilities in a two-species fractional reaction-diffusion system*, SIAM Journal on Applied Mathematics **62** (2002) 870–887.

[9] N. Heymans and J.-C. Bauwens, *Fractal rheological models and fractional differential equations for viscoelastic behavior*, Rheologica Acta **33** (1994) 210–219.

[10] F. Mainardi, *Fractional relaxation-oscillation and fractional phenomena*, Chaos Solitons Fractals **7** (1996) 1461–1477.

[11] J. Palanivel, K. Suresh, S. Sabarathinam, and K. Thamilmaran, *Chaos in a low dimensional fractional order nonautonomous nonlinear oscillator*, Chaos, Solitons & Fractals **95** (2017) 33–41.

[12] I. Redbooks, *IBM System Blue Gene Solution: Blue Gene/P Application Development*, Vervante, 2009.

[13] W. Zhang and X. Cai, *Efficient implementations of the Adams-Bashforth-Moulton method for solving fractional differential equations*, Proceedings of FDA12, (2012).

[14] W. Zhang, W. Wei, and X. Cai, *Performance modeling of serial and parallel implementations of the fractional Adams-Bashforth-Moulton method*, Fractional Calculus and Applied Analysis **17** (2014) 617–637.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

Scaling Probabilistic Record Linkage on Multicore and Multi-GPU Systems

Murilo Boratto[1], Pedro Alonso[2], Clicia Pinto[3], Pedro Melo[3], Marcos Barreto[3] and Spiros Denaxas[4]

[1] *Núcleo de Arquitetura de Computadores e Sistemas Operacionais, Universidade do Estado da Bahia*

[2] *Departamento de Sistemas Informáticos y Computación, Universitat Politècnica de València*

[3] *Laboratório de Sistemas Distribuídos, Universidade Federal da Bahia*

[4] *Institute of Health Informatics Research, University College London*

emails: muriloboratto@uneb.br, palonso@upv.es, cliciasp@ufba.br, pmelo@ufba.br, marcoseb@ufba.br, s.denaxas@ucl.ac.uk

Abstract

Record linkage is a widely used technique to aggregate data stored in disparate sources that presumably pertain to the same real world entity. This should be done probabilistically if there are no common key attributes in all data sources involved. This approach is very time consuming if we consider the amount of data that must be compared, specifically in big data scenarios. In this manuscript, we propose a methodology for exploiting multicore and multi-GPU architectures to probabilistic link large-scale, national administrative data sources of more than 100 million participants from the Brazilian Public Health System.

Key words: Record Linkage, Performance, Multicore, Multi-GPU

ISBN: 978-84-617-8694-7

1 Introduction

The task of linking multiple, disparate database records representing data from the same real world entity is known as *record linkage* [1, 3], being a technique widely used in biomedical and health research, finance, government and other domains. Specifically in Health research, data stored in disparate information systems need to be combined for diverse purposes including aggregation of medical and hospital services, assessment of public health policies, surveillance and monitoring. This is often a challenging task as data quality, complexity and size dramatically differs among the data sources.

There are two main approaches for record linkage: deterministic and probabilistic. Deterministic linkage uses a combination of one or more unique identifying attributes that are common across the data sources to link records, whereas probabilistic linkage is used when these common attributes are absent. In order to improve their accuracy on matching decisions, probabilistic methods must perform a huge number of comparisons, being characterized as complex and highly time-consuming tasks.

Our research involves the probabilistic linkage of several large governmental databases with socioeconomic and health care data from the Brazilian Public Health System. These databases are linked in order to create accurate "data marts" used for epidemiological studies. Specifically, these studies are part of three ongoing Brazil-UK scientific collaborations: 1) *the 100 million cohort project*, in which the effects of a conditional cash transfer programme on health outcomes (e.g. leprosy, tuberculosis, HIV) from 114 million people are investigated; 2) a *long-term surveillance platform for Zika and microcephaly*, a longitudinal study (2001 to 2015) of children diagnosed with microcephaly and other Zika-related illnesses; and 3) a *platform linking data from Malaria transmission, patient care and monitoring* to provide support for analytical methods targeting the elimination of the disease.

Besides the absence of common key attributes, other important issues related to heterogeneity and privacy arise from the sensitive nature of the data stored in health databases. Methods for data harmonization and anonymization must be applied prior to the record linkage step affecting, in turn, the execution time.

Another technical aspect relates to blocking construction: depending on the hardware capabilities, one must group the records into blocks with some similarity criterion, usually the values of a given set of attributes, to avoid unnecessary comparisons. It can affect accuracy if records that must pertain to similar blocks are not correctly grouped.

Given the complexity involved in implementing probabilistic record linkage methods targeted to huge databases, novel algorithmic approaches to fully exploit the processing power of mutliple resources are required. Hybrid parallel architectures can be considered as a viable approach, even introducing challenges to algorithm design and system software.

Our proposal is to balance the workload across all available CPUs and GPUs present in heterogeneous systems to perform the necessary data operations in a timely and accurate manner. We deal with optimal workload distribution for hybrid systems [5], efficient grid

configurations on GPUs, analysis of data transfers, among other hardware specific issues.

2 Experimental Results

Our execution environment comprises 4 Intel Xeon at 2.93 GHz and 130 GB DDR3 main memory. Each one is a quadcore processor with 24 MB of cache memory. It contains two NVIDIA Tesla K40 GPUs. The installed CUDA toolkit is version 4.0.

In our experiments we used a parallel algorithm for the probabilistic linkage model that uses OpenMP [4] and CUDA. Many parameter values were used at installation time to estimate the best values for system parameters. The available range for CPU cores (c) is 1, 2, ..., 32 (Intel Hyper-Threading [2] is set). We checked for GPUs workloads (w) from 10% to 45%. The input sizes of the problem and the number of records (s) for the experiments were $1,000,000$, $2,000,000$, ..., $20,000,000$. The workload distribution used used is (w) = (GPU, GPU, CPU) = (35%, 35%, 30%).

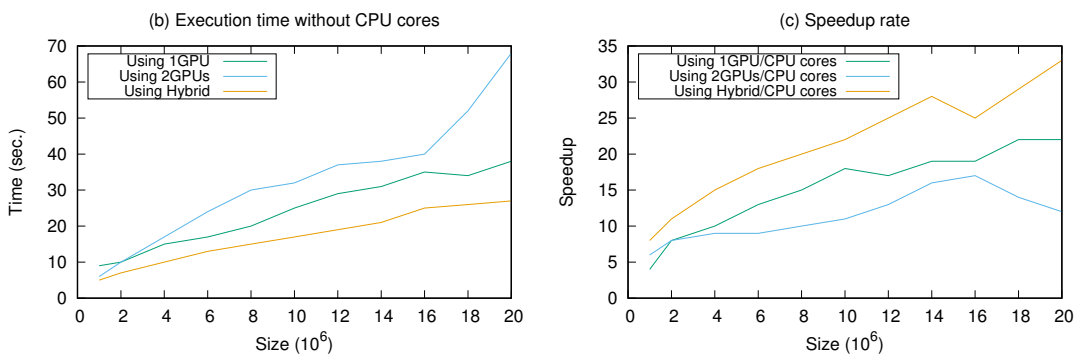


Figure 1: Execution of the parallel probabilistic linkage algorithm. (a) Execution time. (b) Speedup rate.

The execution time with multiple threads is denoted by "CPU cores". It distributes the calculation data mart among the threads and each thread runs exclusively on a CPU core. Versions denoted by "1GPU" and "2GPUs" represent executions in one single and two devices, respectively. The heterogeneous model ("Hybrid") uses all cores available in the heterogeneous system. In this model, the threads are executed by all the elements in the machine with the suitable number of CPU cores and the two GPUs. The results show that the parallel CPU algorithm reduces the execution time significantly.

We show in Fig. 1(a) and Fig. 1(b), the execution time and the speedup, respectively, for the probabilistic linkage algorithms with different sizes ranging from $1,000,000$ to $20,000,000$. The execution was carried out on each subsystem independently (1GPU, 2GPUs and Hybrid) to have a measure for comparison purposes.

 ISBN: 978-84-617-8694-7

Speedup has been obtained with regard to the use of the CPU cores subsystem only. As can be seen in Figure 1(b), the maximum speedup is around 30 with the hybrid subsystem, presenting a difference in performance that can more clearly be observed. Both plots show how the use of GPUs in our system clearly outperforms the computation on the CPU cores. The performance of 1GPU is larger than the performance with 2GPUs (size (s) $<= 20,000,000$). In this case, this is due to the setup time needed for device selection, which is high in our target machine and is not necessary if just one GPU is used. The use of GPU as a standalone tool provides benefits but does not allow to reach the potential performance that could be obtained by adding more GPUs and/or the CPU subsystem.

3 Conclusions and Future Directions

We aim to extend these experiments by exploring multi-GPU with hybrid environments, enjoying the great field of open problems we are dealing with. The parallel algorithm developed enables an efficient computation of the integral that appears in probabilistic record linkage method. Based on this preliminary work and pursuing the challenge of develop more sophisticated solutions, we aim to scale the proposed work to larger databases with more than 100 million records. We claim this proposal is high valuable as it contributes to solve current computational problems related to record linkage of large Brazilian governmental datasets, representing a substantial benefit for public health. Finally, more experiments in systems with larger numbers of cores of different architectures (i.e., multicore + multi-GPU + multi-MIC) and with other linear algebra routines are needed. To achieve this, an adaptation of a technique based on theoretical models should be analyzed.

References

[1] A.H. Doan, A. Halevy, and Z. Ives. *Principles of Data Integration*. Elsevier Science, 2012.

[2] Erik Yama Étienne. *Hyper-threading*. TurbsPublishing, 2012.

[3] I. P. Fellegi and A. B. Sunter. A theory for record linkage. *Journal of the American Statistical Association*, 64:1183–1210, 1969.

[4] OpenMP Architecture Review Board. OpenMP application program interface version 4.0, jul 2013.

[5] Ziming Zhong, Vladimir Rychkov, and Alexey Lastovetsky. Data partitioning on multicore and multi-gpu platforms using functional performance models. *IEEE Transactions on Computers*, 64(9):2506–2518, 2015.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

Modeling CA15-3 longitudinal progression in patients with breast cancer recurrence

Ana Borges[1]

[1] *CIICESI/ESTG, Porto Polytechnic*

[2] *CMAT, DMA-ECUM, University of Minho*

[3] *Senology Unit, Braga's Hospital*

emails: aib@estg.ipp.pt

Abstract

We analyse the progression in time of a tumour marker used on the surveillance of this disease - the Carcinoma Antigen 15-3 (CA15-3) for Braga's Hospital senology unit patients, located in Portugal. Our main purpose is to describe the progression of this tumour marker, for the subset of patients that suffer a recurrence, as a function of possible risk factors. Also, to understand how these risk factors influences that progression. We intend, as well, to detect a possible changing point in the mean progression, that could be used by clinicians as an early detection of the recurrence. The response variable, values of CA15-3, was analysed making use of longitudinal models, testing for different correlation structures. The reference time considered was time from breast cancer recurrence diagnose until blood test data. For diagnostic of the models fitted we have used empirical and theoretical variograms. To evaluate the fixed term of the longitudinal model we have tested for a changing point on the effect of time on the tumour marker progression. Results show that, one year before the detection of breast cancer recurrence there is an abrupt rise on the rate of its progression. The presence of venous vascular invasion of the tumour affects the progression of CA15-3 values, for the subset of patients with breast cancer recurrence.

Key words: Longitudinal models, Breast Cancer, Recurrence, random effects, correlation structure, CA15-3

1 Introduction

As [1] report, breast cancer is the second most common cancer in the world and the most frequent cancer among women. They estimated that 1.67 million new cancer cases were diagnosed in 2012 (25% of all cancers). It is, actually, the most common cancer in women both in more and less developed regions with slightly more cases in less developed (883 000 cases) than in more developed (794 000 cases) regions.

Data were collected directly from the medical records of each patient, listed in the computer system of Braga's Hospital Glintt HS. We therefore have access to baseline and clinical history of each patient (a roll of information such as diagnosis; pre-surgery, post-surgery, group meetings; follow-up and medical exams). The authorization for collect and use of senology data was approved by the Ethical Committee of Braga's Hospital.

We gathered information on all patients diagnosed with breast cancer from 2008 until 2012, in Braga's Hospital, and for all the patients in follow-up at the hospital at the date of 1st January 2008.

Braga's Hospital is situated in the east of the city of Braga, located in northern Portugal. Today it serves a direct area of about 275,000 patients. In 2008, a Senology unit at the Braga's Hospital was created. Currently, it operates an average of 130 new cases of breast cancer per year.

We collected information on 540 patients. The total number of deaths is 74, however the total number of deaths from breast cancer is only 55. Throughout the followup time it was detected a recurrence of breast cancer (local and/or distant) on 81 patients. Of these, 52 died from breast cancer.

From the information gathered in all the medical reports, we were able to collect more than 50 variables that group information at patient level (such as age at diagnosis, residence location, number of births, etc.) and at tumour level (stage, tumour size, estrogen receptor expression, etc.), including measurements of CA15-3 tumour marker values and the date when that measurement was registered.

As [7] explain, Carcinoma Antigen 15-3 (CA15-3) is the most widely used serum marker in breast cancer. It consists in a large transmembrane glycoprotein which is frequently over-expressed and aberrantly glycosylated in cancer. Currently the main uses of CA15-3 are in preclinically detecting recurrent breast cancer and monitoring the treatment of patients with advanced breast cancer, although its clinical value is not validated in a high-level evidence study, as pointed out by the American Society of Clinical Oncology. Nevertheless, as

 ISBN: 978-84-608-6082-2

[8] point out, there are several well-designed studies that show that an increase in CA15-3 after primary and/or adjuvant therapy, can predict recurrence an average of 5 to 6 months before other symptoms or tests.

We only have information on tumour marker CA15-3 values for 534 patients. This translates into a total number of 552 cases analysed, since 18 cases presented bilateral breast cancer. The total number of deaths from breast cancer is 55. There were 5166 measurements of tumour marker CA15-3, with a number of observations per patient varying between 1 and 48 measurements. Being 8 the median number of measurements per person.

Since it is a usual medical procedure to be alert for possible tumor recurrence in the case of detecting a rise in levels of this marker above a certain reference value, our main purpose is to describe the progression of this tumor marker, on patients who were followed and treated in this Unit, that presented a breast cancer recurrence, as a function of possible risk factors. We intend to estimate on average the time to the increase of this tumor marker, as so to characterize the degree of heterogeneity between patients.

In this particular study, the response variable, values of CA15-3 tumour marker, was analysed making use of longitudinal models as defined in [9], where different correlation structures were tested. Also, to evaluate the fixed term of the longitudinal model we have tested for a changing point on the effect of time on the tumour marker progression for the subset of patients that died from breast cancer, and for the subset of patients that presented a breast cancer recurrence.

The analysis presented is an extension of the work published by [10]. Firstly it will be presented the methodology applied, introduced by a summary explanation of the main approaches to analyze longitudinal data, describing the longitudinal models used to infer about the factors that affect the progression of CA15-3 tumour marker values. Subsequently, the main results from the longitudinal analysis of this response will be demonstrated ending with a conclusion section.

The entire analysis was performed using R software, in particular making use of both *nlme* [11] and *JoineR* [12] packages.

2 Methodology

The general longitudinal model described as in [9] is:

$$Y_{ij} = \mu_{ij} + \mathbf{d}_{ij}^{'}\mathbf{U}_i + W_i(t_{ij}) + Z_{ij}. \quad (1)$$

 ISBN: 978-84-608-6082-2

Where $\mathbf{U}_i$ are n i.i.d. realizations of $MVN(0, \nu^2)$, representing the random effects at individual level, and $\mathbf{d}_{ij}^{'}$ is a vector of covariates for the random effects. $W_i(t_{ij})$ is a continuous time Gaussian Process with $E[W_i(t_{ij})] = 0$ and $E[W_i(t_{ij})] = \sigma^2$, representing the variability within subjects, where the correlation between two measurements of an individual is described by: $corr(W_i(t_{ij}), W_i(t_{ik})) = \rho(t_{ij}, t_{ik})$.
Finally, Z_{ij} are N i.i.d. realizations of $N(0, \tau^2)$, representing the measurement error (variability non specified).

Since $W_i(t_{ij})$ is assumed a stationary process we have $\rho(t_{ij}, t_{ik}) = \rho(|t_{ij} - t_{ik}|)$.

We can have different definitions for the function $\rho(|t_{ij} - t_{ik}|)$. That is, if we consider the correlation among $W_i(t_{ij})$, let say between $W_i(t)$ and $W_i(t-u)$, determined by the autocorrelation function $\rho(u)$, we will have for a longitudinal model that accounts for an exponential correlation structure within individuals:

$$\rho(u) = \exp\left(-\frac{1}{\phi}|u|\right), \tag{2}$$

and for a longitudinal model that accounts for a Gaussian correlation structure within individuals:

$$\rho(u) = \exp\left(-\frac{1}{\phi}u^2\right), \tag{3}$$

where ϕ is the range parameter that specifies the rate at which the correlation stables.

To model the fixed term of the longitudinal model, μ_{ij}, we can consider a model with a changing point δ on the effect of time on the response variable. In practice, the changing point is the moment where there is an alteration on the slope of the linear response variable's progression, on average. Considering δ the changing point, we have $E[Y_{ij}] = \mu_{ij}$ with:

$$\mu_{ij} = \begin{cases} X_{ij}\beta + \alpha_1 t_{ij}, & \text{if } t_{ij} < \delta \\ X_{ij}\beta + \alpha_2(t_{ij} - \delta) & \text{if } t_{ij} \geq \delta \end{cases}, \tag{4}$$

where X_{ij} represents the vector of covariates, β the vector of unknown regression coefficients, α_1 and α_2 the coefficients representing the slope before and after the changing point, respectively.

For simplicity, let's consider the complete set of N measurements, y, as a realization of a multivariate Gaussian random vector Y with $Y \sim MVN(X\beta, \eta^2 V)$. Where $\mathbf{X}$ is an $N \times p$ matrix of all values of the p explanatory variables. And $\eta^2 V$ as a block-diagonal matrix with non-zero $n \times n$ blocks $\eta^2 V_0$, each representing the variance matrix for the vector of measurements on a single subjects.

 ISBN: 978-84-608-6082-2

For parameter estimation we adopted the maximum likelihood method, Since we are dealing with unbalance type of data and, also, the missing data detected in our data base does not seem to be related to the progression of the disease, we consider the completely random mechanism of missing data. Which can be ignorable when adopting the likelihood function as the basis for inference [13].

To model the correlation structure for each model we analysed the empirical variogram of OLS residuals from the saturated model for the mean response [9].

The variogram [9] of a stochastic process $Y(t)$ is given by:

$$V(u) = \frac{1}{2}\text{Var}\left\{Y(t) - Y(t-u)\right\},\ u \geq 0. \tag{5}$$

For a stationary process, the autocorrelation function, $\rho(u)$, and the variance of $Y(t)$, σ^2, are related by:

$$\gamma(u) = \sigma^2\{1 - \rho(u)\}. \tag{6}$$

The estimation of the empirical variogram is based on the calculation of the observed half-squared-differences between pair of residuals, $\nu_{ij} = \frac{1}{2}(r_{ij} - r_{ik})^2$, and the corresponding time-differences, $u_{ijk} = t_{ij}t_{ik}$, where $r_{ij} = Y_{ij}\mu_{ij}$, and $j < k = 1, ..., m_i$.

The autocorrelation function at any lag u is estimated from the sample variogram by:

$$\hat{\rho}(u) = 1 - \frac{\hat{\gamma}(u)}{\hat{\sigma}^2}, \tag{7}$$

where $\hat{\gamma}(u)$ is the average of all the ν_{ij} corresponding to that particular value of u, and $\hat{\sigma}^2$ is the estimated process variance.

Diagnostic checking of the longitudinal fitted model can be made by superimposing the fitted mean response profiles on a time-plot of the average observed response within each combination of covariates categories and time, as suggested by [9]. However, this task can be rather difficult if we are dealing with a model that incorporates a great number of covariates. One other form of model fitting diagnose is to superimpose the fitted variogram on a plot of the empirical variogram. Which will allow the validation of the correlation structure by graphical comparison. Another procedure to additionally access which correlation structure best describe data variability, within a longitudinal model, with a predetermined fixed part, is by comparison of the maximized log likelihood values.

Plot of the subject specific residuals versus the fitted responses and a Q-Q plot of the subject-specific residuals can, also, be used for graphical validation of the assumptions of the constant variance and normal distribution of Z_{ij}, respectively.

 ISBN: 978-84-608-6082-2

2.1 Main Results

According to the usual medical procedures, physicians stay alert to a possible recurrence of breast cancer for patients that present values of tumour markers above a certain reference value. We have considered the reference value of 37 U/ml for CA15-3 tumour marker, as this were the values referred in the blood tests analysed.

The normality assumption of both response variables CA15-3 was first tested preforming a Shapiro-Wilk normality test, which returned p-value < 0.0001 and therefore was rejected. It was used a log-transformation of the tumor marker CA15-3 as It is a usual transformation in many biological cancer markers studies.

A longitudinal model was fitted to the subset of patients that presented a breast cancer recurrence. We considered the reference time, in month, from date of recurrence diagnose until blood tests. Note that, doing so, reference time will be on a negative scale. We began with an exploratory analysis and point estimation by modelling a OLS model with all the variables. We then conducted a backward elimination to delete variables not significant, with a limit of 0.08 for the p-value, for inclusion of the variable, until the mean structure was well defined with only significant covariates.

Observing graphical representation of the empirical variogram, patterns suggested the existence of variability between subjects, and a possible variability within subjects. Hence, maintaining the same mean structure we compared two nested models with different covariance structures, such as: (i) random effects, exponential serial correlation and measurement error (REE), and (ii) random effects, Gaussian serial correlation and measurement error (REG).

Figure 1 presents the progression in time, since date of breast cancer recurrence detection, of the CA15-3 values for each patient, with grey lines, against the reference value, and the non-parametric smooth spline line with dash line, indicating the average trend of progression.

The smooth spline of the spaghetti plot suggests that, on average, the marker progression increases at a slow rate until about 20 months before death from breast cancer. In fact, when fitting the saturated general linear model for the subset of patients who died from breast cancer, given by equation 4, we detected a changing point at 1 years before recurrence. Which means that, for the patients which was detected breast cancer recurrence it was detected an abrupt rise of CA15-3 values, above the reference value, 1 years before the detection.

 ISBN: 978-84-608-6082-2

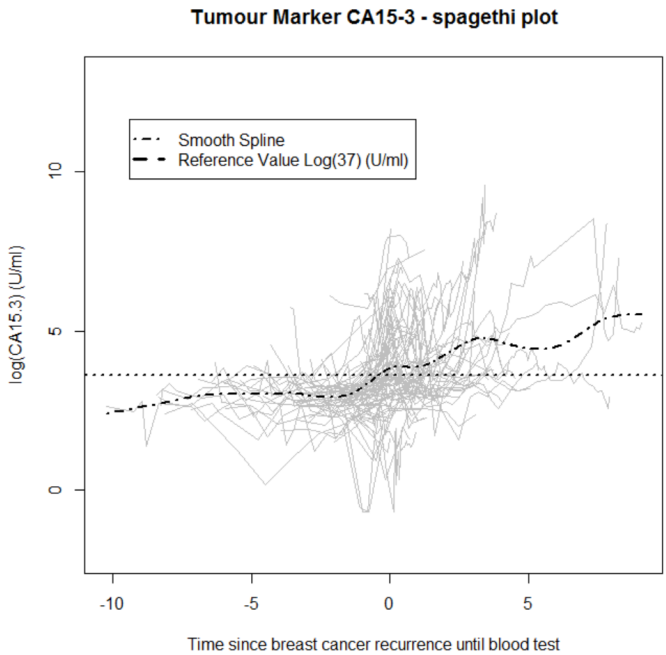


Figure 1: Spaghetti plot for tumor marker CA15-3 values - for the subset of patients with breast cancer recurrence.

Table 1 summarizes and compares the estimated parameters for the two longitudinal models fitted REE and REG - with those of the general linear model (OLS Model). Note that the variable "Time further than 1 year before recurrence" is, in fact, time before the changing point (2 years before death from breast cancer), in the negative scale considered and, equivalently, "Time earlier than 1 year before recurrence" is time after the changing point, in the negative scale considered.

We selected the REE longitudinal model to describe the progression of the tumour marker CA15-3 values in time, since it was the one with higher Log likelihood value.

Results show that time after the changing point, that is before one year prior to the detection of the breast cancer recurrence, has a significant effect (p-value < 0.001) on the mean progression of the tumour marker values, related to an increase of 0.394 per month of the CA15-3 values. However, before changing point the effect of time is not significant (p-value=0.9235) in the mean progression of the tumour marker values.

 ISBN: 978-84-608-6082-2

Table 1: Estimated Parameters Values for General Linear Model and Longitudinal Models For Patients with Breast Cancer Recurrence – CA15-3 Tumour Marker

	REE Model		**REG Model**		**OLS Model**	
	Est	**p-value**	**Est**	**p-value**	**Est**	**p-value**
Intercept	3.110	< 0.001	3.0978098	< 0.001	3.31230	< 0.001
Time further than 1 year before recurrence	0.0055292	0.9235	0.0225484	0.5870	0.07992	0.0168
Time earlier than 1 years before recurrence)	0.3942085	< 0.001	0.3571825	< 0.001	0.31785	< 0.001
Venous Vascular Invasion (Yes)	1.2193974	< 0.001	1.3166430	< 0.001	1.28602	< 0.001
ν^2	0.00000042		0.525		.	
σ^2	0.002		0.091		.	
ϕ	19.958		7.451		.	
τ^2	1.425		0.772		.	
ξ^2	.		.		1.428	
Log Likelihood	-502.163		-523.639		-724.714	

For the subset of patients that presented a breast cancer recurrence, the mean progression of the marker is only composed by one significant covariate on the intercept component of the model - the presence of images of venous vascular invasion (yes versus no). The intercept component of the model, in this particular case, means that a patient with no venous vascular invasion images and a triple negative type of tumour will start the progression of the tumor marker with a value of 3.11, on a logarithmic scale.

As expected, the presence of venous vascular invasion has an increasing effect on the average CA15-3 linear progression in time, as it is related to a worst prognostic case in the previous survival analysis. A case with venous vascular invasion an increment of 1.22 compared to those with no venous vascular invasion.

For this subset, the correlation structure that best represent the variability of the data is the structure that incorporates random effects at individual level with $\hat{\nu}^2 \approx 3 \times 10^{-8}$, an exponential correlation structure to describe the variability within patients with $\hat{\rho}(u) \approx \exp(-\frac{1}{23.247}.|u|)$ and $\hat{\sigma}^2 \approx 0.040$, and a measurement error with variance $\hat{\tau}^2 \approx 1.826$. The superposition of the theoretical variogram of both exponential and exponential correlation structures with the empirical variogram, presented in Figure 2, validates the choice of an exponential correlation structure, since it is the one that best approximates the empirical

 ISBN: 978-84-608-6082-2

variogram curve.

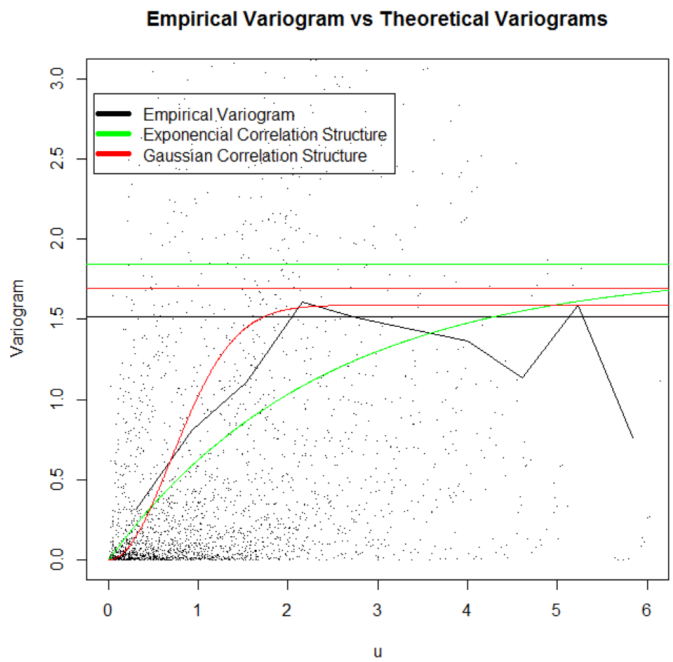


Figure 2: Superposition of empirial variogram and theoretical variograms, for patients with breast cancer recurrence - CA15-3 tumour marker.

2.2 Conclusion

An abrupt rise in values of CA15-3 tumor marker, over a reference value, is an alert sign to a possible recurrence of breast cancer.
When analyzing all patients that were diagnosed with breast cancer, it was detected a changing point on the linear progression of the tumor marker for the subset of patients that died from breast cancer two years before the death. This means that, at that point, there is an abrupt rise on the rate of its progression.
The risk factors for the progression of the marker, for that subset of patients are: bilateral (yes versus no), images of venous vascular invasion (yes versus no) and estrogen receptor expression(positive versus negative). However, for the subset of patients that suffered a breast cancer recurrence it was detected a changing point on the linear progression of the tumour marker one year prior to the detection of the recurrence. And the only covariate with significant effect on the progression was the images of venous vascular invasion. This

 ISBN: 978-84-608-6082-2

is an important result, as it can be used as an alert to a possible state of breast cancer recurrence if the doctor detects an abrupt increase of the tumour marker progression of this nature.
For both models fitted, the fact that the estimated variance of the measurement error is quite lower that the estimated variance of the OLS model means that the fitted REE longitudinal model explains the variability of the data mainly by means of variability between patients and within patients assigning a very low value for measurement error (or *white noise* as usually mentioned in literature).
The fact that, when comparing the REE and the REG models to a longitudinal model with only an intercept random effect, the component the serial correlation was significant stresses the importance incorporating a variability component that translates within subject measurements correlation, in this type of biological data.
The presented longitudinal analysis of this tumor marker, in combination with the previous survival analysis, where the event of interest is breast cancer recurrence, is going to be proceeded, in future work, with a joint modeling of the longitudinal and survival process of the present data.

Acknowledgements

Ana Borges thanks the support by Center for Research and Innovation in Business Sciences and Information Systems (CIICESI), ESTG - P.Porto. and the FCT fundation, Portugal, for the Ph.D. Grant SFRH/BD/74166/2010, for the collection of the data used in the present analysis.

References

[1] J. Ferlay, I. Soerjomataram, M. Ervik, R. Dikshit, S. Eser, C. Mathers, M. Rebelo, DM. Parkin, D. Forman, and F. Bray, (2013), *GLOBOCAN 2012 v1.0, Cancer Incidence and Mortality Worldwide: IARC CancerBase No. 11 [Internet]*; Lyon, France: International Agency for Research on Cancer, source = http://globocan.iarc.fr, accessed on day/month/year

[2] AJ. Chiang, J. Chen, YC. Chung, HJ. Huang, WS. Liou, and C. Chang, (2014), *A longitudinal analysis with CA-125 to predict overall survival in patients with ovarian cancer.*; Clinical Cancer Research, 25, 1, 51-57.

[3] Rodrigues V., (2011) *Chapter 34. In "Manual de Ginecologia"*, Permanyer, Portugal, 175–191.

 ISBN: 978-84-608-6082-2

[4] D. Trichopoulos, H.O. Adami, A. Ebkom, C.C. Hsieh, P. Lagiou, (2008, *Early life events and conditions and breast cancer risk: from epidemioloy to etiology*; Int. J. Cancer, 122, 481-485.

[5] Fitzgibbons P.L.; Page D.L.; Weaver D.; Thor A.D.; Allred D.C., and Clark G.M., (2000), *Prognostic factors in breast cancer, College of American Pathologists Consensus Statement 1999*; Archives of pathology & laboratory medicine, 124, 7, 966-978.

[6] Cianfrocca M., and Goldstein LJ., (2004), *Prognostic and predictive factors in early-stage breast cancer*; The Oncologist, 9, 6, 606-616.

[7] M. Duffy, S. Shering, F. Sherry, E. Mc-Dermott, and N. OHiggins, (2000), *Ca 15-3: a prognostic marker in breast cancer*; The International Journal of Biological Markers, 15, 330-333.

[8] Harris L.; Fritsche H.; Mennel R.; Norton L.; Ravdin P.; Taube S.; Somerfield MR.; Hayes DF., and Bast RC Jr., (2007), *American Society of Clinical Oncology: American Society of Clinical Oncology 2007 update of recommendations for the use of tumor markers in breast cancer*; Journal of Clinical Oncology, 25, 5287-5312.

[9] Diggle PJ; Heagerty P; Liang K-Y, and Zeger SL, (2002), *Analysis of Longitudinal Data*; University Oxford Press.

[10] A. Borges, I. Sousa, L. Castro, (2015), *Longitudinal Analysis of Tumor Marker CEA of Breast Cancer Patients from Braga's Hospital*, REVSTAT Statistical Journal, 13, 1, 63-78.

[11] Pinheiro J, Bates D, DebRoy S, Sarkar D and R Core Team, (2014), *nlme: Linear and Nonlinear Mixed Effects Models*. R package version 3.1-118, http://CRAN.R-project.org/package=nlme.

[12] Philipson P.; Sousa I.; Diggle P., Williamson P.; Kolamunnage-Dona R.; Henderson R.and R Core Team, (2014), *JoineR: Joint modelling of repeated measurements and time-to-event data*. R package version: 1.0-3, http://CRAN.R-project.org/package=joineR.

[13] Little R. J.A., and Rubin DB, (2002), *Statistical Analysis with Missing Data*, 2nd Edition, Wiley-Interscience.

 ISBN: 978-84-608-6082-2

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

Radial and Angular Derivatives of Special Classes of Distributions

Fred Brackx[1]

[1] *Clifford Research Group, Department of Mathematical Analysis*
Faculty of Engineering and Architecture, Ghent University

emails: Freddy.Brackx@UGent.be

Dedicated to Wolfgang Sprößig celebrating his 70th birthday

Abstract

When expressing a distribution in Euclidean space in spherical co-ordinates, derivation with respect to the radial and angular co–ordinates is far from trivial. Exploring the possibilities of defining a radial derivative of the delta distribution $\delta(\underline{x})$ (the angular derivatives of $\delta(\underline{x})$ being zero since the delta distribution is itself radial) led, see [2], to the introduction of a new kind of distributions, the so–called *signumdistributions*, as continuous linear functionals on a space of test functions showing a singularity at the origin. In this paper we search for a definition of the radial and angular derivatives of a general standard distribution and again, as expected, we are inevitably led to consider signumdistributions. Although these signumdistributions provide an adequate framework for the actions on distributions aimed at, it turns out that the derivation with respect to the radial distance of a general (signum)distribution is still not yet unambiguous.

Key words: distribution, radial derivative, angular derivative, signumdistribution
MSC 2000: 46F05, 46F10, 30G35

1 Introduction

Let us consider a scalar–valued distribution $T(\underline{x}) \in \mathcal{D}'(\mathbb{R}^m)$ expressed in terms of spherical co–ordinates: $\underline{x} = r\underline{\omega}$, $r = |\underline{x}|$, $\underline{\omega} = \sum_{j=1}^{m} e_j\,\omega_j \in \mathbb{S}^{m-1}$, $(e_j)_{j=1}^{m}$ being an orthonormal basis of $\mathbb{R}^m$ and $\mathbb{S}^{m-1}$ being the unit sphere in $\mathbb{R}^m$. The aim of this paper is to search for

ISBN: 978-84-617-8694-7

an adequate definition of the radial and angular derivatives $\partial_r T$ and $\partial_{\omega_j} T$, $j = 1, \ldots, m$. This problem was treated in [2] for the special and interesting case of the delta distribution $\delta(\underline{x})$, the following spherical co–ordinates expression of which is often encountered in physics texts:

$$\delta(\underline{x}) = \frac{1}{a_m} \frac{\delta(r)}{r^{m-1}} \tag{1}$$

where $a_m = \frac{2\pi^{\frac{m}{2}}}{\Gamma(\frac{m}{2})}$ is the area of the unit sphere $\mathbb{S}^{m-1}$ in $\mathbb{R}^m$ and $\delta(r)$ is the one–dimensional delta distribution on the real r–axis. Apparently expression (1) can mathematically be explained in the following way. Write the action of the delta distribution as an integral:

$$\begin{aligned} \varphi(0) = \langle\, \delta(\underline{x}), \varphi(\underline{x}) \,\rangle &= \int_{\mathbb{R}^m} \delta(\underline{x})\, \varphi(\underline{x})\, dV(\underline{x}) \\ &= \int_0^\infty r^{m-1} \delta(\underline{x})\, dr \int_{\mathbb{S}^{m-1}} \varphi(r\,\underline{\omega})\, dS_{\underline{\omega}} \\ &= a_m \int_0^\infty r^{m-1}\, \delta(\underline{x})\, \Sigma^0[\varphi](r)\, dr \end{aligned}$$

introducing the so–called *spherical mean* of the test function φ given by

$$\Sigma^0[\varphi](r) = \frac{1}{a_m} \int_{\mathbb{S}^{m-1}} \varphi(r\,\underline{\omega})\, dS_{\underline{\omega}}$$

As it is easily seen that $\Sigma^0[\varphi](0) = \varphi(0)$ it follows that

$$a_m \int_0^\infty r^{m-1}\, \delta(\underline{x})\, \Sigma^0[\varphi](r)\, dr = \int_0^\infty \delta(r)\, \Sigma^0[\varphi](r)\, dr = \langle\, \delta(r), \Sigma^0[\varphi](r) \,\rangle$$

which explains (1). However we prefer to interpret expression (1) mathematically as

$$\varphi(0) = \langle\, \delta(\underline{x}), \varphi(\underline{x}) \,\rangle = \langle\, \delta(r), \Sigma^0[\varphi](r) \,\rangle = \Sigma^0[\varphi](0) \tag{2}$$

Straightforward successive derivation with respect to r of (1) leads to

$$\partial_r^{2\ell}\, \delta(\underline{x}) = \frac{1}{(2\ell)!}(m)(m+1)\cdots(m+2\ell-1)\, \frac{1}{a_m}\, \frac{\delta^{(2\ell)}(r)}{r^{m-1}} \tag{3}$$

$$\partial_r^{2\ell+1}\, \delta(\underline{x}) = \frac{1}{(2\ell+1)!}(m)(m+1)\cdots(m+2\ell)\, \frac{1}{a_m}\, \frac{\delta^{(2\ell+1)}(r)}{r^{m-1}} \tag{4}$$

Expression (3) then is interpreted as

$$\langle\, \partial_r^{2\ell}\, \delta(\underline{x}), \varphi(\underline{x}) \,\rangle = \frac{1}{(2\ell)!}(m)(m+1)\cdots(m+2\ell-1)\, \langle\, \delta^{(2\ell)}(r), \Sigma^0[\varphi](r) \,\rangle$$

which is meaningful and can serve as the definition of the even order derivatives with respect to r of the delta distribution $\delta(\underline{x})$ in $\mathbb{R}^m$. However expression (4) makes no sense at all since

 ISBN: 978-84-617-8694-7

the spherical mean $\Sigma^0[\varphi](r)$ is an even function of r, whose odd order derivatives vanish at the origin:

$$\langle\ -\partial_r^{2\ell+1}\,\delta(r), \Sigma^0[\varphi](r)\ \rangle = \{\partial_r^{2\ell+1}\,\Sigma^0[\varphi](r)\}|_{r=0} = 0$$

How to explain the fact that, proceeding by stepwise derivation with respect to r, the even order derivatives of $\delta(\underline{x})$ apparently make sense, while its odd order derivatives are zero distributions? Let us to that end have a quick look at the functional analytic background of this phenomenon; for a more systematic treatment we refer to [2].

When expressing a scalar–valued test function $\varphi(\underline{x}) \in \mathcal{D}(\mathbb{R}^m)$ in spherical co–ordinates, one obtains a function $\widetilde{\varphi}(r,\underline{\omega}) = \varphi(r\underline{\omega}) \in \mathcal{D}(\mathbb{R}\times\mathbb{S}^{m-1})$, but it is clear that not all functions $\widetilde{\varphi}(r,\underline{\omega}) \in \mathcal{D}(\mathbb{R}\times\mathbb{S}^{m-1})$ stem from a test function in $\mathcal{D}(\mathbb{R}^m)$. However a one–to–one correspondence may be established between the usual space of test functions $\mathcal{D}(\mathbb{R}^m)$ and a specific subspace of $\mathcal{D}(\mathbb{R}\times\mathbb{S}^{m-1})$, which is the contents of the following lemma.

Lemma 1. *(see [4]) There is a one–to–one correspondence $\varphi(\underline{x}) \leftrightarrow \widetilde{\varphi}(r,\underline{\omega}) = \varphi(r\underline{\omega})$ between the spaces $\mathcal{D}(\mathbb{R}^m)$ and $\mathcal{V} = \{\phi(r,\underline{\omega}) \in \mathcal{D}(\mathbb{R}\times\mathbb{S}^{m-1}) : \phi$ is even, i.e. $\phi(-r,-\underline{\omega}) = \phi(r,\underline{\omega})$, and $\{\partial_r^n\,\phi(r,\underline{\omega})\}|_{r=0}$ is a homogeneous polynomial of degree n in $(\omega_1,\ldots,\omega_m), \forall n \in \mathbb{N}\}$.*

Clearly $\mathcal{V}$ is a closed (but not dense) subspace of $\mathcal{D}(\mathbb{R}\times\mathbb{S}^{m-1})$ and even of $\mathcal{D}_E(\mathbb{R}\times\mathbb{S}^{m-1})$, where the subscript E refers to the even character of the test functions in that space; this space $\mathcal{V}$ is endowed with the induced topology of $\mathcal{D}(\mathbb{R}\times\mathbb{S}^{m-1})$. The one–to–one correspondence between the spaces of test functions $\mathcal{D}(\mathbb{R}^m)$ and $\mathcal{V}$ translates into a one–to–one correspondence between the standard distributions $T \in \mathcal{D}'(\mathbb{R}^m)$ and the bounded linear functionals in $\mathcal{V}'$; this correspondence is given by

$$\langle\ T(\underline{x}), \varphi(\underline{x})\ \rangle = \langle\ \widetilde{T}(r,\underline{\omega}), \widetilde{\varphi}(r,\underline{\omega})\ \rangle$$

By Hahn–Banach's theorem the bounded linear functional $\widetilde{T}(r,\underline{\omega}) \in \mathcal{V}'$ may be extended to the distribution $\mathbb{T}(r,\underline{\omega}) \in \mathcal{D}'(\mathbb{R}\times\mathbb{S}^{m-1})$; such an extension is called a *spherical representation* of the distribution T (see e.g. [8]). However as the subspace $\mathcal{V}$ is not dense in $\mathcal{D}(\mathbb{R}\times\mathbb{S}^{m-1})$, the spherical representation of a distribution is *not unique*, but if $\mathbb{T}_1$ and $\mathbb{T}_2$ are two different spherical representations of the same distribution T, their restrictions to $\mathcal{V}$ coincide:

$$\langle\ \mathbb{T}_1(r,\underline{\omega}), \widetilde{\varphi}(r,\underline{\omega})\ \rangle = \langle\ \mathbb{T}_2(r,\underline{\omega}), \widetilde{\varphi}(r,\underline{\omega})\ \rangle = \langle\ \widetilde{T}(r,\underline{\omega}), \varphi(r\underline{\omega})\ \rangle = \langle\ T(\underline{x}), \varphi(\underline{x})\ \rangle$$

For test functions in $\mathcal{D}(\mathbb{R}\times\mathbb{S}^{m-1})$ the spherical variables r and $\underline{\omega}$ are ordinary variables, and thus smooth functions. It follows that for distributions in $\mathcal{D}'(\mathbb{R}\times\mathbb{S}^{m-1})$ multiplication by r and $\omega_j, j = 1,\ldots,m$, and differentiation with respect to r and $\omega_j, j = 1,\ldots,m$, are standard well–defined operations, whence

$$\langle\ \partial_r\,\mathbb{T}(r,\underline{\omega}), \Xi(r,\underline{\omega})\ \rangle = -\langle\ \mathbb{T}(r,\underline{\omega}), \partial_r\,\Xi(r,\underline{\omega})\ \rangle$$

 ISBN: 978-84-617-8694-7

for all test functions $\Xi(r,\underline{\omega}) \in \mathcal{D}(\mathbb{R}\times\mathbb{S}^{m-1})$, and similar expressions for $\partial_{\omega_j}\mathbb{T}$, $r\,\mathbb{T}$ and $\underline{\omega}\,\mathbb{T}$. However if $\mathbb{T}_1$ and $\mathbb{T}_2$ are two different spherical representations of the same distribution $T \in \mathcal{D}'(\mathbb{R}^m)$, then, upon restricting to test functions $\widetilde{\varphi}(r,\underline{\omega}) \in \mathcal{V}$, we are stuck with

$$-\langle\ \mathbb{T}_1(r,\underline{\omega}), \partial_r\,\widetilde{\varphi}(r,\underline{\omega})\ \rangle \neq -\langle\ \mathbb{T}_2(r,\underline{\omega}), \partial_r\,\widetilde{\varphi}(r,\underline{\omega})\ \rangle$$

because $\partial_r\,\widetilde{\varphi}(r,\underline{\omega})$ does no longer belong to $\mathcal{V}$ (and neither do $\partial_{\omega_j}\,\widetilde{\varphi}(r,\underline{\omega})$, $r\,\widetilde{\varphi}(r,\underline{\omega})$ and $\underline{\omega}\,\widetilde{\varphi}(r,\underline{\omega})$) since it is an odd function in the variables $(r,\underline{\omega})$. The conclusion is that the concept of spherical representation of a distribution does *not* allow for an unambiguous definition of the actions proposed. At the same time it becomes clear why even order derivatives with respect to r of the delta distribution and of a standard distribution in general are well-defined instead. Indeed, we have e.g.

$$\langle\ \partial_r^{2\ell}\,\mathbb{T}(r,\underline{\omega}), \Xi(r,\underline{\omega})\ \rangle = \langle\ \mathbb{T}(r,\underline{\omega}), \partial_r^{2\ell}\,\Xi(r,\underline{\omega})\ \rangle$$

where now $\partial_r^{2\ell}\,\Xi(r,\underline{\omega})$ does belong to $\mathcal{D}_E(\mathbb{R}\times\mathbb{S}^{m-1})$ which enables restriction to test functions in $\mathcal{V}$ in an unambiguous way.

2 Signumdistributions

As already remarked in the preceding section, $\underline{\omega}$ is an ordinary (vector) variable in $\mathbb{R}\times\mathbb{S}^{m-1}$, whence it makes sense to consider the following subspace of vector–valued test functions in $\mathbb{R}\times\mathbb{S}^{m-1}$:

$$\mathcal{W} = \underline{\omega}\,\mathcal{V} \subset \mathcal{D}_O(\mathbb{R}\times\mathbb{S}^{m-1};\mathbb{R}^m) \subset \mathcal{D}(\mathbb{R}\times\mathbb{S}^{m-1};\mathbb{R}^m)$$

where now the subscript O refers to the odd character of the test functions under consideration, i.e. $\psi(-r,-\underline{\omega}) = -\,\psi(r,\underline{\omega}), \forall\psi \in \mathcal{D}_O(\mathbb{R}\times\mathbb{S}^{m-1};\mathbb{R}^m)$. This space $\mathcal{W}$ is endowed with the induced topology of $\mathcal{D}(\mathbb{R}\times\mathbb{S}^{m-1};\mathbb{R}^m)$. By definition there is a one–to–one correspondence between the spaces $\mathcal{V}$ and $\mathcal{W}$.

From now on we will interpret vectors in $\mathbb{R}^m$ as Clifford 1–vectors in the Clifford algebra $\mathbb{R}_{0,m}$, where the basis vectors $(e_j, j=1,\ldots,m)$ of $\mathbb{R}^m$, satisfy the relations $e_j^2 = -1$, $e_i \wedge e_j = e_i e_j = -e_j e_i = -e_j \wedge e_i$, $e_i \cdot e_j = 0, i \neq j = 1,\ldots,m$. This allows for the use of the efficient *geometric* or *Clifford product* of Clifford vectors:

$$\underline{x}\,\underline{y} = \underline{x}\cdot\underline{y} + \underline{x}\wedge\underline{y}$$

for which, in particular,

$$\underline{x}\,\underline{x} = \underline{x}\cdot\underline{x} = -\,|\underline{x}|^2$$

$\underline{x}$ being the Clifford 1–vector $\underline{x} = \sum_{j=1}^m e_j\,x_j$, whence

$$\underline{\omega}\,\underline{\omega} = \underline{\omega}\cdot\underline{\omega} = -\,|\underline{\omega}|^2 = -1$$

 ISBN: 978-84-617-8694-7

For more on Clifford algebras we refer to e.g. [5].

For each $\mathbb{U}(r,\underline{\omega}) \in \mathcal{D}'(\mathbb{R}\times\mathbb{S}^{m-1};\mathbb{R}^m)$ we define $\widetilde{U}(r,\underline{\omega}) \in \mathcal{W}'$ by the restriction

$$\langle\, \widetilde{U}(r,\underline{\omega}), \underline{\omega}\,\widetilde{\varphi}(r,\underline{\omega})\,\rangle = \langle\, \mathbb{U}(r,\underline{\omega}), \underline{\omega}\,\widetilde{\varphi}(r,\underline{\omega})\,\rangle, \quad \forall\, \underline{\omega}\,\widetilde{\varphi}(r,\underline{\omega}) \in \mathcal{W}$$

In $\mathbb{R}^m$ we consider the space $\Omega(\mathbb{R}^m) = \{\underline{\omega}\,\varphi(\underline{x}) : \varphi(\underline{x}) \in \mathcal{D}(\mathbb{R}^m)\}$. Clearly the functions in $\Omega(\mathbb{R}^m)$ are no longer differentiable in the whole of $\mathbb{R}^m$, since they are not defined at the origin due to the function $\underline{\omega} = \frac{\underline{x}}{|\underline{x}|}$. By definition there is a one–to–one correspondence between the spaces $\mathcal{D}(\mathbb{R}^m)$ and $\Omega(\mathbb{R}^m)$.
For each $\widetilde{U}(r,\underline{\omega}) \in \mathcal{W}'$ we define $U(\underline{x})$ by

$$\langle\, U(\underline{x}), \underline{\omega}\,\varphi(\underline{x})\,\rangle = \langle\, \widetilde{U}(r,\underline{\omega}), \underline{\omega}\,\widetilde{\varphi}(r,\underline{\omega}))\,\rangle, \quad \forall\, \underline{\omega}\,\varphi(\underline{x}) \in \Omega(\mathbb{R}^m)$$

Clearly $U(\underline{x})$ is a bounded linear functional on $\Omega(\mathbb{R}^m)$, which, in [2], we called a *signumdistribution*.

Now start with a standard distribution $T(\underline{x}) \in \mathcal{D}'(\mathbb{R}^m)$ and let $\mathbb{T}(r,\underline{\omega}) \in \mathcal{D}'(\mathbb{R}\times\mathbb{S}^{m-1})$ be one of its spherical representations. Put $\mathbb{S}(r,\underline{\omega}) = \underline{\omega}\,\mathbb{T}(r,\underline{\omega})$ which in its turn leads to the signumdistribution $S(\underline{x}) \in \Omega(\mathbb{R}^m)$. Then we consecutively have

$$\begin{aligned}\langle\, S(\underline{x}), \underline{\omega}\,\varphi(\underline{x})\,\rangle = \langle\, \mathbb{S}(r,\underline{\omega}), \underline{\omega}\,\widetilde{\varphi}(r,\underline{\omega})\,\rangle &= \langle\, \underline{\omega}\,\mathbb{T}(r,\underline{\omega}), \underline{\omega}\,\widetilde{\varphi}(r,\underline{\omega})\,\rangle \\ &= -\langle\, \mathbb{T}(r,\underline{\omega}), \widetilde{\varphi}(r,\underline{\omega})\,\rangle = -\langle\, T(\underline{x}), \varphi(\underline{x})\,\rangle\end{aligned}$$

since $\underline{\omega}^2 = -1$, and we call $S(\underline{x})$ a signumdistribution associated to the distribution $T(\underline{x})$ and denote it by $T^{\vee}(\underline{x})$. It thus holds that for all test functions $\varphi \in \mathcal{D}(\mathbb{R}^m)$

$$\langle\, T^{\vee}(\underline{x}), \underline{\omega}\,\varphi(\underline{x})\,\rangle = -\langle\, T(\underline{x}), \varphi(\underline{x})\,\rangle$$

It should be emphasized that for a given distribution $T(\underline{x})$ the associated signumdistribution $T^{\vee}(\underline{x})$ is not uniquely defined but instead depends on the spherical representation of $T(\underline{x})$ chosen.
Conversely for a given signumdistribution $U \in \Omega(\mathbb{R}^m)$ we define the associated distribution $U^{\wedge}$ by

$$\langle\, U^{\wedge}(\underline{x}), \varphi(\underline{x})\,\rangle = -\langle\, U(\underline{x}), \underline{\omega}\,\varphi(\underline{x})\,\rangle$$

Clearly it holds that

$$T^{\vee\wedge} = T \qquad \text{and} \qquad U^{\wedge\vee} = U$$

As an example consider the distribution $T(\underline{x}) = \delta(\underline{x})$. Our aim is to define the signumdistribution $\delta^{\vee}(\underline{x})$. A spherical representation of the delta distribution is given by

$$\langle\, \mathbb{T}(r,\omega), \Xi(r,\underline{\omega})\,\rangle = \Sigma^0[\Xi(r,\underline{\omega})]\}|_{r=0}$$

 ISBN: 978-84-617-8694-7

Indeed, when restricting to the space $\mathcal{V}$ and taking into account property (2), we obtain

$$\langle\ \mathbb{T}(r,\omega),\widetilde{\varphi}(r,\underline{\omega})\ \rangle = \Sigma^0[\varphi(r\,\underline{\omega})]\}|_{r=0} = \langle\ \delta(\underline{x}),\varphi(\underline{x})\ \rangle$$

This particular spherical representation of $T(\underline{x})$ induces a signumdistribution associated to $\delta(\underline{x})$, which we define to be $\delta^{\vee}(\underline{x})$. It thus holds that for all test functions $\varphi \in \mathcal{D}(\mathbb{R}^m)$

$$\langle\ \delta^{\vee}(\underline{x}),\underline{\omega}\,\varphi(\underline{x})\ \rangle = -\langle\ \delta(\underline{x}),\varphi(\underline{x})\ \rangle \tag{5}$$

For further examples we refer to [2].

3 The Dirac operator in spherical co–ordinates

The Dirac operator $\underline{\partial} = \sum_{j=1}^{m} e_j\,\partial_{x_j}$, which may be seen as a Stein–Weiss projection of the gradient operator (see e.g. [7]) and which underlies the higher dimensional theory of monogenic functions (see e.g. [3]), linearizes the Laplace operator: $\underline{\partial}^2 = -\Delta$. Its action on a scalar–valued standard distribution $T(\underline{x})$ results into the vector–valued distribution $\underline{\partial}\,T(\underline{x})$ given by

$$\langle\,\underline{\partial}\,T(\underline{x})\,,\varphi(\underline{x})\,\rangle = \sum_{j=1}^{m} e_j\,\langle\,\partial_{x_j}\,T(\underline{x})\,,\varphi(\underline{x})\,\rangle = -\sum_{j=1}^{m} e_j\,\langle\,T(\underline{x})\,,\partial_{x_j}\,\varphi(\underline{x})\,\rangle = -\,\langle\,T(\underline{x})\,,\underline{\partial}\,\varphi(\underline{x})\,\rangle$$

which is a meaningful operation since only derivatives with respect to the cartesian co–ordinates $x_1,\ldots,x_m$ are involved.
Two fundamental formulae in monogenic function theory are

$$\{\underline{x},\underline{\partial}\} = \underline{x}\,\underline{\partial} + \underline{\partial}\,\underline{x} = -\,2\mathbb{E} - m \qquad \text{and} \qquad [\underline{x},\underline{\partial}] = \underline{x}\,\underline{\partial} - \underline{\partial}\,\underline{x} = m - 2\Gamma$$

where $\mathbb{E} = \sum_{j=1}^{m} x_j\,\partial_{x_j}$ is the scalar Euler operator, and $\Gamma = \sum_{j<k} e_j e_k(x_j\partial x_k - x_k\partial_{x_j})$ is the bivector angular momentum operator. It follows that

$$\underline{x}\cdot\underline{\partial} = -\mathbb{E} \qquad \text{and} \qquad \underline{x}\wedge\underline{\partial} = -\Gamma$$

Passing to spherical co–ordinates $\underline{x} = r\underline{\omega}$, $r = |\underline{x}|$, $\underline{\omega} = \sum_{j=1}^{m} e_j\,\omega_j \in \mathbb{S}^{m-1}$, the Dirac operator takes the form

$$\underline{\partial} = \underline{\partial}_{rad} + \underline{\partial}_{ang}$$

with

$$\underline{\partial}_{rad} = \underline{\omega}\,\partial_r \qquad \text{and} \qquad \underline{\partial}_{ang} = \frac{1}{r}\,\partial_{\underline{\omega}}$$

To give an idea what the angular differential operator $\partial_{\underline{\omega}} = \sum_{j=1}^{m} e_j\,\partial_{\omega_j}$ looks like, we mention here its explicit form in dimension $m=2:\partial_{\underline{\omega}} = e_\theta\,\partial_\theta$ and in dimension $m=3:\partial_{\underline{\omega}} =$

 ISBN: 978-84-617-8694-7

$e_\theta\,\partial_\theta + e_\varphi\,\frac{1}{\sin\theta}\,\partial_\varphi$, the meaning of the angular coordinates θ and φ being straightforward.

Taking into account that $\partial_{\underline{\omega}}$ is orthogonal to $\underline{\omega}$, the Euler operator in spherical co–ordinates then reads:

$$\mathbb{E} = -\,\underline{x}\cdot\underline{\partial} = -\,r\underline{\omega}\cdot\underline{\partial}_{rad} = -\,r\underline{\omega}\cdot\underline{\omega}\,\partial_r = r\,\partial_r$$

while the angular momentum operator Γ takes the form

$$\Gamma = -\,\underline{x}\wedge\underline{\partial} = -\,r\underline{\omega}\wedge\underline{\partial}_{ang} = -\,r\underline{\omega}\wedge\frac{1}{r}\,\partial_{\underline{\omega}} = -\,\underline{\omega}\wedge\partial_{\underline{\omega}} = -\underline{\omega}\,\partial_{\underline{\omega}}$$

The question now is how to define, if possible, the action of the $\underline{\partial}_{rad}$ and $\underline{\partial}_{ang}$ operators on a standard distribution. To that end both operators should be expressed in terms of cartesian derivatives, which we achieve by putting

$$\underline{\partial}_{rad} = \underline{\omega}\,\partial_r = -\,\frac{1}{\underline{x}}\,\mathbb{E} \qquad \text{and} \qquad \underline{\partial}_{ang} = \frac{1}{r}\,\partial_{\underline{\omega}} = -\,\frac{1}{\underline{x}}\,\Gamma$$

It becomes clear at once that in this way the actions of $\underline{\partial}_{rad}$ and $\underline{\partial}_{ang}$ on a standard distribution $T(\underline{x})$ are not unambiguously defined. Indeed, due to the division by the analytic function $\underline{x}$, both expressions

$$\underline{\partial}_{rad}\,T(\underline{x}) = \underline{\omega}\,\partial_r\,T(r\underline{\omega}) = -\,\frac{1}{\underline{x}}\,\mathbb{E}\,[T(\underline{x})] \tag{6}$$

and

$$\underline{\partial}_{ang}\,T(\underline{x}) = \frac{1}{r}\,\partial_{\underline{\omega}}\,T(r\underline{\omega}) = -\,\frac{1}{\underline{x}}\,\Gamma\,[T(\underline{x})] \tag{7}$$

represent equivalent classes of distributions each two of which differ by a vector multiple of the delta distribution $\delta(\underline{x})$. But if $S_1 = \underline{\partial}_{rad}\,T(\underline{x})$ and $S_2 = \underline{\partial}_{ang}\,T(\underline{x})$ are distributions arbitrarily chosen in the equivalent classes $\underline{\omega}\,\partial_r\,T(r\underline{\omega}) = -\,\frac{1}{\underline{x}}\,\mathbb{E}\,[T(\underline{x})]$ and $\frac{1}{r}\,\partial_{\underline{\omega}}\,T(r\underline{\omega}) = -\,\frac{1}{\underline{x}}\,\Gamma\,[T(\underline{x})]$ respectively, i.e.

$$\underline{x}\,S_1 = -\,\mathbb{E}\,[T(\underline{x})] \qquad \text{and} \qquad \underline{x}\,S_2 = -\,\Gamma[T(\underline{x})]$$

this choice is not completely arbitrary since S_1 and S_2 always must satisfy the relation

$$S_1 + S_2 = \underline{\partial}_{rad}\,T(\underline{x}) + \underline{\partial}_{ang}\,T(\underline{x}) = \underline{\partial}\,T(\underline{x}) \tag{8}$$

One could say that the differential operators $\underline{\partial}_{rad}$ and $\underline{\partial}_{ang}$ are *entangled* in the sense that the results of their actions on a distribution are always subject to (8).
Let us give an simple example to illustrate this phenomenon. Consider the regular distribution $T(\underline{x}) = \underline{x}$; then $\underline{\partial}\,\underline{x} = -m$, $\mathbb{E}\,\underline{x} = \underline{x}$ and $\Gamma\,\underline{x} = (m-1)\,\underline{x}$, whence

$$\underline{\omega}\,\partial_r\,\underline{x} = -1 + \underline{c}_1\,\delta(\underline{x}) \qquad \text{and} \qquad \frac{1}{r}\,\partial_{\underline{\omega}}\,\underline{x} = 1 - m + \underline{c}_2\,\delta(\underline{x})$$

 ISBN: 978-84-617-8694-7

with the restriction that the vector constants $\underline{c}_1$ and $\underline{c}_2$ always should satisfy $\underline{c}_1 + \underline{c}_2 = 0$.

Apparently there seems to be no possibility to unambiguously define the actions of the $\underline{\partial}_{rad}$ and $\underline{\partial}_{ang}$ operators on a standard distribution by singling out specific distributions in the equivalent classes (6) and (7), except for the following two special cases.
If the distribution $T(\underline{x})$ is *radial*, i.e. only depends on $r = |\underline{x}|$, then we put $\frac{1}{r}\,\partial_{\underline{\omega}}\,T = 0$ and $\underline{\omega}\,\partial_r\,T = \underline{\partial}\,T$, while if the distribution $T(\underline{x})$ is *angular*, i.e. only depends on $\underline{\omega} = \frac{\underline{\omega}}{|\underline{x}|}$, then we put $\underline{\omega}\,\partial_r\,T = 0$ and $\frac{1}{r}\,\partial_{\underline{\omega}}\,T = \underline{\partial}\,T$. The first special case is illustrated by the delta distribution (see also [2]): $\frac{1}{r}\,\partial_{\underline{\omega}}\,\delta(\underline{x}) = 0$ and $\underline{\omega}\,\partial_r\,\delta(\underline{x}) = \underline{\partial}\,\delta(\underline{x})$, the second one by the regular distribution $\underline{\omega}$: $\underline{\omega}\,\partial_r\,\underline{\omega} = 0$ and $\frac{1}{r}\,\partial_{\underline{\omega}}\,\underline{\omega} = \underline{\partial}\,\underline{\omega} = -(m-1)\frac{1}{r}$.
In the next section we will find two other possibilities for an unambiguous definition of the actions of the $\underline{\partial}_{rad}$ and $\underline{\partial}_{ang}$ operators.

4 Radial and angular derivatives of distributions

In Section 1 we explained why it is impossible to define the radial derivative $\partial_r\,T$ and the angular derivative $\partial_{\underline{\omega}}\,T$ of a distribution T within the class of distributions. Neither is it possible to multiply a distribution by the non–analytic functions r and $\underline{\omega}$. For *legitimizing* those *forbidden actions* we have to take the signumdistributions into consideration instead.

Definition 1. *The product of a scalar–valued distribution T by the function $\underline{\omega}$ is the signumdistribution $T^{\vee}$ associated to T, and it holds that*

$$\langle\ \underline{\omega}\,T\ ,\ \underline{\omega}\,\varphi\ \rangle = \langle\ T^{\vee}\ ,\ \underline{\omega}\,\varphi\ \rangle = -\langle\ T\ ,\varphi\ \rangle$$

Similarly, the product of a scalar–valued signumdistribution U by the function $\underline{\omega}$ is the distribution $-U^{\wedge}$ associated to $-U$, and it holds that

$$\langle\ \underline{\omega}\,U\ ,\ \varphi\ \rangle = \langle\ -U^{\wedge}\ ,\ \varphi\ \rangle = \langle\ U\ ,\underline{\omega}\,\varphi\ \rangle$$

Definition 2. *The product of a scalar–valued distribution T by the function r is the signumdistribution $r\,T$ given by*

$$\langle\ r\,T\ ,\ \underline{\omega}\,\varphi\ \rangle = \langle\ \underline{x}\,T\ ,\varphi\ \rangle = \langle\ T\ ,\ \underline{x}\,\varphi\ \rangle$$

The product of a scalar–valued signumdistribution U by the function r is the distribution $r\,U$ given by

$$\langle\ r\,U\ ,\ \varphi\ \rangle = \langle\ -\underline{x}\,U\ ,\ \underline{\omega}\,\varphi\ \rangle = \langle\ -U\ ,\ \underline{\omega}\,(\underline{x}\,\varphi)\ \rangle$$

Definition 3. *The derivative with respect to the radial distance r of a scalar–valued* radial *distribution T is the signumdistribution $\partial_r\,T$ given by*

$$\langle\ \partial_r\,T\ ,\ \underline{\omega}\,\varphi\ \rangle = \langle\ \underline{\omega}\partial_r\,T\ ,\ \varphi\ \rangle = \langle\ \underline{\partial}\,T\ ,\ \varphi\ \rangle$$

 ISBN: 978-84-617-8694-7

The derivative with respect to the radial distance r of a scalar–valued radial *signumdistribution U is the distribution $\partial_r U$ given by*

$$\langle\ \partial_r U\ ,\ \varphi\ \rangle = \langle\ -\underline{\omega}\partial_r U\ ,\ \underline{\omega}\,\varphi\ \rangle = \langle\ -\underline{\partial} U\ ,\ \underline{\omega}\,\varphi\ \rangle$$

Definition 4. *The angular $\partial_{\underline{\omega}}$–derivative of a scalar–valued distribution T is the signumdistribution $\partial_{\underline{\omega}} T$ given by*

$$\langle\ \underline{\omega}\,\varphi\ ,\ \partial_{\underline{\omega}} T\ \rangle = \langle\ \varphi\ ,\ \underline{\omega}\,\partial_{\underline{\omega}} T\ \rangle = \langle\ \varphi\ ,\ -\Gamma\, T\ \rangle$$

The angular $\partial_{\underline{\omega}}$–derivative of a scalar–valued signumdistribution U is the distribution $\partial_{\underline{\omega}} U$ given by

$$\langle\ \varphi\ ,\ \partial_{\underline{\omega}} U\ \rangle = \langle\ \underline{\omega}\,\varphi\ ,\ -\underline{\omega}\,\partial_{\underline{\omega}} U\ \rangle = \langle\ \underline{\omega}\,\varphi\ ,\ \Gamma\, U\ \rangle$$

Notice that only for the derivation with respect to r (Definition 3) is was necessary to confine the action to *radial* distributions. If instead general (signum)distributions are involved for this action, the result is an equivalence class of (signum)distributions seen the result obtained in the preceding section on the action of the differential operator $\underline{\omega}\partial_r$ on a distribution:

$$\langle\ \partial_r T\ ,\ \underline{\omega}\,\varphi\ \rangle = \langle\ \underline{\omega}\partial_r T\ ,\ \varphi\ \rangle = \langle\ S_1\ ,\ \varphi\ \rangle$$

with S_1 any distribution for which $\underline{x} S_1 = -\,\mathbb{E}T$, and a similar result for signumdistributions.

Now it is also possible to unambiguously define the radial derivative of the signumdistribution $T^\vee$ associated to the radial distribution T, $T^\vee$ certainly not being radial, according to the following commutative scheme

$$\begin{array}{ccc} T & \overset{-\underline{\omega}\partial_r = -\underline{\partial}}{\longrightarrow} & (\partial_r T)^\wedge \\ {\scriptstyle -\underline{\omega}}\uparrow\downarrow{\scriptstyle \underline{\omega}} & \nearrow\!\searrow^{-\partial_r}_{\partial_r} & \uparrow\downarrow \\ T^\vee & \underset{-\underline{\omega}\partial_r \neq -\underline{\partial}}{\longrightarrow} & \partial_r T \end{array}$$

Definition 5. *The radial derivative of the signumdistribution $T^\vee$ associated to the distribution T, is given by*

$$\partial_r\, T^\vee = -\,(\partial_r\, T)^\wedge$$

Example 1. Let us illustrate Definition 5 by the following simple example. Take the radial distribution T to be r. Its radial derivative is given by $\partial_r\, r = 1$ since

$$\langle\ \partial_r\, r\ ,\ \underline{\omega}\,\varphi\ \rangle = \langle\ \underline{\omega}\,\partial_r\, r\ ,\ \varphi\ \rangle = \langle\ \underline{\partial}\, r\ ,\ \varphi\ \rangle = \langle\ \underline{\omega}\ ,\ \varphi\ \rangle = \langle\ 1\ ,\ \underline{\omega}\,\varphi\ \rangle$$

 ISBN: 978-84-617-8694-7

It follows that $(\partial_r\, r)^\wedge = 1^\wedge = -\,\underline{\omega}$ and so $\partial_r\, T^\vee = \partial_r\, r^\vee = \partial_r\, r\underline{\omega} = \partial_r\, \underline{x} = -\,(\partial_r\, r)^\wedge = \underline{\omega}$, a result which indeed is quite acceptable!

Now notice that the above scheme implies the definition of the action of the operator $\underline{\omega}\partial_r$ on the signumdistribution $T^\vee$, the distribution T being radial:

$$(\underline{\omega}\,\partial_r)T^\vee = -\,\partial_r\, T = \underline{\omega}(-\,\underline{\omega}\,\partial_r)T = -\,\underline{\omega}\,\underline{\partial}\,T$$

and for all test functions $\underline{\omega}\,\varphi$ it holds that

$$\langle\, -\,\underline{\omega}\,\partial_r\, T^\vee\ ,\ \underline{\omega}\,\varphi\,\rangle = \langle\, (\underline{\omega}\,\partial_r\, T^\vee)^\wedge\ ,\ \varphi\,\rangle = \langle\, \partial_r\, T^\vee\ ,\ \varphi\,\rangle = \langle\, -\,(\partial_r\, T)^\wedge\ ,\ \varphi\,\rangle = \langle\, \partial_r\, T\ ,\ \underline{\omega}\,\varphi\,\rangle$$

In a similar way the definition of the radial derivative of the distribution $U^\wedge$ associated to the radial signumdistribution U, rests on the following commutative scheme

$$\begin{array}{ccc} & \overset{\underline{\omega}\partial_r \neq \underline{\partial}}{} & \\ U^\wedge & \longrightarrow & \partial_r U \\ {\scriptstyle -\underline{\omega}}\uparrow\downarrow{\scriptstyle \underline{\omega}} & \nearrow\searrow\ {\scriptstyle \partial_r,\ -\partial_r} & \uparrow\downarrow \\ U & \longrightarrow & (\partial_r U)^\vee \\ & {\scriptstyle \underline{\omega}\partial_r = \underline{\partial}} & \end{array}$$

Definition 6. *The radial derivative of the distribution $U^\wedge$ associated to the signumdistribution U, is given by*

$$\partial_r\, U^\wedge = -\,(\partial_r\, U)^\vee$$

Example 2. As an illustration of Definition 6 consider the radial signumdistribution $U = r$ for which $\partial_r\, U = \partial_r\, r = 1$, the associated signumdistribution of which is $(\partial_r\, U)^\vee = \underline{\omega}$. The associated distribution of $U = r$ is $U^\wedge = -\,\underline{x}$ and we obtain $\partial_r\, U^\wedge = \partial_r(-\underline{x}) = -\,(\partial_r\, U)^\vee = -\underline{\omega}$.

Notice that the above scheme implies the definition of the action of the operator $\underline{\omega}\partial_r$ on the distribution $U^\wedge$, the signumdistribution U being radial:

$$(\underline{\omega}\,\partial_r)U^\wedge = \partial_r\, U = -\,\underline{\omega}(\underline{\omega}\,\partial_r)U = -\,\underline{\omega}\,\underline{\partial}\,U$$

Finally note that the (signum)distributions appearing in the above examples, are particular cases of two special and useful classes of distributions which were thoroughly studied, see e.g. [1], in the framework of Clifford analysis.

 ISBN: 978-84-617-8694-7

5 Conclusion

In his famous and seminal book [6] Laurent Schwartz writes on page 45: *Using co-ordinate systems other than the cartesian ones should be done with the utmost care* [our translation]. And right he is! Indeed, just consider the delta distribution $\delta(\underline{x})$: it is pointly supported at the origin, it is rotation invariant: $\delta(A\,\underline{x}) = \delta(\underline{x})$, $\forall\, A \in \mathrm{SO}(m)$, it is even: $\delta(-\underline{x}) = \delta(\underline{x})$ and it is homogeneous of order $(-m)$: $\delta(a\underline{x}) = \frac{1}{|a|^m}\,\delta(\underline{x})$. So in a first, naive, approach, one could think of its radial derivative $\partial_r\,\delta(\underline{x})$ as a distribution which remains pointly supported at the origin, rotation invariant, even and homogeneous of degree $(-m-1)$. Temporarily leaving aside the even character, on the basis of the other cited characteristics the distribution $\partial_r\,\delta(\underline{x})$ should take the following form:

$$\partial_r\,\delta(\underline{x}) = c_0\,\partial_{x_1}\delta(\underline{x}) + \cdots + c_m\,\partial_{x_m}\delta(\underline{x})$$

and it becomes immediately clear that this approach to the radial derivation of the delta distribution is impossible since all distributions appearing in the sum at the right–hand side are odd and not rotation invariant, whereas $\partial_r\,\delta(\underline{x})$ is assumed to be even and rotation invariant. It could be that $\partial_r\,\delta(\underline{x})$ is either the zero distribution or is no longer pointly supported at the origin, but both those possibilities are unacceptable. So from the start we are warned by this example that introducing spherical co–ordinates $\underline{x} = r\underline{\omega}$, $r = |\underline{x}|$, $\underline{\omega} \in \mathbb{S}^{m-1}$ makes derivation of distributions in $\mathbb{R}^m$ a far from trivial action, as are, in principle "forbidden", actions such as multiplication by the non–analytic functions r and $\omega_j, j = 1, \ldots, m$. But there is more: functional analytic considerations on the space $\mathcal{D}(\mathbb{R}^m)$ of compactly supported smooth test functions expressed in spherical co–ordinates, forced us to introduce a new space of continuous linear functionals on a auxiliary space of test functions showing a singularity at the origin, for which, in [2], we coined the term *signumdistributions*, bearing in mind that $\underline{\omega} = \frac{\underline{x}}{|\underline{x}|}$ may be interpreted as the higher dimensional counterpart to the *signum* function on the real line. It turns out that the actions by r, $\underline{\omega}$, ∂_r and $\partial_{\underline{\omega}}$ map a distribution to a signumdistribution and vice versa. The basic idea behind the definition of these actions on a distribution $T \in \mathcal{D}'(\mathbb{R}^m)$, is to express the resulting signumdistributions as appropriate and "legal" actions on T. So, for example, we put $\langle rT, \underline{\omega}\varphi\rangle = \langle r\underline{\omega}T, \varphi\rangle = \langle \underline{x}T, \varphi\rangle, \forall\varphi \in \mathcal{D}(\mathbb{R}^m)$. This idea may seem to be rather simple, but it is backed up by the functional analytic considerations mentioned above, which, through the concept of *spherical representation of a distribution*, leave room for a choice, and this choice was made for efficiency's sake, as is amply demonstrated by the easy to handle calculus rules established in [2].

Of the four aforementioned actions only the radial derivative $\partial_r\,T$ escapes, in general, from an unambiguous definition, but leads to an equivalent class of signumdistributions instead. Still we were able to define unambiguously $\partial_r\,T$ in two particular cases: (i) when the given distribution T is radial, i.e. rotation invariant, and (ii) when $T = U^\wedge$ is the associated distribution to a given radial signumdistribution U, these two particular cases being quite

 ISBN: 978-84-617-8694-7

interesting since they correspond to two families of frequently used distributions, such as the fundamental solutions of the Laplace and the Dirac operator, in Clifford analysis.

References

[1] F. Brackx, B. De Knock, H. De Schepper, D. Eelbode, *A Calculus Scheme for Clifford Distributions*, Tokyo Journal of Math. **29(2)**(2006) 495–513.

[2] F. Brackx, F. Sommen, J. Vindas, *On the radial derivative of the delta distribution*, Complex Anal. Oper. Theory (2017) doi:10.1007/s11785-017-0638-8

[3] R. Delanghe, F. Sommen, V. Souček, *Clifford Algebra and Spinor–Valued Functions: A Function Theory for the Dirac Operator*, Kluwer Academic Publishers, Dordrecht, 1992.

[4] S. Helgason, *The Radon transform*, Birkhäuser, Boston, 1999.

[5] I. Porteous, *Clifford Algebras and the Classical groups*, Cambridge University Press, Cambridge, 1995.

[6] L. Schwartz, *Théorie des distributions*, Hermann, Paris, 1966.

[7] E. M. Stein, G. Weiss, *Generalization of the Cauchy-Riemann equations and representations of the rotation group*, Amer. J. Math. **90** (1968) 163–196.

[8] Đ. Vučković, J. Vindas, *Rotation invariant ultradistributions*, in:
Generalized Functions and Fourier Analyis, Operator Theory: Advances and Applications, Springer, Basel, 2017.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

A Cauchy integral formula for Hermitian and quaternionic monogenics

F. Brackx[1], H. De Schepper[1] and D. Eelbode[2]

[1] *Clifford Research Group, Department of Mathematical Analysis, Faculty of Engineering and Architecture, Ghent University, Krijgslaan 281 (S8), 9000 Ghent, Belgium*

[3] *Department of Mathematics and Informatics, University of Antwerp, Middelheimlaan 2, Antwerpen, Belgium*

emails: freddy.brackx@ugent.be, hennie.deschepper@ugent.be, david.eelbode@uantwerpen.be

Abstract

Similarly as hermitian Clifford analysis emerges in Euclidean space $\mathbb{R}^{2n}$ of even dimension as a refinement of euclidean Clifford analysis by the introduction of a complex structure on $\mathbb{R}^{2n}$, quaternionic Clifford analysis arises as a further refinement by the introduction of a so–called hypercomplex structure $\mathbb{Q}$, i.e. three complex structures ($\mathbb{I}$, $\mathbb{J}$, $\mathbb{K}$) which submit to the quaternionic multiplication rules, on Euclidean space $\mathbb{R}^{4p}$, the dimension now being a fourfold. In the hermitian framework, the fundamental symmetry group is $\mathrm{U}(n)$, which is isomorphic to the subgroup of $\mathrm{SO}(2n)$ of matrices which are commuting with the complex structure; in the quaternionic framework, this rôle is taken up by the symplectic group $\mathrm{Sp}(p)$, which is isomorphic with the subgroup of $\mathrm{SO}(4p)$ of matrices which are commuting with all three complex structures. Two, respectively four differential operators are constructed, which are invariant under the action of the corresponding symmetry group. Their simultaneous null solutions are respectively called hermitian and quaternionic monogenic functions. The basics of hermitian monogenicity have been studied in e.g. [2, 3, 9]. Quaternionic monogenicity has been developed in, amongst others, [13, 11, 10, 5, 6]. In this contribution, we compare the ways in which a Cauchy integral representation formula can be established in each of these frameworks.

Key words: Hermitian monogenic, quaternionic monogenic, Cauchy formula
MSC 2000: 30G35

ISBN: 978-84-617-8694-7

1 Hermitian and quaternionic monogenicity

Clifford analysis is centered around the notion of a monogenic function, a continuously differentiable function defined in an open region of Euclidean space $\mathbb{R}^m$, taking its values in the Clifford algebra $\mathbb{R}_{0,m}$, or subspaces thereof, and vanishing under the action of the Dirac operator $\underline{\partial} = \sum_{\alpha=1}^{m} e_\alpha \, \partial_{X_\alpha}$, which is the dual of the Clifford variable $\underline{X} = (X_1, \ldots, X_m)$. This notion is the higher dimensional counterpart of a holomorphic function in the complex plane. The Dirac operator factorizes the Laplacian: $\Delta_m = -\underline{\partial}^2$, and is invariant under the action of the Spin(m)–group which doubly covers the SO(m)–group, whence this framework is usually referred to as euclidean (or orthogonal) Clifford analysis.

Taking the dimension to be even: $m = 2n$, renaming the variables as $(X_1, \ldots, X_{2n}) = (x_1, y_1, x_2, y_2, \ldots, x_n, y_n)$ and considering the standard complex structure $\mathbb{I}_{2n}$, i.e. the complex linear real SO$(2n)$–matrix

$$\mathbb{I}_{2n} = \text{diag} \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix},$$

for which $\mathbb{I}_{2n}^2 = -E_{2n}$, where E_{2n} denotes the identity matrix, we define the rotated vector variable and the corresponding rotated Dirac operator

$$\begin{aligned} \underline{X}_{\mathbb{I}} &= \mathbb{I}_{2n}[\underline{X}] = \sum_{k=1}^{n} (-y_k e_{2k-1} + x_k e_{2k}) \\ \underline{\partial}_{\mathbb{I}} &= \mathbb{I}_{2n}[\underline{\partial}] = \sum_{k=1}^{n} (-\partial_{y_k} e_{2k-1} + \partial_{x_k} e_{2k}) \end{aligned}$$

A differentiable function F then is called hermitian monogenic in some region Ω of $\mathbb{R}^{2n}$, if and only if in that region F is a solution of the system $\underline{\partial} F = 0 = \underline{\partial}_{\mathbb{I}} F$. However, one can also introduce hermitian monogenics by means of the projection operators $\frac{1}{2}(\mathbf{1} \pm i\, \mathbb{I}_{2n})$, involving a complexification, where we consider the Witt basis vectors $\mathfrak{f}_k = -\frac{1}{2}(\mathbf{1} - i\, \mathbb{I}_{2n})[e_{2k-1}]$ and $\mathfrak{f}_k^\dagger = \frac{1}{2}(\mathbf{1} + i\, \mathbb{I}_{2n})[e_{2k-1}]$, $k = 1, \ldots, n$, as well as the vector variables

$$\underline{z} = -\frac{1}{2}(\mathbf{1} - i\, \mathbb{I}_{2n})[\underline{X}] = \sum_{k=1}^{n} (x_k + i y_k)\mathfrak{f}_k = \sum_{k=1}^{n} z_k \mathfrak{f}_k, \qquad \underline{z}^\dagger = \frac{1}{2}(\mathbf{1} + i\, \mathbb{I}_{2n})[\underline{X}] = \sum_{k=1}^{n} \overline{z}_k \mathfrak{f}_k^\dagger$$

and, correspondingly, the hermitian Dirac operators

$$2\, \partial_{\underline{z}}^\dagger = -\frac{1}{2}(\mathbf{1} - i\, \mathbb{I}_{2n})[\underline{\partial}] = 2 \sum_{k=1}^{n} \partial_{\overline{z}_k} \mathfrak{f}_k, \qquad 2\, \partial_{\underline{z}} = \frac{1}{2}(\mathbf{1} + i\, \mathbb{I}_{2n})[\underline{\partial}] = 2 \sum_{k=1}^{n} \partial_{z_k} \mathfrak{f}_k^\dagger$$

It follows that the hermitian monogenic system is equivalent to $\partial_{\underline{z}} F = 0 = \partial_{\underline{z}}^\dagger F$, which can be shown to be invariant under the action of the group U(n).

 ISBN: 978-84-617-8694-7

A refinement of hermitian Clifford analysis is obtained by considering the hypercomplex structure $\mathbb{Q} = (\mathbb{I}_{4p}, \mathbb{J}_{4p}, \mathbb{K}_{4p})$ on $\mathbb{R}^{4p} \simeq \mathbb{C}^{2p} \simeq \mathbb{H}^p$, where the dimension $m = 2n = 4p$ now is assumed to be a 4-fold. This hypercomplex structure arises by introducing, next to the complex structure $\mathbb{I}_{4p}$, a second one, $\mathbb{J}_{4p}$, given by

$$\mathbb{J}_{4p} = \text{diag} \begin{pmatrix} & & 1 & \\ & & & -1 \\ -1 & & & \\ & 1 & & \end{pmatrix}$$

Clearly $\mathbb{J}_{4p} \in \text{SO}(4p)$, with $\mathbb{J}_{4p}^2 = -E_{4p}$, and it anti–commutes with $\mathbb{I}_{4p}$. A third $\text{SO}(4p)$–matrix $\mathbb{K}_{4p} = \mathbb{I}_{4p}\mathbb{J}_{4p} = -\mathbb{J}_{4p}\mathbb{I}_{4p}$ then arises, for which $\mathbb{K}_{4p}^2 = -E_{4p}$ and which anti–commutes with both $\mathbb{I}_{4p}$ and $\mathbb{J}_{4p}$. We introduce the concept of quaternionic monogenicity by means of the additional rotated Dirac operators $\underline{\partial}_{\mathbb{J}} = \mathbb{J}_{4p}[\underline{\partial}]$ and $\underline{\partial}_{\mathbb{K}} = \mathbb{K}_{4p}[\underline{\partial}]$. A differentiable function $F : \mathbb{R}^{4p} \longrightarrow \mathbb{S}$ is called quaternionic monogenic in some region Ω of $\mathbb{R}^{4p}$, if and only if in that region F is a solution of the system $\underline{\partial}F = \underline{\partial}_{\mathbb{I}}F = \underline{\partial}_{\mathbb{J}}F = \underline{\partial}_{\mathbb{K}}F = 0$. Also here an alternative characterization is possible in terms of the hermitian Dirac operators, which in the actual dimension read:

$$\partial_{\underline{z}} = \sum_{j=1}^{p} (\partial_{z_{2j-1}} \mathfrak{f}_{2j-1}^{\dagger} + \partial_{z_{2j}} \mathfrak{f}_{2j}^{\dagger}), \qquad \partial_{\underline{z}}^{\dagger} = \sum_{j=1}^{p} (\partial_{\overline{z}_{2j-1}} \mathfrak{f}_{2j-1} + \partial_{\overline{z}_{2j}} \mathfrak{f}_{2j})$$

and their images under the action of $\mathbb{J}_{4p}$:

$$\partial_{\underline{z}}^{J} = \mathbb{J}_{4p}[\partial_{\underline{z}}] = \sum_{j=1}^{p} (\partial_{z_{2j}} \mathfrak{f}_{2j-1} - \partial_{z_{2j-1}} \mathfrak{f}_{2j}), \qquad \partial_{\underline{z}}^{\dagger J} = \mathbb{J}_{4p}[\partial_{\underline{z}}^{\dagger}] = \sum_{j=1}^{p} (\partial_{\overline{z}_{2j}} \mathfrak{f}_{2j-1}^{\dagger} - \partial_{\overline{z}_{2j-1}} \mathfrak{f}_{2j}^{\dagger})$$

The quaternionic system above then is equivalent to $\partial_{\underline{z}}F = \partial_{\underline{z}}^{\dagger}F = \partial_{\underline{z}}^{J}F = \partial_{\underline{z}}^{\dagger J}F = 0$, which can be shown to be invariant under the action of the symplectic group action of $\text{Sp}(p)$.

2 The Clifford Cauchy formula

The Cauchy integral formula is an important result for holomorphic functions in the complex plane; for a bounded domain D in $\mathbb{C}$ with (piecewise) smooth boundary ∂D it reads

$$f(z) = \frac{1}{2\pi i} \int_{\partial D} \frac{f(\xi)}{\xi - z}\, d\xi, \quad z \in \overset{\circ}{D}$$

It was generalized to monogenic functions in the euclidean Clifford framework in the following way:

$$f(\underline{X}) = \int_{\partial D} E(\underline{\Xi} - \underline{X})\, d\sigma_{\underline{\Xi}}\, f(\underline{\Xi}) \ , \quad \underline{X} \in \overset{\circ}{D}$$

 ISBN: 978-84-617-8694-7

where

$$E(\underline{X}) = \frac{1}{a_m}\,\frac{\overline{\underline{X}}}{|\underline{X}|^m}$$

is the fundamental solution of the Dirac operator, the so-called Cauchy kernel, with a_m denoting the area of the unit sphere S^{m-1} in $\mathbb{R}^m$, $\bar{\cdot}$ being the Clifford conjugation and $d\sigma_{\underline{\Xi}}$ being a Clifford algebra valued differential form of order $m-1$.

For hermitian monogenic functions arriving at a suitable Cauchy formula necessitated a formulation in terms of circulant matrices, see e.g. [4]. The fundamental solutions of the Dirac operators $\underline{\partial}$ and $\underline{\partial}_{\mathbb{I}}$, i.e. the euclidean Cauchy kernels, are respectively given by

$$E(\underline{X}) = \frac{1}{a_{2n}}\,\frac{\overline{\underline{X}}}{|\underline{X}|^{2n}}, \quad E_{\mathbb{I}}(\underline{X}) = \frac{1}{a_{2n}}\,\frac{\overline{\underline{X}}_{\mathbb{I}}}{|\underline{X}_{\mathbb{I}}|^{2n}}$$

where now a_{2n} denotes the area of the unit sphere S^{2n-1} in $\mathbb{R}^{2n}$; they give rise to their hermitian counterparts $\mathcal{E} = -\left(E + i\,E_{\mathbb{I}}\right)$ and $\mathcal{E}^{\dagger} = \left(E - i\,E_{\mathbb{I}}\right)$, or explicitly:

$$E(\underline{Z}) = \frac{2}{a_{2n}}\,\frac{\underline{Z}}{|\underline{Z}|^{2n}}, \quad E^{\dagger}(\underline{Z}) = \frac{2}{a_{2n}}\,\frac{\underline{Z}^{\dagger}}{|\underline{Z}|^{2n}}$$

Introducing the particular circulant (2×2) matrices

$$\boldsymbol{\mathcal{D}}_{(\underline{Z},\underline{Z}^{\dagger})} = \begin{pmatrix} \partial_{\underline{Z}} & \partial_{\underline{Z}^{\dagger}} \\ \partial_{\underline{Z}^{\dagger}} & \partial_{\underline{Z}} \end{pmatrix}, \quad \mathbf{E} = \begin{pmatrix} E & E^{\dagger} \\ E^{\dagger} & E \end{pmatrix}, \quad \text{and} \quad \boldsymbol{\delta} = \begin{pmatrix} \delta & 0 \\ 0 & \delta \end{pmatrix}$$

it was obtained that $\boldsymbol{\mathcal{D}}_{(\underline{Z},\underline{Z}^{\dagger})}\mathbf{E}(\underline{\mathbf{Z}}) = \boldsymbol{\delta}(\underline{\mathbf{Z}})$, whence the concept of a fundamental solution had to be reinterpreted in a matrix context. Consequently, this also turned out to be the case for hermitian monogenicity: a circulant matrix

$$\boldsymbol{G}_2^1 = \begin{pmatrix} g_1 & g_2 \\ g_2 & g_1 \end{pmatrix}$$

with continuously differentiable entries g_1 and g_2 defined in Ω and taking values in $\mathbb{C}_{2n}$ was then called $\boldsymbol{G}_2^1$ hermitian monogenic if and only if it satisfies the system

$$\boldsymbol{\mathcal{D}}_{(\underline{Z},\underline{Z}^{\dagger})}\boldsymbol{G}_2^1 = \boldsymbol{O}$$

where $\boldsymbol{O}$ denotes the matrix with zero entries. A formal Cauchy integral formula for hermitian monogenic circulant matrix functions then reads

$$\int_{\partial D} \mathbf{E}(\underline{Z} - \underline{V})\, \boldsymbol{d\Sigma}_{(\underline{Z},\underline{Z}^{\dagger})}\, \boldsymbol{G}_2^1(\underline{X}) = c_n \boldsymbol{G}_2^1(\underline{Y}), \quad \underline{Y} \in \overset{\circ}{D}$$

where $\boldsymbol{d\Sigma}_{(\underline{Z},\underline{Z}^{\dagger})}$ is a suitable differential form in circulant matrix format, giving rise to a normalization constant c_n at the right hand side, $\underline{Z}$ is the hermitian variable corresponding to $\underline{X}$ and $\underline{V}$ is the one corresponding to $\underline{Y}$. Taking for $\boldsymbol{G}_2^1$ a diagonal matrix, the above formula reduces to a genuine Cauchy formula for hermitian monogenic functions.

 ISBN: 978-84-617-8694-7

Remark 1. *It is well known in complex analysis that an alternative way of generalizing the Cauchy formula to higher dimension is by means of the Martinelli–Bochner kernel, see e.g. [12], which is not holomorphic but still harmonic, in this way establishing a connection between harmonic and holomorphic functions. The above hermitian Cauchy formula reduces to the Martinelli-Bochner formula when the considered functions take their values in a particular homogeneous subspace of complex spinor space, and thus establishes a connection between hermitian Clifford analysis and complex analysis in several variables.*

In the quaternionic context a formally identical formula was obtained in [1], involving the fundamental solutions of the Dirac operators $\underline{\partial}$, $\underline{\partial}_{\mathbb{I}}$, $\underline{\partial}_{\mathbb{J}}$ and $\underline{\partial}_{\mathbb{K}}$ and their quaternionic counterparts , as well as the quaternionic Dirac operators $\partial_{\underline{z}}$, $\partial_{\underline{z}}^{\dagger}$, $\partial_{\underline{z}}^{J}$ and $\partial_{\underline{z}}^{\dagger J}$ themselves, placed in their respective corresponding circulant 4×4 matrices.

3 Future work and ideas

As mentioned in the remark, interesting results were obtained in the hermitian framework by restricting the values of the considered functions to the different homogenous parts of spinor space, which are suggested by the $\mathrm{U}(n)$ symmetry. Thence for quaternionic monogenics the effect of a restriction to the $\mathrm{Sp}(p)$ invariant parts, the so-called symplectic cells arising in the quaternionic setting, will be investigated.

To this end we will need to include in our analysis two additional operators: a scalar Euler operator

$$\mathcal{E} = \sum_{k=1}^{p} z_{2k-1}\, \partial_{\overline{z}_{2k}} - z_{2k}\, \partial_{\overline{z}_{2k-1}}$$

and a multiplication operator $P = \mathfrak{f}_2\mathfrak{f}_1 + \mathfrak{f}_4\mathfrak{f}_3 + \ldots + \mathfrak{f}_{2p}\mathfrak{f}_{2p-1}$, leading to the definition of so-called $\mathfrak{osp}(4|2)$–monogenics, a explained in [7, 8].

References

[1] R. Abreu Blaya *et al.*, Cauchy integral formulae in quaternionic hermitean Clifford analysi, *Comp. Anal. Oper. Theory* **6(5)**, 2012, 971–985.

[2] F. Brackx *et al.*, Fundaments of Hermitean Clifford Analysis. Part I: Complex structure, *Compl. Anal. Oper. Theory* **1(3)**, 2007, 341–365.

[3] F. Brackx *et al.*, Fundaments of Hermitean Clifford Analysis. Part II: Splitting of h-monogenic equations, *Complex Var. Elliptic Eq.* **52(10-11)**, 2007, 1063–1079.

[4] F. Brackx *et al.*, On Cauchy and Martinelli-Bochner integral formulae in Hermitean Clifford analysis, *Bull. Braz. Math. Soc.* **40(3)**, 2009, 395–416.

 ISBN: 978-84-617-8694-7

[5] F. Brackx *et al.*, Fundaments of Quaternionic Clifford Analysis I: Quaternionic Structure, Adv. Appl. Clifford Alg. **24(4)** (2014), 955–980.

[6] F. Brackx *et al.*, Fundaments of Quaternionic Clifford Analysis III: Fischer Decomposition in Symplectic Harmonic Analysis, Ann. Glob. Anal. Geom. **46** (2014), 409–430.

[7] F. Brackx *et al.*, $\mathfrak{osp}$(4—2)monogenicity in Clifford analysis. In: Proceedings of the 15th International Conference on Computational and Mathematical Methods in Science and Engineering I, 2015, 240–243.

[8] F. Brackx *et al.*, Fischer decomposition for $\mathfrak{osp}$(4—2)-monogenics in quaternionic Clifford analysis, *Math. Meth. Appl. Sci.* **39(16)** (2016), 4874–4891.

[9] F. Brackx, H. De Schepper and F. Sommen, The Hermitian Clifford analysis toolbox, *Appl. Clifford Algebras* **18(3-4)**, 2008, 451–487.

[10] A. Damiano, D. Eelbode, I. Sabadini, *Quaternionic Hermitian spinor systems and compatibility conditions*, Adv. Geom. **11** (2011), 169–189.

[11] D. Eelbode, *Irreducible* $\mathfrak{sl}(m)$*–modules of Hermitean monogenics*, Complex Var. Elliptic Equ. **53 (10)** (2008), 975–987.

[12] A. Kytmanov, *The Bochner–Martinelli integral and its applications*, Birkhaüser, Basel–Boston–Berlin, 1995.

[13] D. Peña-Peña, I. Sabadini, F. Sommen, *Quaternionic Clifford analysis: the Hermitian setting*, Complex Anal. Oper. Theory **1** (2007), 97–113.

 ISBN: 978-84-617-8694-7

Volume II

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

Stability analysis of two-component incommensurate fractional-order systems and applications to the FitzHugh-Nagumo model

Oana Brandibur[1,2] and Eva Kaslik[1,2]

[1] *Institute e-Austria Timişoara, Romania*

[2] *Dept. of Mathematics and Computer Science, West University of Timişoara, Romania*

emails: oana.brandibur92@gmail.com, ekaslik@gmail.com

Abstract

For two-dimensional autonomous linear incommensurate fractional-order dynamical systems with Caputo derivatives of different orders, necessary and sufficient conditions are obtained for the asymptotic stability and instability of the null solution. These conditions are expressed in terms of the elements of the system's matrix, as well as of the fractional orders of the Caputo derivatives, leading to a generalization of the well known Routh-Hurwitz conditions. These theoretical results are then used to investigate the stability properties of a two-dimensional fractional-order FitzHugh-Nagumo neuronal model. The occurrence of Hopf bifurcations is also discussed. Numerical simulations are provided with the aim of exemplifying the theoretical results, revealing rich spiking behavior, in comparison with the classical integer-order FitzHugh-Nagumo model.

Key words: Caputo derivative; FitzHugh-Nagumo; mathematical model; fractional order derivative; stability; instability; bifurcation; numerical simulation.

1 Introduction

In many real world applications [5, 8, 10, 11, 15], fractional-order dynamical systems have proven to provide more accurate and realistic results than their classical integer-order counterparts, due to the fact that fractional-order derivatives are able to reflect memory and hereditary properties. However, it is important to emphasize that important qualitative differences may appear when generalizing properties of integer-order dynamical systems to the fractional-order case, and such generalizations have to be done with great care.

 ISBN: 978-84-617-8694-7

Stability analysis is one of the most important research topics of the qualitative theory of fractional-order systems. Two recent surveys [13, 19] provide comprehensive overviews of stability properties of fractional-order systems. In the particular case of linear autonomous commensurate fractional order systems, the most important starting point is Matignon's stability theorem [16], which has been generalized in [20]. Linearization theorems (or analogues of the classical Hartman-Grobman theorem) for fractional-order systems have been recently proved in [12, 21]. Up to this date, incommensurate order systems have not received as much attention as their commensurate order counterparts. Linear incommensurate fractional order systems with rational orders have been analyzed in [17]. Oscillations in two-dimensional incommensurate fractional order systems have been investigated in [6, 18].

The first aim of this paper is to explore necessary and sufficient conditions for the asymptotic stability of two-dimensional linear autonomous incommensurate fractional-order systems with Caputo derivatives of different orders. These results are later applied to investigate stability properties of a fractional-order FitzHugh-Nagumo neuronal model. It is worth emphasizing that fractional-order formulation of neuronal dynamics is strongly justified by experimental results concerning biological neurons [1, 14].

2 Stability results for linear systems with two Caputo derivatives of different orders

Consider the following two-dimensional linear autonomous incommensurate fractional-order system:

$$\begin{cases} {}^cD^{q_1}x(t) = a_{11}x(t) + a_{12}y(t) \\ {}^cD^{q_2}x(t) = a_{21}x(t) + a_{22}y(t) \end{cases} \tag{1}$$

where $A = (a_{ij})$ is a real 2-dimensional matrix and $q_1, q_2 \in (0, 1)$ are the fractional orders of the Caputo derivatives.

Applying the Laplace transform to system (1) we obtain the following system:

$$\begin{bmatrix} s^{q_1}X(s) - s^{q_1-1}x(0) \\ s^{q_2}Y(s) - s^{q_2-1}y(0) \end{bmatrix} = A \cdot \begin{bmatrix} X(s) \\ Y(s) \end{bmatrix},$$

where $X(s) = \mathscr{L}(x)(s)$ and $Y(s) = \mathscr{L}(y)(s)$ represent the Laplace transforms of the functions x and y, and s^{q_1}, s^{q_2} represent the principal values (first branches) of the corresponding complex power functions [7]. Therefore:

$$(\mathrm{diag}(s^{q_1}, s^{q_2}) - A) \cdot \begin{bmatrix} X(s) \\ Y(s) \end{bmatrix} = \begin{bmatrix} s^{q_1-1}x(0) \\ s^{q_2-1}y(0) \end{bmatrix}.$$

Next, we denote

$$\Delta_A(s) = \det\left(\mathrm{diag}(s^{q_1}, s^{q_2}) - A\right) = s^{q_1+q_2} - a_{11}s^{q_2} - a_{22}s^{q_1} + \det(A)$$

 ISBN: 978-84-617-8694-7

and we can easily observe that

$$X(s)=\frac{s^{q_1}(s^{q_2}-a_{22})x(0)+a_{12}s^{q_2}y(0)}{s\Delta_A(s)} \quad \text{and} \quad Y(s)=\frac{s^{q_2}(s^{q_1}-a_{11})y(0)+a_{21}s^{q_1}x(0)}{s\Delta_A(s)} \tag{2}$$

The following result provides necessary and sufficient conditions for the global asymptotic stability of system (1). The proof is based on the Final Value Theorem and asymptotic expansion properties of the Laplace transform [3, 4, 7].

Theorem 2.1.

1. *Denoting $q=\min\{q_1,q_2\}$, system (1) is $\mathcal{O}(t^{-q})$-globally asymptotically stable if and only if all the roots of $\Delta_A(s)$ are in the open left half-plane ($\Re(s)<0$).*
2. *If $\Delta_A(s)$ has a root in the open right half-place ($\Re(s)>0$), system (1) is unstable.*

With the aim of exploring the distribution of the roots of the characteristic function $\Delta_A(s)$ given above, the following result is given, which is a generalization of Proposition 2 from [4].

Proposition 2.2. *Consider the complex-valued function*

$$\Delta(s)=s^{q_1+q_2}+as^{q_2}+bs^{q_1}+c,$$

where $0<q_1\leq q_2<1$, s^{q_1} and s^{q_2} represent the principal values (first branches) of the corresponding complex power functions and $a,b,c\in\mathbb{R}$, $b>0$.

1. *If $c<0$, then $\Delta(s)$ has at least one positive real root.*
2. *$\Delta(0)=0$ if and only if $c=0$.*
3. *Assume that $c>0$.*
   (a) *If $a\geq 0$ then all roots of $\Delta(s)$ have the property that $\Re(s)<0$.*
   (b) *$\Delta(s)$ has a pair of pure imaginary roots if and only if*

$$a=a^{\star}(b,c,q_1,q_2)=-\frac{b^{\frac{q_1}{q_2}}}{\sin\frac{q_2\pi}{2}}\omega^{q_1-q_2}\left[\sin\frac{q_1\pi}{2}+\omega^{q_2}\sin\frac{(q_1+q_2)\pi}{2}\right]. \tag{3}$$

where $\omega=h_{q_1,q_2}^{-1}\left(\dfrac{\sin\frac{q_2\pi}{2}\cdot c}{b^{1+\frac{q_1}{q_2}}}\right)$ and h_{q_1,q_2} is the function defined by

$$h_{q_1,q_2}:\left(\left[\frac{\sin\frac{(q_2-q_1)\pi}{2}}{\sin\frac{q_1\pi}{2}}\right]^{\frac{1}{q_2}},\infty\right)\to(0,\infty)$$

$$h_{q_1,q_2}(\omega)=\omega^{q_1}\Big(\omega^{q_2}\sin\frac{q_1\pi}{2}-\sin\frac{(q_2-q_1)\pi}{2}\Big).$$

 ISBN: 978-84-617-8694-7

(c) If $s(a,b,c,q_1,q_2)$ is one of the roots of $\Delta(s)$ such that

$$\Re(s(a^\star,b,c,q_1,q_2))=0,$$

where $a^\star=a^\star(b,c,q_1,q_2)$ defined at (b), the following transversality condition is satisfied:

$$\frac{\partial\Re(s)}{\partial a}\Big|_{a=a^*}<0.$$

(d) All roots of $\Delta(s)$ are in the left half-plane if and only if $a>a^\star(b,c,q_1,q_2)$.

(e) $\Delta(s)$ has a pair of roots in the right half-plane if and only if $a<a^\star(b,c,q_1,q_2)$.

(f) For any $q_1,q_2\in(0,1)$, the following inequality holds:

$$a^\star(b,c,q_1,q_2)\leq -b^{\frac{q_1}{q_2}}.$$

Furthermore, sufficient stability conditions which do not depend on the fractional orders q_1 and q_2 will be obtained using the following:

Proposition 2.3. *Let $b>0$, $c>0$ and the complex-valued function $\Delta(s)$ defined in Proposition 2.2.*

1. *If $a>-\min\{b,1\}$ then all roots of $\Delta(s)$ are in the open left-half plane, regardless of q_1 and q_2.*
2. *Let $a\leq-\min\{b,1\}$. If $a+b+c+1\leq 0$ or $0<a+b+c+1<(\sqrt{c}-1)^2$ and $c>1$ then the equation $\Delta(s)$ has at least one positive real root, regardless of q_1 and q_2.*

Based on Theorem 2.1 and Propositions 2.2 and 2.3, the following conditions for the stability of system (1) are obtained, with respect to its coefficients and the fractional orders q_1,q_2:

Corrolary 2.4. *We consider the linear system (1) with $q_1,q_2\in(0,1)$ the fractional orders of the Caputo derivatives. If we denote $a=-a_{11}$, $b=-a_{22}$, $c=\det(A)$ and we assume that $b>0$, it results that:*

1. *If $c<0$, system (1) is unstable, regardless of the fractional orders q_1 and q_2.*
2. *We suppose that that $c>0$.*
   (a) System (1) is $\mathcal{O}(t^{-q})$-asymptotically stable if and only if $a>a^\star(b,c,q_1,q_2)$, where $a^\star(b,c,q_1,q_2)$ is defined by (3) and $q=\min\{q_1,q_2\}$.
   (b) If $a>-\min\{b,1\}$, system (1) is asymptotically stable, regardless of the fractional orders q_1 and q_2.
   (c) System (1) is unstable if $a<a^\star(b,c,q_1,q_2)$.
   (d) If $a\leq-\min\{b,1\}$ and $a+b+c+1\leq 0$ or $0<a+b+c+1<(\sqrt{c}-1)^2$ and $c>1$ then system (1) is unstable, regardless of the fractional orders q_1 and q_2.

 ISBN: 978-84-617-8694-7

3 Investigation of a fractional-order FitzHugh-Nagumo model

The FitzHugh-Nagumo neuronal model [9] is a simplification of the well-known Hodgkin-Huxley model, which describes a biological neuron's spiking behavior. In this paper, we consider an extension on the classical FitzHugh-Nagumo model, by replacing the integer-order derivatives by fractional-order Caputo derivatives:

$$\begin{cases} {}^cD^{q_1}v(t) = v - \dfrac{v^3}{3} - w + I \\ {}^cD^{q_2}w(t) = r(v + c - dw) \end{cases} \tag{4}$$

where v represents the membrane potential, w is a recovery variable, I is an external excitation current and $0 < q_1 \leq q_2 \leq 1$. A similar model has been investigated by means of numerical simulations in [2].

The second equation of system (4) can be rewritten as follows:

$${}^cD^{q_2}w(t) = rd\Big(\frac{1}{d}v + \frac{c}{d} - w\Big) = \phi(\alpha v + \beta - w)$$

where $\phi = rd \in (0,1)$, $\alpha = \dfrac{1}{d} > 1$ and $\beta = \dfrac{c}{d}$. Therefore, system (4) is equivalent to the following two-dimensional conductance-based model:

$$\begin{cases} {}^cD^{q_1}v(t) = I - I(v,w) \\ {}^cD^{q_2}w(t) = \phi(w_\infty(v) - w) \end{cases} \tag{5}$$

where $I(v,w) = w - v + \dfrac{v^3}{3}$ and $w_\infty(v) = \alpha v + \beta$ is a linear function.

The equilibrium states of the fractional-ordel neuronal model (5) are the solutions of the algebraic system

$$\begin{cases} I = I_\infty(v) \\ w = w_\infty(v) \end{cases}$$

where

$$I_\infty(v) = I(v, w_\infty(v)) = w_\infty(v) - v + \frac{v^3}{3} = (\alpha - 1)v + \frac{v^3}{3} + \beta.$$

We observe that $I_\infty \in C^1$, $\lim\limits_{v\to-\infty} I_\infty(v) = -\infty$ and $\lim\limits_{v\to\infty} I_\infty(v) = \infty$.

Moreover, $I'_\infty(v) = v^2 + \alpha - 1 > 0$, so the function I_∞ is increasing and, as it is also continuous, it results that I_∞ is bijective. Therefore, there exists an unique solution for the equation $I_\infty(v) = I$, which we denote by $v^* = v^*(I, \alpha, \beta)$.

For the investigation of the stability of the equilibrium states, we consider the Jacobian matrix associated to system (5) at an equilibrium state $(v^*, w^*) = (v^*, w_\infty(v^*))$:

$$J = \begin{bmatrix} 1 - (v^*)^2 & -1 \\ \phi \cdot \alpha & -\phi \end{bmatrix}$$

 ISBN: 978-84-617-8694-7

The characteristic equation at the equilibrium state (v^*, w^*) is

$$s^{q_1+q_2} + a(v^*)s^{q_2} + b(v^*)s^{q_1} + c(v^*) = 0 \tag{6}$$

where

$$\begin{aligned} a(v^*) &= -1 + (v^*)^2 \\ b(v^*) &= \phi > 0 \\ c(v^*) &= \det(J) = \phi \cdot I'_\infty(v^*). \end{aligned}$$

Based on Corrolary 2.4 2.b, we obtain:

Proposition 3.1. *Any equilibrium state (v^*, w^*) with $|v^*| > \sqrt{1-\phi}$ is asymptotically stable, regardless of the fractional orders q_1 and q_2.*

The stability of any equilibrium state (v^*, w^*) with $|v^*| \leq \sqrt{1-\phi}$ depends on the fractional orders q_1 and q_2 (see Figs. 1 and 2).

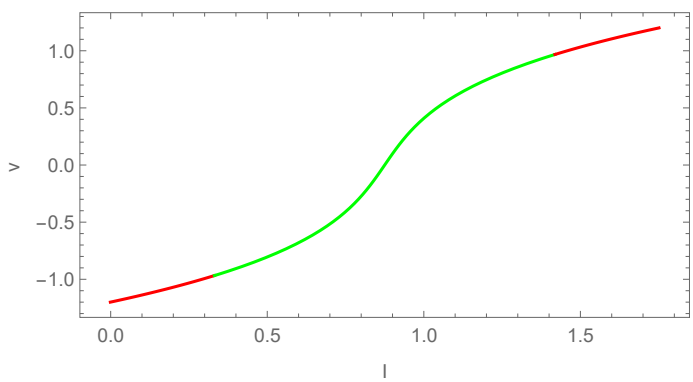


Figure 1: Membrane potential (v^*) of the equilibrium state (v^*, w^*) of system (4) (with parameter values: $r = 0.08$, $c = 0.7$, $d = 0.8$) with respect to the external excitation current I and their stability: red represents asymptotic stability, regardless of the fractional orders q_1 and q_2; green represents equilibrium states whose stability depends on the fractional orders q_1 and q_2.

Let us know consider an arbitrarily fixed equilibrium state (v^*, w^*) of system (4), such that $|v^*| \leq \sqrt{1-\phi}$. According to Proposition 2.2, at the critical values of the fractional orders $(q_1^\star, q_2^\star)$ defined implicitly by the equality

$$a(v^\star) = a^\star(b(v^\star), c(v^\star), q_1, q_2),$$

a Hopf bifurcation is expected to occur (see Fig. 2).

 ISBN: 978-84-617-8694-7

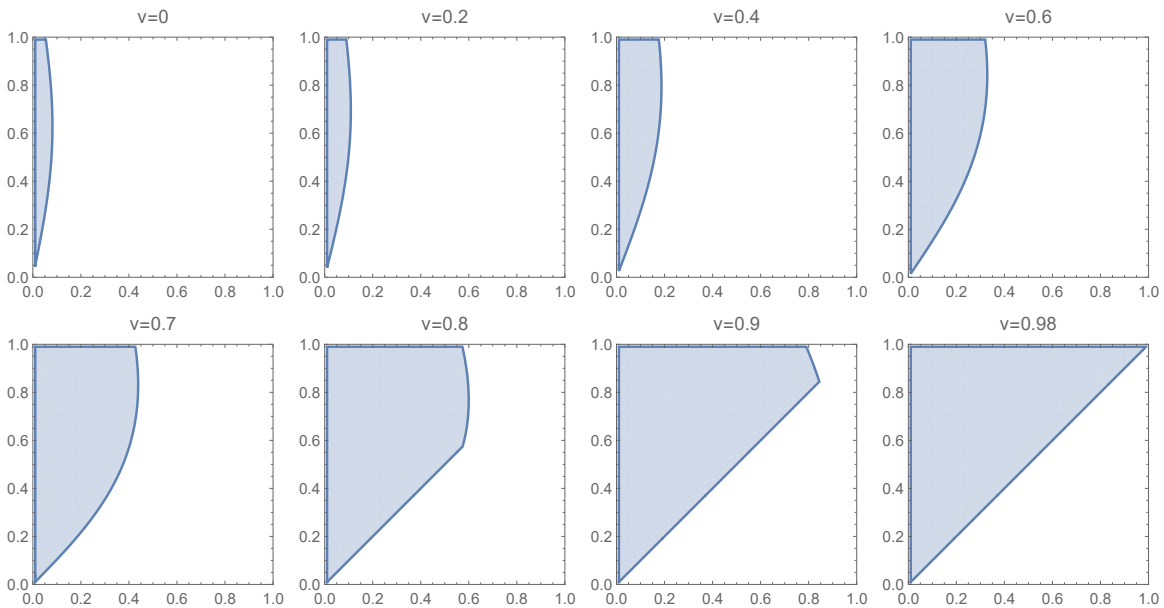


Figure 2: Stability regions (shaded) in the (q_1, q_2)-plane for equilibrium states (v^*, w^*) of system (4) (with parameter values: $r = 0.08$, $c = 0.7$, $d = 0.8$), with different values of the membrane potential v^* satisfying the inequality $|v^*| \leq \sqrt{1-\phi} \approx 0.98$. In each case, the part of the blue curve strictly above the first bisector represents the Hopf bifurcation curve in the (q_1, q_2)-plane.

Indeed, considering the following values for the system parameters: $r = 0.08$, $c = 0.7$, $d = 0.8$ and $I = 1.24567$, the equilibrium state is $(v^*, w^*) = (0.8, 1.875)$. In Fig. 3, the evolution of the state variables is shown, considering an initial condition in a small neighborhood of the equilibrium point. For a fixed value $q_2 = 0.8$, the critical value of the fractional order q_1 for which a Hopf bifurcation occurs is $q_1^* = 0.599$. Indeed, for $q_1 = 0.58$, asymptotically stable behavior is observed. For $q_1 = 0.63$, numerical simulations show quasi-periodic behavior, corresponding to the existence of a stable limit cycle. As q_1 is increased, the frequency of the oscillations increases. Numerical simulations suggest that fractional-order versions of the FitzHugh-Nagumo system provide a more realistic modelling of individual spikes than the corresponding integer-order counterpart (as seen in the last image from Fig. 3).

 ISBN: 978-84-617-8694-7

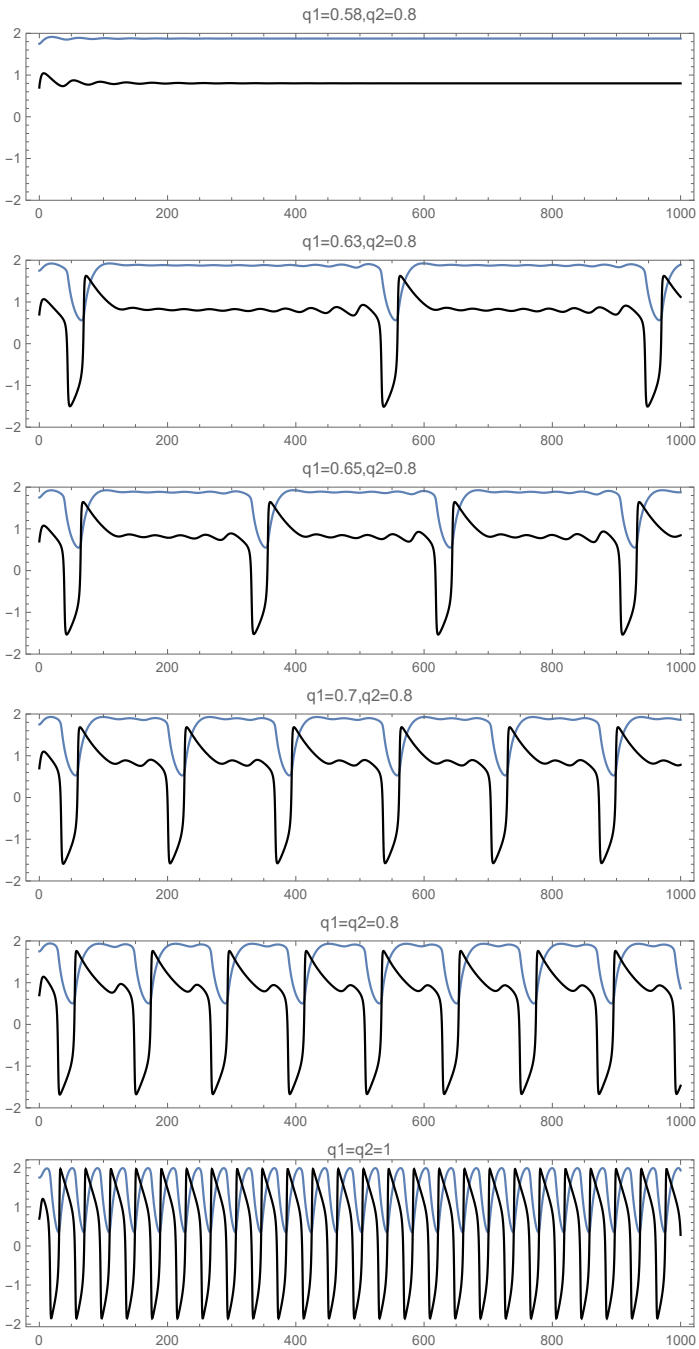


Figure 3: Evolution of the state variables of system (4) (with parameter values: $r = 0.08$, $c = 0.7$, $d = 0.8$ and $I = 1.24567$) for different values of the fractional orders.

 ISBN: 978-84-617-8694-7

4 Conclusions

Necessary and sufficient conditions have been obtained for the asymptotic stability of a two-dimensional incommensurate order linear autonomous system with Caputo derivatives of different fractional orders. These results can be regarded as a generalization of the classical Routh-Hurwitz stability conditions. As an application, the stability properties of a fractional-order FitzHugh-Nagumo system have been explored. Numerical simulations are provided to exemplify the theoretical findings, additionally revealing the occurrence of Hopf bifurcations when critical values of the fractional orders are encountered.

Acknowledgements

This work was supported by a grant of the Romanian National Authority for Scientific Research and Innovation, CNCS-UEFISCDI, project no. PN-II-RU-TE-2014-4-0270.

References

[1] T. Anastasio, *The fractional-order dynamics of brainstem vestibulo-oculomotor neurons*, Biological Cybernetics **72** (1994) 69–79.

[2] M. Armanyos and A. Radwan, *Fractional-order FitzHugh-Nagumo and Izhikevich neuron models*, in Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON), 2016 13th International Conference on, IEEE, (2016) 1–5.

[3] C. Bonnet and J. R. Partington, *Coprime factorizations and stability of fractional differential systems*, Systems & Control Letters **41** (2000) 167–174.

[4] O. Brandibur and E. Kaslik, *Stability properties of a two-dimensional system involving one Caputo derivative and applications to the investigation of a fractional-order Morris-Lecar neuronal model*, arXiv:1612.05389 (2016).

[5] G. Cottone, M. D. Paola, and R. Santoro, *A novel exact representation of stationary colored gaussian processes (fractional differential approach)*, Journal of Physics A: Mathematical and Theoretical **43** (2010) 085002.

[6] B. Datsko and Y. Luchko, *Complex oscillations and limit cycles in autonomous two-component incommensurate fractional dynamical systems*, Mathematica Balkanica, **26** (2012) 65–78.

[7] G. Doetsch, *Introduction to the Theory and Application of the Laplace Transformation*, Springer-Verlag Berlin Heidelberg, 1974.

 ISBN: 978-84-617-8694-7

[8] N. Engheia, *On the role of fractional calculus in electromagnetic theory*, IEEE Antennas and Propagation Magazine, **39** (1997) 35–46.

[9] R. FitzHugh, *Impulses and physiological states in theoretical models of nerve membrane*, Biophysical Journal **1** (1961) 445–466.

[10] B. Henry and S. Wearne, *Existence of Turing instabilities in a two-species fractional reaction-diffusion system*, SIAM Journal on Applied Mathematics **62** (2002) 870–887.

[11] N. Heymans and J.-C. Bauwens, *Fractal rheological models and fractional differential equations for viscoelastic behavior*, Rheologica Acta **33** (1994) 210–219.

[12] C. Li and Y. Ma, *Fractional dynamical system and its linearization theorem*, Nonlinear Dynamics **71** (2013) 621–633.

[13] C. Li and F. Zhang, *A survey on the stability of fractional differential equations*, The European Physical Journal - Special Topics **193** (2011) 27–47.

[14] B. Lundstrom, M. Higgs, W. Spain, and A. Fairhall, *Fractional differentiation by neocortical pyramidal neurons*, Nature Neuroscience **11** (2008) 1335–1342.

[15] F. Mainardi, *Fractional relaxation-oscillation and fractional phenomena*, Chaos Solitons Fractals **7** (1996) 1461–1477.

[16] D. Matignon, *Stability results for fractional differential equations with applications to control processing*, in Computational Engineering in Systems Applications (1996) 963–968.

[17] I. Petras, *Stability of fractional-order systems with rational orders*, arXiv preprint arXiv:0811.4102 (2008).

[18] A. G. Radwan, A. S. Elwakil, and A. M. Soliman, *Fractional-order sinusoidal oscillators: design procedure and practical examples*, IEEE Transactions on Circuits and Systems I: Regular Papers **55** (2008) 2051–2063.

[19] M. Rivero, S. V. Rogosin, J. A. Tenreiro Machado, and J. J. Trujillo, *Stability of fractional order systems*, Mathematical Problems in Engineering **2013** (2013) ID 356215.

[20] J. Sabatier and C. Farges, *On stability of commensurate fractional order systems*, International Journal of Bifurcation and Chaos **22** (2012) 1250084.

[21] Z. Wang, D. Yang, and H. Zhang, *Stability analysis on a class of nonlinear fractional-order systems*, Nonlinear Dynamics **86** (2016) 1023–1033.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

Attraction in network describing systems

Eduard Brokan[1] **and Felix Sadyrbaev**[2]

[1] *Daugavpils University, Latvia,*

[2] *Institute of Mathematics and Computer Science, University of Latvia, Latvia*

emails: Brokan@inbox.lv, felix@latnet.lv

Abstract

Description of attracting sets in specific dynamical systems arising in the gene regulatory theory is provided.

Key words: gene regulatory networks, dynamical systems, attracting sets
MSC 2000: 34B15, 34B23, 34C60, 34D45

1 Introduction

There are different kinds of models of bioregulatory networks. Among them one of widespread mathematical tools are nonlinear ordinary differential equations. Differential relations can be used to describe the regulatory interactions between genes. The time-dependent variables $x(t)$ represent the concentration of gene products mRNAs or protein. These variables are positive valued.

It was noticed by biologists that cells of living organisms are adaptable to unknown and unpredictable changes in environment even if these changes are very rapid. This mechanism was described, for instance, by [2]. It was proposed to use the attractor selection as principal mechanism of adaptation to unknown changes of biological systems [3].

The main idea of attractor selection is that the system is driven by two components, namely, deterministic and stochastic. Attractors are a part of the equilibrium points in the solution space. Conditions of such system are controlled by very simple feedback. When conditions of a system are suitable (close to one of the attractors), it is driven almost only by deterministic behavior, stochastic influence is very limited. When conditions of the systems are poor, deterministic behavior influence is close to zero and in this case system is driven by stochastic behavior. In this case the system randomly fluctuates searching for a

 ISBN: 978-84-617-8694-7

new attractor. When this attractor is found, deterministic behavior again dominates over stochastic [3].

If we use attractor selection mechanism for network resource management, at first we should define regulatory matrix W, which shows relationships between node pairs, that is, how each node pair affects each other including itself. As it was described in [2], three types of influence exist - activation, inhibition and no relation, corresponding to W_{ij} values of 1,-1 and 0.

2 System

Systems used for modeling gene regulatory networks are generally in the form

$$x_i' = f(\Sigma W_{ij}x_j - \Theta)v_g - x_i v_g + \eta,$$

where $f(z)$ is a continuous bounded monotonically increasing function (sigmoidal regulatory function) and matrix W_{ij} consists of entries describing the relation between nodes of the networks. There are various functions f possessing the desired properties. For instance, the function $f(z) = \frac{1}{1+e^{-\mu z}}$ meets the requirements. The argument z is $z = \Sigma W_{ij}x_j - \Theta$ and it represents the input on a gene with threshold Θ for increasing x_i. The parameter μ changes the slope of a function f and it indicates the gain parameter of the sigmoidal function. The term $-x_i v_g$ represents the rate of decrease in the expression level on gene i. The last term η stands for fluctuations and represents stochastic behaviour. The deterministic and stochastic behaviors are controlled by growth rate v_g, which represents the conditions of the metabolic reaction network. We consider this model under the simplifying conditions $\eta = 0$ and $v_g = 1$.

In extended form the system looks

$$\begin{cases} x_1' = f(W_{11}x_1 + \ldots + W_{1n}x_n - \Theta) - x_1, \\ x_2' = f(W_{21}x_1 + \ldots + W_{2n}x_n - \Theta) - x_2, \\ \ldots \quad \ldots \quad \ldots, \\ x_n' = f(W_{n1}x_1 + \ldots + W_{nn}x_n - \Theta) - x_n, \end{cases} \tag{1}$$

where W_{ij} are entries of the regulatory matrix W. We consider the specific case

$$W = \begin{vmatrix} 0 & 1 & \ldots & 1 \\ 1 & 0 & \ldots & 1 \\ \ldots & \ldots & \ldots & \ldots \\ 1 & 1 & \ldots & 0 \end{vmatrix}, \tag{2}$$

and, consequently,

$$\begin{cases} x_1' = f(x_2 + x_3 + \ldots + x_n - \Theta) - x_1, \\ x_2' = f(x_1 + x_3 + \ldots + x_n - \Theta) - x_2, \\ \ldots \quad \ldots \quad \ldots, \\ x_n' = f(x_1 + x_2 + \ldots + x_{n-1} - \Theta) - x_n. \end{cases} \tag{3}$$

ISBN: 978-84-617-8694-7

We study the system (3) following the standard scheme.

First, the critical points are to be found. The critical points are solutions of the system

$$\begin{cases} x_1 = f(x_2 + x_3 + \ldots + x_n - \Theta), \\ x_2 = f(x_1 + x_3 + \ldots + x_n - \Theta), \\ \ldots \quad \ldots \quad \ldots, \\ x_n = f(x_1 + x_2 + \ldots + x_{n-1} - \Theta). \end{cases} \tag{4}$$

Due to properties of a function f all critical points are located in a bounded n-dimensional cube. Moreover, all critical points are in the form $(x, \ldots, x)$, that is, $x_1 = \ldots = x_n$. This is easy to see, considering pairs x_1 and x_2, x_2 and x_3 and so on. Indeed, let a critical point be $x_1, \ldots, x_n$. If $x_2 > x_1$, then due to monotonicity of f one has that

$$x_1 = f(x_2 + x_3 + \ldots + x_n - \Theta) > f(x_1 + x_3 + \ldots + x_n - \Theta) = x_2.$$

Similarly, if $x_2 < x_1$, then $x_1 < x_2$. The contradictions prove that $x_1 = x_2$.

This is true also for regulatory matrices W with equal positive entries everywhere except the main diagonal. This is important in light of the remark from [4, p. 1823] that "An important step will be to move away from the binary idealization in order to analyze experimental data which is measured on a continuous scale."

Therefore the following is true for the case under consideration.

Proposition 2.1 *Any critical point is of the form $(x, \ldots, x)$ (n times), where x satisfies*

$$x = f((n-1)x - \Theta). \tag{5}$$

3 Linearization

Linearization of system (3) at a critical point $(x, \ldots, x)$ yields

$$\begin{cases} u_1' = -u_1 + a(u_2 + u_3 + \ldots + u_n), \\ u_2' = -u_2 + a(u_1 + u_3 + \ldots + u_n), \\ \ldots \quad \ldots \quad \ldots, \\ u_n' = -u_n + a(u_1 + u_3 + \ldots + u_{n-1}), \end{cases} \tag{6}$$

where $a = \frac{df(s)}{ds}|_{s=((n-1)x)}$. The characteristic equation for the linearization is

$$\det(A - \lambda I) = \begin{vmatrix} -1-\lambda & a & \ldots & a \\ a & -1-\lambda & \ldots & a \\ \ldots & \ldots & \ldots & \ldots \\ a & a & \ldots & -1-\lambda \end{vmatrix} = 0. \tag{7}$$

ISBN: 978-84-617-8694-7

It can be solved analytically and the roots are

$$\lambda_1 = \lambda_2 = \ldots = \lambda_{n-1} = -1 - a, \quad \lambda_n = -1 + (n-1)a. \tag{8}$$

Since $f(s)$ is monotonically increasing, a is positive. Therefore all characteristic values but one are negative (and equal). The sign of the last λ_n depends on the dimensionality n and on the value of a for a particular critical point.

We arrived therefore at the following result.

Theorem 3.1 *An attracting set for the system (3) is a complex of critical points of the form $(x, \ldots, x)$, where x solves (5). Any critical point is either fully attractive (all λ-s are negative) or semi-attractive (the last λ is positive). The case of $\lambda_n = 0$ is possible also and the respective critical point is attractive.*

Visualizations and examples for several choices of f are planned for full version of the article.

Acknowledgements

This work has been partially supported by Institute of Mathematics and Computer Science, University of Latvia.

References

[1] F.M. Alakwaa, *Modeling of Gene Regulatory Networks: A Literature Review*, Journal of Computational Systems Biology **1** (2002) 67-103.

[2] C. Furusawa, K. Kaneko, *A generic mechanism for adaptive growth rate regulation*, PLoS Computational Biology **4** (2008) e3.

[3] Y. Koizumi et al., *Adaptive Virtual Network Topology Control Based on Attractor Selection*, Journal of Lightwave Technology **28** (2010) 1720-1731, doi =10.1109/JLT.2010.2048412

[4] R. Somogyi, S. Fuhrman, M. Askenazi, A. Wuensche. *The gene expression matrix: towards the extraction of genetic network architectures*, Nonlinear Analysis: TMA **30** (1997), 1815-1824, doi: 10.1016/S0362-546X(97)00217-4

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

On the similarity solutions and conservation laws of the Cooper-Shepard-Sodano equation

M.S. Bruzón[1], A.P. Márquez[1] and R. de la Rosa [1]

[1] *Department of Mathematics, University of Cádiz*

emails: m.bruzon@uca.es, almudena.marquez@uca.es, rafael.delarosa@uca.es

Abstract

In this paper, we study the Cooper-Shepard-Sodano equation from the point of view of Lie symmetries. We perform an analysis of the symmetry reductions taking into account the similarity variables and the similarity solutions which allow us to transform our equation into ordinary differential equations. Furthermore, some conservation laws are derived by applying the multipliers method.

Key words: Conservation laws, Lie symmetries, Partial differential equations, Similarity solutions.

1 Introduction

In [7] the authors considered the Cooper-Shepard-Sodano equation (CSS) given by

$$u_t - c_0 u_x + \eta u_{xxxx} + \frac{1}{l-1}(u^{l-1})_x - \gamma p(u^{p-1}u_x^2)_x + \frac{2\gamma}{p+2}(u^{p+1})_{xxx} = 0, \tag{1}$$

where $u(x,t)$ is time evolved in the moving frame of reference with velocity c_0, l, $p \neq -2$, and γ are arbitrary constants whereas ηu_{xxxx} represents the presence of an artificial dissipation (hyperviscosity). The discussion on the necessity of adding the hyperviscosity term was given in [4]. The authors argued that this term is needed to damp out explicitly the numerical high-frequency dispersive errors introduced by the lack of smoothness at the edge of the discrete representation of the compacton.

In this paper, we consider a generalization of equation (1)

$$u_t - c_0 u_x + \eta u_{xxxx} + (f(u))_x - \gamma(p+1)(g(u)u_x^2)_x + \frac{2\gamma}{p+2}(g(u)u^2)_{xxx} = 0, \tag{2}$$

 ISBN: 978-84-617-8694-7

where $f(u)$ and $g(u)$ are arbitrary functions. We apply Lie theory to equation (2). Firstly, we perform an analysis of the Lie symmetries. From the infinitesimal generators obtained we determine the similarity variables and the similarity solutions which allow us to reduce our equation into ordinary differential equations. Furthermore, the direct method of the multipliers method is used to determine conservation laws.

2 Basic definitions

Let $x = (x^1, \ldots, x^n)$ be n independent variables, and $u = (u^1, \ldots, u^m)$ be m dependent variables with the partial derivatives $u_{(1)} = \{u_i^\alpha\}$, $u_{(2)} = \{u_{ij}^\alpha\}, \ldots$ of the first, second, etc. orders, where $u_i^\alpha = \partial u^\alpha / \partial x^i$, $u_{ij}^\alpha = \partial^2 u^\alpha / \partial x^i \partial x^j$. Let

$$E_\alpha \left(x, u, u_{(1)}, \ldots, u_{(k)}\right) = 0, \quad \alpha = 1, \ldots, m. \tag{3}$$

Denoting

$$D_i = \frac{\partial}{\partial x^i} + u_i^\alpha \frac{\partial}{\partial u^\alpha} + u_{ij}^\alpha \frac{\partial}{\partial u_j^\alpha} + \cdots \tag{4}$$

the total differentiation with respect to x^i, we have:

$$u_i^\alpha = D_i(u^\alpha), \qquad u_{ij}^\alpha = D_i(u_j^\alpha) = D_i D_j(u^\alpha), \quad \cdots$$

The variables u^α are also known as *differential variables*.

A function $f(x, u, u_{(1)}, \ldots)$ of a finite number of variables x, u, $u_{(1)}$, $u_{(2)}, \ldots$ is called a *differential function* if it is locally analytic. The set of all differential functions of all finite orders is denoted by $\mathcal{A}$.

Let ξ^i, $\eta^\alpha \in \mathcal{A}$ be differential functions depending on any finite number of variables x, u, $u_{(1)}$, $u_{(2)}, \ldots$. A first-order linear differential operator

$$X = \xi^i \frac{\partial}{\partial x^i} + \eta^\alpha \frac{\partial}{\partial u^\alpha} + \zeta_i^\alpha \frac{\partial}{\partial u_i^\alpha} + \zeta_{i_1 i_2}^\alpha \frac{\partial}{\partial u_{i_1 i_2}^\alpha} + \cdots, \tag{5}$$

where

$$\begin{aligned} \zeta_i^\alpha &= D_i(\eta^\alpha - \xi^j u_j^\alpha) + \xi^j u_{ij}^\alpha, \\ \zeta_{i_1 i_2}^\alpha &= D_{i_1} D_{i_2}(\eta^\alpha - \xi^j u_j^\alpha) + \xi^j u_{i_1 i_2}^\alpha, \ \ldots \end{aligned} \tag{6}$$

is called a *Lie-Bäcklund operator* [5, 6].

The vector $T = (T^1, T^2, \ldots, T^n)$ where $T^j \in \mathcal{A}$ and $j = 1, \ldots, n$ is a conserved vector if T^i satisfy

$$D_i T^i = 0. \tag{7}$$

Equation (7) is called conservation law.

 ISBN: 978-84-617-8694-7

From the multiplier method [1, 2, 3], every conservation law arises from multipliers $Q^\alpha(x,u,u_{(1)},\ldots)$ such that

$$Q^\alpha E_\alpha = D_i T^i, \tag{8}$$

all the multipliers can be determined solving the determining equation obtained from the variational derivative of

$$\frac{\delta}{\delta u^\beta}(Q^\alpha E_\alpha) = 0, \tag{9}$$

for arbitrary functions of $u(x^1, x^2, \ldots, x^n)$.

In this paper we prove that for $f(u) = u^l$ and $g(u) = u^p$, $l \neq 0$ and p arbitrary constants, equation (2) admits the multiplier

$$Q_1 = 1.$$

Moreover, if $f(u) = u$ and $g(u) = \frac{1}{u}$ besides Q_1 we obtain the following multipliers

$$Q_2 = c_0 t - t + x, \quad Q_3 = \frac{1}{2}\left(c_0 t - t + x\right)^2.$$

3 Lie point symmetries

A vector field

$$X = \xi(x,t,u)\frac{\partial}{\partial x} + \tau(x,t,u)\frac{\partial}{\partial t} + \phi(x,t,u)\frac{\partial}{\partial u}, \tag{10}$$

where $\xi(x,t,u)$, $\tau(x,t,u)$ and $\phi(x,t,u)$ are the infinitesimals, is a generator of a point symmetry of (2) if

$$X^{[4]}[u_t - c_0 u_x + \eta u_{xxxx} + (f(u))_x - \alpha(p+1)(g(u)u_x^2)_x + \frac{2\alpha}{p+1}(g(u)u^2)_{xxx}] = 0, \tag{11}$$

on (2). Here the operator $X^{[4]}$ is the fourth prolongation of the operator X. Requiring the vanishment of the coefficients of the derivatives of u, we obtain an overdetermined system of linear partial differential equations. Solving this system we deduce that when $f(u)$ and $g(u)$ are arbitrary functions and c_0, l, p, γ are arbitrary constants the infinitesimal generators are

$$X_1 = \partial_x, \qquad X_2 = \partial_t.$$

• If $f = a + bu + cu^{3a+4}$ and $g = bu^a$, where a, b, c are constants, the infinitesimal generators are X_1, X_2 and

$$X_3^1 = (3(b+c_0)t + x)\partial x + 4t\partial t - \left(\frac{u}{a+1}\right)\partial u.$$

 ISBN: 978-84-617-8694-7

• If $f = a + bu + cu\ln(u)$, and $g(u) = \frac{a}{u}$, where a, b, c are constants, the infinitesimal generators are X_1, X_2 and

$$X_3^2 = t\partial x + \left(\frac{u}{c}\right)\partial u.$$

From these generators we obtain the similarity solutions.

Acknowledgements

M.S. Bruzón and R. de la Rosa express their sincere thanks to the Plan Propio de Investigación de la Universidad de Cádiz and A.P. Márquez thanks the support of Vicerrectorado de Alumnos de la Universidad de Cádiz.

References

[1] S. C. Anco and G. W. Bluman, *Direct construction method for conservation laws of partial differential equations Part I: Examples of conservation law classifications*, Eur. J. Appl. Math. **13** (2002), 545–566.

[2] S. C. Anco and G. W. Bluman, *Direct construction method for conservation laws of partial differential equations Part II: General treatment*, Eur. J. Appl. Math. **13** (2002), 567–585.

[3] S. C. Anco, *Generalization of Noether's theorem in modern form to non-variational partial differential equations*, To appear in Fields Institute Communications: Recent progress and Modern Challenges in Applied Mathematics, Modeling and Computational Science.

[4] F. Cooper, J M. Hyman, A. Khare, *Compacton solutions in a class of generalized fifth-order Korteweg-de Vries equations* Phys. Rev. E **64** (2001) 026608.

[5] N. H. Ibragimov, *Transformation groups in mathematical physics. Nauka, Moscow, 1983. English transl. Transformation groups applied to mathematical physics* Riedel, Dordrecht 1985.

[6] N.H. Ibragimov, *Elementary Lie Group Analysis and Ordinary Differential Equations* John Wiley & Sons, Chichester 1999.

[7] B. Mihaila, A. Cardenas, F. Cooper, A. Saxena, *Stability and dynamical properties of Cooper-Shepard-Sodano compactons.*, Physic Rev E Stat Nonlin Soft Matter Phys **82** (2010) 066702.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

Mathematical Analysis of Flows on Contour Networks

Alexander P. Buslaev[1], Pavel A. Sokolov[1] and Marina V. Yashina[2]

[1] *Moscow State Automobile and Road Technical University (MADI), Leningradskii pr. 64, Moscow, Russia, 125319*

[2] *Moscow Technical University of Communications and Informatics (MTUCI), 8-a Aviamotornaya str., Moscow, Russia, 111124*

emails: apal2006@yandex.ru, user7824@gmail.com, yash-marina@yandex.ru

Abstract

Systems, containing a finite set of N cells (vertices) and particles with *deterministic* or *random* movement between the vertices, are the objects of study. This set of vertices is divided into *cycles* (*contours*), which are closed non self-intersecting sequence of *cells.* Contours have fixed direction of particles movement and may have common cells.

Key words: dynamical system, Contour network, Resolution rule of conflict, Self-organization, Collapse, Variation of function

1 Cellular automaton and BML model of traffic

(1.1) The Biham – Middleton - Levine model was published in 1992, [1]. It is a model of cellular automaton on Manhattan network on torus with two fixed types of particles. Particles move in one of two directions – up or right –, and with deterministic or random rules of priority in alternating times or in each step. A main result of this paper is to obtain the effect of self-organization of the system, when the conflicts disappear over some time, for any admissible initial position of particles. It was also obtained the fact that, at certain load of network, the system will stop, i.e. comes to the state of collapse. Thus it can be aware of the authors, *laid down the foundation of the spectral theory for considered systems.*

(1.2) In the next works there are considered analogical models in spaces with different dimensions, [2], [3]. But the results were using computer simulation methods. Exact results in these works are very small.

(1.3) In the original work of 1992 there was announced the exact result about the movement on an elementary contour that later became the object of study in [4].

 ISBN: 978-84-617-8694-7

2 Description of the systems

(2.1) $(k,1)-$ **networks** We introduce the concept of *system with k-structure (k-networks).* Considered networks consist of contours - closed sequences of cells. Standard characteristic of cell (vertices) V of network is *a degree* $deg(V)$, i.e. the number of contours, to which the considered vertex belongs. We consider networks such that the degree of each vertex is equal to 1 or k.

(2.2)Dynamical parameters of network

(2.2.1) Flow capacity of arc. Between each two neighbour contour cells more than one particle can not simultaneously move in both directions.

(2.2.2) The volume of the cell. In each cell there is simultaneously no more than one particle.

(2.3) Rules of particles movement

(2.3.1) Isolated movement. Initially each contour has a set of particles, S which can move only on this contour.

(2.3.2) Individual movement (I). We assume that the particle in the next moment can not go to the next cell, which is occupied at the moment.

(2.3.3)One-way movement (IOW). Each particle of this set S in one cycle makes an attempt to move forward by one cell, if it is impossible, it remains in it's cell.

(2.3.4)Shake-way motion (ISW). Each particle of this set S alternately in one cycle makes an attempt to move forward by one cell, if not possible, then back by one cell and, otherwise it remains in it's cell.

(2.3.5) Total connected one-way motion (TOW). Each cluster of particles of set S in one cycle alternately makes an attempt to move forward by one cell, if it is impossible, it remains in it's cell. If two sets of particles unite in a certain moment of time, subsequently will be considered as a single cluster.

(2.3.6) Control (Collective one-way (CIOW) and Collective shake-way (CISW)). Allocated area of flow selects behavior based on some criterion. For example, a particle moves or stays in accordance with the maximum of so-called *potential velocity.*

(2.4) The interaction of particles of different circuits (Competition and resolution rules). If particles of different contours are trying to move in the same common cell, there is *a conflict (competition)* of the particles, which is resolved in accordance with a given rule, this rule can be *deterministic* or *stochastic.*

(2.4.1) Deterministic local rule. Deterministic rule is, for example, *the priority rule* to resolve the conflict, in which competition always is won by *the particle with a higher priority* or *a particle in a cell with a higher priority.*

(2.4.2) Stochastic local rule. Stochastic rule is, for example, *a fair rule*, in which the particles involved in the competition are winning with equal probability.

(2.4.3) Collective control. The priority rule is determined by the objective function of a collective of particles on the network.

 ISBN: 978-84-617-8694-7

3 Flows on the contours, trajectory on $N-$ simplex and Rolle theorem

(3.1) Flows on the contours and trajectory on the simplex. Cells are located in the vertices of the canonical simplex

$$\left(0,0,\ldots,0,\underbrace{1,}_{i+1}0,\ldots,0\right), \quad i=0,\ldots,N-1 \tag{1}$$

in N-dimensional space, i.e. the location of the particles in the discrete moments of time. Each cell has no more than one particle. The unit of time a particle either remains in place or moves, which is determined by additional conditions.

There are M particles in the vertices of the canonical simplex. Their locations are given by N-vector, such that M coordinates of the vector are equal to 1, and the other is equal to zero. Thus, the flow of particles can be considered as a movement on the vertices of $M-$ simplex in $N-$ dimensional space. Each contour is associated with a cyclic vector of ones and zeros on the simplex $X(N,M)=\{x\}$, $X=(x_1,x_2,\cdots,x_N)$, where $x_i=0\vee 1$, $i=1,\cdots,N$ $\sum_{i=0}^{N}x_i=M$. Cyclicality means that the first and last coordinates are neighbors.

Cluster *is called any maximum connected sequence of ones.*

Variation *$var(X)$ of cyclic vector is called the double number of clusters.* **Each rule of movement gives a mapping** $A:X\to X$.

We assume, that *mapping A does not reduce the variation,* if

$$Var(A(x))\geq Var(x) \tag{2}$$

$\forall x\in X$. *Mapping A does not increase the variation,* if

$$Var(Ax)\leq Var(x) \tag{3}$$

$\forall x\in X$.

(3.2) Rolle theorem for IOW flows on contour.

Theorem 1.*For IOW flows on a contour, mapping A does not reduce variation, and, if there are two clusters with a distance of more than 1, then variation increases. The flow velocity reaches a stable state over a finite time, and coincides with the relative number of clusters, separated by single empty cells.*

(3.3) Rolle theorem for ISW flows on contour.

Theorem 2. *For ISW flows on a contour, mapping A does not reduce the variation, and if there are two clusters with a distance of more than 1, the variation increases. The flow velocity reaches a stable state over a finite time, and coincides with the relative number of clusters, separated by single empty cells. Other things being equal, the time of reaching steady state for ISW flows does not exceed the value of similar value for IOW flows.*

 ISBN: 978-84-617-8694-7

Remark 1. Nagel-Shrekenberg-Blank state function. In [1],[4] the formula was obtained for the dependence of the velocity of OW-movement of particles on a closed contour

$$v = \begin{cases} 1, \ 0 < r \leq \frac{1}{2}, \\ \frac{1}{r} - 1, \ \frac{1}{2} < r \leq 1. \end{cases}$$

In particular, it is proved, that if density r is less than $\frac{1}{2}$, then velocity of particles is equal to 1, i.e. particles move at each step.

Remark 2. The state function of stochastic IOW-flow on infinite contour. On infinite lattice the movements of particles into cell, which is next empty to the right, occur with probability p. If $N \to \infty$, $M \to \infty$ at constant density $r = M/N$, then we have in the limit of velocity, [6]

$$v = \frac{1 - \sqrt{1 - 4pr(1-r)}}{2r}.$$

4 Flows on contour networks and Rolle theorem

(4.1.) Flows on the two contours with one node (eight).

For this example it is clear, that Rolle theorem must be modified, because it becomes wrong in the above formulation. It is not our purpose here to formulate its analog in the most general terms, but we would like to show that the result tested for centuries can be "rejuvenated".

Let OW-priority rule takes place on one of the contours. Then, after a finite number of steps on the priority contour the flow becomes stationary, and corresponding mapping does not reduce the variation. As for the similar value for minor contour, it behaves much more difficult. Since the movement is locked through a common cell for both contours, as soon as the distance between consecutive clusters on the priority contour near the common cell is not greater than 1.

We introduce COW-movement on the minor contour. Then, during a time interval of length at least three steps, when particles on priority contour will not occupy a common cell, we can pass a particle on minor contour through the common cell.

If there exists a time interval of length at least five steps, when particles on priority contour will not occupy a common cell, then we can pass two particles on minor contour through the common cell.

Thus, control on priority contour is become urgent.

(4.2) Rolle theorem for CIOW flows on Eight-figure contour.

Let m be the number of particles on the priority loop length N, k be the number of particles on the minor contour with the same capacity. If we now introduce the collective control on the priority contour, then we will represent all distances between the particles as a linear combination of odd numbers.

 ISBN: 978-84-617-8694-7

For simplicity, let suppose that $m < N/2$. Define $n = N - m$. We express the number n as a sum of m natural numbers as following

$$n = n_1 + n_2 + \cdots + n_m. \tag{4}$$

Suppose that

$$K = \left[\frac{n_1 - 1}{2}\right] + \left[\frac{n_2 - 1}{2}\right] + \cdots + \left[\frac{n_m - 1}{2}\right]. \tag{5}$$

Denote by K^* the maximum possible value of the numbers K (3) among all possible representations of n in the form (4).

Theorem 3. *If $k < K^*$, then there exists a collective control, resulting in a finite number of steps to non conflict CIOW-flow with maximum velocity and maximum variation of the corresponding operator A.*

(4.3) Rolle theorem for SIOW - flows on linear contour graph.

The linear contour graph is a sequence of contours C_i, $i = 1, \cdots, M$ with the same length, such that neighbor contours C_i, C_{i+1} have a common node. Each IOW-flow on the contour C_i, such that has no conflicts with the contour C_{i-1}, defines demands to the IOW-flow on C_{i+1} with the same conditions. Thus, it is possible to obtain constructive conditions for existence of non conflict flow with a maximum variation on the linear sequence of contours.

5 Logistics of flows on networks

(5.1) Reduction of networks with physical measuring devices.

We suppose a graph is given with weights of edges corresponding to physical distances in the real network. Then we want to construct that the number of movement of particles corresponds to traveled distance, and particle velocity corresponds the number of movements per unit time. Therefore we add to each edge the number of vertices proportional to the edge weight, so that each edge of new graph corresponds the same physical distance. This new graph is an abstract graph with vertices corresponding of simplex with equal distances.

(5.2) Abstract model of traffic and scheduling.

In this section, we go from the description of existing flows to the problems of optimization of networks and flows.

(5.3) Logistic plan of particles

Each i-th particle, $1 \leq i \leq M$, has a location plan (movement) onto the vertices of the simplex. The plan is encoded as a sequence of "numbers" from 0 to $N - 1$ as the follow

$$\{b_1 b_2 b_3 \ldots\ldots b_T b_{T+1} \ldots\}.$$

 ISBN: 978-84-617-8694-7

It is a one-to-one mapping to a positional system with base N.

(5.4) Rules of logistic plan realization. If we assume in the original problem formulation that the canonical simplex is flow supporter, then solution of implementation plan problem for an individual particle can be considered obvious. If, however, flow supporter is an arbitrary graph G, then the use of the abstract scheme for graph G contains a one-to-one correspondence between the vertices, connectivity matrix (adjacency, incidence), and finally the weight (distance, price of movements) between adjacent vertices. Then a question of physical non-proportionality of particles moving to the neighbor vertex per the unit of standard time can be appear. In this case we add the necessary number of vertices to the original graph for correspondence between number of time steps and price (distance).

6 Computer exercises for IOW and SOW flows

We consider the contour of which is composed of N cells which is located M particles, where $1 \leq M \leq N$. Using computer simulation we have evaluated variation of mapping. A for IOW and SOW flow on contour by charging the value of density $r = \frac{M}{N}$ brow 0 to 1 with step $\Delta r = \frac{1}{N}$.

(6.1) IOW flows

With computer modeling was developed for IOW-flow on counter cells $N = 200$ during. Time interval $T = 1000$. It is shown in Fig.1,2 the dynamics of IOW-flow variation for $r = \frac{1}{2}$ and variation interval at changing of density r from 0 to 1.

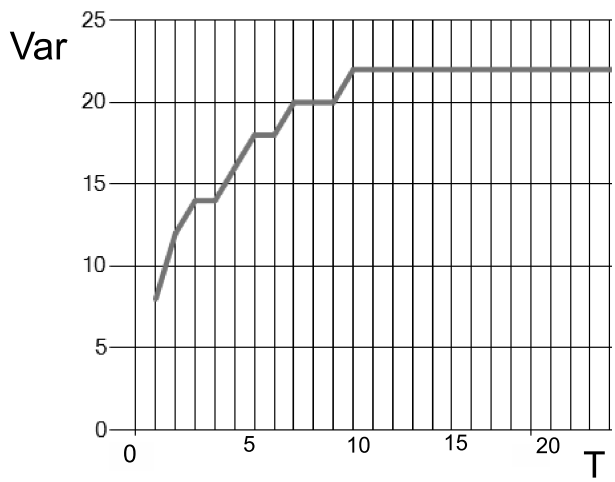


Figure 1: Dynamics of IOW-flow variation, $N = 2000$, $r = \frac{1}{2}$

(6.2) IOW flows

 ISBN: 978-84-617-8694-7

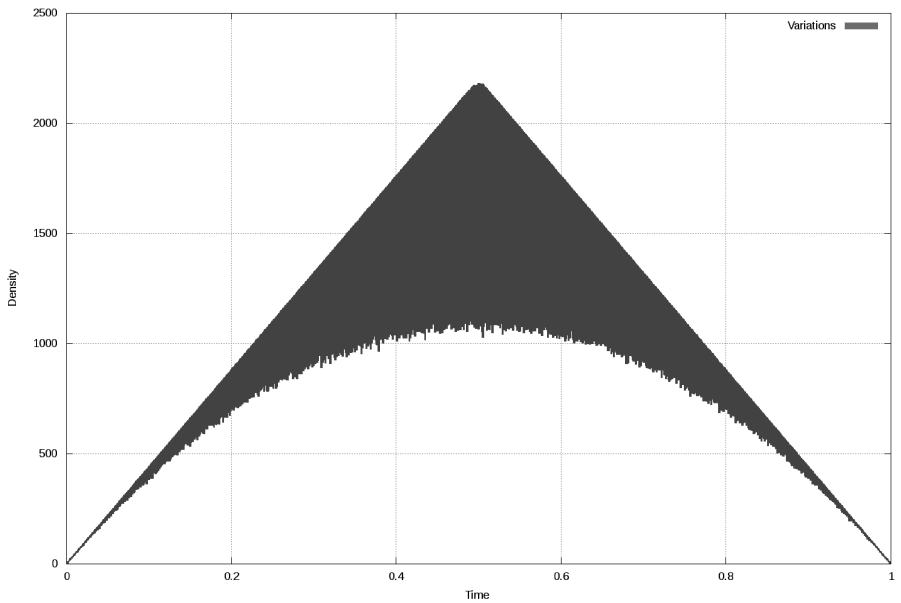


Figure 2: Dependence of IOW-flow variation interval on density r ($N = 2000$)

In Fig.3 it is shown the numerical results of variation interval for ISW-flows on contour.

(6.3) IOW flows at Eight-figure graph

Eight-figure graph consists of two contours with dimensions N_1 and N_2 (numbers of cells) correspondingly, which are connected by one common cell (node). The node connects two cells of each ring and prevents the passage of two particles at the same time. Resolution rule for competition is *stochastic local rule*. In Fig.4 it is shown the dynamics of total variation of the two contours, and variation on each contour with fixed parameters $N_1 = N_2 = 2200$, density $r = 0.5$, modeling time $T = 1000$.

In Fig.5 it is shown the total variation of two contours on dependence the density.

7 Comments. Three-pendulum, Cantor classes and Kolmogorov ideology

Let us assume that a system contains three cells V_0, V_1, V_2 and three particles P_0, P_1, P_2. At any given time there is a single particle in a cell. Particles moving realizes in accordance with their logistic plans given by the number belonging to the interval $(0, 1)$ and represented in the ternary calculus system. Plan of particle P_0 contains only numbers 0 and 1; plan of particle P_1 contains only numbers 1 and 2; plan of particle P_2 contains only numbers 0 and

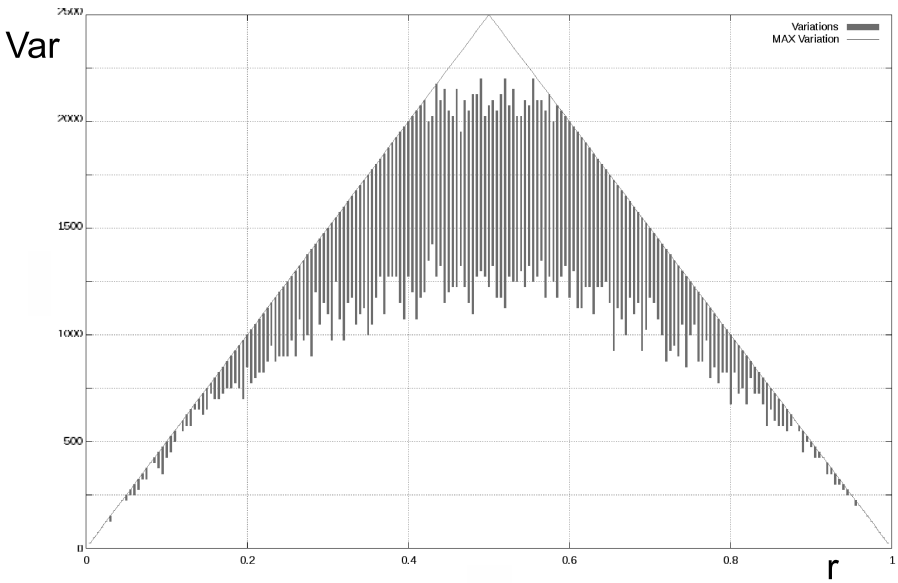


Figure 3: Dependence of SOW-flow variation interval on density r ($N = 2000$)

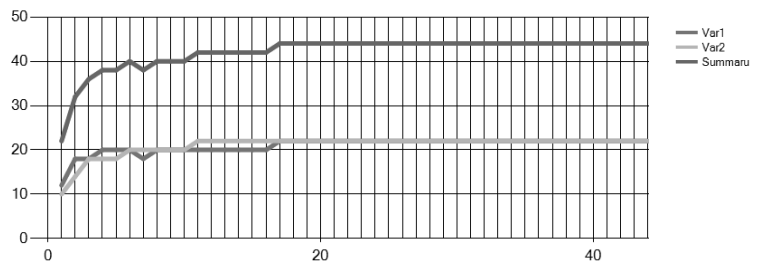


Figure 4: Eight-figure graph: dynamics of variations on each contour $N_1 = N_2 = 2200$ cells, $r = 0.5$, $T = 1000$

2. Thus, the set of admissible plans of each particle is *Cantor set.*

There are only two possible states of the system. In the first state, the particle P_i is located in cell V_i, $i = 0, 1, 2$. In the second state, the particle P_i is located in cell V_{i+1} (addition modulo 3). The system state change occurs, if each particle is scheduled to go to another cell. Kolmogorov ideology is in the follow. We calculate the considered characteristics for classes of plans, which we are interested in extreme values (extremum, "unfashionable" word). In this case, we are interested in the extreme values of coherence or conflict of plans on Cantor subsets of the interval. However, this problem is still open. Assume that each digit of plan takes any of two possible values with a probability $\frac{1}{2}$. Then,

 ISBN: 978-84-617-8694-7

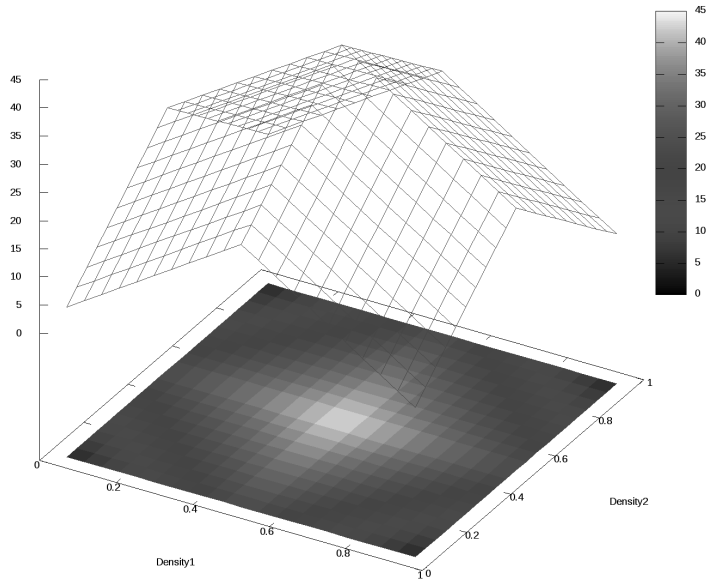


Figure 5: Variation in dependence on densities of both contours of Eight-figure graph

the probability of the event that all three particles have to move to another cell by the plan is equal to $\frac{1}{8}$. Thus, under these conditions the particle velocity is equal to $\frac{1}{8}$.

 ISBN: 978-84-617-8694-7

References

[1] C. W. Misner, K. S. Thorne and J. A. Wheeler, *Gravitation*, Freeman, San Francisco, 1970.

[2] E. Witten, *Supersymmetry and Morse theory*, J. Diff. Geom. **17** (1982) 661–692.

[3] Biham O., Middleton A., Levine D., *Self-organization and a dynamical transitions in traffic-flow models*, Physical Review A. vol. 46, N 10, (1992) R6124-R6127. DOI: 10.1103/PhysRevA.46.R6124

[4] Austin T.D., Benjamint I., *For what number of cars must self organization occur in the Biham –Middleton –Levine traffic model from any possible starting configuration.*

[5] Ding Z., Jiang R., Wang B., *Traffic flow in the Biham – Middleton – Levine model with random update rule.*, Physical Review (2011) 83(4). DOI: 10.1103/PhysRevE.83.04710

[6] Blank M. L., *Exact analysis of dynamical systems arising in models of traffic flow*, Russian Mathematical Surveys. (2000) 562-563.

[7] Buslaev A.P., Tatashev A.G., *Flows on Discrete Traffic Flower*, Journal of Mathematical Research. (2017) 98-108.

[8] Buslaev A.P., Tatashev A.G., *Particles flow on the regular polygon*, Journal of Concrete and Applicable Mathematics. (2011) 290-303.

[9] Buslaev A.P., Tatashev A.G., *Behavior of pendulums on a regular polygon*, Journal of Communication and Computer. (2014) 30-38.

[10] Kozlov V.V., Buslaev A.P., Tatashev A.G., Yashina M.V., *Dynamical systems on honeycombs*, Proceedings TGF' 13. (2015) 441-452.

[11] Kozlov V.V., Buslaev A.P., Tatashev A.G., *On real-valued oscillations of bipendulum*, Applied Mathematics Letters. (2015) 44-49.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

Distributed fusion filtering for multi-sensor systems with correlated random parameter matrices and noises

R. Caballero-Águila[1], I. García-Garrido[1] and J. Linares-Pérez[2]

[1] *Departamento de Estadística e I.O., Universidad de Jaén*

[2] *Departamento de Estadística e I.O., Universidad de Granada*

emails: raguila@ujaen.es, iggarrid@ujaen.es, jlinares@ugr.es

Abstract

This paper addresses the distributed fusion estimation problem for discrete-time linear stochastic systems with multi-sensor measurements including random parameter matrices. It is assumed that the random parameter matrices in the observation equations are one-step autocorrelated and cross-correlated between the different sensors and the additive noises are also correlated. Under these assumptions, a recursive algorithm is proposed to obtain local least squares linear filters based on the measurements of each sensor, and the distributed fusion filter is designed as the matrix-weighted linear combination of these estimators which minimizes the mean squared estimation error. This research is illustrated by a numerical simulation example where a multi-sensor system with randomly delayed measurements is considered and the performance of the proposed estimators is analyzed by comparing the estimation error variances of the distributed and centralized fusion filters.

Key words: distributed fusion filter, correlated random parameter matrices, randomly delayed measurements

MSC 2000: 60G35, 62M20, 93E11

1 Introduction

In recent years, the use of sensor networks has received significant attention in many practical domains, since they usually provide more information than traditional single-sensor communication systems. For this reason, the fusion estimation problem in sensor network stochastic systems has been widely studied in many fields of science, technology and military, such as navigation and detection.

 ISBN: 978-84-617-8694-7

Although there are several information fusion techniques, the most common fusion estimation approaches are the centralized and distributed ones. The former is based on the measurements from all the sensors, which are sent to a fusion centre, and so, it provides the optimal estimator when all the sensors work accurately. However, a sensor error can spoil the performance of the centralized filter and may give rise to heavy computational burden and poor reliability. In the distributed fusion estimation approach, each single sensor sends a local estimator to the fusion centre, where the state is estimated by a combination of all the received local filters using a certain optimality criterion. Thus, the distributed approach has lower estimation accuracy, which is compensated with considerable advantages, such as greater robustness and reliability. Therefore, the distributed fusion method is usually more attractive and has become an interesting research topic (see e.g. [1]-[4] and references therein).

In general, there are many situations with network-induced phenomena, such as multiplicative noise uncertainties, random delays, packet dropouts and missing measurements, in which the state estimation problem can be addressed by transforming the original system into one with random parameter matrices. For example, in [1] and [5] systems with packet dropouts and/or random delays are transformed into systems with random parameter matrices. Also, systems with multiplicative noises in the state and observation equations as those investigated in [3] are special cases of this kind of systems. The optimal filtering problem is addressed for a class of discrete-time stochastic systems with multiplicative noises and random sensor delays in [6] and, later, also with missing measurements in [7], by transforming the original system into one with random parameter matrices.

Accordingly, the study of the estimation problem in systems with random parameter matrices has become an active research field. In [8] a distributed Kalman filtering fusion is proposed for systems with independent random state transition and measurement matrices and white noises. For a class of discrete-time multisensor stochastic systems also with independent random parameter matrices but autocorrelated and cross-correlated noises, the centralized fusion estimation problem has been addressed under the phenomena of fading measurements in [9]. Moreover, by considering this same correlation assumption of the noise processes and one-step correlated and cross-correlated random measurement matrices, the centralized fusion linear filter is obtained in [10], where the results are applied to systems with missing measurements and randomly delayed observations. The distributed fusion estimation problem has also been studied for sensor network systems with independent random parameter matrices and correlated noises in [11]. Centralized and distributed fusion estimation problems are both studied in [12] for networked systems with random parameter matrices, from measurements subject to random delays and packet dropouts during the transmission, using only covariance information.

Motivated by the above discussion, this paper deals with the distributed fusion estimation problem for discrete-time linear stochastic systems with multi-sensor measurements

 ISBN: 978-84-617-8694-7

including correlated and cross-correlated random parameter matrices and noises. On the one hand, it complements [10], where the centralized fusion estimation problem is analyzed for an analogous class of discrete-time stochastic systems with the same correlation assumptions. In this way, the results obtained are applied to multi-sensor systems with correlated randomly delayed measurements as a particular case. On the other, it is different from [11], where the distributed fusion filtering estimators are obtained for networked systems with independent random state transition and measurement matrices, whereas correlation of the random parameter matrices in the observation equation is considered in the current paper.

The rest of the paper is organized as follows. Section 2 describes the system model with correlated and cross-correlated random parameter matrices and noises, specifying the assumptions under which the distributed fusion estimation problem is addressed. In Section 3, the local least squares linear filtering algorithm is given, and the proposed distributed fusion filter is designed by a matrix-weighted linear combination of the local filters using the mean squared error optimality criterion. An illustrative simulation example is given in Section 4 to show the applicability of the proposed filtering algorithm. Finally, some conclusions are drawn in Section 5.

2 System model

Let us consider the following discrete-time linear stochastic system with measurements coming from m different sensors:

$$\begin{aligned} &x_{k+1} = F_k x_k + w_k, \quad k \geq 0, \\ &y_k^{(i)} = H_k^{(i)} x_k + B_k^{(i)} v_k^{(i)}, \quad k \geq 1, \quad i = 1, \ldots, m, \end{aligned} \tag{1}$$

where $x_k \in \mathbb{R}^{n_x}$ is the state vector to be estimated and $y_k^{(i)} \in \mathbb{R}^{n_y}$, $i = 1, \ldots, m$, is the output measurement of the $i-$th sensor at the sampling time k. $w_k \in \mathbb{R}^{n_x}$ and $v_k^{(i)} \in \mathbb{R}^{n_v}$ are the process and measurement noise vectors, respectively. F_k is the state transition matrix and $H_k^{(i)}$ and $B_k^{(i)}$ are the measurement matrices, all of them with random parameters and suitable dimensions.

The following assumptions about the initial state, the random parameter matrices and the noises involved in system (1) are required.

Assumption 1. The initial state x_0 is a zero-mean random vector with $Cov[x_0] = \Sigma_0$. Also, it is assumed to be independent of the random parameter matrices and noise processes.

Assumption 2. $\{F_k;\, k \geq 0\}$, $\{H_k^{(i)};\, k \geq 1\}$ and $\{B_k^{(i)};\, k \geq 1\}$ have known means, which will be denoted $\overline{F}_k \equiv E[F_k]$, $\overline{H}_k^{(i)} \equiv E[H_k^{(i)}]$ and $\overline{B}_k^{(i)} \equiv E[B_k^{(i)}]$, $i = 1, \ldots, m$. Also, for

 ISBN: 978-84-617-8694-7

$i, j = 1, \ldots, m$, the following expectations are assumed to be known

$$\begin{aligned}
&E[f_{pq}(k) f_{p'q'}(s)] = E[f_{pq}(k) f_{p'q'}(k)]\delta_{k,s},\\
&E[h_{pq}^{(i)}(k) h_{p'q'}^{(j)}(s)] = E[h_{pq}^{(i)}(k) h_{p'q'}^{(j)}(k)]\delta_{k,s} + E[h_{pq}^{(i)}(k) h_{p'q'}^{(j)}(k-1)]\delta_{k-1,s}, \quad s \leq k,\\
&E[b_{pq}^{(i)}(k) b_{p'q'}^{(j)}(s)] = E[b_{pq}^{(i)}(k) b_{p'q'}^{(j)}(k)]\delta_{k,s} + E[b_{pq}^{(i)}(k) b_{p'q'}^{(j)}(k-1)]\delta_{k-1,s}, \quad s \leq k,\\
&E[h_{pq}^{(i)}(k) b_{p'q'}^{(j)}(s)] = E[h_{pq}^{(i)}(k) b_{p'q'}^{(j)}(k)]\delta_{k,s} + E[h_{pq}^{(i)}(k) b_{p'q'}^{(j)}(k-1)]\delta_{k-1,s}\\
&\qquad\qquad + E[h_{pq}^{(i)}(k) b_{p'q'}^{(j)}(k+1)]\delta_{k+1,s},
\end{aligned}$$

where $f_{pq}(k)$, $h_{pq}^{(i)}(k)$ and $b_{pq}^{(i)}(k)$ denote the (p,q)-th entries of F_k, $H_k^{(i)}$ and $B_k^{(i)}$, respectively.

Assumption 3. $\{w_k;\, k \geq 0\}$ and $\{v_k^{(i)};\, k \geq 1\}$, $i = 1, \ldots, m$, are zero-mean sequences and the following covariances and cross-covariances are known

$$\begin{aligned}
&E[w_k w_s^T] = Q_k \delta_{k,s} + Q_{k,k-1}\delta_{k-1,s}, \quad s \leq k,\\
&E[v_k^{(i)} v_s^{(j)T}] = R_k^{(ij)}\delta_{k,s} + R_{k,k-1}^{(ij)}\delta_{k-1,s}, \quad s \leq k, \quad i, j = 1, \ldots, m,\\
&E[w_k v_s^{(i)T}] = S_k^{(i)}\delta_{k,s} + S_{k,k+1}^{(i)}\delta_{k+1,s} + S_{k,k+2}^{(i)}\delta_{k+2,s}, \quad i = 1, \ldots, m.
\end{aligned}$$

Assumption 4. Independence assumptions:

- $\{F_k;\, k \geq 0\}$ is independent of $\Big(\{H_k^{(i)};\, k \geq 1\},\, \{B_k^{(i)};\, k \geq 1\},\, \{w_k;\, k \geq 0\},\, \{v_k^{(i)};\, k \geq 1\},\, i = 1, \ldots, m\Big)$
- $\Big(\{H_k^{(i)};\, k \geq 1\},\, \{B_k^{(i)};\, k \geq 1\},\, i = 1, \ldots, m\Big)$ is independent of $\Big(\{F_k;\, k \geq 0\},\, \{w_k;\, k \geq 0\},\, \{v_k^{(i)};\, k \geq 1\},\, i = 1, \ldots, m\Big)$

The following property will be used to calculate some expectations involving the random parameter matrices F_k, $H_k^{(i)}$ and $B_k^{(i)}$, $i = 1, \ldots, m$:

Let $A = \Big(a_{rs}\Big)_{\substack{r=1,\ldots,n_1\\ s=1,\ldots,n_2}}$, $B = \Big(b_{rs}\Big)_{\substack{r=1,\ldots,m_1\\ s=1,\ldots,m_2}}$ *and* $C = \Big(c_{rs}\Big)_{\substack{r=1,\ldots,n_2\\ s=1,\ldots,m_2}}$ *be random parameter matrices, such that* C *is independent of* (A, B)*; then, the* (p,q)*-th entries of the matrix* $E[ACB^T]$ *are given by*

$$\Big(E[ACB^T]\Big)_{pq} = \sum_{r=1}^{n_2}\sum_{s=1}^{m_2} E[a_{pr}b_{qs}]E[c_{rs}], \quad p = 1, \ldots, n_1, \quad q = 1, \ldots, m_1.$$

Our aim is to address the least squares (LS) linear filtering problem in this class of systems with correlated and cross-correlated random parameter matrices and noises using the distributed fusion approach.

 ISBN: 978-84-617-8694-7

3 Distributed fusion filtering estimators

In this section, the distributed fusion filter of the state x_k based on the available measurements $y_1^{(i)}, \ldots, y_k^{(i)}$, $i = 1, \ldots, m$, is designed. First, at each sensor, a local LS linear filtering recursive algorithm is obtained using its own measurement data. Then, these estimators are sent to the fusion centre where the distributed fusion filter is computed.

3.1 Local LS linear filters

In order to simplify the expressions of the local linear filtering estimators and the subsequent calculations, let us first present the following properties [10]:

1. The matrices $\mathcal{D}_{k+1} \equiv E[x_{k+1}x_{k+1}^T]$ and $\mathcal{D}_{k+1,k} \equiv E[x_{k+1}x_k^T]$ are recursively calculated by

$$\begin{array}{l} \mathcal{D}_{k+1} = E[F_k\mathcal{D}_kF_k^T] + \overline{F}_kQ_{k-1,k} + Q_{k,k-1}\overline{F}_k^T + Q_k, \quad k \geq 1; \\ \mathcal{D}_1 = E[F_0\mathcal{D}_0F_0^T] + Q_0; \quad \mathcal{D}_0 = \Sigma_0, \\ \mathcal{D}_{k+1,k} = \overline{F}_k\mathcal{D}_k + Q_{k,k-1}, \quad k \geq 1; \quad \mathcal{D}_{1,0} = \overline{F}_0\mathcal{D}_0. \end{array} \tag{2}$$

2. The noise processes $\{w_k;\, k \geq 0\}$ and $\{v_k^{(i)};\, k \geq 1\}$, $i = 1, \ldots, m$, satisfy the following correlation properties

$$\begin{array}{l} \mathcal{W}_k^{(i)} \equiv E[w_k y_k^{(i)T}] = Q_{k,k-1}\overline{H}_k^{(i)T} + S_k^{(i)}\overline{B}_k^{(i)T}, \quad k \geq 1, \\ \mathcal{E}_k^{(i)} \equiv E[x_k v_k^{(i)T}] = \overline{F}_{k-1}S_{k-2,k}^{(i)} + S_{k-1,k}^{(i)}, \quad k \geq 2; \quad \mathcal{E}_1^{(i)} = S_{0,1}^{(i)}, \\ \mathcal{E}_{k,k-1}^{(i)} \equiv E[x_k v_{k-1}^{(i)T}] = \overline{F}_{k-1}\mathcal{E}_{k-1}^{(i)} + S_{k-1}^{(i)}, \quad k \geq 2, \\ \mathcal{V}_{k,k-1}^{(ij)} \equiv E[B_k^{(i)}v_k^{(i)}y_{k-1}^{(j)T}] = E[B_k^{(i)}S_{k-2,k}^{(i)T}H_{k-1}^{(j)T}] + E[B_k^{(i)}R_{k,k-1}^{(ij)}B_{k-1}^{(j)T}], \quad k \geq 2. \end{array} \tag{3}$$

For each sensor $i = 1, \ldots, m$, a recursive algorithm for the local LS linear filter, $\widehat{x}_{k/k}^{(i)}$, together with its filtering error covariance matrix, $\Sigma_{k/k}^{(i)}$, is given in the following theorem.

Theorem 1 *For system* (1)*, under assumptions 1-4, the local LS linear filter for the* i*-th sensor,* $\widehat{x}_{k/k}^{(i)}$*,* $i = 1, \ldots, m$*, is given by*

$$\widehat{x}_{k/k}^{(i)} = \widehat{x}_{k/k-1}^{(i)} + \mathcal{X}_k^{(i)}\Pi_k^{(i)-1}\mu_k^{(i)}, \quad k \geq 1; \quad \widehat{x}_{0/0}^{(i)} = 0,$$

where the one-step predictor, $\widehat{x}_{k/k-1}^{(i)}$*, is calculated by*

$$\widehat{x}_{k/k-1}^{(i)} = \overline{F}_{k-1}\widehat{x}_{k-1/k-1}^{(i)} + \mathcal{W}_{k-1}^{(i)}\Pi_{k-1}^{(i)-1}\mu_{k-1}^{(i)}, \quad k \geq 2; \quad \widehat{x}_{1/0}^{(i)} = 0.$$

 ISBN: 978-84-617-8694-7

The innovation, $\mu_k^{(i)}$*, is obtained by*

$$\mu_k^{(i)} = y_k^{(i)} - \overline{H}_k^{(i)}\widehat{x}_{k/k-1}^{(i)} - \mathcal{Y}_{k,k-1}^{(i)}\Pi_{k-1}^{(i)-1}\mu_{k-1}^{(i)}, \quad k \geq 2; \quad \mu_1^{(i)} = y_1^{(i)},$$

where, denoting $\widetilde{H}_k^{(i)} \equiv H_k^{(i)} - \overline{H}_k^{(i)}$*, the matrix* $\mathcal{Y}_{k,k-1}^{(i)} \equiv E[y_k^{(i)}\mu_{k-1}^{(i)T}]$ *satisfies*

$$\mathcal{Y}_{k,k-1}^{(i)} = E[\widetilde{H}_k^{(i)}\mathcal{D}_{k,k-1}H_{k-1}^{(i)T}] + E[\widetilde{H}_k^{(i)}\mathcal{E}_{k,k-1}^{(i)}B_{k-1}^{(i)T}] + \mathcal{V}_{k,k-1}^{(i)}, \quad k \geq 2.$$

The matrix $\mathcal{X}_k^{(i)} \equiv E[x_k\mu_k^{(i)T}]$ *is obtained by*

$$\begin{aligned}
\mathcal{X}_k^{(i)} &= \Sigma_{k/k-1}^{(i)}\overline{H}_k^{(i)T} + \mathcal{E}_k^{(i)}\overline{B}_k^{(i)T} - \mathcal{X}_{k,k-1}^{(i)}\Pi_{k-1}^{(i)-1}\mathcal{Y}_{k,k-1}^{(i)T}, \quad k \geq 2;\\
\mathcal{X}_1^{(i)} &= \Sigma_{1/0}^{(i)}\overline{H}_1^{(i)T} + \mathcal{E}_1^{(i)}\overline{B}_1^{(i)T},
\end{aligned}$$

where $\mathcal{X}_{k,k-1}^{(i)} \equiv E[x_k\mu_{k-1}^{(i)T}]$ *is given by* $\mathcal{X}_{k,k-1}^{(i)} = \overline{F}_{k-1}\mathcal{X}_{k-1}^{(i)} + \mathcal{W}_{k-1}^{(i)}$, $k \geq 2$.
The innovation covariance matrix, $\Pi_k^{(i)}$*, satisfies*

$$\begin{aligned}
\Pi_k^{(i)} &= E[\widetilde{H}_k^{(i)}\mathcal{D}_k H_k^{(i)T}] + E[\widetilde{H}_k^{(i)}\mathcal{E}_k^{(i)}B_k^{(i)T}] + E[B_k^{(i)}\mathcal{E}_k^{(i)T}\widetilde{H}_k^{(i)T}] + E[B_k^{(i)}R_k^{(i)}B_k^{(i)T}]\\
&\quad + \overline{H}_k^{(i)}\mathcal{X}_k^{(i)} + \mathcal{X}_k^{(i)T}\overline{H}_k^{(i)T} - \overline{H}_k^{(i)}\Sigma_{k/k-1}^{(i)}\overline{H}_k^{(i)T} - \mathcal{Y}_{k,k-1}^{(i)}\Pi_{k-1}^{(i)-1}\mathcal{Y}_{k,k-1}^{(i)T}, \quad k \geq 2;\\
\Pi_1^{(i)} &= E[\widetilde{H}_1^{(i)}\mathcal{D}_1 H_1^{(i)T}] + E[H_1^{(i)}\mathcal{E}_1^{(i)}B_1^{(i)T}] + E[B_1^{(i)}\mathcal{E}_1^{(i)T}H_1^{(i)T}]\\
&\quad + E[B_1^{(i)}R_1^{(i)}B_1^{(i)T}] + \overline{H}_1^{(i)}\Sigma_{1/0}^{(i)}\overline{H}_1^{(i)T}.
\end{aligned}$$

The filtering error covariance matrix, $\Sigma_{k/k}^{(i)}$*, is computed by*

$$\Sigma_{k/k}^{(i)} = \Sigma_{k/k-1}^{(i)} - \mathcal{X}_k^{(i)}\Pi_k^{(i)-1}\mathcal{X}_k^{(i)T}, \quad k \geq 1; \quad \Sigma_{0/0}^{(i)} = \Sigma_0,$$

where the prediction error covariance matrix, $\Sigma_{k/k-1}^{(i)}$*, is given by*

$$\begin{aligned}
\Sigma_{k/k-1}^{(i)} &= \mathcal{D}_k + \overline{F}_{k-1}(\Sigma_{k-1/k-1}^{(i)} - \mathcal{D}_{k-1})\overline{F}_{k-1}^T + \mathcal{W}_{k-1}^{(i)}\Pi_{k-1}^{(i)-1}\mathcal{W}_{k-1}^{(i)T}\\
&\quad - \mathcal{X}_{k,k-1}^{(i)}\Pi_{k-1}^{(i)-1}\mathcal{W}_{k-1}^{(i)T} - \mathcal{W}_{k-1}^{(i)}\Pi_{k-1}^{(i)-1}\mathcal{X}_{k,k-1}^{(i)T}, \quad k \geq 2;\\
\Sigma_{1/0}^{(i)} &= \mathcal{D}_1.
\end{aligned}$$

Finally, $\mathcal{D}_k$*,* $\mathcal{D}_{k,k-1}$ *and* $\mathcal{W}_k^{(i)}$*,* $\mathcal{E}_k^{(i)}$*,* $\mathcal{E}_{k,k-1}^{(i)}$*,* $\mathcal{V}_{k,k-1}^{(i)}$ *are given in* (2) *and* (3)*, respectively.*

 ISBN: 978-84-617-8694-7

3.2 Distributed fusion filter weighted by matrices

As we have already indicated, our aim is to design a distributed fusion filter, $\widehat{x}_{k/k}^{(D)}$, as a matrix-weighted sum of the local filters, $\widehat{x}_{k/k}^{(i)}$, $i=1,\ldots,m$, such that the optimal weighting matrices are computed to minimize the mean squared estimation error. The following two lemmas provide some expectations required to obtain the proposed fusion estimator. Hereafter, the assumptions and notations used are the same as those in Theorem 1.

Lemma 1 *For $i,j=1,\ldots,m$ and $i\neq j$, the expectation $L_k^{(ij)}\equiv E[\widehat{x}_{k/k-1}^{(i)}\mu_k^{(j)T}]$ satisfies*

$$\begin{aligned}L_k^{(ij)}&=\left(\overline{F}_{k-1}K_{k-1/k-2}^{(i)}\overline{F}_{k-1}^T-K_{k/k-1}^{(ij)}\right)\overline{H}_k^{(j)T}\\&\quad+\mathcal{X}_{k,k-1}^{(i)}\Pi_{k-1}^{(i)-1}\left(\Delta_{k,k-1}^{(ji)T}+\mathcal{V}_{k,k-1}^{(ji)T}\right)-L_{k,k-1}^{(ij)}\Pi_{k-1}^{(j)-1}\mathcal{Y}_{k,k-1}^{(j)T},\quad k\geq 2;\ L_1^{(ij)}=0,\end{aligned}$$

where the expectation $\Delta_{k,k-1}^{(ij)}\equiv E[H_k^{(i)}x_k\mu_{k-1}^{(j)T}]$ is obtained by

$$\begin{aligned}\Delta_{k,k-1}^{(ij)}&=E[H_k^{(i)}\mathcal{D}_{k,k-1}H_{k-1}^{(j)T}]+E[H_k^{(i)}\mathcal{E}_{k,k-1}^{(j)}B_{k-1}^{(j)T}]\\&\quad-\overline{H}_k^{(i)}\overline{F}_{k-1}\left(K_{k-1/k-2}^{(j)}\overline{H}_{k-1}^{(j)T}+\mathcal{X}_{k-1,k-2}^{(j)}\Pi_{k-2}^{(j)-1}\mathcal{Y}_{k-1,k-2}^{(j)T}\right),\quad k\geq 3;\\\Delta_{2,1}^{(ij)}&=E[H_2^{(i)}\mathcal{D}_{2,1}H_1^{(j)T}]+E[H_2^{(i)}\mathcal{E}_{2,1}^{(j)}B_1^{(j)T}]-\overline{H}_2^{(i)}\overline{F}_1K_{1/0}^{(j)}\overline{H}_1^{(j)T},\end{aligned}$$

and $L_{k,k-1}^{(ij)}\equiv E[\widehat{x}_{k/k-1}^{(i)}\mu_{k-1}^{(j)T}]$ is computed by $L_{k,k-1}^{(ij)}=\overline{F}_{k-1}L_{k-1}^{(ij)}+\mathcal{X}_{k,k-1}^{(i)}\Pi_{k-1}^{(i)-1}\Pi_{k-1}^{(ij)}$, $k\geq 2$. The cross-covariance matrices, $K_{k/k-1}^{(ij)}\equiv E[\widehat{x}_{k/k-1}^{(i)}\widehat{x}_{k/k-1}^{(j)T}]$ are computed by

$$\begin{aligned}K_{k/k-1}^{(ij)}&=\overline{F}_{k-1}K_{k-1/k-2}^{(ij)}\overline{F}_{k-1}^T+\overline{F}_{k-1}L_{k-1}^{(ij)}\Pi_{k-1}^{(j)-1}\mathcal{X}_{k,k-1}^{(j)T}\\&\quad+\mathcal{X}_{k,k-1}^{(i)}\Pi_{k-1}^{(i)-1}L_{k-1}^{(ji)T}\overline{F}_{k-1}^T+\mathcal{X}_{k,k-1}^{(i)}\Pi_{k-1}^{(i)-1}\Pi_{k-1}^{(ij)}\Pi_{k-1}^{(j)-1}\mathcal{X}_{k,k-1}^{(j)T},\quad k\geq 2,\ i\neq j;\\K_{k/k-1}^{(i)}&=\mathcal{D}_k-\Sigma_{k/k-1}^{(i)},\quad k\geq 2;\quad K_{1/0}^{(ij)}=0.\end{aligned}$$

Lemma 2 *For $i,j=1,\ldots,m$ and $i\neq j$, the cross-covariance matrices of the innovations, $\Pi_k^{(ij)}\equiv E[\mu_k^{(i)}\mu_k^{(j)T}]$, are given by*

$$\begin{aligned}\Pi_k^{(ij)}&=\Delta_k^{(ij)}+E[B_k^{(i)}\mathcal{E}_k^{(i)T}H_k^{(j)T}]+E[B_k^{(i)}R_k^{(ij)}B_k^{(j)T}]\\&\quad-\mathcal{V}_{k,k-1}^{(ij)}\Pi_{k-1}^{(j)-1}\left(\overline{H}_k^{(j)}\mathcal{X}_{k,k-1}^{(j)}+\mathcal{Y}_{k,k-1}^{(j)}\right)^T-\overline{H}_k^{(i)}L_k^{(ij)}-\mathcal{Y}_{k,k-1}^{(i)}\Pi_{k-1}^{(i)-1}\Pi_{k-1,k}^{(ij)},\quad k\geq 2;\\\Pi_1^{(ij)}&=\Delta_1^{(ij)}+E[B_1^{(i)}\mathcal{E}_1^{(i)T}H_1^{(j)T}]+E[B_1^{(i)}R_1^{(ij)}B_1^{(j)T}],\end{aligned}$$

where the expectations $\Delta_k^{(ij)}\equiv E[H_k^{(i)}x_k\mu_k^{(j)T}]$ are obtained by

$$\begin{aligned}\Delta_k^{(ij)}&=E[H_k^{(i)}\mathcal{D}_kH_k^{(j)T}]+E[H_k^{(i)}\mathcal{E}_k^{(j)}B_k^{(j)T}]-\overline{H}_k^{(i)}\overline{F}_{k-1}K_{k-1/k-2}^{(j)}\overline{F}_{k-1}^T\overline{H}_k^{(j)T}\\&\quad-\Delta_{k,k-1}^{(ij)}\Pi_{k-1}^{(j)-1}\left(\overline{H}_k^{(j)}\mathcal{X}_{k,k-1}^{(j)}+\mathcal{Y}_{k,k-1}^{(j)}\right)^T,\quad k\geq 2;\\\Delta_1^{(ij)}&=E[H_1^{(i)}\mathcal{D}_1H_1^{(j)T}]+E[H_1^{(i)}\mathcal{E}_1^{(j)}B_1^{(j)T}],\end{aligned}$$

 ISBN: 978-84-617-8694-7

and $\Pi_{k-1,k}^{(ij)} \equiv E[\mu_{k-1}^{(i)}\mu_k^{(j)T}]$ *are given by*

$$\Pi_{k-1,k}^{(ij)} = \Delta_{k,k-1}^{(ji)T} + \mathcal{V}_{k,k-1}^{(ji)T} - L_{k,k-1}^{(ji)T}\overline{H}_k^{(j)T} - \Pi_{k-1}^{(ij)}\Pi_{k-1}^{(j)-1}\mathcal{Y}_{k,k-1}^{(j)T}, \quad k \geq 2.$$

The distributed fusion filter, $\widehat{x}_{k/k}^{(D)}$, and its error covariance matrix, $\Sigma_{k/k}^{(D)}$, are given in the following theorem.

Theorem 2 *Let* $\widehat{X}_{k/k} = (\widehat{x}_{k/k}^{(1)T}, \ldots, \widehat{x}_{k/k}^{(m)T})^T$ *be the vector consisting of the local filters obtained in Theorem 1. Then, the distributed fusion filter for the system* (1) *is given by*

$$\widehat{x}_{k/k}^{(D)} = \Xi_{k/k}K_{k/k}^{-1}\widehat{X}_{k/k}, \quad k \geq 1,$$

where $K_{k/k} = \left(K_{k/k}^{(ij)}\right)_{i,j=1,\ldots,m}$ *and* $\Xi_{k/k} = (K_{k/k}^{(1)}, \ldots, K_{k/k}^{(m)})$, *with the cross-covariance matrices,* $K_{k/k}^{(ij)}$, $i,j = 1,\ldots,m$, *between any two local filters calculated as*

$$\begin{aligned} K_{k/k}^{(ij)} &= K_{k/k-1}^{(ij)} + L_k^{(ij)}\Pi_k^{(j)-1}\mathcal{X}_k^{(j)T} + \mathcal{X}_k^{(i)}\Pi_k^{(i)-1}L_k^{(ji)T} \\ &\quad + \mathcal{X}_k^{(i)}\Pi_k^{(i)-1}\Pi_k^{(ij)}\Pi_k^{(j)-1}\mathcal{X}_k^{(j)T}, \quad k \geq 1,\, i \neq j; \quad K_{k/k}^{(i)} = \mathcal{D}_k - \Sigma_{k/k}^{(i)}, \quad k \geq 1. \end{aligned}$$

The error covariance matrices of the distributed fusion filtering estimators are computed by

$$\Sigma_{k/k}^{(D)} = \mathcal{D}_k - \Xi_{k/k}K_{k/k}^{-1}\Xi_{k/k}^T, \quad k \geq 1.$$

Finally, the matrices $L_k^{(ij)}$, $K_{k/k-1}^{(ij)}$ *and* $\Pi_k^{(ij)}$ *are given in Lemmas 1 and 2, respectively.*

4 Numerical simulation example

In this section we show that the results obtained in the current paper for the system model with random measurement matrices (1) can be applied to multi-sensor systems with correlated randomly delayed measurements as a particular case.

Consider a discrete-time linear stochastic system with state-dependent multiplicative noise and scalar randomly delayed measurements coming from two sensors:

$$\begin{aligned} &x_k = (0.95 + 0.2\epsilon_{k-1})x_{k-1} + w_{k-1}, \quad k \geq 1, \\ &z_k^{(i)} = C^{(i)}x_k + v_k^{(i)}, \quad k \geq 1, \quad i = 1,2, \\ &y_k^{(i)} = (1-\gamma_k^{(i)})z_k^{(i)} + \gamma_k^{(i)}z_{k-1}^{(i)}, \quad k \geq 2, \quad y_1^{(i)} = z_1^{(i)}, \quad i = 1,2 \end{aligned} \tag{4}$$

where $\{\epsilon_k;\ k \geq 0\}$ is a zero-mean Gaussian white process with unit variance. Let us assume that $C^{(1)} = 0.7$ and $C^{(2)} = 0.6$. The additive noises are defined as $w_k = 0.6(\eta_k + \eta_{k+1})$ and

 ISBN: 978-84-617-8694-7

$v_k^{(i)} = c^{(i)}\eta_k,\ i = 1, 2$, where $c^{(1)} = 1$, $c^{(2)} = 0.8$ and $\{\eta_k;\, k \geq 0\}$ is a zero-mean Gaussian white process with variance 0.5.

For $i = 1, 2$, the random variables $\gamma_k^{(i)}$, which model the random delays, are defined by two independent sequences of independent Bernoulli random variables, $\{\alpha_k^{(i)};\, k \geq 1\}$, $i = 1, 2$, with probabilities $P[\alpha_k^{(i)} = 1] = \overline{\alpha}^{(i)}$; specifically, for $i = 1, 2$, $\gamma_k^{(i)} = \alpha_{k+1}^{(i)}(1 - \alpha_k^{(i)})$.

Taking into account the previous definition, at each sensor the variables $\gamma_k^{(i)}$ and $\gamma_s^{(i)}$ are independent for $|k - s| \neq 0, 1$ but correlated at consecutive sampling times. The mean of these variables is $\overline{\gamma}^{(i)} = \overline{\alpha}^{(i)}(1 - \overline{\alpha}^{(i)})$, $i = 1, 2$ and the correlation function is given by $E[\gamma_k^{(i)}\gamma_s^{(i)}] = \begin{cases} \overline{\gamma}^{(i)}, & |k-s| = 0 \\ 0, & |k-s| = 1 \end{cases}$.

In order to apply the theoretical results established in Section 3, system (4) can be equivalently rewritten as the following one, with random parameter matrices:

$$\begin{aligned} &X_{k+1} = \mathcal{F}_k X_k + W_k, \quad k \geq 0, \\ &y_k^{(i)} = H_k^{(i)} X_k + B_k^{(i)} V_k^{(i)}, \quad k \geq 1, \quad i = 1, 2, \end{aligned} \tag{5}$$

where

$$X_k = \begin{pmatrix} x_k \\ x_{k-1} \end{pmatrix}, \quad k \geq 1, \quad X_0 = \begin{pmatrix} x_0 \\ 0 \end{pmatrix}, \quad \mathcal{F}_k = \begin{pmatrix} 0.95 + 0.2\epsilon_k & 0 \\ 1 & 0 \end{pmatrix}, \quad W_k = \begin{pmatrix} w_k \\ 0 \end{pmatrix},\ k \geq 0$$

$$H_k^{(i)} = \begin{cases} \left(C^{(i)}, 0\right), & k = 1 \\ \left((1 - \gamma_k^{(i)})C^{(i)}, \gamma_k^{(i)} C^{(i)}\right), & k \geq 2 \end{cases}, \quad B_k^{(i)} = \begin{cases} \left(1, 0\right), & k = 1 \\ \left(1 - \gamma_k^{(i)}, \gamma_k^{(i)}\right), & k \geq 2 \end{cases},$$

$$V_k^{(i)} = \begin{cases} \left(v_1^{(i)}, 0\right)^T, & k = 1 \\ \left(v_k^{(i)}, v_{k-1}^{(i)}\right)^T, & k \geq 2 \end{cases}.$$

The new noise processes and the random parameter matrices of system (5) satisfy the assumptions 1-4 to apply the algorithm proposed in this paper. Specifically, we have:

- $\{H_k^{(i)};\, k \geq 1\}$ and $\{B_k^{(i)};\, k \geq 1\}$ are correlated and cross-correlated at consecutive sampling times.
- The additive noises $\{W_k;\, k \geq 0\}$ and $\{V_k^{(i)};\, k \geq 1\}$ are correlated, with

$$\begin{aligned} &Q_k = \begin{pmatrix} 0.36 & 0 \\ 0 & 0 \end{pmatrix},\ k \geq 0, \quad Q_{k,k-1} = \begin{pmatrix} 0.18 & 0 \\ 0 & 0 \end{pmatrix},\ k \geq 1, \\ &R_1^{(ij)} = \begin{pmatrix} 0.5c^{(i)}c^{(j)} & 0 \\ 0 & 0 \end{pmatrix}, \quad R_k^{(ij)} = \begin{pmatrix} 0.5c^{(i)}c^{(j)} & 0 \\ 0 & 0.5c^{(i)}c^{(j)} \end{pmatrix}, \\ &R_{k,k-1}^{(ij)} = \begin{pmatrix} 0 & 0 \\ 0.5c^{(i)}c^{(j)} & 0 \end{pmatrix},\ k \geq 2, \end{aligned}$$

 ISBN: 978-84-617-8694-7

$$S_k^{(i)} = \begin{pmatrix} 0.3c^{(i)} & 0 \\ 0 & 0 \end{pmatrix}, \; k \geq 1,$$

$$S_{0,1}^{(i)} = S_1^{(i)}, \quad S_{k-1,k}^{(i)} = \begin{pmatrix} 0.3c^{(i)} & 0.3c^{(i)} \\ 0 & 0 \end{pmatrix}, \quad S_{k-2,k}^{(i)} = \begin{pmatrix} 0 & 0.3c^{(i)} \\ 0 & 0 \end{pmatrix}, \; k \geq 2.$$

To illustrate the feasibility of the proposed estimators, the corresponding algorithms were implemented in MATLAB, and one hundred iterations of the centralized linear filtering algorithm ([10]) and the proposed distributed filtering algorithm were run. The error variances of both distributed and centralized fusion filters were calculated for several values of $\overline{\alpha}^{(i)}$ which provide different values of the delay probabilities $\overline{\gamma}^{(i)}$, $i = 1, 2$. Only the cases $\overline{\alpha}^{(i)} \leq 0.5$ are displayed due to the symmetry of $\overline{\gamma}^{(i)} = \overline{\alpha}^{(i)}(1 - \overline{\alpha}^{(i)})$.

Figure 1 shows local, centralized, and distributed filtering error variances considering $\overline{\alpha}^{(1)} = 0.5$ and $\overline{\alpha}^{(2)} = 0.1$. This figure confirms that the distributed fusion filter outperforms each local filter and that the centralized method is more accurate than the distributed one.

Next, in order to show the dependence of the error variances upon the values $\overline{\alpha}^{(1)}$ and $\overline{\alpha}^{(2)}$, Figure 2 displays the filtering error variances, at a fixed iteration (namely, $k = 100$), when both $\overline{\alpha}^{(1)}$ and $\overline{\alpha}^{(2)}$ are varied from 0.1 to 0.5, which provide different values of the delay probabilities $\overline{\gamma}^{(1)}$ and $\overline{\gamma}^{(2)}$ ($\overline{\gamma}^{(i)} = 0.09, 0.16, 0.21, 0.24, 0.25$). In both graphs we can see that worse estimations are obtained as the delay probabilities increase. Also, this figure confirms the similar accuracy of both methods, centralized and distributed.

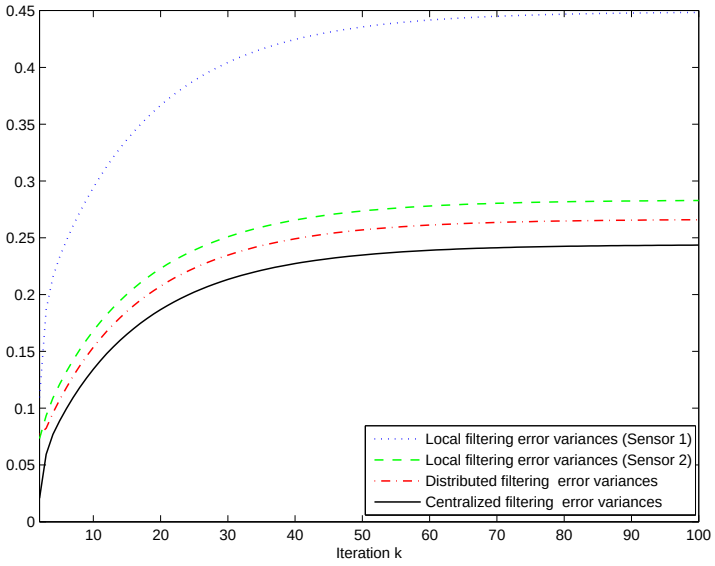


Figure 1: Filtering error variances for $\overline{\alpha}^{(1)} = 0.5$ and $\overline{\alpha}^{(2)} = 0.1$.

 ISBN: 978-84-617-8694-7

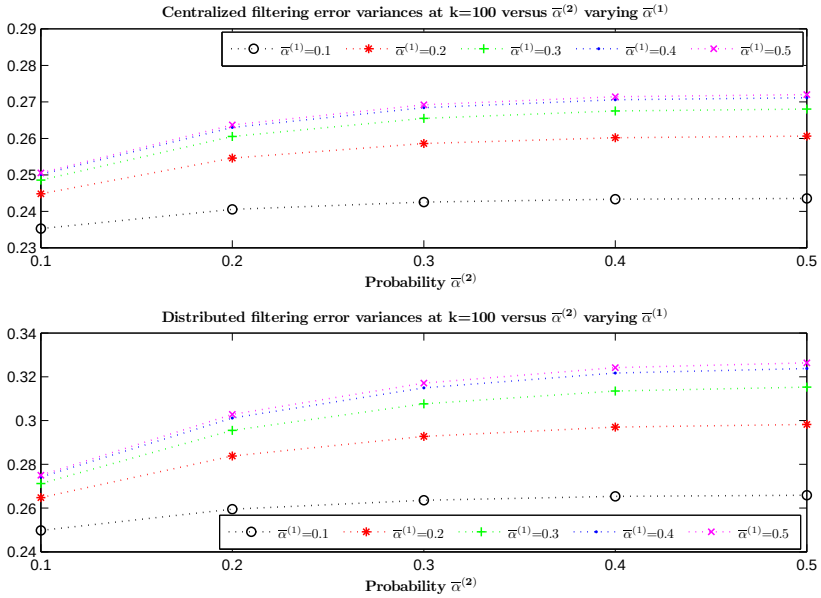


Figure 2: Filtering error variances at $k = 100$ versus $\overline{\alpha}^{(2)}$ with $\overline{\alpha}^{(1)}$ varying from 0.1 to 0.5.

5 Conclusion

The distributed fusion filtering problem has been addressed for multi-sensor stochastic systems with correlated random parameter matrices and additive noises. Firstly, recursive algorithms for the local LS linear filters of the system state based on the measurements coming from each sensor have been obtained. Next, a distributed fusion filter has been designed as a matrix-weighted linear combination of such local estimators by minimizing the mean squared estimation error. The applicability of the proposed estimators has been illustrated by a numerical simulation example, where an error variance comparison has been carried out to show the performance of the centralized and distributed fusion estimators.

Acknowledgements

This work is supported by *Ministerio de Economía y Competitividad* and *Fondo Europeo de Desarrollo Regional FEDER* (grant no. MTM2014-52291-P).

 ISBN: 978-84-617-8694-7

References

[1] J. Ma and S. Sun, *Information fusion estimators for systems with multiple sensors of different packet dropout rates*, Inf. Fusion **12** (2011) 213–222.

[2] R. Caballero-Águila, I. García-Garrido and J. Linares-Pérez, *Optimal fusion filtering in multisensor stochastic systems with missing measurements and correlated noises*, Math. Probl. Eng. **2013** (2013) Article ID 418678, 14 pages.

[3] T. Tian, S. Sun and N. Li, *Multi-sensor information fusion estimators for stochastic uncertain systems with correlated noises*, Inf. Fusion **27** (2016) 126–137.

[4] J. Ma and S. Sun, *Distributed fusion filter for networked stochastic uncertain systems with transmission delays and packet dropouts*, Signal Process. **130** (2017) 268–278.

[5] S. Wang, H. Fang and X. Tian, *Recursive estimation for nonlinear stochastic systems with multi-step transmission delays, multiple packet dropouts and correlated noises*, Signal Process. **115** (2015) 164–175.

[6] D. Chen, Y. Yu, L. Xu and X. Liu, *Kalman filtering for discrete stochastic systems with multiplicative noises and random two-step sensor delays*, Discret. Dyn. Nat. Soc. **2015** (2015) Article ID 809734, 11 pages.

[7] D. Chen, L. Xu and J. Du, *Optimal filtering for systems with finite-step autocorrelated process noise, random one-step sensor delay and missing measurements*, Commun. Nonlinear Sci. Numer. Simul. **32** (2016) 211–224.

[8] Y. Luo, Y. Zhu, D. Luo, J. Zhou, E. Song and D. Wang, *Globally optimal multisensor distributed random parameter matrices Kalman filtering fusion with applications*, Sensors **8** (2008) 8086-8103.

[9] J. Hu, Z. Wang and H. Gao, *Recursive filtering with random parameter matrices, multiple fading measurements and correlated noises*, Automatica **49** (2013) 3440–3448.

[10] J. Linares-Pérez, R. Caballero-Águila and I. García-Garrido, *Optimal linear filter design for systems with correlation in the measurement matrices and noises: recursive algorithm and applications*, Int. J. Syst. Sci. **45** (2014) 1548–1562.

[11] R. Caballero-Águila, I. García-Garrido and J. Linares-Pérez, *Distributed fusion filtering in networked systems with random measurement matrices and correlated noises*, Discret. Dyn. Nat. Soc. **2015** (2015) Article ID 398605, 10 pages.

[12] R. Caballero-Águila, A. Hermoso-Carazo and J. Linares-Pérez, *Fusion estimation using measured outputs with random parameter matrices subject to random delays and packet dropouts*, Signal Process. **127** (2016) 12–23.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

Centralized fusion estimation with random one-step delays and non-consecutive packet dropouts in transmission

R. Caballero-Águila[1], A. Hermoso-Carazo[2] and J. Linares-Pérez[2]

[1] *Dpto. de Estadística e Investigación Operativa, Universidad de Jaén (Spain)*

[2] *Dpto. de Estadística e Investigación Operativa, Universidad de Granada (Spain)*

emails: raguila@ujaen.es, ahermoso@ugr.es, jlinares@ugr.es

Abstract

The centralized fusion estimation problem is addressed for a class of discrete-time multisensor networked systems subject to random transmission one-step delays and non-consecutive packet dropouts with different rates. For each sensor, a different sequence of Bernoulli variables is used to model these random transmission failures, and the measured outputs are perturbed by both random parameter matrices and white noises. Using an innovation approach and without requiring full knowledge of the signal evolution model, but only the first and second order moments of the processes involved, a computationally simple and easily implementable recursive algorithm is obtained for the centralized prediction and filtering problems. The proposed estimators depend on the delay probabilities at each sampling time, but do not need to know if a particular measurement is delayed or well-timed. A numerical example is given which support our analysis and shows how the random delays influence the estimation accuracy.

Key words: centralized fusion estimation, random delays and packet dropouts
MSC 2000: 60G35, 62M20, 93E10, 93E11

1 Introduction

Over the past few decades, the scientific community has been concerned with networked systems, where the observations provided by all the network sensors are transmitted to a fusion center for being processed, and considerable attention has been paid to the fusion estimation of signals over multisensor systems with networked-induced random phenomena, such as missing and fading measurements, sensor gain degradation or multiplicative noise uncertainties, among others (see e.g. [1]-[5]).

 ISBN: 978-84-617-8694-7

Also, due to the limited bandwidths of the communication channels, random delays and/or packet losses are inevitable during the data transmission through the network and, hence, the measurements received by the processing center may be imperfect. Standard observation models are not appropriate under these transmission uncertainties, and classical estimation algorithms cannot be applied directly. For this reason, random delays and packet dropouts are two important issues which have received significant attention in the research of the estimation problem in networked systems (see e.g. [6]-[10]). Due to the random nature of transmission delays and packet dropouts, they can be described by Bernoulli random variables; however, different observation models are employed depending on whether each packet at sensor side is transmitted several times, in order to avoid losses as far as possible (see, e.g. [6]), or just once in order to avoid the network congestion (see e.g. [8]).

In relation to the mentioned networked-induced random phenomena, it is worth noting that random parameter measurement matrices can be used to describe, for example, missing or fading sensor measurements, and the networked systems involving stochastic multiplicative noises can be rewritten by using random parameter measurement matrices. Also, based on an augmentation approach, systems with random delays and packet dropouts, or systems with two-step random delays, can be transformed into systems with random parameter matrices. Consequently, this kind of systems with random parameter matrices provide an appropriate unified framework to model some of the aforementioned networked-induced random phenomena; this fact has encouraged an increasing research interest in the fusion estimation problems in this class of systems (see e.g. [11]- [16], and references therein).

Driven by the above considerations, this paper is concerned with the centralized fusion problem over multisensor networked systems using measurements perturbed by random parameter matrices and additive white noises, which are transmitted to the fusion center and one-step random delays and non-consecutive losses, with different rates, may occur during the transmission process. In order to avoid the network congestion, at each sampling time, the measured output of each sensor is sent just once and only one packet or no packet is available at the fusion center. The main contributions of this paper can be highlighted as follows: (1) the observation model considers random parameter matrices in the measured outputs and, simultaneously, random one-step delays and non-consecutive packet dropouts with different rates in the transmission; (2) an optimal least-squares linear recursive prediction and filtering algorithm is obtained without requiring full knowledge of the state-space model generating the signal process; and (3) the innovation technique is used, simplifying substantially the derivation of the algorithm, due to the fact that the innovation process is a white noise.

The rest of the paper is organized as follows. Section 2 presents the observation model with random parameter matrices and random one-step delays and non-consecutive packet dropouts to be considered. In Section 3, a recursive algorithm for the centralized least-squares linear prediction and filtering estimators is derived. In Section 4, the performance

 ISBN: 978-84-617-8694-7

of the proposed algorithm is illustrated by a numerical simulation example. Section 5 summarizes the main conclusions of the current study and points out some future research lines.

Notation: The notation used throughout the paper is standard. $\mathbb{R}^n$ denotes the n-dimensional Euclidean space. A^T and A^{-1} denote the transpose and inverse of a matrix A, respectively. The shorthand $Diag(a_1, \ldots, a_m)$ denotes a diagonal matrix whose diagonal entries are $a_1, \ldots, a_m$. $\mathbf{1} = (1, \ldots, 1)^T$ denotes the all-ones vector and I the identity matrix. If the dimensions of matrices are not explicitly stated, they are assumed to be compatible for algebraic operations. The notation $\otimes$ and $\circ$ represent the Kronecker and Hadamard products, respectively. Finally, $\delta_{k,s}$ represents the Kronecker delta function.

2 Problem formulation

Our aim is to find a recursive algorithm for the optimal least-squares (LS) linear prediction and filtering problem of a multidimensional discrete-time random signal using measurements perturbed by random parameter matrices and additive white noises, which are transmitted by multiple sensors, assuming that one-step random delays and non-consecutive losses may occur during the transmission process, with different rates at the different sensors. Next, we present the observation model and the assumptions under which the estimation problem will be addressed.

Observation model. Consider a networked system with m sensor nodes, which provide measured outputs, $z_k^{(i)} \in \mathbb{R}^{n_z}$, of the signal vector, $x_k \in \mathbb{R}^{n_x}$, according to the following model:

$$z_k^{(i)} = H_k^{(i)} x_k + v_k^{(i)}, \quad k \geq 1; \quad i = 1, \ldots, m, \tag{1}$$

where:

(H1) $\{x_k;\ k \geq 1\}$, the signal process, has zero mean and its autocovariance function is given by: $E[x_k x_s^T] = A_k B_s^T,\ s \leq k$, where $A_k, B_k,\ k \geq 1$, are known matrices.

(H2) $\{H_k^{(i)};\ k \geq 1\}$, $i = 1, \ldots, m$, are independent sequences of independent random parameter matrices, whose entries have known means and second-order moments; we will denote $\overline{H}_k^{(i)} \equiv E[H_k^{(i)}],\ k \geq 1$.

(H3) $\{v_k^{(i)};\ k \geq 1\},\ i = 1, \ldots, m$, are white noise sequences with zero mean and known second-order moments, satisfying $E[v_k^{(i)} v_s^{(j)T}] = R_k^{(ij)} \delta_{k,s}, \quad i, j = 1, \ldots, m$.

It is assumed that, at any sampling time, the outputs are transmitted from the m different sensors to a data processing center, where the signal estimation is performed and,

 ISBN: 978-84-617-8694-7

as a consequence of possible failures during the transmission process, one-step delays and non-consecutive packet dropouts may occur randomly in the transmission. In order to avoid the network congestion, for each $i = 1, \ldots, m$, the measured output $z_k^{(i)}$ of the i-th sensor is sent just once and only one packet or no packet is available at the processing center. Namely, for $i = 1, \ldots, m$, the measurements received are modelled as follows:

$$y_k^{(i)} = (1 - \gamma_k^{(i)}) z_k^{(i)} + \gamma_k^{(i)} \gamma_{k-1}^{(i)} z_{k-1}^{(i)}, \quad k \geq 2, \qquad y_1^{(i)} = (1 - \gamma_1^{(i)}) z_1^{(i)}, \tag{2}$$

where:

(H4) $\{\gamma_k^{(i)};\ k \geq 1\},\ i = 1, \ldots, m$, are independent sequences of independent Bernoulli random variables with known means, $\overline{\gamma}_k^{(i)} \equiv E[\gamma_k^{(i)}],\ k \geq 1$.

Also, it is assumed that,

(H5) For $i = 1, \ldots, m$, the processes $\{x_k;\ k \geq 1\}$, $\{H_k^{(i)};\ k \geq 1\}$, $\{v_k^{(i)};\ k \geq 1\}$, $\{\gamma_k^{(i)};\ k \geq 1\}$ are mutually independent.

Stacked observation model. To address the estimation problem through the centralized fusion method, the observations of the different sensors $\{y_1^{(i)}, \ldots, y_k^{(i)},\ i = 1, \ldots, m\}$, are gathered and jointly processed at each sampling time; for this purpose, the observation equations (1) and (2) are combined yielding the following observation model:

$$\begin{array}{l} z_k = H_k x_k + v_k, \quad k \geq 1. \\ y_k = (I - \Gamma_k) z_k + \Gamma_k \Gamma_{k-1} z_{k-1}, \quad k \geq 2; \qquad y_1 = (I - \Gamma_1) z_1, \end{array} \tag{3}$$

where $z_k = (z_k^{(1)T}, \ldots, z_k^{(m)T})^T$, $H_k = (H_k^{(1)T}, \ldots, H_k^{(m)T})^T$, $v_k = (v_k^{(1)T}, \ldots, v_k^{(m)T})^T$ and $\Gamma_k = Diag(\gamma_k^{(1)}, \ldots, \gamma_k^{(m)}) \otimes I$.

Hence, the problem is to obtain the LS linear estimator of the signal, x_k, based on the randomly delayed observations $\{y_1, \ldots, y_L\}$, $L \leq k$, given in (3). Next, we present the statistical properties of the processes involved in the observation model (3), from which the LS linear prediction and filtering algorithm of the signal x_k will be derived; these properties are easily inferred from the model hypotheses *(H1)-(H5)* previously established.

(P1) $\{H_k;\ k \geq 1\}$ is a sequence of independent random parameter matrices with known means, $\overline{H}_k \equiv E[H_k] = \left(\overline{H}_k^{(1)T}, \ldots, \overline{H}_k^{(m)T}\right)^T$, and

$$E[H_k x_k x_s^T H_s^T] = E[H_k A_k B_s^T H_s^T] = \left(E[H_k^{(i)} A_k B_s^T H_s^{(j)T}]\right)_{i,j=1\ldots,m}, \quad s \leq k,$$

 ISBN: 978-84-617-8694-7

where $E[H_k^{(i)}A_kB_s^TH_s^{(j)T}] = \overline{H}_k^{(i)}A_kB_s^T\overline{H}_s^{(j)T}$, for $j \neq i$ or $s \neq k$, and the entries of $E[H_k^{(i)}A_kB_k^TH_k^{(i)T}]$ are computed as follows:

$$\left(E[H_k^{(i)}A_kB_k^TH_k^{(i)T}]\right)_{pq} = \sum_{a=1}^{n_x}\sum_{b=1}^{n_x} E[h_{pa}^{(i)}(k)h_{qb}^{(i)}(k)](A_kB_k^T)_{ab}, \quad p,q = 1,\ldots,n_z,$$

where $h_{pq}^{(i)}(k)$ denotes the $(p,q)-$entry of the matrix $H_k^{(i)}$.

(P2) The noise $\{v_k;\ k \geq 1\}$ is a zero-mean sequence with known second-order moments defined by the matrices $R_k \equiv (R_k^{(ij)})_{i,j=1,\ldots,m}$.

(P3) The random matrices $\{\Gamma_k;\ k \geq 1\}$ are independent or, equivalently, the process $\{\gamma_k;\ k \geq 1\}$, where $\gamma_k = (\gamma_k^{(1)},\ldots,\gamma_k^{(m)})^T \otimes \mathbf{1}$, is a white sequence; their first and second-order moments are known, and the following notation will be used

- $\overline{\Gamma}_k \equiv E[\Gamma_k] = Diag\left(\overline{\gamma}_k^{(1)},\ldots,\overline{\gamma}_k^{(m)}\right) \otimes I.$
- $K_k^{\gamma} \equiv E\left[\gamma_k\gamma_k^T\right], \quad K_k^{\mathbf{1}-\gamma} \equiv E\left[(\mathbf{1}-\gamma_k)(\mathbf{1}-\gamma_k)^T\right], \quad K_k^{\gamma,\mathbf{1}-\gamma} \equiv E\left[\gamma_k(\mathbf{1}-\gamma_k)^T\right].$

(P4) The processes $\{x_k;\ k \geq 1\}$, $\{H_k;\ k \geq 1\}$, $\{v_k;\ k \geq 1\}$ and $\{\Gamma_k;\ k \geq 1\}$ are mutually independent.

Remark 1: From the previous properties, it is clear that $\{z_k;\ k \geq 1\}$ and $\{y_k;\ k \geq 1\}$ are zero-mean processes whose correlation functions, $\Sigma_{k,s}^z \equiv E[z_kz_s^T]$ and $\Sigma_{k,s}^y \equiv E[y_ky_s^T]$, for $s = k, k-1$, are obtained by the following expressions:

$$\begin{array}{l}
\Sigma_{k,s}^z = E[H_kA_kB_s^TH_s^T] + R_k\delta_{k,s}, \quad s \leq k. \\
\Sigma_k^y = K_k^{\mathbf{1}-\gamma} \circ \Sigma_k^z + K_k^{\mathbf{1}-\gamma,\gamma} \circ \Sigma_{k,k-1}^z\overline{\Gamma}_{k-1} + K_k^{\gamma,\mathbf{1}-\gamma} \circ \overline{\Gamma}_{k-1}\Sigma_{k-1,k}^z + K_k^{\gamma} \circ K_{k-1}^{\gamma} \circ \Sigma_{k-1}^z, \ k \geq 2; \\
\Sigma_1^y = K_1^{\mathbf{1}-\gamma} \circ \Sigma_1^z. \\
\Sigma_{k,k-1}^y = (I-\overline{\Gamma}_k)\Sigma_{k,k-1}^z(I-\overline{\Gamma}_{k-1}) + (I-\overline{\Gamma}_k)\Sigma_{k,k-2}^z\overline{\Gamma}_{k-2}\overline{\Gamma}_{k-1} + \overline{\Gamma}_kK_{k-1}^{\gamma,\mathbf{1}-\gamma} \circ \Sigma_{k-1}^z \\
\qquad\qquad + \overline{\Gamma}_kK_{k-1}^{\gamma} \circ \Sigma_{k-1,k-2}^z\overline{\Gamma}_{k-2}, \ k \geq 3; \\
\Sigma_{2,1}^y = (I-\overline{\Gamma}_2)\Sigma_{2,1}^z(I-\overline{\Gamma}_1) + \overline{\Gamma}_2K_1^{\gamma,\mathbf{1}-\gamma} \circ \Sigma_1^z.
\end{array} \tag{4}$$

3 Centralized fusion estimators

In this section, a recursive algorithm for the LS linear centralized fusion prediction and filtering estimators of the signal is obtained by an innovation approach. According to such approach, the observation process $\{y_k;\ k \geq 1\}$ is transformed into an equivalent one (innovation process) of orthogonal vectors $\{\mu_k;\ k \geq 1\}$, defined by $\mu_k = y_k - \widehat{y}_{k/k-1}$, where

$\widehat{y}_{k/k-1}$ is the orthogonal projection of y_k onto the linear space generated by $\{\mu_1,\ldots,\mu_{k-1}\}$. So, the LS linear estimator of any random vector w_k based on the observations $\{y_1,\ldots,y_L\}$, denoted as $\widehat{w}_{k/L}$, agrees with that based on the innovations $\{\mu_1,\ldots,\mu_L\}$, and, denoting $\Pi_h = E[\mu_h\mu_h^T]$, the following general expression for the LS linear estimators of w_k is obtained

$$\widehat{w}_{k/L} = \sum_{h=1}^{L} E[w_k\mu_h^T]\Pi_h^{-1}\mu_h. \tag{5}$$

One-stage observation predictor. To simplify future formulas and expressions, the observation model (3) will be equivalently written as follows:

$$y_k = (I-\Gamma_k)H_kx_k + \Gamma_k\overline{\Gamma}_{k-1}\overline{H}_{k-1}x_{k-1} + V_k, \quad k\geq 2,$$

where $V_k = (I-\Gamma_k)v_k + \Gamma_k\Gamma_{k-1}v_{k-1} + \Gamma_k(\Gamma_{k-1}H_{k-1} - \overline{\Gamma}_{k-1}\overline{H}_{k-1})x_{k-1}$.

From the general expression (5), denoting $\mathcal{V}_{k,h} \equiv E[V_k\mu_h^T]$, $h\leq k-1$, and since $\mathcal{V}_{k,h}=0$ for $h\leq k-2$, we obtain that $\widehat{V}_{k/k-1} = \mathcal{V}_{k,k-1}\Pi_{k-1}^{-1}\mu_{k-1}$, $k\geq 2$; hence, from the orthogonal projection lemma (OPL), the observation predictor is given by

$$\widehat{y}_{k/k-1} = (I-\overline{\Gamma}_k)\overline{H}_k\widehat{x}_{k/k-1} + \overline{\Gamma}_k\overline{\Gamma}_{k-1}\overline{H}_{k-1}\widehat{x}_{k-1/k-1} + \mathcal{V}_{k,k-1}\Pi_{k-1}^{-1}\mu_{k-1}, \quad k\geq 2 \tag{6}$$

and, consequently, the one-stage predictor and filter of the signal x_k are needed .

Centralized prediction and filtering recursive algorithm. The following theorem presents a recursive algorithm for the optimal LS linear centralized fusion estimators $\widehat{x}_{k/L}$, $L\leq k$, of the signal x_k based on the observations $\{y_1,\ldots,y_L\}$ given by (3).

Theorem 1. *The centralized predictor and filter, $\widehat{x}_{k/L}$, $L\leq k$, and the corresponding error covariance matrices, $\widehat{\Sigma}_{k/L} \equiv E[(x_k-\widehat{x}_{k/L})(x_k-\widehat{x}_{k/L}^T)]$, are obtained by*

$$\widehat{x}_{k/L} = A_ke_L, \quad L\leq k, \tag{7}$$

$$\widehat{\Sigma}_{k/L} = A_k\left(B_k - A_k\Sigma_L^e\right)^T, \quad k\geq 1, \tag{8}$$

where the vectors e_L and the matrices $\Sigma_L^e \equiv E[e_Le_L^T]$ are recursively obtained from

$$e_L = e_{L-1} + \mathcal{E}_L\Pi_L^{-1}\mu_L, \quad L\geq 1; \qquad e_0 = 0, \tag{9}$$

$$\Sigma_L^e = \Sigma_{L-1}^e + \mathcal{E}_L\Pi_L^{-1}\mathcal{E}_L^T, \quad L\geq 1; \quad \Sigma_0^e = 0. \tag{10}$$

The innovation, μ_L, satisfies

$$\mu_L = y_L - \overline{\mathcal{H}}_{A_L}e_{L-1} - \mathcal{V}_{L,L-1}\Pi_{L-1}^{-1}\mu_{L-1}, \quad L\geq 2; \quad \mu_1 = y_1, \tag{11}$$

 ISBN: 978-84-617-8694-7

where $\mathcal{V}_{L,L-1} = \Sigma^y_{L,L-1} - \mathcal{H}_{A_L}\mathcal{H}^T_{B_{L-1}}$, *and the innovation covariance,* Π_L, *is given by*

$$\Pi_L = \Sigma^y_L - \overline{\mathcal{H}}_{A_L}(\overline{\mathcal{H}}^T_{B_L} - \mathcal{E}_L) - \mathcal{V}_{L,L-1}\Pi^{-1}_{L-1}(\overline{\mathcal{H}}_{A_L}\mathcal{E}_{L-1} + \mathcal{V}_{L,L-1})^T, \quad L \geq 2; \quad \Pi_1 = \Sigma^y_1. \tag{12}$$

The matrices $\mathcal{E}_L$ *are given by*

$$\mathcal{E}_L = \overline{\mathcal{H}}^T_{B_L} - \Sigma^e_{L-1}\overline{\mathcal{H}}^T_{A_L} - \mathcal{E}_{L-1}\Pi^{-1}_{L-1}\mathcal{V}^T_{L,L-1}, \quad L \geq 2; \quad \mathcal{E}_1 = \overline{\mathcal{H}}^T_{B_1}. \tag{13}$$

Finally, the matrices Σ^y_L *and* $\Sigma^y_{L,L-1}$ *are given in (4), and the matrices* $\overline{\mathcal{H}}_{\Psi_L}$ *with* $\Psi_L = A_L,\ B_L$, *are defined by*

$$\overline{\mathcal{H}}_{\Psi_L} = (I - \overline{\Gamma}_L)\overline{H}_L\Psi_L + \overline{\Gamma}_L\overline{\Gamma}_{L-1}\overline{H}_{L-1}\Psi_{L-1}, \quad L \geq 2; \quad \overline{\mathcal{H}}_{\Psi_1} = (I - \overline{\Gamma}_1)\overline{H}_1\Psi_1. \tag{14}$$

Proof. From the general expression (5), to obtain the LS linear estimators $\widehat{x}_{k/L}$, $L \leq k$, it is necessary to calculate the coefficients $\mathcal{X}_{k,h} \equiv E[x_k\mu^T_h] = E[x_k y^T_h] - E[x_k\widehat{y}^T_{h/h-1}], \quad h \leq k$.

From *(H.1)* and the independence hypotheses, we have that $E[x_k y^T_h] = A_k\overline{\mathcal{H}}^T_{B_h}$, with $\overline{\mathcal{H}}_{B_h}$ given in (14). Now, using expression (6) for $\widehat{y}_{h/h-1}$, together with (5) for $\widehat{x}_{h/h-1}$ and $\widehat{x}_{h-1/h-1}$, we obtain that the filter coefficients are expressed as $\mathcal{X}_{k,h} = A_k\mathcal{E}_h,\ h \leq k$, where

$$\mathcal{E}_h = \overline{\mathcal{H}}^T_{B_h} - \sum_{j=1}^{h-1}\mathcal{E}_j\Pi^{-1}_j\mathcal{E}^T_j\overline{\mathcal{H}}^T_{A_h} - \mathcal{E}_{h-1}\Pi^{-1}_{h-1}\mathcal{V}^T_{h,h-1}, \quad h \geq 2; \quad \mathcal{E}_1 = \overline{\mathcal{H}}^T_{B_1}.$$

Then, by defining

$$e_L = \sum_{h=1}^{L}\mathcal{E}_h\Pi^{-1}_h\mu_h,\ L \geq 1; \quad e_0 = 0 \text{ and } \quad \Sigma^e_L = \sum_{h=1}^{L}\mathcal{E}_h\Pi^{-1}_h\mathcal{E}_h,\ L \geq 1; \quad \Sigma^e_0 = 0,$$

and taking into account that $E[e_L\mu^{(i)T}_h] = \mathcal{E}_h$, for $h \leq L$, and $E[e_L y^T_L] = \overline{\mathcal{H}}^T_{B_L}$, it is easy to obtain expressions (7)-(11) and (13) of Theorem 1.

Finally, expression (12) for Π_L is obtained taking into account that from the OPL, the innovation covariance matrix can be expressed as $\Pi_L = E[y_L y^T_L] - E[\widehat{y}_{L/L-1}\widehat{y}^T_{L/L-1}]$, and using that $\widehat{y}_{L/L-1} = \overline{\mathcal{H}}_{A_L}e_{L-1} + \mathcal{V}_{L,L-1}\Pi^{-1}_{L-1}\mu_{L-1}$. Then the proof is complete. □

4 Numerical simulation example

Consider a zero-mean scalar signal $\{x_k;\ k \geq 1\}$ with autocovariance function given by $E[x_k x_s] = 1.025641 \times 0.95^{k-s}, \quad s \leq k$, which is factorizable according to *(H1)* taking for example $A_k = 1.025641 \times 0.95^k$ and $B_k = 0.95^{-k}$. Measured outputs perturbed by white additive noises coming from two sensors are considered, and two different types of uncertainty are assumed in both sensors: missing measurements in sensor 1, and gain degradation in sensor 2. Specifically,

$$z^{(i)}_k = H^{(i)}_k x_k + v^{(i)}_k, \quad k \geq 1, \quad i = 1, 2,$$

 ISBN: 978-84-617-8694-7

with $H_k^{(i)} = 0.8\lambda_k^{(i)}$, where $\{\lambda_k^{(i)};\ k \geq 1\}$, $i = 1, 2$, are sequences of independent random variables, such that, for all $k \geq 1$, $\lambda_k^{(1)}$ a Bernoulli variable with $P[\lambda_k^{(1)} = 1] = p$, and $\lambda_k^{(2)}$ is uniformly distributed on the interval $[0.1, 0.9]$. The additive noises are defined as $v_k^{(i)} = c_i\eta_k$, $i = 1, 2$, where $c_1 = 0.5$, $c_2 = 0.75$, and $\{\eta_k;\ k \geq 1\}$ is a zero-mean Gaussian white process with unit variance. Also, we assume that the sequences $\{\eta_k;\ k \geq 1\}$ and $\{\lambda_k^{(i)};\ k \geq 1\}$, $i = 1, 2$, are mutually independent.

Now, according to the proposed observation model, it is assumed that, at any sampling time $k \geq 1$, the measured output from the i-th sensor, can be randomly delayed by one sampling period or lost during network transmission; that is,

$$y_k^{(i)} = (1 - \gamma_k^{(i)})z_k^{(i)} + \gamma_k^{(i)}\gamma_{k-1}^{(i)}z_{k-1}^{(i)}, \quad k \geq 2, \quad y_1^{(i)} = (1 - \gamma_1^{(i)})z_1^{(i)}; \qquad i = 1, 2,$$

where $\{\gamma_k^{(i)};\ k > 1\}$, $i = 1,\ 2$, are independent sequences of independent Bernoulli random variables with constant probabilities, $P[\gamma_k^{(i)} = 1] = \overline{\gamma}^{(i)}$.

To illustrate the effectiveness and applicability of the proposed estimators, the centralized prediction and filtering algorithm was implemented in MATLAB and a hundred iterations were performed.

Firstly, in order to compare the performance of the predictors, $\widehat{z}_{k/L}$, $L = k - 3,\ k - 2,\ k - 1$ and the filter, $\widehat{z}_{k/k}$, their error variances are calculated considering constant values $p = 0.5$ and $\overline{\gamma}^{(1)} = 0.2$, $\overline{\gamma}^{(2)} = 0.4$. The results are displayed in Figure 1, which shows that the filtering error variances are smaller than the prediction ones, thus confirming that the filter outperforms the predictor. Also, this figure shows that the predictor becomes more accurate as the number of available observations increases. Analogous results are obtained for other values of the probabilities p, $\overline{\gamma}^{(1)}$ and $\overline{\gamma}^{(2)}$.

Next, we analyze the sensitivity of the estimation performance to the variation of the probability p that the signal is present in the measurements of the first sensor. For this purpose, the filtering error variances are calculated for different values of p, when $\overline{\gamma}^{(1)} = 0.2$ and $\overline{\gamma}^{(2)} = 0.4$. The results are given in Figure 2, which shows that, as p increases, the filtering error variances become smaller and, hence, better estimations are obtained. Analogous conclusions are deduced for other values of $\overline{\gamma}^{(1)}$, $\overline{\gamma}^{(2)}$ and p.

Finally, considering a fixed value of p, namely $p = 0.5$, the filtering error variances have been calculated for different values of the delay probabilities $\overline{\gamma}^{(1)}$ and $\overline{\gamma}^{(2)}$. Specifically, the values $\overline{\gamma}^{(1)} = 0.1,\ 0.2,\ 0.3,\ 0.4,\ 0.5$, and $\overline{\gamma}^{(2)} = 0.1,\ 0.3,\ 0.5,\ 0.6$, have been used. The results are displayed in Figure 3, which shows that the filtering error variances become greater (and, consequently, worse estimations are obtained) as the delay probability $\overline{\gamma}^{(1)}$ or $\overline{\gamma}^{(2)}$ increases.

 ISBN: 978-84-617-8694-7

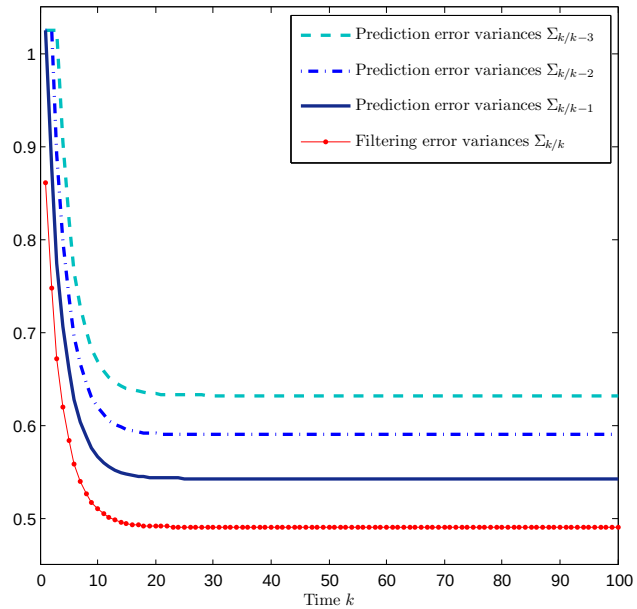


Figure 1: Prediction and filtering error variances, when $p = 0.5$ and $\overline{\gamma}^{(1)} = 0.2$, $\overline{\gamma}^{(2)} = 0.4$.

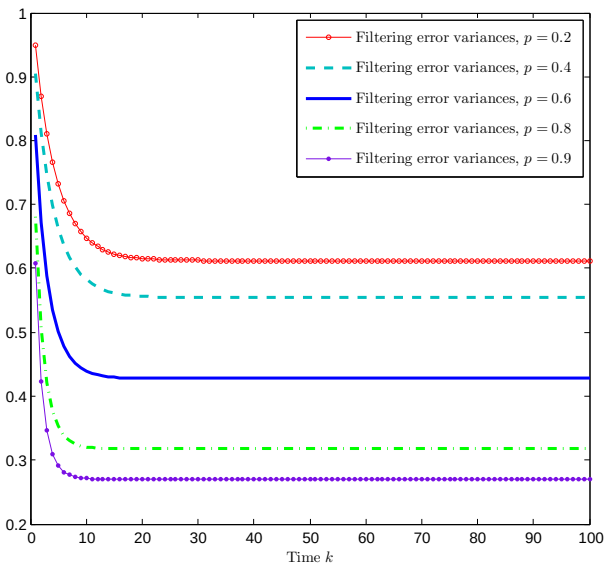


Figure 2: Filtering error variances for different values of p, when $\overline{\gamma}^{(1)} = 0.2$ and $\overline{\gamma}^{(2)} = 0.4$.

 ISBN: 978-84-617-8694-7

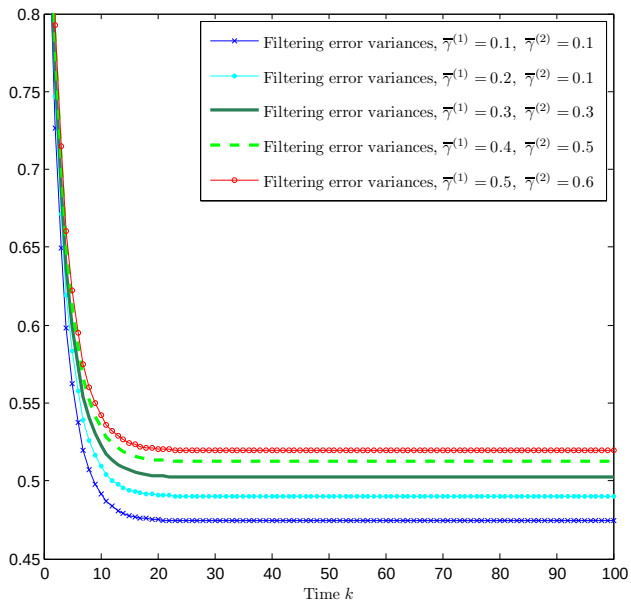


Figure 3: Filtering error variances for different values of $\overline{\gamma}^{(1)}$ and $\overline{\gamma}^{(2)}$, when $p = 0.5$

5 Concluding remark

In this paper, we have studied the recursive LS centralized fusion estimation problem for a class of multisensor uncertain systems with random parameter matrices, under the assumption that the sensor output transmissions to the processing center may experience random one-step delays and non-consecutive packet losses, when these transmission uncertainties occur with different rates at the different sensors. For the addressed model, we have derived an optimal recursive algorithm for the filtering and prediction problems, using the innovation technique. To measure the estimation accuracy, recursive formulas for the error covariance matrices have been also derived. The current results are applicable when the signal evolution model is not available, since the algorithm needs only the mean and covariance functions of the processes involved, but it can be also applied to the conventional state-space model formulation. The usefulness of the developed estimation algorithm has been illustrated by a numerical simulation example.

Future lines of research will attempt to: (i) Consider noise correlation, since many practical situations involve non-white measurement noises that can also be correlated among the sensors, and this correlation should not be ignored. (ii) Explore the design of distributed fusion estimators which, as it is known, despite being slightly less accurate than the centralized ones, are more robust and save time and storage resources in the fusion center.

 ISBN: 978-84-617-8694-7

Acknowledgements

This research is supported by *Ministerio de Economía y Competitividad* and *Fondo Europeo de Desarrollo Regional FEDER* (grant no. MTM2014-52291-P).

References

[1] R. Caballero-Águila, I. García-Garrido and J. Linares-Pérez, *Information fusion algorithms for state estimation in multi-sensor systems with correlated missing measurements*, Appl. Math. Comput. **226** (2014) 548–563.

[2] P. Fangfang and S. Sun, *Distributed fusion estimation for multisensor multirate systems with stochastic observation multiplicative noises*, Math. Probl. Eng. **2014** (2014) (art. no.373270).

[3] Y. Liu, X. He, Z. Wang and D. Zhou, *Optimal filtering for networked systems with stochastic sensor gain degradation*, Automatica **50(5)** (2014) 1521–1525.

[4] F. Zhou, L. Wu and X. Feng, *Sequential fusion for asynchronous multi-sensor fading measurements*, Int. J. Control Autom. **9(4)** (2016) 197–208.

[5] Y. Liu, Z. Wang, X. He and D.H. Zhou, *Minimum-variance recursive filtering over sensor networks with stochastic sensor gain degradation: Algorithms and performance analysis*, IEEE Trans. Control Netw. Syst. **3(3)** (2016) 265–274.

[6] J. Hu, Z. Wang, B. Shen and H. Gao, *Gain-constrained recursive filtering with stochastic nonlinearities and probabilistic sensor delays*, IEEE Trans. Signal Process. **61(5)** (2013) 1230–1238.

[7] J. Ma and S. Sun, *Centralized fusion estimators for multisensor systems with random sensor delays, multiple packet dropouts and uncertain observations*, IEEE Sensors J. **13(4)** (2013) 1228–1235.

[8] N. Li, S. Sun and J. Ma, *Multi-sensor distributed fusion filtering for networked systems with different delay and loss rates*, Digit. Signal Process. **34** (2014) 29–38.

[9] B. Chen, W. Zhang and L. Yu, *Distributed fusion estimation with missing measurements, random transmission delays and packet dropouts*, IEEE Trans. Automat. Contr. **59(7)** (2014) 1961–1967.

[10] J. Ma and S. Sun, *Distributed fusion filter for networked stochastic uncertain systems with transmission delays and packet dropouts*, Signal Process. **130** (2017) 268–278.

 ISBN: 978-84-617-8694-7

[11] Y. Luo, Y. Zhu, D. Luo, J. Zhou, E. Song and D. Wang, *Globally optimal multisensor distributed random parameter matrices Kalman filtering fusion with applications*, Sensors **8(12)** (2008) 8086–8103.

[12] X.J. Shen, Y.T. Luo, Y.M. Zhu and E.B. Song, *Globally optimal distributed Kalman filtering fusion*, *Sci. China Inf. Sci.* **55(3)** (2012) 512–529.

[13] J. Hu, Z. Wang and H. Gao, *Recursive filtering with random parameter matrices, multiple fading measurements and correlated noises*, Automatica **49** (2013) 3440-3448.

[14] Y. Yang, Y. Liang, Q. Pan, Y. Qin and F. Yang, *Distributed fusion estimation with square-root array implementation for Markovian jump linear systems with random parameter matrices and cross-correlated noises*, Inf. Sci. **370–371** (2016) 446–462.

[15] R. Caballero-Águila, A. Hermoso-Carazo and J. Linares-Pérez, *Networked Fusion Filtering from Outputs with Stochastic Uncertainties and Correlated Random Transmission Delays*, Sensors **16** (2016) 847.

[16] R. Caballero-Águila, A. Hermoso-Carazo and J. Linares-Pérez, *Fusion estimation using measured outputs with random parameter matrices subject to random delays and packet dropouts*, Signal Process. **127** (2016) 12–23.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

Multi-sensor distributed fusion filtering from observations with different random transmission failures

R. Caballero-Águila[1], A. Hermoso-Carazo[2] and J. Linares-Pérez[2]

[1] *Dpto. de Estadística e Investigación Operativa, Universidad de Jaén (Spain)*

[2] *Dpto. de Estadística e Investigación Operativa, Universidad de Granada (Spain)*

emails: raguila@ujaen.es, ahermoso@ugr.es, jlinares@ugr.es

Abstract

The distributed fusion filtering problem is addressed for discrete-time random signals from multi-sensor noisy measurements, which are transmitted to local processors through different communication channel links. Assume that, due to random transmission failures, some of the data packet processed for the estimation may either contain only noise (uncertain observations), be delayed (randomly delayed observations) or even be definitely lost (random packet dropouts). Using only covariance information, without requiring the evolution model of the signal process, local least-squares linear estimators based on the measurements received by the local processor of each individual sensor are obtained, and the distributed fusion method is then used to generate an optimal fusion filter by a matrix-weighted linear combination of such local estimators using the mean squared error as optimality criterion. The accuracy of the proposed estimators, which is measured by the estimation error covariances, is examined by a simulation example.

Key words: distributed fusion estimation, uncertain observations, random delays, packet dropouts

MSC 2000: 60G35, 62M20, 93E10, 93E11

1 Introduction

Sensor networks are currently one of the most important technologies and information fusion has become a challenging issue within the study of the estimation problem in networked stochastic systems. Many of the existing fusion estimation algorithms are concerned with conventional systems, where each sensor transmits its outputs to the fusion center (FC) over perfect connections (see e.g. [1] and [2], and references therein).

However, usually the network characteristics may not be utterly reliable and, when the sensors send their measurements to the FC, some problems may arise as, for example, uncertain observations or missing measurements, random delays and/or packet dropouts, thus causing the deterioration of the quality of the fusion estimators designed without considering these drawbacks. For this reason, the design of new fusion estimation algorithms for systems featuring one of the aforementioned uncertainties (see e.g.[3]-[5] and references therein), or even several of them simultaneously (see e.g. [6]- [13] and references therein), has become an active research topic of growing interest.

There are two fundamental kinds of fusion filtering algorithms: (a) The centralized fusion filtering algorithms (see e.g. [6]), where the sensors simply forward all the measured data to the FC and then they are processed to provide optimal estimators; hence, when all the sensors work correctly, centralized fusion estimators have the best accuracy. (b) The distributed fusion filtering algorithms (see e.g. [12]), where local estimators are obtained from the measurements of each single sensor, and these local estimators are then combined according to a certain information fusion criterion; the distributed estimators, despite being slightly less accurate than the centralized ones, have better robustness and reliability.

In this paper, we address the distributed fusion estimation problem in networked systems with multiple uncertainties during transmission, which include random delays, packet dropouts and/or uncertain observations. To the best of the authors' knowledge, the simultaneous consideration of these uncertainties has not yet been investigated in the framework of covariance information and, therefore, it constitutes an interesting research challenge. The main contributions of the present paper include: (1) Our approach, based on covariance information, does not require the evolution model generating the signal process. (2) The estimators are obtained without the necessity of augmenting the state; so, the dimension of the designed estimators is the same as that of the original state, thus reducing the computational cost compared with the augmentation method.

The rest of the paper is structured as follows. In Section 2, we present the measurement model to be considered and the assumptions under which the distributed estimation problem is addressed. In Section 3, a local least-squares linear filtering algorithm is derived, and the proposed distributed estimators are generated by a matrix-weighted linear combination of the local estimators using the mean squared error as optimality criterion. A simulation example is given in Section 4 to show the performance of the proposed estimators. Finally, some conclusions are drawn in Section 5.

Notations. The notations used throughout the paper are standard. $\mathbb{R}^n$ denotes the n-dimensional Euclidean space. For a matrix A, A^T and A^{-1} denote its transpose and inverse, respectively. If a matrix dimension is not specified, it is assumed to be compatible with algebraic operations. $\delta_{k,s}$ denotes the Kronecker delta function. Finally, for any function $G_{k,s}$, depending on the instants k and s, we will write $G_k = G_{k,k}$ for simplicity; analogously, $K^{(i)} = K^{(ii)}$ will be written for any function $K^{(ij)}$, depending on sensors i and j.

 ISBN: 978-84-617-8694-7

2 Observation model

This paper deals with the fusion estimation problem of discrete-time random signals from multi-sensor noisy measurements transmitted through different channels using the distributed fusion method. Each sensor is assumed to transmit its outputs to a local processor over imperfect networks, which yield mixed uncertainties. Specifically, the observations processed for the estimation may either contain only noise, be one-step randomly delayed or dropped out, in which case the last observation that successfully arrived will be used for the estimation.

Signal process. The distributed fusion estimators will be obtained under the assumption that the evolution model of the signal to be estimated is unknown and only information about its mean and covariance functions is available; this information is specified in the following assumption: *The n_x-dimensional signal process $\{x_k;\ k \geq 1\}$ has zero mean and its autocovariance function is expressed in a separable form, $E[x_k x_s^T] = A_k B_s^T,\ s \leq k$, where A_k, B_s are known matrices.*

Multi-sensor measured outputs. Consider m sensors, whose measurements are described by the following equations:

$$z_k^{(i)} = H_k^{(i)} x_k + v_k^{(i)}, \quad k \geq 1, \quad i = 1, \ldots, m, \tag{1}$$

where $z_k^{(i)} \in \mathbb{R}^{n_z}$ is the measured output of the i-th sensor at time k, and $v_k^{(i)}$ is the noise vector. Assume that the measurement noises $\{v_k^{(i)};\ k \geq 1\}$, $i = 1, \ldots, m$, are zero-mean white processes with known covariances $E[v_k^{(i)} v_s^{(j)T}] = R_k^{(ij)} \delta_{k,s}$, for $i, j = 1, \ldots, m$.

From this assumption and the separable form of the signal autocovariance function, it is clear that, for $i, j = 1, \ldots, m$, the correlation matrices $\Sigma_{s,k}^{z^{(ij)}} \equiv E[z_s^{(i)} z_k^{(j)T}]$ are given by

$$\Sigma_{s,k}^{z^{(ij)}} = H_s^{(i)} B_s A_k^T H_k^{(j)T} + R_k^{(ij)} \delta_{s,k}, \quad s \leq k. \tag{2}$$

Observation model with mixed uncertainties. As it has been indicated, random one-step delays, packet dropouts and uncertain observations are supposed to exist in data transmissions from the individual sensors to the local processors. Namely, the following model is considered for the measurement of the i-th local processor, $y_k^{(i)}$, $i = 1, \ldots, m$:

$$\begin{array}{l} y_k^{(i)} = \gamma_{0,k}^{(i)} z_k^{(i)} + \gamma_{1,k}^{(i)} z_{k-1}^{(i)} + \gamma_{2,k}^{(i)} v_k^{(i)} + \gamma_{3,k}^{(i)} y_{k-1}^{(i)},\ k \geq 2; \\ y_1^{(i)} = \gamma_{0,1}^{(i)} z_1^{(i)} + \gamma_{2,1}^{(i)} v_1^{(i)}, \end{array} \tag{3}$$

where $\gamma_{3,k}^{(i)} = 1 - \sum_{d=0}^{2} \gamma_{d,k}^{(i)},\ k \geq 2$, and $\gamma_{2,1}^{(i)} = 1 - \gamma_{0,1}^{(i)}$. We denote $\gamma_{1,1}^{(i)} = \gamma_{3,1}^{(i)} = 0$, and we assume that, for $i = 1, \ldots, m$, and $d = 0, 1, 2, 3$, the process $\{\gamma_{d,k}^{(i)};\ k \geq 1\}$ is a sequence of

 ISBN: 978-84-617-8694-7

independent Bernoulli variables with known probabilities $P[\gamma_{d,k}^{(i)}=1]=\overline{\gamma}_{d,k}^{(i)}$, $\forall k\geq 1$. Also, we assume that $\{\gamma_{d,k}^{(i)};\ k\geq 1\}$ is independent of the sequences $\{\gamma_{d',k}^{(j)};\ k\geq 1\}$, $d'=0,1,2,3$, for any $j\neq i$.

Finally, the following independence hypothesis is also assumed: *For $i=1,\ldots,m$, and $d=0,1,2,3$, the signal, $\{x_k;\ k\geq 1\}$, and the processes $\{v_k^{(i)};\ k\geq 1\}$ and $\{\gamma_{d,k}^{(i)};\ k\geq 1\}$ are mutually independent.*

3 Distributed fusion filtering estimators

In this section a recursive filter is designed using the distributed fusion method. For this purpose, for each $i=1,\ldots,m$ a local least-squares (LS) linear estimator, $\widehat{x}_{k/k}^{(i)}$, $k\geq 1$, of the signal x_k, based on the measurements $\left\{y_1^{(i)},\ldots,y_k^{(i)}\right\}$, is obtained by a recursive algorithm and then a fusion distributed estimator, $\widehat{x}_{k/k}$, is generated by a matrix-weighted linear combination of the local estimators, $\widehat{x}_{k/k}^{(i)}$, $i=1,\ldots,m$, using the mean squared error as optimality criterion.

3.1 Local LS linear filtering recursive algorithm

For each $i=1,\ldots,m$, the following theorem presents a recursive algorithm for the optimal LS linear filtering estimators $\widehat{x}_{k/k}^{(i)}$, $k\geq 1$.

Theorem 1. *For $i=1,\ldots,m$, the local LS linear filtering estimators, $\widehat{x}_{k/k}^{(i)}$, $k\geq 1$, and their error covariance matrices, $P_{k/k}^{(i)}$, $k\geq 1$, are obtained by*

$$\widehat{x}_{k/k}^{(i)}=A_kO_k^{(i)},\quad k\geq 1;\qquad P_{k/k}^{(i)}=A_k\Big(B_k-A_kr_k^{(i)}\Big)^T,\quad k\geq 1. \tag{4}$$

where

$$O_k^{(i)}=O_{k-1}^{(i)}+J_k^{(i)}\Pi_k^{(i)-1}\mu_k^{(i)},\quad k\geq 1;\qquad O_0^{(i)}=0, \tag{5}$$

with

$$J_k^{(i)}=\mathcal{H}_{B_k}^{(i)T}-r_{k-1}^{(i)}\mathcal{H}_{A_k}^{(i)T}-J_{k-1}^{(i)}\Pi_{k-1}^{(i)-1}\mathcal{R}_{k,k-1}^{(i)T},\quad k\geq 2;\quad J_1^{(i)}=\mathcal{H}_{B_1}^{(i)T}, \tag{6}$$

where $\mathcal{R}_{k,k-1}^{(i)}=\overline{\gamma}_{1,k}^{(i)}\left(\overline{\gamma}_{0,k-1}^{(i)}+\overline{\gamma}_{2,k-1}^{(i)}\right)R_{k-1}^{(i)}$, $k\geq 2$, and $r_k^{(i)}=E[O_k^{(i)}O_k^{(i)T}]$ is obtained from

$$r_k^{(i)}=r_{k-1}^{(i)}+J_k^{(i)}\Pi_k^{(i)-1}J_k^{(i)T},\quad k\geq 1;\qquad r_0^{(i)}=0. \tag{7}$$

The innovation, $\mu_k^{(i)}$, is calculated by

$$\mu_k^{(i)}=y_k^{(i)}-\mathcal{H}_{A_k}^{(i)}O_{k-1}^{(i)}-\mathcal{R}_{k,k-1}^{(i)}\Pi_{k-1}^{(i)-1}\mu_{k-1}^{(i)}-\overline{\gamma}_{3,k}^{(i)}y_{k-1}^{(i)},\quad k\geq 2;\quad \mu_1=y_1^{(i)} \tag{8}$$

 ISBN: 978-84-617-8694-7

and its covariance matrix, $\Pi_k^{(i)}$*, is given by*

$$\begin{array}{l}\Pi_k^{(i)}=\Sigma_k^{y^{(i)}}-(\overline{\gamma}_{3,k}^{(i)})^2\Sigma_{k-1}^{y^{(i)}}-\overline{\mathcal{H}}_{A_k}^{(i)}\left(\mathcal{H}_{B_k}^{(i)T}-J_k^{(i)}\right)-\mathcal{R}_{k,k-1}^{(i)}\Pi_{k-1}^{(i)-1}\left(\mathcal{H}_{A_k}^{(i)}J_{k-1}^{(i)}+\mathcal{R}_{k,k-1}^{(i)}\right)^T,\ k\geq 2;\\ \Pi_1^{(i)}=\Sigma_1^{y^{(i)}}\end{array} \tag{9}$$

The matrices $\Sigma_k^{y^{(i)}}=E[y_k^{(i)}y_k^{(i)T}]$ *satisfy*

$$\Sigma_k^{y^{(i)}}=\overline{\gamma}_{0,k}^{(i)}\Sigma_k^{z^{(i)}}+\overline{\gamma}_{1,k}^{(i)}\Sigma_{k-1}^{z^{(i)}}+\overline{\gamma}_{2,k}^{(i)}R_k^{(i)}+\overline{\gamma}_{3,k}^{(i)}\Sigma_{k-1}^{y^{(i)}},\ k\geq 2;\quad \Sigma_1^{y^{(i)}}=\overline{\gamma}_{0,1}^{(i)}\Sigma_1^{z^{(i)}}+\overline{\gamma}_{2,1}^{(i)}R_1^{(i)},$$

where $\Sigma_k^{z^{(i)}}$ *is given in (2), and the matrices* $\mathcal{H}_{\Psi_k}^{(i)}$*, for* $\Psi_k=A_k,\ B_k$*, are defined by*

$$\mathcal{H}_{\Psi_k}^{(i)}=\overline{\gamma}_{0,k}^{(i)}H_k^{(i)}\Psi_k+\overline{\gamma}_{1,k}^{(i)}H_{k-1}^{(i)}\Upsilon_{k-1},\quad k\geq 2;\qquad \mathcal{H}_{\Psi_1}^{(i)}=\overline{\gamma}_{0,1}^{(i)}H_1^{(i)}\Psi_1.$$

Proof. Theorem 1 will be proved by an innovation approach. For the i-th sensor, the innovation at time h is defined as $\mu_h^{(i)}=y_h^{(i)}-\widehat{y}_{h/h-1}^{(i)}$, where $\widehat{y}_{h/h-1}^{(i)}$ is the LS one-stage linear predictor of $y_h^{(i)}$. To simplify the notation we write

$$\xi_h^{(i)}=\gamma_{0,h}^{(i)}z_h^{(i)}+\gamma_{1,h}^{(i)}z_{h-1}^{(i)}+\gamma_{2,h}^{(i)}v_h^{(i)},\quad h\geq 2;\quad \xi_1^{(i)}=y_1^{(i)}. \tag{10}$$

Then, from (3), the innovation process can be expressed by

$$\mu_h^{(i)}=\xi_h^{(i)}-\widehat{\xi}_{h/h-1}^{(i)}-(\gamma_{3,h}^{(i)}-\overline{\gamma}_{3,h}^{(i)})y_{h-1}^{(i)},\quad h\geq 2;\qquad \mu_1^{(i)}=\xi_1^{(i)}. \tag{11}$$

Now, by denoting $\mathcal{X}_{k,h}^{(i)}=E[x_k\mu_h^{(i)T}]$, the local LS linear estimators $\widehat{x}_{k/s}^{(i)},\ s\leq k$, are expressed as linear combination of the innovations as follows:

$$\widehat{x}_{k/s}^{(i)}=\sum_{h=1}^{s}\mathcal{X}_{k,h}^{(i)}\Pi_h^{(i)-1}\mu_h^{(i)},\quad s\leq k,$$

and we start by calculating the coefficients $\mathcal{X}_{k,h}^{(i)}$ which, from (11), are given by

$$\mathcal{X}_{k,h}^{(i)}=E[x_k\xi_h^{(i)T}]-E[x_k\widehat{\xi}_{h/h-1}^{(i)T}],\quad 2\leq h\leq k.$$

- From the model assumptions, we have that $E[x_k\xi_h^{(i)T}]=A_k\mathcal{H}_{B_h}^T,\quad 1\leq h\leq k.$
- Using that $\widehat{\xi}_{h/h-1}^{(i)}=\sum_{l=1}^{h-1}E[\xi_h^{(i)}\mu_l^{(i)T}]\Pi_l^{(i)-1}\mu_l^{(i)}$, the following identity holds

$$E[x_k\widehat{\xi}_{h/h-1}^{(i)T}]=\sum_{l=1}^{h-1}\mathcal{X}_{k,l}^{(i)}\Pi_l^{(i)-1}[\overline{\gamma}_{0,h}^{(i)}\mathcal{X}_{h,l}^{(i)T}H_h^{(i)T}+\overline{\gamma}_{1,h}^{(i)}\mathcal{X}_{h-1,l}^{(i)T}H_{h-1}^{(i)T}]-\mathcal{X}_{k,h-1}^{(i)}\Pi_{h-1}^{(i)-1}\mathcal{R}_{h,h-1}^{(i)T}.$$

 ISBN: 978-84-617-8694-7

Hence, $\mathcal{X}_{k,h}^{(i)} = A_k J_h^{(i)}$, $h \leq k$, where $J_h^{(i)}$ is a function satisfying

$$J_h^{(i)} = \mathcal{H}_{B_h}^{(i)T} - \sum_{l=1}^{h-1} J_l^{(i)} \Pi_l^{(i)-1} J_l^{(i)T} \mathcal{H}_{A_h}^{(i)T} - J_{h-1}^{(i)} \Pi_{h-1}^{(i)-1} \mathcal{R}_{h,h-1}^{(i)T}, \quad h \geq 2; \quad J_1^{(i)} = \mathcal{H}_{B_1}^{(i)T}.$$

Then, by defining

$$O_k^{(i)} = \sum_{h=1}^{k} J_h^{(i)} \Pi_h^{(i)-1} \mu_h^{(i)}, \ k \geq 1; \ \ O_0^{(i)} = 0, \text{ and } r_k^{(i)} = \sum_{h=1}^{k} J_h^{(i)} \Pi_h^{(i)-1} J_h^{(i)T}, \ k \geq 1; \ \ r_0^{(i)} = 0,$$

it is easy to obtain expressions (4)-(8) of Theorem 1.

Finally, expression (9) for $\Pi_k^{(i)}$ is obtained taking into account that the innovation covariance matrix can be expressed as

$$\Pi_k^{(i)} = E[(y_k^{(i)} - \overline{\gamma}_{3,k}^{(i)} y_{k-1}^{(i)})(y_k^{(i)} - \overline{\gamma}_{3,k}^{(i)} y_{k-1}^{(i)})^T] - E[\widehat{\xi}_{k/k-1}^{(i)} \widehat{\xi}_{k/k-1}^{(i)T}],$$

and using that $\widehat{\xi}_{k/k-1}^{(i)} = \mathcal{H}_{A_k}^{(i)} O_{k-1}^{(i)} + \mathcal{R}_{k,k-1}^{(i)} \Pi_{k-1}^{(i)-1} \mu_{k-1}^{(i)}$. Then the proof is complete. □

3.2 Matrix-weighted distributed fusion filtering estimators design

Once the local LS linear filters, $\widehat{x}_{k/k}^{(i)}$ for $i = 1, \ldots, m$, have been obtained, our goal is to design a distributed fusion filter, $\widehat{x}_{k/k}$, as the matrix-weighted linear combination of such local estimators that minimizes the mean squared estimation error. The assumptions and notation in this section are those of Theorem 1.

By denoting $\widehat{X}_{k/k} = \left(\widehat{x}_{k/k}^{(1)T}, \ldots, \widehat{x}_{k/k}^{(m)T}\right)^T$ and $\mathcal{F}_k = \left(F_k^{(1)}, \ldots, F_k^{(m)}\right)$, the aim is to find $\mathcal{F}_k$ such that the estimator $\mathcal{F}_k \widehat{X}_{k/k}$ minimizes $E\left[(x_k - \mathcal{F}_k \widehat{X}_{k/k})(x_k - \mathcal{F}_k \widehat{X}_{k/k})^T\right]$. As it is well known, the solution of this problem is given by the matrix

$$\mathcal{F}_k^{opt} = E\left[x_k \widehat{X}_{k/k}^T\right] \left(E\left[\widehat{X}_{k/k} \widehat{X}_{k/k}^T\right]\right)^{-1}, \quad k \geq 1. \tag{12}$$

The following theorem provides the proposed distributed fusion filtering estimators, $\widehat{x}_{k/k}$, and their error covariance matrices, $P_{k/k}$.

Theorem 2. *Let $\widehat{X}_{k/k} = \left(\widehat{x}_{k/k}^{(1)T}, \ldots, \widehat{x}_{k/k}^{(m)T}\right)^T$ be the vector formed by the local LS estimators calculated in Theorem 1. Then, the distributed fusion filtering estimators $\widehat{x}_{k/k}$, $k \geq 1$, and their error covariance matrices, $P_{k/k}$, $k \geq 1$, are given by*

$$\widehat{x}_{k/k} = \left(\Sigma_{k/k}^{\widehat{x}^{(1)}}, \ldots, \Sigma_{k/k}^{\widehat{x}^{(m)}}\right) \left(\Sigma_{k/k}^{\widehat{X}}\right)^{-1} \widehat{X}_{k/k}, \quad k \geq 1, \tag{13}$$

 ISBN: 978-84-617-8694-7

$$P_{k/k} = A_k B_k^T - \left(\Sigma_{k/k}^{\widehat{x}^{(1)}}, \ldots, \Sigma_{k/k}^{\widehat{x}^{(m)}}\right) \left(\Sigma_{k/k}^{\widehat{X}}\right)^{-1} \left(\Sigma_{k/k}^{\widehat{x}^{(1)}}, \ldots, \Sigma_{k/k}^{\widehat{x}^{(m)}}\right)^T, \quad k \geq 1. \tag{14}$$

where $\Sigma_{k/k}^{\widehat{X}} = \left(\Sigma_{k/k}^{\widehat{x}^{(ij)}}\right)_{i,j=1,\ldots,m}$, *and the cross-correlation matrices* $\Sigma_{k/k}^{\widehat{x}^{(ij)}} = E\left[\widehat{x}_{k/k}^{(i)} \widehat{x}_{k/k}^{(j)T}\right]$, $i, j = 1, \ldots, m$, *are obtained by*

$$\Sigma_{k/k}^{\widehat{x}^{(ij)}} = A_k r_k^{(ij)} A_k^T, \quad k \geq 1, \tag{15}$$

with $r_k^{(ij)} = E[O_k^{(i)} O_k^{(j)T}]$ *satisfying*

$$r_k^{(ij)} = r_{k-1}^{(ij)} + J_{k-1,k}^{(ij)} \Pi_k^{(j)-1} J_k^{(j)T} + J_k^{(i)} \Pi_k^{(i)-1} J_k^{(ji)T}, \quad k \geq 1; \quad r_0^{(ij)} = 0. \tag{16}$$

The matrices $J_k^{(ij)} = E[O_k^{(i)} \mu_k^{(j)T}]$ *are given by*

$$J_k^{(ij)} = J_{k-1,k}^{(ij)} + J_k^{(i)} \Pi_k^{(i)-1} \Pi_k^{(ij)}, \quad k \geq 1, \tag{17}$$

and $J_{k-1,k}^{(ij)} = E[O_{k-1}^{(i)} \mu_k^{(j)T}]$ *are calculated by*

$$\begin{array}{l} J_{k-1,k}^{(ij)} = \left(r_{k-1}^{(i)} - r_{k-1}^{(ij)}\right) \overline{\mathcal{H}}_{A_k}^{(j)T} + J_{k-1}^{(i)} \Pi_{k-1}^{(i)-1} \mathcal{R}_{k,k-1}^{(ji)T} - J_{k-1}^{(ij)} \Pi_{k-1}^{(j)-1} \mathcal{R}_{k,k-1}^{(j)T}, \quad k \geq 2; \\ J_{0,1}^{(ij)} = 0. \end{array} \tag{18}$$

The innovation cross-correlation matrices, $\Pi_k^{(ij)} = E[\mu_k^{(i)} \mu_k^{(j)T}]$, *are given by*

$$\begin{array}{l} \Pi_k^{(ij)} = \Sigma_k^{\xi^{(ij)}} - \overline{\mathcal{H}}_{A_k}^{(i)} \left(\overline{\mathcal{H}}_{B_k}^{(j)T} - J_k^{(j)} - J_{k-1,k}^{(ij)}\right) \\ \qquad\qquad - \mathcal{R}_{k,k-1}^{(ij)} \Pi_{k-1}^{(j)-1} \left(\overline{\mathcal{H}}_{A_k}^{(j)} J_{k-1}^{(j)} + \mathcal{R}_{k,k-1}^{(j)}\right)^T - \mathcal{R}_{k,k-1}^{(i)} \Pi_{k-1}^{(i)-1} \Pi_{k-1,k}^{(ij)}, \quad k \geq 2; \\ \Pi_1^{(ij)} = \Sigma_1^{\xi^{(ij)}}, \end{array} \tag{19}$$

where $\Sigma_k^{\xi^{(ij)}} = E[\xi_k^{(i)} \xi_k^{(j)T}]$ *satisfy*

$$\begin{array}{l} \Sigma_k^{\xi^{(ij)}} = \overline{\gamma}_{0,k}^{(i)} \Sigma_k^{z\xi^{(ij)}} + \overline{\gamma}_{1,k}^{(i)} \Sigma_{k-1,k}^{z\xi^{(ij)}} + \overline{\gamma}_{2,k}^{(i)} (\overline{\gamma}_{0,k}^{(j)} + \overline{\gamma}_{2,k}^{(j)}) R_k^{(ij)}, \quad k \geq 2; \\ \Sigma_1^{\xi^{(ij)}} = \overline{\gamma}_{0,1}^{(i)} \Sigma_1^{z\xi^{(ij)}} + \overline{\gamma}_{2,1}^{(i)} (\overline{\gamma}_{0,1}^{(j)} + \overline{\gamma}_{2,1}^{(j)}) R_1^{(ij)}, \end{array}$$

with $\Sigma_{s,k}^{z\xi^{(ij)}} = E[z_s^{(i)} \xi_k^{(j)T}]$, *for* $s = k-1,\ k$, *given by*

$$\Sigma_{s,k}^{z\xi^{(ij)}} = \overline{\gamma}_{0,k}^{(j)} \Sigma_{s,k}^{z^{(ij)}} + \overline{\gamma}_{1,k}^{(j)} \Sigma_{s,k-1}^{z^{(ij)}} + \overline{\gamma}_{2,k}^{(j)} R_k^{(ij)} \delta_{s,k}, \quad k \geq 2; \qquad \Sigma_1^{z\xi^{(ij)}} = \overline{\gamma}_{0,1}^{(j)} \Sigma_1^{z^{(ij)}} + \overline{\gamma}_{2,1}^{(j)} R_1^{(ij)}.$$

The matrices $\Pi_{k-1,k}^{(ij)} = E[\mu_{k-1}^{(i)} \mu_k^{(j)T}]$ *are calculated by*

$$\Pi_{k-1,k}^{(ij)} = \overline{\mathcal{H}}_{A_k}^{(i)} \left(J_{k-1}^{(j)} - J_{k-1}^{(ij)}\right) + \mathcal{R}_{k,k-1}^{(ij)} - \mathcal{R}_{k-1}^{(i)} \Pi_{k-1}^{(i)-1} \Pi_{k-1}^{(ij)}, \quad k \geq 2. \tag{20}$$

 ISBN: 978-84-617-8694-7

Finally, $\mathcal{R}_{k,k-1}^{(ij)} = \overline{\gamma}_{1,k}^{(i)}(\overline{\gamma}_{0,k-1}^{(j)} + \overline{\gamma}_{2,k-1}^{(j)})R_{k-1}^{(ij)},\ k \geq 2$, *and* $\Sigma_{k,s}^{z^{(ij)}}$ *are given in (2).*

Proof. From Theorem 1, it is not difficult to check that $\widehat{\xi}_{k/k-1}^{(i/j)}$, the estimator of $\xi_k^{(i)}$ based on the measurements $\left\{y_1^{(j)}, \ldots, y_{k-1}^{(j)}\right\}$, is given by

$$\widehat{\xi}_{k/k-1}^{(i/j)} = \mathcal{H}_{A_k}^{(i)} O_{k-1}^{(j)} + \mathcal{R}_{k,k-1}^{(ij)} \Pi_{k-1}^{(j)-1} \mu_{k-1}^{(j)}, \quad k \geq 2. \tag{21}$$

- Expressions (13) and (14) for the distributed estimators and their error covariance matrices, respectively, are immediately derived from (12).

- Expression (15) for the cross-correlation matrices between local estimators, $\Sigma_{k/k}^{\widehat{x}^{(ij)}}$, follows easily from (4), since $r_k^{(ij)} = E[O_k^{(i)} O_k^{(j)T}]$; moreover, using (5) and taking into account that $J_{s,k}^{(ij)} = E[O_s^{(i)} \mu_k^{(j)T}]$, for $s = k-1,\ k$, we get (16) for $r_k^{(ij)}$.

- Using (5) for $O_k^{(i)}$, expression (17) for $J_k^{(ij)} = E[O_k^{(i)} \mu_k^{(j)T}]$ is directly obtained.

- To derive (18) for $J_{k-1,k}^{(ij)} = E[O_{k-1}^{(i)} \mu_k^{(j)T}]$, we use (11) for $\mu_k^{(j)}$, which leads to $J_{k-1,k}^{(ij)} = E[O_{k-1}^{(i)} \xi_k^{(j)T}] - E[O_{k-1}^{(i)} \widehat{\xi}_{k/k-1}^{(j)T}]$. Now, from the Orthogonal Projection Lemma (OPL), we have $E[O_{k-1}^{(i)} \xi_k^{(j)T}] = E[O_{k-1}^{(i)} \widehat{\xi}_{k/k-1}^{(j/i)T}]$ and, using (21), it is easy to prove (18).

- If we use expression (11) in $\Pi_k^{(ij)} = E[\mu_k^{(i)} \mu_k^{(j)}]$, it is not difficult to see that

$$\Pi_k^{(ij)} = \Sigma_k^{\xi^{(ij)}} - E[\xi_k^{(i)} \widehat{\xi}_{k/k-1}^{(j)T}] - E[\widehat{\xi}_{k/k-1}^{(i)} \mu_k^{(j)T}], \quad k \geq 2. \tag{22}$$

  From the OPL, we have $E[\xi_k^{(i)} \widehat{\xi}_{k/k-1}^{(j)T}] = E[\widehat{\xi}_{k/k-1}^{(i/j)} \xi_k^{(j)T}]$ and, from (21) and (6), it follows that

$$E[\widehat{\xi}_{k/k-1}^{(i/j)} \xi_k^{(j)T}] = \overline{\mathcal{H}}_{A_k}^{(i)} \left(\overline{\mathcal{H}}_{B_k}^{(j)T} - J_k^{(j)}\right) + \mathcal{R}_{k,k-1}^{(ij)} \Pi_{k-1}^{(j)-1} \left(\overline{\mathcal{H}}_{A_k}^{(j)} J_{k-1}^{(j)} + \mathcal{R}_{k,k-1}^{(j)}\right)^T.$$

  Also from (21), the following identity holds

$$E[\widehat{\xi}_{k/k-1}^{(i)} \mu_k^{(j)T}] = \overline{\mathcal{H}}_{A_k}^{(i)} J_{k-1,k}^{(ij)} + \mathcal{R}_{k,k-1}^{(i)} \Pi_{k-1}^{(i)-1} \Pi_{k-1,k}^{(ij)}.$$

  Substituting the above expectations into (22), we immediately get (19).

- Using again expression (11) for $\mu_k^{(j)}$ in $\Pi_{k-1,k}^{(ij)} = E[\mu_{k-1}^{(i)} \mu_k^{(j)}]$, we have $\Pi_{k-1,k}^{(ij)} = E[\mu_{k-1}^{(i)} \xi_k^{(j)T}] - E[\mu_{k-1}^{(i)} \widehat{\xi}_{k/k-1}^{(j)T}]$, and applying the OPL and (21), we can show that expression (20) for $\Pi_{k-1,k}^{(ij)}$ holds true.

The proof of Theorem 3 is then complete. □

4 Numerical simulation example

Consider a zero-mean two-dimensional signal $\{x_k;\ k \geq 1\}$ whose autocovariance function is given by

$$E[x_k x_s^T] = \begin{pmatrix} 0.8^{k-s} & 1.02 \times 0.8^{k-s} \\ 0.9 \times 0.8^{k-s-1} & 0.918 \times 0.8^{k-s-1} \end{pmatrix}, \quad s < k,$$

which is factorizable just taking

$$A_k = \begin{pmatrix} 0.8^k & 1.02 \times 0.8^k \\ 0.9 \times 0.8^{k-1} & 0.918 \times 0.8^{k-1} \end{pmatrix} \text{ and } B_s = \begin{pmatrix} 0.8^{-s} & 0 \\ 0 & 0.8^{-s} \end{pmatrix}.$$

Consider four sensors which provide scalar measurements of the signal according to measured outputs model (1): $z_k^{(i)} = H_k^{(i)} x_k + v_k^{(i)}, \quad k \geq 1, \quad i = 1, 2, 3, 4$, where $H_k^{(1)} = (0.74,\ 0.75)$, $H_k^{(2)} = (0.65,\ 0.67)$, $H_k^{(3)} = (0.75,\ 0.7)$ and $H_k^{(4)} = (0.95,\ 0.9)$. The additive noise processes $\{v_k^{(i)};\ k \geq 1\}$, $i = 1, 2, 3, 4$, are defined as $v_k^{(i)} = c_i \eta_k$, where $c_1 = c_2 = 1$, $c_3 = 0.75$ and $c_4 = 1.5$, and $\{\eta_k; k \geq 1\}$ is a zero-mean Gaussian white process with unit variance. Clearly, these noises $\{v_k^{(i)}; k \geq 1\}$, $i = 1, 2, 3$, are correlated at any sampling time, with $R_k^{(ij)} = c_i c_j$.

Next, according to the theoretical observation model, suppose that random one-step delays, packet dropouts and uncertain observations, with different rates exist in the data transmissions from the individual sensors to the local processors. Specifically, let us consider the observation model (3):

$$\begin{aligned} y_k^{(i)} &= \gamma_{0,k}^{(i)} z_k^{(i)} + \gamma_{1,k}^{(i)} z_{k-1}^{(i)} + \gamma_{2,k}^{(i)} v_k^{(i)} + \gamma_{3,k}^{(i)} y_{k-1}^{(i)},\ k \geq 2; \\ y_1^{(i)} &= \gamma_{0,1}^{(i)} z_1^{(i)} + \gamma_{2,1}^{(i)} v_1^{(i)}, \end{aligned}$$

where, for $i = 1, 2, 3, 4$, and $d = 0, 1, 2, 3$, $\{\gamma_{d,k}^{(i)};\ k \geq 1\}$ are sequences of independent Bernoulli variables with the following probabilities:

$k = 1$

sensor	$\overline{\gamma}_{0,1}^{(i)}$	$\overline{\gamma}_{1,1}^{(i)}$	$\overline{\gamma}_{2,1}^{(i)}$	$\overline{\gamma}_{3,1}^{(i)}$
$i = 1$	0.8	0	0	0.2
$i = 2$	0.75	0	0	0.25
$i = 3$	0.75	0	0	0.25
$i = 4$	0.8	0	0	0.2

$k \geq 2$

sensor	$\overline{\gamma}_{0,k}^{(i)}$	$\overline{\gamma}_{1,k}^{(i)}$	$\overline{\gamma}_{2,k}^{(i)}$	$\overline{\gamma}_{3,k}^{(i)}$
$i = 1$	0.3	0.25	0.25	0.2
$i = 2$	0.25	0.3	0.2	0.25
$i = 3$	0.3	0.2	0.25	0.25
$i = 4$	0.25	0.25	0.3	0.2

To illustrate the feasibility and effectiveness of the estimators proposed in the current paper, the algorithms were implemented in MATLAB, and fifty iterations were run. Using simulated values, both the centralized and the proposed distributed fusion estimates were calculated, as well as the corresponding error covariance matrices in order to measure the

 ISBN: 978-84-617-8694-7

estimators accuracy. The derivation of the centralized filtering algorithm is omitted since it is totally analogous to that of the local filtering algorithm (Therorem 1).

For the first and second signal components, Figure 1 displays simulated trajectories along with the centralized and distributed fusion filtering estimations, and Figure 2 the error variances of the different filtering estimators (local, centralized and distributed).

From Figure 1, a satisfactory and efficient tracking performance of the proposed distributed fusion filtering estimators is observed. Figure 2 compares the filtering error variances of the local filters and the fusion filters and shows that the error variances of the distributed fusion filtering estimators are significantly smaller than those of every local estimators, but lightly greater than those of the centralized ones; nevertheless, this slight difference is compensated by the fact that the distributed fusion structure reduces the computational cost and has better robustness and fault tolerance. Consequently, agreeing with the theoretical results, the distributed fusion filter has better accuracy than the local filters and the centralized fusion filter outperforms them all, as it is the optimal one.

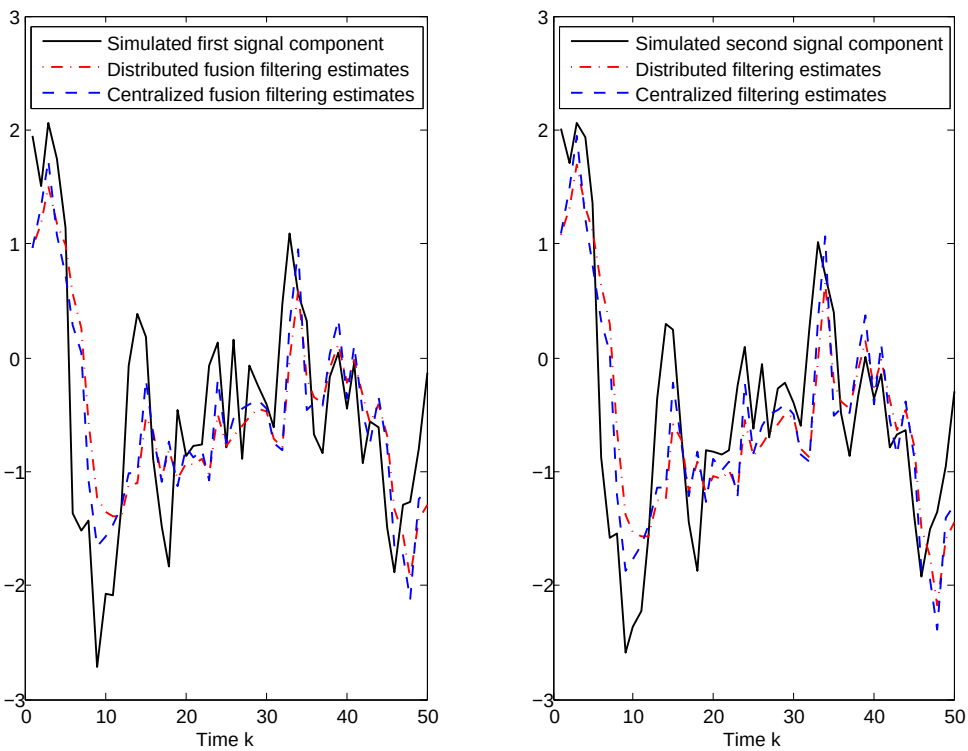


Figure 1: Simulated first and second signal components and centralized and distributed fusion filtering estimates.

 ISBN: 978-84-617-8694-7

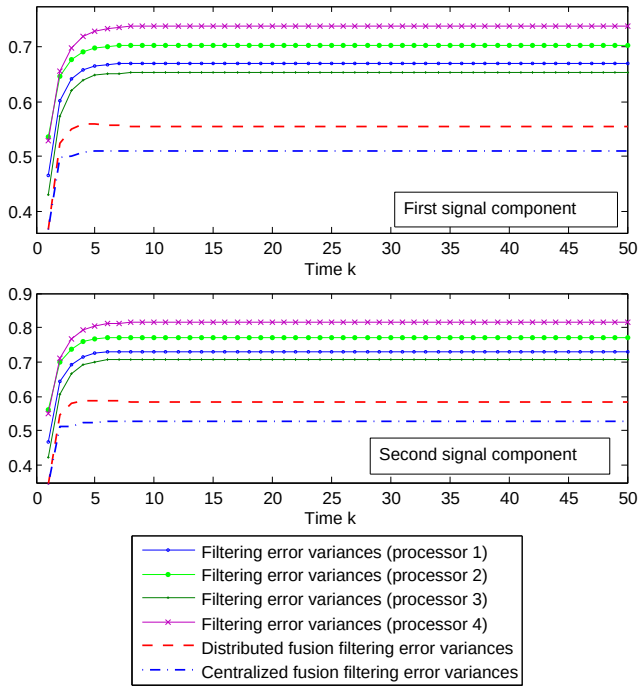


Figure 2: Error variance comparison of the centralized, distributed and local filters.

5 Conclusion

This paper makes valuable contributions to the estimation problem in networked stochastic systems featuring three types of random transmission uncertainties: missing measurements, one-step delays and packet dropouts. By an innovation approach, a recursive distributed filtering algorithm has been designed, which is easily implementable and does not require the signal evolution model, but only the mean and covariance functions of the system processes.

Acknowledgements

This research is supported by *Ministerio de Economía y Competitividad* and *Fondo Europeo de Desarrollo Regional FEDER* (grant no. MTM2014-52291-P).

References

[1] J. Feng and M. Zeng, *Optimal distributed Kalman filtering fusion for a linear dynamic system with cross-correlated noises*, Int. J. Syst. Sci. **43(2)** (2012) 385–398.

[2] L. Yan, X. Rong Li, Y. Xia and M. Fu, *Optimal sequential and distributed fusion for state estimation in cross-correlated noise*, Automatica **49** (2013) 3607–3612.

[3] R. Caballero-Águila, A. Hermoso-Carazo and J. Linares-Pérez, *Least-squares linear estimators using measurements transmitted by different sensors with packet dropouts*, Digit. Signal Process. **22(6)** (2012) 1118–1125.

[4] S. Gao and P. Chen, *Suboptimal filtering of networked discrete-time systems with random observation losses*, Math. Probl. Eng. **2014** (2014) art. no.151836.

[5] Y. Liu, X. He, Z. Wang and D. Zhou, *Optimal filtering for networked systems with stochastic sensor gain degradation*, Automatica **50(5)** (2014) 1521–1525.

[6] J. Ma and S. Sun, *Centralized fusion estimators for multisensor systems with random sensor delays, multiple packet dropouts and uncertain observations*, IEEE Sensors J. **13(4)** (2013) 1228–1235.

[7] N. Li, S. Sun and J. Ma, *Multi-sensor distributed fusion filtering for networked systems with different delay and loss rates*, Digit. Signal Process. **34** (2014) 29–38.

[8] B. Chen, W. Zhang and L. Yu, *Distributed fusion estimation with missing measurements, random transmission delays and packet dropouts*, IEEE Trans. Automat. Contr. **59(7)** (2014) 1961–1967.

[9] B. Chen, W. Zhang and L. Yu, *Networked fusion Kalman filtering with multiple uncertainties*, IEEE Trans. Aerosp. Electron. Syst. **51(3)** (2015) 2332–2349.

[10] W. Li, Z. Wang, G. Wei, L. Ma, J. Hu and D. Ding, *A Survey on multisensor fusion and consensus filtering for sensor networks*, Discrete Dyn. Nat. Soc. **2015** (2015) Article ID 683701.

[11] R. Caballero-Águila, A. Hermoso-Carazo and J. Linares-Pérez, *Fusion estimation using measured outputs with random parameter matrices subject to random delays and packet dropouts*, Signal Process. **127** (2016) 12–23.

[12] R. Caballero-Águila, A. Hermoso-Carazo and J. Linares-Pérez, *Distributed fusion filters from uncertain measured outputs in sensor networks with random packet losses*, Inform. Fusion, **34** (2017) 70–79.

[13] J. Ma and S. Sun, *Distributed fusion filter for networked stochastic uncertain systems with transmission delays and packet dropouts*, Signal Process. **130** (2017) 268–278.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

Relation-based Galois-connections: towards the residual of a relation

Inma P. Cabrera[1], Pablo Cordero[1] and Manuel Ojeda-Aciego[1]

[1] *Departamento de Matemática Aplicada, Universidad de Málaga. Spain*

emails: ipcabrera@uma.es, pcordero@uma.es, aciego@uma.es

Abstract

We explore a suitable generalization of the notion of Galois connection in which their components are binary relations. Many different approaches are possible depending both on the (pre-)order relation between subsets in the underlying powerdomain and the chosen type of relational composition.

Key words: Galois connection, relations

1 Introduction

Galois connections can be identified in lots of situations, and they have shown to be an interesting tool both for theory and for applications [6, 9]. Recent applications can still be found in different topics, specially in the realm of Formal Concept Analysis and the foundations of Fuzzy Set Theory, see for instance [1, 7].

In all the different notions of generalized Galois connection the problem of its construction is of capital importance: specifically, given a mapping $f\colon A \to B$ how can one obtain its residual (aka right adjoint). Freyd's adjoint theorem characterizes when such a residual exists when both A and B have the same structure. One of our recent research topics has been to study what happens if A and B are differently structured, and a number of results have been obtained considering different underlying settings. Namely, in [10] we worked with crisp functions between a poset (resp. preordered set) and an unstructured set; later, in [2] we entered in the fuzzy arena, considering the case in which A is a fuzzy preposet; then, in [3], we extended the previous results by allowing fuzzy equivalence relations as an adequate substitute to equality. In [4] we introduced the notion of relational fuzzy Galois connection, in which the components of the connection are not fuzzy functions but fuzzy

 ISBN: 978-84-617-8694-7

relations satisfying certain properties. Before proceeding to further generalizations, it is worth to start again from the crisp case and consider a more adequate notion of Galois connection whose components are crisp relations, and that is the topic of the present paper.

2 Preliminary definitions

A binary relation R between two sets A and B is a subset of the Cartesian product $A \times B$ and it can be also seen as a multivalued function R from the set A to the powerset $\mathcal{P}(B)$. For an element $(a, b) \in R$, it is said that a is related to b and denoted aRb.

Definition 1 *Given a binary relation $R \subseteq A \times B$,*

- *the* afterset *aR of an element $a \in A$ is defined as $\{b \in B \colon aRb\}$*
- *the* foreset *Rb of an element $b \in B$ is defined as $\{a \in A \colon aRb\}$*
- *the* domain *of R is the set* $\mathrm{Dom}(R) = \{a \in A \colon aR \neq \varnothing\}$
- *the* range *of R is the set* $\mathrm{Rng}(R) = \{b \in B \colon Rb \neq \varnothing\}$

Definition 2 *Given a binary relation $R \subseteq A \times B$ and a subset $X \subseteq A$,*

- *the* direct image *of X under R is*

$$R(X) = XR = \{b \in B \colon \text{ there exists } x \in X \text{ such that } xRb\} = \bigcup_{x \in X} xR$$

  The direct image of a subset X is the set of those elements of B that are related to at least *one element of X.*

- *the* subdirect image *of X under R is*

$$R^{\triangleleft}(X) = X^{\triangleleft}R = \{b \in B \colon xRb \text{ for all } x \in X\} = \bigcap_{x \in X} xR$$

  The subdirect image of a subset X is the set of those elements of B that are related to all *the elements of X.*

Observe that if $X = \{a\}$, the direct image and the subdirect image of X coincides with the afterset, thus aR coincides with $R(a)$ should we interpret R as a multivalued mapping.

In the realm of ordered structures, Ore [12] introduced in 1944 the so-called *Galois connections* as a pair of antitone mappings for which both possible compositions are inflationary. In order to extend this concept to binary relations it is necessary to fix the meaning of antitone and inflationary relation. One possibility is to consider the relation as a multivalued function and extend it naturally, but both the preordered structure in the powerset and the composition of relations admit different approaches.

 ISBN: 978-84-617-8694-7

Definition 3 *Given P an arbitrary set and $\leq$ a preorder (reflexive and transitive relation) defined over P, it is possible to lift the preorder structure to the powerset $\mathcal{P}(P)$ by defining*

$$X \ll Y \iff \text{for all } x \in X \text{ there exists } y \in Y \text{ such that } x \leq y \quad (1)$$

$$X \Subset Y \iff \text{for all } y \in Y \text{ there exists } x \in X \text{ such that } x \leq y \quad (2)$$

$$X \ltimes Y \iff \text{for all } x \in X \text{ there exists } y \in Y \text{ such that } x \leq y \text{ and} \quad (3)$$
$$\text{for all } y \in Y \text{ there exists } x \in X \text{ such that } x \leq y$$

Note that the relations defined above are the preorder relations used in the construction of the, respectively, Hoare, Smyth, and Plotkin powerdomains.

The concept of antitone multivalued function between two preordered sets depends on the preorder relation considered over the powerset. We explore three possibilities, being one approach more restrictive than the others:

Definition 4 *Let $\mathbb{A} = (A, \leq_A), \mathbb{B} = (B, \leq_B)$ be two preordered sets and $R \subseteq A \times B$ a binary relation between A and B.*

- *R is said to be* h-antitone *if $a_1 \leq_A a_2$ implies $a_2R \ll a_1R$, for all $a_1, a_2 \in A$.*
- *R is said to be* s-antitone *if $a_1 \leq_A a_2$ implies $a_2R \Subset a_1R$, for all $a_1, a_2 \in A$.*
- *R is said to be* p-antitone *if $a_1 \leq_A a_2$ implies $a_2R \ltimes a_1R$, for all $a_1, a_2 \in A$, equivalently, if it is h-antitone and s-antitone.*

Analogously, the definition of inflationary multivalued function also admits three possibilities.

Definition 5 *Let $\mathbb{A} = (A, \leq_A)$ be a preordered set and $R \subseteq A \times A$ a binary relation on A.*

- *R is said to be* h-inflationary *if $\{a\} \ll aR$, for all $a \in A$, that is, there exists (at least) $x \in aR$ such that $a \leq x$.*
- *R is said to be* s-inflationary *if $\{a\} \Subset aR$, for all $a \in A$, that is, $a \leq x$ for all $x \in aR$.*
- *R is said to be* p-inflationary *if it is h-inflationary and s-inflationary.*

Remark 1 *Notice that if a relation is s-inflationary, then it is also h-inflationary, therefore p-inflationary and s-inflationary are equivalent notions.*

The condition of the composition of two multivalued mappings being inflationary in some sense requires also to fix which definition of composition will be used.

Definition 6 *Let R be a binary relation between A and B and S be a binary relation between B and C.*

 ISBN: 978-84-617-8694-7

- *The* classical *composition of R and S is defined as follows*

$$\begin{aligned} R \circ S &= \{(x,z) \in A \times C\colon \text{ there exists } b \in B \text{ such that } xRb \text{ and } bSz\} \\ &= \{(x,z) \in A \times C\colon xR \cap Sz \neq \varnothing\} \end{aligned}$$

- *The $\triangleleft$ composition of R and S is defined as follows*

$$\begin{aligned} R \triangleleft S &= \{(x,z) \in A \times C\colon \text{ for all } b \in B \text{ such that } xRb \text{ it holds that } bSz\} \\ &= \{(x,z) \in A \times C\colon xR \subseteq Sz\} \end{aligned}$$

Observe that for an element $a \in A$, the afterset $a[R \circ S]$ coincides with the direct image of the afterset aR under S, that is

$$a[R \circ S] = (aR)S = \bigcup_{b \in aR} bS$$

Analogously, for an element $a \in A$, the afterset $a[R \triangleleft S]$ coincides with the subdirect image of the afterset aR under S, that is

$$a[R \triangleleft S] = (aR)^{\triangleleft}S = \bigcap_{b \in aR} bS$$

3 Two types of relational Galois connections

Definition 7 *Let $\mathbb{A} = (A, \leq_A), \mathbb{B} = (B, \leq_B)$ be two posets, $R \subseteq A \times B$ a binary relation between A and B and $S \subseteq B \times A$ a binary relation between B and A.*

The pair (R, S) is said to be an s-Galois connection *between $\mathbb{A}$ and $\mathbb{B}$ if*

i) R and S are s-antitone.

ii) $R \circ S$ and $S \circ R$ are s-inflationary.

Proposition 1 *Let $\mathbb{A} = (A, \leq_A), \mathbb{B} = (B, \leq_B)$ be two posets and (R, S) be an* s-Galois connection *between $\mathbb{A}$ and $\mathbb{B}$. Then if $b \in \mathrm{Rng}(R)$ then bS is at most a singleton, so, the restriction of S to $\mathrm{Rng}(R)$ is a (partial) single-valued function.*

Proof: If $b \in \mathrm{Rng}(R) \smallsetminus \mathrm{Dom}(S)$, then there is nothing to prove; therefore, let us assume that $b \in \mathrm{Rng}(R) \cap \mathrm{Dom}(S)$.

1. As $b \in \mathrm{Rng}(R)$, there exists $a \in A$ such that $b \in aR$ and, as $b \in \mathrm{Dom}(S)$, we have that bS is nonempty. We will now see that, $b \in xR$ for all $x \in bS$.

   Since $\{a\} \Subset a[R \circ S]$, taking into account that $b \in aR$, we have $a \leq x$ for all $x \in bS$. Now, as R is s-antitone, there exists $b' \in xR$ such that $b' \leq b$. The other inequality $b \leq b'$ follows because of $\{b\} \Subset b[S \circ R]$. As a result, $b = b' \in xR$.

 ISBN: 978-84-617-8694-7

2. Consider two elements $x, x^\star \in bS$, by definition of composition and the previous item, we have $x^\star \in x[R \circ S]$; since, by hypothesis, we have $\{x\} \Subset x[R \circ S]$, it turns out that $x \leq x^\star$. Applying that R is s-antitone, there exists $b^\star \in x^\star R$ such that $b^\star \leq b$. Again, the hypothesis $\{b\} \Subset b[S \circ R]$ implies $b \leq b^\star$, and we obtain $b = b^\star \in x^\star R$. This, together with $x \in bS$, proves that $x \in x^\star[R \circ S]$; finally, applying once again $x^\star \Subset x^\star[R \circ S]$, we obtain $x^\star \leq x$ and, therefore $x^\star = x$ and bS is a singleton. □

Notice that the previous result shows that the definition of s-Galois connection necessarily collapses the relations R and S to be (partial) functions in the case of posets. If we drop the antisymmetry and consider the more general case of preordered sets, we obtain a similar result in that the images of the relations are clusters.

As a result, it seems more convenient to consider alternative approaches either by changing the ordering between subsets and/or slightly modifying the notions of antitone or inflationary relation.

A promising definition seems to be the following:

Definition 8 *Let $\mathbb{A} = (A, \leq_A), \mathbb{B} = (B, \leq_B)$ be two preordered sets, $R \subseteq A \times B$ a binary relation between A and B and $S \subseteq B \times A$ a binary relation between B and A. The pair (R, S) is said to be an* h-Galois connection *between $\mathbb{A}$ and $\mathbb{B}$ if*

i) R and S are h-antitone.

ii) For all $a \in A$ and all $b \in B$ the following conditions hold:

$$\{a\} \ll yS \text{ for all } y \in aR \quad \text{and} \quad \{b\} \ll xR \text{ for all } x \in bS. \tag{4}$$

It is not difficult to check that condition (4) above is a consequence of the property of $R \triangleleft S$ and $S \triangleleft R$ being h-inflationary but, in general, are not equivalent.

The following result shows a necessary condition for a pair (R, S) to be an h-Galois connection.

Lemma 1 *If the pair (R, S) is an h-Galois connection between $\mathbb{A}$ and $\mathbb{B}$, then the following inclusions hold:* $\mathrm{Rng}(R) \subseteq \mathrm{Dom}(S)$ *and* $\mathrm{Rng}(S) \subseteq \mathrm{Dom}(R)$.

Proof: Given $b \in \mathrm{Rng}(R)$, there exists $a \in A$ such that $b \in aR$. Now, by condition (4) above, we obtain that $bS \neq \varnothing$ and, therefore, $b \in \mathrm{Dom}(S)$.

The other inclusion can be proved similarly. □

Notice that, in fact, the proof of the previous lemma does not use the antitonicity of either R or S.

With the condition of Lemma 1 in mind, we obtain an equivalence with the usual notion of Galois connection, as stated below:

 ISBN: 978-84-617-8694-7

Theorem 1 *Let $\mathbb{A} = (A, \leq_A), \mathbb{B} = (B, \leq_B)$ be two preordered sets, $R \subseteq A \times B$ a binary relation between A and B and $S \subseteq B \times A$ a binary relation between B and A. The pair (R, S) is an h-Galois connection between $\mathbb{A}$ and $\mathbb{B}$ if and only if the following holds:*

$$\mathrm{Rng}(R) \subseteq \mathrm{Dom}(S) \quad \text{and} \quad \mathrm{Rng}(S) \subseteq \mathrm{Dom}(R) \tag{5}$$

$$\{a\} \ll bS \iff \{b\} \ll aR \tag{6}$$

Proof: Given (R, S) an h-Galois connection between $\mathbb{A}$ and $\mathbb{B}$, condition (5) follows by Lemma 1. Now, for condition (6), assume that $\{a\} \ll bS$. Then, there exists $x \in bS$ such that $a \leq x$ and, by R h-antitone, we obtain $xR \ll aR$. On the other hand, by condition (4) we have $\{b\} \ll xR$. Now, by transitivity of $\ll$, we obtain that $\{b\} \ll aR$. The proof that $\{b\} \ll aR$ implies $\{a\} \ll bS$ is similar.

Conversely, assume that equivalence (6) holds, and let us prove that (R, S) is an h-Galois connection.

Firstly, we will show condition (4): Given $a \in A$ and $y \in aR$, since $y \leq y$, then $\{y\} \ll aR$ which by (6) implies that $\{a\} \ll yS$ for all $y \in aR$. The other part is similar.

Now, consider $a_1 \leq a_2$ in A. Then, since $\{a_2\} \ll yS$ for all $y \in a_2R$, it also holds that $\{a_1\} \ll yS$ for all $y \in a_2R$. Hence, by (6), we have $\{y\} \ll a_1R$, for all $y \in a_2R$ which means that $a_2R \ll a_1R$. The antitonicity of S follows analogously. □

4 Conclusions and further work

The problem of considering relations within the notion of Galois connection is not new, since it can be dated back to [8], nor outdated, since one can still find recent references dealing with different aspects of the integration of relations and Galois connections, see for instance [5, 11, 13].

We have obtained some prospective results on the notion of relational-based Galois connection, in which the components of the connection are relations between posets. There are several possibilities depending both on the (pre-)order relation between subsets in the underlying powerdomain and the chosen type of relational composition. We have just scratched the surface of the problem, and shown that one of the most reasonable approaches collapses in that the involved relations R and S should actually be functions. The second proposed definition uses a different approach in that, apart from considering an alternative ordering in the underlying powerdomain for the definition of antitonicity, it also generalizes the notion of being inflationary. This way, we have obtained a promising result in the form of Theorem 1.

As future work, we are planning to continue the line initiated in [2, 3] and attempt the construction of the residual, in the sense of relation-based (fuzzy) Galois connections, to a given mapping between differently structured domain and codomain, as stated in the introduction.

 ISBN: 978-84-617-8694-7

References

[1] L. Antoni, S. Krajči, and O. Krídlo, Representation of fuzzy subsets by Galois connections. *Fuzzy Sets and Systems*, 2017. https://doi.org/10.1016/j.fss.2017.05.020.

[2] I.P. Cabrera, P. Cordero, F. García-Pardo, M. Ojeda-Aciego, and B. De Baets. On the construction of adjunctions between a fuzzy preposet and an unstructured set, *Fuzzy Sets and Systems* 2017. To appear. http://dx.doi.org/10.1016/j.fss.2016.09.013

[3] ——, Adjunctions between a fuzzy preposet and an unstructured set with underlying fuzzy equivalence relations, 2017, submitted.

[4] I.P. Cabrera, P. Cordero, and M. Ojeda-Aciego. Relational fuzzy Galois connections. In *Proc. of the 17th World Congress of Intl Fuzzy Systems Association (IFSA-SCIS'17)*, 2017. To appear.

[5] M. Couceiro. Galois connections for generalized functions and relational constraints. In *Contributions to General Algebra*, volume 16, pages 35–54. Verlag Johannes Heyn, 2005.

[6] K. Denecke, M. Erné, and S. Wismath. *Galois connections and applications*, ser. Mathematics and its Applications. Springer, 2004, vol. 565.

[7] J.T. Denniston, A. Melton, S.E. Rodabaugh. Formal Contexts, Formal Concept Analysis, and Galois Connections. *Electr. Proc. on Theoretical Computer Science* 129: 105–120, 2013.

[8] I. Fleischer and I. Rosenberg. The Galois connection between partial functions and relations. *Pacific Journal of Mathematics*, 79(1):93–97, 1978.

[9] F. García-Pardo, I.P. Cabrera, P. Cordero, and M. Ojeda Aciego. On Galois connections and soft computing, *Lect. Notes in Computer Science*, vol. 7903, pp. 224–235, 2013.

[10] F. García-Pardo, I.P. Cabrera, P. Cordero, M. Ojeda Aciego, and F. Rodríguez. On the definition of suitable orderings to generate adjunctions over an unstructured codomain, *Information Sciences*, vol. 286, pp. 173–187, 2014.

[11] E. Jeřábek. Galois connection for multiple-output operations. https://arxiv.org/abs/1612.04353v1, dec 2016.

[12] O. Ore. Galois Connexions. *Transactions of the American Mathematical Society*, 55(3):493–513, 1944.

[13] Á. Száz. A particular Galois connection between relations and set functions. *Acta Univ. Sapientiae, Mathematica*, 6(1):73–91, 2014.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

Matrices related to orthogonal hypercomplex polynomial systems

I. Cação[1], H. R. Malonek[1] and G. Tomaz[1,2]

[1] *CIDMA-Center for Research and Development in Mathematics and Applications, University of Aveiro, Portugal*

[2] *UDI-Research Unit for Inland Development, Polythecnic Institute of Guarda, Portugal*

emails: isabel.cacao@ua.pt, hrmalon@ua.pt, gtomaz@ipg.pt

Abstract

We show how special well known matrices, namely, the creation and shift matrices play an important role on a matrix representation of orthogonal systems of polynomials with a hypercomplex variable and values in a Clifford algebra.

Key words: hypercomplex polynomials, matrix representation, creation matrix, shift matrix
MSC 2000: 30G35, 65F60, 11B83.

1 Introduction

In [7] the author constructs orthogonal bases of polynomials in the space of square integrable functions that are in the kernel of a generalized Cauchy-Riemann operator in the unit ball of $\mathbb{R}^{n+1}$. The construction process relies on building blocks that do not belong, in general, to the kernel of the referred operator. By using results established in [3] for these building blocks, we generalize the algebraic approach developed in [1] and stress the role of the well-known *creation matrix* and *shift matrix* in this representation. We recall that the so-called creation matrix H and the shift matrix J are defined by

$$(H)_{il} = \begin{cases} i, & i = l+1 \\ 0, & i \neq l+1 \end{cases} \qquad \text{and} \qquad (J)_{il} = \begin{cases} 1, & i = l+1 \\ 0, & i \neq l+1, \end{cases}$$

$i, l = 0, \ldots, m$, respectively. Although their simple structure, these matrices appear naturally in a matrix decomposition that represents polynomials in arbitrary dimension and in the framework of non-commutative algebras.

2 Basic concepts

Let $\{e_1, e_2, \ldots, e_n\}$ be an orthonormal basis of the Euclidean vector space $\mathbb{R}^n$ endowed with a non-commutative product according to the multiplication rules $e_k e_l + e_l e_k = -2\delta_{kl},\ k, l = 1, \ldots, n$, where δ_{kl} is the Kronecker symbol. The associative 2^n-dimensional Clifford algebra $\mathcal{C}\ell_{0,n}$ over $\mathbb{R}$ is the set of numbers of the form $a = \sum_A a_A e_A,\ a_A \in \mathbb{R}$, with $A \subseteq \{1, \cdots, n\}$, $e_A = e_{l_1} e_{l_2} \cdots e_{l_r}$, where $1 \leq l_1 < \cdots < l_r \leq n$ and $e_\emptyset =: e_0 =: 1$. The vector space $\mathbb{R}^{n+1}$ is embedded in $\mathcal{C}\ell_{0,n}$ by identifying $(x_0, x_1, \ldots, x_n) \in \mathbb{R}^{n+1}$ with the so-called *paravectors* $x = x_0 + \underline{x} \in \mathcal{A}_n := \mathrm{span}_{\mathbb{R}}\{1, e_1, \ldots, e_n\} \subset \mathcal{C}\ell_{0,n}$, where $\underline{x} = x_1 e_1 + \ldots + x_n e_n$ is called a *vector*. The conjugate $\bar{x}$ and the norm $|x|$ of x are given by $\bar{x} = x_0 - \underline{x}$ and $|x| = (x\bar{x})^{1/2} = (\bar{x}x)^{1/2} = \left(\sum_{k=0}^n x_k^2\right)^{1/2}$, respectively. The generalized Cauchy-Riemann operator and its conjugate are given, respectively, by $\overline{\partial} := \frac{1}{2}(\partial_0 + \partial_{\underline{x}})$ and $\partial := \frac{1}{2}(\partial_0 - \partial_{\underline{x}})$, where $\partial_0 := \frac{\partial}{\partial x_0}$ and $\partial_{\underline{x}} := e_1 \frac{\partial}{\partial x_1} + \cdots + e_n \frac{\partial}{\partial x_n}$.

We consider $\mathcal{C}\ell_{0,n}-$valued functions defined in an open subset $\Omega \subseteq \mathbb{R}^{n+1} \cong \mathcal{A}_n$, i.e. functions of the form $f(z) = \sum_A f_A(z) e_A$ with $f_A(z)$ real valued. Continuously differentiable functions f that satisfy the equation $\overline{\partial} f = 0$ (resp. $f\overline{\partial} = 0$) are called (left) *monogenic* (resp. right monogenic) and constitute the analogue of the class of holomorphic functions in higher dimensions. For more details, see [2, 5].

Let f be a monogenic function that is hypercomplex-differentiable in some domain $\Omega \subset \mathbb{R}^{n+1}$ in the sense of [6]. Then f is real-differentiable and its (hypercomplex) derivative is given by $f' = \partial f$ in Ω.

3 Matrix representation of orthogonal Clifford algebra-valued polynomials

Branching techniques combined with Gelfand-Tsetlin bases approach yield to the monogenic polynomials

$$f_{k,\mu} = X_{n+1,k_n}^{(k-k_n)} X_{n,k_{n-1}}^{(k_n - k_{n-1})} \cdots X_{3,k_2}^{(k_3-k_2)} \zeta^{k_2},$$

where $\zeta := x_1 - x_2\, e_1 e_2$ and μ is an arbitrary sequence of integers $(k_{n+1}, k_n, \ldots, k_3, k_2)$ such that $k = k_{n+1} \geq k_n \geq \ldots \geq k_3 \geq k_2 \geq 0$. These polynomials form an orthogonal basis with respect to a suitable Clifford algebra valued inner product of the space of monogenic polynomials of degree k (see [7]). The building blocks $X_{n+1,j}^{(k-j)}$, $j = 0, \ldots, k$, are, in general, non-monogenic polynomials and can be expressed as

$$X_{n+1,j}^{(k-j)}(x) = \sum_{s=0}^{k-j} \binom{k}{j+s} d_{j,s}(n)\, x_0^{k-j-s}\, \underline{x}^s, \quad x \in \mathcal{A}_n \tag{1}$$

 ISBN: 978-84-617-8694-7

where $d_{j,s}(n)$ are suitable real constants (cf. [3]). Given an arbitrary monogenic polynomial $P_j(\underline{x})$, in $\mathbb{R}^n$, of degree j, the system

$$\left\{\widetilde{X}^{(k)}_{n+1,j}(x) := X^{(k-j)}_{n+1,j}(x)\, P_j(\underline{x}), j = 0, \ldots, k,\ x \in \mathcal{A}_n\right\}_{k\in\mathbb{N}} \tag{2}$$

is formed by monogenic polynomials where again (1) appear as building blocks.

The matrix representation of the building blocks relies on the so-called *shifted generalized Pascal matrix* $S_r(t) = e^{(H+rJ)t}$, $t \in \mathbb{R}, r \in \mathbb{N}_0$, that combines in a single expression both H and J matrices.

For this representation we restrict ourselves to vectors of polynomials up to a certain degree $m \in \mathbb{N}_0$. Therefore, for each $j = 0, \ldots, k$, we consider the vectors $\xi(\underline{x}) = [1\ \underline{x}\ \cdots\ \underline{x}^m]^T$, $\boldsymbol{X}_j(x) = [X^{(0)}_{n+1,j}(x)\ X^{(1)}_{n+1,j}(x)\ \cdots\ X^{(m)}_{n+1,j}(x)]^T$ and the diagonal matrix $\mathcal{D}_j = \mathrm{diag}[d_{j,0}(n)\ d_{j,1}(n) \cdots d_{j,m}(n)]$.

Theorem 3.1. *For each fixed j, the vector $\boldsymbol{X}_j(x)$ can be decomposed in the form*

$$\boldsymbol{X}_j(x) = S_j(x_0)\ \mathcal{D}_j\ \xi(\underline{x}). \tag{3}$$

Proof. For each fixed j and considering $l = k - j$, it holds

$$\partial_0 X^{(0)}_{n+1,j}(x) = 0 \quad \text{and} \quad \partial_0 X^{(l)}_{n+1,j}(x) = (l + j)X^{(l-1)}_{n+1,j}(x), \ \ l > 0.$$

This property leads to the vector differential equation

$$\partial_0 \boldsymbol{X}_j(x) = (H + jJ)\boldsymbol{X}_j(x)$$

whose general solution is

$$\boldsymbol{X}_j(x) = e^{(H+jJ)x_0}\boldsymbol{X}_j(0, \underline{x}).$$

The result follows immediately noting that $X^{(l)}_{n+1,j}(0, \underline{x}) = d_{j,l}(n)\underline{x}^l$. □

The proposed decomposition (3) can be generalized for the whole orthogonal system (2) with the help of block matrices. The result highlights the role of the shift matrix J and its powers. Moreover, the generalized Pascal matrix $P(x_0) = S_0(x_0)$ comes into play as well (see [4]), connecting special simple matrices with real entries with hypercomplex entities.

Acknowledgements

This work was supported by Portuguese funds through the CIDMA - Center for Research and Development in Mathematics and Applications, and the Portuguese Foundation for Science and Technology ("FCT-Fundação para a Ciência e Tecnologia"), within project PEst-OE/MAT/UI4106/2013.

References

[1] L. ACETO, H. R. MALONEK AND G. TOMAZ, *A unified matrix approach to the representation of Appell polynomials*, Integral Transforms Spec. Funct. **26** (2015) 426-441.

[2] F. BRACKX, R. DELANGHE AND F. SOMMEN, *Clifford Analysis*, Pitman, Boston-London-Melbourne, 1982.

[3] I. CAÇÃO, M. I. FALCÃO AND H. R. MALONEK, *Three term recurrence relations for systems of Clifford algebra-valued orthogonal polynomials*, Adv. Appl. Clifford Algebras **27** (2017) 71-85.

[4] I. CAÇÃO, H. R. MALONEK AND G. TOMAZ, *Shifted generalized Pascal matrices in the context of Clifford algebra-valued polynomial sequences*, accepted for publication.

[5] K. GÜRLEBECK, K. HABETHA AND W. SPRÖßIG, *Holomorphic Functions in the Plane and n-Dimensional Space*, (Translated from the 2006 German original), Birkhäuser Verlag, Basel, 2008.

[6] K. GÜRLEBECK AND H. R. MALONEK, *A hypercomplex derivative of monogenic functions in* $\mathbb{R}^{m+1}$ *and its applications*, Complex Variables **39** (1999) 199-228.

[7] R. LÁVIČKA, *Complete Orthogonal Appell Systems for Spherical Monogenics*, Complex Anal. Oper. Theory **6** (2012) 477-489.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

On Vietoris' number sequence and combinatorial identities with quaternions

I. Cação[1], M. I. Falcão[2] and H. R. Malonek[1]

[1] *Center for Research and Development in Mathematics and Applications, University of Aveiro, Portugal*

[2] *Centre of Mathematics, University of Minho, Portugal*

emails: isabel.cacao@ua.pt, mif@math.uminho.pt, hrmalon@ua.pt

Abstract

Ruscheweyh and Salinas showed in 2004 the relationship of a celebrated theorem of Vietoris (1958) about the positivity of certain sine and cosine sums with the function theoretic concept of stable holomorphic functions in the unit disc. The present paper reveals that the coefficient sequence in Vietoris' theorem is identical to a number sequence obtained by a new combinatorial identity which involves generators of quaternions. In this sense Vietoris' sequence of rational numbers combines seemingly disperse subjects in Real, Complex and Hypercomplex Analysis. Thereby we show that a non-standard application of Clifford algebra tools is able to reveal new insights in objects of combinatorial nature.

Key words: Vietoris' number sequence, quaternions, combinatorial identities
MSC 2000: 30G35; 11B83; 05A19.

1 Introduction

In the center of our attention lies the sequence of rational numbers

$$1,\ \tfrac{1}{2},\ \tfrac{1}{2},\ \tfrac{3}{8},\ \tfrac{3}{8},\ \tfrac{5}{16},\ \tfrac{5}{16}, \tfrac{35}{128}, \tfrac{35}{128}, \tfrac{63}{256}, \tfrac{63}{256}, \tfrac{231}{1024}, \tfrac{231}{1024}, \ldots. \tag{1}$$

which by means of the *generalized central binomial coefficient* $\binom{k}{\lfloor \frac{k}{2} \rfloor}$ can be written in compact form (cf. [4]) as

$$\mathcal{S} = (c_k)_{k\geq 0}\,, \quad \text{where } \ c_k = \frac{1}{2^k}\binom{k}{\lfloor \frac{k}{2} \rfloor}, \ \ k \geq 0. \tag{2}$$

 ISBN: 978-84-617-8694-7

Seemingly for the first time this sequence appeared in the context of positive trigonometric sums in a celebrated paper of L. Vietoris [17]. Askey's version ([2, p. 5]) of Vietoris' theorem is the following:

Theorem 1 (L. Vietoris)

$$\sum_{k=1}^{n} a_k \sin k\theta > 0, \;\; 0 < \theta < \pi, \;\; \textit{and} \;\; \sum_{k=0}^{n} a_k \cos k\theta > 0, \;\; 0 \leq \theta < \pi,$$

where

$$a_{2k} = a_{2k+1} = \frac{(\frac{1}{2})_k}{k!}, \qquad k = 0, 1, \ldots, \tag{3}$$

with $(\cdot)_k$ *as the raising factorial in the classical form of the Pochhammer symbol.*

We call attention to the fact that because of (3), the coefficients in the *sine* sum used in Askey's as well as in Vietoris' original version are exactly the elements of $\mathcal{S}$ in (2) or, explicitly, in (1). Obviously, demanding in (3) that a_{2k} and a_{2k+1} coincide, the sequence of coefficients in the *cosine* sum differs from (1) by the inclusion of $a_0 = 1$ and the shift of the indices by one to the left, i.e. $a_0 = 1 \quad \text{and} \quad a_{k+1} = c_k, \; k \geq 0$. Even though this small difference, we call $\mathcal{S}$ in the sequel simply *Vietoris' number sequence*. Compared with the traditional way of defining the coefficient sequence by (3), the use of the properties of the *generalized central binomial coefficient* allowed at least a unique representation (2) with consecutively running index k.

Before continuing with the specific task of the present paper, it seems worthwhile to mention the other areas in which Vietoris' theorem played an important role. Using the arsenal of real analysis methods in positivity theory, Askey and Steinig showed in [3] the embedding of Vietoris' results in general problems for Jacobi polynomials, including their relation to other subjects in Harmonic Analysis. Later on, Ruscheweyh and Salinas showed in [15] an interesting relationship of Vietoris' theorem with the function theoretic concept of stable holomorphic functions in the unit disc. The common origin of the present paper with others like, for example, [9, 5] where the sequence $\mathcal{S}$ was already mentioned in different contexts, is the field of Hypercomplex Analysis, particularly the study of monogenic (or Clifford-holomorphic) Appell polynomials [1, 8, 10]. Recently in [6], the authors obtained even some number theoretic results for a related to $\mathcal{S}$ integer number sequence (sequence A283208 in The On-Line Encyclopedia of Integer Sequences, published electronically at https://oeis.org).

The goal of the present paper concerns the surprising appearance of $\mathcal{S}$ in a relation between the generators of Hamilton's well known non-commutative algebra $\mathbb{H}$ of quaternions (see e.g. [10]), relying only on elementary properties of a adequately generalized binomial formula for the quaternions. Taking into account that $\mathbb{H}$ can be considered as a Clifford algebra $\mathcal{C}\ell_{0,n}$, for $n = 2$, the generalization of our results for an arbitrary $n \geq 2$ will be

 ISBN: 978-84-617-8694-7

treated in an extended version of this paper by applying intrinsic properties of monogenic Appell polynomials in terms of several hypercomplex variables.

2 Hamilton's quaternions come into the play

Consider a quaternion $q \in \mathbb{H}$ written as

$$q = x_0 + x_1\mathbf{i} + x_2\mathbf{j} + x_3\mathbf{k}, \text{ where } \mathbf{i}^2 = \mathbf{j}^2 = \mathbf{k}^2 = \mathbf{ijk} = -1.$$

Due to non-commutativity the formal expansion of a binomial with two imaginary units (quaternion generators) $(\mathbf{i}+\mathbf{j})^k$, $k \geq 0$, will not directly lead to Pascal's triangle, as the case $k = 3$ shows:

$$(\mathbf{i}+\mathbf{j})^3 = \mathbf{i}^3 + (\mathbf{iij} + \mathbf{iji} + \mathbf{jii}) + (\mathbf{ijj} + \mathbf{jij} + \mathbf{jji}) + \mathbf{j}^3. \tag{4}$$

But that will happen if we try to embed the non-commutative multiplication into the concept of a $k-nary$ symmetric (or permutative) operation. Therefore let a_i stay for one of the generators $\mathbf{i}$ or $\mathbf{j}$ and write the quaternionic k-fold product of $k-s$ generators $\mathbf{i}$ and s generators $\mathbf{j}$, respectively, in the general form of a symmetric " $\times$ " product ([13]), i.e.

$$\mathbf{i}^{k-s} \times \mathbf{j}^s := \frac{1}{k!} \sum_{\pi(i_1,\ldots,i_n)} a_{i_1} a_{i_2} \cdots a_{i_k} \tag{5}$$

where the sum runs over ***all*** permutations of all $(i_1, \ldots, i_n)$. Then, by taking into account the repeated use of $\mathbf{i}$ and $\mathbf{j}$ on the right hand side of (5), we can write

$$\mathbf{i}^{k-s} \times \mathbf{j}^s = \frac{(k-s)!s!}{k!} \sum_{\pi(i_1,\ldots,i_n)} a_{i_1} a_{i_2} \cdots a_{i_k} = \left[\binom{k}{s}\right]^{-1} \sum_{\pi(i_1,\ldots,i_n)} a_{i_1} a_{i_2} \cdots a_{i_k}$$

where now the sum runs only over ***all distinguished*** permutations of all $(i_1, \ldots, i_n)$. Applying, for example, the convention (5) to (4) we obtain now, for $k = 3$ the expansion written with binomial coefficients in the form

$$(\mathbf{i}+\mathbf{j})^3 = \binom{3}{0}\mathbf{i}^3 + \binom{3}{1}\mathbf{i}^2 \times \mathbf{j} + \binom{3}{2}\mathbf{i} \times \mathbf{j}^2 + \binom{3}{3}\mathbf{j}^3. \tag{6}$$

Analogously, the expansion of $(\mathbf{i}+\mathbf{j})^k$ for any $k \geq 0$ follows now the rules of the ordinary binomial expansion in an evident way and leads to[1]

$$(\mathbf{i}+\mathbf{j})^k = \left[\sum_{s=0}^{k} \binom{k}{s} (\mathbf{i}^{k-s} \times \mathbf{j}^s)\right], \quad k \geq 0. \tag{7}$$

[1]An obvious generalization of (5) to the case of more than two generators used in the general case of $\mathcal{C}\ell_{0,n}$ for $n \geq 2$ leads to a polynomial formula.

 ISBN: 978-84-617-8694-7

Needless to say that the generalized binomial formula (7) is a key for studying combinatorial relations with quaternions in the following sections.

We will show now that another step towards our goal is the evaluation of expressions of the form $(\mathbf{i}^{k-s} \times \mathbf{j}^s)$ $k \geq 0,\ s = 0, 1, \ldots k$.

3 Evaluating symmetric products of quaternion generators

Notice that for $k \geq 2$ the influence of the non-commutativity of the ordinary quaternionic product is evident. This can be illustrated by the following examples:

$$\mathbf{i} \times \mathbf{j} = \frac{1!1!}{2!}(\mathbf{ij} + \mathbf{ij}) = 0,$$

$$\mathbf{i}^2 \times \mathbf{j} = \binom{3}{1}^{-1}(\mathbf{iij} + \mathbf{iji} + \mathbf{jii}) = -\frac{1}{3}\mathbf{j}, \tag{8}$$

$$\mathbf{i} \times \mathbf{j}^2 = \binom{3}{1}^{-1}(\mathbf{ijj} + \mathbf{jij} + \mathbf{jji}) = -\frac{1}{3}\mathbf{i}. \tag{9}$$

To obtain a general rule for those products we refer to an early version of the famous *Faá di Bruno formula* for the derivative of a composed function (see [11] and [12]) as it was used in [7].

T. Abadie's formula
If f and g are real functions of λ, with a sufficient number of derivatives, then

$$(g \circ f)^{(s)}(\lambda) = \sum_{l=0}^{s} \binom{s}{l} g^{(l)}\big(f(\lambda)\big)\Big\{\frac{d^{s-l}}{dh^{s-l}}\big(\Delta_h f(\lambda)\big)^l\Big\}_{h=0},$$

where $\Delta_h f(\lambda) := \frac{f(\lambda+h)-f(\lambda)}{h}$ is the difference quotient of f.

Consider now the polynomial $F_k(\lambda)$ of degree k in the real parameter λ,

$$F_k(\lambda) = (\mathbf{i} + \lambda\mathbf{j})^k = \sum_{s=0}^{k} \binom{k}{s} \lambda^s \mathbf{i}^{k-s} \times \mathbf{j}^s$$

and note that

$$\mathbf{i}^{k-s} \times \mathbf{j}^s = \frac{F_k^{(s)}(0)}{s!\binom{k}{s}}. \tag{10}$$

Since

$$F_k(\lambda) = \begin{cases} (-1-\lambda^2)^{\frac{k}{2}}, & \text{if} \quad k \text{ even;} \\ (-1-\lambda^2)^{\frac{k-1}{2}}(\mathbf{i} + \lambda\mathbf{j}), & \text{if} \quad k \text{ odd,} \end{cases} \tag{11}$$

 ISBN: 978-84-617-8694-7

it can be composed, for even k, in the form $F_k(\lambda) = (g \circ f)(\lambda)$ with suitably chosen functions

$$g(\lambda) = (-1-\lambda)^{\frac{k}{2}} \text{ and } f(\lambda) = \lambda^2.$$

whereas the case of an odd k can be reduced to the previous case by the relation

$$F_k(\lambda) = F_{k-1}(\lambda)(\mathbf{i} + \lambda\mathbf{j}).$$

Applying *T. Abadie's formula* and following the proof of *Proposition 1* in [7, p. 1730], about generalized powers of hypercomplex variables, one gets finally the values of (10) in the form:

$$\mathbf{i}^{k-s} \times \mathbf{j}^s = \begin{cases} (-1)^{\frac{k}{2}} \binom{\frac{k}{2}}{\frac{s}{2}} \binom{k}{s}^{-1}, & k \text{ even and } s \text{ even;} \\ 0, & k \text{ even and } s \text{ odd;} \\ (-1)^{\frac{k-1}{2}} \binom{\frac{k-1}{2}}{\frac{s}{2}} \binom{k}{s}^{-1} \mathbf{i}, & k \text{ odd and } s \text{ even;} \\ (-1)^{\frac{k-1}{2}} \binom{\frac{k-1}{2}}{\frac{s-1}{2}} \binom{k}{s}^{-1} \mathbf{j}, & k \text{ odd and } s \text{ odd.} \end{cases} \tag{12}$$

Remark 1 The examples in the beginning of this section, in particular relations (8) and (9), confirm very well the last three equalities in (12). It is evident, that there are, depending on the relative parities of k resp. s only four types of values of $\mathbf{i}^{k-s} \times \mathbf{j}^s$, namely (i) real and different from zero, (ii) equal to zero, (iii) a real multiple of $\mathbf{i}$ and (iv) a real multiple of $\mathbf{j}$. The reason for this, at first glance, surprising result is based on the following facts. Obviously, an integer power of the type of a so-called *reduced purely imaginary quaternion* $\underline{q} = \mathbf{i}+\mathbf{j}$, is either a real number (if k is even) or again a reduced purely imaginary quaternion (if k is odd; cf. (11)). Besides this, the symmetric products in which such a power (7) is additively decomposed avoid the appearance of ordinary mixed products like, for example, $\mathbf{i} \cdot \mathbf{j}$ as mutually annihilating summands. This becomes directly plausible if we look for example to the binomial expansion for even $k = 2$ or $k = 4$ where only entries of type (i) and type (ii) are present. An example for an odd k with $2 \times \frac{k+1}{2} = k+1$ alternating entries of type (iii) and (iv) is (6). In both cases we can recognize the symmetric structure of the corresponding lines in a Pascal triangle with quaternionic entries.

4 A combinatorial identity for Vietoris' number sequence

Before coming to our main result, let us still remember a well known combinatorial identity (cf. [14, p. 130] or [16, p. 44]) that we need for its proof, namely

$$\sum_{t=0}^{m} \binom{2t}{t} \binom{2m-2t}{m-t} = 4^m, \tag{13}$$

 ISBN: 978-84-617-8694-7

which in turn can be proved by evaluating the square of the generating function of the central binomial coefficients and its derivatives at $x = 0$.

Now we can prove

Theorem 2 *Let* $\mathbf{i}$ *and* $\mathbf{j}$ *be two generators of a reduced purely imaginary quaternion. Then the following combinatorial identity holds*

$$\binom{k}{\lfloor \frac{k}{2} \rfloor} \left[\sum_{s=0}^{k} \binom{k}{s} (\mathbf{i}^{k-s} \times \mathbf{j}^{s})^2 \right] = (-2)^k. \tag{14}$$

Taking into account the form of the elements of Vietoris' number sequence (2), formula (14) can be rewritten in order to obtain the representation of Vietoris' number sequence by symmetric products of the generators $\mathbf{i}$ and $\mathbf{j}$.

Corollary [*Representation of Vietoris' number sequence*]

$$c_k = (-1)^k \left[\sum_{s=0}^{k} \binom{k}{s} (\mathbf{i}^{k-s} \times \mathbf{j}^{s})^2 \right]^{-1}. \tag{15}$$

Proof of Theorem 2
As an auxiliary calculation we determine the square of the symmetric products in (12) multiplied by $\binom{k}{s}$ and distinguish between $k = 2m$ and $k = 2m + 1$ resp. $s = 2t$ and $s = 2t + 1$ for the different parities. We get immediately

$$\binom{k}{s} (\mathbf{i}^{k-s} \times \mathbf{j}^{s})^2 = \begin{cases} \dfrac{\binom{m}{t}^2}{\binom{2m}{2t}}, & k = 2m \quad \text{and } s = 2t; \\ 0, & k = 2m \quad \text{and } s = 2t + 1; \\ \dfrac{-\binom{m}{t}^2}{\binom{2m+1}{2t}}, & k = 2m + 1 \quad \text{and } s = 2t; \\ \dfrac{-\binom{m}{t}^2}{\binom{2m+1}{2t+1}}, & k = 2m + 1 \quad \text{and } s = 2t + 1. \end{cases} \tag{16}$$

Now we consider two cases corresponding to the parity of k.

I. k even

Denote by A_k the left-hand side of (14). The use of (16) (note that the second case

 ISBN: 978-84-617-8694-7

implies that the sum over odd values of s completely vanishes) together with (13) allows to write

$$\begin{aligned} A_{2m} &= \binom{2m}{m}\Big[\sum_{s=0}^{2m}\binom{2m}{s}(\mathbf{i}^{2m-s}\times\mathbf{j}^{s})^2\Big] = \binom{2m}{m}\Big[\sum_{t=0}^{m}\binom{2m}{2t}(\mathbf{i}^{2m-2t}\times\mathbf{j}^{2t})^2+0\Big] \\ &= \binom{2m}{m}\left[\sum_{t=0}^{m}\frac{\binom{m}{t}^2}{\binom{2m}{2t}}\right] = \sum_{t=0}^{m}\frac{(2t)!}{t!t!}\cdot\frac{(2m-2t)!}{(m-t)!(m-t)!} \\ &= 4^m = (-2)^{2m} \end{aligned}$$

II. k odd

In this case we apply the third and the fourth case of (16) and proceed analogously to the former case.

$$\begin{aligned} A_{2m+1} &= \binom{2m+1}{m}\Big[\sum_{s=0}^{2m+1}\binom{2m}{s}(\mathbf{i}^{2m+1-s}\times\mathbf{j}^{s})^2\Big] = \binom{2m+1}{m}\left[\sum_{t=0}^{m}\frac{-\binom{m}{t}^2}{\binom{2m+1}{2t}}-\sum_{t=0}^{m}\frac{\binom{m}{t}^2}{\binom{2m+1}{2t+1}}\right] \\ &= \binom{2m+1}{m}\sum_{t=0}^{m}\binom{m}{t}^2\left[\frac{(2t)!(2m-2t+1)!}{(2m+1)!}+\frac{(2m-2t)!(2t+1)!}{(2m+1)!}\right] \\ &= -2\sum_{t=0}^{m}\frac{(2t)!}{t!t!}\cdot\frac{(2m-2t)!}{(m-t)!(m-t)!} \\ &= -2\cdot 4^m = (-2)^{2m+1}. \end{aligned}$$

□

We finish with examples for the first values of c_k in (15) resp. (1),

$$\begin{aligned} c_0 &= 1 \\ c_1 &= (-1)^1\left[\mathbf{i}^2+\mathbf{j}^2\right]^{-1} = \frac{1}{2} \\ c_2 &= (-1)^2\left[(\mathbf{i}^2)^2+\binom{2}{1}(\mathbf{i}\times\mathbf{j})^2+(\mathbf{j}^2)^2\right]^{-1} = \frac{1}{2} \\ c_3 &= (-1)^3\left[(\mathbf{i}^3)^2+\binom{3}{1}(\mathbf{i}^2\times\mathbf{j})^2+\binom{3}{2}(\mathbf{i}\times\mathbf{j}^2)^2+(\mathbf{j}^3)^2\right]^{-1} = \frac{3}{8} \\ c_4 &= (-1)^4\left[(\mathbf{i}^4)^2+\binom{4}{1}(\mathbf{i}^3\times\mathbf{j})^2+\binom{4}{2}(\mathbf{i}^2\times\mathbf{j}^2)^2+\binom{4}{3}(\mathbf{i}\times\mathbf{j}^3)^2+(\mathbf{j}^4)^2\right]^{-1} = \frac{3}{8}. \end{aligned}$$

 ISBN: 978-84-617-8694-7

Acknowledgements

The work of the first and third authors was supported by Portuguese funds through the CIDMA - Center for Research and Development in Mathematics and Applications, and the Portuguese Foundation for Science and Technology ("FCT-Fundação para a Ciência e Tecnologia"), within project PEst-OE/MAT/UI4106/2013. The work of the second author was supported by Portuguese funds through the CMAT - Centre of Mathematics and FCT within the Project UID/MAT/00013/2013.

References

[1] P. Appell, *Sur une classe de polynomes*, Ann. Sci. Ècole Norm. Sup. **9** (2) (1880) 119–144.

[2] R. Askey, *Orthogonal polynomials and special functions*, Society for Industrial and Applied Mathematics, Philadelphia, 1975, 2nd printing 1994.

[3] R. Askey and J. Steinig, *Some positive trigonometric sums*, Transactions AMS **187** (1) (1974) 295–307.

[4] I. Cação, M. I. Falcão and H. R. Malonek, *Matrix representations of a basic polynomial sequence in arbitrary dimension*, Comput. Methods Funct. Theory **12** (2) (2012) 371–391.

[5] I. Cação, M. I. Falcão and H. R. Malonek, *Three-term recurrence relations for systems of Clifford algebra-valued orthogonal polynomials*, Adv. Appl. Clifford Algebr. **27** (1) (2017) 71–85.

[6] I. Cação, M. I. Falcão and H. R Malonek, *Hypercomplex polynomials, Vietoris' rational numbers and a related integer numbers sequence*, Complex Anal. Oper. Theor. (2017) 1–18, Article in Press.

[7] C. Cruz, M. I. Falcão and H. R. Malonek, *Monogenic pseudo-complex power functions and their applications*, Math. Meth. App. Sci. **37** (2014) 1723–1735.

[8] M. I. Falcão and H. R. Malonek, *Generalized exponentials through Appell sets in* $\mathbb{R}^{n+1}$ *and Bessel functions*. In: T. E. Simos, G. Psihoyios, C. Tsitouras (Eds.), AIP Conference Proceedings **936** (2007) 738–741.

[9] M. I. Falcão and H. R. Malonek, *A note on a one-parameter family of non-symmetric number triangles*, Opuscula Mathematica **32** (4) (2012) 661–673.

 ISBN: 978-84-617-8694-7

[10] K. Gürlebeck, K. Habetha and W. Sprössig, *Holomorphic Functions in the Plane and n-Dimensional Space*, Translated from the 2006 German original. Birkhäuser Verlag, Basel, 2008.

[11] W. P. Johnson, *The curious history of Faà di Bruno's formula*, The Mathematical Association of America Monthly **109** (3) (2002) 217–234.

[12] C. F. Faà di Bruno, *Note sur une nouvelle formule de calcul différentiel*, Quarterly J. Pure Appl. Math. **1** (1857) 359–360.

[13] H. R. Malonek, *Power series representation for monogenic functions in $\mathbb{R}^{n+1}$ based on a permutational product*, Complex Variables, Theory Appl. **15** (1990) 181–191.

[14] J. Riordan, *Combinatorial Identities*, John Wiley & Sons Inc., New York, 1968.

[15] S. Ruscheweyh and L. Salinas, *Stable functions and Vietoris' theorem*, J. Math. Anal. Appl. **291** (2004) 596–604.

[16] R. Stanley, *Enumerative Combinatorics*, Vol. 1, Cambridge Studies in Advanced Mathematics 49 Cambridge University Press, Cambridge, 1997.

[17] L. Vietoris, *Über das Vorzeichen gewisser trigonometrischer Summen*, Sitzungsber. Österr. Akad. Wiss **167** (1958) 125–135.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

Convergence and stability of a modification of Jungck-Ishikawa iteration sequence

K. Calderón[1], J. Martínez-Moreno[2] and E. Rojas[3]

[1] *Facultad de Ciencias, Universidad de Ciencias Aplicadas y Ambientales U.D.C.A. Colombia*

[2] *Department of Mathematics, Universidad de Jaén. Spain*

[3] *Department of Mathematics, Universidad Nacional de Colombia. Colombia*

emails: kencalderon@udca.edu.co, jmmoreno@ujaen.es, emrojass@unal.edu.co

Abstract

The purpose of this paper is to study the problem of (S,T)-stability and convergence of the θ-modified Jungck-Ishikawa.

Key words: Coincidence point; Ishikawa iteration; stability.

1 Introduction and Preliminaries

Throughout this paper, we assume that E is a real Banach space.

An operator $T : E \to E$ is said to be Lipschizian if there exists a constant $L > 0$ such that

$$\| Tx - Ty \| \leq L \| x - y \|, \tag{1}$$

for all $x, y \in E$.

An operator $T : E \to E$ is said to be strongly pseudocontractive if there exists $t > 1$ such that

$$\| x - y \| \leq \| (1 + r)(x - y) - rt(Tx - Ty) \|, \tag{2}$$

for all $x, y \in E$ and $r > 0$.

Let $S, T : E \longrightarrow E$ be two non-self mappings such that $T(E) \subseteq S(E)$. Then the sequence $\{Sx_n\}_{n=0}^{\infty}$ defined by

$$Sx_{n+1} = Tx_n; \qquad n = 0, 1, 2, ... \tag{3}$$

 ISBN: 978-84-617-8694-7

is called Jungck-Picard iteration scheme [3].

Definition 1.1 *Let $T : E \to E$ be a self mapping. Then T has a fixed point if there is an $x \in E$ such that $Tx = x$. The point x is called a fixed point of T. The set of fixed point of T will be denoted by $F(T)$.*

Definition 1.2 *Let $S, T : E \longrightarrow E$. Then x is called a coincidence (common fixed) point of T and S, respectively, if there exists $x \in X$ such that*

$$(x =)Tx = Sx$$

The set of all coincidence points of T and S will be denoted by $C(T, S)$.

Let $S, T : E \longrightarrow E$ are two non-self mappings such that $T(E) \subseteq S(E)$. Then the sequence $\{Sx_n\}_{n=0}^{\infty}$ defined by

$$Sx_{n+1} = (1 - \alpha_n)Sx_n + \alpha_n Tx_n \qquad n = 0, 1, 2, ..., \tag{4}$$

where $\{\alpha_n\}$ is introduced by Singh et al. [7] to establish some stability results and it is called Jungck-Mann iteration scheme.

On the other hand, the sequence $\{Sx_n\}_{n=0}^{\infty}$ defined by

$$\begin{aligned} Sx_{n+1} &= (1 - \alpha_n)Sx_n + \alpha_n Tz_n \\ Sz_n &= (1 - \beta_n)Sx_n + \beta_n Tx_n, \qquad n = 0, 1, 2, ..., \end{aligned} \tag{5}$$

where $\{\alpha_n\}$ and $\{\beta_n\}$ are the sequences in [0, 1), is called Jungck Ishikawa iteration scheme [5]. If S is the identity mapping, then Jungck Picard, Jungck Mann and Jungck Ishikawa iteration schemes becomes Picard, Mann and Ishikawa iteration schemes.

The Ishikawa iterative process was first introduced by Ishikawa [2] in 1974, in order to approximate fixed point of Lipschitzian pseudocontractive operators, because in the case T is only pseudocontractive, the Mann iteration does not converge generally to the fixed point of T. In the literature, there are many convergence results on the Ishikawa iterations. For example, Chidume [1] proved the following result:

Theorem 1.1 *If E is a real Banach space with a uniformly convex dual E^*, and K is a nonempty closed convex and bounded subset of E, and $T : E \to E$ is a Lipschitz strongly pseudocontractive mapping, the Ishikawa iterative sequence $\{x_n\}$ converges strongly to the unique fixed point of T.*

 ISBN: 978-84-617-8694-7

Consider the nonlinear equation

$$f(x) = 0, \quad x \in \mathbb{R}. \tag{6}$$

Let $S, T : E \to E$, $T(E) \subset S(E)$, S is onto and T and S are differentiable. Suppose ξ is a simple zero of f and x_0 is an initial guess nearer ξ. The equation (6) can be written as

$$Sx = Tx. \tag{7}$$

Following the approach of [4] (see also [6]), if $T'x \neq 1$, we can modifify (7) by multiplying $\theta \neq -1$ on both sides as follows

$$\theta Sx = \theta Tx,$$

which implies that

$$Sx = \frac{\theta Tx + Sx}{\theta + 1} := T_\theta x, \tag{8}$$

where θ is an arbitrary number.

For a given x_0, we can find the approximate solution of x_{n+1} by using the iteration scheme named as modified Jungck iteration scheme

$$Sx_{n+1} = \frac{\theta Tx_x + Sx_x}{\theta + 1}, \quad \theta \neq -1, \tag{9}$$

or

$$Tx_{n+1} = \frac{\theta Tx_x + Sx_x}{\theta + 1}, \quad \theta \neq -1. \tag{10}$$

Using (9) in Jungck-Ishikawa iteration scheme, we develop a modified one scheme as follow:

Definition 1.3 *(See [6]). Let $T : E \longrightarrow E$ be a mapping. For given $\theta \neq -1$, $x_0 \in E$ and $\{\alpha_n\}$, $\{\beta_n\}$ sequences in [0,1], we denote $\{Sx_n\} \subset E$ the sequence defined by*

$$\begin{aligned} Sx_{n+1} =& (1 - \alpha_n)Sx_n + \alpha_n T_\theta z_n, \\ Sz_n =& (1 - \beta_n)Sx_n + \beta_n T_\theta x_n, \end{aligned} \tag{11}$$

where $T_\theta x = \frac{\theta Tx + Sx}{1+\theta}$. The iteration process (11) will be called modified Jungck-Ishikawa iteration.

The following lemma will be needed in proving our main results.

Lemma 1.1 *(See [8]). Let $\{\lambda_n\}$,$\{\mu_n\}$, $\{d_n\}$ be nonnegative real sequences satisfying*

$$\lambda_{n+1} \leq (1 + t_n)\lambda_n + \mu_n\lambda_n + c_n + d_n \qquad \forall n \geq 1$$

If $\{t_n\}$ is a sequence in [0,1] such that $\sum_{n=1}^{\infty} t_n = \infty$, $\sum_{n=1}^{\infty} \mu_n < \infty$, $\sum_{n=1}^{\infty} c_n < \infty$ and $d_n = o(t_n)$. Then $\lambda \longrightarrow 0$ as $n \longrightarrow \infty$

 ISBN: 978-84-617-8694-7

Using the following definition of the stability of an iteration process, Olantiwo [5] established some stability results as well as some strong convergence results for a pair of nonselfmappings using a Jungck-Ishikawa iteration process and some general contractive conditions.

Definition 1.4 *See ([5]). Let $S, T : E \to E$, $T(E) \subseteq S(E)$ and $l \in C(T, S)$. For any $x_0 \in E$, let the sequence $\{Sx_{n+1}\}_{n=0}^{\infty}$ generated by iteration $Sx_{n+1} = f(T, x_n)$, $n = 0, 1, 2, 3, ...$ converges to w. Let $\{Sy_n\}_{n=0}^{\infty} \subset E$ be an arbitrary sequence, and $\varepsilon_n = d(Sy_{n+1}, f(T, y_n))$, $n = 0, 1, 2, ...$ Then, the iteration $\{Sx_{n+1}\}_{n=0}^{\infty}$ will be called (S, T)-stable if only if $\lim_{n\to\infty} \varepsilon_n = 0$ implies that $\lim_{n\longrightarrow\infty} Sy_n = w$*

2 Main results

Definition 2.1 *Given $S : E \to E$ and $T : E \to E$.*

T is said to be S-Lipschitzian if there exists a constant $L > 0$ such that

$$\| Tx - Ty \| \leq L \| Sx - Sy \|, \tag{12}$$

for all $x, y \in E$.

T is S-strongly pseudocontractive is there exists $k > 0$ such that

$$\| Sx - Sy \| \leq \| Sx - Sy + q[(S - T - kS)x - (S - T - kS)y] \| \tag{13}$$

for all $x, y \in E$ and $q > 0$.

Let $S, T : E \to E$. Suppose $x_0 \in E$ and $x_{n+1} = f(x_n, \alpha_n, T, S)$ defines an iteration procedure which yields a sequence of points $\{Sx_n\} \subset E$ converging to a coincidence point w of T and S. Let $\{y_n\}$ and let $\varepsilon_n = \| Sy_{n+1} - f(x_n, \alpha_n, T, S) \|$ be a sequence in $[0, \infty)$.

Definition 2.2 *If $\sum_{n=1}^{\infty} \varepsilon_n < \infty$, implies that $\lim_{n\to\infty} Sy_n = w$, then the procedure defined by $x_{n+1} = f(x_n, \alpha_n, T, S)$ is said to be almost (S, T)-stable.*

If $\lim_{n\to\infty} \frac{\varepsilon_n}{\alpha_n} = 0$, implies that $\lim_{n\to\infty} Sy_n = w$, then the procedure defined by $x_{n+1} = f(x_n, \alpha_n, T, S)$ is said to be weakly (S, T)-stable.

Theorem 2.1 *Let $T, S : E \to E$ are two non-self mappings such that $T(E) \subseteq S(E)$, where $T_\theta = \frac{\theta Tx + Sx}{\theta + 1}$ is a S-Lipschitzian and S-strongly pseudocontractive mapping with $C(T, S) \neq \emptyset$. Let $\{x_n\}$ be defined by (11) and $\{\alpha_n\}, \{\beta_n\} \subset [0, 1]$ satisfying the conditions:*

i . $\sum_{k=1}^{\infty} \alpha_k = \infty$

ii . $\lim_{n\longrightarrow\infty} \sup \alpha_n < k/((L + 1)^3 + 2)$ *and* $\sum_{k=1}^{\infty} \alpha_k \beta_k < \infty$

 ISBN: 978-84-617-8694-7

Suppose $\{y_n\} \subset E$ *and define* $\{\varepsilon_n\}$ *by*

$$\begin{aligned} Ss_n =&(1-\beta_n)y_n + \beta_n T_\theta y_n \\ \varepsilon_n =& \parallel Sy_{n+1} - (1-\alpha_n)Sy_n - \alpha_n T_\theta s_n \parallel, \qquad n \geq 0 \end{aligned}$$

Then the following assertions hold:

1. $\{Sx_n\}_{n=0}^{\infty}$ *converges strongly to a coincidente point of* S *and* T.
2. $\{Sx_n\}_{n=0}^{\infty}$ *is both almost* (S,T)*-stable and weakly* (S,T)*-stable.*

Similarly to Theorem 2.2, we have the following

Theorem 2.2 *Let* $T, S : E \to E$ *are two non-self mappings such that* $T(E) \subseteq S(E)$*, where* $T_\theta = \frac{\theta Tx + Sx}{\theta+1}$ *is a* S*-Lipschitzian and* S*-strictly pseudocontractive mapping with* $C(T,S) \neq \emptyset$*. Let* $\{x_n\}$ *be defined by (11) and* $\{\alpha_n\}, \{\beta_n\} \subset [0,1]$ *satisfying the conditions:*

i . $\sum_{k=1}^{\infty} \alpha_k = \infty$

ii . *There exists* $\delta \in (0, 1-k)$ *such that* $\lim\limits_{n \longrightarrow \infty} \sup \alpha_n < \delta/((L+1)^3 + 2)$

iii. $\sum_{k=1}^{\infty} \alpha_k \beta_k < \infty$

Suppose $\{y_n\} \subset E$ *and define* $\{\varepsilon_n\}$ *by*

$$\begin{aligned} Ss_n =&(1-\beta_n)y_n + \beta_n T_\theta y_n \\ \varepsilon_n =& \parallel Sy_{n+1} - (1-\alpha_n)Sy_n - \alpha_n T_\theta s_n \parallel, \qquad n \geq 0 \end{aligned}$$

Then the following assertions hold:

1. $\{Sx_n\}_{n=0}^{\infty}$ *converges strongly to a coincidente point of* S *and* T.
2. $\{Sx_n\}_{n=0}^{\infty}$ *is both almost* (S,T)*-stable and weakly* (S,T)*-stable.*

References

[1] C. E. Chidume, Approximation of fixed points of strongly pseudocontractive mappings, Proc. Amer. Math. Soc. 120, No.2 (1994), 545-551. MR 94d:47056

[2] Ishikawa, S. *Fixed points by a new iteration method*, Proceedings of the American Mathematical Society, 44(1), 147-150. (1974).

[3] Jungck, G. *Communting and fixed points* The American Mathematical Monthly, 85(4), 261-263. (1976)

 ISBN: 978-84-617-8694-7

[4] Kang, S.M., Rafiq, A., Kwun, Y.C. *A new second-order iteration method for solving nonlinear equations*, Abstract and Applied Analysis, 2013, art. no. 487062

[5] Olatinwo, M.O. *Some stability and strong convergence results for the Jungck-Ishikawa iteration process*. Creative Mathematics and Informatics, 17, 33-42. (2008)

[6] Raqif A, Tanveer M, & Kang, S. *The Modified Jungck Mann and modified Jungck Ishikawa Iteration Schemes For Zamifirescu Operators*, International Journal of pure and applied Mathematics,102(2), 357-382. (2015)

[7] Singh, S.L. Bhatnagar,C. & Mishra, S. N., *Stability of Jungck-type iterative procedures*, International Journal Mathematics and Mathematical Sciences.2005(19), 3035-3043. (2005).

[8] Yang, L, & Peng, S. *Convergence and stability of modified Ishikawa iteration sequence with errors*, Fixed point Theory and aplications, 2014(1),224. (2014)

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference on Computational and Mathematical Methods in Science and Engineering, CMMSE 2017 4–8 July, 2017.

On the effect of a polynomial field prescribed in an unbounded domain with an elliptical isolated inhomogeneity

Carmen Calvo-Jurado[1] **and William J. Parnell**[2]

[1] *Department of Mathematics, University of Extremadura*

[2] *School of Mathematics, Oxford Road, Manchester, M13 9PL, University of Manchester*

emails: ccalvo@unex.es, William.Parnell@manchester.ac.uk

Abstract

The Eshelby inhomogeneity or inclusion problem is known to play a crucial role in the micromechanical analysis of the mechanical behavior of composites [10] since leads to predict the overall behaviour of heterogeneous mediums. The so call inclusion problem has attracted the interest of several authors in the last two centuries. In particular, Poisson studied the perturbed field due to an isolated ellipsoid in the context of the Newtonian potential problem [1]. Later, Maxwell [12] obtained explicit uniform expressions for the induced electric field inside the ellipsoidal inclusion under an uniform electric polarization. Donnell [2] studied the problem of a homogeneous elliptical inclusion with the same elastic properties as the surrounding matrix and undergoes uniform stress-free transformation strains (also called *eigenstrains* in the literature). Afterward, Hardiman [3] considered the case of an elliptical inhomodeneity having elastic properties different from those of the matrix.

In the context of linear elasticity Eshelby [4], [5], [7] provided a systematic investigation of the corresponding three-dimensional problems. He introduced an interesting method known as the equivalent inclusion method for dealing with 3D inhomogeneity problems and showed that given an isolated ellipsoidal inclusion embedded into an infinitely homogeneous material, for any uniform loading imposed in the far field, the strain inside it is also uniform. This property, also called the *Eshelby uniformity property*, has been studied by many since then because it can model numerous phenomena in materials science, such as phase transformations in solids or the thermal expansion problem, as well as being useful as a simple modelling tool for inhomogeneous media [14].

In the case of non-uniform far field conditions, Eshelby showed that if the loading is a polynomial of order n, the interior field is characterized by a polynomial of the same

 ISBN: 978-84-617-8694-7

order. This is often called *Eshelby's polynomial conservation theorem*. Mura [10] gave further details of this theory and developed a method of solution based on multipole expansions in order to evaluate the induced strain field in an ellipsoidal inhomogeneity using the eigenstrain concept.

For ellipsoidal inhomogeneity problems the so-called Eshelby tensor arises naturally and it can be shown to be uniform for under uniform far fields conditions. However, for a *non-uniform* loading, Eshelby's solutions in general involve difficult integral expressions and then, the prediction of its influence is not a natural quantity to work with. Therefore, several methods based on complex variable [11] (due to uniform interior eigenstrain) or some circle theorems [6], [9], [8] have been developed in recent years for the study of interior fields. Recently, in [14] other one has been defined in order to determine such Eshelby tensors, even for general inclusion shapes. Despite these difficulties, it occurs that in many practical applications, e.g. permittivity, conductivity, elasticity, the influence on interior fields under *non-uniform* far field conditions also needs to be treated. As consequence, new better predictive methods are needed to be developed in order to characterize them in a direct manner. For these reasons, in this work for prescribed polynomial far field conditions, following the approximation to the Eshelby tensors developed in [14] and the integral operator method introduced in [13], and [15], we present a scheme to approximate interior fields inside isolated inhomogeneities of elliptical shape since many composites belong to this class (including layered and fiber reinforced composites). Moreover, we also verify that the obtained results ascertain the *Eshelby's polynomial conservation property* and are agree with some others provided by using complex conformal mapping [11] or the classical circle inclusion theorem ([6], [9], [8]). This result constitutes an useful scheme in the development of predictive material models that undergoes general non-uniform eigenstrains since they also can be uniformly approximated by Taylor series expansions, and even because they can be extended in forthcoming schemes to non elliptical shapes.

Key words: eigenstrains, elliptical inclusions, heterogeneities, conformal mapping, Eshelby's tensor, circle theorem, polynomial conservation property

MSC 2000: 30E10, 42A10, 65D99, 35Q74

Acknowledgements

Parnell is grateful to the Engineering and Physical Sciences Research Council for funding his fellowship (EP/L018039/1). Calvo has been partially supported by the project MTM2014-53309-P of the "Ministerio de Economía y Competitividad (Plan Estatal 2013-2016 Excelencia - Proyectos I+D)" of Spain and the research groups FQM-309 of the "Junta de Andalucía and FQM-022 of the Junta de Extremadura".

References

[1] S.D. Poisson, *Second mémoire sur la thorie de magnetisme*, Mém. Acad. R. Sci. Inst.

 ISBN: 978-84-617-8694-7

Fr. **5** (1826) 488-533.

[2] L.H., Donell, *Stress concentration due to elliptical discontinuities in plates under edge stresses*. In: Theodore Von Karman Anniversary Volume, 293–309, California Institute of Technology, 1941.

[3] N.J., Hardiman, *Elliptical elastic inclusion in an infinite elastic plane*, Q. J. Mech. Appl. Math., **7** (1954) 226–230.

[4] J.D. Eshelby, *The determination of the elastic field of an ellipsoidal inclusion and related problems*, Proc. R. Soc. London A, **241** (1957) 376-396.

[5] J.D. Eshelby, *The Elastic Field Outside an Ellipsoidal Inclusion*, Proc. R. Soc. London, Ser. A, **252** (1959) 561-569.

[6] L.M. Milne-Thromson, *Theoretical Hydrodynamics*, Macmillan, 1960.

[7] J.D. Eshelby, *Elastic Inclusions and Inhomogeneities*, Progress in Solid Mechanics, 2, I. N. Sneddon and R. Hill, eds., North-Holland, Amsterdam, 89140, 1960.

[8] E. Smith, *The interaction between dislocations and inhomogeneities*, Int. J. Eng. Sci. **6**, 3 (1968) 129–143.

[9] G.P. Sendeckyj, *Longitudinal shear modulus of filamentary composite containing curvilinear fibers*, Fibre Sci. Technol., **2**, 3 (1970) 211-222.

[10] T. Mura, *Micromechanics of Defects in Solids*, Martinus Nijhoff Publishers, Dordrecht, 1982.

[11] C.Q. Ru and P. Schiavone, *On the elliptic inclusion in anti-plane shear*, Math. Mech. Solids, **1** (1996) 327-333.

[12] J.C. Maxwell, *A Treatise on Electricity and Magnetism*, Vols. 1 and 2, Oxford University Press, Oxford, 1998.

[13] W.J. Parnell and I.D. Abrahams, *A new integral equation approach to elastodynamic homogenization*, Proc. Roy. Soc. A, **464** (2008), 1461–1482.

[14] W.J. Parnell, *The Eshelby, Hill, Moment and Concentration Tensors for Ellipsoidal Inhomogeneities in the Newtonian Potential Problem and Linear Elastostatics*, J. Elast., **125** (2016), 231-294.

[15] D. Joyce and W.J. Parnell, *The Newtonian potential inhomogeneity problem: non-uniform eigenstrains in cylinders of non-elliptical cross-section*, J. Eng. Math. Submitted.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

Symmetry analysis for a generalized dissipative Dullin-Gottwald-Holm equation with arbitrary coefficients

J.C. Camacho[1], M. Rosa[1], M.L. Gandarias[1] and M.S. Bruzón[1]

[1] *Department of Mathematics, University of Cádiz*

emails: josecarlos.camacho@uca.es, maria.rosa@uca.es, marialuz.gandarias@uca.es, m.bruzon@uca.es

Abstract

In this work we study the classical Lie symmetries of a generalized Dullin-Gottwald-Holm equation with arbitrary coefficients. We consider the model (GDGH)

$$u_t - \alpha^2 u_{xxt} + (f(u))_x + \gamma u_{xxx} + \lambda(u - \alpha^2 u_{xx}) = \alpha^2(2u_x u_{xx} + u u_{xxx}). \qquad (1)$$

The equation GDGH is a nonlinear partial differential equation describing the unidirectional propagation of shallow water waves. The function $u(x,t)$ is the fluid velocity at time $t > 0$ in the direction x, γ is the coefficient of the linear dispersion term, $f(u)$ is a polynomial which derivative can be used to model the linear wave speed for undisturbed water at rest at spatial infinity, α^2 and $\frac{\gamma}{k}$ are squares of length scales where k is the coefficient of u in $f(u)$, this is, the linear wave speed for undisturbed water resting at spatial infinity, and λ is a coefficient of the weakly dissipative term $(u - \alpha^2 u_{xx})$.

When $f(u) = \frac{3}{2}u^2 + c_1 u + c_0$ and $\lambda > 0$ the GDGH equation is the disipate Dullin-Gottwald-Holm equation studied by Guo and Ni in [6], and later by Novruzov in [11].

When $f(u) = \frac{3}{2}u^2 + c_1 u + c_0$, $\lambda = 0$ and $\gamma \neq 0$ (1) becomes the Dullin-Gottwald-Holm equation (DGH) whith was orignally derived as an approximation to the incompressible Euler equations [4]. Also, when $\alpha = 0$ it becomes the Korteweg-de Vries (KdV) equation.

When $f(u) = \frac{3}{2}u^2 + c_1 u + c_0$, $\lambda = 0$, $\gamma = 0$ and $\alpha = 1$ (1) becomes the classical Camassa-Holm (CH) equation, introduced for Camassa and Holm in [3].

The GDGH equation when $\lambda = 0$ was proposed by Shaoyoung Lai and Meng We in [9]. This equation is also generalized as

$$u_t - u_{txx} + (h(u))_x + b u_{xxx} = a\left(\frac{g'(u)}{2}u_x^2 + g(u)u_{xx}\right)_x,$$

see [2].

 ISBN: 978-84-617-8694-7

By using the classical method of symmetry reductions of partial differential equations (PDEs) we obtain transformations under which differential equations are invariants, such transformations bring the GDGH (1) to itself.

When we have an invariant solution, which can be found by symmetry reduction, by using new variables, dependents and independents, we obtain a new differential equation considerably simpler. In the case of two independent variables and one dependent variable, we obtain ordinary differential equations (ODEs). The description of the method can be found in [12, 13].

There are several papers that study the symmetries of particular cases of equation GDGH [1, 2, 5, 7, 8, 10, 14]

The Lie symmetry group of the GDGH (1) will be generated by a vector field of the form

$$X = \xi(x,t,u)\partial_x + \tau(x,t,u)\partial_t + \phi(x,t,u)\partial_u.$$

When we require that the infinitesimal generator leaves invariant the set of solutions of the equation, we obtain, in the general case, the space and time translations, $\mathbf{v}_1 = \partial_x,\ \mathbf{v}_2 = \partial_t$.

The generators of an optimal system of one-dimensional symmetries in the general case are the set: $\{\mathbf{v}_1, \quad \lambda\mathbf{v}_1 + \mathbf{v}_2\}$. With the second generator, we have travelling waves, where the constant λ represents the speed of the wave. In order to reduce (1) to ODEs we obtain the similarity reductions by using the characteristic equations

$$\frac{dx}{\xi(x,t,u)} = \frac{dt}{\tau(x,t,u)} = \frac{du}{\phi(x,t,u)}.$$

The similarity reductions are

$$\left\{ \begin{array}{rcl} z & = & x + \mu\, t, \\ u & = & h(z), \end{array} \right.$$

and the reduced ODE is

$$\alpha^2\, h\, h''' + \mu\, \alpha^2\, h''' - \gamma\, h''' + 2\,\alpha^2\, h'\, h'' + \lambda\,\alpha^2\, h'' - f\, h' - \mu\, h' - \lambda\, h = 0. \tag{2}$$

If we considered h the independent variable, and $y = y(h)$, we can reduce the equation (2) to

$$\left(\alpha^2\, h + \mu\,\alpha^2 - \gamma\right)\, y\, \left(y\, y'' + (y')^2\right) + \alpha^2\, y\, \left(2\, y + \lambda\right)\, y' - \left(f + \mu\right)\, y - \lambda\, h = 0. \tag{3}$$

Solving (2) we obtain travelling waves solutions of the equation GDGH

For some special choices of the constants, other generators are admitted.

- Case 1. $f(u) = \frac{c_1}{2}u^2 - \frac{\gamma}{\alpha^2}u + c_2$.

$$\mathbf{v}_{31} = \gamma \mathrm{e}^{\lambda\, t}\partial_x - \alpha^2 \mathrm{e}^{\lambda\, t}\partial_t + \alpha^2 \lambda \mathrm{e}^{\lambda\, t} u \partial_u.$$

- Case 2. $f(u) = \frac{1}{2}u^2 + c_1 u + c_2$.

$$\mathbf{v}_{32} = \mathrm{e}^{-\lambda\, t}\partial_x - \lambda \mathrm{e}^{-\lambda\, t}\partial_u.$$

 ISBN: 978-84-617-8694-7

- Case 3. $f(u) = \frac{1}{2}u^2 - \frac{\gamma}{\alpha^2}u + c_1$.

$$\mathbf{v}_{32}\ ,\ \mathbf{v}_{43} = \gamma \mathrm{e}^{\lambda\, t}\partial_x - \alpha^2 \mathrm{e}^{\lambda\, t}\partial_t + \lambda\alpha^2 \mathrm{e}^{\lambda\, t} u\partial_u.$$

- Case 4. $f(u) = \frac{c_1}{2}u^2 + c_2 u + c_3$, $\lambda = 0$.

$$\mathbf{v}_{34} = \alpha^2 c_2 + c_1\gamma t\partial_x - \alpha^2(c_1 - 1)t\partial_t + \left(\alpha^2(c_1 - 1)u + \alpha^2 c_2 + \gamma\right)\partial_u.$$

- Case 5. $f(u) = c_2 + c_1\left(u - \frac{\gamma}{\alpha^2}\right)^2$, $\lambda = 0$.

$$\mathbf{v}_{35} = t\partial_t - \left(u - \frac{\gamma}{\alpha^2}\right)\partial_u.$$

- Case 6. $f(u) = \frac{1}{2}u^2 - \frac{\gamma}{\alpha^2}u + c_1$, $\lambda = 0$.

$$\mathbf{v}_{35}\ ,\ \mathbf{v}_{46} = t\partial_x + \partial_u.$$

we generate the optimal system in each case

- Case 1. $\{\mathbf{v}_1,\ \mu\mathbf{v}_1 + \mathbf{v}_2,\ \mu\mathbf{v}_1 + \mathbf{v}_{31}\}$
- Case 2. $\{\mathbf{v}_1,\ \mu\mathbf{v}_1 + \mathbf{v}_2,\ \mu\mathbf{v}_1 + \mathbf{v}_{32}\}$
- Case 3. $\{\mathbf{v}_1,\ \mu\mathbf{v}_1 + \mathbf{v}_2,\ \mathbf{v}_{33},\ \mu\mathbf{v}_{33} + \mathbf{v}_{43}\}$
- Case 4. $\{\mathbf{v}_1,\ \mu\mathbf{v}_1 + \mathbf{v}_2, \mu\mathbf{v}_2 + \mathbf{v}_{34}\}$
- Case 5. $\{\mathbf{v}_1,\ \mu\mathbf{v}_1 + \mathbf{v}_2, \mu\mathbf{v}_1 + \mathbf{v}_{35}\}$
- Case 6. $\{\mathbf{v}_1,\ \mu\mathbf{v}_1 + \mathbf{v}_2, \mathbf{v}_{36}, \mathbf{v}_{46}\}$

From the other generators of the optimal system in the different cases we obtain the reduced equations.

Key words: partial differential equations, symmetries

Acknowledgements

The support of Andalucia FQM-201 group and University of Cádiz is gratefully acknowledged.

References

[1] M. S. Bruzón, M. L. Gandarias, *Classical And Nonclassical Symmetries for the Krichever-Novikov Equation*, Theoretical and Mathematical Physics, 168 (1) (2011) 875-885.

[2] M. S. Bruzón, M. L. Gandarias, J.C. Camacho, J. Ramrez, *Symmetry reductions for a generalized Dullin-Gottwald-Holm equation*, AIP Conference Proceedings , 1479, 1365 (2012)

[3] R. Camassa, D.D. Holm, *An integrable shallow water equation with peaked solitons*, Phys. Rev. Lett, 71 (1993) 1661–1664

[4] H.R. Dullin, G. Gottwald, D.D. Holm, *An integrable shallow water equation with linear and nonlinear dispersion*, Phys. Rev Lett (87),(2001) 194501-194504.

[5] Fuchssteiner B., *Some tricks from the symmetry-toolbox for nonlinear equations: Generalizations of the Camassa-Holm equation*, Physica D: Nonlinear Phenomena, Volume 95, Issue 3, 1996, 229-243,

[6] Guo Z., Ni L., *Wave breaking for the periodic weakly dissipative Dullin-Gottwald-Holm equation*, Nonlinear Analisis. (74) (2011) 965-973.

[7] Gupta, R.K., Anupma, *The Dullin-Gottwald-Holm Equation: Classical Lie Approach and Exact Solutions*, International J. of Nonlinear Science. 10, 146-152 (2010)

[8] R.A. Kraenkel R.A, Senthilvelan M., Zenchuk A.I., *Lie symmetry analysis and reductions of a two-dimensional integrable generalization of the CamassaHolm equation*, Physics Letters A, 273, 3, (2000),

[9] S. Lai M. Wu , *Global weak solutions for a generalized Dullin-Gottwald-Holm equation in the space $H^1(R)$*, Bound Value Probl (2014) 203.

[10] H. Liu, J. Li, Q. Zhang, *Lie symmetry analysis and exact explicit solutions for general Burgerséquation*, Journal of Computational and Applied Mathematics, 228(1) (1996) 1-9.

[11] Novruzov E., *Blow-up of solutions for the dissipative Dullin-Gottwald-Holm equation with arbitrary coefficients*, Jorunal of Differential Equation. (261) (2016) 1115–1127

[12] P.J. Olver, *Applications of Lie groups to differential equations*, Springer-Verlag, 1986.

[13] L.V. Ovsyannikov, *Group Analysis of Differential Equations*, Academic, New York, 1982.

[14] Singh K.,Gupta, R.K., Kumar S., *Exact Solutions of b -family Equation: Classical Lie Approach and Direct Method*, International J. of Nonlinear Science. 11 (1) (2011) 59-67.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

A class of matrices having a set of block diagonal Lyapunov solutions satisfying $\mathcal{R}-$contractivity

A.C. Carapito[1]

[1] *Department of Mathematics, University of Beira Interior*

emails: carapito@ubi.pt

Abstract

We consider a set of square real matrices $\mathcal{A} = \{A_1, A_2, \cdots, A_N\}$, where each matrix is partitioned, in the same way, into blocks such that the diagonal ones are square matrices. Under the assumption that the block (1,1) in the same position have a common Lyapunov solution, a sufficient condition for the existence of a contractive set of block Lyapunov solutions for $\mathcal{A}$ is presented.

Key words: Block matrices, Common Lyapunov solution, Stability

1 Introduction

Consider a finite set $\mathcal{A} = \{A_1, \ldots, A_N\}$ of matrices in $\mathbb{R}^{n\times n}$, $\mathcal{A}$ is said to be *simultaneously stable* if there exists a positive definite and symmetric matrix P which is a Lyapunov solution for every $A_p \in \mathcal{A}$, i.e., if there exists P such that $-(A_p^T P + P A_p)$ is positive definite, for every $A_p \in \mathcal{A}$. The matrix P is called a *common Lyapunov solution* for $\mathcal{A}$, [9]. In practice, simultaneously stable sets of matrices play an important role, for instance, in the study of the stability of a class of hybrid dynamical systems, called linear switched systems where the state evolution is continuous, i.e., the state components are not subject to jumps during switching, see for instance [2, 1, 8]. However, when state discontinuities are allowed at the switching instants, see for instance [7], [5], [6], the stability of a switched system, for every switching law, can be assured by means of the existence of a $\mathcal{R}$-contractive set of Lyapunov solutions, [3], [4].
In this paper we address the problem of the existence of a $\mathcal{R}-$contractive set of Lyapunov solutions $\{P_p : p \in \mathcal{P}\}$ using an analysis in terms of block matrices such that the matrices P_p have a block diagonal structure.

 ISBN: 978-84-617-8694-7

2 Preliminaries

Let $\mathcal{A} = \{A_1, \dots A_N\}$ be a set of matrices in $\mathbb{R}^{n\times n}$. Consider that each matrix $A_p \in \mathcal{A}$ is similarly partitioned into 2×2 blocks as follows:

$$A_p = \begin{bmatrix} A_{11}^p & A_{12}^p \\ A_{21}^p & A_{22}^p \end{bmatrix}, p \in \mathcal{P} = \{1, \cdots, N\}, \tag{1}$$

where $A_{ii}^p \in \mathbb{R}^{n_i \times n_i}$, $i = 1, 2$, for each p and $n_1 + +n_2 = n$.

Let $\mathbb{S}$ a set of symmetric and definite matrices P_p, $p \in \mathcal{P}$ defined as follows

$$P = \begin{bmatrix} P_{11} & P_{12} \\ P_{21} & P_{22} \end{bmatrix} \tag{2}$$

where the block P_{22} is an invertible matrix of order $z \in \{1, 2, \dots, n-1\}$.

Definition 1. *Let $\mathcal{A}$ be a set of matrices in $\mathbb{R}^{n\times n}$ defined as in (1). A set of symmetric matrices $\mathbb{S} = \{P_p, p \in \mathcal{P}\}$ is said a $\mathcal{R}-$contractive set of Lyapunov solutions for a switched system if, for the set of matrices $\mathcal{R} = \left\{ R_{(q,p)} = \begin{bmatrix} I_{n-z} & 0 \\ R_{21}^{(q,p)} & R_{22}^{(q,p)} \end{bmatrix} : p, q \in \mathcal{P} \right\}$, $\mathbb{S}$ satisfies:*

1. $A_p^T P_p + P_p A_p < 0$;
2. $R_{(q,p)}^T P_p R_{(q,p)} \leq P_p$, *for all* $p, q \in \mathcal{P}$

where I_{n-z} denotes the identity matrix of order $n - z$ and $R_{22}^{(q,p)}$, $p, q \in \mathcal{P}$,are invertible matrices.

A necessary condition for the $\mathcal{R}-$contractivity of a set of Lyapunov solutions can be establish in terms of Schur complement, [3].

Theorem 2. *If $\mathbb{S}$ is a $\mathcal{R}-$contractive set of Lyapunov solutions for $\mathcal{A}$, then P_p have comom Schur Complement of order $n - z$, i.e, $P_{11}^p - P_{12}^p (P_{22}^p)^{-1} P_{21}^p = C$ not depends on $p \in \mathcal{P}$.*

Notice that if $\mathbb{S}$ is a set of QLFs with common Shur complement of order n, C, then $P_p = C$, for all p. So, the problem of existence of a set of Lyapunov solutions with common Schur complement of order n reduces to the problem of existence of a common Lyapunov solution. The next theorem gives the relation between the contractivity and the common complement Schur property, [3].

Theorem 3. *Let $\mathcal{A}$ a set of matrices in $\mathbb{R}^{n\times n}$ and $\mathbb{S}$ be a set of symmetric and positive definite matrices of order n. The following statements are equivalent:*

1. *There exists a set of matrices $\mathcal{R} = \left\{ R_{(q,p)} = \begin{bmatrix} I_{n-z} & 0 \\ R_{21}^{(q,p)} & R_{22}^{(q,p)} \end{bmatrix} : p, q \in \mathcal{P} \right\}$ such that $\mathbb{S}$ is a $\mathcal{R}-$contractive set of Lyapunov solutions for $\mathcal{A}$.*
2. *$\mathbb{S}$ is a set of Lyapunov solutions for $\mathcal{A}$ with common Schur complement of order z.*

 ISBN: 978-84-617-8694-7

3 Existence of block diagonal Lyapunov solutions with common Schur complement

In this section, we identify a class of matrices $\mathcal{A} = A_p = \begin{bmatrix} A_{11}^p & A_{12}^p \\ A_{21}^p & A_{22}^p \end{bmatrix}$, $p \in \mathcal{P} = \{1, \cdots, N\}$, having a $\mathcal{R}-$contractive set of Lyapunov solutions, more precisely a set $\{P_p \colon p \in \mathcal{P}\}$ where the matrices P_p have a block diagonal structure with all blocks (1,1), P_{11}^p, equals to some matrix C, this is the same as saying that P_p share a common Schur complement, C. Next we present the following necessary condition,[3].

Theorem 4. *Let $\mathcal{A} = \{A_1, \dots A_N\}$ be a set of stable matrices in $\mathbb{R}^{n\times n}$ defined as in (1) where the blocks (2,2) are square matrices of order z. If there exists a set of block diagonal solutions for $\mathcal{A}$ with common Schur complement C, of order $n - z$, then C is a common Lyapunov solution for the blocks (1,1) and the blocks (2,2) are stable.*

In the sequel, we shall use the previous necessary condition and we suppose that C is a common Lyapunov solution for the blocks (1,1) of A_p and P_{22}^p are Lyapunov solutions for the blocks (2,2) of A_p, $p \in \mathcal{P}$, respectively. Then,

$$\begin{array}{ll} (A_{11}^p)^T C + C A_{11} & = -Q_1^p < \ 0, \text{ for all } p \in \mathcal{P}; \\ (A_{22}^p)^T P_{22}^p + P_{22}^p A_{22}^p & = -Q_2^p < \ 0, \ p \in \mathcal{P}, \end{array} , \tag{3}$$

respectively.

Theorem 5. *Let $\mathcal{A}$ a set of stable matrices partioned into 2×2 blocks satisfying (3). Then, $\{diag(P, P_{22}^p) : p \in \mathcal{P}\}$ is a set of Lyapunov solutions for $\mathcal{A}$ if, for each $p \in \mathcal{P}$, one of the following conditions is satisfied:*

(A) $16\lambda_{\max}\left((A_{12}^p)^T C (Q_1^p)^{-1} C A_{12}^p\right) \lambda_{\max}\left(P_{22}^p A_{21}^p (Q_1^p)^{-1} (A_{21}^p)^T P_{22}^p\right) < \lambda_{\min}^2 (Q_2^p)$

(B) $\lambda_{\min}(Q_1^p) > 4\|C A_{12}^p\|_s \|{A_{21}^p}^T P_{22}^p\| \|(Q_2^p)^{-1}\|$,

where $\|.\|$ denotes the spectral norm.

Proof. Suppose that for each $p \in \mathcal{P}$, (A) is satisfied. If $A_{12}^p = 0$ and $A_{21}^p = 0$ for all $p \in \mathcal{P}$, i.e, $\mathcal{A}$ is a set of block diagonal matrices, it is obvious that $\{\text{diag}(P, P_{22}^p) : p \in \mathcal{P}\}$ is a set of Lyapunov solutions for $\mathcal{A}$. If $A_{12}^p = 0$ or $A_{21}^p = 0$ for all $p \in \mathcal{P}$, i.e, $\mathcal{A}$ is a set of block triangular matrices, then easily we prove that $\{\text{diag}(P, \epsilon_p P_{22}^p) : p \in \mathcal{P}\}$ and $\{\text{diag}(P, \beta_p P_{22}^p) : p \in \mathcal{P}\}$, where

$$0 < \epsilon_p \ < \ \frac{\lambda_{\min}(Q_2^p)}{\lambda_{\max}(P_{22}^p A_{21}^p (Q_1^p)^{-1} (A_{21}^p)^T P_{22}^p)}.$$

and

$$\beta_p \ > \ \frac{\lambda_{\max}\left((A_{12}^p)^T C (Q_1^p)^{-1} C A_{12}^p\right)}{\lambda_{\min}(Q_2^p)},$$

 ISBN: 978-84-617-8694-7

are sets of Lyapunov solutions for $\mathcal{A}$ with respect to lower or upper case.

Otherwise, since each A_p is a sum of a lower block triangular T_p matrix with an upper block triangular matrix W_p

$$A_p = \begin{bmatrix} \frac{1}{2}A_{11}^p & 0 \\ A_{21}^p & \frac{1}{2}A_{22}^p \end{bmatrix} + \begin{bmatrix} \frac{1}{2}A_{11}^p & A_{12}^p \\ 0 & \frac{1}{2}A_{22}^p \end{bmatrix}, p \in \mathcal{P} \tag{4}$$

and the previous cases, we conclude that, for each p, diag $(C, t_p P_{22}^p)$ with

$$0 < t_p \quad < \quad \frac{\lambda_{\min}\left(Q_2^p\right)}{4\lambda_{\max}\left(P_{22}^p A_{21}^p (Q_1^p)^{-1} (A_{21}^p)^T P_{22}^p\right)}$$

is a Lyapunov solution for T_p and diag $(C, w_p P_{22}^p)$

$$w_p \quad > \quad \frac{4\lambda_{\max}\left((A_{12}^p)^T C (Q_1^p)^{-1} C A_{12}^p\right)}{\lambda_{\min}\left(Q_2^p\right)}.$$

is a Lyapunov solution for W_p. Therefore, taking account that (A) is verified, we can choose for each $p \in \mathcal{P}$, $t_p = w_p$. Taking $\alpha_p := t_p = w_p$ we conclude that $\{\text{diag}\,(C, \alpha_p P_{22}^p) : p \in \mathcal{P}\}$ is a set of block diagonal Lyapunov solutions for $\mathcal{A}$. □

Example 6. *The matrices*

$$A_1 = \left[\begin{array}{cc:cc} -2 & -1 & 1 & -1 \\ -0.5 & -2 & 0.5 & 2 \\ \hdashline -1 & 1 & -6 & -9 \\ -1 & -2 & 9 & -6 \end{array}\right] \quad \textit{and} \quad A_2 = \left[\begin{array}{cc:cc} -100 & 0 & -2 & 1 \\ 1 & -120 & -1 & 0 \\ \hdashline 1 & 0 & -10 & -11 \\ -1 & 1 & 150 & 0 \end{array}\right]$$

have a block diagonal Lyapunov solutions with common Schur complement of order 2.

Notice that the block (1,1) of both matrices share a common Lyapunov solution:

$$C = \begin{bmatrix} 1 & -0.2 \\ -0.2 & 1 \end{bmatrix}$$

and the blocks (2,2) are stable (but not admite a common Lyapunov solution). Considering the Lyapunov solutions

$$P_{22}^1 = \textit{diag}(0.08, 0.08) \ \textit{and} \ P_{22}^2 = \begin{bmatrix} 0.7 & 0.04 \\ 0.04 & 0.05 \end{bmatrix}$$

for the blocks (2,2), respectively, we obtain

$$Q_1^1 = \begin{bmatrix} 3.8 & 0.7 \\ 0.7 & 3.6 \end{bmatrix}; \qquad Q_1^2 = \begin{bmatrix} 200.4 & -45 \\ -45 & 240 \end{bmatrix}$$

$$Q_2^1 = \textit{diag}(0.96, 0.96) \qquad (\lambda_{\min}\left(Q_2^1\right))^2 \approx 0.9216;$$
$$Q_2^2 = \begin{bmatrix} 2 & 0.6 \\ 0.6 & 0.88 \end{bmatrix} \qquad (\lambda_{\min}\left(Q_2^2\right))^2 \approx 0.3835.$$

 ISBN: 978-84-617-8694-7

Easily we verify that Condition (A) of previous theorem is satisfied. So, we conclude that

$$\left\{ diag\left(\begin{bmatrix} 1 & -0.2 \\ -0.2 & 1 \end{bmatrix},\ diag(0.08, 0.08) \right), diag\left(\begin{bmatrix} 1 & -0.2 \\ -0.2 & 1 \end{bmatrix}, \begin{bmatrix} 0.7 & 0.04 \\ 0.04 & 0.05 \end{bmatrix} \right) \right\}$$

is a set of block diagonal Lyapunov solutions with common Schur complement of order 2.

Acknowledgements

This work was supported by Portuguese Foundation for Science and Technology ("FCT–Fundação para a Ciência e a Tecnologia"), within project PEst-C/MAT/UI4106/2011 with COMPETE number FCOMP-01-0124-FEDER-022690.

References

[1] D. LIBERZON, *Switching in Systems and Control*, Birkhauser, Boston, 2003.

[2] D. LIBERZON AND A. S. MORSE, *Basic problems in stability and design of switched systems*, IEEE Control Systems **19** (1999) 59–70.

[3] I. BRÁS, A. C. CARAPITO AND P. ROCHA, *Stability of Switched Systems With Partial State Reset*, IEEE Transactions on Automatic Control **58** (2013) 1008–1012.

[4] I. BRÁS, A. C. CARAPITO AND P. ROCHA, *Stability of simultaneously block triangularisable switched systems with partial state reset*, International Journal of Control **90** (2017) 428–437.

[5] J. P. HESPANHA AND A. S. MORSE, *Switching between stabilizing controllers*, Automatica **38** (2002) 1905-1917.

[6] J. P. HESPANHA, P. SANTESSO, AND G. STEWART, *Optimal controller initialization for switching between stabilizing controllers*, 46th IEEE conference on decision and control (2007) 5634–5639.

[7] J. PAXMAN AND G. VINNICOMBE, *Stability of reset switching systems*, Pro. Conf. Decision and Control (2003).

[8] R. SHORTEN, F. WIRTH, O. MASON, K. WULLF, AND C. KING, *Stability criteria for switched and hybrid systems*, SIAM Rev. **49** (2007) 545–592.

[9] T. ANDO, *Set of matrices with common lyapunov solution*, Archiv der Mathematik **77** (2001) 76–84.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

Hyers-Ulam and Hyers-Ulam-Rassias Stability of a Class of Integral Equations on Finite Intervals

L. P. Castro[1] and A. M. Simões[2]

[1] *Center for Research and Development in Mathematics and Applications (CIDMA), University of Aveiro, Aveiro, Portugal*

[2] *Center of Mathematics and Applications of University of Beira Interior (CMA-UBI), University of Beira Interior, Covilhã, Portugal*

emails: castro@ua.pt, asimoes@ubi.pt

Abstract

The purpose of this work is to study different kinds of stability for a class of integral equations defined on a finite interval. Sufficient conditions are derived in view to obtain Hyers-Ulam stability and Hyers-Ulam-Rassias stability by using fixed point techniques and the Bielecki metric.

Key words: Hyers-Ulam stability, Hyers-Ulam-Rassias stability, Banach fixed point theorem, integral equation
MSC 2000: 45M10, 34K20, 47H10

1 Introduction

During the last seven decades the concepts of Hyers-Ulam stability and Hyers-Ulam-Rassias stability for different kinds of functional equations, differential equations, integral equations and others has been studied in a quite extensive way due to their great number of applications e.g. in elasticity, semiconductors, heat conduction, fluid flow, scattering theory, chemical reactions and population dynamic, among others (see [1, 2, 3, 4, 5, 6, 7, 8]). Originated in 1940 from a famous question raised by S. M. Ulam, the first results of stability of this type were about to discover when a solution of an equation differing "slightly" from a given one must be somehow near to the solution of the given equation. A first parcial answer to this question was given by D. H. Hyers, introducing therefore the so-called Hyers-Ulam

 ISBN: 978-84-617-8694-7

stability. New directions were introduced by Th. M. Rassias, see [9], introducing therefore the so-called Hyers-Ulam-Rassias stability.

In this work, we will be devoted to analyse Hyers-Ulam and Hyers-Ulam-Rassias stability for the following class of integral equations:

$$y(x) = f\left(x, y(x), \int_a^b k(x, \tau, y(\tau), y(\alpha(\tau)))d\tau\right), \qquad x \in [a, b], \tag{1}$$

and

$$y(x) = f\left(x, y(x), \int_a^x k(x, \tau, y(\tau), y(\alpha(\tau)))d\tau\right), \qquad x \in [a, b], \tag{2}$$

where a and b are fixed real numbers, $f : [a, b] \times \mathbb{C} \times \mathbb{C} \to \mathbb{C}$ and $k : [a, b] \times [a, b] \times \mathbb{C} \times \mathbb{C} \to \mathbb{C}$ are continuous functions, and $\alpha : [a, b] \to [a, b]$ is a continuous delay function which therefore fulfills $\alpha(\tau) \leq \tau$ for all $\tau \in [a, b]$.

The formal definition of the above mentioned Hyers-Ulam-Rassias stability and Hyers-Ulam stability are now introduced for the integral equation (1).

If for each function y satisfying

$$\left|y(x) - f\left(x, y(x), \int_a^b k(x, \tau, y(\tau), y(\alpha(\tau)))d\tau\right)\right| \leq \sigma(x), \qquad x \in [a, b], \tag{3}$$

where σ is a non-negative function, there is a solution y_0 of the integral equation and a constant $C > 0$ independent of y and y_0 such that $|y(x) - y_0(x)| \leq C\sigma(x)$, for all $x \in [a, b]$, then we say that the integral equation (1) has the Hyers-Ulam-Rassias stability.

If for each function y satisfying

$$\left|y(x) - f\left(x, y(x), \int_a^b k(x, \tau, y(\tau), y(\alpha(\tau)))d\tau\right)\right| \leq \theta, \qquad x \in [a, b], \tag{4}$$

where $\theta \geq 0$, there is a solution y_0 of the integral equation and a constant $C > 0$ independent of y and y_0 such that $|y(x) - y_0(x)| \leq C\theta$, for all $x \in [a, b]$, then we say that the integral equation has the Hyers-Ulam stability.

Some of the present techniques to study the stability of functional equations use a combination of the following well-known Banach Fixed Point Theorem with a generalized metric in appropriate settings.

Theorem 1 *Let (X, d) be a generalized complete metric space and $T : X \to X$ a strictly contractive operator with a Lipschitz constant $L < 1$. If there exists a nonnegative integer k such that $d(T^{k+1}x, T^k x) < \infty$ for some $x \in X$, then the following three propositions hold true:*

i) the sequence $(T^n x)_{n\in\mathbb{N}}$ converges to a fixed point x^* of T;

 ISBN: 978-84-617-8694-7

ii) x^* *is the unique fixed point of* T *in*

$$X^* = \{y \in X : d(T^k x, y) < \infty\}; \tag{5}$$

iii) if $y \in X^*$*, then*

$$d(y, x^*) \leq \frac{1}{1-L}\, d(Ty, y). \tag{6}$$

Let $p > 0$ be a constant, we will be using the space $C_p([a,b])$ of continuous functions $u : [a,b] \to \mathbb{C}$ endowed with the generalized Bielecki metric

$$d_p(u,v) = \sup_{x\in[a,b]} \frac{|u(x) - v(x)|}{e^{p(x-a)}}. \tag{7}$$

We recall that $(C_p([a,b]), d_p)$ is a complete metric spaces (cf., [10]).

2 Hyers-Ulam-Rassias Stability

The present section is devoted to present sufficient conditions for the Hyers-Ulam-Rassias stability of the integral equations (1) and (2).

Theorem 2 *Let* $\alpha : [a,b] \to [a,b]$ *a continuous delay function with* $\alpha(t) \leq t$ *for all* $t \in [a,b]$ *and* $\sigma : [a,b] \to (0,\infty)$ *a non-negative function. Moreover, suppose that* $f : [a,b]\times\mathbb{C}\times\mathbb{C} \to \mathbb{C}$ *is a continuous function satisfying the Lipschitz condition*

$$|f(x,u(x),g(x)) - f(x,v(x),h(x))| \leq M\,(|u(x) - v(x)| + |g(x) - h(x)|) \tag{8}$$

with $M > 0$ *and the kernel* $k : [a,b] \times [a,b] \times \mathbb{C} \times \mathbb{C} \to \mathbb{C}$ *is a continuous kernel function satisfying the Lipschitz condition*

$$|k(x,t,u(t),u(\alpha(t))) - k(x,t,v(t),v(\alpha(t)))| \leq L|u(t) - v(t)| \tag{9}$$

with $L > 0$*.*

If $y \in C_p([a,b])$ *is such that*

$$\left| y(x) - f\left(x, y(x), \int_a^b k(x,\tau,y(\tau),y(\alpha(\tau)))d\tau\right)\right| \leq \sigma(x), \qquad x \in [a,b], \tag{10}$$

and $M\left(1 + \frac{L}{p}\left(e^{p(b-a)} - 1\right)\right) < 1$*, then there is a unique function* $y_0 \in C_p([a,b])$ *such that*

$$y_0(x) = f\left(x, y_0(x), \int_a^b k(x,\tau,y_0(\tau),y_0(\alpha(\tau)))d\tau\right) \tag{11}$$

 ISBN: 978-84-617-8694-7

and

$$|u(x)-y_0(x)| \leq \frac{p\sigma(x)}{p-Mp-ML(e^{p(b-a)}-1)} \tag{12}$$

for all $x \in [a,b]$.

This means that under the above conditions, the integral equation (1) has the Hyers-Ulam-Rassias stability.

Proof. We will consider the operator $T : C_p([a,b]) \to C_p([a,b])$, defined by

$$(Tu)(x) = f\left(x, u(x), \int_a^b k(x,\tau,u(\tau),u(\alpha(\tau)))d\tau\right), \tag{13}$$

for all $x \in [a,b]$ and $u \in C_p([a,b])$.

Under the present conditions, we will deduce that the operator T is strictly contractive with respect to the metric (7). Indeed, for all $u, v \in C_p([a,b])$, we have,

$$\begin{aligned}
d_p(Tu,Tv) &= \sup_{x\in[a,b]} \frac{|(Tu)(x)-(Tv)(x)|}{e^{p(x-a)}} \\
&\leq M \sup_{x\in[a,b]} \frac{1}{e^{p(x-a)}} \{|u(x)-v(x)| + \\
&\qquad \left|\int_a^b k(x,\tau,u(\tau),u(\alpha(\tau)))d\tau - \int_a^b k(x,\tau,v(\tau),v(\alpha(\tau)))d\tau\right|\Big\} \\
&\leq M \sup_{x\in[a,b]} \frac{1}{e^{p(x-a)}} \{|u(x)-v(x)| + \\
&\qquad \int_a^b |k(x,\tau,u(\tau),u(\alpha(\tau))) - k(x,\tau,v(\tau),v(\alpha(\tau)))|\, d\tau\Big\} \\
&\leq M \sup_{x\in[a,b]} \frac{1}{e^{p(x-a)}} \left\{|u(x)-v(x)| + L\int_a^b |u(\tau)-v(\tau)|\, d\tau\right\} \\
&\leq M \left\{\sup_{x\in[a,b]} \frac{|u(x)-v(x)|}{e^{p(x-a)}} + L \sup_{x\in[a,b]} \frac{1}{e^{p(x-a)}} \int_a^b |u(\tau)-v(\tau)|\, d\tau\right\} \\
&= M \left\{\sup_{x\in[a,b]} \frac{|u(x)-v(x)|}{e^{p(x-a)}} + L \sup_{x\in[a,b]} \frac{1}{e^{p(x-a)}} \int_a^b e^{p(\tau-a)} \frac{|u(\tau)-v(\tau)|}{e^{p(\tau-a)}} d\tau\right\} \\
&\leq M \left\{\sup_{x\in[a,b]} \frac{|u(x)-v(x)|}{e^{p(x-a)}} + L \sup_{\tau\in[a,b]} \frac{|u(\tau)-v(\tau)|}{e^{p(\tau-a)}} \sup_{x\in[a,b]} \frac{1}{e^{p(x-a)}} \int_a^b e^{p(\tau-a)} d\tau\right\} \\
&= M \left\{d_p(u,v) + L d_p(u,v) \sup_{x\in[a,b]} \frac{1}{e^{p(x-a)}} \frac{e^{p(b-a)}-1}{p}\right\}
\end{aligned}$$

 ISBN: 978-84-617-8694-7

$$= M\left(1+\frac{L}{p}\left(e^{p(b-a)}-1\right)\right)d_p(u,v). \tag{14}$$

Due to the fact that $M\left(1+\frac{L}{p}\left(e^{p(b-a)}-1\right)\right)<1$ it follows that T is strictly contractive. Thus, we can apply the above mentioned Banach Fixed Point Theorem, which ensures that we have the Hyers-Ulam-Rassias stability for the integral equation (1). Additionally, (12) follows from (6) and (10).

For the Volterra integral equation (2) we have the following result.

Theorem 3 *Let $\alpha:[a,b]\to[a,b]$ a continuous delay function with $\alpha(t)\leq t$ for all $t\in[a,b]$ and $\sigma:[a,b]\to(0,\infty)$ a non-negative function. Moreover, suppose that $f:[a,b]\times\mathbb{C}\times\mathbb{C}\to\mathbb{C}$ is a continuous function satisfying the Lipschitz condition*

$$|f(x,u(x),g(x))-f(x,v(x),h(x))|\leq M\left(|u(x)-v(x)|+|g(x)-h(x)|\right) \tag{15}$$

with $M>0$ and the kernel $k:[a,b]\times[a,b]\times\mathbb{C}\times\mathbb{C}\to\mathbb{C}$ is a continuous kernel function satisfying the Lipschitz condition

$$|k(x,t,u(t),u(\alpha(t)))-k(x,t,v(t),v(\alpha(t)))|\leq L|u(t)-v(t)| \tag{16}$$

with $L>0$.

If $y\in C_p([a,b])$ is such that

$$\left|y(x)-f\left(x,y(x),\int_a^x k(x,\tau,y(\tau),y(\alpha(\tau)))d\tau\right)\right|\leq\sigma(x),\qquad x\in[a,b], \tag{17}$$

and $M\left(1+\frac{L}{p}\left(\frac{e^{p(b-a)}-1}{e^{p(b-a)}}\right)\right)<1$, then there is a unique function $y_0\in C_p([a,b])$ such that

$$y_0(x)=f\left(x,y_0(x),\int_a^x k(x,\tau,y_0(\tau),y_0(\alpha(\tau)))d\tau\right) \tag{18}$$

and

$$|u(x)-y_0(x)|\leq\frac{pe^{p(b-a)}\sigma(x)}{e^{p(b-a)}(p-Mp)-ML(e^{p(b-a)}-1)} \tag{19}$$

for all $x\in[a,b]$.

This means that under the above conditions, the Volterra integral equation (2) has the Hyers-Ulam-Rassias stability.

 ISBN: 978-84-617-8694-7

Proof. We will consider the operator $T : C_p([a,b]) \to C_p([a,b])$, defined by

$$(Tu)(x) = f\left(x, u(x), \int_a^x k(x,\tau,u(\tau),u(\alpha(\tau)))d\tau\right), \tag{20}$$

for all $x \in [a,b]$ and $u \in C_p([a,b])$.

Under the present conditions, we will deduce that the operator T is strictly contractive (with respect to the metric under consideration). Indeed, for all $u, v \in C_p([a,b])$, we have,

$$\begin{aligned}
d_p(Tu,Tv) &= \sup_{x\in[a,b]} \frac{|(Tu)(x)-(Tv)(x)|}{e^{p(x-a)}} \\
&\le M \sup_{x\in[a,b]} \frac{1}{e^{p(x-a)}} \{|u(x)-v(x)| \\
&\qquad + \left|\int_a^x k(x,\tau,u(\tau),u(\alpha(\tau)))d\tau - \int_a^x k(x,\tau,v(\tau),v(\alpha(\tau)))d\tau\right|\Big\} \\
&\le M \sup_{x\in[a,b]} \frac{1}{e^{p(x-a)}} \{|u(x)-v(x)| \\
&\qquad + \int_a^x |k(x,\tau,u(\tau),u(\alpha(\tau))) - k(x,\tau,v(\tau),v(\alpha(\tau)))|\, d\tau\Big\} \\
&\le M \sup_{x\in[a,b]} \frac{1}{e^{p(x-a)}} \left\{|u(x)-v(x)| + L\int_a^x |u(\tau)-v(\tau)|\, d\tau\right\} \\
&\le M \left\{\sup_{x\in[a,b]} \frac{|u(x)-v(x)|}{e^{p(x-a)}} + L \sup_{x\in[a,b]} \frac{1}{e^{p(x-a)}} \int_a^x |u(\tau)-v(\tau)|\, d\tau\right\} \\
&= M \left\{\sup_{x\in[a,b]} \frac{|u(x)-v(x)|}{e^{p(x-a)}} + L \sup_{x\in[a,b]} \frac{1}{e^{p(x-a)}} \int_a^x e^{p(\tau-a)} \frac{|u(\tau)-v(\tau)|}{e^{p(\tau-a)}} d\tau\right\} \\
&\le M \left\{\sup_{x\in[a,b]} \frac{|u(x)-v(x)|}{e^{p(x-a)}} + L \sup_{\tau\in[a,b]} \frac{|u(\tau)-v(\tau)|}{e^{p(\tau-a)}} \sup_{x\in[a,b]} \frac{1}{e^{p(x-a)}} \int_a^x e^{p(\tau-a)} d\tau\right\} \\
&= M \left\{d_p(u,v) + L d_p(u,v) \sup_{x\in[a,b]} \frac{1}{e^{p(x-a)}} \frac{e^{p(b-a)}-1}{p}\right\} \\
&= M \left(1 + \frac{L}{p}\left(\frac{e^{p(b-a)}-1}{e^{p(b-a)}}\right)\right) d_p(u,v).
\end{aligned} \tag{21}$$

Due to the fact that $M\left(1+\frac{L}{p}\left(\frac{e^{p(b-a)}-1}{e^{p(b-a)}}\right)\right) < 1$ it follows that T is strictly contractive. Thus, we can apply the above mentioned Banach Fixed Point Theorem, which ensures that we have the Hyers-Ulam-Rassias stability for the Volterra integral equation (2). Additionally, (19) follows from (6) and (17).

 ISBN: 978-84-617-8694-7

3 Hyers-Ulam Stability

The present section is devoted to present sufficient conditions for the Hyers-Ulam stability of the integral equations (1) and (2).

Theorem 4 *Let $\alpha : [a,b] \to [a,b]$ a continuous delay function with $\alpha(t) \leq t$ for all $t \in [a,b]$. Moreover, suppose that $f : [a,b]\times\mathbb{C}\times\mathbb{C} \to \mathbb{C}$ is a continuous function satisfying the Lipschitz condition*

$$|f(x,u(x),g(x)) - f(x,v(x),h(x))| \leq M\left(|u(x)-v(x)| + |g(x)-h(x)|\right) \tag{22}$$

with $M > 0$ and the kernel $k : [a,b]\times[a,b]\times\mathbb{C}\times\mathbb{C} \to \mathbb{C}$ is a continuous kernel function satisfying the Lipschitz condition

$$|k(x,t,u(t),u(\alpha(t))) - k(x,t,v(t),v(\alpha(t)))| \leq L|u(t)-v(t)| \tag{23}$$

with $L > 0$.

If $y \in C_p([a,b])$ is such that

$$\left| y(x) - f\left(x, y(x), \int_a^b k(x,\tau,y(\tau),y(\alpha(\tau)))d\tau\right)\right| \leq \theta, \qquad x \in [a,b], \tag{24}$$

where $\theta > 0$ and $M\left(1+\frac{L}{p}\left(e^{p(b-a)}-1\right)\right) < 1$, then there is a unique function $y_0 \in C_p([a,b])$ such that

$$y_0(x) = f\left(x, y_0(x), \int_a^b k(x,t,y_0(t),y_0(\alpha(t)))dt\right) \tag{25}$$

and

$$|u(x) - y_0(x)| \leq \frac{p\theta}{p - Mp - ML(e^{p(b-a)}-1)} \tag{26}$$

for all $x \in [a,b]$

This means that under the above conditions, the integral equation (1) has the Hyers-Ulam stability.

Proof. We will consider the operator $T : C_p([a,b]) \to C_p([a,b])$, defined by

$$(Tu)(x) = f\left(x, u(x), \int_a^b k(x,\tau,u(\tau),u(\alpha(\tau)))d\tau\right), \tag{27}$$

for all $x \in [a,b]$ and $u \in C_p([a,b])$.

By the same above procedure we have T strictly contractive with respect to the metric (7) due to the fact that $M\left(1+\frac{L}{p}\left(e^{p(b-a)}-1\right)\right) < 1$. Thus, we can again apply the Banach Fixed Point Theorem, which ensures that we have the Hyers-Ulam stability for the integral equation with (26) being obtained by using (6) and (24).

 ISBN: 978-84-617-8694-7

Now, we consider the Volterra integral equation (2).

Theorem 5 *Let $\alpha : [a,b] \to [a,b]$ a continuous delay function with $\alpha(t) \leq t$ for all $t \in [a,b]$. Moreover, suppose that $f : [a,b]\times\mathbb{C}\times\mathbb{C} \to \mathbb{C}$ is a continuous function satisfying the Lipschitz condition*

$$|f(x,u(x),g(x)) - f(x,v(x),h(x))| \leq M\left(|u(x)-v(x)| + |g(x)-h(x)|\right) \tag{28}$$

with $M > 0$ and the kernel $k : [a,b]\times[a,b]\times\mathbb{C}\times\mathbb{C} \to \mathbb{C}$ is a continuous kernel function satisfying the Lipschitz condition

$$|k(x,t,u(t),u(\alpha(t))) - k(x,t,v(t),v(\alpha(t)))| \leq L|u(t)-v(t)| \tag{29}$$

with $L > 0$.

If $y \in C_p([a,b])$ is such that

$$\left|y(x) - f\left(x, y(x), \int_a^x k(x,\tau,y(\tau),y(\alpha(\tau)))d\tau\right)\right| \leq \theta, \qquad x \in [a,b], \tag{30}$$

where $\theta > 0$ and $M\left(1+\frac{L}{p}\left(\frac{e^{p(b-a)}-1}{e^{p(b-a)}}\right)\right) < 1$, then there is a unique function $y_0 \in C_p([a,b])$ such that

$$y_0(x) = f\left(x, y_0(x), \int_a^x k(x,t,y_0(t),y_0(\alpha(t)))dt\right) \tag{31}$$

and

$$|u(x) - y_0(x)| \leq \frac{pe^{p(b-a)}\theta}{e^{p(b-a)}(p-Mp) - ML(e^{p(b-a)}-1)} \tag{32}$$

for all $x \in [a,b]$

This means that under the above conditions, the Volterra integral equation (2) has the Hyers-Ulam stability.

Proof. We will consider the operator $T : C_p([a,b]) \to C_p([a,b])$, defined by

$$(Tu)(x) = f\left(x, u(x), \int_a^x k(x,\tau,u(\tau),u(\alpha(\tau)))d\tau\right), \tag{33}$$

for all $x \in [a,b]$ and $u \in C_p([a,b])$.

By the same above procedure we have T strictly contractive (with respect to the metric under consideration) due to the fact that $M\left(1+\frac{L}{p}\left(\frac{e^{p(b-a)}-1}{e^{p(b-a)}}\right)\right) < 1$. Thus, we can again apply the Banach Fixed Point Theorem, which ensures that we have the Hyers-Ulam stability for the integral equation with (32) being obtained by using (6) and (30).

Remark 6 *Is possible analyse the Hyers-Ulam-Rassias stability of the integral equation but defined on infinite intervals. These results will be presented in a future work.*

 ISBN: 978-84-617-8694-7

Acknowledgements

This work was supported in part by FCT–*Portuguese Foundation for Science and Technology* through the *Center for Research and Development in Mathematics and Applications (CIDMA)* of University of Aveiro, within UID/MAT/04106/2013, and through the *Center of Mathematics and Applications* of University of Beira Interior, within project UID/MAT/00212/2013.

References

[1] J. Brzdek, D. Popa and I. Rasa, *HyersUlam stability with respect to gauges*, J. of Math. Anal. and Appl. **453**(1) (2017) 620–628.

[2] L. P. Castro and R. C. Guerra, *Hyers-Ulam-Rassias stability of Volterra integral equations within weighted spaces*, Lib. Math. (N.S.) **33**(2) (2013) 21–35.

[3] L. P. Castro and A. Ramos, *Hyers-Ulam and Hyers-Ulam-Rassias stability of Volterra integral equations with a delay*, Integral Methods in Science and Engineering **1** (2010) 85–94.

[4] L. P. Castro and A. M. Simões, *Hyers-Ulam and Hyers-Ulam-Rassias stability of a class of Hammerstein integral equations*, AIP Conference Proceedings **1798:020036** (2017) 1–10.

[5] Y. J. Cho, C. Park, T. M. Rassias and R. Saadati, *Stability of Functional Equations in Banach Algebras*, Springer International Publishing, Switzerland, 2015.

[6] G.-L. Forti, *Hyers-Ulam stability of functional equations in several variables*, Aequationes Math. **50** (1995) 143–190.

[7] S.-M. Jung, *A fixed point approach to the stability of an integral equation related to the wave equation*, Abstr. Appl. Anal. **2013** (2013) 4 pp.

[8] S.-M. Jung, *Hyers-Ulam-Rassias Stability of Functional Equations in Mathematical Analysis*, Hadronic Press, Palm Harbor, 2001.

[9] Th. M. Rassias, *On the stability of the linear mapping in Banach spaces*, Proc. Amer. Math. Soc. **72** (1978) 297–300.

[10] Christopher C. Tisdell and Atiya Zaidi, *Basic qualitative and quantitative results for solutions to nonlinear, dynamic equations on time scales with an application to economic modelling*, Nonlinear Analysis **68** (2008) 3504–3524.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

An efficient technique for the interpolation on compact triangulations

Roberto Cavoretto[1], Alessandra De Rossi[1], Francesco Dell'Accio[2] and Filomena Di Tommaso[2]

[1] *Department of Mathematics "Giuseppe Peano", University of Torino*

[2] *Department of Mathematics and Computer Science, University of Calabria*

emails: roberto.cavoretto@unito.it, alessandra.derossi@unito.it, francesco.dellaccio@unical.it, ditommaso@mat.unical.it

Abstract

In this paper we present an efficient scheme for the computation of triangular Shepard method. More precisely, it is well known that the triangular Shepard method reaches an approximation order better than the Shepard one [4], but it needs to identify useful general triangulation of the node set. Here we propose a searching technique used to detect and select the nearest neighbor points in the interpolation scheme [2, 3]. It consists in determining the closest points belonging to the different neighborhoods and consequently applies to the triangulation-based approach. Numerical results show efficiency of the interpolation procedure.

Key words: scattered data interpolation, triangular Shepard method, fast computation, approximation algorithms
MSC 2000: 65D05, 65D15, 41A05

1 Introduction

Scattered data consists of a set of points $X_n = \{x_1, \ldots, x_n\}$ and corresponding functional values $f_1, \ldots, f_n$, where the points have no structure or order between their relative locations. Among the various approaches to interpolating scattered data, the Shepard method [8] is one of the earliest techniques. It defines an interpolating function as a convex combination of the functional values, that is a linear combination of them with non negative coefficients (or weight functions or basis functions) which are inverse distances to the scattered points and form a partition of unity. The main drawback of the Shepard method is

 ISBN: 978-84-617-8694-7

its low polynomial precision (only constants) that badly affects the reconstructed surface. Several variants of the Shepard method have been considered to overcome this drawback. Among them, the triangular Shepard method [7] is a convex combination of local linear interpolants with triangle-based weight functions, which are the product of inverse distances from the vertices of triangles and form a partition of unity. The main feature of the triangular Shepard method is its linear precision without using derivative data, although, for achieving a good accuracy of approximation, it needs to identify useful general triangulation of the node set.

An efficient organization of the scattered data, when local interpolation is used, turns out to be crucial. To this aim, in literature, techniques known as kd-trees, which are not specifically implemented for a specific interpolation scheme, have already been designed, [1, 6]. In this paper, we propose the use of a versatile partitioning structure, called the block-based partitioning structure, given in [3], which is suitably adapted to triangular Shepard interpolation.

The paper is organized as follows. In Section 2 we consider interpolation using the triangular Shepard method and the selection of the compact triangulation. Section 3 is devoted to present the searching technique used to detect and select the nearest neighbor points in our interpolation scheme. Finally, Section 4 shows some numerical results.

2 Interpolation on compact triangulations

2.1 Triangular Shepard method

In 1983 Little [7] introduced a variant of the Shepard method, called triangular Shepard, which is a Shepard-like convex combination of the linear interpolants of a set of triangles. More precisely, if we denote by $X_n = \{x_1, x_2, \ldots, x_n\}$ a set of nodes of $\mathbb{R}^2$ with associated function data $f_1, \ldots, f_n$ and by $T_m = \{t_1, t_2, \ldots, t_m\}$ a set of triangles which vertices are points of X_n, the triangular Shepard operator is defined by

$$K_\mu[f](x) = \sum_{j=1}^{m} B_{\mu,j}(x) L_j(x), \quad \mu > 0, \tag{1}$$

where $L_j(x)$ is the linear interpolant on the vertices of t_j, $j = 1, \ldots, m$, and the weight functions $B_{\mu,j}(x)$ are defined by

$$B_{\mu,j}(x) = \frac{\prod_{\ell=1}^{3} \frac{1}{\|x - x_{j_\ell}\|^\mu}}{\sum_{k=1}^{m} \prod_{\ell=1}^{3} \frac{1}{\|x - x_{k_\ell}\|^\mu}}, \qquad j = 1, \ldots, m. \tag{2}$$

 ISBN: 978-84-617-8694-7

As noticed by Little himself, the triangular Shepard operator (1) exceeds the Shepard method both in polynomial precision and esthetic behaviour. In fact, it has linear precision while Shepard method achieves only constant precision. The better polynomial precision of the triangular Shepard method reflects on a higher order of approximation. As shown in [4] the triangular Shepard method reaches quadratic approximation order while the Shepard method achieves at most linear approximation order [5]. We remark that the definition of the triangular Shepard operator (1) requires an appropriate list of triangles to be identified. However, these triangles can realize only a general triangulation of the node set X_n, that is a triangulation in which some triangles may overlap or be disjoint.

2.2 Selection of the compact triangulation

In order to identify useful general triangulation of the node set X_n we take into account theoretical results achieved in [4], which link the bound for the remainder term of the linear interpolant $L_j(x)$ with the bound for the remainder of the triangular Shepard operator (1). In particular, all triangles should have a rather regular form (as near as possible to equilateral triangles) and two quantities, denoted by h' and h'', play a key role in the choice of these triangles: the first one is the fill distance in the maximum norm and controls the uniformity of the triangle distribution, the second one excludes the presence of large triangles.

For each node x_i of the node set X_n we choose, among the 15 triangles with a vertex in x_i and other 2 vertices among its 6 nearest neighbors in X_n, the one which locally reduces the bound

$$2\|x - x_{j_1}\|^2 + 4h_j C_j \|x - x_{j_1}\| \tag{3}$$

for the error of the local linear interpolant, where x_{j_1} denotes the first vertex of t_j, h_j denotes the maximum edge length of t_j and C_j is a constant which depends only on the shape of t_j. We omit duplicate triangles and we get a triangulation T_m of the nodes with $m \leq n$ triangles, in which some of the triangles may overlap or be disjoint.

3 Localizing searching technique

In this section we present the searching technique used to detect and select the nearest neighbor points in our interpolation scheme [2, 3]. Though this procedure can be applied on a generic domain $R \subseteq \mathbb{R}^2$ and in higher dimensions, for our purpose we here pursue this description focusing on the unit square, i.e. $R = [0, 1] \times [0, 1]$.

Firstly, we define a circular neighborhood of radius

$$\delta = \frac{2}{d}, \tag{4}$$

 ISBN: 978-84-617-8694-7

with

$$d = \left\lfloor \frac{\sqrt{n}}{2} \right\rfloor, \tag{5}$$

where each neighborhood is centred at a data point belonging to R. To localize points in our method, we set $n/d^2 = 4$. Note that the larger (smaller) the value of d is, the more (less) localizing the scheme is.

In order to determine the closest points belonging to the different neighborhoods and consequently apply our triangulation-based approach, we propose a new structure that partitions the domain in blocks of square shape. Such technique results in an effective searching procedure which is quite efficient from a computational viewpoint. For this scope we partition the domain R with b^2 square blocks, b being the number of blocks along one side of the unit square defined as

$$b = \left\lceil \frac{1}{\delta} \right\rceil. \tag{6}$$

It follows that the side of each square block turns out to be equal to the neighborhood radius. Hence, such a choice (seemingly trivial) allows us to examine in the searching process only a small number of blocks, significantly reducing the computational effort compared to standard or more advanced searching procedures such kd-trees [1, 9]. In fact, our searching routine is performed in a constant time, independently from the initial number of nodes considered.

In our partitioning technique square blocks are numbered from 1 to b^2, following the lexicographic order "bottom to top, left to right". By a repeated use of a quicksort routine the set X_n is thus partitioned by the block-based partitioning structure into b^2 subsets X_{n_k}, $k = 1, \ldots, b^2$, where X_{n_k} are the points belonging to nine blocks: the k-th block and its eight neighboring blocks, see Figure 1. In such framework, we are able to get an optimal procedure to find the interpolation nodes closest to each of points.

4 Numerical results

In our numerical experiments we report results obtained by using the triangular Shepard interpolant (1). All the experiments are carried out by means of a grid of $n_e = 21 \times 21$ evaluation points on $R = [0, 1] \times [0, 1]$ by using several sets of Halton points [6]. The used test function is the Franke's one,

$$\begin{aligned} f(x,y) = 0.75 \exp\left(-\frac{(9x-2)^2 + (9y-2)^2}{4}\right) + 0.50 \exp\left(-\frac{(9x-7)^2 + (9y-3)^2}{4}\right) \\ +0.75 \exp\left(-\frac{(9x+1)^2}{49} - \frac{(9y+1)}{4}\right) - 0.20 \exp\left(-(9x-4)^2 - (9y-7)^2\right). \end{aligned}$$

 ISBN: 978-84-617-8694-7

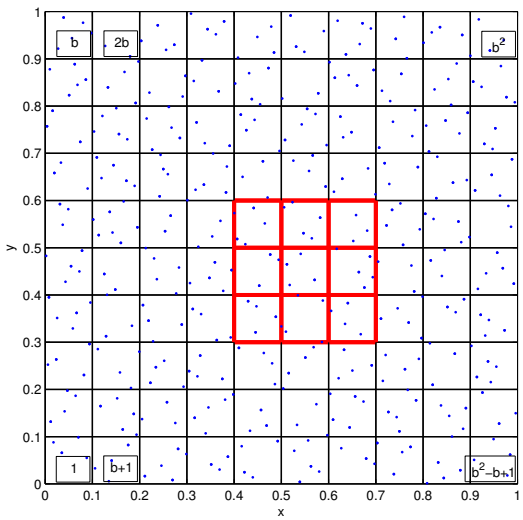


Figure 1: Example of 2D block-based partitioning structure with a data point set: in red the k-th block and its eight neighboring blocks, in blue a set of scattered points contained in the unit square R.

n	RMSE	t_{old}	t_{new}
10000	2.87e-4	27.3056	7.0784
20000	1.53e-4	242.5568	18.1962
40000	7.51e-5	–	49.3938
80000	3.59e-5	–	245.2871

Table 1: CPU times (in seconds) and RMSEs on a grid of $n_e = 21 \times 21$ evaluation points for the Franke's test function using different sets of n Halton points.

In Table 1 we show the CPU times computed in seconds and the root mean square error (RMSE)

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n_e} {e_i}^2}{n_e}}$$

with

$$e_i = |f(P_i) - K_2[f](P_i)|, \tag{7}$$

P_i being an evaluation point in R. In particular, we compare performance of the procedure which computes all the distances between the scattered points (t_{old}) with the one using the block-based searching technique (t_{new}).

 ISBN: 978-84-617-8694-7

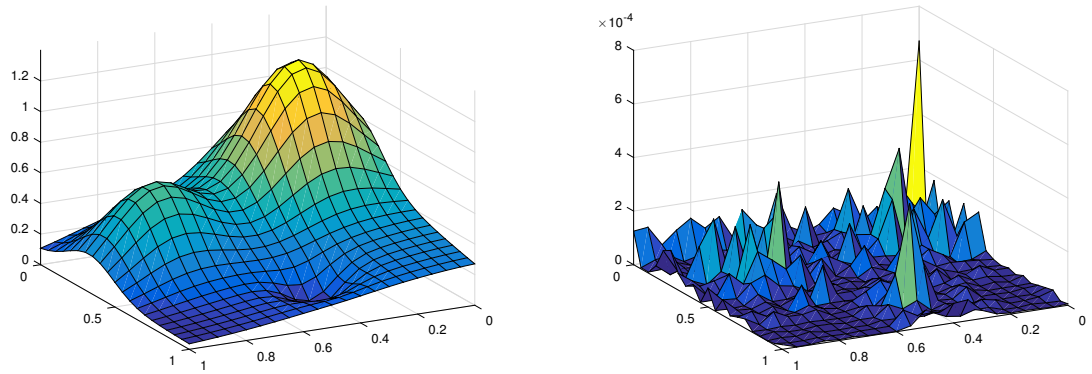


Figure 2: Approximation of Franke's test function (left) and the absolute errors (right) on a grid of $n_e = 21 \times 21$ evaluation points using $n = 40000$ Halton points.

Figure 2 shows the reconstruction of Franke's test function and the absolute values e_i on $n = 40000$ Halton points.

Acknowledgements

This research has been accomplished within the RITA "Research ITalian network on Approximation". This work was partially supported by the projects "Metodi e modelli numerici per le scienze applicate" and "Approssimazione multivariata e algoritmi efficienti con applicazioni a problemi algebrici, differenziali e integrali" of the Department of Mathematics of the University of Torino. Moreover, this research was supported by GNCS–INdAM 2016-2017 project and by a research fellow of the Centro Universitario Cattolico.

References

[1] S. Arya, D.M. Mount, N.S. Netanyahu, R. Silverman, A.Y. Wu, *An optimal algorithm for approximate nearest neighbor searching in fixed dimensions*, J. ACM **45** (1998) 891–923.

[2] R. Cavoretto, A. De Rossi, *A meshless interpolation algorithm using a cell-based searching procedure*, Comput. Math. Appl. **67** (2014) 1024–1038.

[3] R. Cavoretto, A. De Rossi, E. Perracchione, *Efficient computation of partition of unity interpolants through a block-based searching technique*, Comput. Math. Appl. **71** (2016) 2568–2584.

[4] F. Dell'Accio, F. Di Tommaso, K. Hormann, *On the approximation order of the triangular Shepard interpolation*, IMA J. Numer. Anal. **36** (2016) 359–379.

[5] R. Farwig, *Rate of convergence of Shepard's global interpolation formula*, Math. Comp **46** (1986) 577–590.

[6] G.E. Fasshauer, *Meshfree approximation methods with MATLAB*, World Scientific, Singapore, 2007.

[7] F. Little, *Convex combination surfaces*, in R.E. Barnhill, W. Boehm (eds.), *Surfaces in Computer Aided Geometric Design*, North-Holland, Amsterdam, 1983.

[8] D. Shepard, *A two-dimensional interpolation function for irregularly-spaced data.*, in *Proceedings of the 23rd ACM National Conference*, New York: ACM Press, pp. 517524.

[9] H. Wendland, *Scattered data approximation*, Cambridge University Press, Cambridge, 2005.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

Surface approximation of basins of attraction through RBF interpolation schemes

Roberto Cavoretto[1], Alessandra De Rossi[1] and Emma Perracchione[2]

[1] *Department of Mathematics "G. Peano", University of Torino*

[2] *Department of Mathematics "T. Levi-Civita", University of Padova*

emails: roberto.cavoretto@unito.it, alessandra.derossi@unito.it, emma.perracchione@math.unipd.it

Abstract

In this paper we present the problem of determining the so-called *basins of attraction* of dynamical systems. They are found out so that there exist manifolds partitioning the phase space into different regions. The reconstruction of such surfaces is carried out by means of meshfree interpolation tools and specifically we apply the Partition of Unity (PU) method with local Radial Basis Function (RBF) interpolants.

Key words: scattered data approximation, partition of unity method, radial basis functions, dynamical systems, competition population models, basins of attraction
MSC 2000: 65D05, 65D17, 92D25, 37M20

1 Introduction

Over the last years the topic of numerical approximation of multivariate data has gained popularity in various disciplines, such as numerical solution of PDEs, image registration, neural networks, optimization, statistics and finance. In what follows we investigate an application to population dynamics [1].

Nowadays, mathematical modeling is commonly applied to major disciplines and by these models the prediction of the temporal evolution of the considered quantities, i.e. populations, cancer, divorces, is sought [9]. This is obtained in general via dynamical systems. Here we present a reliable algorithm for the reconstruction of unknown manifolds partitioning the phase state of dynamical systems into disjoint sets. Indeed, in an initial value problem, involving a set of ordinary differential equations, a particular solution of the

system is completely determined by the Initial Condition (IC). Depending on the initial state of the system and on conditions involving the model parameters, the trajectories may in fact tend towards different equilibria.

The phase state of the dynamical system is thus partitioned into different regions, called the basins of attraction of each equilibrium, depending on where the trajectories originating in them will ultimately stabilize. In such cases, the final outcome of a mathematical model depends on the IC. If it lies in the basin of attraction of a certain equilibrium point, the system will finally settle to this specific steady state. To establish the ultimate system behavior, it is therefore important to assess for each possible attractor its domain of attraction.

Thus, we present a tool that allows to reconstruct the basin of attraction of each equilibrium, providing a graphical representation of the separatrix manifold [4, 5]. First, a suitable scheme is constructed for the generation of these manifolds. It provides points that, within a certain tolerance, lie on these sought manifolds. This is obtained via a suitable bisection-like routine that employs pairs of points belonging to two different sets of the partition. Then, since an attraction basin can be described by an implicit equation, we interpolate such points with the implicit PU method using local RBF approximants [6, 8, 10].

The model we consider throughout this paper, involving three populations P, Q and R, reads as follows

$$\begin{aligned}
\frac{dP}{dt} &= p\left(1-\frac{P}{u}\right)P - aPQ - bPR,\\
\frac{dQ}{dt} &= q\left(1-\frac{Q}{v}\right)Q - cPQ - eQR,\\
\frac{dR}{dt} &= r\left(1-\frac{R}{w}\right)R - fPR - gQR,
\end{aligned} \tag{1}$$

where p, q and r are the growth rates of P, Q and R, respectively, a, b, c, e, f and g are the competition rates, u, v and w are the carrying capacities of the three populations. The model describes the interaction of three competing populations within the same environment and has eight equilibria. For simplicity, we list here only those playing a role in this investigation, i.e. the origin $E_0 = (0, 0, 0)$, the points associated with the survival of only one population

$$E_1 = (u, 0, 0), \qquad E_2 = (0, v, 0), \qquad E_3 = (0, 0, w)$$

 ISBN: 978-84-617-8694-7

and the coexistence equilibrium

$$E_* = \left(\frac{u[p(gvwe - qr) - avr(we - q) - bwq(vg - r)]}{p(gvwe - qr) + uva(rc - fwe) + uwb(fq - gcv)}, \right.$$
$$\frac{v[q(fuwb - pr) - rcu(wb - p) - pew(fu - r)]}{q(fuwb - pr) + cuv(ra - gwb) + evw(gp - afu)},$$
$$\left. \frac{r[(cuva - pq) - gpv(cu - q) - ufq(va - p)]}{r(cuva - pq) + bwu(fq - vcg) + evw(gp - fua)} \right).$$

With the parameter setting $p = 1$, $q = 2$, $r = 2$, $a = 5$, $b = 4$, $c = 3$, $e = 7$, $f = 7$, $g = 10$, $u = 3$, $v = 2$, $w = 1$, the points associated with the survival of only one population are stable, the origin E_0 is an unstable equilibrium and the coexistence equilibrium E_* is a saddle point. The manifolds that partition the phase space into the different domains of attraction intersect only at the coexistence saddle point E_*.

2 Implicit RBF-PU approximation

In this section we present the method used to approximate the basins of attraction, i.e. the separatrix manifolds. Since they are often described by implicit surfaces, we consider the implicit PU approximation using locally radial kelnels or RBFs [2]. Such surfaces are defined by a point cloud data set $X_N = \{\boldsymbol{x}_i \in \mathbb{R}^3,\, i = 1, \ldots, N\}$, which belongs to a surface in $\mathbb{R}^3$.

2.1 Radial kernels

Generally, to recover a function $f : \Omega \to \mathbb{R}$ on a bounded domain $\Omega \subset \mathbb{R}^3$, we use a set of samples of f on N pairwise distinct data points or nodes $X_N \subset \Omega$, i.e. $\boldsymbol{f} = [f_1, \ldots, f_N]^T$, with $f_i = f(\boldsymbol{x}_i)$, $\boldsymbol{x}_i \in X_N$. For this aim, we define a positive definite and symmetric kernel $\Phi : \Omega \times \Omega \to \mathbb{R}$, obtaining the interpolant expressed as follows

$$u(\boldsymbol{x}) = \sum_{j=1}^{N} c_j \Phi(\boldsymbol{x}, \boldsymbol{x}_j), \qquad \boldsymbol{x} \in \Omega. \tag{2}$$

Here Φ is a radial kernel depending on a positive *shape parameter* ε for all $\boldsymbol{x}, \boldsymbol{z} \in \Omega$, i.e.

$$\Phi(\boldsymbol{x}, \boldsymbol{z}) = \phi_\varepsilon(||\boldsymbol{x} - \boldsymbol{z}||_2) = \phi(\varepsilon||\boldsymbol{x} - \boldsymbol{z}||_2),$$

where $\phi : \mathbb{R}_{\geq 0} \to \mathbb{R}$ defines a radial basis function. Table 1 shows a list of a few strictly positive definite radial kernels with their smoothness degrees [10].

The coefficients $\boldsymbol{c} = [c_1, \ldots, c_N]^T$ in (2) are found by solving the system of linear equations

$$A\boldsymbol{c} = \boldsymbol{f}, \tag{3}$$

 ISBN: 978-84-617-8694-7

RBF	$\phi_\varepsilon(r)$
Inverse MultiQuadric C^∞ (IMQ)	$(1+\varepsilon^2 r^2)^{-1/2}$
Matérn C^6 (M6)	$\mathrm{e}^{-\varepsilon r}(\varepsilon^3 r^3 + 6\varepsilon^2 r^2 + 15\varepsilon r + 15)$
Matérn C^4 (M4)	$\mathrm{e}^{-\varepsilon r}(\varepsilon^2 r^2 + 3\varepsilon r + 3)$
Wendland C^6 (W6)	$(1-\varepsilon r)_+^8\,(32\varepsilon^3 r^3 + 25\varepsilon^2 r^2 + 8\varepsilon r + 1)$
Wendland C^4 (W4)	$(1-\varepsilon r)_+^6\,(35\varepsilon^2 r^2 + 18\varepsilon r + 3)$

Table 1: Some examples of strictly positive definite radial kernels in $\mathbb{R}^3$.

where entries of the interpolation matrix are given by

$$A_{ij} = \Phi(\boldsymbol{x}_i, \boldsymbol{x}_j), \quad i,j = 1,\ldots,N.$$

The resulting solution u is a function of the *native Hilbert space* $\mathcal{N}_\Phi(\Omega)$ uniquely associated with the kernel, and, if $f \in \mathcal{N}_\Phi(\Omega)$, it is in particular the $\mathcal{N}_\Phi(\Omega)$-projection of f into the subspace

$$\mathcal{N}_\Phi(X_N) = \mathrm{span}\{\Phi(\boldsymbol{x}, \boldsymbol{x}_j), \boldsymbol{x}_j \in X_N\}$$

spanned by the standard basis of translates $\mathcal{T}_{X_N} = \{\Phi(\boldsymbol{x}, \boldsymbol{x}_j), 1 \le j \le N\}$, see [6, Ch. 14].

However, since in some situations the matrix A in (3) might turn out to be very ill-conditioned and, accordingly, the interpolant (2) unstable, many efforts have recently been made to derive more stable bases (see [7] for an overview).

2.2 Implicit PU interpolation

In order to determine the implicit PU interpolant, we consider additional interpolation conditions taking an extra set of off-surface points. Hence we generate the extra off-surface points referring to a small step away along the surface normals $\boldsymbol{n}_i$. This enables us to obtain for each node $\boldsymbol{x}_i$ two additional off-surface points: the former defined as $\boldsymbol{x}_{N+i} = \boldsymbol{x}_i + \delta \boldsymbol{n}_i$ is located outside the surface, the latter expressed in the form $\boldsymbol{x}_{2N+i} = \boldsymbol{x}_i - \delta \boldsymbol{n}_i$ is set inside, δ being the stepsize [6, Ch. 30].

After creating the data point set, we can construct the PU interpolant, whose zero contour or iso-surface interpolates three sets, i.e., X_N, $X_\delta^+ = \{\boldsymbol{x}_{N+1}, \ldots, \boldsymbol{x}_{2N}\}$ and $X_\delta^- = \{\boldsymbol{x}_{2N+1}, \ldots, \boldsymbol{x}_{3N}\}$.

The idea of the PU method is to decompose a (usually) large problem or domain $\Omega \subseteq \mathbb{R}^3$ into d small problems or subdomains Ω_j such that $\Omega \subseteq \bigcup_{j=1}^d \Omega_j$ with some mild overlap among the subdomains. Associated with these subdomains we construct a partition of

 ISBN: 978-84-617-8694-7

unity, i.e. a family of compactly supported, non-negative, continuous functions w_j with $\text{supp}(w_j) \subseteq \Omega_j$ such that

$$\sum_{j=1}^{d} w_j(\boldsymbol{x}) = 1, \qquad \boldsymbol{x} \in \Omega.$$

Therefore the global approximant may be expressed as follows

$$P(\boldsymbol{x}) = \sum_{j=1}^{d} u_j(\boldsymbol{x}) w_j(\boldsymbol{x}), \qquad \boldsymbol{x} \in \Omega, \tag{4}$$

where $w_j : \Omega_j \to \mathbb{R}$ defines the *Shepard weight*

$$w_j(\boldsymbol{x}) = \frac{\varphi_j(\boldsymbol{x})}{\sum_{k=1}^{d} \varphi_k(\boldsymbol{x})},$$

φ_j being the compactly supported Wendland C^{2k}, $k \geq 1$ functions [10]. The local RBF interpolant $u_j : \Omega_j \to \mathbb{R}$ in (4) assumes the form

$$u_j(\boldsymbol{x}) = \sum_{i=1}^{N_j} c_i^j \phi(||\boldsymbol{x} - \boldsymbol{x}_i^j||_2),$$

where N_j indicates the number of data points in Ω_j, i.e. the points $\boldsymbol{x}_i^j \in X_j = X_N \cap \Omega_j$.

The PU approach is thus a simple and effective computational method that allows us to decompose a large problem into many small subproblems, ensuring that the accuracy obtained for the local fits is carried over to the global one (for further details see [3, 10]).

3 Numerical experiments

In this numerical section we show how the implicit PU method can be applied to reconstruct the domains of attraction of dynamical systems presenting three stable equilibria.

Specifically, using the routine given in [5], we can approximate the basins of attraction of the system (1). In fact, considering n equispaced points on each edge of the cube $[0, \gamma]^3$, with $\gamma \in \mathbb{R}^+$, we can define a set of ICs, i.e.

$$\begin{array}{llll} P^1_{i_1,i_2} = (x_{i_1}, y_{i_2}, 0) & \text{and} & P^2_{i_1,i_2} = (x_{i_1}, y_{i_2}, \gamma), & i_1, i_2 = 1, \ldots, n, \\ P^3_{i_1,i_2} = (x_{i_1}, 0, z_{i_2}) & \text{and} & P^4_{i_1,i_2} = (x_{i_1}, \gamma, z_{i_2}), & i_1, i_2 = 1, \ldots, n, \\ P^5_{i_1,i_2} = (0, y_{i_1}, z_{i_2}) & \text{and} & P^6_{i_1,i_2} = (\gamma, y_{i_1}, z_{i_2}), & i_1, i_2 = 1, \ldots, n. \end{array}$$

Taking then such points in pairs and applying a bisection algorithm, we find a certain number of separatrix points that lie on the basins of attraction. As an example, the points

 ISBN: 978-84-617-8694-7

lying on the separatrix manifolds depicted in Figure 1 have been found assuming $n = 15$ and $\gamma = 6$.

In Figure 1 we report the plot of the three surfaces showing the domains of attraction. The latter have been approximated by using the method (4) described in Section 2. Such surfaces have been obtained by taking $d = 4$ subdomains and the W6 function in Table 1 with $\varepsilon = 0.1$. In this case the parameters used are $p = 1$, $q = 2$, $r = 2$, $a = 5$, $b = 4$, $c = 3$, $e = 7$, $f = 7$, $g = 10$, $u = 3$, $v = 2$ and $w = 1$.

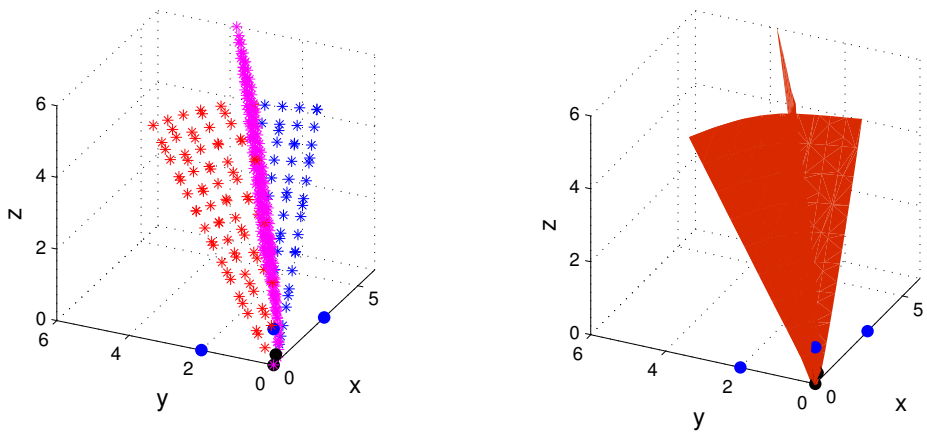


Figure 1: Detection of points on the surfaces determining the basins of attraction (left) and approximation of the domains of E_1, E_2 and E_3 (right). Stable equilibria are marked by a blue dot, unstable saddle points E_0 and E_* are represented by a black dot.

Acknowledgements

This research has been accomplished within the RITA "Research ITalian network on Approximation". This work was partially supported by the projects "Metodi e modelli numerici per le scienze applicate" and "Approssimazione multivariata e algoritmi efficienti con applicazioni a problemi algebrici, differenziali e integrali" of the Department of Mathematics of the University of Torino. Moreover, the authors acknowledge support from the GNCS–INdAM.

References

[1] D. K. Arrowsmith, C. K. Place, *An Introduction to Dynamical Systems*, Cambridge Univ. Press, Cambridge, 1990.

 ISBN: 978-84-617-8694-7

[2] M. D. Buhmann, *Radial Basis Functions: Theory and Implementation*, Cambridge Monogr. Appl. Comput. Math., vol. 12, Cambridge Univ. Press, Cambridge, 2003.

[3] R. Cavoretto, A. De Rossi, E. Perracchione, *Efficient computation of partition of unity interpolants through a block-based searching technique*, Comput. Math. Appl. **71** (2016), 2568–2584.

[4] R. Cavoretto, A. De Rossi, E. Perracchione, E. Venturino, *Robust approximation algorithms for the detection of attraction basins in dynamical systems*, J. Sci. Comput. **68** (2016), 395–415.

[5] R. Cavoretto, A. De Rossi, E. Perracchione, E. Venturino, *Graphical representation of separatrices of attraction basins in two and three dimensional dynamical systems*, Int. J. Comput. Methods 14 (2017), 1750008, 16 pp.

[6] G. E. Fasshauer, *Meshfree Approximation Methods with* Matlab, World Scientific, Singapore, 2007.

[7] G. Fasshauer, M. McCourt, *Kernel-based Approximation Methods using* Matlab, World Scientific, Singapore, 2015.

[8] H. Hoppe, *Surface Reconstruction from Unorganized Points*, PhD thesis, University of Washington, 1994.

[9] J. D. Murray, *Mathematical Biology*, Springer-Verlag, Berlin, 2002.

[10] H. Wendland, *Scattered Data Approximation*, Cambridge Monogr. Appl. Comput. Math., vol. 17, Cambridge Univ. Press, Cambridge, 2005.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

(Pseudo)digraphs and Leibniz algebra isomorphisms

Manuel Ceballos[1], Juan Núñez[2] and Ángel F. Tenorio[3]

[1] *Departamento de Ingeniería, Universidad Loyola Andalucía*

[2] *Departamento de Geometría y Topología, Facultad de Matemáticas. Universidad de Sevilla.*

[3] *Dpto. de Economía, Métodos Cuantitativos e Historia Económica, Escuela Politécnica Superior. Universidad Pablo de Olavide.*

emails: mceballos@uloyola.es, jnvaldes@us.es, aftenorio@upo.es

Abstract

In this paper we study the link between isomorphic digraphs and isomorphic Leibniz algebras, determining in detail this fact when using (pseudo)digraphs of 2 and 3 vertices associated with Leibniz algebras according to their isomorphism classes. Moreover, we introduce and implement an algorithmic procedure which allows us to decide if a given combinatorial structure is associated or not with a Leibniz algebra. Finally, and as application of this algorithm, we give the complete list with all the non-isomorphic combinatorial structures of 3 vertices associated with Leibniz algebras, studying the Leibniz-algebra structure associated with each configuration.

Key words: (Pseudo)digraph, Combinatorial structure, Leibniz algebra, Isomorphism class, Algorithm
MSC 2000: 17A32, 17A60, 05C25, 05C20, 05C90, 68W30, 68W40, 68Q25.

1 Introduction

Research on non-associative algebras is very extensive due to both its own theoretical relevance and its applications to many different fields, like Engineering, Physics or Applied Mathematics. Within these algebras, we will study Leibniz algebras. These algebras were introduced at the beginning of the 1990s by Loday [3] as a particular type of non-associative algebras providing a non-commutative generalization of Lie algebras. However, many general questions about them have not been solved at present by means of traditional techniques, such as obtaining their classification.

Nowadays, Graph Theory has become an essential tool to solve a wide range of problems in different research fields. In this way, we think that graphs and simplicial complexes (their generalization to higher dimensions) may be used to study non-associative algebras and solve open problems like the above-mentioned problem of classifying Leibniz algebras.

The main goal of this paper is to study the link between combinatorial structures and Leibniz algebras, taking into account some previous papers like [1, 2]. More concretely, we determine all the isomorphism classes of Leibniz algebras associated with combinatorial structures. We also introduce and implement an algorithm to decide if a given combinatorial structure is associated or not with a Leibniz algebra. Finally, we give a complete list with all non-isomorphic combinatorial structures of 3 vertices associated with each isomorphism class of Leibniz algebras, indicating the structure of each algebra.

2 Preliminaries

We recall some preliminary concepts on Leibniz algebras, bearing in mind that the reader can consult [3] as an introductory paper.

Definition 1 *A* Leibniz algebra $\mathcal{L}$ *over a field* $\mathbb{K}$ *is a vector space with a second inner bilinear composition law* $[\cdot,\cdot]$*, which verifies the so-called Leibniz identity*

$$[[X,Y],Z]-[[X,Z],Y]-[X,[Y,Z]]=0,\ \forall\, X,Y,Z\in\mathcal{L}$$

From now on, we will denote $L(X,Y,Z)=[[X,Y],Z]-[[X,Z],Y]-[X,[Y,Z]]$.

If, in addition, it is verified that $[X,X]=0$*, for all* $X\in\mathcal{L}$*, the Leibniz algebra is also a* Lie algebra*. In this case, it is satisfied that* $[X,Y]=-[Y,X]$ *and the Leibniz identity is equivalent to the Jacobi identity.*

Definition 2 *Given a basis* $\{e_i\}_{i=1}^n$ *of an* n*-dimensional Leibniz algebra* $\mathcal{L}$*, the* structure constants *of* $\mathcal{L}$ *are defined by* $[e_i,e_j]=\sum_{h=1}^n c_{i,j}^h e_h$*, for* $1\leq i,j\leq n$*.*

Definition 3 *The* derived *and* central series *of a finite-dimensional Leibniz algebra* $\mathcal{L}$ *are*

$\mathcal{L}_1=\mathcal{L},\ \mathcal{L}_2=[\mathcal{L},\mathcal{L}],\ \ldots,\ \mathcal{L}_k=[\mathcal{L}_{k-1},\mathcal{L}_{k-1}],\ \ldots$ and $\mathcal{L}^1=\mathcal{L},\ \mathcal{L}^2=[\mathcal{L},\mathcal{L}],\ \ldots,\ \mathcal{L}^k=[\mathcal{L}^{k-1},\mathcal{L}],\ \ldots$

So, $\mathcal{L}$ *is called* $(m-1)$*-step* solvable *(resp.* nilpotent*) if there exists* $m\in\mathbb{N}$ *such that* $\mathcal{L}_m=\{0\}$ *and* $\mathcal{L}_{m-1}\neq\{0\}$ *(resp.* $\mathcal{L}^m=\{0\}$ *and* $\mathcal{L}^{m-1}\neq\{0\}$*).*

3 Associating combinatorial structures with Leibniz algebras

Let $\mathcal{L}$ be a n-dimensional Leibniz algebra with basis $\mathcal{B}=\{e_i\}_{i=1}^n$. Its structure constants correspond to $[e_i,e_j]=\sum_{h=1}^n c_{i,j}^h e_h$ and, hence, the pair $(\mathcal{L},\mathcal{B})$ is associated with a combinatorial structure by the following procedure

 ISBN: 978-84-617-8694-7

a) For each $e_i \in \mathcal{B}$, we draw a vertex i.

b) For every vertex i verifying $[e_i, e_i] \neq 0$, we draw a loop such that its weight is an n-tuple given by $(c^1_{i,i}, c^2_{i,i}, \ldots, c^n_{i,i})$.

c) Given two vertices i, j verifying $(c^j_{i,j}, c^j_{j,i}) \neq (0,0)$, we draw a directed edge from vertex i to j whose weight is given by the pair $(c^j_{i,j}, c^j_{j,i})$.

d) Given three vertices $i < j < k$ such that $(c^k_{i,j}, c^k_{j,i}, c^i_{j,k}, c^i_{k,j}, c^j_{i,k}, c^j_{k,i}) \neq (0,0,0,0,0,0)$, we draw a full triangle ijk such that the edges ij, jk and ik have weights $(c^k_{i,j}, c^k_{j,i})$, $(c^i_{j,k}, c^i_{k,j})$ and $(c^j_{i,k}, c^j_{k,i})$, respectively. Moreover,

   d1) we use a discontinuous line (named *ghost edge*) for edges with weight $(0,0)$.

   d2) If two triangles ijk and ijl satisfy $(c^k_{i,j}, c^k_{j,i}) = (c^l_{i,j}, c^l_{j,i})$, draw only one edge between vertices i and j shared by both triangles.

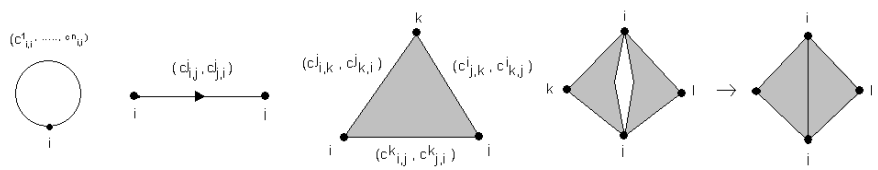


Figure 1: Loop, directed edge, full triangle and two triangles sharing an edge.

4 (Pseudo)digraphs associated with Leibniz algebras

This section is devoted to state some general results on the isomorphism classes of Leibniz algebras admitting configurations with 2 and 3 vertices, including the link of isomorphism between Leibniz algebras associated with different configurations. Throughout the paper we have used the results obtained in [2]. The main results that we have obtained are the following.

Theorem 1 *There exist only three isomorphism classes of 2-dimensional, non-abelian Leibniz algebras associated with (pseudo)digraphs, namely:* $\mathcal{L}^2_1 : [e_1, e_1] = e_2$, $\mathcal{L}^2_{2,1} : [e_2, e_1] = e_2$ *and* $\mathcal{L}^2_{2,2} : [e_1, e_2] = -[e_2, e_1] = e_2$.

 ISBN: 978-84-617-8694-7

Theorem 2 *There exist six isomorphism classes of 3-dimensional, non-abelian Leibniz algebras associated with non-connected (pseudo)digraphs. They are:* $\mathcal{L}_1^3 : [e_1, e_1] = e_2$, $\mathcal{L}_2^3 : [e_1, e_1] = e_2, [e_3, e_3] = e_2$, $\mathcal{L}_{3,1}^3 : [e_3, e_2] = e_3$, $\mathcal{L}_{3,2}^3 : [e_2, e_3] = -[e_3, e_2] = e_3$, $\mathcal{L}_{4,1}^3 : [e_2, e_2] = e_1, [e_3, e_2] = e_3$, $\mathcal{L}_{4,2}^3 : [e_2, e_2] = e_1, [e_2, e_3] = -[e_3, e_2] = e_3$.

5 Algorithmic procedure

In this section, we show an algorithmic method to determine if a given combinatorial structure is associated or not with a Leibniz algebra.

We have implemented this algorithm by using the symbolic computation package Maple, working the implementation in version 12 or higher. To do this, we have used the libraries `linalg` and `combinat` to activate commands related to Linear and Combinatorial Algebra. This algorithmic procedure consists of the following three steps

a) Defining the values of the structure constants according to the combinatorial structure.

b) Generating the law which should be satisfied by the Leibniz algebra, starting from the structure constants.

c) Checking if the Leibniz identities are satisfied for this law.

Acknowledgements

This work has been partially supported by MTM2013-40455-P and FEDER.

References

[1] J. Cáceres, M. Ceballos, J. Núñez, M.L. Puertas and A.F. Tenorio, *Combinatorial structures of three vertices and Lie algebras*, Int. J. Comput. Math. 89 (2012), 1879–1900.

[2] M. Ceballos, J. Núñez, A. F. Tenorio, *Finite-dimensional Leibniz algebras and combinatorial structures*, Communications in Contemporary Mathematics In press. doi: 10.1142/S0219199717500043

[3] J.L. Loday, *Une version non commutative des algèbres de Lie: les algèbres de Leibniz*, Enseign. Math. 39:2 (1993), 269–293.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference on Computational and Mathematical Methods in Science and Engineering, CMMSE 2017 4–8 July, 2017.

Heterogeneous CPU Plus GPU Tile-Based Approach for HEVC

Gabriel Cebrián-Márquez[1], Vicente Galiano[2], Héctor Migallón[2], José Luis Martínez[1], Pedro Cuenca[1] and Otoniel López-Granado[2]

[1] *High-Performance Networks and Architectures (RAAP), Albacete Research Institute of Informatics (I3A), University of Castilla-La Mancha, 02071 Albacete, Spain*

[2] *Department of Physics and Computer Architecture, Miguel Hernández University, 03202 Elche, Spain*

emails: Gabriel.Cebrian@uclm.es, vgaliano@umh.es, hmigallon@umh.es, JoseLuis.Martinez@uclm.es, Pedro.Cuenca@uclm.es, otoniel@umh.es

Abstract

The High Efficiency Video Coding (HEVC) standard has opened the door to high quality multimedia contents, enabling the advent of new formats such as Ultra High Definition (UHD), as a result of the unceasing demands of the market. This standard is able to outperform prior standards by up to 50% in terms of perceptual video quality, but at the cost of extremely large computational complexities. For this reason, the development of fast coding algorithms is now a requirement to make HEVC an adequate candidate for real-world scenarios. In this regard, this paper proposes a collaborative CPU+GPU coding architecture for this standard, in which the host performs a coarse-grained parallelization of the encoder using tile partitions, while the device carries out a fast motion estimation. As a result, an average speed-up of 8.41× can be obtained with 12 threads, at the cost of reasonable coding efficiency penalties.

Key words: HEVC, H.265, Heterogeneous, Parallel Encoding, GPU, Tiles

1 Introduction

Since its ratification more than a decade ago, the H.264/Advanced Video Coding (AVC) [1] standard established itself as the most widespread video compression standard for many types of applications and scenarios, including Blu-ray and various high-definition (HD)

 ISBN: 978-84-617-8694-7

television broadcasts. Nevertheless, the consumption pattern of video contents has changed in recent years, with higher preference for larger resolutions such as ultra high definition (UHD), and higher quality of contents. This, however, has led to an increase in bit rate that could become a relevant problem from the point of view of communication networks and storage services. For these reasons, the Joint Collaborative Team on Video Coding (JCT-VC) defined the High Efficiency Video Coding (HEVC) standard in early 2013 [2]. This standard has been designed to support very large resolutions and formats while improving the coding efficiency of previous standards. In fact, HEVC roughly doubles the rate-distortion (R-D) performance of H.264/AVC. In other words, it is able to reduce nearly 50% bit rate for the same perceptual video quality [3]. This notable improvement in coding efficiency comes, however, at the expense of extremely high computational complexities [4].

The HEVC standard is based on the same block-based scheme as its predecessors. The improved efficiency is derived from the enhancement of many existing coding tools, as well as the introduction of new ones. Among others, these tools include a highly flexible quadtree partitioning scheme that divides the picture into square blocks named coding tree units (CTU). While these tools enable good coding efficiency, they imply a huge increase in encoding time. To reduce the processing time, HEVC introduces some high-level tools, such as tiles and wavefronts, with the aim of allowing the parallel encoding and decoding of a video sequence. These parallelization techniques rely on creating partitions that are processed concurrently. However, some dependencies are broken to achieve this concurrency.

The parallel techniques defined in the standard operate at a coarse-grained level, while the rest of encoding operations are sequential. In fact, some of these operations can involve large computation times. In particular, the inter prediction module can take more than 60% of the encoding time in Random Access configurations [5]. This module is responsible for determining the motion vectors (MV) that minimize the rate-distortion (R-D) cost of matching each of the blocks in which the picture is divided with respect to temporal neighbors. Given the nature of this operation and the large number of repetitive operations performed, an auxiliary device such as a graphic processing unit (GPU) could help reduce the time devoted to the inter prediction. In this regard, this paper proposes a GPU-based heterogeneous coding architecture in which several picture partitions are processed in parallel in the CPU, and the integer motion estimation (ME) algorithm of the inter prediction module is performed in a single GPU, creating a collaborative framework between both devices. The main aim of the proposed architecture is to reduce the total processing time of the encoder, which is achieved at the cost of minimal coding efficiency losses.

The rest of this paper is organized as follows. Section 2 presents a short summary of HEVC. Section 3 covers some of the most relevant related works in the topic. Then, Section 4 provides an overview of the proposed GPU-based heterogeneous coding architecture. An experimental evaluation is carried out in Section 5, showing the results of this architecture in terms of time reduction and coding efficiency. Finally, conclusions are drawn in Section 6.

 ISBN: 978-84-617-8694-7

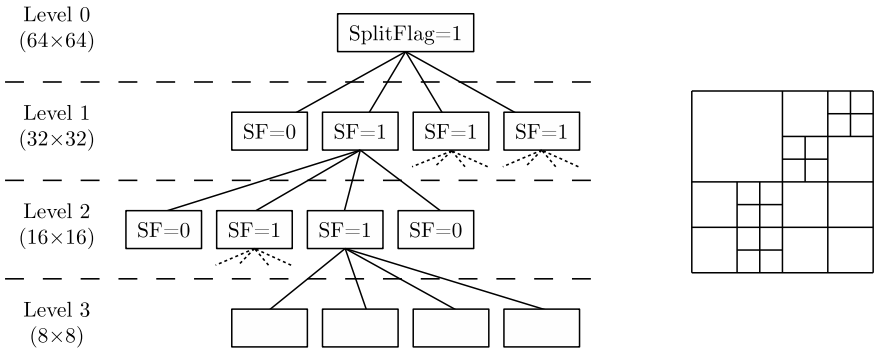


Figure 1: Example of CTU quadtree structure defined in HEVC

2 Technical Background

As mentioned in the introduction of this paper, HEVC achieves considerably higher coding efficiency compared to previous standards by enhancing existing coding tools and introducing new ones. In this regard, one of the most important novelties introduced by HEVC is the new picture partitioning scheme, which is now based on a quadtree structure as shown in Fig. 1. Each input picture is partitioned into square regions called coding tree units (CTU), whose size is typically 64×64 pixels. These structures can be partitioned, in turn, into coding units (CU), prediction units (PU) and transform units (TU). PUs store prediction information such as MVs, and can range from 64×64 to 8×8 using either symmetrical or asymmetrical sizes. TUs, in turn, contain residual information and range from 32×32 to 4×4, adopting a tree structure named residual quadtree (RQT). Other new features in HEVC include a total of 35 different intra coding modes, or the new in-loop filters.

With regard to the prediction, HEVC defines eight possible partitions for each CU size: 2N×2N, 2N×N, N×2N, N×N, 2N×nU, 2N×nD, nL×2N and nR×2N. The last four PU types correspond to the asymmetric motion partitioning (AMP) introduced as a novelty in this standard. These new sizes involve a large increase in the complexity of the encoder, but it also enables larger adaptability for edges, especially with larger CU sizes. To get more accurate motion vector predictors (MVP), the standard defines the advanced motion vector prediction (AMVP) algorithm, which enables to derive several most probable candidates based on data from adjacent PUs. As a result, less bits are necessary to encode the MVs, which enables the use of the merge mode more often, omitting the transmission of MVs.

All these new coding tools introduced in HEVC enable a very flexible encoding. Nevertheless, they also involve a notable increase in the computational complexity of both the encoder and the decoder. With the aim of reducing this complexity, the standard defines two different parallelization strategies known as tiles [6] and wavefront parallel processing (WPP) [7]. On the one hand, tiles are rectangular partitions where dependencies are par-

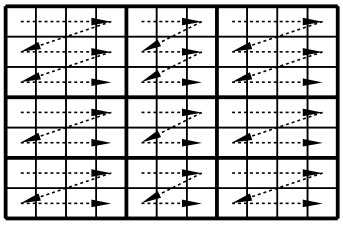


Figure 2: Example of picture partitioned in tiles, and coding order

tially broken across boundaries, making it possible to process them independently. The in-loop filters can still cross these boundaries, however, to improve the overall coding efficiency. An example of tile-based processing can be seen in Fig. 2, in which the input picture has been divided into 9 tiles. On the other hand, WPP enables the creation of rows that can be processed in parallel. In this case, unlike tiles, the entropy coding is allowed to cross partitions with the aim of minimizing coding losses, but at the cost of introducing a delay of two CTUs between consecutive rows, which limits the parallelism of the algorithm. As a design decision, both techniques cannot coexist at the same time.

3 Related Work

Provided the need for parallel techniques in HEVC, several state-of-art works address the complexity analysis and parallelization strategies of this standard [5, 8]. As for other parallelization techniques, authors in [9] present a variation of WPP called Overlapped Wavefront (OWF), which enables processing two consecutive pictures at the same time to avoid ramping inefficiencies. In this way, when a thread finishes processing a CTU row and there is no more available rows in the current frame, it can begin processing the next picture instead of idling. However, this paper applies this technique to the decoder side. With regard to the encoder, authors in [10] propose a fine-grained parallel optimization of the ME module for multi-core and many-core platforms, allowing to perform the MV prediction of all PUs available in a CU at the same time. Therefore, the maximum parallelism will be dependent of the platform. In a similar way to this work, a ME algorithm for many-core platforms is presented in [11].

Other works make use of GPU devices in the encoding process, especially to perform inter prediction. Authors in [12] propose a scheme designed for GPU plus multi-core CPU platforms, in a similar way to this paper. However, the encoder used in the tests does not include all the coding tools, and thus results are not comparable. In [13], authors propose a GPU-based ME algorithm that makes use of a diamond search pattern, which might not obtain the best candidate in some cases. This work was later extended for bi-prediction in [14], but still used the same search pattern. Additionally, the achieved speed-up is

limited due to the portion of time devoted to the ME, which could be solved with the use of a coarse-grained algorithm.

4 CPU Plus GPU Tile-Based Coding Architecture

As detailed in the previous sections, it is possible to parallelize both the encoder and the decoder using the techniques defined in the standard: tiles and WPP. While this is an effective way to reduce the total encoding time, some modules still involve large computation times, which should be reduced to achieve reasonable processing times. The module in which it is devoted the largest amount of time is the inter prediction module, which basically consists in subtracting the current PU to different positions in the reference frame, in order to reduce the R-D cost generated by a given MV. The nature of this operation makes the GPU a very adequate auxiliary device to perform the whole processing in little time. For this reason, in this paper we propose a GPU-based heterogeneous coding architecture in which the CPU performs a coarse-grained parallelization of the encoding, while the GPU carries out a fine-grained parallelization of the ME algorithm of the inter prediction module. In this way, it is possible to archive much larger speed-ups than using a homogeneous parallel architecture, i.e. with each of the algorithms separately. In this regard, the tile-based algorithm defined in the standard has been selected to perform the coarse-grained parallelization provided that it results in larger speed-ups. The next subsections will show the approach followed by the proposed architecture.

4.1 Coarse-Grained Parallelization Algorithm: Tiles

As mentioned in Section 2, a tile is a partition that can be independently encoded or decoded with regard to the other tiles in the same frame. As an exception, the filtering operation needs to be performed after all the tiles in a frame have been encoded or decoded, and thus cannot be parallelized. Consequently, a synchronization point is required at the end of the frame. Despite this, tiles offer huge parallelism that can be exploited so that every tile is encoded by a different processing unit, which in our case corresponds to a thread.

Our proposed approach is to divide each frame into as many tiles as available processing units. In this way, if a hardware platform is composed of an octa-core CPU, each frame would be divided into eight different tiles. While tiles can be assigned any rectangular shape defined by parameter, the approach followed in this paper is to utilize only horizontal tiles. The reason for this is to improve the data locality in the device, as frames are usually stored in raster scan order. Additionally, in order to maximize the parallel efficiency of the algorithm, each partition is assigned the same number of CTUs, if possible.

Apart from the filtering operation mentioned before, the synchronization point located at the end of the frame also enables to update the decoded picture buffer (DPB), which stores the reconstructed version of the frames after their encoding. In order to avoid costly

memory transfers, a shared memory scheme has been utilized, so that the DPB is shared across processing units.

4.2 GPU-Based Inter Prediction Algorithm

Whereas the tile-based algorithm is able to achieve coarse-grained parallelism of the encoding process, it is still possible to further parallelize the encoding with the proposed heterogeneous architecture. As already mentioned, one of the most computationally expensive operations of the encoder is the inter prediction, and the ME algorithm in particular. This operation consists in estimating the MVs that perform the transformation of the PUs in the current frame with regard to the reference frames with the lowest R-D cost. It is performed by subtracting the corresponding partitions from the reference pictures, which involves a large number of repetitive operations. Considering the nature of this algorithm, it should be noted the convenience of the GPU architecture.

CPU and GPU can work together in a synchronous or an asynchronous way. However, to avoid delays caused by data transfers and kernel executions, the GPU needs to perform its operations asynchronously. The way in which this asynchronism is achieved by the proposed algorithm is shown in Fig. 3. As can be seen, the original frames are copied to the device as soon as a group of pictures (GOP) starts being processed. In this way, they are always available for prediction in the device. Given that the ME requires the reconstructed version of the frames to perform the prediction, these are copied and replaced in the device once they are encoded. With regard to the assignation of workload to the device, the host queues the ME of the following CTU to the one that is currently being processed. In other words, the device is always processing a CTU ahead of the host. This is possible thanks to a double buffer that holds the motion information of the current CTU, while it allocates the space necessary for next one. For this reason, the host might need to wait for the device to finish processing the first CTU as an exceptional case.

In a similar way to how the encoder would perform the ME in the host, the device calculates the MVs for all the possible PUs in which a CTU is divided, and their associated distortion costs. However, the device is capable of evaluating all the possible MVs in the search area, while the host requires to use fast algorithms in order to perform the estimation in an assumable period of time. The ME is performed in two steps, which corresponds to two different GPU kernels. On the one hand, the first kernel executes the operations required to calculate the distortion costs, i.e. the sum of absolute difference (SAD) residuals across a search area in the reference frame. On the other hand, the second one consists in a reduction algorithm that determines the best MV in terms of SAD cost.

The first kernel relies on the fact that the dimensions of every PU defined by the standard are divisible by four, and thus it is possible to calculate the SAD residual information of a PU from the composition of its 4×4 partitions. Under this assumption, the previously mentioned kernel distributes a device thread to each sample in the reference search area,

 ISBN: 978-84-617-8694-7

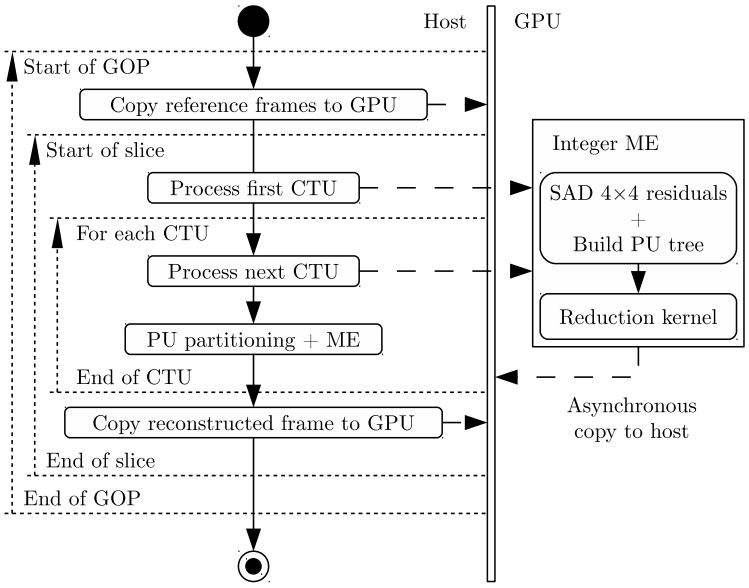


Figure 3: GPU-based inter prediction algorithm diagram

which is equivalent to a MV. Each thread is responsible for calculating all the 4×4 SAD blocks in a CTU. After all the threads have finished calculating these costs, each of them puts them together to calculate the distortion of the PUs into which a CTU might be divided. At this point, the device has at its disposal the distortion cost of each PU in each position in the search area. Therefore, the second kernel consists in a reduction algorithm that selects the best MV in terms of SAD. The result of this step is the optimal MV from the point of view of the distortion for each PU and reference frame. This information is sent to the host.

With the information calculated in the GPU, the encoder is able to skip the integer motion estimation process, which is the first part of the inter prediction algorithm. Instead of performing this estimation, the host only needs to obtain the MV that the device asynchronously copied beforehand. As a result, the computational complexity of the inter prediction module is reduced.

4.3 Joint GPU-Based Heterogeneous Architecture

While the tile-based algorithm defined in the standard performs a coarse-grained parallelization of the encoding, the GPU-based ME algorithm is able to speed up one of the most computationally expensive modules of the encoder. As they both work at different levels, they can cooperate simultaneously. The way in which this collaboration takes place

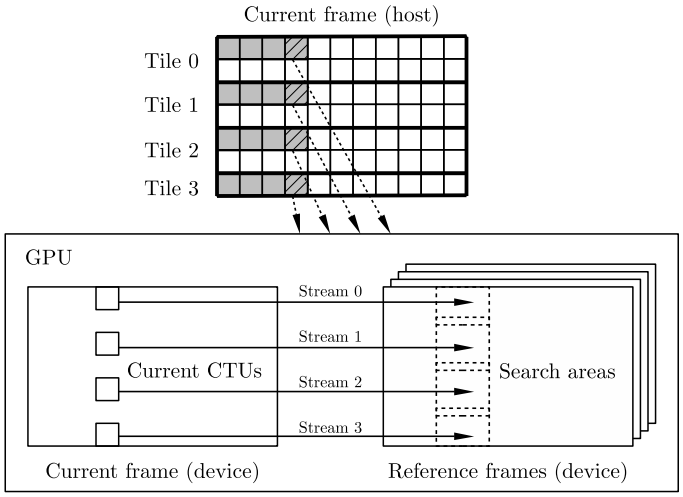


Figure 4: Combination of the tile-based algorithm and the GPU-based proposal (4 threads)

is shown in Fig. 4. As can be seen, the input pictures are partitioned into horizontal tiles as described in Section 4.1. Each of these tiles is assigned to a processing unit, which is equivalent to a thread in our model. Every processing unit queues, in turn, the GPU-based ME algorithm into different streams of the same device, making it possible to execute the kernels concurrently with respect to the host. As a result, the device utilization increases, and the total power consumption decreases.

The scalability of the proposed architecture depends on two different factors. The first one is inherent to the design of the tiles algorithm, and limits the maximum parallelism to the number of rows in a frame. However, this issue becomes less important the larger the frame is. For example, a UHD sequence could hold 34 different horizontal tiles. The second factor that might limit the parallelism is the hardware platform itself, as the device could limit the amount of parallelism if is not able to process all the required CTUs simultaneously. Nevertheless, this has not happened in our experimental evaluation of the architecture.

5 Performance Evaluation

The experiments have been executed on the HEVC Test Model (HM-16.3) reference software [15], following the guidelines and coding conditions provided by the JCT-VC in [16]. In this regard, Random Access has been the selected configuration, and the QP values are 22, 27, 32 and 37. The encodings have been carried out using the Main profile and 8-bit depth. Sequences of classes A to D according to the JCT-VC classification have been used,

Table 1: Per-sequence results of the proposed GPU-based heterogeneous architecture varying number of threads

Class	Video sequence	BD-rate (%)					Speed-up				
		1	2	4	8	12	1	2	4	8	12
A	Traffic	0.45	0.64	1.03	1.84	2.67	1.05	2.01	3.82	6.88	9.17
	PeopleOnStreet	−0.13	0.06	0.42	1.07	1.72	1.12	2.14	4.09	7.24	9.57
B	Kimono	0.50	1.07	2.19	4.15	6.03	1.09	2.09	3.95	7.03	8.26
	ParkScene	0.22	0.47	1.07	2.16	3.21	1.06	2.03	3.69	6.48	7.99
	Cactus	0.32	0.71	1.64	3.18	4.54	1.08	2.04	3.68	6.62	7.98
	BasketballDrive	2.99	3.68	4.78	7.07	9.12	1.13	1.99	3.73	6.76	7.89
	BQTerrace	−0.24	0.14	0.96	2.33	3.59	1.06	2.01	3.77	6.69	7.99
C	BasketballDrill	−0.15	0.79	2.66	6.06		1.10	2.06	3.79	6.99	
	BQMall	0.79	1.88	3.76	7.14		1.08	2.05	3.53	6.66	
	PartyScene	−0.20	0.23	0.98	2.34		1.07	1.97	3.68	6.95	
	RaceHorses	0.52	1.38	3.18	6.25		1.14	2.03	3.87	7.28	
D	BasketballPass	0.38	2.47	5.25			1.11	2.05	3.38		
	BQSquare	0.06	1.23	2.93			1.05	1.85	3.35		
	BlowingBubbles	−0.26	0.65	2.33			1.05	1.89	3.34		
	RaceHorses	−0.30	1.30	4.05			1.11	1.89	3.49		
Mean values (A & B)		0.59	0.97	1.73	3.12	4.41	1.09	2.04	3.82	6.82	8.41
Mean values (all)		0.33	1.11	2.48	3.96	4.41	1.09	2.01	3.68	6.87	8.41

which include the following resolutions: 2560×1600 (A), 1920×1080 (B), 832×480 (C), and 416×240 (D).

The hardware platform used in the experiments is composed of an Intel® Xeon® E5-2650 v2 CPU running at 2.60 GHz and an NVIDIA® Tesla® K40 GPU. The encoder has been compiled with GCC 4.8.5 and NVIDIA® CUDA 7.0, and executed on CentOS 7. Turbo Boost was disabled to achieve the reproducibility of the results.

With regard to the parametrization of the encoder, the sample adaptive offset (SAO) filter has been disabled, as well as the *LFCrossSliceBoundary* flag. The reason for this decision is to achieve full independence between tiles, as the SAO filter might introduce some dependencies in the encoding. The executions have been performed using 1, 2, 4, 8 and 12 threads.

The results will be provided in terms of speed-up and Bjøntegaard delta rate (BD-rate). The latter represents the percentage of bit rate variation between two sequences with the same objective quality [17]. Therefore, a negative value implies that the proposal improves the coding efficiency of the baseline encoder.

 ISBN: 978-84-617-8694-7

Table 1 shows the results obtained by the proposed architecture when different number of threads are used for the execution. The results in which only one thread is used correspond to the case in which only the GPU is enabled, but no tile-based parallelization is performed. Also, it can be seen that the tests were not carried out for class D when 8 and 12 threads were used, and also for class C with 12 threads. This is caused by the fact that the maximum parallelism of the tile-based coarse-grained algorithm is limited by the total number of rows in the frame. As these classes are composed of low resolution sequences, they do not admit as many threads as used in the tests, so the obtained results would be the same as using a lower number of processing units: 4 threads for class D and 8 threads for class C. Considering that the average values shown in the table might not comprise the same set of results, the mean values for classes A and B are also shown to better reflect the effect of the proposed architecture.

First, it is worth detailing the results obtained by the GPU-based ME algorithm itself. It can be seen that the achieved speed-up is limited to 1.09×, which is derived from the fact that this is the portion of time devoted by the encoder to the integer ME in the inter prediction module. Nevertheless, this time reduction is achieved at the expense of very low coding efficiency losses of 0.33% on average. In fact, it can be seen that in some sequences the coding efficiency improved with respect to the baseline encoder, which is caused by the fact that the GPU-based algorithm performs a full search that covers all the possible MVs in the search area, while the baseline algorithm carries out a simplified search. Additionally, it has to be taken into account that this speed-up will perform as a multiplier factor for the tile-based parallelization algorithm.

With regard to the joint algorithm, it can be seen that the coding efficiency experiences larger losses when the number of processing units increases. This is the expected behavior of the algorithm, as more dependencies are broken. In fact, it is inherent to the design of the tile-based scheme defined in the standard. The table also displays the speed-up obtained by the proposed joint algorithm, which can reach a factor of up to 8.41× with 12 threads. This means that the proposed architecture does not fully utilize the hardware platform. At first, it might seem that this is caused by a bottleneck in the device, but in fact, the real cause is that the pictures cannot be evenly partitioned, which causes that some tiles might be larger than others. Also, it might happen that two tiles with the same number of CTUs take different times to process depending on the complexity of the picture. Nevertheless, the achieved speed-up is notable, and denotes that the proposed architecture is scalable as long as the frames are divided adequately. Regarding the contribution of the GPU in the joint algorithm, the benefits of the device can be observed in the results obtained for 2 threads, in which it can be seen that the achieved speed-up can be larger than 2×, which would not be feasible with only the host. The same behavior is displayed for the sequence PeopleOnStreet for 4 threads. As for the rest, the multiplying factor of the GPU is occluded by the inefficiencies derived from the tile-based algorithm.

 ISBN: 978-84-617-8694-7

6 Conclusions and Future Work

The HEVC standard has enabled the advent of a great number of applications and video formats such as UHD, as a result of the demands of the market for higher quality of video contents. Nevertheless, the improved coding efficiency of HEVC has also led to a relevant increase in computational complexity of both the encoder and the decoder. The research community is making large efforts to develop algorithms and techniques that enable to reduce this complexity with the aim of making this standard feasible for real-world scenarios. Accordingly, this paper proposes a CPU+GPU collaborative framework for HEVC in which the CPU performs a coarse-grained parallelization of the encoding, while the GPU aids in carrying out the ME. Having into consideration the intrinsic limitations of tiles, the experimental evaluation of the algorithm showed that the algorithm is scalable, reaching a speed-up of 8.41% with 12 threads for reasonable coding efficiency losses.

As future work, the most immediate step would be to extend the GPU algorithm to perform the fractional ME as well, so larger time reductions would be obtained. Additionally, it would be also of interest to analyze other coarse-grained algorithms that solve the limitations of horizontal tiles.

Acknowledgments

This work was jointly supported by the Spanish Ministry of Economy and Competitiveness and the European Commission (FEDER funds) under the projects TIN2015-66972-C5-2-R and TIN2015-66972-C5-4-R, and by the Spanish Ministry of Education, Culture and Sports under the grant FPU13/04601.

References

[1] ISO/IEC and ITU-T, *Advanced Video Coding for Generic Audiovisual Services. ITU-T Recommendation H.264 and ISO/IEC 14496-10 (version 12)*, April 2017.

[2] ISO/IEC and ITU-T, *High Efficiency Video Coding (HEVC). ITU-T Recommendation H.265 and ISO/IEC 23008-2 (version 4)*, December 2016.

[3] G. J. Sullivan, J.-R. Ohm, W.-J. Han, and T. Wiegand, *Overview of the High Efficiency Video Coding (HEVC) Standard*, IEEE Trans. Circuits Syst. Video Technol., 22 (2012), pp. 1649–1668.

[4] J.-R. Ohm, G. J. Sullivan, H. Schwarz, T. K. Tan, and T. Wiegand, *Comparison of the Coding Efficiency of Video Coding Standards - Including High Efficiency Video Coding (HEVC)*, IEEE Trans. Circuits Syst. Video Technol., 22 (2012), pp. 1669–1684.

 ISBN: 978-84-617-8694-7

[5] F. Bossen, B. Bross, K. Suhring, and D. Flynn, *HEVC Complexity and Implementation Analysis*, IEEE Trans. Circuits Syst. Video Technol., 22 (2012), pp. 1685–1696.

[6] K. Misra, A. Segall, M. Horowitz, S. Xu, A. Fuldseth, and M. Zhou, *An Overview of Tiles in HEVC*, IEEE J. Sel. Topics Signal Process., 7 (2013), pp. 969–977.

[7] F. Henry and S. Pateux, *Wavefront Parallel Processing*, Tech. Rep. JCTVC-E196, March 2011.

[8] M. Álvarez-Mesa, C. C. Chi, B. Juurlink, V. George, and T. Schierl, *Parallel Video Decoding in the Emerging HEVC Standard*, in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), March 2012, pp. 1545–1548.

[9] C. C. Chi, M. Álvarez-Mesa, J. Lucas, B. Juurlink, and T. Schierl, *Parallel HEVC Decoding on Multi- and Many-core Architectures*, J. Sign. Process. Syst., 71 (2013), pp. 247–260.

[10] Q. Yu, L. Zhao, and S. Ma, *Parallel AMVP Candidate List Construction for HEVC*, in IEEE Visual Communications and Image Processing (VCIP), November 2012, pp. 1–6.

[11] C. Yan, Y. Zhang, J. Xu, F. Dai, J. Zhang, Q. Dai, and F. Wu, *Efficient Parallel Framework for HEVC Motion Estimation on Many-Core Processors*, IEEE Trans. Circuits Syst. Video Technol., 24 (2014), pp. 2077–2089.

[12] X. Wang, L. Song, M. Chen, and J. Yang, *Paralleling Variable Block Size Motion Estimation of HEVC on CPU Plus GPU Platform*, in IEEE International Conference on Multimedia and Expo Workshops (ICMEW), July 2013, pp. 1–5.

[13] S. Radicke, J.-U. Hahn, C. Grecos, and Q. Wang, *A Highly-Parallel Approach on Motion Estimation for High Efficiency Video Coding (HEVC)*, in IEEE International Conference on Consumer Electronics (ICCE), January 2014, pp. 187–188.

[14] S. Radicke, J.-U. Hahn, Q. Wang, and C. Grecos, *Bi-predictive motion estimation for HEVC on a graphics processing unit (GPU)*, IEEE Trans. Consum. Electron., 60 (2014), pp. 728–736.

[15] *HEVC Test Model (HM) Reference Software*. `https://hevc.hhi.fraunhofer.de/`.

[16] F. Bossen, *Common Test Conditions and Software Reference Configurations*, Tech. Rep. JCTVC-L1100, January 2013.

[17] G. Bjøntegaard, *Calculation of average PSNR differences between RD-curves*, Tech. Rep. VCEG-M33, ITU-T Video Coding Experts Group (VCEG), 2001.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

Preconditioners for rank-deficient least squares problems

J. Cerdán[1], D. Guerrero[2], J. Marín[1] and J. Mas[1]

[1] *Institut de Matemàtica Multidisciplinar, Universitat Politècnica de València, València, España*

[2] *Departamento de Ciencias Matemáticas, Universidad Pedagógica Nacional Francisco Morazán, Tegucigalpa, Honduras*

emails: jcerdan@mat.upv.es, danguefl@doctor.upv.es, jmarinma@mat.upv.es, jmasm@mat.upv.es

Abstract

In this paper we present a method for computing sparse preconditioners for solving rank deficient least squares (LS) arising from practical problems. The main idea of the method proposed is updating an incomplete factorization computed for a regularized problem to recover the solution of the original one. The numerical experiments show that the preconditioner proposed can be successfully applied to accelerate the convergence of iterative Krylov subspace methods.

Key words: Iterative methods, preconditioners, least squares.

1 Introduction

Linear least squares (LS) problems arise in many large-scale applications of the science and engineering as neural networks, linear programming, exploration seismology or image processing to name a few. The LS problem considered is formulated as

$$\min_x \|b - Ax\|_2, \tag{1}$$

where $A \in \mathbb{R}^{m\times n}$ $(m \geq n)$ is large and sparse and $b \in \mathbb{R}^m$. This problem can be solved iteratively using the conjugate gradient like methods [1]. Basically, this methods implicitly apply the conjugate gradient or minimal residual method to the normal equations

$$A^T Ax = A^T b. \tag{2}$$

 ISBN: 978-84-617-8694-7

When the system matrix is large and sparse iterative methods may be preferred because they often require much less storage than their direct counterparts. Their successful application often needs a good preconditioner in order to achieve fast convergence rates. When the matrix A has full rank, the preferred method of choice is computing an incomplete Cholesky factorization (IC) of the symmetric and positive definite matrix $C = A^T A$, see [2]. If the matrix A is rank deficient then, the matrix C is a semidefinite positive matrix and the Cholesky factorization suffers breakdown because negative or zero pivots are encountered. Thus, rank deficient LS problems are in general much more harder to solve. Basically, the are two types of approaches for solving this case iteratively. The first one consists of computing an incomplete factorization of a regularized matrix which can be used as a preconditioner for the original LS problem [6]. The second type is solving a mathematically equivalen augmented linear system of order $m+n$ [1]. The technique proposed in this work belongs to the first type.

The idea is to compute an IC factorization of the normal equations associated to the regularized matrix

$$\begin{bmatrix} A \\ \alpha^{1/2} I \end{bmatrix}. \tag{3}$$

Thouse are given by

$$C_\alpha = A^T A + \alpha I. \tag{4}$$

The shift α is known as Tikhonov regularization parameter. If α is choosen large enough the computation of an IC for the matrix C_α can be done easily. On the other hand, since the final purpose is to use this incomplete factorization as a preconditioner for the original (unregularized) linear system, the parameter α should be chosen as small as possible. Both requirements make difficult the choice of the appropriate α. In practice, the factorization is restarted more than once, increasing α on each restart until breakdown is avoided.

We propose a method that simplifies the choice of the regularization parameter. It is based on the work presented in [3] in which the authors study how to update a preconditioner for LS problems when the linear system is modified by adding or removing equations. Consider the matrix

$$(C_\alpha - \beta I) \tag{5}$$

which is an update of the shifted matrix C_α in (3). Clearly, the closer β is to α, the closer this update is to the normal equations of the original system. Our technique consists of updating an incomplete Cholesky factorization obtained for C_α using the augmented matrix

$$\begin{bmatrix} C_\alpha & \beta^{1/2} I \\ \beta^{1/2} I & I \end{bmatrix}. \tag{6}$$

Observe that between equations (5) and (6) one can establish the relations

$$C_\alpha - \beta I = \begin{bmatrix} I & O \end{bmatrix} \begin{bmatrix} C_\alpha & \beta^{1/2} I \\ \beta^{1/2} I & I \end{bmatrix} \begin{bmatrix} I \\ \beta^{1/2} I \end{bmatrix}, \tag{7}$$

 ISBN: 978-84-617-8694-7

and

$$(C_\alpha - \beta I)^{-1} = \begin{bmatrix} I & O \end{bmatrix} \begin{bmatrix} C_\alpha & \beta^{1/2} I \\ \beta^{1/2} I & I \end{bmatrix}^{-1} \begin{bmatrix} I \\ O \end{bmatrix}. \quad (8)$$

Thus, an IC factorization computed for (6) can be used as a preconditioner for the original normal equations using these relations. Also note that the modification of the shifted matrix in equation (5) allows for the choice of shift α values large enough to avoid breakdown during the IC of the matrix C_α. The subsequent update of the computed IC factorization should keep the updated preconditioner close to the original system.

2 Preconditioner computation

A block IC factorization of the augmented matrix (6) can be computed as follows. Since

$$\begin{pmatrix} C_\alpha & \beta^{1/2} I \\ \beta^{1/2} I & I \end{pmatrix} = \begin{pmatrix} L_\alpha & 0 \\ \beta^{1/2} L_\alpha^{-T} & I \end{pmatrix} \begin{pmatrix} I & 0 \\ 0 & R \end{pmatrix} \begin{pmatrix} L_\alpha^T & \beta^{1/2} L_\alpha^{-1} \\ 0 & I \end{pmatrix}$$

where $R = I - \beta L_\alpha^{-T} L_\alpha^{-1})$, the preconditioner is computed in four steps:

1. Compute $C_\alpha \approx L_\alpha L_\alpha^T$.
2. Compute $T = \beta^{1/2} L_\alpha^{-1}$.
3. Compute $R = I - T^T T$.
4. Compute $L_R L_R^T \approx R$.

3 Preconditioner application

The preconditioning step for a Krylov subspace iterative method typically consists of obtaining the preconditioned vector $s = M^{-1} r$ where M^{-1} is the preconditioner and r is the residual. Thus, the preconditioning strategy proposed computes the preconditioned residual by applying equation (8) with an incomplete factorization of the augmented matrix. That is,

$$\begin{pmatrix} L_\alpha & 0 \\ \beta^{1/2} L_\alpha^{-T} & I \end{pmatrix} \begin{pmatrix} I & 0 \\ 0 & R \end{pmatrix} \begin{pmatrix} L_\alpha^T & \beta^{1/2} L_\alpha^{-1} \\ 0 & I \end{pmatrix} \begin{pmatrix} s \\ s_1 \end{pmatrix} = \begin{pmatrix} r \\ 0 \end{pmatrix}.$$

The preconditioning step is done as follows:

1. Solve $L_\alpha r_1 = r$.
2. Solve $L_R (L_R^T r_2) = -T^T r_1$.
3. Solve $L_\alpha^T s = r_1 - T r_2$.

These three steps will be referenced as updated preconditioned method.

 ISBN: 978-84-617-8694-7

4 Numerical experiments

In this section we study the numerical performance of the preconditioner update method proposed. We present results obtained with matrices arising in different areas of scientific computing from the Florida Sparse Matrix Collection [4].

The technique proposed was used as preconditioner for the conjugate gradient method applied to the normal equations. It was compared with the IC factorization of C_α used as preconditioner for the LSMR method [5].

Table 1 shows the set of matrices tested. The matrices were cleaned by removing the null rows and columns before solving the LS problem. The number of rows and columns after removing the null ones, number of nonzeros, nnz, and nullity of the matrix (estimated null space rank) are reported.

Matrix	m	n	nnz	nullity	Application
BAXTER	27441	30733	111576	3055	Linear programming
DBIR1	18804	45775	1077025	2	Linear programming
DBIR2	18906	45877	1158159	2	Linear programming
NSCT1	22901	37461	678739	1	Linear programming
beaflw	492	500	53403	32	Economic
Pd_rhs	5804	4371	6323	3	Counter-example
162bit	3606	3476	37118	16	Combinatorial
176bit	7441	7150	82270	40	Combinatorial
192bit	13691	13093	154303	87	Combinatorial
208bit	24430	23191	299756	210	Combinatorial
kneser_1041	349651	330751	922252	7350	Combinatorial
wheel_601	902103	723605	2170814	600	Combinatorial
12month1	12471	872622	22624727	53	Bipartite graph
ND_actors	383640	127823	1470404	13061	Bipartite graph
Maragal_6	21251	10144	537694	92	Least squares
Maragal_7	46845	26525	1200537	659	Least squares
Maragal_8	33093	60845	1308415	14637	Least squares
mri1	65536	114637	589824	1019	graphics/vision
mri2	63240	104597	569160	14919	graphics/vision
tomographic1	142752	1014301	11537419	3700	graphics/vision

Table 1: Set of tested matrices

 ISBN: 978-84-617-8694-7

The matrices with $m < n$ were transposed. Moreover, the Dulmage-Mendelson permutation was applied to A. The columns of the matrix corresponding to the normal equations $C = A^T A$ were normalised by their 2-norm. As result, an IC factorization $L_\alpha L_\alpha^T$ was computed for the shifted matrix,

$$C_\alpha = SPCP^T S^T + \alpha I.$$

For the updated preconditioner, a value of $\alpha = \beta = 1$ was used for all the matrices, except for the matrix BEAFLW for which a value of $\alpha = 10^{-3}$ was needed. The right-hand side b provided with the problem was used. Otherwise, the vector of all ones was used.

Table 2 show the results. In this table, $time_t$, $||r||_2$ and itn represent the total time (in seconds), residual norm ($||b-Ax||_2$) of the original LS problem and the number of iterations needed to converge the iterative method. As mentioned above, LSMR correspond to the results obtained with the IC factorization of C_α while PCG correspond to ones obtained with the proposed updated preconditioning technique. The iterative methods were stopped when the relative initial residual, i.e., $||b - Ax||_2/||b||$, was reduced to 10^{-6}. L_α was calculated with drop tolerance equal to 0.1, except for the matrices MRI1 and MRI2 for which a value of 0.2 was used. The experiments were performed using MATLAB.

The results show that the updated preconditioner method is competitive and robust for solving rank deficient least square problems. There are some problem for which the total solution time is reduced considerably. Thus, we think that the preconditioner proposed can be successfully applied to accelerate the convergence of iterative Krylov subspace methods.

Acknowledgements

This research is supported by the Spanish Ministerio de Economa y Competitividad under grants MTM2014-58159-P and MTM2015-68805-REDT.

References

[1] Björck, Å.: Numerical methods for Least Squares Problems. SIAM, Philadelphia (1996)

[2] Bru, R., Marín, J., Mas, J., Tůma, M.: Preconditioned iterative methods for solving linear least squares problems. SIAM J. Sci. Comput. **36**(4), A2002–A2022 (2014)

[3] Marín, J. , Mas, J., Guerrero D., Hayami K., : Updating preconditioners for modified leas squares problems. Numer. Algorithms, published online, (2017).

[4] T. A. Davis, Y. Hu, The university of florida sparse matrix collection, ACM Trans. Math. Softw. 38 (1) (2011) 1:1–1:25.

 ISBN: 978-84-617-8694-7

[5] David Chin-Lung Fong and Michael Saunders. Lsmr: An iterative algorithm for sparse least-squares problems. *SIAM J. Sci. Comput.*, 33(5):2950–2971, October 2011.

[6] J. Scott. On using cholesky-based factorizations for solving rank-deficient sparse linear least-squares problems. *Prepint RAL-P-2016-005, SIAM J. Sci. Comput.*, 2016.

 ISBN: 978-84-617-8694-7

Matrix $time_t/\|\|r\|\|_2/itn$	PCG $time_t/\|\|r\|\|_2/itn$	LSMR
BAXTER	4.2/2.90/1701	2.6/2.90/1574
DBIR1	1.0/3.65/198	59.5/15.00/12918
DBIR2	0.8/11.35/138	40.8/11.35/8404
NSCT1	0.4/13.03/82	6.7/13.03/1957
beaflw	4.5/4.35/8332	18.8/4.92/15341
Pd_rhs	0.4/34.62/188	0.3/34.62/911
162bit	0.1/0.62/255	0.1/0.62/279
176bit	0.2/0.80/254	0.3/0.80/277
192bit	0.4/1.28/246	0.4/1.28/270
208bit	0.7/1.62/215	0.7/1.62/240
kneser_1041	4.6/504.50/105	5.6/504.50/114
wheel_601	13.0/497.51/68	13.7/497.51/140
12month1	37.7/679.32/190	41.4/679.32/245
ND_actors	133.4/301.65/6491	140.5/301.65/7188
Maragal_6	$1.7/4.8e^{-4}/568$	$1.8/4.8e^{-4}/609$
Maragal_7	$4.4/1.3e^{-4}/465$	$3.8/1.3e^{-4}/493$
Maragal_8	$6.0/1.4e^{-3}/778$	$7.0/1.4e^{-3}/1002$
mri1	1.5/209.2854/128	1.7/208.1454/175
mri2	$0.8/3.6e^{-3}/55$	$1.1/3.6e^{-3}/99$
tomographic1	$3.9/3.4e^{-3}/616$	$3.5/3.4e^{-3}/653$

Table 2: Results for the PCG and LSMR methods.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

An Efficient Numerical Method for Two Parameter Singularly Perturbed Problem with Discontinuous Convection Coefficient and Source Term

M. Chandru[1], T. Prabha[1], P. Das[2], V. Shanthi[1] and H. Ramos[3]

[1] *Department of Mathematics, National Institute of Technology, Tiruchirappalli-620 015, Tamilnadu, INDIA.*

[2] *Department of Mathematics, Indian Institute of Technology, Patna- 801 103, Bihar, INDIA.*

[3] *Department of Applied Mathematics, Scientific Computing Group, University of Salamanca, Plaza de la Merced 37008, SPAIN.*

emails: leochandru@gmail.com, prabha.thevaraj@gmail.com, pratibhamoy@iitp.ac.in, vshanthi@nitt.edu, higra@usal.es

Abstract

In this paper, we discuss a higher order convergent numerical method for two-parameter singularly perturbed problem having boundary and interior layers due to a discontinuous convection coefficient and discontinuous source term. A hybrid monotone scheme which provides almost second order uniform convergent solution, is constructed on a piecewise uniform mesh. The current scheme is compared with the standard upwind scheme, used at the point of discontinuity. Numerical experiments show that the hybrid scheme provides a higher order (almost second order) accuracy compared to the standard upwind scheme which is almost first order accurate.

Key words: singular perturbation, boundary and interior layers, hybrid scheme
MSC 2000: 65N12, 65N30 and 65N06

1 Introduction

Singularly perturbed problems (SPPs) can be characterized by the presence of a (or few) small positive parameter(s) which follows the boundary or interior layers. Adaptive mesh

generation [3, 4] is necessary to identify these layers. The simplest adaptive mesh on this context is due to Shishkin [1], who proposed a piecewise uniform mesh to capture the layer behavior. In today's literature, numerical methods for SPPs (involving only diffusion parameter) with smooth data [1], and non-smooth data [2, 5, 6, 7] are available on Shishkin meshes. Two-parameter problems, involving convection (ε_c) and diffusion (ε_d) parameters, extend the convection and reaction dominated model. In recent years, several numerical methods are observed for two-parameter problems with smooth data [8, 9, 10, 13] and non-smooth [11]. The papers [2, 6, 7] are dedicated to the numerical analysis of interior layers due to the discontinuous convection coefficient. Motivated by the works, we consider the following two-parameter singularly perturbed problem with discontinuous convection coefficient and source term:

$$Lu(x) \equiv \varepsilon_d u''(x) + \varepsilon_c a(x)u'(x) - b(x)u(x) = f(x),\ \forall\ x \in (\Gamma^- \cup \Gamma^+), \tag{1}$$

$$u(0) = u_0, \quad u(1) = u_1, \tag{2}$$

$$\text{where } a(x) \leq -\alpha_1 < 0,\ \text{for } x \in \Gamma^- \text{ and } a(x) \geq \alpha_2 > 0,\ \text{for } x \in \Gamma^+, \tag{3}$$

$$|\ [a](d)\ | \leq C,\ \ |\ [f](d)\ | \leq C. \tag{4}$$

Here ε_d and ε_c are known as singular perturbation parameters, where $0 < \varepsilon_d \ll 1$, $0 \leq \varepsilon_c \leq 1$. For simplicity, we consider the domain as $\overline{\Gamma} = [0, 1]$ with $\Gamma = (0, 1)$ and $\Gamma^- = (0, d)$, $\Gamma^+ = (d, 1)$. Here $b(x)$ is assumed to be sufficiently smooth function in $\overline{\Gamma}$ and satisfies $b(x) \geq \beta > 0$ and $a(x)$, $f(x)$ are sufficiently smooth in $(\Gamma^- \cup \Gamma^+) \cup \{0, 1\}$. Also $a(x), f(x)$ and its derivatives have a jump discontinuity at $d \in \Gamma$, where the jump of $\omega(x)$ at $x = d$ is denoted as $[\omega](d) = \omega(d+) - \omega(d-)$. These assumptions ensure that the SPP (1)-(2) has a solution $u(x) \in C^0(\overline{\Gamma}) \cap C^1(\Gamma) \cap C^2(\Gamma^- \cup \Gamma^+)$.

Note that $\varepsilon_c = 0$ reduces (1) to a reaction diffusion problem [5] and $\varepsilon_c = 1$ reduces (1) to a convection diffusion problem [6]. The solution behaves similar to dissipative form, when $\varepsilon_d/\varepsilon_c^2 \to 0$ as $\varepsilon_c \to 0$ and it acts in the dispersive form, when $\varepsilon_c^2/\varepsilon_d \to 0$ as $\varepsilon_d \to 0$ [9]. Hence, we consider the following two cases for the numerical discretization:
Case (i): $\sqrt{\alpha}\varepsilon_c \leq \sqrt{\gamma\varepsilon_d}$,
Case (ii): $\sqrt{\alpha}\varepsilon_c \geq \sqrt{\gamma\varepsilon_d}$,
where $\gamma = \min_{\overline{\Gamma}}\{b(x)/\alpha(x)\}$ with $\alpha(x) = -\alpha_1,\ x < d$ and $\alpha(x) = \alpha_2,\ x > d$.

For two-parameter problems with smooth data, the upwind scheme leads to $O(N^{-1}\ln^2 N)$ accuracy on Shishkin mesh [10], where N defines the number of partitions in the domain. Later in [8], the authors established a higher order numerical method by combining the upwind, central difference and mid-point schemes in various partitions of the domain. Their numerical method provides $O(N^{-2}\ln^3 N)$ accuracy, if $\sqrt{\alpha}\varepsilon_c \leq \sqrt{\gamma\varepsilon_d}$ and $O(N^{-2}\ln^2 N)$, if $\sqrt{\alpha}\varepsilon_c \geq \sqrt{\gamma\varepsilon_d}$ on Shishkin mesh. In [11]a discontinuity in source term is introduced and obtained $O(N^{-1}\ln^2 N)$, for $\sqrt{\alpha}\varepsilon_c \leq \sqrt{\gamma\varepsilon_d}$ and $O(N^{-1}\ln^3 N)$, for $\sqrt{\alpha}\varepsilon_c \geq \sqrt{\gamma\varepsilon_d}$ on the Shishkin mesh. Several other adaptive meshes, based on equidistribution principle [4] can be observed for two-parameter problems which leads to first order accuracy.

 ISBN: 978-84-617-8694-7

The above articles motivate us to consider the higher order numerical analysis for two-parameter problems with discontinuous convection and source term which lead to boundary and interior layers.

Throughout this article, C denotes a generic positive constant independent of mesh points, N (assumed to be divisible by 2) and ε_d, ε_c. The infinity norm is denoted as $\|u\|_{\overline{\Omega}} = \max_{x\in\overline{\Omega}} |u(x)|$ for a function $u(x)$ defined on a general domain $\overline{\Omega}$. We denote it by $\|.\| = \|.\|_{\overline{\Omega}}$. Accordingly, the corresponding discrete norm is denoted as $\|.\| = \|.\|_{\overline{\Omega}^N}$.

The paper is arranged as follows. Section 2 presents a discrete problem based on hybrid finite difference scheme corresponding to the continuous problem. Numerical examples in Section 3, validates these theoretical findings. The conclusion is drawn in Section 4.

2 Numerical Approximation

This section introduces the discretization of (1)-(2). The discrete problem will be defined on an *a priori* adaptive piecewise uniform mesh, which is dense inside the boundary and interior layer regions. To construct this mesh, we first divide the domain $\overline{\Gamma}$ into six subintervals:

$$\overline{\Gamma} = [0,\tau_1] \cup [\tau_1, d-\tau_2] \cup [d-\tau_2, d] \cup [d, d+\tau_3] \cup [d+\tau_3, 1-\tau_4] \cup [1-\tau_4, 1],$$

for some τ_1, τ_2, τ_3 and τ_4. The mesh points are denoted as $\overline{\Gamma}^N = \{x_i\}_0^N$ where $x_{N/2}$ denotes the point of discontinuity $x_{N/2} = d$. On these mesh points, we define the discrete solution as U_i. The transitions parameters τ_1, τ_2, τ_3 and τ_4 in $\overline{\Gamma}$ are chosen as follows:

$$\left\{\begin{array}{l} \tau_1 = \min\left\{\dfrac{d}{4}, \dfrac{2}{\theta_2}\ln N\right\}, \quad \tau_2 = \min\left\{\dfrac{d}{4}, \dfrac{2}{\theta_1}\ln N\right\}, \\ \tau_3 = \min\left\{\dfrac{1-d}{4}, \dfrac{2}{\theta_1}\ln N\right\}, \quad \tau_4 = \min\left\{\dfrac{1-d}{4}, \dfrac{2}{\theta_2}\ln N\right\}, \end{array}\right. \tag{5}$$

where

$$\theta_1 = \left\{\begin{array}{ll} \dfrac{\sqrt{\rho\alpha}}{2\sqrt{\varepsilon}}, & \text{if } \sqrt{\alpha}\mu \le \sqrt{\rho\varepsilon}, \\ \dfrac{\alpha\mu}{2\varepsilon}, & \text{if } \sqrt{\alpha}\mu \ge \sqrt{\rho\varepsilon}, \end{array}\right. \text{and } \theta_2 = \left\{\begin{array}{ll} \dfrac{\sqrt{\rho\alpha}}{2\sqrt{\varepsilon}}, & \text{if } \sqrt{\alpha}\mu \le \sqrt{\rho\varepsilon}, \\ \dfrac{\rho}{2\mu}, & \text{if } \sqrt{\alpha}\mu \ge \sqrt{\rho\varepsilon}. \end{array}\right.$$

Now we construct a uniform mesh on each subintervals $[0,\tau_1]$, $[d-\tau_2, d]$, $[d, d+\tau_3]$ and $[1-\tau_4, 1]$, so that each subintervals contains $N/8+1$ uniform mesh points and the subintervals $[\tau_1, d-\tau_2]$ and $[d+\tau_3, 1-\tau_4]$ contains $N/4+1$ uniform mesh points respectively. The mesh sizes in each of the subintervals from left to right of $\overline{\Gamma}$ are denoted as $h_1 = 8\tau_1/N$, $h_2 = 4(d-\tau_2-\tau_1)/N$, $h_3 = 8\tau_2/N$, $h_4 = 8\tau_3/N$, $h_5 = 4(1-d-\tau_3-\tau_4)/N$ and $h_6 = 8\tau_4/N$. On the above adaptive mesh $\overline{\Gamma}^N$, we discretize the BVP (1)-(2) as

$$\begin{array}{l} L^N U_i \equiv \varepsilon_d \delta^2 U_i + \varepsilon_c a_i D^* U_i - b_i U_i = f_i, \quad \text{for } i = 1, 2, ..., N-1, \\ D^+ U_{N/2} - D^- U_{N/2} = 0 \ \text{ with } \ U_0 = u(0), \ U_N = u(1). \end{array} \tag{6}$$

 ISBN: 978-84-617-8694-7

The upwind discretization (6) is almost first order accurate [11] (see the numerical Section 3). Now, we construct an almost second order accurate hybrid scheme for (1)-(2) by combining the following central, upwind and mid-point difference schemes with a five-point scheme:

$$\begin{aligned} L_c^N U_i &\equiv \varepsilon_d \delta^2 U_i + \varepsilon_c a_i D^0 U_i - b_i U_i = f_i, \\ L_u^N U_i &\equiv \varepsilon_d \delta^2 U_i + \varepsilon_c a_i D^* U_i - b_i U_i = f_i, \\ L_m^N U_i &\equiv \varepsilon_d \delta^2 U_i + \varepsilon_c \overline{a}_i D^* U_i - \overline{b_i U_i} = \overline{f}_i, \end{aligned}$$

where, $D^0 U_i = \dfrac{U_{i+1} - U_{i-1}}{h_i + h_{i+1}}$, $D^+ U_i = \dfrac{U_{i+1} - U_i}{h_{i+1}}$, $D^- U_i = \dfrac{U_i - U_{i-1}}{h_i}$,
$\delta^2 U_i = \dfrac{1}{\overline{h}_i}\left(D^+ U_i - D^- U_i\right)$, $\overline{z}_i = \dfrac{z_i + z_{i+1}}{2}$ and $D^* = \begin{cases} D^-, & \text{if } i < N/2, \\ D^+, & \text{if } i > N/2. \end{cases}$

At $x_{N/2} = d$, we use a five-point difference scheme by combining the second order accurate one sided difference approximation $u'(x) \approx (-3U(x)+4U(x+h_3)-U(x+2h_3))/2h_3$, based on forward difference operator and $u'(x) \approx (3U(x) - 4U(x-h_4) + U(x-2h_4))/2h_4$, based on backward difference operator. At the point of discontinuity, we define the scheme

$$L_t^N U_{N/2} \equiv \frac{-U_{N/2+2} + 4U_{N/2+1} - 3U_{N/2}}{2h_4} - \frac{U_{N/2-2} - 4U_{N/2-1} + 3U_{N/2}}{2h_3} = 0, \qquad (7)$$

Now, we define the finite difference scheme (L^N), which involves the central, upwind, mid-point and a five-point scheme on the piecewise uniform mesh which is as follows:
when $\sqrt{\alpha}\varepsilon_c \le \sqrt{\gamma\varepsilon_d}$

$$L^N \equiv \begin{cases} L_c^N, \text{ if } x_i \in (0,\tau_1) \cup (d-\tau_2, d) \cup (d, d+\tau_3) \cup (1-\tau_4, 1), \\ L_c^N, \text{ if } x_i \in (\tau_1, d-\tau_2) \cup (d+\tau_3, 1-\tau_4), \text{ with } 2\varepsilon_c \|a\| h_k < \varepsilon_d, \ k = 2, 5, \\ L_t^N, \text{ if } x_i = d, \end{cases}$$

and for $\sqrt{\alpha}\varepsilon_c \ge \sqrt{\gamma\varepsilon_d}$

$$L^N \equiv \begin{cases} L_m^N, \text{ if } x_i \in (0,\tau_1) \cup (1-\tau_4, 1), \\ L_m^N, \text{ if } x_i \in (\tau_1, d-\tau_2) \cup (d+\tau_3, 1-\tau_4), \text{ with } \begin{cases} 2\varepsilon_c \|a\| h_k \ge \varepsilon_d, \\ 2\|b\| h_k < \varepsilon_c \alpha, \ k = 2, 5, \end{cases} \\ L_u^N, \text{ if } x_i \in (\tau_1, d-\tau_2) \cup (d+\tau_3, 1-\tau_4), \text{ with } \begin{cases} 2\varepsilon_c \|a\| h_k \ge \varepsilon_d, \\ 2\|b\| h_k \ge \varepsilon_c \alpha, \ k = 2, 5, \end{cases} \\ L_c^N, \text{ if } x_i \in (d-\tau_2, d) \cup (d, d+\tau_3), \\ L_t^N, \text{ if } x_i = d. \end{cases}$$

 ISBN: 978-84-617-8694-7

At the transition points τ_1 and $(d+\tau_3)$, the scheme is defined as

$$L^N \equiv \begin{cases} L_c^N, \text{ if } x_i = \begin{cases} \tau_1 = \dfrac{1}{8}, \\ d+\tau_3 = d+\dfrac{1}{8}, \end{cases} \\ L_m^N, \text{ if } x_i = \begin{cases} \tau_1, \text{ where } \tau_1 < \dfrac{1}{8}, \text{ for } 2\|b\|h_2 < \varepsilon_c\alpha, \\ d+\tau_3, \text{ where } d+\tau_3 < d+\dfrac{1}{8}, \text{ for } 2\|b\|h_5 < \varepsilon_c\alpha, \end{cases} \\ L_u^N, \text{ otherwise.} \end{cases}$$

At the transition points $(d-\tau_2)$ and $(1-\tau_4)$, we define the scheme as

$$L^N \equiv \begin{cases} L_c^N, \text{ if } x_i = \begin{cases} d-\tau_2 = d-\dfrac{1}{8}, \text{ for } 2\varepsilon_c\|a\|h_3 < \varepsilon_d, \\ 1-\tau_4 = 1-\dfrac{1}{8}, \text{ for } 2\varepsilon_c\|a\|h_6 < \varepsilon_d, \end{cases} \\ L_m^N, \text{ if } x_i = \begin{cases} d-\tau_2 = d-\dfrac{1}{8}, \text{ for } 2\varepsilon_c\|a\|h_3 \ \geq \varepsilon_d, \\ 1-\tau_4 = 1-\dfrac{1}{8}, \text{ for } 2\varepsilon_c\|a\|h_6 \geq \varepsilon_d, \end{cases} \\ L_m^N, \text{ if } x_i = \begin{cases} d-\tau_2, \text{ where } d-\tau_2 > d-\dfrac{1}{8}, \text{ for } 2\|b\|h_3 < \varepsilon_c\alpha, \\ 1-\tau_4, \text{ where } 1-\tau_4 > 1-\dfrac{1}{8}, \text{ for } 2\|b\|h_6 < \varepsilon_c\alpha, \end{cases} \\ L_u^N, \text{ otherwise.} \end{cases}$$

Now, we define the discrete problem as

$$\begin{aligned} & L^N U_i = Q^N f_i, \text{ for } i = 1, ..., N-1, \text{ with } U_0 = u(0),\ U_N = u(1), \\ & \text{where} \qquad Q^N f_i = \begin{cases} f_i, & \text{if } L^N \equiv L_c^N \text{ or } L_u^N, \\ \overline{f}_i, & \text{if } L^N \equiv L_m^N, \\ 0, & \text{if } L^N \equiv L_t^N. \end{cases} \end{aligned} \tag{8}$$

The matrix associated with (8) does not satisfy M-matrix condition at the point of discontinuity $x_i = d$. But, (without loss of generality) we can convert this five-point difference scheme into a three-point difference scheme (say, $L_T^N U_i$) by estimating $U_{N/2-2}$, $U_{N/2+2}$ from $L_c^N U_i$, so that the new equations do have the monotonicity property. To do this, note

$$\begin{aligned} U_{N/2-2} \ = & \frac{2h_3}{2\varepsilon_d - h_3\varepsilon_c a_{N/2-1}} \left[h_3 f_{N/2-1} + \left(\frac{2\varepsilon_d}{h_3} + h_3 b_{N/2-1} \right) U_{N/2-1} \right. \\ & \left. - \left(\frac{2\varepsilon_d + h_3\varepsilon_c a_{N/2-1}}{2h_3} \right) U_{N/2} \right], \end{aligned}$$

 ISBN: 978-84-617-8694-7

$$U_{N/2+2} = \frac{2h_4}{2\varepsilon_d + h_4\varepsilon_c a_{N/2+1}}\left[h_4 f_{N/2+1} + \left(\frac{2\varepsilon_d}{h_4} + h_4 b_{N/2+1}\right)U_{N/2+1} - \left(\frac{2\varepsilon_d - h_4\varepsilon_c a_{N/2+1}}{2h_4}\right)U_{N/2}\right].$$

Now we replace the above expressions of $U_{N/2-2}$, $U_{N/2+2}$ of five-point difference scheme $(L_t^N U_{N/2})$ to construct a three point scheme $(L_T^N U_{N/2})$ which preserves the monotonicity property and leads to an higher order accuracy at the point of discontinuity.

$$\begin{aligned} L_T^N U_{N/2} \equiv & \left(\frac{2\varepsilon_d - h_4\varepsilon_c a_{N/2+1}}{2\varepsilon_d + h_4\varepsilon_c a_{N/2+1}} - 6 + \frac{2\varepsilon_d - h_3\varepsilon_c a_{N/2-1}}{2\varepsilon_d + h_3\varepsilon_c a_{N/2-1}}\right)U_{N/2} \\ & + \left(\frac{-4\varepsilon_d - 2h_4^2 b_{N/2+1}}{2\varepsilon_d + h_4\varepsilon_c a_{N/2+1}} + 4\right)U_{N/2+1} + \left(\frac{-4\varepsilon_d - 2h_3^2 b_{N/2-1}}{2\varepsilon_d + h_3\varepsilon_c a_{N/2-1}} + 4\right)U_{N/2-1} \\ = & \frac{2h_3^2 f_{N/2-1}}{2\varepsilon_d + h_3\varepsilon_c a_{N/2-1}} + \frac{2h_4^2 f_{N/2+1}}{2\varepsilon_d + h_4\varepsilon_c a_{N/2+1}}. \end{aligned}$$

So, the reformulated discrete operator (say $L_*^N U_i$) of (8) can be written as

$$L_*^N U_i = Q_*^N f_i, \text{ for } i = 1, 2, ..., N-1, \tag{9}$$

$$U_0 = u(0),\ U_N = u(1), \tag{10}$$

where

$$L_*^N U_i = \begin{cases} L_T^N U_i, & \text{for } i = N/2, \\ L^N U_i, & \text{for } i \neq N/2, \end{cases}$$

and

$$Q_*^N f_i = \begin{cases} \dfrac{2h_3^2 f_{N/2-1}}{2\varepsilon_d + h_3\varepsilon_c a_{N/2-1}} + \dfrac{2h_4^2 f_{N/2+1}}{2\varepsilon_d + h_4\varepsilon_c a_{N/2+1}}, & \text{if } i = N/2, \\ Q^N f_i, & \text{if } i \neq N/2. \end{cases}$$

Finally the discrete problem (9)-(10) form a traditional system of equation which will be solved using a TDMA Solver to obtain the numerical solution. The parameter uniform higher order error estimate of the computed solution for sufficiently large N satisfy the following bound

$$\|U - u\| \leq \begin{cases} C(N^{-1}\ln N)^2, & \text{if } \sqrt{\alpha}\varepsilon_c \leq \sqrt{\gamma\varepsilon_d}, \\ CN^{-2}(\ln N)^3, & \text{if } \sqrt{\alpha}\varepsilon_c \geq \sqrt{\gamma\varepsilon_d}, \end{cases} \quad \forall\ 0 \leq i \leq N.$$

where $u(x_i)$ be the solution of the continuous problem (1)-(2) and $U(x_i)$ be the solution of the discrete problem (9)-(10). Further a study on stability and error analysis of the proposed scheme will be presented in nearest future.

 ISBN: 978-84-617-8694-7

3 Numerical Examples

This section experimentally demonstrates the applicability of the hybrid scheme (9)-(10) and compare it with the existed upwind scheme (6).

Example 1 *Consider the two-parameter problem (1)-(2) with the following discontinuous convection coefficient and source term:*

$$a(x) = \begin{cases} -(1+x(1-x)), & for \quad 0 \le x \le 0.5, \\ (1+x(1-x)), & for \quad 0.5 < x \le 1, \end{cases} \quad b(x) = 1,$$

$$f(x) = \begin{cases} -2(1+x^2), & for \quad 0 \le x \le 0.5, \\ 3(1+x^2), & for \quad 0.5 < x \le 1, \end{cases} \quad and\ u(0) = u(1) = 0.$$

As the exact solution of Example 1 is unknown, we use the double mesh principle [1, 3, 4, 12] to calculate the maximum pointwise error $E^N_{\varepsilon_d,\varepsilon_c}$ and corresponding order of convergence $\rho^N_{\varepsilon_d,\varepsilon_c}$ of the numerical solution provided by the scheme (9) which is as follows:

$$E^N_{\varepsilon_d,\varepsilon_c} = \max_{0\le i\le N} |U^N_i - U^{2N}_i| \quad \text{and} \quad \rho^N_{\varepsilon_d,\varepsilon_c} = \log_2 \left(E^N_{\varepsilon_d,\varepsilon_c} / E^{2N}_{\varepsilon_d,\varepsilon_c} \right).$$

Here U^N_i denotes the numerical solution obtained with N number of mesh intervals and U^{2N}_i denotes the solution on $2N$ number of mesh intervals obtained by bisecting the previous original mesh. Similarly, we find the uniform error E^N and order of convergence ρ^N of (6) for a fixed ε_c and various values of ε_d, taken from the set $S = \{\varepsilon_d | \varepsilon_d = 10^{-2}, 10^{-4}, \cdots, 10^{-14}\}$:

$$E^N = \max_{\varepsilon_d \in S} E^N_{\varepsilon_d,\varepsilon_c} \quad \text{and} \quad \rho^N = \log_2 \left(E^N / E^{2N} \right).$$

Here, the numerical experiment is performed by choosing the constant values α=2.5, β=1.0 and γ=0.2 for Example 1.

We present $E^N_{\varepsilon_d,\varepsilon_c}$ and $\rho^N_{\varepsilon_d,\varepsilon_c}$ in Tables 1 and 2 respectively for Example 1. Table 2 shows that the order of convergence is almost second order $O(N^{-2}\ln^2 N)$ when $\sqrt{\alpha}\varepsilon_c \le \sqrt{\gamma\varepsilon_d}$ and $O(N^{-2}\ln^3 N)$ when $\sqrt{\alpha}\varepsilon_c \ge \sqrt{\gamma\varepsilon_d}$ for the above example. Table 3 presents the maximum error E^N and order of convergence ρ^N using the standard upwind scheme (6). This table shows that the existing scheme gives first order parameter uniform convergence, while the Table 2 shows almost second order convergence using the hybrid difference scheme. Note that the errors are also lower in Table 1 compared to Table 3.

Loglog error plot in Figure 1 also demonstrates that the predicted order of convergence, i.e., $O(N^{-2}\ln^3 N)$, if $\sqrt{\alpha}\varepsilon_c \le \sqrt{\gamma\varepsilon_d}$ and $O(N^{-2}\ln^2 N)$, if $\sqrt{\alpha}\varepsilon_c \ge \sqrt{\gamma\varepsilon_d}$ are correct.

4 Conclusion

In this paper, an almost second order uniformly convergent numerical solution is obtained for a two-parameter singularly perturbed problem where the convection coefficient and source

Table 1: Maximum Pointwise Errors ($E^N_{\varepsilon_d,\varepsilon_c}$) with $\varepsilon_c = 10^{-4}$ for Example 1

ε_d	Number of mesh points N				
	128	256	512	1024	2048
10^{-2}	3.01720E-05	7.03000E-06	1.21160E-06	1.77190E-07	2.39400E-08
10^{-4}	1.55370E-03	7.67900E-04	2.02340E-04	5.52550E-05	1.32040E-05
10^{-6}	1.30140E-03	6.03150E-04	2.63820E-04	1.09680E-04	3.83720E-05
10^{-8}	1.45400E-03	3.22960E-04	1.05520E-04	2.74780E-05	8.60100E-06
10^{-10}	4.45260E-03	2.01520E-03	7.84000E-04	2.43460E-04	7.18400E-05
10^{-12}	4.48230E-03	2.03110E-03	7.90700E-04	2.45440E-04	7.25100E-05
10^{-14}	4.48260E-03	2.03130E-03	7.90800E-04	2.45460E-04	7.25150E-05

Table 2: Orders of Convergence ($\rho^N_{\varepsilon_d,\varepsilon_c}$) with $\varepsilon_c = 10^{-4}$ for Example 1

ε_d	Number of mesh points N				
	128	256	512	1024	2048
10^{-2}	2.101589829	2.536670744	2.773525566	2.887791549	2.944105039
10^{-4}	1.016671183	1.924101248	1.872640276	2.065184608	2.476653267
10^{-6}	1.109475717	1.192935556	1.266280804	1.515174609	1.550321966
10^{-8}	2.170649483	1.613839022	1.941167564	1.675700690	1.594048784
10^{-10}	1.143741186	1.361997468	1.687198524	1.760795875	2.076963082
10^{-12}	1.141926332	1.361094457	1.687760037	1.759118509	2.075855289
10^{-14}	1.141916350	1.361018549	1.687824928	1.759136585	2.075787013

Table 3: Maximum Pointwise Errors (E^N) and Orders of Convergence (ρ^N) with $\varepsilon_c = 10^{-4}$ for Example 1

$\varepsilon_d \in S$	Number of mesh points N				
	128	256	512	1024	2048
E^N	1.16250E-01	1.00500E-01	7.58210E-02	5.27330E-02	3.30160E-02
ρ^N	0.210035215	0.406526112	0.523891410	0.675540731	0.748636031

 ISBN: 978-84-617-8694-7

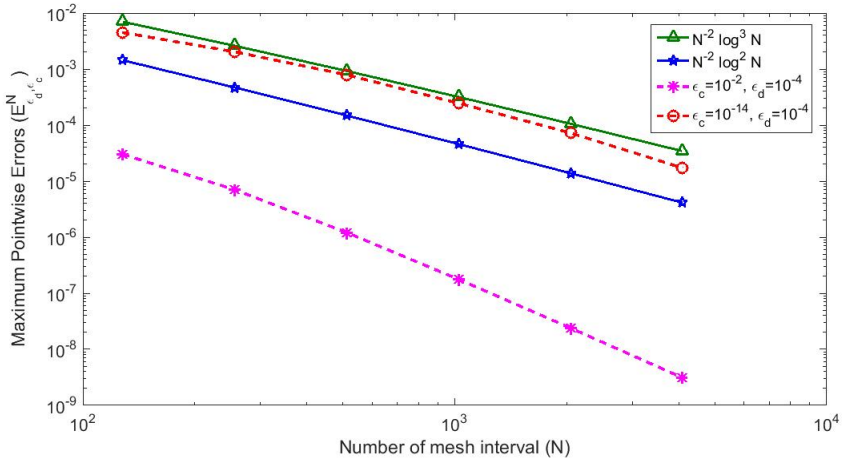


Figure 1: Loglog Plot of the Maximum Pointwise Errors for Example 1

term have a jump discontinuity at an interior point of the domain. The present hybrid difference scheme is a combination of upwind, midpoint and central difference schemes with a five point scheme at the point of discontinuity which preserves the monotonicity property on Shishkin mesh. The computational result shows that the proposed scheme is uniformly convergent and provides better accuracy than standard upwind difference. Therefore, looking towards the better numerical performance of the proposed scheme, it can be concluded that one can take the above computational analysis as a motivation for further theoretical analysis of the proposed scheme.

Acknowledgements

The first and fourth authors wish to thank Department of Science and Technology(SERB), New Delhi for financial support of project SR/FTP/MS-039/2012.

References

[1] P. A. Farrell, A. Hegarty, J. J. H. Miller, E. O'Riordan and G. I. Shishkin, *Robust computational techniques for boundary layers*, CRC Press, 2000.

[2] Z. Cen, *A hybrid difference scheme for a singularly perturbed convection-diffusion problem with discontinuous convection coefficient*, Applied Mathematics and Computation **169(1)** (2005) 689–699.

[3] P. Das, S. Natesan, *Numerical solution of a system of singularly perturbed convection-diffusion boundary-value problems using mesh equidistribution technique*, The Australian Journal of Mathematical Analysis and Applications **10** (2013) 1–17.

[4] P. Das and V. Mehrmann, *Numerical solution of singularly perturbed convection-diffusion-reaction problems with two small parameters*, BIT Numerical Mathematics **56(1)** (2016) 51–76.

[5] P. A. Farrell, J. J. H. Miller, E. O'Riordan and G. I. Shishkin, *Singularly perturbed differential equations with discontinuous source terms*, appeared in proceedings of Workshop'98, Lozenetz, Bulgaria, Aug. 27-31 (2000) 147–156.

[6] P. A. Farrell, A. Hegarty, J. J. H. Miller, E. O'Riordan and G. I. Shishkin, *Singularly perturbed convection–diffusion problems with boundary and weak interior layers*, Journal of Computational and Applied Mathematics **166(1)** (2004) 133–151.

[7] P. A. Farrell, A. Hegarty, J. J. H. Miller, E. O'Riordan and G. I. Shishkin, *Global maximum norm parameter-uniform numerical method for a singularly perturbed convection-diffusion problem with discontinuous convection coefficient*, Mathematical and Computer Modelling **40(11)** (2004) 1375–1392.

[8] J. L. Gracia, E. O'Riordan and M. L. Pickett, *A parameter robust second order numerical method for a singularly perturbed two-parameter problem*, Applied Numerical Mathematics **56(7)** (2006) 962–980.

[9] K. C. Patidar, *A robust fitted operator finite difference method for a two-parameter singular perturbation problem*, Journal of Difference Equations and Applications **14(12)** (2000) 1197–1214.

[10] E. O'Riordan, M. L. Pickett and G. I. Shishkin, *Singularly perturbed problems modeling reaction-convection-diffusion processes*, Computational Methods in Applied Mathematics **3(3)** (2003) 424–442.

[11] V. Shanthi, N. Ramanujam and S. Natesan, *Fitted mesh method for singularly perturbed reaction-convection-diffusion problems with boundary and interior layers*, Journal of Applied Mathematics and Computing **22(1-2)** (2006) 49–65.

[12] P. Das, S. Natesan, *Adaptive mesh generation for singularly perturbed fourth-order ordinary differential equations*, International Journal of Computer Mathematics **92** (2015) 562-578.

[13] R. Vulanović, *A higher-order scheme for quasilinear boundary value problems with two small parameters*, Computing **67(4)** (2001) 287–303.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

Accelerated POD least-squares approach for missing data reconstruction

Saifon Chaturantabut[1]

[1] *Department of Mathematics and Statistics, Faculty of Science and Technology, Thammasat University, Thailand*

emails: saifon@mathstat.sci.tu.ac.th

Abstract

This work introduces a technique that reduces the computational complexity, and therefore decreases the simulation time, for approximating missing data components through the least-squares (LS) approximation with the basis from proper orthogonal decomposition (POD). In the existing standard approach, POD basis is first computed to extract dominant trends of data from existing set of complete samples. Then, the approximation of each missing data is obtained optimally in the least-squares sense. However, in the case of high-dimensional data, the reconstruction process might require high-complexity computation. To accelerate this approximation process, this work employs certain *important* and *relevant* available data components, which are selected through a greedy iterative procedure, called discrete empirical interpolation method (DEIM). This work introduces two different variants of the acceleration, which are tested on the subsurface flow data. The numerical results demonstrate that the proposed technique can be as accurate as the standard POD-LS method with much less computational time.

Key words: Data reconstruction, Proper orthogonal decomposition, Least-squares method

1 Introduction

Missing data estimation has been an important problem in many engineering applications, such as in the reconstruction of unavailable experimental data or in image processing research. A popular approach for solving this problem is based on proper orthogonal decomposition(POD). The approach using POD for the purpose of data reconstruction is often called gappy POD (GPOD). GPOD method essentially uses POD basis in the least-squares

 ISBN: 978-84-617-8694-7

approximation, and it will be also called *POD-LS* method in this work. In the application of aerodynamic flow fields, GPOD was formally introduced in [7] and it was later used to calibrate and illustrate air flow past a wing [22]. GPOD has been recently used in many other engineering applications, such as in chemical engineering [5, 11], in mechanical engineering [6, 27], in image processing [19], and in optimization of water flooding reservoir [15].

Proper orthogonal decomposition (POD) was introduced in 1937 by Lumley in the context of inhomogeneous structure turbulent flows [21] and stochastic tools in turbulence [20]. POD is also known as, for example, Karhunen-Love decomposition (KLD), principal component analysis (PCA), or singular value decomposition(SVD). POD has been successfully used in many applications, e.g. [4, 14, 18, 24], since it can provide an approximation from the basis that extracts the dominant characteristic of the existing data. Discrete empirical interpolation method (DEIM) [10], a variant of [3], was originally introduced to estimate nonlinear term in dynamical systems by selecting the interpolation indices using a greedy algorithm. The index selection process in this algorithm is based on trying to capture most variation of the sample set heuristically. It has been used in many applications, such as morphological structure spiking neurons [17], 2-D shallow-water equations [12], Navier-Stokes equations [1], four-dimensional variational data assimilation [26], three-dimensional nonlinear aeroelasticity model [13], electrical, thermal, and microelectromechanical models [16].

To derive the accelerated POD-LS method for decreasing computational time in reconstructing missing data components, this work first reviews some background of POD and GPOD in Section 2. Then, DEIM approximation and the corresponding greedy process for selecting components are discussed in Section 3.1. The combination of DEIM with the POD-LS method or Gappy POD to accelerate the approximation procedure is presented in Section 3.2. The numerical results in Section 4 demonstrate the accuracy of the proposed method through the reconstruction of missing components in concentration flow data with much less computational time. Finally, the conclusion and future extension are discussed in Section 5.

2 Standard POD-LS method (Gappy POD)

This section reviews some relevant background for constructing missing data based on proper orthogonal decomposition (POD) and least-squares (LS) approximations. The combination of these approaches is also known as gappy POD (GPOD) method.

2.1 Proper Orthogonal Decomposition (POD)

Let $\{\mathbf{y}_j\}_{j=1}^{n_s} \subset \mathbb{R}^n$ be the set of snapshots or samples. Proper orthogonal decomposition (POD) can be used to extract important features of the sample sets. In particular, consider an approximation of $\mathbf{y}_j$ by using orthogonal projection onto the space spanned by the set of vectors $\{\mathbf{v}_i\}_{i=1}^{k} \subset \mathbb{R}^n$ in the form $\mathbf{y}_j \approx \sum_{i=1}^{k} \mathbf{v}_i(\mathbf{v}_i^T\mathbf{y}_j) = \mathbf{V}\mathbf{V}^T\mathbf{y}_j$. Then, if $\{\mathbf{v}_i\}_{i=1}^{k}$ is the

 ISBN: 978-84-617-8694-7

POD basis, it is an orthonormal basis that minimizes the approximation error in $2-$norm for a given fixed basis rank $k < n$, i.e. $\{\mathbf{v}_i\}_{i=1}^k$ solves the following minimization problem:

$$\min_{\{\phi_i\}_{i=1}^k} \sum_{j=1}^{n_s} \|\mathbf{y}_j - \sum_{i=1}^{k} \phi_i(\phi_i^T \mathbf{y}_j)\|_2^2, \qquad \phi_i^T \phi_j = \delta_{ij},$$

where $\delta_{ij} = 0$ if $i \neq j$ and $\delta_{ij} = 1$ if $i = j$. Singular value decomposition (SVD) turns out to be corresponding to this solution of POD basis. That is, POD basis of rank k consists of the first k *left* singular vectors from SVD of the snapshot matrix: $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_{n_s}] \in \mathbb{R}^{n \times n_s}$. Suppose the SVD of a rectangular matrix $\mathbf{Y} \in \mathbb{R}^{n \times n_s}$ is given by $\mathbf{Y} = \hat{\mathbf{V}} \Sigma \mathbf{Z}^T$, where r is the rank of $\mathbf{Y}$, $\hat{\mathbf{V}} = [\mathbf{v}_1, \ldots, \mathbf{v}_r] \in \mathbb{R}^{n \times r}$ and $\mathbf{Z} \in \mathbb{R}^{n_s \times r}$ are orthogonal matrices and $\Sigma = diag(\sigma_1, \ldots, \sigma_r) \in \mathbb{R}^{r \times r}$ with singular values in decreasing order: $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0$.

It can be shown [23] that the minimum error of the approximation by using POD basis is given by the sum of the neglected singular values $\sigma_{k+1}, \ldots, \sigma_r$. I.e.,

$$\sum_{j=1}^{n_s} \|\mathbf{y}_j - \mathbf{V}\mathbf{V}^T \mathbf{y}_j\|_2^2 = \sum_{\ell=k+1}^{r} \sigma_\ell^2. \tag{2.1}$$

Note that, besides using SVD, the POD basis can be computed by using the method of snapshots based on eigenvalue decomposition of correlation matrix of the snapshots [25]. The procedure for computing POD basis is shown in Algorithm 1.

Algorithm 1 Algorithm for constructing POD basis

INPUT: Snapshots $\{\mathbf{y}_j\}_{j=1}^{n_s} \subset \mathbb{R}^n$

OUTPUT: POD basis $\mathbf{V}_k$.

1: Create snapshot matrix : $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_{n_s}] \in \mathbb{R}^{n \times n_s}$and let $r = rank(\mathbf{Y})$
2: Compute SVD: $\mathbf{Y} = \hat{\mathbf{V}} \Sigma \mathbf{Z}^T$ and choose dimension $k \leq r$
3: POD basis of rank k : $\mathbf{V} = [\mathbf{v}_1, \ldots, \mathbf{v}_k] = \hat{\mathbf{V}}(:, 1:k)$

2.2 Gappy POD

Gappy POD (GPOD) can be used to approximate or reconstruct missing data from the available partial data, that obtained, e.g. from experimental measurements or numerical simulations.

Define $\mathcal{Y} := \{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_{n_s}\} \subset \mathbb{R}^n$. Let $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_{n_s}] \in \mathbb{R}^{n \times n_s}$. Suppose $\widehat{\mathbf{y}}$ is an incomplete sample which consists of n_c known components and $n_g = n - n_c$ unknown components. Let $\mathcal{C} := \{\wp_1, \wp_2, \ldots, c_{n_c}\} \subset \{1, 2, \ldots, n\}$ be the indices of the *known* components in $\widehat{\mathbf{y}}$ and define $\mathbf{C} = [\mathbf{e}_{c_1}, \ldots, \mathbf{e}_{c_{n_c}}] \in \mathbb{R}^{n \times n_c}$, where $\mathbf{e}_{c_i} = [0, \ldots, 0, 1, 0, \ldots, 0]^T \in \mathbb{R}^n$ is the c_i-th column of the identity matrix $\mathbf{I}_n \in \mathbb{R}^{n \times n}$, for $i = 1, \ldots, m$. Note that, pre-multiplying

 ISBN: 978-84-617-8694-7

$\mathbf{C}^T$ is equivalent to extracting the n_c rows corresponding to the indices $c_1, \ldots, c_{n_c},$. Similarly, let $\mathcal{G} = \{g_1, g_2, \ldots, g_{n_g}\} \subset \{1, 2, \ldots, n\}$ be the indices of the *unknown* components in $\widehat{\mathbf{y}}$ and define $\mathbf{G} = [\mathbf{e}_{g_1}, \ldots, \mathbf{e}_{g_{n_g}}] \in \mathbb{R}^{n \times n_g}$. I.e. the known components and the unknown components are given in the following two vectors $\widehat{\mathbf{y}}_c \in \mathbb{R}^{n_c}$ and $\widehat{\mathbf{y}}_g \in \mathbb{R}^{n_g}$, respectively. The missing components contained in $\widehat{\mathbf{y}}_g$ will be approximated by first performing a projection onto the column span of POD basis matrix $\mathbf{V}$ with rank k from Algorithm 1, i.e.

$$\widehat{\mathbf{y}} \;\approx\; \mathbf{V}\mathbf{a}, \qquad \text{or} \qquad \widehat{\mathbf{y}}_c \approx \mathbf{V}_c\mathbf{a} \;\text{ and }\; \widehat{\mathbf{y}}_g \approx \mathbf{V}_g\mathbf{a},$$

for some coefficient vector $\mathbf{a} \in \mathbb{R}^k$, and where $\mathbf{V}_c := \mathbf{C}^T\mathbf{V} \in \mathbb{R}^{n_c \times k}$, $\mathbf{V}_g := \mathbf{G}^T\mathbf{V} \in \mathbb{R}^{n_g \times k}$.

The known components contained in $\widehat{\mathbf{y}}_c = \mathbf{C}^T\widehat{\mathbf{y}}$ are then used to determine the coefficient vector $\mathbf{a}$ through the approximation $\widehat{\mathbf{y}}_c \approx \mathbf{V}_c\mathbf{a}$ from the following least-squares problem:

$$\min_{\mathbf{a}\in\mathbb{R}^k} \|\widehat{\mathbf{y}}_c - \mathbf{V}_c\mathbf{a}\|_2^2. \tag{2.2}$$

The solution obtained from the corresponding normal equation of the above problem is given by $\mathbf{a} = (\mathbf{V}_c^T\mathbf{V}_c)^{-1}\mathbf{V}_c^T\widehat{\mathbf{y}}_c$. That is,

$$\widehat{\mathbf{y}}_g \;\approx\; \mathbf{V}_g\mathbf{a} \;=\; \mathbf{V}_g(\mathbf{V}_c^T\mathbf{V}_c)^{-1}\mathbf{V}_c^T\widehat{\mathbf{y}}_c. \tag{2.3}$$

The steps described above are summarized in Algorithm 2.

Algorithm 2 Gappy POD method for approximating missing data

INPUT:

- Complete snapshot set $\{\mathbf{y}_j\}_{j=1}^{n_s} \subset \mathbb{R}^n$ and, dimension $k \leq rank(\{\mathbf{y}_j\}_{j=1}^{n_s})$
- Incomplete data $\widehat{\mathbf{y}} \in \mathbb{R}^n$ with known entries $\widehat{y}_j$, $j \in \mathcal{C}$ and unknown entries $\widehat{y}_j$, $j \in \mathcal{G}$

OUTPUT:

- Approximation: $\widehat{\mathbf{y}}_g = [\widehat{y}_j]$, $j \in \mathcal{G} = \{g_1, g_2, \ldots, g_{n_g}\}$

1: Create snapshot matrix : $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_{n_s}] \in \mathbb{R}^{n \times n_s}$and let $r = rank(\mathbf{Y})$.
2: Compute POD basis $\mathbf{V}$ of rank $k \leq r$ for $\mathbf{Y}$ from Algorithm 1.
3: Find coefficient vector $\mathbf{a}$ from $\widehat{\mathbf{y}}_c$ using least-squares problem in (2.2): $\min_{\mathbf{a}\in\mathbb{R}^k} \|\widehat{\mathbf{y}}_c - \mathbf{V}_c\mathbf{a}\|_2^2$.
4: Compute the approximation $\widehat{\mathbf{y}}_g \;\approx\; \mathbf{V}_g\mathbf{a}$.

3 Accelerated POD-LS method (Gappy POD)

First, this section describes a greedy-based technique called discrete empirical interpolation method (DEIM), which is used in this work to reduce computational complexity of the standard GPOD (LS-POD) approach. Two small variants of missing data approximations using DEIM is then later derived.

 ISBN: 978-84-617-8694-7

3.1 Discrete Empirical Interpolation Method (DEIM)

DEIM was first introduced for the purpose of approximating the nonlinear term in the differential equations [9]. In this work, the interpolation indices from this method will be used to reduce the computational complexity of GPOD in Step 3 of Algorithm 2. This section describes DEIM in a general setting and provides the corresponding greedy algorithm for selecting *important* components that can be used in the approximation.

Consider a vector $\mathbf{f} \in \mathbb{R}^n$ by projecting on a low-dimensional subspace, span$\{\mathbf{U}\}$ where $\mathbf{U} \in \mathbb{R}^{n\times m}$ is the matrix of rank $m \leq n$ with orthogonal columns, i.e. $\mathbf{f} \approx \mathbf{U}\mathbf{d}$ for $\mathbf{d} \in \mathbb{R}^m$. The m components of the nonlinear vector $\mathbf{f}$ are selected by a greedy algorithm [2,9] to specify $\mathbf{d}$ using the 2-norm least-squares approximation. In particular, suppose $\wp_1, \wp_2, \ldots, \wp_m$ are the indices of the selected components in $\mathbf{f}$ and define $\mathbf{P} = [\mathbf{e}_{\wp_1}, \ldots, \mathbf{e}_{\wp_m}] \in \mathbb{R}^{n\times m}$, where $\mathbf{e}_{\wp_i} = [0, \ldots, 0, 1, 0, \ldots, 0]^T \in \mathbb{R}^n$ is the $\wp_i$-th column of the identity matrix $\mathbf{I}_n \in \mathbb{R}^{n\times n}$, for $i = 1, \ldots, m$. Note that, as in the previous section, pre-multiplying $\mathbf{P}^T$ is equivalent to extracting the m rows corresponding to the interpolation indices $\wp_1, \ldots, \wp_m$,. The coefficient vector $\mathbf{d}$ in the DEIM approximation solves the following minimization problem

$$\min_{\mathbf{d}\in\mathbb{R}^m} \|\mathbf{P}^T\mathbf{f} - \mathbf{P}^T\mathbf{U}\mathbf{d}\|_2^2$$

which gives $\mathbf{d} = (\mathbf{P}^T\mathbf{U})^+\mathbf{P}^T\mathbf{f}$, where $(\mathbf{P}^T\mathbf{U})^+ = [(\mathbf{P}^T\mathbf{U})^T(\mathbf{P}\mathbf{U})]^{-1}(\mathbf{P}^T\mathbf{U})^T$ is the pseudo-inverse of $\mathbf{P}^T\mathbf{U}$. Since $\mathbf{P}^T\mathbf{U}$ is a square matrix and it is shown to be invertible [9], $(\mathbf{P}^T\mathbf{U})^+ = (\mathbf{P}^T\mathbf{U})^{-1}$ and the DEIM approximation for $\mathbf{f}$ becomes

$$\mathbf{f} \approx \mathbf{U}(\mathbf{P}\mathbf{U})^{-1}\mathbf{P}^T\mathbf{f}.$$

The sets of indices $\{\wp_1, \wp_2, \ldots, \wp_m\}$ is obtained by DEIM index selection algorithm [9], shown in Algorithm 3, which is a greedy procedure that aim to capture the variation of the spatial behavior of the input basis using the infinity norm.

From Algorithm 3, DEIM selects the interpolation indices so that the approximation has smallest error $\mathbf{r} = \mathbf{u}_j - \mathbf{U}\mathbf{c}$ in each iteration j. The procedure of DEIM Algorithm 3 can be described as follows. First, the input a basis of rank m, which can be obtained by using POD of nonlinear term. Then, it selects the first index of a component in the first basis vector $\mathbf{u}_1$ with the largest absolute value. Next, each of the other indices is selected from the component with largest absolute residual error $\mathbf{r} = \mathbf{u}_\ell - \mathbf{U}\mathbf{c}$ in each step.

The corresponding error of DEIM approximation was proposed in [9] and an extension of this error bound of to the state-space error estimate can be found in [8].

3.2 Accelerated POD-LS method by using DEIM

This work particularly focuses on the situation when the incomplete data $\widehat{\mathbf{y}}$ is in a high-dimensional space, i.e. the value of n is large, which may result in a large number of known

 ISBN: 978-84-617-8694-7

Algorithm 3 Algorithm to create for Interpolation Indices DEIM

INPUT: $\{\mathbf{u}_\ell\}_{\ell=1}^m \subset \mathbb{R}^n$ linearly independent
OUTPUT: $\vec{\wp} = [\wp_1, \ldots, \wp_m]^T \in \mathbb{R}^m$ and $\mathbf{P}$

1: $\wp_1 = \arg\max_{i=1,2,\ldots,n}\{|\mathbf{u}_{i1}|\}$
2: $\mathbf{U} = [\mathbf{u}_1], \mathbf{P} = [\mathbf{e}_{\wp_1}], \vec{\wp} = [\wp_1]$;
3: **for** $j = 2$ to m **do**
4: Solve $(\mathbf{P}^T\mathbf{U})\mathbf{c} = \mathbf{P}^T\mathbf{u}_j$;
5: $\mathbf{r} = \mathbf{u}_j - \mathbf{U}\mathbf{c}$
6: $\wp_j = \arg\max_{i=1,\ldots,n}\{|\mathbf{r}_i|\}$
7: $\mathbf{U} \leftarrow [\mathbf{U} \quad \mathbf{u}_j], \mathbf{P} \leftarrow [\mathbf{P} \quad \mathbf{e}_{\wp_j}], \vec{\wp} \leftarrow \begin{bmatrix} \vec{\wp} \\ \wp_j \end{bmatrix}$
8: **end for**

components n_c, even though there are a lot of unknown components n_g. In this case, the main computational work in approximating missing components may occur while solving for the coefficient vector $\mathbf{a}$ in (2.2) or in Step 3 of Algorithm 2. One possible way to reduce this computational work is to use only small number known components to specify $\mathbf{a} \in \mathbb{R}^k$. These components have to be carefully selected so that they can represent all other n_c components and maintain the same accuracy in the approximation. For this purpose, the procedure for selecting DEIM indices of *important* components from Algorithm 3 will be used as described next.

Recall from Step 3 in Algorithm 2 that, it has to solve for $\mathbf{a}$ from

$$\min_{\mathbf{a}\in\mathbb{R}^k} \|\widehat{\mathbf{y}}_c - \mathbf{V}_c\mathbf{a}\|_2^2.$$

Suppose $\mathbf{V}_c \in \mathbb{R}^{n_c\times k}$ has linearly independent columns. Then, for a given DEIM dimension $m \leq \min\{n_c, k\}$, the columns of $\mathbf{V}_c$ can be used as an input of Algorithm 3 to select m DEIM indices that can cover the variations of n_c components.

In many applications when n is much larger than the number of available complete data n_s , i.e. $n > n_s \geq k$, we have $\min\{n, k\} = k$. If we want to use DEIM dimension $m > k$, using $\mathbf{V}_c = \mathbf{C}^T\mathbf{V} \in \mathbb{R}^{n_c\times k}$ will not be enough and this could limit the accuracy of the approximation. Therefore, in this work, all columns of the left singular matrix $\bar{\mathbf{V}} \in \mathbb{R}^{n\times r}$ from SVD of complete snapshot matrix $\mathbf{Y}$ will be used, where $r = rank(\mathbf{Y})$. In particular, consider the matrix

$$\bar{\mathbf{V}}_c := \mathbf{C}^T\bar{\mathbf{V}} \in \mathbb{R}^{n_c\times r}, \tag{3.1}$$

which comes from $\bar{\mathbf{V}}$ with selected rows corresponding to the n_c known components in $\mathcal{C}$.

Suppose $\bar{\mathbf{V}}_c \in \mathbb{R}^{n_c\times r}$ has linearly independent columns. Then, the columns of $\bar{\mathbf{V}}_c$ can be used in Algorithm 3 to select m DEIM indices, where the largest dimension m is $\min\{n_c, r\}$.

 ISBN: 978-84-617-8694-7

In this case, suppose $\wp_1, \wp_2, \ldots, \wp_m$ are the output indices of DEIM Algorithm 3 and let $\mathbf{P} = [\mathbf{e}_{\wp_1}, \ldots, \mathbf{e}_{\wp_m}] \in \mathbb{R}^{n_c \times m}$ as defined earlier in this section. The vector $\mathbf{a}$ in (2.2) can be computed from a smaller least-squares problem:

$$\min_{\mathbf{a}\in\mathbb{R}^k} \|\mathbf{P}^T\widehat{\mathbf{y}}_c - \mathbf{P}^T\bar{\mathbf{V}}_c\mathbf{a}\|_2^2. \tag{3.2}$$

Note that, we have assumed that $\bar{\mathbf{V}}_c \in \mathbb{R}^{n_c \times r}$ has linearly independent columns, which may not hold in general cases. To avoid this assumption, we can apply SVD on the matrix $\bar{\mathbf{V}}_c$ and use the set of first m corresponding left singular vectors as an input of DEIM algorithm 3. The accelerated POD-LS steps are summarized in Algorithm 4 for both cases of (i) using $\bar{\mathbf{V}}_c$ directly as an input (ii) using the set of left singular vectors of $\bar{\mathbf{V}}_c$ as an input to DEIM algorithm 3.

Algorithm 4 Accelerated POD-LS method for approximating missing data

INPUT:

- Complete snapshot set $\{\mathbf{y}_j\}_{j=1}^{n_s} \subset \mathbb{R}^n$ and, dimension $k \leq rank(\{\mathbf{y}_j\}_{j=1}^{n_s})$
- Incomplete data $\widehat{\mathbf{y}} \in \mathbb{R}^n$ with known entries $\widehat{y}_j$, $j \in \mathcal{C}$ and unknown entries $\widehat{y}_j$, $j \in \mathcal{G}$

OUTPUT:

- Approximation: $\widehat{\mathbf{y}}_g = [\widehat{y}_j]$, $j \in \mathcal{G} = \{g_1, g_2, \ldots, g_{n_g}\}$

1: Create snapshot matrix : $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_{n_s}] \in \mathbb{R}^{n\times n_s}$and let $r = rank(\mathbf{Y})$.

2: Compute POD basis $\mathbf{V}$ of rank $k \leq r$ for $\mathbf{Y}$ from Algorithm 1.

3: Find coefficient vector $\mathbf{a}$

3.1 Find indices $\vec{\wp} = [\wp_1, \ldots, \wp_m]^T \in \mathbb{R}^m$ and $\mathbf{P}$ from Algorithm 3 by using input from either
(i) $\bar{\mathbf{V}}_c$ defined in (3.1), or
(ii) Left singular vectors of $\bar{\mathbf{V}}_c$ from Algorithm 1.

3.2 Solve $\mathbf{a}$ from (3.2): $\min_{\mathbf{a}\in\mathbb{R}^k} \|\mathbf{P}^T\widehat{\mathbf{y}}_c - \mathbf{P}^T\bar{\mathbf{V}}_c\mathbf{a}\|_2^2$..

4: Compute the approximation $\widehat{\mathbf{y}}_g \approx \mathbf{V}_g\mathbf{a}$.

Remarks: In practice, the additional computation for DEIM indices can be done in advance and reused for many incomplete snapshots. Once the DEIM indices are found from Algorithm 3, the term $\mathbf{P}^T\widehat{\mathbf{y}}_c$ and $\mathbf{P}^T\bar{\mathbf{V}}_c$ can be computed without actually performing matrix multiplication, since this can be done through selected row indices.

The next section applies the techniques introduced in Algorithm 4 on the miscible flow data and compares with the standard Gappy POD approach in Algorithm 2.

 ISBN: 978-84-617-8694-7

4 Numerical Results

This section considers two numerical tests that compare the proposed algorithm with the standard Gappy POD method when applied to miscible flow data with spatial dimension $n = 15000$. The first test considers the effects of POD dimension k and DEIM dimension m on accuracy and computation time of the approximations. The other numerical test considers different amount of missing components in the data. An example of the incomplete snapshot for the concentration flow is shown in the first plot of Figure 1.

4.1 Numerical Test 1

This numerical test investigates different dimensions of POD and DEIM. Using the same notations defined in the previous sections, this numerical experiment uses $n_s = 200$ complete snapshots and tests the algorithms on 50 incomplete snapshots. Each of these incomplete snapshots consists of 50% unknown components (missing or gappy data). I.e. there are 7500 known and unknown components, since the total dimension is $n = 15000$. This section compares the accelerated POD-LS approaches in Algorithm 4 both cases (i) and (ii) with the standard least-squares (LS) or Gappy POD.

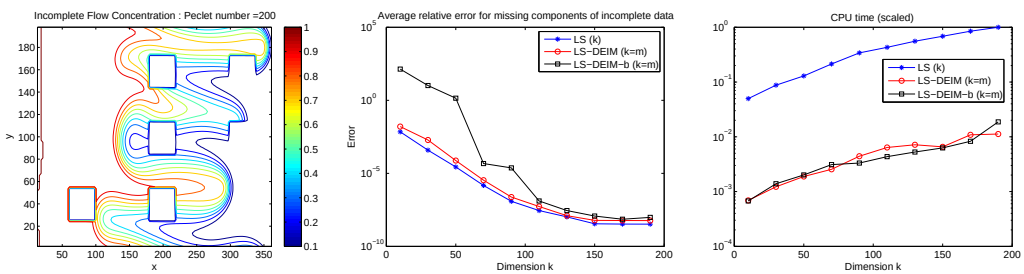


Figure 1: Comparison of average relative errors and CPU time for LS (Gappy POD)Algorithm 2, LS-DEIM: Algorithm 4(i), and LS-DEIM-b Algorithm 4(ii), when using different POD dimension (k) and DEIM dimension (m) with k =m.

The middle plot in Figure 1 shows that the approximations from the proposed accelerated POD-LS method Algorithm 4 (i) are very accurate when compared to the Gappy POD approach, for different POD dimension k and DEIM dimension m with $k = m$. However, the results from Algorithm 4(ii) are less accurate for $k, m < 100$. The CPU times are equivalent for Algorithm 4 in both cases (i) and (ii). However, these two cases use much less computational time than the standard LS, i.e. roughly 100 times less CPU time.

 ISBN: 978-84-617-8694-7

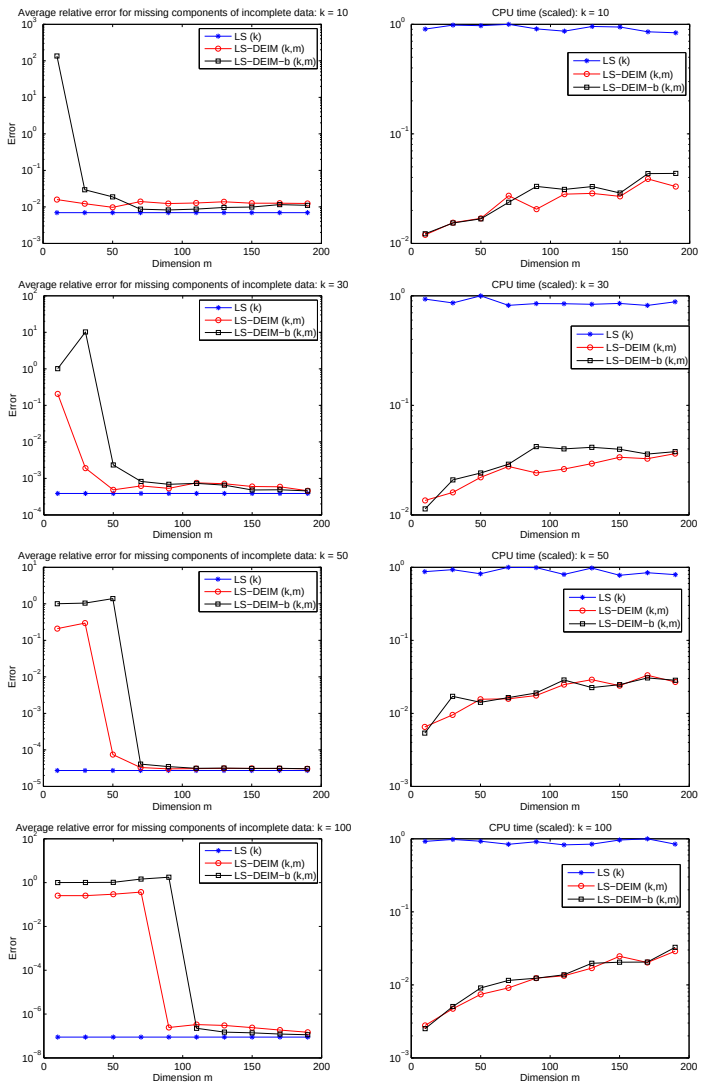


Figure 2: Comparison of average relative errors and CPU time (scaled) for LS (Gappy POD)Algorithm 2, LS-DEIM: Algorithm 4(i), and LS-DEIM-b Algorithm 4(ii), when different DEIM dimensions are used for a fixed dimension of POD (k = 10, 30, 50, 100).

Figure 2 considers different DEIM dimensions for each of the 4 fixed dimensions of POD $k = 10, 30, 50, 100$. The reduction in the CPU time is similar to the results given in Figure 1. From the 4 error plots in Figure 2, as expected, the accuracy is shown to be increased as the POD dimension k gets larger. For a fixed dimension k, increasing the DEIM dimension m for Algorithm 4 in both cases (i) and (ii) may not always improve the accuracy, which is

shown through the flat portion of the error plots. The accuracy depends on both dimensions k and m that are used together. In addition, there might not be an obvious convergence trend, e.g. for $k = 50$ and $k = 100$ there are jumps of the errors at around $m = 60$ and $m = 100$, respectively. This suggests that the appropriate dimension of m is roughly equal to k.

4.2 Numerical Test 2

This numerical test considers the effect of having different amount of missing data in the incomplete snapshots. This numerical experiment uses $n_s = 200$ complete snapshots and tests the algorithms on 50 incomplete snapshots. Each of these incomplete snapshots may contain 10% to 90% unknown components (missing or gappy data) of the total dimension is $n = 15000$. This section compares the accelerated POD-LS approaches in Algorithm 4 (i) with the standard least-squares (LS) or Gappy POD. The results are shown in Figure 3 when $m = k$ for $k = 20, 30$. Notice that, when the amount of missing data is less than 70%, there is no significant different in the accuracy. When 90% of data components are missing, the approximation becomes inaccurate.

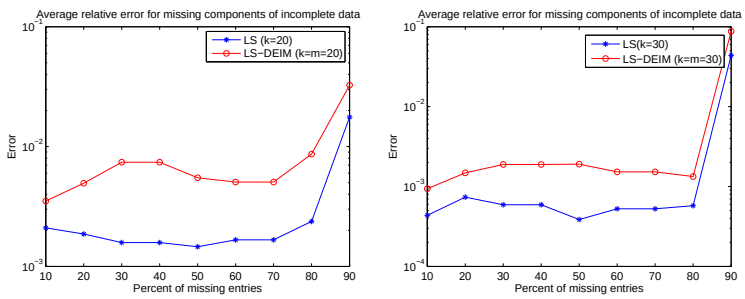


Figure 3: Comparison of average relative errors for LS (Gappy POD)Algorithm 2 and LS-DEIM: Algorithm 4(i) when different percentages of unknown (missing) components when $m = k$ for $k = 20, 30$.

5 Conclusion

This work has presented an approach, called accelerated POD-LS method, for decreasing computational time for reconstructing missing components of incomplete samples. This approach is based on the existing Gappy-POD method, which uses POD basis with least-squares approximation. The proposed method can reduce the computational complexity by using a greedy algorithm from DEIM index selection process to choose only crucial available components used in the approximation. The numerical results demonstrate the accuracy and the efficiency of the proposed method through the reconstruction of missing components

in the concentration flow data. The CPU time is shown to be around $\mathcal{O}(10^{-1})$ to $\mathcal{O}(10^{-2})$ times reduction while the accuracy is of the same order as the standard POD-LS method. Hence, the proposed approach has also shown the potential to reduce approximation time for high-dimensional data. Theoretical analysis of this approach can be considered in the future to provide a rigorous error bound for the approximation of missing data.

Acknowledgements

This work has been partially supported by Thammasat University.

References

[1] D. Xiao A, F. Fang A, A. G. Buchan A, C. C. Pain A, I. M. Navon C, J. Du D, and G. Hu B. Non-linear model reduction for the navier-stokes equations using the residual deim method.

[2] M. Barrault, Y. Maday, N. C. Nguyen, and A. T. Patera. An 'Empirical Interpolation' Method: Application to Efficient Reduced-Basis Discretization Of Partial Differential Equations. *Comptes Rendus Mathematique*, 339(9):667–672, 2004.

[3] Maxime Barrault, Yvon Maday, Ngoc Cuong Nguyen, and Anthony T. Patera. An empirical interpolation method: application to efficient reduced-basis discretization of partial differential equations. *Comptes Rendus Mathematique*, 339(9):667 – 672, 2004.

[4] Gal Berkooz, Philip Holmes, and John L. Lumley. The proper orthogonal decomposition in the analysis of turbulent flows. *Annual Rev. Fluid Mech*, pages 539–575, 1993.

[5] Katarzyna Bizon, Gaetano Continillo, Simona S. Merola, and Bianca M. Vaglieco. Reconstruction of flame kinematics and analysis of cycle variation in a spark ignition engine by means of proper orthogonal decomposition. *Computer Aided Chemical Engineering*, 26:1039 – 1043, 2009.

[6] Elkhadim Bouhoubeiny and Philippe Druault. Note on the pod-based time interpolation from successive piv images. *Comptes Rendus Mcanique*, 337(11):776 – 780, 2009.

[7] T. Bui-Thanh, M. Damodaran, and K. Willcox. Aerodynamic data reconstruction and inverse design using proper orthogonal decomposition. *AIAA*, 42(8):1505 – 1516, 2004.

[8] S. Chaturantabut and D. Sorensen. A state space error estimate for pod-deim nonlinear model reduction. *SIAM Journal on Numerical Analysis*, 50(1):46–63, 2012.

 ISBN: 978-84-617-8694-7

[9] S. Chaturantabut and D.C. Sorensen. Discrete empirical interpolation for nonlinear model reduction. *SIAM J. Sci. Comput*, 32(5):2737–2764, 2010.

[10] Saifon Chaturantabut and Danny C. Sorensen. Nonlinear model reduction via discrete empirical interpolation. *SIAM J. Sci. Comput.*, 32(5):2737–2764, September 2010.

[11] Ouk Choi and Min Chul Lee. Investigation into the combustion instability of synthetic natural gases using high speed flame images and their proper orthogonal decomposition. *International Journal of Hydrogen Energy*, 41(45):20731 – 20743, 2016.

[12] R. Ştefănescu and I. M. Navon. POD/DEIM nonlinear model order reduction of an ADI implicit shallow water equations model. *Journal of Computational Physics*, 237:95–114, March 2013.

[13] Zhengkun Feng and Azzeddine Soulaimani. Reduced order modelling based on pod method for 3d nonlinear aeroelasticity. In *The 18th IASTED International Conference on Modelling and Simulation*, MS '07, pages 489–494, Anaheim, CA, USA, 2007. ACTA Press.

[14] Roi Gurka, Alexander Liberzon, and Gad Hetsroni. {POD} of vorticity fields: A method for spatial characterization of coherent structures. *International Journal of Heat and Fluid Flow*, 27(3):416 – 423, 2006.

[15] Xian hang Sun and Ming hai Xu. Optimal control of water flooding reservoir using proper orthogonal decomposition. *Journal of Computational and Applied Mathematics*, 320:120 – 137, 2017.

[16] Amit Hochman, Bradley N. Bond, and Jacob K. White. A stabilized discrete empirical interpolation method for model reduction of electrical, thermal, and microelectromechanical systems. In *Proceedings of the 48th Design Automation Conference*, DAC '11, pages 540–545, New York, NY, USA, 2011. ACM.

[17] Anthony R. Kellems, Saifon Chaturantabut, and Steven J. Cox. Morphologically accurate reduced order modeling of spiking neurons. *Journal of Computational Neuroscience*, 28:477–494, 2010.

[18] F Lanata and A Del Grosso. Damage detection and localization for continuous static monitoring of structures using a proper orthogonal decomposition of signals. *Smart Materials and Structures*, 15(6):1811, 2006.

[19] J. Lei, J.H. Qiu, and S. Liu. Dynamic reconstruction algorithm for electrical capacitance tomography based on the proper orthogonal decomposition. *Applied Mathematical Modelling*, 39(22):6925 – 6940, 2015.

 ISBN: 978-84-617-8694-7

[20] J. L. Lumley. Stochastic Tools in Turbulence. *Academic Press, New York*, 1970.

[21] J.L. Lumley. The structure of inhomogeneous turbulent flows. *in Atmospheric Turbulence and Radio Wave Propagation (A. M. Yaglom and V. I. Tararsky, eds.), (Nauka, Moscow)*, 1967.

[22] Ana I. Moreno, Artur A. Jarzabek, Jos M. Perales, and Jos M. Vega. Aerodynamic database reconstruction via gappy high order singular value decomposition. *Aerospace Science and Technology*, 52:115 – 128, 2016.

[23] K. Kunisch S. Volkwein, M. Kahlbacher and F. Troltzsch. *Proper Orthogonal Decomposition: Applications in Optimization and Control*.

[24] Elisa Schenone. *Reduced Order Models, Forward and Inverse Problems in Cardiac Electrophysiology*. Theses, Université Pierre et Marie Curie - Paris VI, November 2014.

[25] L Sirovich. Turbulence and the dynamics of coherent structures. i. coherent structures. *Quart. Appl. Math.*, 45(3):561–571, 1987.

[26] Razvan Stefanescu, Adrian Sandu, and Ionel Michael Navon. POD/DEIM strategies for reduced data assimilation systems. *CoRR*, abs/1402.5992, 2014.

[27] Mengyu Wang, Debaditya Dutta, Kang Kim, and John C. Brigham. A computationally efficient approach for inverse material characterization combining gappy {POD} with direct inversion. *Computer Methods in Applied Mechanics and Engineering*, 286:373 – 393, 2015.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

Memory and Dynamics for a family of King-type iterative methods

Francisco I. Chicharro[1], Alicia Cordero[1] and Juan R. Torregrosa[1]

[1] *Instituto de Matemáticas Multidisciplinar, Universitat Politècnica de València*

emails: frachilo@upvnet.upv.es, acordero@mat.upv.es, jrtorre@mat.upv.es

Abstract

A bi-parametric family of derivative-free optimal iterative methods of order four, for solving nonlinear equations, is presented. From its error equation, different iterative schemes with memory can be designed increasing the order of convergence up to six. This family applied on quadratic polynomials gives us a rational operator whose dynamics is studied. The stability of its fixed points, in terms of the values of the parameters, its critical points and their associated parameter planes, etc. give us important information about which members of the family have good properties of stability and whether in any of them appear chaos in the iterative process.

Key words: Iterative method, stability, parameter plane, dynamical plane, chaos

1 Introduction

Solving nonlinear equations $f(x) = 0$, $f : I \subseteq \mathbb{R} \to \mathbb{R}$, or nonlinear systems $F(x) = 0$, $F : D \subseteq \mathbb{R}^n \to \mathbb{R}^n$, are important problems with interesting applications in many fields of Science and Engineering. Analytical methods for solving such problems are hardly available and so, it is only possible to obtain approximate solutions by applying numerical methods based on iterative algorithms.

Due to the limitations of one-point methods (Chebyshev, Halley, etc.), the multi-point schemes appear in the literature and have and spectacular development in the last years. Multi-point iterative methods are defined as methods that require evaluation of f and its derivatives at a number of values of the independent variable. These methods are divided into two classes: without and with memory. The methods of the first class have as iterative expression

$$x_{k+1} = \Phi(x_k), \quad k = 0, 1, 2, \ldots$$

 ISBN: 978-84-617-8694-7

whilst for the second one

$$x_{k+1} = \Phi(x_k, x_{k-1}, \ldots), \quad k = 0, 1, 2, \ldots$$

being $\Phi(x)$ the fixed point function.

The main motivation in the construction of new methods is to achieve the highest computational efficiency; that is, it is desirable to attain as high as possible convergence order with a fixed number of function evaluations per iteration. In this context, Kung and Traub presented in [3] a conjecture that says: the order of an iterative method without memory, which needs d functional evaluations per iteration, is at most 2^{d-1}. When the order reaches this bound, the method is called optimal.

The basic idea for the construction of multi-point methods with memory was introduced by Traub [6], who presented a version with memory from the Steffensen's method. Recently, based on this method, some schemes with memory have been developed by several authors. We can see an interesting overview in [5].

In this work, our starting point is the King's family of fourth-order schemes

$$\begin{array}{rcl} y_k & = & x_k - \frac{f(x_k)}{f'(x_k)}, \\ x_{k+1} & = & y_k - \frac{f(x_k)+\alpha f(y_k)}{f(x_k)+(\alpha-2)f(y_k)} \frac{f(y_k)}{f'(x_k)}, \end{array}$$

where α is a disposable parameter. From this class, we design a derivative-free new family of fourth-order methods, whose error equation allows us to introduce memory in the iterative expression that increase the order of convergence up to two units. The parameters of the family give us the possibility to analyze the stability of the different members in terms of the values of these parameters. By using tools of complex dynamics we analyze the stability of the fixed points of the rational operator that appears when our family is applied on a second degree polynomial. The parameter plane associated to each critical point gives us important information about the stability of the elements of the family and which of them have unstable behavior.

2 Parametric families of iterative schemes

Following the structure of the King's family and replacing the derivative by a first order divided difference we present the following schemes:

$$\begin{array}{rcl} y_k & = & x_k - \frac{f(x_k)}{f[x_k, w_k])}, \\ x_{k+1} & = & y_k - \frac{f(x_k)+\alpha f(y_k)}{\mu f(x_k)+\beta f(y_k)} \frac{f(y_k)}{f[y_k, w_k]}, \end{array} \tag{1}$$

where $w_k = x_k + \gamma f(x_k)$, with α, β, γ and μ real parameters, $\gamma \neq 0$.

The order of convergence of the methods (1) is established in the following result. Its proof only requires the development in Taylor series of the elements of the iterative expression and some algebraic manipulations.

 ISBN: 978-84-617-8694-7

Theorem 1 *Let us suppose that $f : I \subseteq \mathbb{R} \to \mathbb{R}$ is a sufficiently differentiable function in an open interval I and $x^* \in I$ is a simple root of $f(x) = 0$. If the initial approximation x_0 is close enough of x^*, then the iterative schemes (1) have optimal fourth-order convergence when $\mu = 1$, $\beta = \alpha - 1$ and for all nonzero γ, being in this case the error equation*

$$e_{k+1} = (1 + \gamma f'(x^*))^2 c_2 (2 + \alpha + \alpha\gamma f'(x^*) c_2^2 - c_3) e_k^4 + O(e_k^5), \tag{2}$$

where $e_k = x_k - x^*$, $k = 0, 1, \ldots$ and $c_j = \frac{1}{j!}\frac{f^{(k)}(x^*)}{f'(x^*)}$, $k \geq 2$.

Let us observe that the first factor in the second step of (1) can be considered as a particular case of a weight function $H(t)$, where $t = f(y)/f(x)$. So, we consider the family

$$\begin{array}{rcl} y_k & = & x_k - \frac{f(x_k)}{f[x_k, w_k])}, \\ x_{k+1} & = & y_k - H(t_k)\frac{f(y_k)}{f[y_k, w_k]}, \end{array} \tag{3}$$

where $H(t)$ is a weight function of variable $t = f(y)/f(x)$.

Theorem 1 can be generalized in the following way:

Theorem 2 *Let us suppose that $f : I \subseteq \mathbb{R} \to \mathbb{R}$ is a sufficiently differentiable function in an open interval I and $x^* \in I$ is a simple root of $f(x) = 0$. If the initial approximation x_0 is close enough of x^* and function $H(t)$ satisfies the conditions $H(0) = H'(0) = 1$ and $H''(0) < \infty$, then the iterative schemes (3) have optimal fourth-order convergence, for all nonzero γ, being in this case the error equation*

$$e_{k+1} = (-1/2)(1 + \gamma f'(x^*))^2 c_2 \left((-6 + \gamma f'(x^*)(-2 + H''(0)) + H''(0))c_2^2 + 2c_3\right) e_k^4 + O(e_k^5), \tag{4}$$

where $e_k = x_k - x^*$, $k = 0, 1, \ldots$ and $c_j = \frac{1}{j!}\frac{f^{(k)}(x^*)}{f'(x^*)}$, $k \geq 2$.

Let us observe that family (3) supports the Kung-Traub conjecture, having optimal efficiency index $I = 4^{1/3} \approx 1.587$. On the other hand, by observing the expression of the error equation (4) we can choose different values of the free disposable parameters in order to obtain iterative methods with memory, increasing the order of convergence.

Although we can work with both expressions (1) and (3), we are going to introduce memory in the first of them.

3 Iterative methods with memory

We are going to design derivative-free schemes with memory based on the proposed methods of the family (1).

From equation (2) we can assure that the order of convergence of family (1) increase up to six if $\gamma = \frac{-1}{f'(x^*)}$, but the value of $f'(x^*)$ is not available in practice and such acceleration

 ISBN: 978-84-617-8694-7

is not possible. However, we can use an approximation $\bar{f}'(x^*) \approx f'(x^*)$, calculated by using known information. Therefore, by setting $\gamma = -1/\bar{f}'(x^*)$ we can increase the convergence order without using new functional evaluations. The main idea in constructing methods with memory consists of the calculation of the parameter $\gamma = \gamma_k$ as the iteration proceeds by $\gamma_k = -1/\bar{f}'(x^*)$, $k = 1, 2, \ldots$. We are going to consider different approximations of $f'(x^*)$.

(1) Let $N_1(t) = N_1(t, x_k, x_{k-1})$ be the Newton's interpolation polynomial of first degree through two available approximations x_k, x_{k-1}, that is $N_1(t) = f(x_k) + f[x_k, x_{k-1}](t - x_k)$, so

$$\gamma_k = \frac{-1}{N_1'(x_k)} = \frac{-1}{f[x_k, x_{k-1}]} = -\frac{x_k - x_{k-1}}{f(x_k) - f(x_{k-1})},$$

and the algorithm denoted by MM1(α) can be presented in the following way:

- x_0, γ_0 are given,
- $w_k = x_k + \gamma_k f(x_k)$, $k = 0, 1, 2, \ldots$
- $y_k = x_k - \frac{f(x_k)}{f[x_k, w_k]}$,
- $x_{k+1} = y_k - \frac{f(x_k)+\alpha f(y_k)}{f(x_k)+(\alpha-1)f(y_k)} \frac{f(x_k)}{f[y_k, w_k]}$,

where $\gamma_k = \frac{-1}{f[x_k, x_{k-1}]}$ and α is a free parameter.

By using Taylor expansions we obtain the following error expression

$$e_{k+1} = (2c_2^5 - c_2^3 c_3) e_{k-1}^2 e_k^4 + O_7(e_k e_{k-1}),$$

where $O_7(e_k e_{k-1})$ indicates that the sum of the exponents of e_{k-1} and e_k in the rejected terms is at least 7. By applying Theorem 9.2.9 of [4], we establish the following result

Theorem 3 *Let x^* be a simple zero of a sufficiently differentiable function $f : I \subseteq \mathbb{R} \to \mathbb{R}$ in an open interval I. If x_0 is close enough to x^* and γ_0 is given, then the R-order of family MM1(α) is at least $2 + \sqrt{6} \approx 4.45$ that corresponds to the positive root of polynomial $p^2 - 4p - 2$.*

(2) Let $N_2(t) = N_2(t, x_k, x_{k-1}, y_{k-1})$ (can be also used $N_2(t) = N_2(t, x_k, x_{k-1}, u_{k-1})$) be the Newton's interpolation polynomial of second degree, that is $N_2(t) = f(x_k) + f[x_k, x_{k-1}](t - x_k) + f[x_k, x_{k-1}, y_{k-1}](t - x_k)(t - x_{k-1})$, therefore

$$\gamma_k = \frac{-1}{N_2'(x_k)} = \frac{-1}{f[x_k, x_{k-1}] + f[x_k, x_{k-1}, y_{k-1}](x_k - x_{k-1})},$$

and a similar algorithm to the previous one, denoted by MM2(α), can be presented.

 ISBN: 978-84-617-8694-7

(3) Let $N_3(t) = N_3(t, x_k, x_{k-1}, y_{k-1}, u_{k-1})$ be the Newton's interpolation polynomial of third degree, $N_3(t) = f(x_k) + f[x_k, x_{k-1}](t - x_k) + f[x_k, x_{k-1}, y_{k-1}](t - x_k)(t - x_{k-1}) + f[x_k, x_{k-1}, y_{k-1}, u_{k-1}](t - x_k)(t - x_{k-1})(t - y_{k-1})$, then

$$\begin{aligned} N_3'(x_k) = & \quad f[x_k, x_{k-1}] + f[x_k, x_{k-1}, y_{k-1}](x_k - x_{k-1}) \\ + & \quad f[x_k, x_{k-1}, y_{k-1}, u_{k-1}](x_k - x_{k-1})(x_k - y_{k-1}) \end{aligned}$$

and

$$\gamma_k = \frac{-1}{N_3'(x_k)}.$$

The algorithm denoted by MM3(α) can be presented in the following way:

- x_0, γ_0 are given,
- $w_k = x_k + \gamma_k f(x_k)$, $k = 0, 1, 2, \ldots$
- $y_k = x_k - \frac{f(x_k)}{f[x_k, w_k]}$,
- $x_{k+1} = y_k - \frac{f(x_k)+\alpha f(y_k)}{f(x_k)+(\alpha-1)f(y_k)} \frac{f(x_k)}{f[y_k, w_k]}$,

where $\gamma_k = \frac{-1}{N_3'(x_k)}$ and α is a free parameter.

Theorem 4 *Let x^* be a simple zero of a sufficiently differentiable function $f : I \subseteq \mathbb{R} \to \mathbb{R}$ in an open interval I. If x_0 is close enough to x^* and γ_0 is given, then the R-order of family MM3 is at least* 6.

Many other approximations of $f'(x^*)$ are possible, but they either are of sixth order (with more computational cost) or lower than six.

Let us observe that family (1) has two free parameters γ and α. In the next section, we analyze the stability of the elements of the family taking into account the values of these parameters.

4 Dynamical analysis

In order to study the dynamical behavior of the proposed schemes, the recall of some fundamentals is mandatory. Further explanations can be found in [1, 2].

Let $M : \mathbb{R} \to \mathbb{R}$ be a rational function. The orbit of a point $x_0 \in \mathbb{R}$ is defined as the set $\{x_0, M(x_0), M^2(x_0), \ldots, M^n(x_0), \ldots\}$. A point x_0 is a fixed point, x_0^F, of M if $M(x_0^F) = x_0^F$. The multiplier $|M'(x_0^F)|$ classifies the fixed points in attracting, repelling, or neutral if its value is lower than, greater than, or equal to 1, respectively. x_0^F is called superattracting when $M'(x_0^F) = 0$.

 ISBN: 978-84-617-8694-7

The basin of attraction of an attracting fixed point x^*, $\mathcal{A}(x^*)$, is defined as the set of pre-images of any order such that

$$\mathcal{A}(x^*) = \{x_0 \in \mathbb{R} : M^n(x_0) \to x^*, n \to \infty\}. \tag{5}$$

The Fatou set, $\mathcal{F}(M)$, includes the points whose orbits tend to an attracting point x^*. The Julia set, $\mathcal{J}(M)$, is its complementary. It covers the repelling points and sets the borders between the basins of attraction.

The fixed point operator of the biparametric family (1) is

$$M_f(x) = y - \frac{f(x) + \alpha f(y)}{f(x) + (\alpha - 1)f(y)} \frac{f(y)}{f[y, w]}, \tag{6}$$

where $w = x + \gamma f(x)$ and $y = x - \frac{f(x)}{f[x,w]}$.

We are going to analyze the dynamical behavior of the rational function obtained when family (1) is applied on $f(x) = x^2 - 1$. In this case, expression (6) is $M_f(x) = \frac{N_{11}(x)}{D_{10}(x)}$, where $N_{11}(x)$ and $D_{10}(x)$ are polynomials of degrees 11 and 10, respectively, depending on x, α and γ.

There are 10 fixed points of $M_f(x)$. $x^F_{1,2} = x^*_{1,2} = \pm 1$ are superattracting points, while $x^F_{3-10}(\alpha, \gamma)$ are the roots of an 8th-degree polynomial: $x^F_{3,4} \in \mathbb{R}$, $x^F_{5-10} \in \mathbb{C}$. Since our purpose is the study of the real dynamics, x^F_{5-10} are rejected. The evaluation of $|M'_f(x^F_{3,4})|$ establishes the behavior of these points. For different values of α and γ, $x^F_{3,4}$ have different dynamical features, as the stability plane of Figure 1 represents. A mesh of 100×100 points covers the values of $\alpha \in [-2, 2], \gamma \in [-2, 2]$. The white area represents where the multiplier is lower than 1, while the corresponding black represents where the multiplier is greater than 1. Its dynamical meaning is immediate, since white and black regions represent attracting and repelling behavior, respectively.

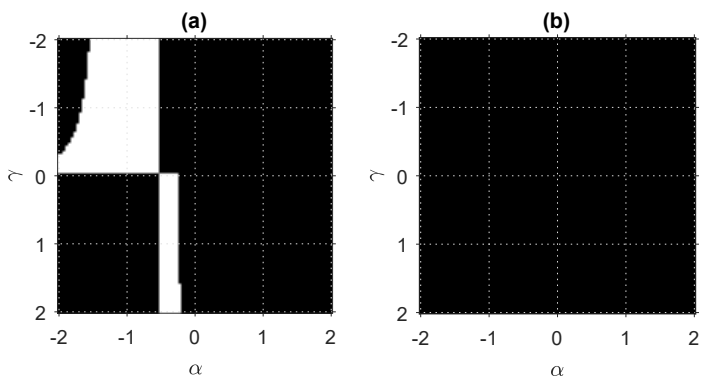


Figure 1: Stability plane of (a) x^F_3 and (b) x^F_4.

 ISBN: 978-84-617-8694-7

Therefore, x_3^F has both behaviors, depending on the value of α and γ. For assuring the convergence of the method to $x_{1,2}^*$, white areas must be avoided. In addition, it is not difficult to prove that the infinity is also an attracting fixed point.

Computing $|M_f'(x)| = 0$, 16 critical points can be found. As expected, $x_{1,2}^C = x_{1,2}^*$. $x_{3,4}^C = \frac{-1\pm\gamma}{\gamma}$ are pre-images of $x_{1,2}^*$. $x_{5-10}^C \in \mathbb{R}$ depend on the values of α and γ. Finally, $x_{11-16}^C \in \mathbb{C}$ are left out of the study since we are only interested in real dynamics.

A classical result (due to Fatou and Julia) establishes that there is at least one critical point associated with each invariant Fatou component, so it is interesting to analyze the behavior of each free critical point (critical point different to the roots of the polynomial) used as initial estimation for all the elements of the family.

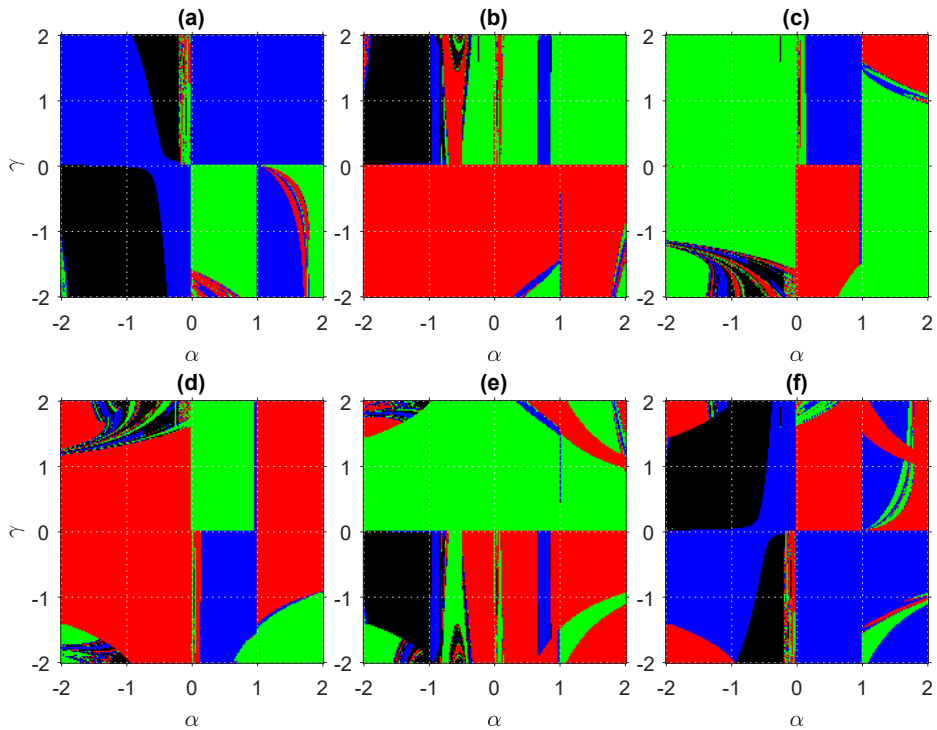


Figure 2: Parameter planes of (a) x_5^C, (b) x_6^C, (c) x_7^C, (d) x_8^C, (e) x_9^C, and (f) x_{10}^C (red: x_1^*, green: x_2^*, blue: x_∞^*, black: other).

The parameter plane represents the family of methods $M_f(x)$, where each point (α, γ) stands for an individual method. If the orbit of the free critical point tends to $-1, 1$, or ∞, the point is colored in red, green, or blue, respectively. If x^C does not converge to any of those points, the point is colored in black. Figure 2 shows the parameter planes of x_{5-10}^C.

The unified parameter plane represented in Figure 3 gathers in one image the main

information of Figure 2. It is composed by the superposition of the black regions. In this way, the election of a point in the white region guarantees that the corresponding scheme tends to ± 1 or ∞, while the set of methods in the black region do not.

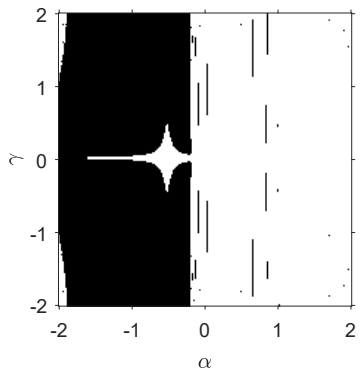


Figure 3: Unified parameter plane of x_{5-10}^{C}.

In order to visualize the behavior of each method, the dynamical line is introduced. Analogously to the dynamical plane for complex variable, the dynamical line represents the basins of attraction, plotting in different colors where the orbit of each initial estimation tends. In coherence to Figure 2, the red color is assigned to those initial approximations that tends to $x_1^* = -1$. Analogously, the green color to $x_2^* = 1$, the blue color to $x^* = \infty$ and the black one to those initial guess that tends to a point different than the previous three. The final orbit of 1001 initial values of $x \in [-2, 2]$ has been calculated, and the convergence to a point is set when its distance is lower than 10^{-3}.

Figure 4 represents the dynamical line of a set of methods whose orbits tend to $-1, 1$, or ∞, i.e., methods that belong to the white area of the unified parameter plane. On the contrary, Figure 5 shows methods whose points also tend to another fixed point.

Note that every point in Figure 4 tends to one of the attracting points. However, when a value of (α, γ) of the black region of Figure 3 is chosen, some points tend to a different point. For instance, these points can be found in Figures 5(a-d) for $x_0 = -0.2$, $x_0 = 0.2$, $x_0 = -0.28$ and $x_0 = 0.28$, respectively.

As expected from the attracting area of x_3^F and the unified parameter plane, the election of methods with $\alpha > 0$ is capital, as proved in Figures 4 and 5.

Acknowledgements

This research was partially supported by Ministerio de Economía y Competitividad under grants MTM2014-52016-C2-2-P and Generalitat Valenciana PROMETEO/2016/089.

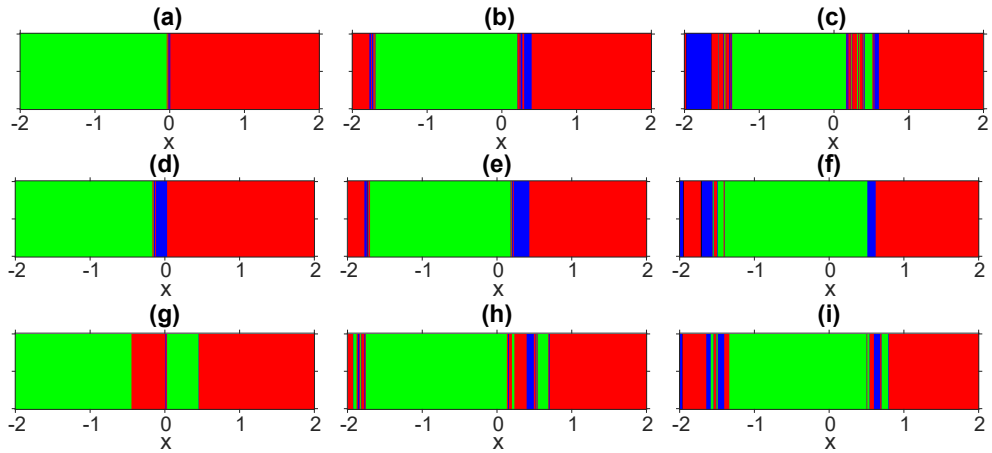


Figure 4: Dynamical lines of methods with (a) $(\alpha,\gamma) = (0.5, 0.05)$, (b) $(\alpha,\gamma) = (0.5, 1)$, (c) $(\alpha,\gamma) = (0.5, 2)$, (d) $(\alpha,\gamma) = (1, 0.05)$, (e) $(\alpha,\gamma) = (1, 1)$, (f) $(\alpha,\gamma) = (1, 2)$, (g) $(\alpha,\gamma) = (2, 0.05)$, (h) $(\alpha,\gamma) = (2, 1)$, (i) $(\alpha,\gamma) = (2, 2)$ (red: x_1^*, green: x_2^*, blue: x_∞^*).

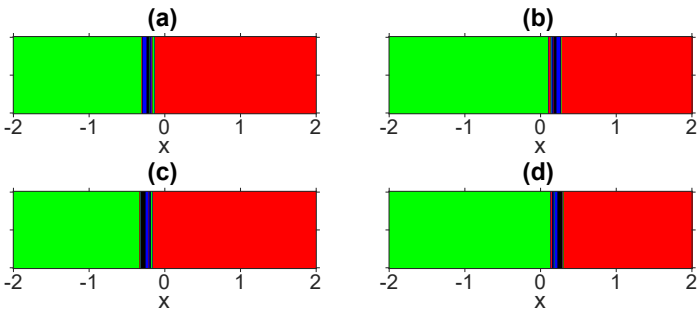


Figure 5: Dynamical lines of methods with (a) $(\alpha,\gamma) = (-0.5, -0.5)$, (b) $(\alpha,\gamma) = (-0.5, 0.5)$, (c) $(\alpha,\gamma) = (-1, -0.5)$, (d) $(\alpha,\gamma) = (-1, 0.5)$ (red: x_1^*, green: x_2^*, blue: x_∞^*, black: other).

References

[1] P.Blanchard, *Complex analytic dynamics on the Riemann sphere*, Bull. Amer. Math. Soc. **11** (1984) 85–141.

[2] R. L. Devaney, *An Introduction to Chaotic Dynamical Systems*, Addison-Wesley Publishing Company, 1989.

[3] H.T. Kung, J.F. Traub, *Optimal order of one-point and multi-point iteration*, J. Assoc. Comput. Math. **21** (1974) 643–651.

[4] J.M. Ortega, W.C. Rheinboldt, *Iterative solutions of nonlinear equations in several variables*, Academic Press, New York, 1970.

[5] M.S. Petković, B. Neta, L.D. Petković, J. Dzunić, *Multipoint methods for solving nonlinear equations*, Academic Press, Elsevier, 2013.

[6] J.F. Traub, *Iterative methods of the solution of equations*, Prentice-Hall, New York, 1984.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

Special discontinuities in models of continuum mechanics

Anna Chugaynova[1]

[1] *Steklov Mathematical Institute, Russian Academy of Sciences, Moscow, Russia*

emails: anna_ch@mi.ras.ru

Abstract

Solutions of the problems of disintegration of an arbitrary discontinuity of the generalized Hopf equation are under analysis. These solutions are constructed from the sequence of non-tipping Riemann waves and shock waves having the stable stationary or non-stationary structure.

Key words: special discontinuity, Hopf equation, problems of disintegration of an arbitrary discontinuity

1 Introduction

In [1, 2], devoted to investigation of the solutions of the Hopf equations with complex nonlinearity, for selection of discontinuities, which have been used for construction of the solution, the request of existence of the stationary structure of the discontinuity has been posed. The structure of discontinuities has been described by the generalized (in the sense of nonlinearity) Korteweg-de Vries-Burgers equation. Appearance of the resent works [3, 4], in which spectral stability of the solutions describing the structure is investigated, makes it possible to include effectively in the notion of the permissible discontinuity the claim of stability of its structure and from this point of view revise before obtained results. We call admissible (i. e. realizable in practice for disintegration of an arbitrary discontinuity) discontinuities with structure, having stability property.

Introduction of the request of stability of the structure in the notion of admissibility of discontinuities results in cutting down the set of admissible discontinuities, described in [1, 2], and eliminate non-uniqueness of the solution of the problem about disintegration of the arbitrary shock, discovered in previous investigations [1]. Furthermore, for construction

 ISBN: 978-84-617-8694-7

of the solution of the problem we have used the discontinuities with structure, containing the internal periodic oscillations (non-stationary structures). Variation of the quantities in such discontinuities may not coincide with variation of the quantities in any discontinuities with stationary structure. It has been shown [5] that the solution of the problem of disintegration of the arbitrary discontinuity in this setting always uniquely exists.

2 Discontinuities with a stationary structure

RWe consider the generalized KdVB equation

$$\frac{\partial v}{\partial t} + \frac{\partial \varphi(v)}{\partial x} = \mu \frac{\partial^2 v}{\partial x^2} - m \frac{\partial^3 v}{\partial x^3}, \tag{1}$$

$$m, \mu = \text{const}, \qquad v = v(x, t).$$

On the right hand side of the equation (1) the term, containing the coefficient m, describes dispersion effects (the coefficient m is the parameter of dispersion). We assume everywhere throughout the paper that $m > 0$. The term, containing the coefficient μ, takes into account viscous effects and it determines dissipation in the system (the coefficient μ is the parameter of dissipation).

In this paper we use the following potential

$$\varphi(v) = v^4 - v^2. \tag{2}$$

The characteristic property of this potential appears to be the presence of two points where the second derivative changes its sign. We study stability of traveling waves representing discontinuity structures obeying the equation

$$\frac{\partial v}{\partial t} + \frac{\partial \varphi(v)}{\partial x} = 0, \tag{3}$$

which is the formal limit of (1), when we consider processes characterized by a large spatial scale L (then both terms on the right hand side of (1) become small in comparison with the terms on the left hand side).

Equation (3) can be called the generalized Hopf equation; it can be transformed into the known Hopf equation when $\varphi(v)$ is a quadratic function of v.

Equation (3) (as well as (1)) expresses the conservation law, consequently the corresponding relation on the discontinuity can be written in the following form:

$$W = \frac{[\varphi(v)]}{[v]}. \tag{4}$$

 ISBN: 978-84-617-8694-7

Here W is the speed of the discontinuity, and quadratic brackets denote the difference of functions in front of the discontinuity and behind it.

Make the change of variables in (1) for $\mu \neq 0$

$$t \to t\sqrt{m}, \quad x \to x\sqrt{m}, \quad \gamma = \sqrt{m}/\mu, \quad \varphi(v) = f(u-1), \quad v = u - 1. \tag{5}$$

Equation (1) takes the form

$$\frac{\partial u}{\partial t} + \frac{\partial f(u)}{\partial x} + \frac{\partial^3 u}{\partial x^3} = \frac{1}{\gamma}\frac{\partial^2 u}{\partial x^2}. \tag{6}$$

The stationary structure of the discontinuity is described by the following equations:

$$u = u(\xi), \qquad \xi = x - Wt,$$

$$\frac{d^2u}{d\xi^2} - \frac{1}{\gamma}\frac{du}{d\xi} = Wu - f(u), \tag{7}$$

$$\lim_{\xi \to -\infty} u(\xi) = u_l, \quad \lim_{\xi \to +\infty} u(\xi) = u_r.$$

If a solution of (7) exists, then from the point of view of the large scale of the length it has to represent a discontinuity with a stationary structure, where $u = u_r$ and $u = u_l$ are the states in front of the discontinuity ($\xi > 0$) and behind the discontinuity ($\xi < 0$). The states u_l and u_r satisfy the relations (4) and they represent, therefore, the states, in front of a discontinuity and behind it corresponding to the conservation law. We assume that the state in front of a discontinuity is taken in the form $u_r = 0$.

Construction of numerous discontinuities, having the stationary structure is specified, which are caused, besides dispersion and dissipation, also by the complex nonlinearity which is given by the potential (2). When the function $f(u)$ is given in such a way, among the solutions of the equation (1) special discontinuities are possible. A discontinuity is called special when its structure represents a heteroclinic phase curve, connecting in the $(u, du/d\xi)$ -plane two singular points of the saddle type (one of these points corresponds to the state in front of the discontinuity, and the other corresponds to the state behind the discontinuity). The number of special discontinuities increases when γ grows [6]. In this setting, the entire set of discontinuities with stationary structure varies and becomes more complex. This set consists of special and non-special discontinuities. The structure of the non-special discontinuity represents the phase curve, connecting two stationary points: the saddle (corresponds to the state in front of a discontinuity) and the focus or the node (corresponds to the state behind a discontinuity).

 ISBN: 978-84-617-8694-7

3 Linear stability of heteroclinic solutions

For investigation of linear stability of heteroclinic stationary solutions of (6), we look for a solution of the form [7, 8, 9]

$$u(x,t) = U(\xi) + w(\xi,t). \tag{8}$$

The function $w(\xi,t)$ satisfies the linearization of equation (7), i. e.,

$$\frac{\partial w}{\partial t} = \frac{\partial}{\partial \xi}\Big(- \frac{\partial^2 w}{\partial \xi^2} + \frac{1}{\gamma}\frac{\partial w}{\partial \xi} + (W - f'(U(\xi)))w\Big), \quad f'(U) = \frac{df(U)}{dU}. \tag{9}$$

The function $U(\xi)$ satisfies the equation (7) and obeys the conditions

$$\lim_{\xi\to-\infty} U(\xi) = u_l, \quad \lim_{\xi\to+\infty} U(\xi) = 0.$$

D e f i n i t i o n 2. The solution $U(\xi)$ of equation (7) is called spectrally (linearly) unstable if there exists a solution of (9) of the form

$$w(\xi,t) = \mathrm{e}^{\lambda t} Y(\xi) \tag{10}$$

with $\mathrm{Re}\,\lambda > 0$, $Y(\xi) \to 0$ for $\xi \to \pm\infty$.

4 Conclusion and Discussion

Solution of problems which are described by systems of nonlinear hyperbolic equations necessarily involve discontinuities, and their proper treatment can cause considerable difficulties. A problem of this type is considered in the present work. The mentioned difficulties are associated with theoretical selection of the admissible discontinuities, i. e. discontinuities that can exist in real media. The naturally occurring requirement for discontinuous solutions obeying nonlinear hyperbolic equations (here we consider the Hopf equation (3)) is the convention that the discontinuities represent the singular limit of solutions to complicated equations, when terms describing small scale processes are taken into consideration (in concrete physical problems, these terms have to correspond to the real processes). In this paper the complicated problem is represented by the equation (1). Previously, the existence of a stationary discontinuity structure has been used as an admissability requirement in many works (see, for example, [1, 6, 10, 11]). However, solutions of standard self-similar problems constructed with the help of so defined admissible discontinuities turned out to be non-unique [1].

In this work we modified the notion of an admissible discontinuity and considered the solutions of the equations (3), (1) from a new point of view. Specifically, admissible discontinuities were treated as discontinuities with a stable structure, which can be stationary

 ISBN: 978-84-617-8694-7

or periodic in time. Accordingly, discontinuities with stationary unstable structures were excluded from the set of discontinuities regarded previously as admissible, while discontinuities with stable and time-periodic structures were added to this set (their stability and periodicity were verified by direct numerical computation). As a result we showed that the solution of the problem of arbitrary discontinuity disintegration constructed in this work uniquely exists for all parameter values. This was done using an example of the KdVB equation with the potential providing the existence of a rich set of front solutions to the equation. Non-uniqueness in the problem of arbitrary discontinuity disintegration more or less frequently occurs in continuum mechanics and physics. Gas dynamics with non-standard equations of state gives such an example [12]. The other famous example is an elastic media. Wave processes there are described, as a rule, by a hyperbolic system of equations and hence, inevitably discontinuities appear, which, unlike usual gas dynamics, occur to be non unique, giving thereby a fundamental example of non-uniqueness of shock wave in mechanics of continuous media [2]. Therefore, in this field the question about selection of admissible discontinuous solutions (shock waves) is extremely important. In this sense, the criterion of uniqueness we obtained can be used for the unique selection of front solutions and simple waves) which solve the problem of arbitrary discontinuity disintegration in a wide class of media which are described by hyperbolic systems of equations.

The results of our study were achieved in in several steps.

- The form of the discontinuous solutions with a structure dependending on a physical parameter of the problem (namely, the relation between dispersion and dissipation) was described. The types of these solutions and their properties were examined.

- The dynamical (spectral) stability of these solutions was analysed.

- The problem of arbitrary discontinuity disintegration was solved using admissible discontinuities with either a stationary or non-stationary structure in time asymptotics of the initial discontinuity.

Acknowledgements

This work was supported by the Russian Science Foundation under grant No 14-50-00005.

References

[1] A. G. Kulikovskii, A. P. Chugainova, *Modeling the influence of small-scale dispersion processes in a continuum on the formation of large-scale phenomena*, Comput. Math. Math. Phys. **44**:6 (2004) 1062–1068.

 ISBN: 978-84-617-8694-7

[2] A. G. Kulikovskii, A. P. Chugainova, *Classical and non-classical discontinuities in solutions of equations of non-linear elasticity theory*, Russian Math. Surveys **63**:2 (2008) 283–350.

[3] A. T. Il'ichev, A. P. Chugainova, V. A. Shargatov, *Spectral stability of special discontinuities*, Dokl. Math. **91**:3 (2015) 347–351.

[4] A. P. Chugainova, V. A. Shargatov, *Stability of discontinuity structures described by a generalized KdV-Burgers equation*, Comput. Math. Math. Phys. **56**:2 (2016) 263–277.

[5] A. G. Kulikovskii, A. P. Chugainova, V. A. Shargatov, *Uniqueness of self-similar solutions to the Riemann problem for the Hopf equation with complex nonlinearity*, Comput. Math. Math. Phys. **56**:7 (2016) 1355–1362.

[6] A. G. Kulikovskii, *The possible effect of oscillations in a discontinuity structure on the set of admissible discontinuities*, Soviet Phys. Dokl. **29** (1984) 283-285.

[7] R. L. Pego, & M. I. Weinstein, *Eigenvalues, and instabilities of solitary waves*, Phil. Trans. R. Soc. Lond. A **340** (1992) 47–94.

[8] R. L. Pego, P. Smereka & M. I. Weinstein, *Oscillatory instability of traveling waves for a KdV-Burgers equation*, Physica D **67** (1993) 45–65.

[9] A. T. Il'ichev, A. P. Chugainova, *Spectral stability theory of heteroclinic solutions to the Kortewegde VriesBurgers equation with an arbitrary potential*, Proc. Steklov Inst. Math. **295** (2016) 148-157.

[10] I. M. Gelfand, *Some problems in the theory of quasilinear equations*, Transl. Ser.2. Am. Math. Soc **29** (1963) 295–381.

[11] S. K. Godunov, *Non-unique blurrings of discontinuities in solutions of quasi-linear systems*, Soviet Math. Dokl. **2** (1961) 43–44.

[12] G. Ya. Galin, *hock waves in media with arbitrary equations of state*, Soviet Physics Dokl. **119** (1958) 244–247.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

An efficient numerical method for 2D systems of singularly perturbed parabolic reaction-diffusion equations

C. Clavero[1] and J.L. Gracia[1]

[1] *Department of Applied Mathematics and IUMA, University of Zaragoza, Spain*

emails: clavero@unizar.es, jlgracia@unizar.es

Abstract

In this paper we propose an efficient numerical method to solve parabolic initial-boundary two dimensional coupled singularly perturbed systems of reaction-diffusion type. The diffusion parameter, which can be sufficiently small, is the same in all the equations of the system. In general, the exact solution of the problem has parabolic boundary layers at the boundary of the spatial domain. The fully discrete scheme combines a splitting or additive scheme, to discretize in time, and the classical central finite difference scheme, to discretize in space. Then, if the time derivatives are discretized on a uniform mesh and the spatial derivatives on a special piecewise uniform mesh of Shishkin type, the method is uniformly convergent, having first order in time and almost second order in space. Some numerical results are showed, which corroborate in practice the good theoretical properties of the method.

Key words: 2D parabolic systems, reaction-diffusion, additive methods, piecewise uniform meshes, uniform convergence
MSC 2000: 65N05, 65M06, 65N06, 65N12

1 Introduction

In this work we are interested in approximating the solution of two dimensional parabolic singularly perturbed coupled reaction-diffusion systems of ℓ equations, given by

$$\left\{ \begin{array}{l} L_\varepsilon \mathbf{u} \equiv \dfrac{\partial \mathbf{u}}{\partial t}(\mathbf{x},t) + \mathcal{L}_{\mathbf{x},\varepsilon}\mathbf{u}(\mathbf{x},t) = \mathbf{f}(\mathbf{x},t),\ (\mathbf{x},t) \in Q = \Omega \times (0,T], \\ \mathbf{u}(\mathbf{x},t) = \mathbf{0},\ \mathbf{x} \in \partial\Omega,\ t \in (0,T],\ \mathbf{u}(\mathbf{x},0) = \mathbf{0},\ \mathbf{x} \in \overline{\Omega}, \end{array} \right. \tag{1}$$

 ISBN: 978-84-617-8694-7

where $\Omega = (0,1)^2$ and the spatial differential operator $\mathcal{L}_{\mathbf{x},\varepsilon}$ is defined as

$$\mathcal{L}_{\mathbf{x},\varepsilon}\mathbf{u} \equiv -\mathcal{D}\Delta\mathbf{u} + \mathcal{A}\mathbf{u}, \tag{2}$$

with $\mathcal{D} = \mathrm{diag}(\varepsilon,\ldots,\varepsilon)$, $\mathcal{A}(\mathbf{x},t) = (a_{ij}(\mathbf{x},t))$, $i,j = 1,2,\ldots,\ell$. We assume that the diffusion parameter ε, $0 < \varepsilon \leq 1$, can be sufficiently small, that the reaction matrix $\mathcal{A}$ is an M-matrix, i.e., for $(\mathbf{x},t) \in \overline{Q}$ it satisfies

$$\sum_{j=1}^{\ell} a_{ij} \geq \alpha > 0,\ a_{ii} > 0,\ i = 1,2,\ldots,\ell,\ a_{ij} \leq 0, \quad \text{if} \quad i \neq j, \tag{3}$$

and also that the components of the right-hand side of the differential equation, $\mathbf{f}(\mathbf{x},t) = (f_1(\mathbf{x},t), f_2(\mathbf{x},t),\ldots,f_\ell(\mathbf{x},t))^T$, and the reaction matrix $\mathcal{A}$ are sufficiently smooth functions, which satisfy sufficient compatibility conditions, in order to guarantee that the exact solution $\mathbf{u} \in C^{4,2}(\overline{Q})$.

There exists many works in the literature (see for instance [2, 3, 7, 8] and references therein), where 1D singularly perturbed systems of reaction-diffusion type, in both elliptic or parabolic case, are analyzed. In those papers, the time variable is discretized by using the backward Euler method, on a uniform mesh, and the spatial variable is discretized by the classical central finite difference scheme, defined on a piecewise uniform mesh of Shishkin type. Then, the resulting schemes are uniformly convergent in both variables. At each time level of the time discretization, the numerical approximation is obtained by solving a linear system which requires a high computational cost, due to the components of the discrete solution are coupled. In order to reduce this computational cost, it is very convenient to use techniques that decouple the components. This idea is used in [1], where a splitting (or additive) scheme, defined on a uniform mesh, is used to discretize in time. Here, we are interested in extending this idea to the class of problems (1), where the dimension of the spatial variable makes considerably more difficult the resolution of the continuous problem.

In [5, 6, 9] the case of 2D elliptic singularly perturbed systems is analyzed; from those papers, it follows that parabolic boundary layers, of width $\mathcal{O}(\sqrt{\varepsilon})$, appear at the boundary $\partial\Omega$ of the spatial domain. Similarly to the case of parabolic problems with a single equation ($\ell = 1$), parabolic boundary layers are expected in the exact solution of problem (1). Therefore, uniformly convergent methods are necessary to find accurate approximations to the solution for any value of the diffusion parameter, with a number of grid points which is also independent of ε.

The paper is organized as follows. In Section 2 we construct the fully discrete scheme, which combines an additive method to discretize in time and the central finite difference scheme to discretize in space; we also give the result proving the uniform convergence, with respect to the diffusion parameter, of the numerical method. In Section 3, some results obtained for different test problems are showed, which corroborate, from a numerical point of view, the order of uniform convergence of the method.

 ISBN: 978-84-617-8694-7

2 The fully discrete scheme: uniform convergence

The analysis of the asymptotic behavior of the exact solution, following similar ideas and techniques as in [5, 6] for systems of elliptic reaction-diffusion equations, shows that the solution of (1) has parabolic boundary layers at the boundary of the spatial domain. Then, to approximate that solution we need an efficient numerical scheme.

The first step to construct this method is the time discretization. For that, we consider a uniform mesh, $\overline{\omega}^M = \{t_m = m\tau,\ 0 \leq m \leq M,\ \tau = T/M\}$, and the discretization is given by

$$\begin{array}{l} \mathbf{z}^0 = \mathbf{u}(\mathbf{x},0) = \mathbf{0}, \\ \left\{ \begin{array}{l} \text{For } m = 0, 1, \cdots, M-1, \\ \tau^{-1}(\mathbf{z}^{m+1} - \mathbf{z}^m) - \mathcal{D}\Delta \mathbf{z}^{m+1} + \mathcal{M}^{m+1}\mathbf{z}^{m+1} - \mathcal{N}^{m+1}\mathbf{z}^m = \mathbf{f}^{m+1}, \text{ in } \Omega, \\ \mathbf{z}^{m+1} = \mathbf{0}, \quad \text{on } \partial\Omega, \end{array} \right. \end{array} \tag{4}$$

where $\mathbf{f}^{m+1} = \mathbf{f}(\mathbf{x}, t_{m+1})$, $m = 0, 1, \cdots, M-1$, the operator $\mathcal{M}^{m+1}$ is given by

$$\mathcal{M}^{m+1}(\mathbf{x}) = \begin{pmatrix} a_{11}(\mathbf{x}, t_{m+1}) & 0 & \cdots & \cdots & 0 \\ a_{21}(\mathbf{x}, t_{m+1}) & a_{22}(\mathbf{x}, t_{m+1}) & 0 & \cdots & 0 \\ \vdots & \vdots & & \ddots & \vdots \\ a_{m1}(\mathbf{x}, t_{m+1}) & a_{m2}(\mathbf{x}, t_{m+1}) & \cdots & \cdots & a_{mm}(\mathbf{x}, t_{m+1}) \end{pmatrix} \tag{5}$$

and $\mathcal{N}^{m+1}(\mathbf{x}) = \mathcal{M}^{m+1}(\mathbf{x}) - \mathcal{A}^{m+1}(\mathbf{x})$. In this way, at the time level t_{m+1} the components of the vector unknown $\mathbf{z}^{m+1}$ are decoupled.

To obtain the fully discrete method, we discretize (4) with the classical central difference scheme, which is defined on a piecewise uniform mesh, $\overline{\Omega}^N \equiv I_{x,\varepsilon,N} \times I_{y,\varepsilon,N}$, given as a tensor product of one dimensional piecewise uniform Shishkin meshes, $I_{x,\varepsilon,N} = \{0 = x_0 < \ldots < x_N = 1\}$, $I_{y,\varepsilon,N} = \{0 = y_0 < \ldots < y_N = 1\}$, where N, the discretization parameter, is a positive integer. We only give the details of the construction of $I_{x,\varepsilon,N}$ and similarly can be done for $I_{y,\varepsilon,N}$. For simplicity in the presentation, we take the same value N for both spatial variables, but a similar result follows in the case that the number of grid points at each spatial direction is different.

We know that parabolic boundary layers of width $\mathcal{O}(\sqrt{\varepsilon})$ appear in $\partial\Omega$; so, the grid points must condense in the boundary layer regions. In this paper, we choose a piecewise uniform mesh of Shishkin type. Then, in the x-spatial variable, the grid points of the mesh are given by (see [4])

$$x_j = \begin{cases} jh, & j = 0, \ldots, N/4, \\ x_{N/4} + (j - N/4)H, & j = N/4+1, \ldots, 3N/4, \\ x_{3N/4} + (j - 3N/4)h, & j = 3N/4+1, \ldots, N, \end{cases} \tag{6a}$$

where $h = 4\sigma/N,\ H = 2(1-2\sigma)/N$, and the transition parameter σ is defined by

$$\sigma = \min\left\{1/4, \sqrt{\varepsilon}\ln N\right\}. \tag{6b}$$

 ISBN: 978-84-617-8694-7

We denote by $\overline{Q}^{N,M} = \overline{\Omega}^N \times \overline{\omega}^M$ the corresponding grid for the $(\mathbf{x}, t)$-variables, by $Q^{N,M} = \overline{Q}^{N,M} \cap Q$, $\partial\Omega^{N,M} = \overline{Q}^{N,M} \backslash Q^{N,M}$, and by $\mathbf{U} = \{\mathbf{U}^0, \ldots, \mathbf{U}^M\}$ the vector numerical approximation on the grid $\overline{Q}^{N,M}$. Thus, $\mathbf{u}(\mathbf{x}, t_m) \approx \mathbf{U}^m(\mathbf{x})$ with $\mathbf{x} \in \overline{Q}^{N,M}$.

The fully discrete scheme is defined as

$$\begin{array}{l} \mathbf{U}^0 = \mathbf{0}, \\ \left\{ \begin{array}{l} \text{For } m = 0, 1, \cdots, M-1, \\ \tau^{-1}(\mathbf{U}^{m+1} - \mathbf{U}^m) - \mathcal{D}\left(\delta_x^2 + \delta_y^2\right)\mathbf{U}^{m+1} + \mathcal{M}^{m+1}\mathbf{U}^{m+1} - \mathcal{N}^{m+1}\mathbf{U}^m = \mathbf{f}^{m+1}, \text{in } Q^{N,M}, \\ \mathbf{U}^{m+1} = \mathbf{U}^{m+1} = \mathbf{0}, \text{ on } \partial\Omega^{N,M}, \end{array} \right. \end{array} \tag{7}$$

where

$$\delta_x^2 Z_{i,j} = \frac{2}{h_i + h_{i+1}} \left(\frac{Z_{i+1,j} - Z_{i,j}}{h_{i+1}} - \frac{Z_{i,j} - Z_{i-1,j}}{h_i} \right),$$
$$\delta_y^2 Z_{i,j} = \frac{2}{h_j + h_{j+1}} \left(\frac{Z_{i,j+1} - Z_{i,j}}{h_{j+1}} - \frac{Z_{i,j} - Z_{i,j-1}}{h_j} \right),$$

are the standard approximations of the second order derivative, at each spatial variable on a nonuniform mesh, with $h_i = x_i - x_{i-1},\ h_j = y_j - y_{j-1},\ i, j = 1, \ldots, N$.

The proposed numerical scheme decouples the ℓ components of the vector problem and then ℓ discrete problems are solved at each time level. Each discrete problem approximates one of the unknowns using a five-point approximation whose associated $N \times N$ matrix is a tridiagonal block matrix.

The following result proves the uniform convergence of the fully discrete scheme, having first order in time and almost second order in space.

Theorem 1 *Let* $\mathbf{U}$ *be the numerical solution of (7) on* $\overline{Q}^{N,M}$ *using a uniform mesh in time and the piecewise uniform Shishkin mesh* (6) *in space, and let* $\mathbf{u}$ *be the solution of the continuous problem (1). Then, the following error bound is satisfied*

$$\max_{0 \le m \le M} \max_{\mathbf{x} \in \overline{Q}^{N,M}} |\mathbf{U}^m(\mathbf{x}) - \mathbf{u}(\mathbf{x}, t_m)| \le C(M^{-1} + (N^{-1} \ln N)^2), \tag{8}$$

where C *is a positive constant independent of the diffusion parameter* ε *and the discretization parameters* N *and* M.

3 Numerical results

In this section we show the numerical results obtained for some test problem of type (1). The first example is defined when the reaction matrix and the right-hand side are

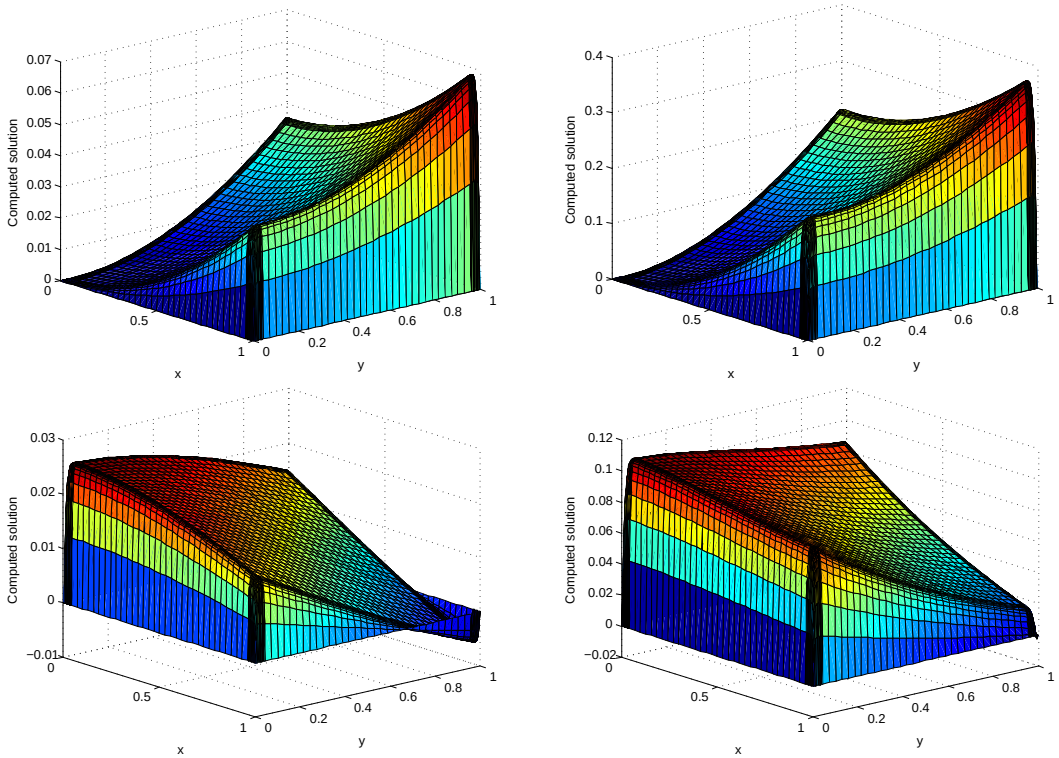
$$\mathcal{A} = \begin{pmatrix} 1 + t(x+y) & -t\sin(x+y) \\ t(\cos(xy) - 1) & (t+1)(1+xy) \end{pmatrix}, \quad \mathbf{f} = \begin{pmatrix} t^2(x^2+y^2) \\ (1 - e^{-t})\cos(x+y) \end{pmatrix}, \tag{9}$$

 ISBN: 978-84-617-8694-7

and the final time is $T = 1$. The exact solution of problem (1) and (9) is unknown.

Figure 1 displays the numerical solution, at $t = 0.5$ and $t = 1$, using the scheme (7) when the discretization parameters are $N = 64$, $M = 32$ and the diffusion parameter is $\varepsilon = 10^{-4}$. From it, we clearly see the boundary layers at the four sides of the spatial domain.

Figure 1: Components u_1 (first row) and u_2 (second row) at $t = 0.5$ (left figures) and $t = 1$ (right figures) of problem (1) and (9) for $\varepsilon = 10^{-4}$ with $N = 64$ and $M = 32$



As the exact solution is unknown, we cannot calculate exactly the errors; to approximate them, we use a variant of the double-mesh principle (see [4]). Then, the maximum errors for each value of ε are approximated by

$$\mathbf{d}_{\varepsilon}^{N,M} = \max_{0\leq m\leq M} \max_{0\leq i,j\leq N} |\mathbf{U}_{i,j}^{m} - \widehat{\mathbf{U}}_{2i,2j}^{2m}|,$$

where $\{\widehat{\mathbf{U}}_{i,j}^{m}\}$ is the numerical solution on a finer mesh $\{(\hat{x}_i, \hat{y}_j, \hat{t}_m)\}$, which has the mesh

 ISBN: 978-84-617-8694-7

points of the coarse mesh and their midpoints, i.e.,

$$\begin{array}{l} \hat{x}_{2i} = x_i, \quad i = 0, \ldots, N, \quad \hat{x}_{2i+1} = (x_i + x_{i+1})/2, \quad i = 0, \ldots, N-1, \\ \hat{y}_{2j} = y_j, \quad j = 0, \ldots, N, \quad \hat{y}_{2j+1} = (y_j + y_{j+1})/2, \quad j = 0, \ldots, N-1, \\ \hat{t}_{2m} = t_m, \quad m = 0, \ldots, M, \quad \hat{t}_{2m+1} = (t_m + t_{m+1})/2, \quad m = 0, \ldots, M-1. \end{array} \tag{10}$$

From the maximum two-mesh differences $\mathbf{d}_\varepsilon^{N,M}$, we obtain the ε-uniform two-mesh differences by

$$\mathbf{d}^{N,M} = \max_\varepsilon \mathbf{d}_\varepsilon^{N,M}.$$

From the approximated maximum errors $\mathbf{d}_\varepsilon^{N,M}$, in a standard way, the numerical orders of convergence, for each value of ε, are calculated by

$$p_\varepsilon^{N,M} = \log\left(\mathbf{d}_\varepsilon^{N,M}/\mathbf{d}_\varepsilon^{2N,2M}\right)/\log 2,$$

and from the approximated uniform maximum errors $\mathbf{d}^{N,M}$, the numerical uniform orders of convergence are calculated by

$$p^{N,M} = \log\left(\mathbf{d}^{N,M}/\mathbf{d}^{2N,2M}\right)/\log 2.$$

Tables 1 and 2 show the maximum two-mesh differences and the orders of convergence for components u_1 and u_2 respectively; from them, we clearly deduce that the method is first-order uniformly convergent. Moreover, we can conclude that the errors associated with the time discretization dominate into the global error of the numerical method.

In Tables 3 and 4 the discretization parameters are multiplied by different factors so that the errors associated with the space discretization dominate into the global errors. The space and time discretization parameters are multiplied by a factor of 2 and 4, respectively. The computed orders of convergence now show almost second order of convergence, in agreement with Theorem 1.

In the second example the reaction matrix and the right-hand side are

$$\mathcal{A} = \begin{pmatrix} e^{x+y} & -(x+y) & -tx \\ -(x+y) & (t+1)(3+x+y) & -t\sin(y) \\ -tx & -t\sin(y) & e^t(2+\cos(x+y)) \end{pmatrix}, \quad \mathbf{f} = \begin{pmatrix} txy \\ (1-e^{-t})\sin(xy) \\ (t+1)\cos(x+y) \end{pmatrix}, \tag{11}$$

and the final time is again $T = 1$. The computed solutions with the finite difference scheme (7) at $t = 0.5$ and $t = 1$ are displayed in Figure 2. These surfaces show the presence of boundary layers in the three components of the solution.

Similarly to the previous example, we show the numerical results for each component in separate tables. The maximum two-mesh differences and the orders of convergence for u_i, $i = 1, 2, 3$ are given in Tables 5-7 respectively, where the discretization parameters N and M are multiplied by a factor of 2. The ε-uniform computed orders of convergence shows first order, and it again agrees with Theorem 1.

 ISBN: 978-84-617-8694-7

Table 1: Test problem (1) and (9): Maximum and uniform two-mesh differences and their orders of convergence for the component u_1

	N=16	N=32	N=64	N=128
	M=12	M=24	M=48	M=96
$\varepsilon = 1$	1.240E-4	4.184E-5	2.772E-5	1.637E-5
	1.568	0.594	0.759	
$\varepsilon = 10^{-1}$	2.419E-3	1.353E-3	7.130E-4	3.659E-4
	0.838	0.924	0.962	
$\varepsilon = 10^{-2}$	8.373E-3	3.616E-3	1.857E-3	9.448E-4
	1.211	0.962	0.975	
$\varepsilon = 10^{-3}$	1.101E-2	5.750E-3	2.594E-3	1.296E-3
	0.937	1.148	1.001	
$\varepsilon = 10^{-4}$	1.251E-2	6.035E-3	2.946E-3	1.464E-3
	1.052	1.035	1.009	
$\varepsilon = 10^{-5}$	1.299E-2	6.283E-3	3.084E-3	1.528E-3
	1.047	1.027	1.013	
$\varepsilon = 10^{-6}$	1.313E-2	6.361E-3	3.129E-3	1.551E-3
	1.046	1.023	1.012	
$\varepsilon = 10^{-7}$	1.318E-2	6.386E-3	3.143E-3	1.559E-3
	1.045	1.023	1.011	
$\varepsilon = 10^{-8}$	1.320E-2	6.393E-3	3.148E-3	1.562E-3
	1.045	1.022	1.011	
$\varepsilon = 10^{-9}$	1.320E-2	6.396E-3	3.149E-3	1.563E-3
	1.045	1.022	1.011	
$\varepsilon = 10^{-10}$	1.320E-2	6.397E-3	3.150E-3	1.563E-3
	1.045	1.022	1.011	
$d_1^{N,M}$	1.320E-2	6.397E-3	3.150E-3	1.563E-3
$p_1^{N,M}$	1.045	1.022	1.011	

Acknowledgements

This research was partially supported by the Instituto Universitario de Investigación en Matemáticas y Aplicaciones (IUMA), the projects MTM2014-52859 and MTM2016-75139-R and the Diputación General de Aragón.

References

[1] C. Clavero, J.L. Gracia, *Uniformly convergent additive finite difference schemes for singularly perturbed parabolic reaction-diffusion system*, Computer and Mathematics with Applications **67** (2014) 655–670.

[2] C. Clavero, J.L. Gracia, F. Lisbona, *Second order uniform approximations for the solution of time dependent singularly perturbed reaction-diffusion systems*, Int. J. Numer. Anal. Mod. **7** (2010) 428–443.

[3] J.L. Gracia, F. Lisbona, E. O'Riordan, *A coupled system of singularly perturbed parabolic reaction-diffusion equations*, Adv. Comput. Math. **32** (2010) 43–61.

Table 2: Test problem (1) and (9): Maximum and uniform two-mesh differences and their orders of convergence for the component u_2

	N=16	N=32	N=64	N=128
	M=12	M=24	M=48	M=96
$\varepsilon = 1$	4.892E-5	2.314E-5	1.223E-5	6.289E-6
	1.080	0.920	0.960	
$\varepsilon = 10^{-1}$	8.118E-4	4.393E-4	2.280E-4	1.161E-4
	0.886	0.946	0.974	
$\varepsilon = 10^{-2}$	2.251E-3	1.052E-3	5.320E-4	2.677E-4
	1.098	0.983	0.991	
$\varepsilon = 10^{-3}$	2.633E-3	1.313E-3	6.497E-4	3.285E-4
	1.004	1.015	0.984	
$\varepsilon = 10^{-4}$	2.916E-3	1.494E-3	7.490E-4	3.758E-4
	0.964	0.996	0.995	
$\varepsilon = 10^{-5}$	3.046E-3	1.564E-3	7.900E-4	3.962E-4
	0.961	0.986	0.996	
$\varepsilon = 10^{-6}$	3.088E-3	1.587E-3	8.036E-4	4.039E-4
	0.960	0.982	0.992	
$\varepsilon = 10^{-7}$	3.101E-3	1.594E-3	8.079E-4	4.065E-4
	0.960	0.981	0.991	
$\varepsilon = 10^{-8}$	3.105E-3	1.597E-3	8.093E-4	4.073E-4
	0.959	0.980	0.990	
$\varepsilon = 10^{-9}$	3.106E-3	1.597E-3	8.097E-4	4.076E-4
	0.959	0.980	0.990	
$\varepsilon = 10^{-10}$	3.107E-3	1.598E-3	8.099E-4	4.077E-4
	0.959	0.980	0.990	
$d_2^{N,M}$	3.107E-3	1.598E-3	8.099E-4	4.077E-4
$p_2^{N,M}$	0.959	0.980	0.990	

[4] P.A. FARRELL, A.F. HEGARTY, J.J.H. MILLER, E. O'RIORDAN, G.I. SHISHKIN, *Robust Computational Techniques for Boundary Layers, Applied Mathematics*, **16**. Chapman and Hall/CRC, 2000.

[5] R.B. KELLOGG, T. LINSS, M. STYNES, *A finite difference method on layer-adapted meshes for an elliptic reaction-diffusion system in two dimensions*, Math. Comput. **774** (2008) 2085–2096.

[6] R.B. KELLOGG, N. MADDEN, M. STYNES, *A parameter robust numerical method for a system of reaction-diffusion equations in two dimensions*, Num. Meth. Part. Diff. Equa. **24** (2007) 312–334.

[7] T. LINSS, M. STYNES, *Numerical solution of systems of singularly perturbed differential equations,* Comput. Methods Appl. Math. **9** (2009) 165–191.

[8] N. MADDEN, M. STYNES, *A uniformly convergent numerical method for a coupled system of two singularly perturbed linear reaction-diffusion problems*, IMA J. Numer. Anal. **23** (2003) 627–644.

Table 3: Test problem (1) and (9): Maximum and uniform two-mesh differences and their orders of convergence for the component u_1

	N=16	N=32	N=64	N=128
	M=12	M=48	M=192	M=768
$\varepsilon = 1$	1.240E-4	3.513E-5	9.251E-6	2.373E-6
	1.820	1.925	1.963	
$\varepsilon = 10^{-1}$	2.419E-3	5.957E-4	1.483E-4	3.704E-5
	2.022	2.006	2.001	
$\varepsilon = 10^{-2}$	8.373E-3	3.388E-3	9.724E-4	2.564E-4
	1.305	1.801	1.923	
$\varepsilon = 10^{-3}$	1.101E-2	6.316E-3	2.552E-3	9.826E-4
	0.801	1.307	1.377	
$\varepsilon = 10^{-4}$	1.251E-2	6.357E-3	2.573E-3	9.875E-4
	0.977	1.305	1.382	
$\varepsilon = 10^{-5}$	1.299E-2	6.371E-3	2.580E-3	9.891E-4
	1.027	1.304	1.383	
$\varepsilon = 10^{-6}$	1.313E-2	6.375E-3	2.582E-3	9.896E-4
	1.043	1.304	1.384	
$\varepsilon = 10^{-7}$	1.318E-2	6.376E-3	2.583E-3	9.898E-4
	1.048	1.304	1.384	
$\varepsilon = 10^{-8}$	1.320E-2	6.377E-3	2.583E-3	9.898E-4
	1.049	1.304	1.384	
$\varepsilon = 10^{-9}$	1.320E-2	6.377E-3	2.583E-3	9.898E-4
	1.050	1.304	1.384	
$\varepsilon = 10^{-10}$	1.320E-2	6.377E-3	2.583E-3	9.898E-4
	1.050	1.304	1.384	
$d_1^{N,M}$	1.320E-2	6.377E-3	2.583E-3	9.898E-4
$p_1^{N,M}$	1.050	1.304	1.384	

[9] G.I. Shishkin, *Approximation of systems of singularly perurbed elliptic reaction-diffusion equations with two parameters*, Comput. Math. Math. Phys. **47** (2007) 797–828.

 ISBN: 978-84-617-8694-7

Table 4: Test problem (1) and (9): Maximum and uniform two-mesh differences and their orders of convergence for the component u_2

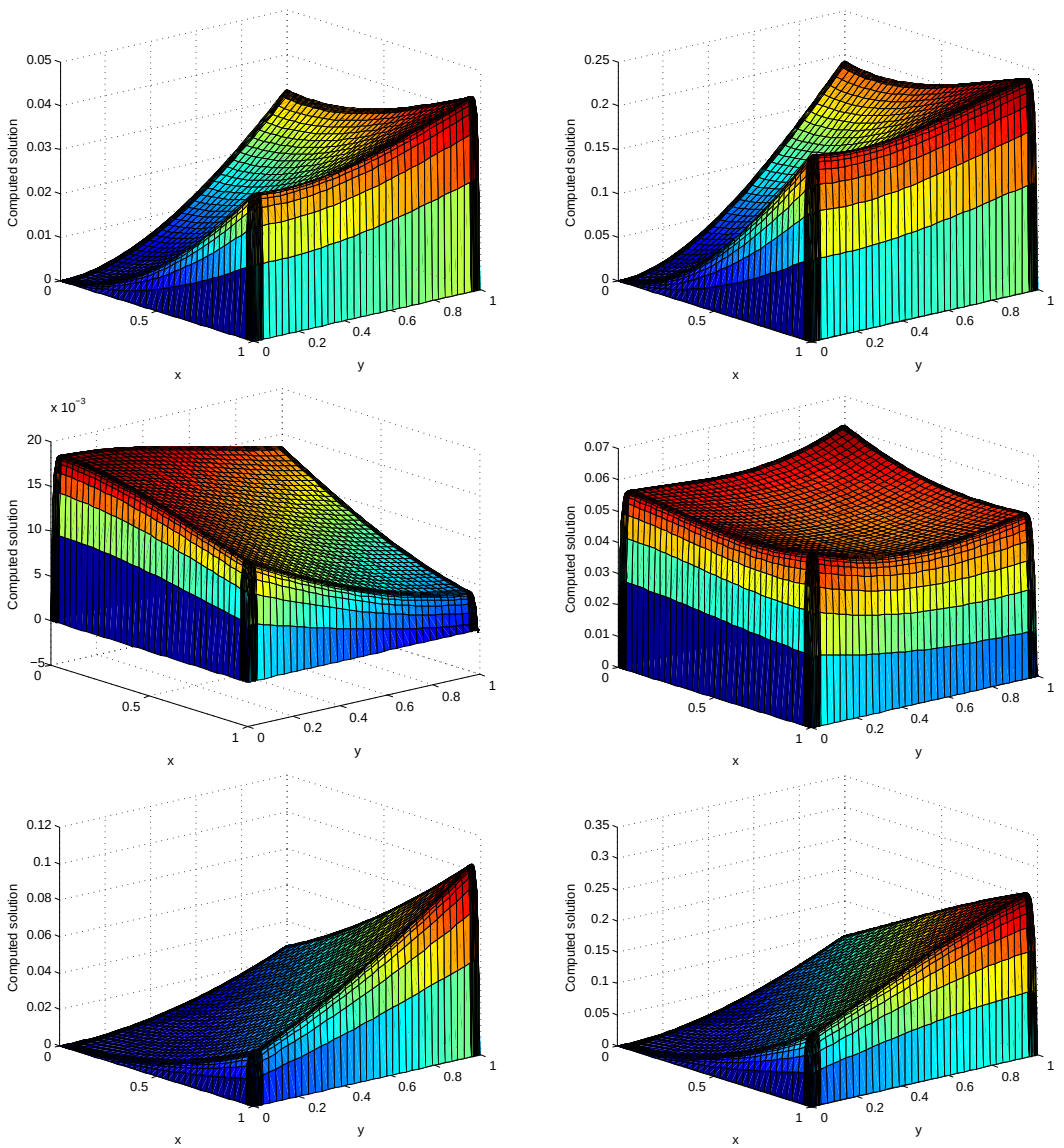
	N=16	N=32	N=64	N=128
	M=12	M=48	M=192	M=768
$\varepsilon = 1$	4.892E-5	1.227E-5	3.079E-6	7.824E-7
	1.996	1.994	1.976	
$\varepsilon = 10^{-1}$	8.118E-4	2.140E-4	5.427E-5	1.363E-5
	1.923	1.979	1.994	
$\varepsilon = 10^{-2}$	2.251E-3	7.215E-4	2.041E-4	5.299E-5
	1.641	1.822	1.945	
$\varepsilon = 10^{-3}$	2.633E-3	1.385E-3	5.644E-4	2.039E-4
	0.927	1.295	1.469	
$\varepsilon = 10^{-4}$	2.916E-3	1.386E-3	5.646E-4	2.040E-4
	1.074	1.295	1.469	
$\varepsilon = 10^{-5}$	3.046E-3	1.386E-3	5.644E-4	2.040E-4
	1.137	1.296	1.468	
$\varepsilon = 10^{-6}$	3.088E-3	1.386E-3	5.643E-4	2.040E-4
	1.156	1.296	1.468	
$\varepsilon = 10^{-7}$	3.101E-3	1.386E-3	5.643E-4	2.040E-4
	1.162	1.296	1.468	
$\varepsilon = 10^{-8}$	3.105E-3	1.386E-3	5.643E-4	2.040E-4
	1.164	1.296	1.468	
$\varepsilon = 10^{-9}$	3.106E-3	1.386E-3	5.643E-4	2.040E-4
	1.165	1.296	1.468	
$\varepsilon = 10^{-10}$	3.107E-3	1.386E-3	5.643E-4	2.040E-4
	1.165	1.296	1.468	
$d_2^{N,M}$	3.107E-3	1.386E-3	5.646E-4	2.040E-4
$p_2^{N,M}$	1.165	1.295	1.469	

Table 5: Test problem (1) and (11): Maximum and uniform two-mesh differences and their orders of convergence for the component u_1

	N=16	N=32	N=64	N=128
	M=8	M=16	M=32	M=64
$\varepsilon = 1$	2.026E-4	6.121E-5	2.165E-5	1.117E-5
	1.727	1.499	0.955	
$\varepsilon = 10^{-1}$	1.567E-3	5.514E-4	2.875E-4	1.468E-4
	1.507	0.939	0.970	
$\varepsilon = 10^{-2}$	9.990E-3	3.609E-3	1.049E-3	3.086E-4
	1.469	1.783	1.765	
$\varepsilon = 10^{-3}$	1.125E-2	6.274E-3	2.585E-3	9.934E-4
	0.842	1.280	1.379	
$\varepsilon = 10^{-4}$	1.131E-2	6.306E-3	2.593E-3	9.970E-4
	0.842	1.282	1.379	
$\varepsilon = 10^{-5}$	1.132E-2	6.315E-3	2.596E-3	9.980E-4
	0.843	1.282	1.379	
$\varepsilon = 10^{-6}$	1.133E-2	6.318E-3	2.597E-3	9.984E-4
	0.843	1.283	1.379	
$\varepsilon = 10^{-7}$	1.133E-2	6.319E-3	2.597E-3	9.985E-4
	0.843	1.283	1.379	
$\varepsilon = 10^{-8}$	1.133E-2	6.319E-3	2.597E-3	9.985E-4
	0.843	1.283	1.379	
$\varepsilon = 10^{-9}$	1.133E-2	6.320E-3	2.597E-3	9.985E-4
	0.843	1.283	1.379	
$\varepsilon = 10^{-10}$	1.133E-2	6.320E-3	2.597E-3	9.985E-4
	0.843	1.283	1.379	
$d_1^{N,M}$	1.133E-2	6.320E-3	2.597E-3	9.985E-4
$p_1^{N,M}$	0.843	1.283	1.379	

 ISBN: 978-84-617-8694-7

Figure 2: Components u_1 (first row), u_2 (second row) and u_3 (third row) at $t = 0.5$ (left figures) and $t = 1$ (right figures) of problem (1) and (11) for $\varepsilon = 10^{-4}$ with $N = 64$ and $M = 32$



 ISBN: 978-84-617-8694-7

Table 6: Test problem (1) and (11): Maximum and uniform two-mesh differences and their orders of convergence for the component u_2

	N=16	N=32	N=64	N=128
	M=8	M=16	M=32	M=64
$\varepsilon = 1$	6.031E-5	3.204E-5	1.673E-5	8.587E-6
	0.913	0.938	0.962	
$\varepsilon = 10^{-1}$	7.733E-4	3.863E-4	1.933E-4	9.683E-5
	1.001	0.999	0.998	
$\varepsilon = 10^{-2}$	1.578E-3	6.477E-4	3.220E-4	1.613E-4
	1.284	1.008	0.998	
$\varepsilon = 10^{-3}$	2.045E-3	1.135E-3	4.834E-4	2.302E-4
	0.850	1.231	1.071	
$\varepsilon = 10^{-4}$	2.392E-3	1.135E-3	5.473E-4	2.711E-4
	1.076	1.052	1.013	
$\varepsilon = 10^{-5}$	2.509E-3	1.198E-3	5.834E-4	2.880E-4
	1.066	1.038	1.018	
$\varepsilon = 10^{-6}$	2.546E-3	1.219E-3	5.960E-4	2.944E-4
	1.062	1.033	1.017	
$\varepsilon = 10^{-7}$	2.558E-3	1.226E-3	6.000E-4	2.967E-4
	1.061	1.031	1.016	
$\varepsilon = 10^{-8}$	2.562E-3	1.228E-3	6.013E-4	2.975E-4
	1.061	1.030	1.015	
$\varepsilon = 10^{-9}$	2.563E-3	1.229E-3	6.017E-4	2.977E-4
	1.061	1.030	1.015	
$\varepsilon = 10^{-10}$	2.563E-3	1.229E-3	6.018E-4	2.978E-4
	1.060	1.030	1.015	
$d_2^{N,M}$	2.563E-3	1.229E-3	6.018E-4	2.978E-4
$p_2^{N,M}$	1.060	1.030	1.015	

Table 7: Test problem (1) and (11): Maximum and uniform two-mesh differences and their orders of convergence for the component u_3

	N=16	N=32	N=64	N=128
	M=8	M=16	M=32	M=64
$\varepsilon = 1$	3.183E-4	1.844E-4	9.879E-5	5.106E-5
	0.787	0.901	0.952	
$\varepsilon = 10^{-1}$	3.082E-3	1.544E-3	7.713E-4	3.853E-4
	0.997	1.002	1.001	
$\varepsilon = 10^{-2}$	6.155E-3	3.176E-3	1.624E-3	8.220E-4
	0.955	0.967	0.983	
$\varepsilon = 10^{-3}$	8.837E-3	4.377E-3	2.210E-3	1.110E-3
	1.014	0.986	0.993	
$\varepsilon = 10^{-4}$	1.001E-2	4.927E-3	2.472E-3	1.238E-3
	1.023	0.995	0.997	
$\varepsilon = 10^{-5}$	1.039E-2	5.128E-3	2.570E-3	1.286E-3
	1.019	0.997	0.999	
$\varepsilon = 10^{-6}$	1.051E-2	5.199E-3	2.605E-3	1.302E-3
	1.016	0.997	1.000	
$\varepsilon = 10^{-7}$	1.055E-2	5.222E-3	2.617E-3	1.308E-3
	1.015	0.997	1.000	
$\varepsilon = 10^{-8}$	1.056E-2	5.229E-3	2.621E-3	1.310E-3
	1.015	0.996	1.000	
$\varepsilon = 10^{-9}$	1.057E-2	5.231E-3	2.622E-3	1.311E-3
	1.015	0.996	1.000	
$\varepsilon = 10^{-10}$	1.057E-2	5.232E-3	2.623E-3	1.311E-3
	1.014	0.996	1.000	
$d_3^{N,M}$	1.057E-2	5.232E-3	2.623E-3	1.311E-3
$p_3^{N,M}$	1.014	0.996	1.000	

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

Differential systems with reflection and matrix invariants

Santiago Codesido[1] and F. Adrián F. Tojo[2]

[1] *Département de Physique Théorique et Section de Mathématiques, Université de Genève*

[2] *Instituto de Matemáticas, Universidade de Santiago de Compostela*

emails: santiago.codesido@unige.ch, fernandoadrian.fernandez@usc.es

Abstract

In this work we derive important properties regarding matrix invariants which occur in the theory of differential equations with reflection.

Key words: differential equations with reflection, matrix invariants.

1 Introduction

In recent works regarding the solution and Green's functions of Differential Equations with Reflection (see for instance [1–3, 8]) the strong relation between linear analysis and linear algebra is highlighted. In particular, in the most recent of the aforementioned works, the authors obtain an explicit fundamental matrix for the system of differential equations with reflection

$$Hu(t) := Fu'(t) + Gu'(-t) + Au(t) + Bu(-t) = 0, t \in \mathbb{R}, \tag{1.1}$$

where $n \in \mathbb{N}$, $A, B, F, G \in \mathcal{M}_n(\mathbb{R})$ and $u : \mathbb{R} \to \mathbb{R}^n$. To be precise, they prove the following result.

Theorem 1.1 ([3])**.** *Assume $F - G$ and $F + G$ are invertible. Then*

$$X(t) := \sum_{k=0}^{\infty} \frac{E^k t^{2k}}{(2k)!} - (F+G)^{-1}(A+B) \sum_{k=0}^{\infty} \frac{E^k t^{2k+1}}{(2k+1)!},$$

where $E = (F-G)^{-1}(A-B)(F+G)^{-1}(A+B)$, is a fundamental matrix of problem (1.1)*. If we further assume $A-B$ and $A+B$ are invertible, then E is invertible and we can consider a square root Ω of E. Then,*

$$X(t) = \cosh \Omega t - (F+G)^{-1}(A+B)\Omega^{-1} \sinh \Omega t.$$

 ISBN: 978-84-617-8694-7

What is more, in another recent work the authors proved an analog of the Liouville's formula for the case with reflections in systems of order two.

Theorem 1.2 (Abel-Jacobi-Liouville Identity [5])**.** *Let $n = 2$ in equation (1.1). Then $(|X|, |X'|)$ is the unique solution of the system of differential equations*

$$\begin{aligned} x'' &= \operatorname{tr}(E)x - 2y, \\ y'' &= -2|E|x + \operatorname{tr}(E)y, \end{aligned}$$

subject to the one point conditions

$$x(0) = 1, \quad y(0) = |M_+|, \quad x'(0) = -\operatorname{tr}(M_+), \quad y'(0) = \operatorname{tr}(\operatorname{Adj}(M_+)E).$$

The authors also presented in that work the following conjecture:

Conjecture 1.3. *For any $n \geq 1$, if $X(t)$ is a fundamental matrix of problem (1.1), then $|X(t)|$ can be obtained as a component of the solution of a linear system of differential equations with constant coefficients, those coefficients depending only on the different matrix invariants of E, which is defined as in Theorem 1.1.*

In order to attempt proving this conjecture, and taking into account the proof of Theorem 1.2, we need to study the different matrix invariants of the matrices appearing in the theory.

2 The Y matrix

For $X(t)$ the fundamental matrix of the problem, define

$$Y(t) := X(t)^{-1}X'(t). \tag{2.1}$$

We have that $X = S_1 - M_+S_2$ where S_1 and S_2 s are power series in E which we can formally give, by using $\Omega = \sqrt{E}$, as $S_1 = \cosh(\Omega t)$, $S_2 = \Omega^{-1}\sinh(\Omega t)$.

Notice both are indeed power series in $\Omega^2 = E$. Since $X'' = XE$ and $(Y^{-1})' = -Y^{-1}Y'Y^{-1}$ we have that

$$Y' = E - Y^2. \tag{2.2}$$

Using the construction of [6] we build an associated ODE system

$$z' = \begin{pmatrix} 0 & E \\ I & 0 \end{pmatrix} z.$$

The system has as fundamental matrix

$$\begin{pmatrix} \cosh(\Omega t) & \Omega^{-1}\sinh(\Omega t) \\ \Omega\sinh(\Omega t) & \cosh(\Omega t) \end{pmatrix}.$$

 ISBN: 978-84-617-8694-7

The solution of equation (2.2) is then given by

$$Y(t) = [\cosh(\Omega t)Y(0) + \Omega \sinh(\Omega t)] \left[\Omega^{-1} \sinh(\Omega t)Y(0) + \cosh(\Omega t)\right]^{-1},$$

which in terms of the S functions is $Y(t) = [-S_1 M_+ + E S_2] [-S_2 M_+ + S_1]^{-1}$, where we fix the initial condition with $Y(0) = X'(0) = -M_+$. This seems like a commuted version of expression (2.1), but it is nothing more than the hypergeometric identity.

Consider the Liouville equation for Y itself, that is,

$$(\log |Y|)' = \mathrm{Tr}\left(Y^{-1}Y'\right) = \mathrm{Tr}\left(Y - Y^{-1}E\right).$$

Then we have $\mathrm{Tr}(Y^{-1}E) = \mathrm{Tr}(Y) - (\log |Y|)'$, which can be calculated in terms of invariants of $|X|$.

3 Complex systems

The main involution occurring in the theory of complex variable is the complex conjugation $\mathcal{C} : \mathbb{C} \to \mathbb{C}$, $\mathcal{C}(z) = \overline{z}$. It is, in fact, a reflection with respect to the second variable if we write $z = (x, y) \in \mathbb{R}^2$: $\mathcal{C}(x, y) = (x, -y)$.

We consider now an operator L acting on $z(t)$ as

$$A_0 z(t) + A_1 \overline{z(t)} + B_0 z(t) + B_1 \overline{z'(t)}, \tag{3.1}$$

where $z : \mathbb{R} \to \mathbb{C}^n$, and $A_i, B_i \in \mathcal{M}_{n\times n}(\mathbb{C}) := \mathcal{M}$.

We can consider a extended algebra $\mathcal{M}^*$ by with the linear operation of complex conjugation which acts as $\mathcal{C}z = \overline{z}, z \in \mathbb{C}^n$.

It is easy to see that the following properties hold[1],

$$\mathcal{C}^2 = I, \mathcal{C}A = \overline{A}\mathcal{C}, \tag{3.2}$$

where $\overline{A}$ is the complex conjugate of A.

The we consider the free product quotiented by these relations,

$$\mathcal{M}^* = \mathcal{M} \star \{\mathcal{C}\} \,/ \left(\mathcal{C}^2 = I, \mathcal{C}A = \overline{A}\mathcal{C}\right).$$

Now, we can see that this is in fact a $\mathbb{Z}_2$-graded algebra. Due to the conditions (3.2), we can move any $\mathcal{C}$'s to the right, and any power of it is reduced modulo 2. Therefore, any element of $A \in \mathcal{M}^*$ can be written as $\mathbf{A} = A_0 + A_1\mathcal{C}$ with $A_0, A_1 \in \mathcal{M}$. As a vector space, $\mathcal{M}^* = \mathcal{M} \oplus \mathcal{M}$.

[1] In fact, one could use here any involution for which matrix conjugation verifies $\mathcal{C}A\mathcal{C} \in \mathcal{M}$ by defining $\overline{A}$ suitably.

 ISBN: 978-84-617-8694-7

The grading is clear by looking at the product of two generic elements,

$$\mathbf{AB} = (A_0 + A_1\mathcal{C})(B_0 + B_1\mathcal{C}) = \left(A_0B_0 + A_1\overline{B_1}\right) + \left(A_0B_1 + A_1\overline{B_0}\right)\mathcal{C}. \tag{3.3}$$

We can also calculate a explicit inverse, using $(I - A\mathcal{C})(I + A\mathcal{C}) = I - A\overline{A}$, from where $(I + A\mathcal{C})^{-1} = \left(I - A\overline{A}\right)(I - A\mathcal{C})$.

Since $(AB)^{-1} = B^{-1}A^{-1}$, we can generalize it to

$$(A_0 + A_1\mathcal{C})^{-1} = \left(A_1^{-1}A_0 - \overline{A_0^{-1}A_1}\right)^{-1}\left(A_1^{-1} - \overline{A_0^{-1}}\mathcal{C}\right).$$

As it is, the expression is unclear when either $A_i = 0$. We rewrite

$$\mathbf{A}^{-1} = \Delta(A_0, A_1) + \Delta\left(\overline{A_1}, \overline{A_0}\right)\mathcal{C}, \tag{3.4}$$

with

$$\Delta(A_0, A_1) = \begin{cases} 0, & A_0 = 0, A_1 \neq 0, \\ \left(A_0 - A_1\overline{A_0^{-1}A_1}\right)^{-1}, & A_0 \neq 0, A_1 \neq 0. \end{cases}$$

As a last note, this can of course be realized as a matrix algebra over $\mathbb{R}^{2n}$, although it does not lead to anything new (other than clutter). For the record, one can take a representation

$$\rho(z) = \left(\begin{array}{c} \Re z \\ \hline \Im z \end{array}\right), \quad \rho(A) = \left(\begin{array}{c|c} \Re A & -\Im A \\ \hline \Im A & \Re A \end{array}\right), \quad \rho(\mathcal{C}) = \left(\begin{array}{c|c} I & 0 \\ \hline 0 & -I \end{array}\right),$$

for which it is easy to see that properties such as (3.2) or (3.3) hold.

3.1 Equation reduction

With these tools, one can rewrite equation (3.1) as $\mathbf{B}z'(t) + \mathbf{A}z(t) = 0$, which can be reduced to $z'(t) + \left(\mathbf{B}^{-1}\mathbf{A}\right)z(t) = 0$. This means we can just focus on the study of

$$z'(t) + A_0z(t) + A_1\overline{z(t)} = 0.$$

Of course, one could now look for a fundamental operator inside the $\mathcal{M}^*$ algebra, such that solutions fulfill

$$z(t) = \mathbf{X}(t)z(0),$$

which is

$$\mathbf{X} = \cosh(\mathbf{A}t) - \sinh(\mathbf{A}t).$$

Unfortunately, it does not seem easy to calculate terms like $\mathbf{A}^n$ at the moment. We could in principle directly get an explicit fundamental matrix if we could write a manageable

 ISBN: 978-84-617-8694-7

expression. However, we do learn something important. Since $\mathbf{A}^n \in \mathcal{M}^*$, then $\cosh(\mathbf{A}t) \in \mathcal{M}^*$ too. Hence, if we want to find fundamental matrices for the problem, the ansatz must be of the form

$$z(t) = \left(X_0(t) + X_1(t)\mathcal{C}\right) z(0) = X_0(t)z(0) + X_1(t)\overline{z(0)},$$

where $X(0) = I$ and $Y(0) = 0$ as to agree with $\mathbf{X}(0) = I$.

3.2 $\mathbf{A^n}$ generating function

The components of $\mathbf{A}^n$ can still be algorithmically computed by using the expression

$$(I - t\mathbf{A})^{-1} = \sum_{n=0}^{\infty} (t\mathbf{A}).$$

By writing the explicit inverse of $I - t\mathbf{A} = (I - tA_0) - tA_1\mathcal{C}$, we get

$$\mathbf{A}^n = \frac{1}{n!}\frac{\mathrm{d}}{\mathrm{d}t}\left[\left(A_1^{-1}(I - tA_0) - \overline{(I - tA_0)^{-1}A_1}\right)^{-1}\left(A_1^{-1} - \overline{(t^{-1}I - A_0)^{-1}}\mathcal{C}\right)\right]_{t=0}.$$

3.3 Ansatz

We take the system

$$z' + Az + \overline{Bz} = 0, z : \mathbb{R} \to \mathbb{C}^n, A, B \in \mathcal{M}$$

and introduce the ansatz

$$z = Xz_0 + \overline{Yz_0}, z_0 = z(0), X, Y : \mathbb{R} \to \mathcal{M},$$

$$\left(X' + AX + \overline{B}Y\right) z_0 + \left(Y' + \overline{A}Y + BX\right)\overline{z_0} = 0,$$

$$\begin{cases} X' + AX + \overline{B}Y = 0, \\ Y' + \overline{A}Y + BX = 0. \end{cases} \tag{3.5}$$

This a ordinary system. Take X'' and substitute Y' and Y through equation (3.5),

$$X'' + AX' + \overline{B}\left(-BX - \overline{A\overline{B^{-1}}}\left(-X' - AX\right)\right) = 0,$$
$$X'' + \left(A + \overline{BAB^{-1}}\right)X' + \left(\overline{BAB^{-1}}A - \overline{B}B\right)X = 0.$$

Unsurprisingly, we get a very similar structure to the inverses in expression (3.4).For the sake of notation, we will rename the coefficients as

$$X'' + FX' + GX = 0.$$

Repeating the process, we get

$$Y'' + \overline{F}Y' + \overline{G}Y = 0.$$

The initial conditions for this second order problem are given by $\mathbf{X}(0) = I$ and equation (3.5). That is,

$$X(0) = I, \quad Y(0) = 0, \quad X'(0) = -A, \quad Y'(0) = -\overline{A}.$$

In principle, we could now take as an ansatz

$$X = \alpha e^{(\Gamma+\Omega)t} + \beta e^{(\Gamma-\Omega)t},$$

subject to the conditions

$$X(0) = \alpha + \beta, X'(0) = \alpha(\Gamma + \Omega) + \beta(\Gamma - \Omega),$$

which can be inverted into

$$\begin{aligned}\alpha =& \frac{1}{2}\left[X'(0) - X(0)\left(\Gamma - \Omega\right)\right]\Omega^{-1},\\ \beta =& -\frac{1}{2}\left[X'(0) - X(0)\left(\Gamma + \Omega\right)\right]\Omega^{-1}.\end{aligned}$$

4 Generalized matrix invariants

In the following section we use the concept of crossed or generalized matrix invariants which can be found in [7] and [4] among others.

4.1 Definition and basic properties

Let $X_1, \ldots, X_N \in GL\left(n\right)$. Define

$$Z\left(X_1, \ldots, X_N\right) := \det\left(I + \sum_{i=1}^{N} \alpha_i X_i\right) := \sum_{m_i} \alpha_1^{m_1} \ldots \alpha_N^{m_N} Z_{m_1,\ldots,m_N}\left(X_1, \ldots, X_N\right). \quad (4.1)$$

Since det is an algebraic combination of matrix entries, the expansion is a polynomial in the α_i variables. We can however take the sum to be over all integer values of m_i by suitably defining its α-coefficients, $Z_{m_1,\ldots,m_N}$, as zero when not corresponding to any power that appears in the det expansion. In particular,

$$Z_{m_1,\ldots,m_N}\left(X_1, \ldots, X_n\right) = 0 \text{ if } \min\left\{m_i\right\} < 0. \quad (4.2)$$

 ISBN: 978-84-617-8694-7

These Z coefficients then give us the *generalized matrix invariants*, which reduce to the usual ones when we only consider one matrix (or set the other indices to 0). We can get explicit expressions in terms of traces via

$$\det(I+\alpha X) = e^{\mathrm{Tr}\log(I+\alpha X)} \tag{4.3}$$

by expanding the Taylor series around $\alpha = 0$ and using the linearity of the trace. This already gives, looking at the leading order of the exponential expansion,

$$Z_{0,\dots,0}(X_1,\dots,X_n) = 1. \tag{4.4}$$

In the same way, we can reduce any expression with a 0 index,

$$Z_{n_1,\dots,n_{N-1},0}\left(X_1,\dots,X_N\right) = Z_{n_1,\dots,n_{N-1}}\left(X_1,\dots,X_{N-1}\right).$$

Looking at higher coefficients upon expanding the exponential returns higher invariants. For instance,

$$\begin{aligned}
\mathrm{Tr}\log\left(I+\alpha X\right) =& \alpha \mathrm{Tr}\left(X\right) - \frac{\alpha^2}{2}\mathrm{Tr}\left(X^2\right) + \frac{\alpha^3}{3}\mathrm{Tr}\left(X^3\right) + O\left(\alpha^4\right),\\
Z_1(X) =& \mathrm{Tr}(X),\\
Z_2(X) =& \frac{1}{2}\left(\mathrm{Tr}(X)^2 - \mathrm{Tr}(X^2)\right),\\
Z_3(X) =& \frac{1}{6}\left(\mathrm{Tr}(X)^3 - 3\mathrm{Tr}(X^2)\mathrm{Tr}(X) + \mathrm{Tr}(X^3)\right),
\end{aligned}$$

etc, but also

$$Z_{1,1}(X,Y) = \mathrm{Tr}(X)\mathrm{Tr}(Y) - \mathrm{Tr}(XY).$$

Of course, equation (4.3) is usually proven using Liouville's formula. We will make contact with it again later, when looking at the derivatives of the Z invariants themselves.

4.2 Factorization

Consider

$$\det\left(1+\alpha A + \sum_i \beta_i B\right),$$

and the fact that

$$\det\left(\alpha A\right) = \alpha^n \det A.$$

Extract this determinant from the original expansion,

$$\begin{aligned}
\det\left(I+\alpha A + \sum_i \beta_i B\right) &= \sum_{l,m_i} \alpha^l \left(\prod \beta^{m_i}\right) Z_{l,m_1,\dots,m_N}\left(A,B_1,\dots,B_N\right)\\
&= \det\left(A\right) \sum_{l,m_i} \alpha^l \left(\prod \beta_i^{m_i}\right) Z_{n-l-\sum m_i,m_1,\dots,m_N}\left(A^{-1},A^{-1}B_1,\dots,A^{-1}B_N\right).
\end{aligned}$$

 ISBN: 978-84-617-8694-7

To equate the two polynomials, we equate every coefficient and get a **duality** relationship

$$Z_{l,m_1,\ldots,m_N}\left(A,B_1,\ldots,B_N\right)=\det\left(A\right)Z_{n-l-\sum m_i,m_1,\ldots,m_N}\left(A^{-1},A^{-1}B_1,\ldots,A^{-1}B_N\right). \tag{4.5}$$

Already an interesting property comes from the fact that any Z with negative indices must be 0, by equation (4.2). The dual of this statement is then

$$Z_{m_1,\ldots,m_N}\left(X_1,\ldots,X_N\right)=0 \text{ if } \sum_i m_i > n.$$

We will call the sum of all indices $\sum m_i$ the **order** of the trace $Z_{m_1,\ldots,m_N}$. That an invariant of order higher than the size of the matrix is zero reduces, as expected, to the usual property of matrix invariants when we have a single matrix, and together with expression (4.2) ensures that only a finite number of Z invariants for any given set of X_i is non-zero.

We can also take a dual of equation (4.4), which is the well known

$$Z_n\left(X\right)=\det\left(X\right)$$

or

$$1=\det\left(X\right)Z_n\left(X^{-1}\right).$$

Now, this statement gets interesting when we introduce more matrices. Consider the two matrix case,

$$Z_{l,m}\left(A,B\right)=\det\left(A\right)Z_{n-l-m,m}\left(A^{-1},A^{-1}B\right)$$

and set $A=X$, $B=XY$, and $l+m=n$,

$$Z_{n-m,m}\left(X,XY\right)=\det\left(X\right)Z_{0,m}\left(X^{-1},Y\right)=\det\left(X\right)Z_m\left(Y\right).$$

This, which is the generalization of

$$\det\left(XY\right)=\det\left(X\right)\det\left(Y\right),$$

allows us to decompose order n invariants of a product into products of invariants. In particular, we get the $n=2$ expression with which we built the ODE system.

More generally,

$$Z_{n-\sum m_i,m_1,\ldots,m_N}\left(X,XY_1,\ldots,XY_N\right)=\det\left(X\right)Z_{m_1,\ldots,m_N}\left(Y_1,\ldots,Y_N\right). \tag{4.6}$$

 ISBN: 978-84-617-8694-7

4.3 Small-ε expansion

By using expression (4.1) we can easily derive distributivity properties, which can be applied to calculate

$$\begin{aligned}
&Z_{l,m_1,\dots,m_N}\left(A_1+\varepsilon A_2+O\left(\varepsilon^2\right),B_1,\dots,B_N\right)\\
=&\sum_{i=0}^{l}\varepsilon^i Z_{l-i,i,m_1,\dots,m_N}\left(A_1,A_2+O\left(\varepsilon\right),B_1,\dots,B_N\right)\\
=&\sum_{i=0}^{l}\varepsilon^i \left[Z_{l-i,i,m_1,\dots,m_N}\left(A_1,A_2,B_1,\dots,B_N\right)+O\left(\varepsilon\right)\right]\\
=&\sum_{i=0}^{1}\varepsilon^i Z_{l-i,i,m_1,\dots,m_N}\left(A_1,A_2,B_1,\dots,B_N\right)+O\left(\varepsilon^2\right)\\
=&Z_{l,m_1,\dots,m_N}\left(A_1,B_1,\dots,B_N\right)+\varepsilon Z_{l-1,1,m_1,\dots,m_N}\left(A_1,A_2,B_1,\dots,B_N\right)+O\left(\varepsilon^2\right).
\end{aligned}$$

4.4 Derivatives

As we have seen before, the derivatives of the invariants play an essential role in the theory. We would now like to have a formula for derivatives of the form

$$\frac{\mathrm{d}}{\mathrm{d}t}Z_m\left(X\left(t\right)\right).$$

Consider

$$Z^{(m_0,m_1,m_2,\dots)}\left(X\right):=Z_{m_0,m_1,m_2,\dots}\left(X,X',X'',\dots\right),$$

such that for some N we have $m_i=0$ for every $i>N$.

We can retrieve its first derivative from its Taylor series, which we can in turn get from its small ε expansion.

$$\begin{aligned}
&Z^{(m_0,m_1,\dots)}\left(X+\varepsilon X'+O\left(\varepsilon^2\right)\right)\\
=&Z^{(m_0,m_1,m_2,\dots)}\left(X\right)+\varepsilon\left(m_1+1\right)Z_{m_0-1,m_1+1,m_2,\dots}\left(X,X',X'',\dots\right)\\
&+\varepsilon\left(m_2+1\right)Z_{m_0,m_1-1,m_2+1,\dots}\left(X,X',X'',\dots\right)+\cdots
\end{aligned}$$

Taking the ε term we get the first coefficient of the Taylor series, i. e., the first derivative,

$$\left(Z^{(m_0,m_1,m_2,\dots)}\left(X\right)\right)'=\sum_{i=1}^{\infty}\left(m_i+1\right)Z^{(m_0,\dots,m_{i-1}-1,m_i+1,\dots)}\left(X\right).$$

Notice that the infinite sum is merely formal, since by equation (4.2) it is guaranteed to terminate as soon as all the remaining m_i are 0, due to the $m_{i-1}-1$ index at every term.

 ISBN: 978-84-617-8694-7

For the first few derivatives, we find via recursion the general expressions

$$\begin{aligned} Z_m(X)' &= \left(Z^{(m)}\right)' = Z^{(m-1,1)}, \\ Z_m(X)'' &= \left(Z^{(m)}\right)'' = 2Z^{(m-2,2)} + Z^{(m-1,0,1)}, \\ Z_m(X)''' &= \left(Z^{(m)}\right)''' = 6Z^{(m-3,3)} + 3Z^{(m-2,1,1)} + Z^{(m-1,0,0,1)}. \end{aligned}$$

Something very important (albeit somehow obvious, following Leibniz's rule for matrices), is that the order of the invariants involved in the expressions is preserved.

This allows us to use the factorization formula (4.6) over the derivatives of the determinant, which corresponds to Z_n.

As a small note, if we take in the first derivative $m = n$ together with expression (4.5), we get

$$\det(X)' = Z_{n-1,1}\left(X, X'\right) = \det(X)\, Z_1\left(X^{-1}X'\right) = \det(X)\,\mathrm{Tr}\left(X^{-1}X'\right),$$

the usual Liouville's Formula.

4.5 Application to the differential system of invariants for n > 2

In the matrix dimension $m = 2$ case, taking derivatives of the determinant eventually closes, since $X'' = XE$. This follows from

$$\det(X)'' = Z_m(X)'' = 2Z^{(m-2,2)}(X) + Z^{(m-1,0,1)}(X).$$

The $Z^{(m-1,0,1)}(X)$ can be immediately rewritten as a determinant by using the duality formula,

$$Z_{m-1,0,1}(X, X', X'') = \det(X) Z_1(X^{-1}X'') = \det(X)\,\mathrm{tr}(E),$$

and, when $m = 2$,

$$Z^{(m-2,2)}(X) = Z^{(0,2)}(X) = \det(X').$$

Of course, now we can do the same for X',

$$\det(X')'' = 2Z^{(m-2,2)}(X) + Z^{(m-1,0,1)}(X)$$

and, for $m = 2$,

$$\det(X')'' = 2\det(X'') + \det(X')\,\mathrm{tr}(E) = 2\det(E)\det(X) + \det(X')\,\mathrm{tr}(E),$$

closing the system as we had found in Theorem 1.2. The problem is now obvious, since for $m > 2$, $Z^{(m-2,2)}(X)$ will involve a non trivial product between X and X'. One could consider this as a new variable for the system, but its derivatives will now concern objects

 ISBN: 978-84-617-8694-7

of the form $Z^{(m-2,1,1)}(X)$ which, if understood as yet another variable of the system, would yield upon derivation

$$Z^{(m-2,1,0,1)}(X),\ Z^{(m-2,1,0,0,1)}(X),\ Z^{(m-2,1,0,\ldots,0,1)},\ldots$$

Notice that this will always involve a term in X', and a term in X, so that we cannot perform the same trick as we did for $Z^{(m-1,0,1)}(X)$ –namely, using $X'' = XE$ to factor the determinant out. Hence, the system of second derivatives of invariants for $m > 2$ does not close.

Acknowledgements

Prof. F. Adrián F. Tojo is partially supported by project MTM2016-75140-P (AEI/FEDER, UE) and Xunta de Galicia (Spain), project EM2014/032.

References

[1] Cabada, A., Tojo, F.A.F.: *Solutions and Green's function of the first order linear equation with reflection and initial conditions.* Bound. Value Probl. **2014**(1), 99 (2014)

[2] Cabada, A., Tojo, F.A.F.: *Green's functions for reducible functional differential equations.* Bull. Malays. Math. Sci. Soc. pp. 1–22 (2016)

[3] Cabada, A., Tojo, F.A.F.: *On linear differential equations and systems with reflection.* Applied Mathematics and Computation **305**, 84–102 (2017)

[4] Codesido, S., Grassi, A., Marino, M.: *Spectral theory and mirror curves of higher genus.* In: Annales Henri Poincaré, pp. 1–64. Springer (2015)

[5] Codesido, S., Tojo, F.A.F.: *A Liouville's Formula for systems with reflection* (preprint)

[6] Levin, J.: *On the matrix Riccati equation.* Proceedings of the American Mathematical Society **10**(4), 519–524 (1959)

[7] Simon, B.: *Notes on infinite determinants of Hilbert space operators.* Advances in Mathematics **24**(3), 244–273 (1977)

[8] Tojo, F.A.F.: *Computation of Green's functions through algebraic decomposition of operators.* Boundary Value Problems **2016**(1), 167 (2016)

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

Stochastic liquidity horizon in market risk

Gemma Colldeforns-Papiol[1] and Luis Ortiz-Gracia[2]

[1] *Financial Mathematics and Risk Control, Centre de Recerca Matemàtica*

[2] *School of Economics, University of Barcelona*

emails: gcolldeforns@crm.cat, luis.ortiz-gracia@ub.edu

Abstract

The Basel Committee of Banking Supervision has recently set out the revised standards for minimum capital requirements for market risk. The Committee has focused, among others, on the two key areas of moving from Value-at-Risk (VaR) to Expected Shortfall (ES) and considering a comprehensive incorporation of the risk of market illiquidity by extending the risk measurement horizon. The estimation of the ES for several trading desks and taking into account different liquidity horizons is computationally very involved. We present a novel numerical method to compute the VaR and ES of a given portfolio within the stochastic holding period framework. Two approaches are considered, the delta-gamma approximation, for modelling the change in value of the portfolio as a quadratic approximation of the change in value of the risk factors, and some of the state-of-the-art stochastic processes for driving the dynamics of the log-value change of the portfolio like the Merton jump-diffusion model and the Kou model. Central to this procedure is the application of the SWIFT method developed for option pricing, that appears to be a very efficient and robust Fourier inversion method for risk management purposes.

Key words: market risk, liquidity risk, stochastic liquidity horizon, value-at-risk, expected shortfall, Fourier transform inversion, Shannon wavelets.

MSC 2000: 91G60, 62P05, 60E10, 65T60

1 Introduction

The Basel Committee of Banking Supervision states in the consultative documents [1, 2] that *"the financial crisis exposed material weaknesses in the overall design of the framework for capitalising trading activities. The level of capital required against trading book*

exposures proved insufficient to absorb losses". Within the mentioned documents, the Basel Committee initiated a fundamental review of the trading book regime, beginning with an assessment of those things that went wrong. The revised standards for minimum capital requirements for market risk were recently established in [3].

The Committee has focused, among others, on the two key areas of moving from VaR to ES and considering a comprehensive incorporation of the risk of market illiquidity. In regards to the first issue, a number of weaknesses have been identified with using VaR for determining regulatory capital requirements, including its inability to capture the risk in the tail. For this reason, the Committee has considered alternative risk metrics like, in particular, the ES, which measures the riskiness of a position by considering both the size and the likelihood of losses above a certain confidence level. The second issue relies on the importance of incorporating the risk of market illiquidity as a key consideration in banks' regulatory capital requirements for trading portfolios. The assumption that trading book risk positions where liquid, i.e., that banks could exit or hedge these positions over a ten-day horizon proved to be false during the recent crisis.

The estimation of the ES for several trading desks and taking into account different liquidity horizons is computationally very involved. In this work we present some of the results obtained in [5]. We compute the VaR and ES risk measures of a market portfolio and we assume that the holding period follows a certain positive stochastic process to account for liquidity risk. We will therefore measure the risk in the situation where the holding period is the liquidity horizon. To our best knowledge, this idea was first introduced in [4] as a proposal to open a research effort in stochastic holding period models for risk measures. In that paper the authors assume that the log-return on the portfolio value is normally distributed, which facilitates the calculation of the risk measures. Within this work, we give a step further by considering more realistic models for the log-value of the portfolio. On the one hand we propose the use of the delta-gamma approach, where it is assumed that the change in portfolio value is a quadratic function of the changes in the risk factors. On the other hand, we consider the Merton jump diffusion model and the Kou model to drive the log-return on the portfolio value. Under any of these scenarios, the closed formulae to compute the risk measures within the Gaussian setting in [4] are not available anymore.

2 Risk measures and stochastic liquidity horizon

Let us assume that the liquidity horizon follows a certain stochastic process $\{H(t)\}_{t\geq 0}$ where $H(t)$ is a positive random variable associated to the liquidity horizon at time $t \geq 0$. Let $V(t)$ be the value of the portfolio under consideration at time t. We are interested in measuring the change in value of the portfolio within the stochastic liquidity horizon framework. To do this, we consider two different approaches. The first one is the well-known delta-gamma approximation, which assumes that the change in value of the portfolio is a quadratic

 ISBN: 978-84-617-8694-7

function of the change in value of the risk factors. Within the present context of stochastic liquidity horizon, the change in value of the portfolio under the delta-gamma approach is defined as $\Delta V := V(t + H(t)) - V(t)$. To our best knowledge, this is the first time that the delta-gamma approach is considered with an stochastic holding period. The second approach consists of assuming that the value of the portfolio follows a certain stochastic process and we are therefore interested in measuring the change in the log-value of the portfolio rather than in the value itself. Then, we define $X := \ln(V(t + H(t))) - \ln(V(t))$. Let $f_{\Delta V}$ (respectively f_X) be the probability density function (PDF) of ΔV (respectively X) and $F_{\Delta V}$ (respectively F_X) its cumulative distribution function (CDF). If we assume that we short the portfolio, then the right tail of f_X represents losses. Given some confidence level $\alpha \in (0, 1)$, the VaR to measure the risk of holding the portfolio during the stochastic period $H(t)$ is given by the smallest number l such that the probability that the loss X exceeds l is no larger than $1 - \alpha$, where typically $\alpha \geq 0.95$. Formally,

$$\mathrm{VaR}(\alpha) := \inf\{l \in \mathbb{R} : \mathbb{P}(X > l) \leq 1 - \alpha\} = \inf\{l \in \mathbb{R} : F_X(l) \geq \alpha\}. \tag{1}$$

By definition, ES is related to VaR by,

$$\mathrm{ES}(\alpha) := \frac{1}{1-\alpha} \int_\alpha^1 \mathrm{VaR}(u)\, du.$$

Instead of fixing a particular level α, we average VaR over all levels $u \geq \alpha$. Obviously, $\mathrm{ES}(\alpha)$ depends only on the distribution of X, and $\mathrm{ES}(\alpha) \geq \mathrm{VaR}(\alpha)$. For continuous loss distributions an even more intuitive expression can be derived that shows that ES can be interpreted as the expected loss that is incurred in the event that VaR is exceeded. For an integrable loss X with continuous distribution function F_X and for any $\alpha \in (0, 1)$ we have,

$$\mathrm{ES}(\alpha) = \mathbb{E}(X | X \geq \mathrm{VaR}(\alpha)),$$

or, in integral form,

$$\mathrm{ES}(\alpha) = \frac{1}{1-\alpha} \int_{\mathrm{VaR}(\alpha)}^{+\infty} x f_X(x)\, dx. \tag{2}$$

Note that when we work under the delta-gamma approach, then we replace X, F_X, f_X by $\Delta V, F_{\Delta V}, f_{\Delta V}$ in (1) and (2). It is worth remarking that ES is a coherent measure of risk, satisfying in particular, the axiom of sub-additivity in line with the concept of diversification.

In practice, analytical expressions are not available, and Monte Carlo (MC) simulation is often used to compute the risk measures, being the main drawback the computational effort. From this point of view, the situation worsens when we consider a stochastic liquidity horizon $H(t)$, since an extra source of randomness is introduced and must be simulated as well. For this reason, there is an increasing interest in looking for alternative and more efficient methods. Here we propose the SWIFT method, which was originally developed for

 ISBN: 978-84-617-8694-7

European options pricing in [6]. The SWIFT method gives us an accurate and extremely fast recovery of the density function and we give a prescription on how to select the parameters appearing in the numerical method. All these features make our proposal efficient, robust and reliable for practical implementations.

Acknowledgements

The research leading to these results has received funding from La Caixa Foundation. G. C.-P. acknowledges AGAUR-Generalitat de Catalunya for funding under its doctoral scholarship programme. G. C.-P. and L. O.-G. acknowledge the Spanish Ministry of Economy and Competitiveness (MINECO) for funding under grant MTM2013-40782-P.

References

[1] Basel Committee on Banking Supervision, *Fundamental review of the trading book*, Bank for International Settlements (2012).

[2] Basel Committee on Banking Supervision, *Fundamental review of the trading book: a revised market risk framework*, Bank for International Settlements (2013).

[3] Basel Committee on Banking Supervision, *Minimum capital requirements for market risk*, Bank for International Settlements (2016).

[4] D. Brigo and C. Nodio, *A random holding period approach for liquidity-inclusive risk management. In: Innovations in Quantitative Risk Management*, volume 99, Springer Proceedings in Mathematics and Statistics 99, 2015.

[5] G. Colldeforns-Papiol and L. Ortiz-Gracia, *Computation of market risk measures with stochastic liquidity horizon*, Submitted for publication (2016).

[6] L. Ortiz-Gracia and C.W. Oosterlee, *A highly efficient Shannon wavelet inverse Fourier technique for pricing European options*, SIAM Journal on Scientific Computing **38** (2016) B18–B143.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

The Neumann Problem for Bending of Elastic Plates

Christian Constanda[1] **and Dale Doty**[1]

[1] *Department of Mathematics, The University of Tulsa*

emails: christian-constanda@utulsa.edu, dale-doty@utulsa.edu

Abstract

A generalized Fourier series method is constructed to approximate the solution of the Neumann problem in a finite domain for the system of equations governing the bending of elastic plates with transverse shear deformation. The method is illustrated by an example with computation performed by four different techniques that are contrasted and compared for efficiency, accuracy, and stability.

Key words: elastic plates, Neumann problem, generalized Fourier series

Solutions of boundary value problems for an elliptic mathematical model can be approximated by expanding them in a complete set of functions in a conveniently chosen space—for example, L^2. The technique becomes more user-friendly when these functions are closely connected with the structure of the layer potentials constructed for the problem. Computation based on such expansions, however, is hampered by the fact that the Gram–Schmidt process used to orthonormalize the selected set is numerically unstable and, to our knowledge, has not been properly brought under control, particularly in the case of Neumann-type boundary conditions. In what follows, we aim to indicate a procedure that circumvents this obstacle and leads to excellent numerical results to within a prescribed accuracy. The technique is illustrated in application to an interior Neumann problem for the system governing the bending of elastic plates with transverse shear deformation.

Let S be a finite domain in $\mathbb{R}^2$ bounded by a simple, closed, C^2-curve ∂S, let x and y be generic points in S or on ∂S, and let $h_0 = \text{const} > 0,\ h_0 \ll \text{diam}\, S$. We assume that the three-dimensional region $(S \cup \partial S) \times [-h_0/2, h_0/2]$ is occupied by a homogeneous and isotropic elastic material with Lamé constants λ and μ, and write $h^2 = h_0^2/12$.

The Neumann problem for the process of bending is

$$A(\partial_1, \partial_2)u(x) = 0,\ \ x \in S, \quad T(\partial_1, \partial_2)u(x) = \mathcal{Q}(x),\ \ x \in \partial S, \tag{1}$$

ISBN: 978-84-617-8694-7

where the partial differential operators A and T are

$$A(\partial_1,\partial_2)=\begin{pmatrix} h^2\mu\Delta+h^2(\lambda+\mu)\partial_1^2-\mu & h^2(\lambda+\mu)\partial_1\partial_2 & -\mu\partial_1\\ h^2(\lambda+\mu)\partial_1\partial_2 & h^2\mu\Delta+h^2(\lambda+\mu)\partial_2^2-\mu & -\mu\partial_2\\ \mu\partial_1 & \mu\partial_2 & \mu\Delta \end{pmatrix},$$

$$T(\partial_1,\partial_2)=\begin{pmatrix} h^2(\lambda+2\mu)\nu_1\partial_1+h^2\mu\nu_2\partial_2 & h^2\mu\nu_2\partial_1+h^2\lambda\nu_1\partial_2 & 0\\ h^2\lambda\nu_2\partial_1+h^2\mu\nu_1\partial_2 & h^2\mu\nu_1\partial_1+h^2(\lambda+2\mu)\nu_2\partial_2 & 0\\ \mu\nu_1 & \mu\nu_2 & \mu(\nu_1\partial_1+\nu_2\partial_2) \end{pmatrix},$$

$u=(u_1,u_2,u_3)^{\mathrm{T}}$ is a vector characterizing the displacements and $\nu=(\nu_1,\nu_2)^{\mathrm{T}}$ is the unit vector of the outward normal to ∂S. We also consider the matrix of fundamental solutions $D(x,y)$ computed in [1] and the associated matrix $P(x,y)=(T(\partial_y)D(y,x))^{\mathrm{T}}$.

Problem (1) is solvable for any $\mathcal{Q}\in C^{(0,\alpha)}(\partial S)$, $\alpha\in(0,1)$, iff

$$\int\limits_{\partial S}(\mathcal{Q}_\alpha-x_\alpha\mathcal{Q}_3)\,ds=0,\ \ \alpha=1,2,\quad \int\limits_{\partial S}\mathcal{Q}_3\,ds=0. \tag{2}$$

The solution $u\in C^2(S)\cap C^1(S\cup\partial S)$ is unique up to a rigid displacement [1].

According to Somigliana's representation formula,

$$\phi(x)u(x)=-\int\limits_{\partial S}P(x,y)\rho(y)\,ds(y)+\mathcal{L}(x),\quad x\in\mathbb{R}^2, \tag{3}$$

where

$$\phi(x)=\begin{cases}1, & x\in S,\\ 1/2, & x\in\partial S,\\ 0, & x\in\mathbb{R}^2\setminus(S\cup\partial S),\end{cases}$$

ρ is the trace of u on ∂S, and

$$\mathcal{L}(x)=\int\limits_{\partial S}D(x,y)\mathcal{Q}(y)\,ds(y),\quad x\in\mathbb{R}^2. \tag{4}$$

Let S be a disk of radius r centered at origin, let ∂S_* be a circle of radius $r_*>r$, also centered at the origin, and let $\{x^{(k)}\}_{k=1}^{\infty}$ be a set of points densely distributed on ∂S_*. We construct the vector functions

$$\tau^{(jk)}(x)=T(\partial_x)D^{(j)}(x,x^{(k)}),\quad k=1,2,\dots,$$

where $D^{(j)}$ are the columns of D. The set of these functions, which are the rows of $P(x^{(k)},x)$, augmented with a basis for the space of rigid displacements, is linearly independent on ∂S

 ISBN: 978-84-617-8694-7

and complete in $L^2(\partial S)$. We re-order the sequence $\{\tau^{(11)}, \tau^{(21)}, \tau^{(31)}, \tau^{(12)}, \tau^{(22)}, \tau^{(32)}, \ldots\}$ as $\{\tau^{(1)}, \tau^{(2)}, \tau^{(3)}, \tau^{(4)}, \tau^{(5)}, \tau^{(6)}, \ldots\}$ and orthonormalize it to obtain a sequence $\{\eta^{(k)}\}_{k=1}^{\infty}$.

Seeking an approximation of the unknown vector function ρ in the form

$$\rho^{(n)} = \sum_{k=1}^{n} \langle \eta^{(k)}, \rho \rangle \eta^{(k)} = \sum_{k=1}^{n} \sum_{i=1}^{k} \tilde{r}_{ik} \langle \tau^{(i)}, \rho \rangle \left(\sum_{j=1}^{k} \tilde{r}_{jk} \tau^{(j)} \right), \quad n = 1, 2, \ldots, \tag{5}$$

where the $\tilde{r}_{ik}$ are the orthonormalization coefficients and $\langle \cdot, \cdot \rangle$ is the inner product on $L^2(\partial S)$, we use (3) to construct (up to an arbitrary rigid displacement) the approximate solution

$$u^{(n)}(x) = -\int_{\partial S} P(x, y) \rho^{(n)}(y)\, ds(y) + \mathcal{L}(x), \quad x \in S. \tag{6}$$

From (3) with $x \in \mathbb{R}^2 \setminus (S \cup \partial S)$ replaced by $x^{(k)}$ it follows that

$$\int_{\partial S} P(x^{(k)}, x) \rho(x)\, ds(x) = \mathcal{L}(x^{(k)}),$$

so, given the definitions of P and $\tau^{(jk)}$, we have $\langle \tau^{(jk)}, \rho \rangle = \mathcal{L}_j(x^{(k)})$ for $j = 1, 2, 3$ and $k = 1, 2, \ldots$, or, after re-indexing,

$$\langle \tau^{(3(k-1)+j)}, \rho \rangle = \mathcal{L}_j(x^{(k)}), \quad j = 1, 2, 3, \quad k = 1, 2, \ldots. \tag{7}$$

In conclusion, $u^{(n)}$ is given by (6), with $\rho^{(n)}$ computed from (5), $\langle \tau^{(i)}, \rho \rangle$ from (7), and $\mathcal{L}_j(x^{(k)})$ from (4). The sequence $u^{(n)}$ converges uniformly in the L^2-norm on any closed subdomain of S to a solution u of the problem. This solution may contain a specific rigid displacement, which can be easily identified.

As an illustration of the method, we choose $\lambda = \mu = 1$, $h = 0.5$, $r = 1$, $r_* = 2$,

$$\{x^{(1)}, x^{(2)}, x^{(3)}, \ldots\}_{\text{Cartesian}} = \{(2, 0), (2, \pi), (2, \pi/2), (2, 3\pi/2), (2, \pi/4), \ldots\}_{\text{Polar}},$$
$$\mathcal{Q}(x)|_{x_1 = \cos\theta,\, x_2 = \sin\theta} = \big(6\cos\theta + 3\cos(3\theta) - 20\cos(4\theta),$$
$$- 6\sin\theta + 3\sin(3\theta) + 20\sin(4\theta),\, 18\cos(2\theta)\big)^{\mathrm{T}}.$$

It is easy to verify that $\mathcal{Q}$ satisfies the solvability conditions (2).

The set $\{\tau^{(k)}\}_{k=1}^{\infty}$ is orthonormalized by three different procedures: the classical Gram–Schmidt (CGS), the modified Gram–Schmidt (MGS), and the Householder reflections (HR) (see [2] and [3]). We also use a fourth technique, which, based on row reduction, obviates the need for orthonormalization. Below are the results obtained by the MGS method.

Fig. 1 displays the three components of $u^{(150)}$ in S, and the approximate computed traces of these components on ∂S. Fig. 2 shows the corresponding computational errors. The curves on the left in Fig. 3 are the graphs (in terms of the polar angle) of the computed

 ISBN: 978-84-617-8694-7

traces of the components of $u^{(150)}$ on ∂S. The graph on the right in Fig. 3 (drawn in a logarithmic scale) indicates the exponential convergence in the L^2-norm of the sequence of approximations to the exact solution as the number of the sample points $x^{(k)}$ increases.

The solution u constructed in this example contains no rigid displacements.

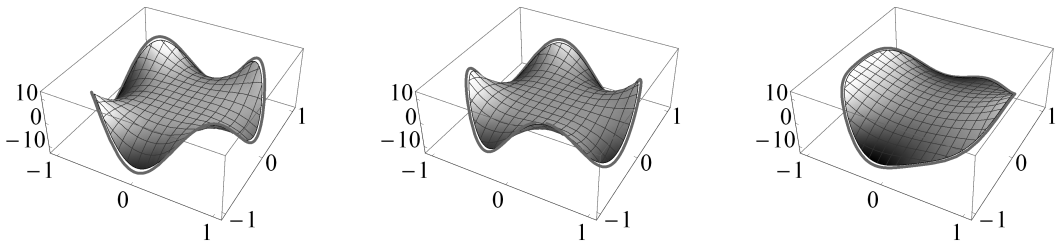


Figure 1

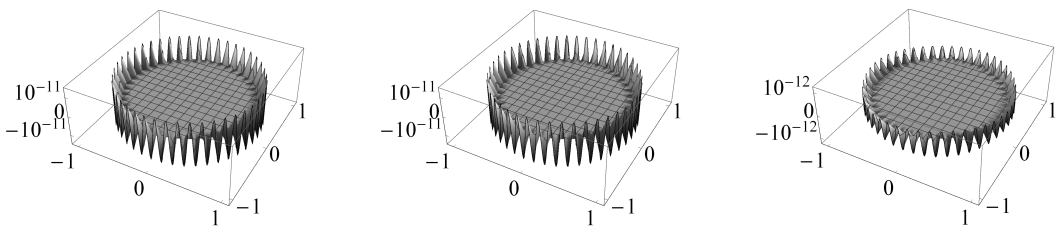


Figure 2

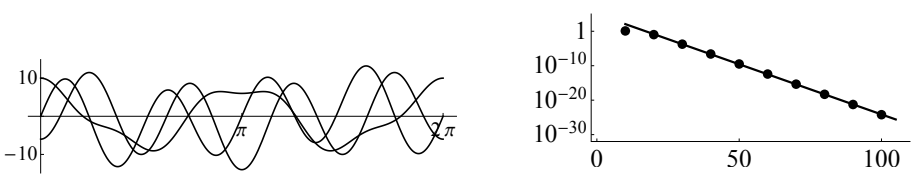


Figure 3

References

[1] C. Constanda, *Mathematical Methods for Elastic Plates*, Springer, London, 2014.

[2] L. N. Trefethen, *Householder triangularization of a quasimatrix*, IMA J. Numer. Anal. **30** (2010) 887–897.

[3] A. Gorodetsky, S. Karaman and Y. Marzouk, *Function–train: a continuous analogue of the tensor–train decomposition*, preprint arXiv: 1510.09088 (2015).

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

Stability study of a parametric class of iterative methods for solving nonlinear models

Alicia Cordero[1], Lucía Guasp[1] and Juan R. Torregrosa[1]

[1] *Instituto de Matemáticas Multidisciplinar, Universitat Politècnica de València*

emails: acordero@mat.upv.es, luguaal@ade.upv.es, jrtorre@mat.upv.es

Abstract

The dynamical analysis of iterative methods without memory for solving nonlinear models, by using complex dynamics tools, is a very useful technique to study their stability and reliability. A parametric family of order four, applied on quadratic polynomials, gives us a rational operator whose dynamics is studied. The stability of its fixed points, in terms of the value of the parameter, its critical points and their associated parameter planes, etc. give us important information about which members of the family have good properties of stability and whether in any of them appear chaos in the iterative process.

Key words: Iterative methods, nonlinear equations, stability, parameter plane, basin of attraction, rational operator

1 Introduction

The design of fixed point iterative schemes for solving equations and systems of nonlinear equations is an important and challenging task in the field of Numerical Analysis. To find the solution x^* of a nonlinear equation $f(x) = 0$, $f : I \subseteq \mathbb{R} \to \mathbb{R}$, or a nonlinear system $F(x) = 0$, $F : D \subseteq \mathbb{R}^n \to \mathbb{R}^n$, is a classical and difficult problem with many applications in Science and Engineering.

In the last years, many iterative methods have been constructed for solving these type of problems, see for example [1] and [2] and the references therein. Many times, the researchers design a parametric family of iterative methods, all of them with the same order of convergence but, all members have the same stability? Does the width of the basins of attraction depend on the value of the parameter?, Which elements of the family have

 ISBN: 978-84-617-8694-7

chaotic behavior?, ... We can found the answer of these and other questions by analyzing the dynamical behavior of the rational operator associated to the iterative method on low degree polynomials.

Our goal in this paper is to carry out a dynamical study of the parametric family of iterative methods designed for solving nonlinear equations $f(x) = 0$. The idea for constructing this class appears in [3]. By using tools of complex dynamics we analyze the stability of the fixed points of the rational operator that appears when our family is applied on an arbitrary second degree polynomial. The parameter plane associated to each critical point gives us important information about the stability of the elements of the family and which of them have unstable behavior.

Now, we are going to recall some dynamical concepts of complex dynamics that we use in this work. Given a rational function $R : \hat{\mathbb{C}} \to \hat{\mathbb{C}}$, where $\hat{\mathbb{C}}$ is the Riemann sphere, the *orbit of a point* $z_0 \in \hat{\mathbb{C}}$ is defined as:

$$\{z_0,\, R(z_0),\, R^2(z_0), \ldots, R^n(z_0), \ldots\}.$$

We analyze the phase plane of the map R by classifying the starting points from the asymptotic behavior of their orbits. A $z_0 \in \hat{\mathbb{C}}$ is called a *fixed point* if $R(z_0) = z_0$. A *periodic point* z_0 of period $p > 1$ is a point such that $R^p(z_0) = z_0$ and $R^k(z_0) \neq z_0$, for $k < p$. A *pre-periodic point* is a point z_0 that is not periodic but there exists a $k > 0$ such that $R^k(z_0)$ is periodic. A *critical point* z_0 is a point where the derivative of the rational function vanishes, $R'(z_0) = 0$. Moreover, a fixed point z_0 is called *attractor* if $|R'(z_0)| < 1$, *superattractor* if $|R'(z_0)| = 0$, *repulsor* if $|R'(z_0)| > 1$ and *parabolic* if $|R'(z_0)| = 1$.

The basin of attraction of an attractor α is defined as:

$$\mathcal{A}(\alpha) = \{z_0 \in \hat{\mathbb{C}} \,:\, R^n(z_0) \to \alpha,\ n \to \infty\}.$$

The *immediate basin of attraction* of an attractor is the connected component of its basin of attraction that holds the attractor.

The *Fatou set* of the rational function R, $\mathcal{F}(R)$, is the set of points $z \in \hat{\mathbb{C}}$ whose orbits tend to an attractor (fixed point, periodic orbit or infinity). Its complement in $\hat{\mathbb{C}}$ is the *Julia set*, $\mathcal{J}(R)$. That means that the basin of attraction of any fixed point belongs to the Fatou set and the boundaries of these basins of attraction belong to the Julia set.

The following theorem establishes a classical result of Fatou and Julia that we use in the study of parameter space associated to the family.

Theorem 1 *Let R be a rational function. The immediate basin of attraction of an attracting fixed or periodic point holds, at least, a critical point.*

By using this result, one can be sure to find all the stable behavior associated to a rational function R, by analyzing the performance of R on the set of critical points.

 ISBN: 978-84-617-8694-7

1.1 The parametric family

By adding a new step to Newton's method, we construct the following two-step scheme

$$\begin{aligned} y_k &= x_k - \frac{f(x_k)}{f'(x_k)}, \\ x_{k+1} &= y_k - \left(\alpha_1 + \alpha_2 \frac{f'(x_k)}{f'(y_k)} + \alpha_3 \left(\frac{f'(x_k)}{f'(y_k)}\right)^2\right) \frac{f(y_k)}{f'(y_k)}, \quad k = 0, 1, \ldots, \end{aligned} \tag{1}$$

where α_1, α_2 and α_3 are free parameters.

The following result establishes the convergence of family (1).

Theorem 2 *Let $f : I \subseteq \mathbb{R} \to \mathbb{R}$ be a sufficiently differentiable function in an open interval I and $x^* \in I$ a root of equation $f(x) = 0$. Choosing an initial approximation x_0 close enough to x^*, the iterative scheme defined by (1) has fourth-order convergence when $\alpha_2 = 3 - 2\alpha_1$ and $\alpha_3 = -2 + \alpha_1$, being α_1 a free parameter. In particular, if $\alpha_1 = \frac{5}{4}$ then method (1) has order five.*

Which is the relation between the values of the free parameter α_1 and the stability of the corresponding iterative method? Under the point of view of complex dynamics, we are going to study the general convergence of the families on quadratic polynomials. It is known that, if the iterative method satisfies the scaling theorem (and family (1) doesw), the roots of a polynomial can be transformed by an affine map with no qualitative changes on the dynamics of the family. So, we can use a generic quadratic polynomial $p(z) = (z-a)(z-b)$. The rational operator obtained when family (1) is applied on $p(z)$ has the expression:

$$\begin{aligned} T_{p,\alpha_1,a,b}(z) &= \frac{(a-z)(b-z)}{a+b-2z} + z \\ &+(a-z)^2(b-z)^2 \left[\frac{(a^4+b^4-4a^3z-4b^3z+4(1+\alpha_1)b^2z^2-8\alpha_1 bz^3}{(a+b-2z)(a^2+b^2-2az-2bz+2z^2)^3}\right. \\ &+\frac{4\alpha_1 z^4 - 4az((-1+2\alpha_1)b^2+(2-4\alpha_1)bz+2\alpha_1 z^2)}{(a+b-2z)(a^2+b^2-2az-2bz+2z^2)^3} \\ &\left.+\frac{a^2((-2+4\alpha_1)b^2+(4-8\alpha_1)bz+4(1+\alpha_1)z^2))}{(a+b-2z)(a^2+b^2-2az-2bz+2z^2)^3}\right], \end{aligned}$$

depending on parameter α_1 and also on the roots of the polynomial a and b.

Blanchard in [4] considered the conjugacy map $h(z) = \dfrac{z-a}{z-b}$, (a Möbius transformation) with the following properties:

$$\text{i)}\ \ h(\infty) = 1, \qquad \text{ii)}\ \ h(a) = 0, \qquad \text{iii)}\ \ h(b) = \infty,$$

and proved that, for quadratic polynomials, Newton's operator is conjugate to the rational map z^2. In an analogous way, operator $T_{p,\alpha_1,a,b}(z)$ on quadratic polynomials is conjugated to operator $O_{\alpha_1}(z)$,

 ISBN: 978-84-617-8694-7

$$O_{\alpha_1}(z) = \left(h \circ T_{p,\alpha_1,a,b} \circ h^{-1}\right)(z) = -z^4 \frac{5 - 4\alpha_1 + 2z^2 + z^4}{-1 - 2z^2 + -5z^4 + 4\alpha_1 z^4}, \tag{2}$$

We observe that parameters a and b have been obviated in $O_{\alpha_1}(z)$.

In the following sections we study the fixed and critical points of operator $O_{\alpha_1}(z)$, the stability of the fixed points, the parameter plane associated to the family and some dynamical planes describing different behavior: stability, periodic orbits, ...

2 Analysis of the fixed and critical points

Firstly, we study the fixed points of the rational function $O_{\alpha_1}(z)$ that are not related with the original roots of the polynomial $p(z)$ (called *strange fixed points*), and the *free critical points*, that is, the critical points of $O_{\alpha_1}(z)$ different from 0 and ∞, which are associated to the roots of $p(z)$.

Fixed points of $O_{\alpha_1}(z)$ are the roots of equation $O_{\alpha_1}(z) = z$, that is, $z = 0$, $z = \infty$ and the strange fixed points $ex_1(\alpha_1) = 1$ and the roots of the polynomial

$$r(\alpha_1, z) = 1 + z + 3z^2 + (-2 + 4\alpha_1)z^3 + 3z^4 + z^5 + z^6,$$

that are denoted by $ex_i(\alpha_1)$, $i = 2, 3, 4, 5, 6, 7$.

Therefore, there are seven strange fixed points, except in the following cases:

i) If $\alpha_1 = 1$, then the operator is $O_1(z) = z^4$, so there are no strange fixed points.

ii) If $\alpha_1 = 2$, then the operator is $O_2(z) = -z^4 \frac{3+z^2}{1+3z^2}$. There are only six strange fixed points as $ex_2(\alpha_1) = ex_3(\alpha_1) = -1$.

iii) If $\alpha_1 = -2$, there are only five strange fixed point as $ex_2(\alpha_1) = ex_3(\alpha_1) = 1$.

On the other hand, in order to determine the critical points, we calculate the first derivative of $O_{\alpha_1}(z)$:

$$O'_{\alpha_1}(z) = -4z^3 \frac{(1+z^2)^2(-5 + 4\alpha_1 + 2(1 - 2\alpha_1)z^2 + (-5 + 4\alpha_1)z^4)}{(1 + 2z^2 + (5 - 4\alpha_1)z^4)^2}.$$

As we have said, a classical result establishes that there is, at least, one critical point associated with each invariant Fatou component. Due to the order of convergence of the methods under study, it is clear that $z = 0$ and $z = \infty$ (related to the roots of the polynomial by means of Möbius map) are critical points and give rise to their respective Fatou components, but there exist in the family some free critical points, some of them depending on the value of the parameter.

Proposition 1 *By analyzing the equation $O'_{\alpha_1}(z) = 0$ of the family, we obtain:*

 ISBN: 978-84-617-8694-7

a) If $\alpha_1 = 1$ there is no free critical points of operator $O_{\alpha_1}(z)$.

b) If $\alpha_1 = 2$ or $\alpha_1 = \frac{5}{4}$, then $z = -i$ and $z = i$ are the only free critical points.

c) In any other case,

$$cr_1(\alpha_1) = -i,$$

$$cr_2(\alpha_1) = i,$$

$$cr_3(\alpha_1) = -\sqrt{\frac{1 - 2\alpha_1 + 2\sqrt{3}\sqrt{-2 + 3\alpha_1 - \alpha_1^2}}{5 - 4\alpha_1}},$$

$$cr_4(\alpha_1) = \sqrt{\frac{1 - 2\alpha_1 + 2\sqrt{3}\sqrt{-2 + 3\alpha_1 - \alpha_1^2}}{5 - 4\alpha_1}},$$

$$cr_5(\alpha_1) = -\sqrt{\frac{-1 + 2\alpha_1 + 2\sqrt{3}\sqrt{-2 + 3\alpha_1 - \alpha_1^2}}{-5 + 4\alpha_1}},$$

and

$$cr_6(\alpha_1) = \sqrt{\frac{-1 + 2\alpha_1 + 2\sqrt{3}\sqrt{-2 + 3\alpha_1 - \alpha_1^2}}{-5 + 4\alpha_1}},$$

are free critical points.

Let us remark that $cr_1(\alpha_1)$ and $cr_2(\alpha_1)$ are pre-images of $z = 1$ and $cr_3(\alpha_1)$ and $cr_5(\alpha_1)$ are conjugated, as well as $cr_4(\alpha_1)$ and $cr_6(\alpha_1)$. Therefore, we only have two independent free critical points.

3 Stability of the fixed points

Of course, $z = 0$ and $z = \infty$ are superattracting fixed points but, which is the character of the rest of fixed points? The following results answer this question.

The character of the strange fixed point $ex_1(\alpha_1) = 1$ of the family, $\alpha_1 \neq 2$, is as follows:

i) If $|\alpha_1 - 2| > 4$, then $ex_1(\alpha_1) = 1$ is an attractor.

ii) When $|\alpha_1 - 2| = 4$, $ex_1(\alpha_1) = 1$ is a parabolic point.

 ISBN: 978-84-617-8694-7

iii) If $|\alpha_1 - 2| < 4$, then $ex_1(\alpha_1) = 1$ is a repulsor.

The analysis of the stability of strange fixed points $ex_i(\alpha_1), i = 2, 3, 4$ shows that they are repulsors for any value of the parameter. Finally, the character of the strange fixed points $ex_i(\alpha_1), i = 5, 6$ is as follows:

i) If $|\alpha_1 + 1| < 1$, then both points are attractors.

ii) When $|\alpha_1 + 1| = 1$, $ex_5(\alpha_1)$ and $ex_6(\alpha_1)$ are parabolic points.

iii) If $|\alpha_1 + 1| > 1$, then both points are repulsors.

In Figure 1, we represent the stability regions of all strange fixed points $ex_i(\alpha_1)$, $i = 1, 2, 3, 4, 5, 6$.

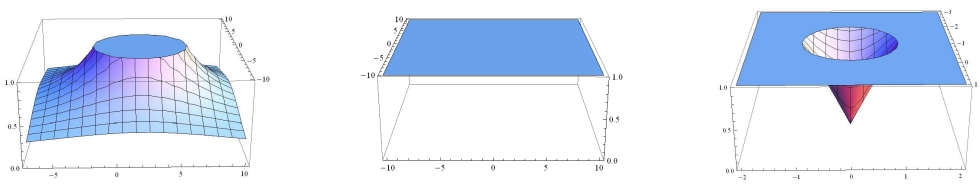


Figure 1: Stability regions of $ex_1(\alpha_1)$ (left),$ex_i(\alpha_1)$, $i = 2, 3, 4$ (middle) and $ex_i(\alpha_1)$, $i = 5, 6$ (right).

4 The parameter space

The parameter space associated with an independent free critical point of operator is obtained by associating each point of the complex plane with a value of α_1, i.e., with an element of family. Every value of the parameter belonging to the same connected component of the parameter space gives rise to subsets of schemes of the family with similar dynamical behavior. So, it is interesting to find regions of the parameter plane as much stable as possible, because these values of the parameter will give us the best members of the family in terms of numerical stability.

When we consider the independent free critical points of operator $O_{\alpha_1}(z)$ as a starting point of the iterative scheme of the family associated to each complex value of α_1, we paint this point of the complex plane in red if the method converges to any of the roots (zero and infinity) and they are black in other cases. The color used is brighter when the number of iterations is lower. Then, the parameter plane P_1 is obtained. A mesh of 500×500 points has been used, 250 has been the maximum number of iterations involved and 10^{-3} the tolerance used as a stopping criterium.

 ISBN: 978-84-617-8694-7

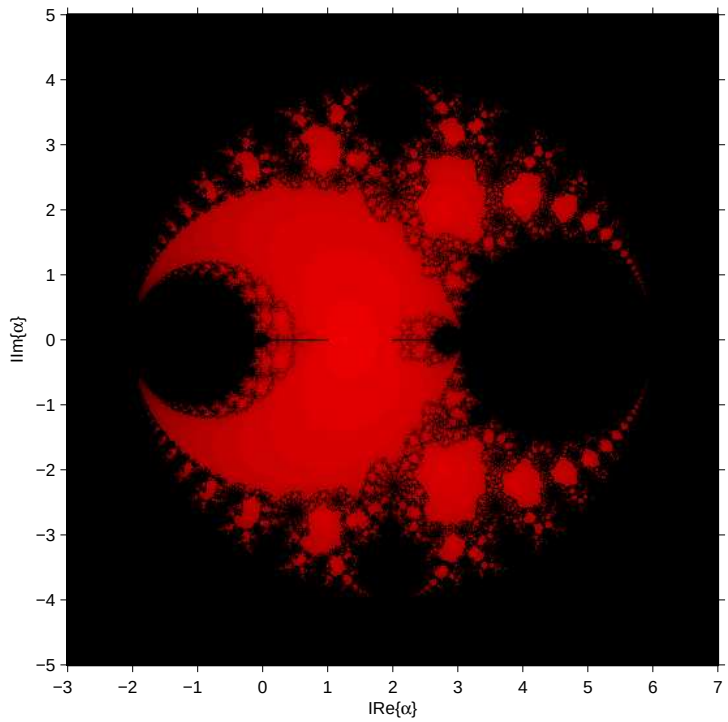


Figure 2: Parameter plane P_1 associated to $cr_i(\alpha_1)$, $i = 3, 4, 5, 6$

We obtain an only parameter plane due to the fact that $cr_4(\alpha_1)$ is equal in module to $cr_6(\alpha_1)$ and the operator's powers are even numbers. We can observe that the best real values of the parameter α_1 are between 1 and 2.

4.1 Dynamical Planes

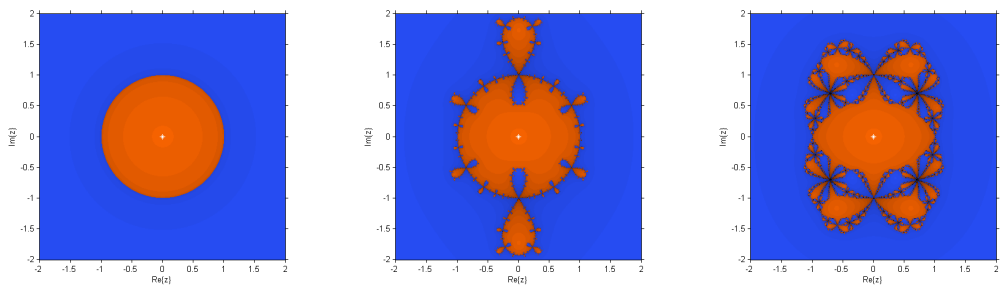
In this section we show, by means of dynamical planes, the qualitative behavior of the different elements of family. We select these elements by using the conclusions obtained by analyzing the parameter plane of the family and the stability analysis made on fixed points.

The dynamical plane associated to a value of the parameter, that is, obtained by iterating an element of family, is generated by using each point of the complex plane as initial estimation (we have used a mesh of 400×400 points). We paint in blue the points whose orbit converges to infinity, in orange the points converging to zero (with a tolerance of 10^{-3}), in other colors (green, red, etc.) those points whose orbit converges to one of the strange fixed points (all fixed points appear marked as a white star in the figures) and in black if it

 ISBN: 978-84-617-8694-7

reaches the maximum number of 40 iterations without converging to any of the fixed points.

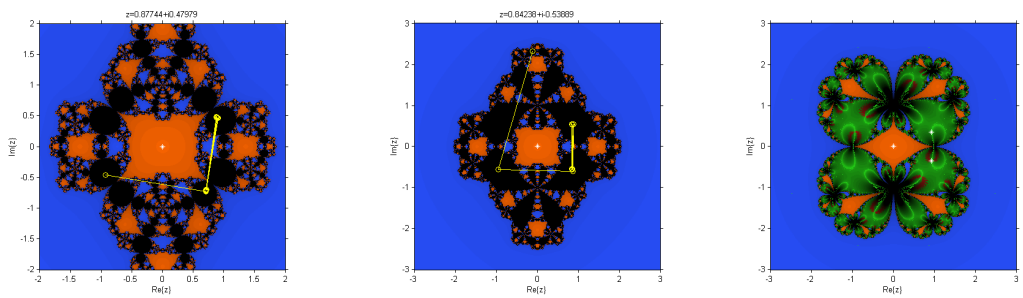
There are some regions in parameter spaces whose corresponding iterative methods have good numerical behavior, in terms of stability and efficiency. They correspond to values of the parameter painted in red. In Figure 3 we show the dynamical planes corresponding to stable values of the parameter, specifically $\alpha_1 = 1$, $\alpha_1 = 2$ and $\alpha_1 = 0.5$.



(a) $\alpha_1 = 1$ (b) $\alpha_1 = 2$ (c) $\alpha_1 = 0.5$

Figure 3: Some dynamical planes with stable behavior

On the other hand, unstable behavior is found when we choose values of α_1 in the black region of parameter plane. In Figure 4, dynamical planes corresponding to values of parameter $\alpha_1 = 3$, $\alpha_1 = 3.5$ and $\alpha_1 = -1.5$ are presented.



(a) $\alpha_1 = 3$ (b) $\alpha_1 = 3.5$ (c) $\alpha_1 = -1.5$

Figure 4: Dynamical planes with unstable behavior

In Figure 4a and 4b we can observe periodic orbits of period two, while in Figure 4c four basins of attraction appear, two of them corresponding to the roots of $p(z)$ and the other ones are the basins of attraction of the strange fixed points $ex_i(\alpha_1)$, $i = 5, 6$.

 ISBN: 978-84-617-8694-7

Acknowledgements

This work has been partially supported by Ministerio de Economía y Competitividad, MTM2014-52016-C2-2-P, and Generalitat Valenciana PROMETEO/2016/089.

References

[1] S. Amat and S. Busquier, *Advances in iterative methods for nonlinear equations*, Springer, 2016.

[2] M.S. Petković, B. Neta, L.D. Petković and J. Džunić, *Multipoint methods for solving nonlinear equations*, Elsevier, Amsterdam, 2013.

[3] A. Cordero, E. Gómez and J.R. Torregrosa, *Efficient high-order iterative methods for solving nonlinear systems and their application on heat conduction problems*, Complexity Volume 2017, Artivle ID 6457532, 11 pages.

[4] P. Blanchard, *The dynamics of Newton's method*, Proc. Symp. Appl. Math. **49** (1994) 139–154.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

Mathematical model for predicting the biomass growth of *Mytilus chilensis* (Hupe 1954) in suspension cultures

F. Córdova-Lepe[1], K. Vilches [11], B. Martel [2 2] and H. Plaza [3 2]

[1] *Departamento de Matemática, Física y Estadística, Universidad Católica del Maule, Facultad de Ciencias Básicas*

[2] *Departmento de Investigación y Desarrollo, Fishing Partners Ltda.*

emails: fcordova@ucm.cl, kvilches@ucm.cl, bmartel@fishingpartners.cl , hplaza@fishingpartners.cl

Abstract

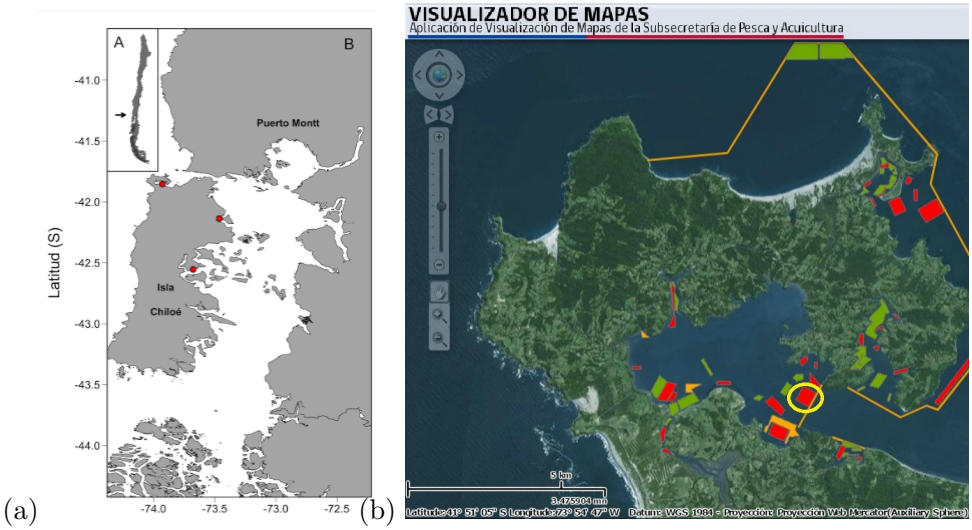
A growth mathematical model (individual-population) of biomass for *Mytilus chilensis* is introduced. This incorporates the energy variable when assuming a predation component by means of the filtering. We will assume that the prey is the phytoplankton one and the predators are the filtering organisms. A new generalized Monod type biomass gain function is considered. Some simulations based in biogeographic data are included.

Key words: biomass curve, Mytilus chilensis, mathematical model
MSC 2000: 92B05

1 Introduction

The aquaculture is a important productive sector in Chilean economy [4]. Chile is the third world producer of mollusc with with 78,500 tons of annual production [4]. The Mytiliculture is a part of this productive areas, which is concentrated in the South of Chile in Los Lagos region [4]. The *Mylitus chilensis* (Hupe 1854) is a filtering bivalve of the family Mylidae. This molluscs is widely distributed along the Chilean coast from Arica (18 degrees South) to Cape Horn (56 degrees South) [6] to depths of 25 [m] [1]. Along the latitudinal range of

 ISBN: 978-84-617-8694-7



(a) (b)

Figure 1: (a) Chiloé island. (b) Satellite image of cultures in Chiloé island, Quetalmahue sector.

Chile, the shell-shape differences in the population of *Mytilus chilensis* are highly significant [5, 12] and these are related with the latitude and origin [12].

The main objective of this work is to propose a mathematical model to predict the individual biomass growth (in average) of the *Mytilus chilensis* in suspensions cultures, in a such manner that the consulting Fishing Partners company can estimate the production in each of the growing areas (see figure 1-(b)). In mathematical terms, the main result of this work, it is the formulation of an ordinary differential equations system, which model the relation between individual growth and food, considering a set of environmental and biological parameters.

The relation between *Mytilus chilensis* and fitoplankton is assumed as an interaction prey-predator, supposing that the fouling effect is nule. The spatial-temporal scale was choice including biotics and abiotics factors. In this first approach, we have supposed that the culture area is shared among some producers. In addition, we have assumed that there exists a temporal horizon, which is the average time between sowing and spawning.

The biological and environmental main assumptions are: (i) The volume of water in the culture area is constant. (ii) The phytoplankton growth is external to the culture area. (iii) On the culture area only exists consumption of phytoplankton. (iv) The fouling is neglected. (v) The consumption of phytoplankton is proportional to its concentration. (vi)

[1]Founded by CONICYT PAI/Academia 79150021 2016-2018.
[2]Founded by Portafolio de Innovación Corfo 16GPI723
[3]Founded by Portafolio de Innovación Corfo 16GPI723

 ISBN: 978-84-617-8694-7

Assimilated energy is proportional to consumption. And, (v) the gain of biomass from the potential energy (assimilated less metabolic) is determined by a generalized Monod type function.

The resulting biomass curves (in time) are visualized with the sigmoid shape, which is similar to the individual growth curves obtained by means of statistical methods applied to predict the growth in bivalve, see [3].

2 Model deduction

2.1 Population growth

Let us consider that the culture process presents the following stage: seeding, individual growth and spawning. The model is limited at individual growth stage. We also assume that the number of individuals is only affected by detachment of the rope, since the death implies detachment. Then, denoting by δ_k *the detachment rate* at time k, it follows that

$$\delta_k = \frac{1}{n_k}\frac{n_k - n_{k+1}}{t_{k+1} - t_k}\,, \quad k = 1, \cdots, m-1\,,$$

where n_k is the number of individuals at time k. By now, the relation, in the extremes of $[t, t+\Delta t]$, is given by

$$n(t+\Delta t) = n(t)[1 - \delta\,\Delta t]\,.$$

In conclusion, the daily detachment is proportional to the numbers of individuals. Taking the limit as $\Delta t \to 0$ we obtain the differential equation

$$n'(t) = -\delta\, n(t)\,.$$

Therefore, after a intial time $t_0 > 0$ where the number of individuals is n_0, we have exponential decay given by

$$n(t) = n_0\, \exp\{\delta(t - t_0)\}, \quad \text{for} \quad t \geq t_0,$$

where t_0 is the first time of count.

Denoting by $N(t)$ the total number of individual in the culture area at t day, we can conclude that

$$N(t) = L \cdot n(t) = L\, n_0\, \exp\{\delta(t - t_0)\}, \quad \text{for} \quad t \geq t_0, \tag{1}$$

where L is the length of culture rope.

 ISBN: 978-84-617-8694-7

2.2 Phytoplankton Abundance

Let us denote the total concentration of phytoplankton in the culture area by $C(t)$ [cel/m^3]. We assume that the phytoplankton abundance is affected by the consumption of individuals and by that enters through the flow. We denote by Ω the culture zone defined as a body of water with constant volume $V\,[m^3]$ with an external flux of fresh water $Q[m^3/day]$ assuming that the same amount of water leaves Ω. Let us represent by $C^e[cel/m^3]$ the external flux of phytoplankton entering to Ω. Notice that $A(t)[cel] = C(t)[cel/m^3]\,V[m^3]$, describe the total phytoplankton on Ω at t-day and the variation during the time interval $[t, t+\Delta t]$ is given by,

$$A(t+\Delta t) - A(t) = \left[\left\{ \begin{array}{c} \text{Entering} \\ \text{for day} \end{array} \right\} - \left\{ \begin{array}{c} \text{Consumption} \\ \text{for day} \end{array} \right\} - \left\{ \begin{array}{c} \text{Outgoing} \\ \text{for day} \end{array} \right\} \right] \Delta t\,.$$

We have that,

$$\text{Entering for day} =: C^e[cel/m^3] \cdot Q[m^3/day]$$

and

$$\text{Outgoing for day} =: C(t)[cel/m^3] \cdot Q[m^3/day]\,.$$

Then, taking the limit as $\Delta t \to 0$ we obtain

$$A'(t) = C'(t)\,V = (C^e - C(t))Q - \{\text{consumption for day}\}\ .$$

Therefore,

$$C'(t) = (C^e - C(t))\,D - \frac{1}{V}\{\text{consumption for day}\}. \tag{2}$$

Noticing that $D = (Q/V)[day^{-1}]$ is the *dilution*.

2.3 Phytoplankton consumption

Let us consider that the phytoplankton consumption on Ω is dependent on: (1) The phytoplankton concentration $C(t)$ $[cel/m^3]$. (2) The total numbers of individuals $N(t)[ind]$. And, (3) The biomass $b(t)$ $[gr]$. We also assume that one unity of biomass consumes $F(C)$ $[cel/day]$ of phytoplankton and the fouling consumption is negligible. Thus, we can define $F(\cdot)$ as follows,

$$F(C)[cel/day] = \nu(t)[m^3/day] \cdot C[cel/m^3]\,.$$

where $\nu(t)$ is the number of cubic meters a unit of biomass is capable of filtering in one day and it changes with the age. We define $\nu(t)$ by

$$v(t) = \nu\frac{t}{T}\,, \tag{3}$$

where ν is the maximal consumption $[gr/m^3]$ for 1 [gr] of biomass at spawning time and T is the spawinng time.

ISBN: 978-84-617-8694-7

2.4 Energy assimilation

Let us suppose that one biomass unit of *Mytilus chilensis* transforms its consumption $F(C)[cel/day]$ into Energy represented by $E(C)\,[cal/day]$ and it is assumed proportional to $F(C)[cel/day]$

$$E(C)[cal/day] = q[cal/cel] \cdot F(C)[cel/day]\,,$$

where q denotes the amount of calories containing in one cell of phytoplankton.

2.5 Biomass gain

In [11] the author proposes that the assimilated energy is divided into biomass gain and metabolism maintenance. We denote by $\mathcal{E} = E - M$, which is the difference between the assimilated energy (E) and the one that is spent to maintain its metabolism (M).

Here, we introduce the generalized Monod function,

$$G_r(\mathcal{E}) = G_\infty \left\{ \frac{\mathcal{E}}{\mathcal{E}+a} \right\}^\alpha , \tag{4}$$

where G_∞ is a parameter of maximal efficiency. The metabolims is modeled as follows,

$$M(t) = (M_{max} - M_{min}) \left(1 - \frac{t}{T}\right) + M_{\min},$$

where M_{max} and M_{min} are the maximal and minimal metabolic spent, respectively, given in the literature [10], assuming that the metabolism is increasing respect to the age.

3 The model

The prey-predator interaction between one specie of phytoplankton (prey) and the *Mytilus chilensis* (predator) is modeled by the no linear system, non-autonomous of ordinary differential equations, that follows

$$\left\{ \begin{array}{rcl} C'(t) & = & \{C^e - C(t)\} \cdot D - \frac{L}{V} n_0 e^{-\delta t} \nu(t) C(t) b(t)\,, \\ \\ b'(t) & = & G_\infty \left\{ \frac{qM(t)\nu(t)C(t)}{qM(t)\nu(t)C(t)+a} \right\}^\alpha b(t)\,, \end{array} \right. \tag{5}$$

where $C(t)$ is the phytoplankton concentration $[gr/m^3]$ at time t [day] and $b(t)$ is the individual biomass of one *Mytilus chilensis* $[gr]$ at time t [day].

 ISBN: 978-84-617-8694-7

4 Simulations

(a)

Parameter	value
C_e	5
V	3419514, 25
Q	162, 18
L	150
N	491
δ	0, 0008
q	4, 618
T	400
G	0, 025
a	0, 7
α	0, 5
ν	2, 5
M_{min}	0, 25
M_{max}	0, 55

(b)

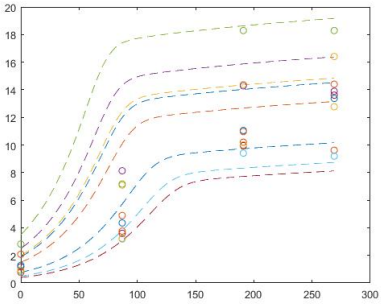
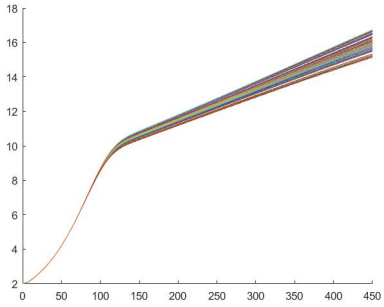


Figure 2: (a) Simulated parameters values. (b) Simulations (- -) Real data of Quetalmahue (o) (2015).

(a)



(b)

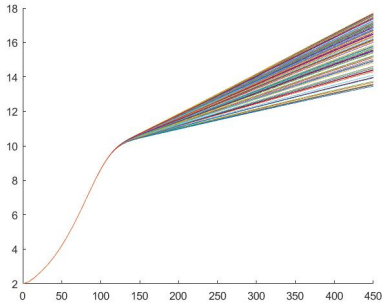


Figure 3: (a) Sensibility at parameter δ. (b) Sensibility at parameter Q.

ISBN: 978-84-617-8694-7

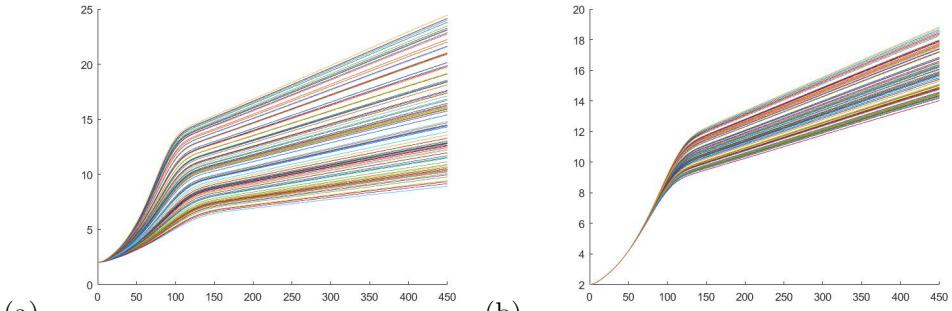


Figure 4: (a) Sensibility at parameter G. (b) Sensibility at parameter N.

References

[1] Brattstrm, H., & Johanssen, A. *Ecological and regional zoogeography of the marine benthic fauna of Chile Report no. 49 of the Lund University Chile Expedition 194849.*Sarsia, **68(4)**, (1983) 289-339.

[2] Díaz E. *Mitílidos en la regin de Los Lagos. Condiciones de trabajo en la industria del chorito.* Cuaderno de Investigacin No. 38. Dirección del Trabajo. (2010) Gobierno de Chile.

[3] Hawkins, A. J. S., Duarte, P., Fang, J. G., Pascoe, P. L., Zhang, J. H., Zhang, X. L., & Zhu, M. Y. *A functional model of responsive suspension-feeding and growth in bivalve shellfish, configured and validated for the scallop Chlamys farreri during culture in China.*Journal of Experimental Marine Biology and Ecology **281(1)** (2002) 13-40.

[4] Uriarte, I.*Estado actual del cultivo de moluscos bivalvos en Chile.* FAO Actas de Pesca y Acuicultura. (2008)

[5] Krapivka, S., Toro, J. E., Alcapán, A. C., Astorga, M., Presa, P., Pérez, M., & Guiñez, R.*Shellshape variation along the latitudinal range of the Chilean blue mussel Mytilus chilensis (Hupe 1854).*Aquaculture Research, **38(16)**, (2007) 1770-1777.

[6] Lancellotti, D. A., & Vásquez, J. A.*Zoogeografía de macroinvertebrados bentónicos de la costa de Chile: contribución para la conservacin marina.* Revista chilena de historia natural, **73(1)**, (2000) 99-129.

[7] Marambio J, Maturana S, Campos B. *Modelo dinmico de crecimiento de la biomasa para Mytilus chilensis en sistemas de cultivo en líneas.* Rev. Biol. Mar Ocenog. **47**, (2012)51-64.

 ISBN: 978-84-617-8694-7

[8] Molinet C., Navarro J., Marín S., Daz M., Seguel M., Niklitschek E., Olivares G., Toledo P., Rosales S. Evaluación de los factores limitantes en el desarrollo de cultivos de mitílidos, para anlisis de capacidad de carga, X Región de Los Lagos**(1ra Etapa)** (2013).

[9] Molinet C., Navarro J., Marín S., Díaz M., Seguel M., Niklitschek E., Olivares G., Toledo P., Rosales S. *Evaluación de los factores limitantes en el desarrollo de cultivos de mitlidos, para anlisis de capacidad de carga, X Región de Los Lagos* **(2da Etapa)**(2014).

[10] Navarro, J. M., & Winter, J. E. *Ingestion rate, assimilation efficiency and energy balance in Mytilus chilensis in relation to body size and different algal concentrations.* Marine Biology, **67(3)**, (1982) 255-266.

[11] Orellana-Torres, A. *Modelacin de la tasa de crecimiento de organismos filtradores en cultivo bajo limitacin de alimento.*Revista de biología marina y oceanografía, **49(1)**, (2014) 43-54.

[12] Valladares, A., Manríquez, G., & Suárez-Isla, B. A. *Shell shape variation in populations of Mytilus chilensis (Hupe 1854) from southern Chile: a geometric morphometric approach.* Marine biology, **157(12)**, (2010) 2731-2738.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

Linearity and its algebra in the bi-geometrical context

Fernando Córdova-Lepe[1], Rodrigo del Valle[1] and Karina Vilches Ponce[1]

[1] *Departamento de Matemática, Física y Estadística, Universidad Católica del Maule*

emails: fcordova@ucm.cl, rvalle@ucm.cl, kvilches@ucm.cl

Abstract

The multiplicative arithmetic is the base of the Bigeometric Calculus. In this work, we present the fundamental elements to construct a multiplicative linear algebra. Some important concepts and examples are given to visualized the effects of the multiplicative arithmetic properties over about what has traditionally been understood as linearity.
Key words: Non-newtonian arithmetic, Linearity, Matrix-matrix exponentiation

1 Introduction

The complete ordered field that gives arithmetic support to the bi-geometrical calculus is that of the positive real numbers endowed with two operations. The usual multiplication, as the first operation, and the *e-exponentiation* as the second one, which is defined by: $a \star b = a^{\ln(b)}$, $a, b \in \mathbb{R}^+$, see [3, 4, 5]. In other words, the operation $\star$ makes the calculation of powers a commutative operation.

If the intention is to advance towards generalizations or expansions of the bi-geometric or multiplicative calculus, for example towards larger dimensions, it is necessary to rescue the expressions that would have a linear algebra of multiplicative base. Notice that, just as the family of functions $y = ax$, $a \in \mathbb{R}$, corresponds to the functions that have constant multiplicative derivative, the lineal ones, it is of waiting, that one of the type $z = x^a y^b$, $a, b \in \mathbb{R}$, responds to the re-signification of the concept of linearity.

The multiplicative calculation in several variables has proved to be useful, for example, in the treatment of images with multiplicative noise, see [6], however, we know that a fundamental theoretical support to work in larger dimensions is the need for a linear algebra, in this work, that necessity guides our purpose.

As a novelty, in this paper we explore the traditional concepts of linear algebra, such as: vector space, linear transformations and matrix representations. In particular new

 ISBN: 978-84-617-8694-7

operations are introduced between matrices, and as a main result, we gives meaning to the idea of matrix raised to a matrix. Unlike other existing ones, see [1, 2, 7], this has an easier definition.

The present work (summary) is organized as follows: In Section 2, we give some preliminary of mathematical base. Section 3 is oriented toward a reconstruction of the the linear algebra. Finally, in Section 4, we introduce power between matrices and its properties are presented.

2 Mathematical preliminary

Let be $(\mathbb{R}^+, \cdot, \star)$ the field of the positive real numbers with the operations given by: (i) The usual product $a \cdot b$ and (ii) the *exponentiation* defined by $a \star b = a^{ln(b)}$. As we know, this field is an isomorphic copy of $(\mathbb{R}, +, \cdot)$. Notice that, if $\mathcal{P} =]1, \infty[$, then $\mathcal{P}$ divides the positives into three mutually exclusive categories, in effect, for each $a \in \Re$, we have $a \in \mathcal{P}$, $a = 1$ or $a^{-1} \in \mathcal{P}$. The elements of $\mathcal{P}$ are named *e-positive* numbers and if a, b are *e-positive* numbers, then the same with $a \cdot b$ and $a \star b$. Moreover, $\mathbb{R}^+$, with the *relative value* given by

$$[x] = \begin{cases} x\,, & x \geq 1\,, \\ x^{-1}\,, & x < 1\,, \end{cases} \tag{1}$$

is a *relative metric space*. In general, $\mathbb{R}^+$ is a complete ordered field. For $x, y \in \mathbb{R}^+$, some immediate properties are: (i) $[x] \geq 1$, (ii) $[x] = 1$ iff $x = 1$, (iii) $[x \star y] = [x] \star [y]$, and (iv) $[x \cdot y] \leq [x] \cdot [y]$, the last one is the triangular inequality.

3 Fundamental elements for a Linear Algebra

Let us define the pair $(V, \mathbb{R}^+)$ as an *e-vector space*, if we have an internal operations $v \cdot u$, for all v, $u \in V$ and an external one $\alpha \star v$, for all $\alpha \in \mathbb{R}^+$ and for all $v \in V$, that satisfy the following list of properties:

- Commutative: If u, $v \in V$, then $v \cdot u = u \cdot v$.
- Associative: If u, v and $w \in V$, then $v \cdot (u \cdot w) = (v \cdot u) \cdot w$.
- Neutral element: There exists $1 \in V$, such that $v \cdot 1 = v = 1 \cdot v$, for all $v \in V$.
- Inverse element: For each $v \in V$ there is a unique vector $v^{-1} \in V$, such that $v \cdot v^{-1} = 1 = v^{-1} \cdot v$.
- Compatibility of scalar multiplication with vector multiplication: Given $\alpha, \beta \in \mathbb{R}^+$, $(\alpha \star \beta) \star v = \alpha \star (\beta \star v)$, for all $v \in V$.

ISBN: 978-84-617-8694-7

- Distributive with respect to scalars: Given $\alpha, \beta \in \mathbb{R}^+$, then $(\alpha \cdot \beta) \star v = (\alpha \star v) \cdot (\beta \star v)$, for all $v \in V$.
- Distributive with respect to vectors: Given $\alpha \in \mathbb{R}^+$, $\alpha \star (u \cdot v) = (\alpha \star u) \cdot (\alpha \star v)$, for all u, $v \in V$.
- Identity element of scalar multiplication: The element $e \in V$ (e is the euler number) satisfies $e \star v = v$, for all $v \in V$.

It is possible to think of many examples of vector spaces from the multiplicative perspective, all in analogy with those of the additive approximation of linear algebra. Here are a few examples of *e-vector space*:

Euclidean Space: Let us consider, for any natural number n, the space $\mathbb{R}^n_+ = \mathbb{R}^+ \times \cdots \times \mathbb{R}^+$, n-times, of the $(x_1, \cdots, x_n)$, such that $x_i \in \mathbb{R}^+$ for each $i \in \{1, \cdots, n\}$. We consider the following operations:

- *Vector multiplication*: Given $u = (x_1, \cdots, x_n)$ and $v = (y_1, \cdots, y_n)$ in $\mathbb{R}^n_+$, we define the multiplication between vectors by

$$u \cdot v = (x_1 y_1, \cdots, x_n y_n)\,.$$

- *Scalar multiplication*: Given $u = (x_1, \cdots, x_n)$ in $\mathbb{R}^n_+$ and $\alpha \in \mathbb{R}^+$, the scalar multiplication is defined by

$$\alpha \star u = (\alpha \star x_1, \cdots, \alpha \star x_n) = (x_1^{\ln(\alpha)}, \cdots, x_n^{\ln(\alpha)})\,.$$

It is a simple and tedious calculation to test all the properties, listed above, that make of $(V, \mathbb{R}^+)$ with $V = \mathbb{R}^n_+$ as a e-vector space. We only emphasize that the neutral element of V for the vector multiplication is $e_V = (1, \cdots, 1)$ and the inverse element for each vector $u = (x_1, \cdots, x_n)$ is given by $u_{-1} = (x_1^{-1}, \cdots, x_n^{-1})$.

In addition, we note that a *linear combination* of two elements u and v of V, with parameter $\alpha, \beta \in \mathbb{R}^+$, looks like,

$$(\alpha \star u) \cdot (\beta \star v) = \left(x_1^{\ln(\alpha)} y_1^{\ln(\beta)}, \cdots, x_n^{\ln(\alpha)} y_n^{\ln(\beta)}\right).$$

In this way, the analogue to the *canonical base* in this context is defined by $\mathcal{C} = \{e_1, \cdots, e_n\}$, where each vector in this base have the Euler number e in the i-th position. Precisely,

$$e_i = (1, \cdots, 1, e, 1, \cdots, 1)\,,$$

 ISBN: 978-84-617-8694-7

for all $i \in \{1, \cdots, n\}$. Then, the decomposition of any vector $u = (x_1, \cdots, x_n)$ in the canonical base $\mathcal{C}$ is given by,

$$u = \prod_{i=1}^{n} x_i \star e_i = (x_1 \star e_1) \cdots (x_n \star e_n) .$$

In effect, we observe that

$$\prod_{i=1}^{n} x_i \star e_i = (x_1^{\ln(e)}, x_1^{\ln(1)}, \cdots, x_1^{\ln(1)}) \cdots (x_n^{\ln(1)}, \cdots, x_n^{\ln(1)}, x_n^{\ln(e)}) = (x_1, 1, \cdots, 1) \cdots (1, \cdots, 1, x_n) .$$

Matrix Space: Let us consider, for any natural number n, the space $M_n(\mathbb{R}^+)$ of matrices with positive real numbers inputs. We proceed to define the following operations,

- *Matrix multiplication*: Given $A = (a_{ij})$ and $B = (b_{ij})$ in $M_n(\mathbb{R}^+)$, the matrix product is defined by $A \cdot B = C$, where $c_{ij} = a_{ij} \cdot b_{ij}$, for $i, j \in \{1, \cdots, n\}$.
- *Scalar multiplication*: Given $A = (a_{ij})$ in $M_n(\mathbb{R}^+)$ and $\alpha \in \mathbb{R}^+$, the scalar product is defined by $\alpha \star A = (c_{ij})$, with $c_{ij} = \alpha \star a_{ij} = a_{ij}^{\ln(\alpha)}$ for any $i, j \in \{1, \cdots, n\}$.

With the pair of multiplicative operations defined above, the set $M_n(\mathbb{R}^+)$ is a e-vector space on the field $\mathbb{R}^+$.

4 Matrix-matrix exponentiation

There exists some definitions in the literature for matrix-matrix exponentiation. In [2], for example, it is defined as $A^B = e^{\ln(A)B}$, where $\exp(X) = \sum_{n=0}^{\infty}(1/n!)X^n$ and $\ln(Y)$ denotes the unique solution of the matrix equation $e^X = Y$ with eigenvalues $\lambda \in \mathbb{C}$ such that $-\pi < Im(\lambda) < \pi$.

In this section, a novel and simple operation between matrices is defined. It is named as *matrix exponentiation* and the objective is generalize the usual a^b for $a \in \mathbb{R}^+$ and $b \in \mathbb{R}$. First we will introduce, by simple analogies sum-product and product-exponentiation, the star operation $A \star B$ between matrices $A = (a_{ij}) \in M_{m\times n}(\mathbb{R}^+)$ and $B = (b_{ij}) \in M_{n\times p}(\mathcal{K})$ as:

$$A \star B = (c_{ij}) \in M_{m\times p}(\mathbb{R}^+), \quad \text{where} \quad c_{ij} = (a_{i1} \star b_{1j}) \cdots (a_{n1} \star b_{nj}).$$

This product is not commutative, but associative and distributed with respect to the traditional product of matrices (*i.e.* $A \star (B \cdot C) = (A \star B) \cdot (A \star C)$, the matrices with the dimensions that allow the operation). The neutral element will be denoted by $E = (a_{ij})$ and is defined as $a_{ij} = e$ if $i = j$ and $a_{ij} = 1$ if $i \neq j$.

 ISBN: 978-84-617-8694-7

On the other hand, given a matrix $B = (b_{ij}) \in M(\mathbb{R})$, we will denote by e_B to the matrix $(e^{b_{ij}}) \in M(\mathbb{R}^+)$. An obvious property is $e_{B+C} = e_B \cdot e_C$, for any B, $C \in M(\mathbb{R})$.

With this previous operations, we have the conditions to introduce the following definition of powers between matrices:

Definition: *Given matrices $A \in M(\mathbb{R}^+)$ and $B \in M(\mathbb{R})$, the operation power of base A and exponent B is defined by:*

$$A^B = A \star e_B.$$

It is important to remark that intuitively $A^B = A \star e_B = A^{\ln(e_B)}$.

Example: Let us consider matrices

$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in M(\mathbb{R}^+) \quad \text{and} \quad B = \begin{pmatrix} \alpha & \beta \\ \gamma & \delta \end{pmatrix} \in M(\mathbb{R}),$$

then

$$A^B = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \star \begin{pmatrix} e^{\alpha} & e^{\beta} \\ e^{\gamma} & e^{\delta} \end{pmatrix} = \begin{pmatrix} (a \star e^{\alpha})(b \star e^{\gamma}) & (a \star e^{\beta})(b \star e^{\delta}) \\ (c \star e^{\alpha})(d \star e^{\gamma}) & (c \star e^{\beta})(d \star e^{\delta}) \end{pmatrix},$$

this is

$$A^B = \begin{pmatrix} a^{\alpha} b^{\gamma} & a^{\beta} b^{\delta} \\ c^{\alpha} d^{\gamma} & c^{\beta} d^{\delta} \end{pmatrix}.$$

Some properties:

1. Given $A \in M(\mathbb{R}^+)$ and B, $C \in M(\mathbb{R})$, then $A^{(B+C)} = A^B \cdot A^C$.
2. Given $A \in M(\mathbb{R}^+)$ and B, $C \in M(\mathbb{R})$, then $(A^B)^C = A^{B \cdot C}$.
3. Given $A \in M(\mathbb{R}^+)$ and O the null matrix of $M(\mathbb{R})$, then $A^O = 1$, where 1 is the matrix for which all its entries are equal to 1.
4. Given $A \in M(\mathbb{R}^+)$ and Id the identity matrix of $M(\mathbb{R})$, then $A^{Id} = A$.
5. Given $B \in M(\mathbb{R})$, then $E^B = e_B$.
6. Given $1 \in M(\mathbb{R}^+)$, then $1^B = 1$.

5 Conclusions and perspectives

The multiplicative arithmetic, base of the bi-geometric calculus, allows us to introduce novel concepts in the traditional Linear Algebra and the idea of linearity is resignified.

 ISBN: 978-84-617-8694-7

Acknowledgements

This work is partially founded by CONICYT PAI-Academia 79150021.

References

[1] Barradas, I., & Cohen, J. E. *Iterated exponentiation, matrix-matrix exponentiation, and entropy*. Journal of mathematical analysis and applications, 183(1), (1994), 76-88.

[2] Cardoso, J. R., & Sadeghi, A. *On the conditioning of the matrix-matrix exponentiation*. arXiv preprint arXiv:1703.08804. (2017).

[3] Córdova-Lepe, F., & Pinto, M. *From quotient operation toward a proportional calculus*. International Journal of Mathematics, Game Theory and Algebra, **18(6)**, (2009) 527-536.

[4] Córdova-Lepe, F., & Pinto, M. *From quotient operation toward a proportional calculus*. In Mathematics Research Perspectives. Editors: Michael C. Leung (2012) 253-266.

[5] Córdova-Lepe, F. *The multiplicative derivative as a measure of elasticity in economics*. TMAT Revista Latinoamericana de Ciencias e Ingeniera,**2(3)**, (2006).

[6] Mora, M., Córdova-Lepe, F., & Del-Valle, R. *A non-Newtonian gradient for contour detection in images with multiplicative noise*. Pattern Recognition Letters, 33(10), 1245-1256, (2012).

[7] Wilcox, R. M. Exponential operators and parameter differentiation in quantum physics. Journal of Mathematical Physics, 8(4), 962-982, (1967).

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

Multi-adjoint object-oriented concept lattices in the resolution of multi-adjoint relation equations

M. Eugenia Cornejo, J. Carlos Díaz-Moreno, Jesús Medina[1]

[1] *Department of Mathematics, University of Cádiz. Spain.*

emails: {mariaeugenia.cornejo,juancarlos.diaz,jesus.medina}@uca.es

Abstract

This paper provides a characterization for the solvability of multi-adjoint relation equations by using multi-adjoint object-oriented concept lattices theory. Moreover, when multi-adjoint relation equations are solvable, we give an interpretation of their solutions as the intensions of the concepts in a multi-adjoint object-oriented concept lattice. Finally, we introduce a procedure in order to obtain two approximate solutions for unsolvable multi-adjoint relation equations.

Key words: Multi-adjoint object-oriented concept lattice; Fuzzy relation equations; Approximate solutions.

1 Introduction

The study of fuzzy relation equations has become an important research topic, both from theoretical and applicational perspective, since these equations were introduced by Sanchez in [23]. To the best of our knowledge, the most relevant papers focused on the resolution of fuzzy relation equations are [1, 4, 8, 9, 21, 22]. Other interesting approaches are [14, 15, 17, 18]. A generalization of fuzzy relation equations was given by Díaz-Moreno and Medina [10], who apply the multi-adjoint paradigm in order to obtain a more flexible framework than the existence environments [2, 3, 4, 20]. These general equations are known as multi-adjoint relation equations and an interesting study containing their most important properties can be found in [5, 10, 11, 12, 13, 19].

In this paper, we will present sufficient and necessary conditions in terms of concepts of multi-adjoint object-oriented concept lattices [16] to guarantee when a multi-adjoint fuzzy relation equation is solvable. Furthermore, we will provide an optimistic approximate solution and a pessimistic approximate solution based on concept lattice theory for unsolvable

multi-adjoint relation equations. These approximations are interesting when a solution of the problem is needed and the multi-adjoint relation equation is not solvable. To finish with, some illustrative examples together with conclusions and prospects for future work are given.

2 Multi-adjoint object-oriented concept lattice framework

Following the philosophy of the multi-adjoint paradigm, a generalization of the classical property and object-oriented concept lattices to a fuzzy environment was introduced in [16]. Later, the concepts of a multi-adjoint property-oriented concept lattice were related to the solutions of multi-adjoint fuzzy relation equations in [10]. Now, we are interested in interpreting the solutions of multi-adjoint fuzzy relation equations as the intensions of concepts in a multi-adjoint object-oriented concept lattice. For that reason, we need to recall some preliminary definitions and results associated with multi-adjoint object-oriented concept lattices theory.

To begin with, we will present the notion of adjoint triple which is composed by a non-commutative conjunctor and two residuated implications verifying the well-known adjoint property. Adjoint triples have been widely studied in [6] and their comparison with other general operators have been introduced in [7].

Definition 1 ([6]) *Let $(P_1, \leq_1)$, $(P_2, \leq_2)$, $(P_3, \leq_3)$ be three posets and $\&\colon P_1 \times P_2 \to P_3$, $\swarrow\colon P_3 \times P_2 \to P_1$, $\nwarrow\colon P_3 \times P_1 \to P_2$ be mappings, then $(\&, \swarrow, \nwarrow)$ is called an* adjoint triple *with respect to P_1, P_2, P_3, if the double equivalence:*

$$x \leq_1 z \swarrow y \quad \textit{iff} \quad x \,\&\, y \leq_3 z \quad \textit{iff} \quad y \leq_2 z \nwarrow x$$

is satisfied, for all $x \in P_1$, $y \in P_2$ and $z \in P_3$.

Adjoint triples play an important role as basic operators to make the computations in multi-adjoint object-oriented concept lattices environment, as it is shown in the following.

Definition 2 ([16])

- *A* multi-adjoint object-oriented frame *is a tuple $(L_1, L_2, P, \preceq_1, \preceq_2, \leq, \&_1, \swarrow^1, \nwarrow_1, \ldots, \&_n, \swarrow^n, \nwarrow_n)$ where $(L_1, \preceq_1)$ and $(L_2, \preceq_2)$ are two complete lattices, $(P, \leq)$ is a poset and $(\&_i, \swarrow^i, \nwarrow_i)$ is an adjoint triple, for all $i \in \{1, \ldots, n\}$. Multi-adjoint object-oriented frames will be denoted as $(L_1, L_2, P, \&_1, \ldots, \&_n)$.*
- *A* multi-adjoint object-oriented context *is a tuple (A, B, R, σ) such that A and B are non-empty sets, R is a P-fuzzy relation $R\colon A \times B \to P$ and $\sigma\colon A \times B \to \{1, \ldots, n\}$ is a mapping which associates any element in $A \times B$ with some particular adjoint triple in the frame.*

 ISBN: 978-84-617-8694-7

Considering a fixed multi-adjoint object-oriented frame and context, we can introduce the concept-forming operators ${}^{\uparrow_N}\colon L_2^B \to L_1^A$ and ${}^{\downarrow^\Pi}\colon L_1^A \to L_2^B$ which are defined as:

$$g^{\uparrow_N}(a) = \inf\{g(b) \swarrow^{a,b} R(a,b) \mid b \in B\} \quad (1)$$

$$f^{\downarrow^\Pi}(b) = \sup\{f(a) \,\&_{a,b}\, R(a,b) \mid a \in A\} \quad (2)$$

As usual, we define a *multi-adjoint object-oriented formal concept* $\langle g, f\rangle$ as a pair of mappings $g \in L_2^B$ and $f \in L_1^A$ satisfying that $g^{\uparrow_N} = f$ and $f^{\downarrow^\Pi} = g$. It is convenient to mention that the pair $({}^{\uparrow_N}, {}^{\downarrow^\Pi})$ forms an isotone Galois connection, the composition mapping ${}^{\uparrow_N\downarrow^\Pi}\colon L_2^B \to L_2^B$ is an interior operator and ${}^{\downarrow^\Pi\uparrow_N}\colon L_1^A \to L_1^A$ is an closure operator. These considerations are very important in order to obtain the elements of a *multi-adjoint object-oriented concept lattice*.

Theorem 1 ([16]) *Let $(L_1, L_2, P, \&_1, \ldots, \&_n)$ be a multi-adjoint object-oriented frame and (A, B, R, σ) be a multi-adjoint object-oriented context. The set of all multi-adjoint object-oriented formal concepts, which are denoted as $M_{N\Pi}(A, B, R, \sigma)$, together with the ordering $\langle g_1, f_1\rangle \preceq \langle g_2, f_2\rangle$ if and only if $g_1 \preceq_2 g_2$, or equivalently, if and only if $f_1 \preceq_1 f_2$, forms a complete lattice. This complete lattice will be called* multi-adjoint object-oriented formal concept lattice *and will be denoted as $(M_{N\Pi}(A, B, R, \sigma), \preceq)$.*

As we mentioned above, the main notions and results on the solvability of multi-adjoint fuzzy relations equations included in this paper will be based on the previous concepts corresponding to the multi-adjoint object-oriented concept lattices theory. As a consequence, a multi-adjoint object-oriented frame, denoted as $(L_1, L_2, P, \&_1, \ldots, \&_s)$, will be fixed in order to present our advances in the resolution of multi-adjoint fuzzy relations equations.

3 Solving multi-adjoint fuzzy relation equations

Multi-adjoint relation equations were introduced by Díaz-Moreno and Medina in order to generalize the usual fuzzy relation equations given in the literature [8, 9, 22]. In particular, a detailed study on multi-adjoint relation equations related to their definition, properties and solutions using multi-adjoint property-oriented concept lattices was presented in [10]. Now, we will introduce new definitions and will develop a mechanism based on multi-adjoint object-oriented (instead of property-oriented) concept lattices theory to solve multi-adjoint relation equations. This mechanism will not be studied in depth because it is very similar to the one given in [10]. We will start including the notions of multi-adjoint relation equation with sup-&-composition.

Definition 3 *Given the universes U, V, W, the fuzzy relations $S\colon V \times W \to P$, $T\colon U \times W \to L_2$, an unknown fuzzy relation $R\colon U \times V \to L_1$, and a mapping $\sigma\colon V \to \{1, \ldots, s\}$ that relates*

 ISBN: 978-84-617-8694-7

each element in V to an adjoint triple of the frame. A multi-adjoint relation equation with sup-&-composition *is the equation*

$$R \odot_\sigma S = T \tag{3}$$

that is to say,

$$\bigvee_{v \in V} (R(u,v) \,\&_v\, S(v,w)) = T(u,w), \quad u \in U, w \in W \tag{4}$$

where $\&_v$ represents the adjoint conjunctor associated with v by σ.

A mechanism to know when the multi-adjoint relation Equation (3) is solvable is given in the following theorem. It is important to emphasize that the solvability of these equations is characterized by the concepts of a multi-adjoint object-oriented concept lattice.

Theorem 2 *Let $u \in U$ and the fuzzy subset $g_u \in L_2^W$, defined as $g_u(w) = T(u,w)$, for all $w \in W$. Equation (3) can be solved if and only if $\langle g_u, g_u^{\uparrow_N} \rangle$ is a concept of $\mathcal{M}_{N\Pi}(V,W,S,\sigma)$, for all $u \in U$. In this case, the matrix R, defined by $R(u,v) = g_u^{\uparrow_N}(v)$ for all $u \in U$, $v \in V$, is the greatest solution.*

Note that, this procedure is very similar to the one given by Díaz-Moreno and Medina in the multi-adjoint property-oriented concept lattice framework [10]. For that reason, the proof will be omitted.

An illustrative example will be included in order to clarify the notions and properties on multi-adjoint relation equations previously presented.

Example 1 *Let us consider the multi-adjoint lattice $\langle [0,1], \leq, \&_1, \swarrow^1, \&_2, \swarrow^2 \rangle$ such that $(\&_1, \swarrow^1) = (\&_G, \swarrow^{\mathrm{G}})$ and $(\&_2, \swarrow^2) = (\&_P, \swarrow^{\mathrm{P}})$ are the Gödel and product adjoint pairs, respectively (see [6, 7] for more details). Given the sets $U = \{u_1, u_2\}, V = \{v_1, v_2\}, W = \{w_1, w_2, w_3\}$, the mapping $\sigma: V \to \{1,2\}$ such that $\sigma(v_1) = 1, \sigma(v_2) = 2$ and the fuzzy relations $S: V \times W \to [0,1]$ and $T: U \times W \to [0,1]$ defined by the following tables:*

Figure 1: Relation S.

	w_1	w_2	w_3
v_1	0.4	0.6	0.6
v_2	0.4	0.8	0.3

Figure 2: Relation T.

	w_1	w_2	w_3
u_1	0.4	0.7	0.4
u_2	0.4	0.5	0.3

We want to solve the multi-adjoint relation equation $(x,y) \odot_\sigma S(v,w) = T(u_1,w)$ being $v \in V, w \in W$ and x, y unknown variables. This equation is equivalent to the following system:

$$\begin{aligned} 0.4 &= (x \,\&_G\, 0.4) \vee (y \,\&_P\, 0.4) \\ 0.7 &= (x \,\&_G\, 0.6) \vee (y \,\&_P\, 0.8) \\ 0.4 &= (x \,\&_G\, 0.6) \vee (y \,\&_P\, 0.3) \end{aligned}$$

 ISBN: 978-84-617-8694-7

In order to analyze the solvability of this system, we will consider the fuzzy subset $g_{u_1} \in [0,1]^W$ *defined as* $g_{u_1}(w) = T(u_1, w)$*, for all* $w \in W$*, which is equivalent to say that* $g_{u_1}(w_1) = 0.4$*,* $g_{u_1}(w_2) = 0.7$*,* $g_{u_1}(w_3) = 0.4$*. To apply Theorem 2, we need to prove that* $\langle g_{u_1}, g_{u_1}^{\uparrow_N} \rangle$ *is a concept of* $\mathcal{M}_{N\Pi}(V, W, S, \sigma)$*, that is, we need to check if the equality* $g_{u_1}^{\uparrow_N\downarrow^\Pi} = g_{u_1}$ *holds. First of all, we will compute* $(g_{u_1})^{\uparrow_N}$*:*

$$\begin{aligned}(g_{u_1})^{\uparrow_N}(v_1) &= \inf\{0.4 \swarrow^{\mathrm{G}} 0.4, 0.7 \swarrow^{\mathrm{G}} 0.6, 0.4 \swarrow^{\mathrm{G}} 0.6\} = \inf\{1, 1, 0.4\} = 0.4\\ (g_{u_1})^{\uparrow_N}(v_2) &= \inf\{0.4 \swarrow^{\mathrm{P}} 0.4, 0.7 \swarrow^{\mathrm{P}} 0.8, 0.4 \swarrow^{\mathrm{P}} 0.3\} = \inf\{1, 0.875, 1\} = 0.875\end{aligned}$$

The fuzzy subset $(g_{\mathtt{flu}})^{\uparrow_N\downarrow^\Pi}$ *is obtained as follows:*

$$\begin{aligned}(g_{u_1})^{\uparrow_N\downarrow^\Pi}(w_1) &= \sup\{0.4 \,\&_{\mathrm{G}}\, 0.4, 0.875 \,\&_{\mathrm{P}}\, 0.4\} = \sup\{0.4, 0.35\} = 0.4\\ (g_{u_1})^{\uparrow_N\downarrow^\Pi}(w_2) &= \sup\{0.4 \,\&_{\mathrm{G}}\, 0.6, 0.875 \,\&_{\mathrm{P}}\, 0.8\} = \sup\{0.4, 0.7\} = 0.7\\ (g_{u_1})^{\uparrow_N\downarrow^\Pi}(w_3) &= \sup\{0.4 \,\&_{\mathrm{G}}\, 0.6, 0.875 \,\&_{\mathrm{P}}\, 0.3\} = \sup\{0.4, 0.2625\} = 0.4\end{aligned}$$

It is clear that the equality $g_{u_1}^{\uparrow_N\downarrow^\Pi} = g_{u_1}$ *is satisfied and therefore, we can ensure that* $\langle g_{u_1}, g_{u_1}^{\uparrow_N} \rangle$ *is a concept of* $\mathcal{M}_{N\Pi}(V, W, S, \sigma)$*. Hence, the considered system is solvable and its greatest solution is* $(g_{u_1})^{\uparrow_N} = (0.4, 0.875)$*.*

There exist different applications in which a solution is needed and the multi-adjoint fuzzy relation equations system is not solvable. In order to overcome this drawback, we can compute two approximations of the unknown variables as it is shown in the following section.

4 Optimistic and pessimistic approximations

We have provided a characterization for the solvability of multi-adjoint relations equations by means of the concepts of a multi-adjoint object-oriented concept lattice. Now, our task is focused on knowing what can we do when we need the solution of an unsolvable multi-adjoint relation equation. The answer to this question is given by the procedure developed in this section, which offer approximate solutions (an optimistic and a pessimistic approach) for unsolvable equations.

We suppose that for certain fuzzy subset $g_u \in L_2^W$, defined as $g_u(w) = T(u, w)$, for all $w \in W$, Equation (3) has not solution. Then, we present two possible alternatives in order to obtain approximate solutions for Equation (3):

- An approximation can be computed interchanging the row $T(u, w)$ by the fuzzy set $g_u^{\uparrow_N\downarrow^\Pi}$ in Equation (3). This approximation can be understand as a *pessimistic approximation*. Observe that, there exist other possibilities to obtain an approximate solution but the complexity for the computations is major.

 ISBN: 978-84-617-8694-7

- Another different approximation can be obtained by using a procedure based on the concepts whose extension is greater than g_u. In this case, we need to consider a finite object-oriented concept lattice which implies that the set

$$G = \{g \in L_2^B \mid g_u \preceq_2 g \text{ and } \langle g, g^{\uparrow_N} \rangle \in \mathcal{M}_{N\Pi}(V, W, S, \sigma)\}$$

either is empty or has minimal elements. It is easy to see that if g_m is a minimal element of G, then Equation (3) with $T(u, w) = g_m(w)$, for all $w \in W$, is solvable and $g_u \preceq_2 g_m$. As a consequence, $g_m^{\uparrow_N}$ can be considered as an *optimistic approximation.*

The following example will be useful to clarify the developed procedure previously.

Example 2 *Coming back Example 1 and considering the fuzzy subset $g_{u_2} \in [0,1]^W$ defined as $g_{u_2}(w) = (0.4, 0.5, 0.3)$, we can easily see that $\langle g_{u_2}, g_{u_2}^{\uparrow_N} \rangle$ is not a multi-adjoint object-oriented concept. Therefore, the system $(x, y) \odot_\sigma S(v, w) = g_{u_2}(w)$ is not solvable. Now, we will compute the pessimistic and optimistic approximations following the ideas introduced above:*

- *We will consider the fuzzy set $(g_{u_2})^{\uparrow_N \downarrow^\Pi}(w) = (0.3, 0.5, 0.3)$ instead of $g_{u_2}(w)$ in order to obtain the equation $(x, y) \odot_\sigma S(v, w) = (g_{u_2})^{\uparrow_N \downarrow^\Pi}(w)$. This equation is solvable and its greatest solution is $(g_{u_2})^{\uparrow_N} = (0.3, 0.625)$, which will be considered as a pessimistic approximation.*

- *In order to obtain an optimistic approximation, we will consider a finite multi-adjoint object-oriented concept lattice whose frame is composed by the regular partition $[0,1]_{10}$ and the discretizations of the Gödel and product conjunctors with respect to $[0,1]_{10}$ (see [6] for more details). We will fix the context (V, W, S, σ) given in Example 1, but in this occasion the mapping σ is associated with the discretization conjunctors. The Hasse diagram of the concept lattice $(M_{N\Pi}(A, B, R, \sigma), \preceq)$ together with the concepts is shown in Figure 3. According to Figure 3, we can ensure that the set of concepts whose extensions are greater than $g_{u_2} = (0.4, 0.5, 0.3)$ is:*

$$\{C_{13}, C_{14}, C_{15}, C_{16}, C_{17}, C_{18}, C_{19}, C_{20}, C_{21}, C_{22}, C_{23}, C_{24}, C_{25}\}$$

The previous set has two minimal elements C_{13} and C_{14} and therefore, we have two optimistic approximations. Considering the extension of C_{13}, we obtain that equation $(x, y) \odot_\sigma S(v, w) = (0.4, 0.5, 0.4)$ is solvable and its greatest solution is the intension of the concept C_{13}. Hence, $(0.4, 0.6)$ is an optimistic approximate solution of the original system. Following an analogous reasoning with the concept C_{14}, we have that $(0.3, 0.8)$ is another optimistic approximate solution of the original system.

 ISBN: 978-84-617-8694-7

$$
\begin{aligned}
C_0 &= \langle (0,0,0),(0,0)\rangle \\
C_1 &= \langle (0.1,0.1,0.1),(0.1,0.1)\rangle \\
C_2 &= \langle (0.1,0.2,0.1),(0.1,0.2)\rangle \\
C_3 &= \langle (0.2,0.3,0.1),(0.1,0.3)\rangle \\
C_4 &= \langle (0.2,0.2,0.2),(0.2,0.2)\rangle \\
C_5 &= \langle (0.2,0.3,0.2),(0.2,0.3)\rangle \\
C_6 &= \langle (0.2,0.4,0.2),(0.2,0.5)\rangle \\
C_7 &= \langle (0.3,0.3,0.3),(0.3,0.3)\rangle \\
C_8 &= \langle (0.3,0.5,0.2),(0.2,0.6)\rangle \\
C_9 &= \langle (0.3,0.4,0.3),(0.3,0.5)\rangle \\
C_{10} &= \langle (0.3,0.5,0.3),(0.3,0.6)\rangle \\
C_{11} &= \langle (0.4,0.4,0.4),(0.4,0.5)\rangle \\
C_{12} &= \langle (0.3,0.6,0.3),(0.3,0.7)\rangle \\
C_{13} &= \langle (0.4,0.5,0.4),(0.4,0.6)\rangle \\
C_{14} &= \langle (0.4,0.7,0.3),(0.3,0.8)\rangle \\
C_{15} &= \langle (0.4,0.6,0.4),(0.4,0.7)\rangle \\
C_{16} &= \langle (0.4,0.5,0.5),(0.5,0.6)\rangle \\
C_{17} &= \langle (0.4,0.8,0.3),(0.3,1.0)\rangle \\
C_{18} &= \langle (0.4,0.7,0.4),(0.4,0.8)\rangle \\
C_{19} &= \langle (0.4,0.6,0.5),(0.5,0.7)\rangle \\
C_{20} &= \langle (0.4,0.8,0.4),(0.4,1.0)\rangle \\
C_{21} &= \langle (0.4,0.7,0.5),(0.5,0.8)\rangle \\
C_{22} &= \langle (0.4,0.6,0.6),(1.0,0.7)\rangle \\
C_{23} &= \langle (0.4,0.8,0.5),(0.5,1.0)\rangle \\
C_{24} &= \langle (0.4,0.7,0.6),(1.0,0.8)\rangle \\
C_{25} &= \langle (0.4,0.8,0.6),(1.0,1.0)\rangle
\end{aligned}
$$

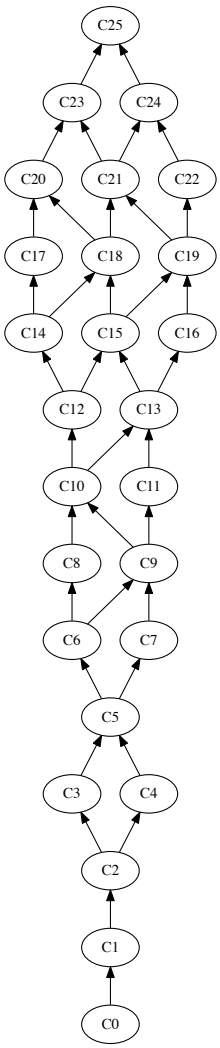


Figure 3: Hasse diagram of $(M_{N\Pi}(A,B,R,\sigma),\preceq)$ and its concepts (Example 2).

If we are interested in comparing the optimistic and the pessimistic approximations, it is necessary to compute the pessimistic approximation in the considered finite case. Solving the equation $(x,y)\odot_\sigma S(v,w)=(0.3,0.5,0.3)$ *where* $(0.3,0.5,0.3)$ *is the extension of concept* C_{11}*, we have that the pessimistic approximate solution is* $(0.3,0.6)$*. Clearly,* $(0.3,0.6)$ *is less than the optimistic approximations.*

After computing the pessimistic and optimistic approximate solutions, one can choose what approximate solution is the best in the $\wedge$*-semilattice given by the least element* $(0.3,0.6)$

 ISBN: 978-84-617-8694-7

and the maximal elements $(0.4, 0.6)$ *and* $(0.3, 0.8)$ *of the lattice* $[0,1]_{10} \times [0,1]_{10}$. *In this case, the* $\wedge$*-semilattice only has four elements:* $\{(0.3, 0.6), (0.4, 0.6), (0.3, 0.8), (0.3, 0.7)\}$.

5 Conclusions and future work

We have interpreted the solutions of multi-adjoint relation equations by using the concepst in a multi-adjoint object-oriented concept lattice. Furthermore, we have presented a mechanism in order to compute approximate solutions for unsolvable multi-adjoint relation equations. This method provides optimistic and pessimistic approximations of the considered equations based on concept lattice theory.

As a future work, we will study the interpretation of the minimal solutions of multi-adjoint relation equations and we will present real-life applications of these equations.

Acknowledgements

Partially supported by the State Research Agency (AEI) and the European Regional Development Fund (FEDER) project TIN2016-76653-P.

References

[1] W. Bandler and L. Kohout. Semantics of implication operators and fuzzy relational products. *Int. J. Man-Machine Studies*, 12:89–116, 1980.

[2] E. Bartl and R. Bělohlávek. Sup-t-norm and inf-residuum are a single type of relational equations. *International Journal of General Systems*, 40(6):599–609, 2011.

[3] R. Bělohlávek. Concept equations. *Journal of Logic and Computation*, 14(3):395–403, 2004.

[4] R. Bělohlávek. Sup-t-norm and inf-residuum are one type of relational product: Unifying framework and consequences. *Fuzzy Sets and Systems*, 197:45–58, 2012.

[5] M. E. Cornejo, J. C. Daz-Moreno, and J. Medina. Multi-adjoint relation equations: A decision support system for fuzzy logic. *International Journal of Intelligent Systems*, pages n/a–n/a, 2017.

[6] M. E. Cornejo, J. Medina, and E. Ramírez-Poussa. A comparative study of adjoint triples. *Fuzzy Sets and Systems*, 211:1–14, 2013.

 ISBN: 978-84-617-8694-7

[7] M. E. Cornejo, J. Medina, and E. Ramírez-Poussa. Multi-adjoint algebras versus non-commutative residuated structures. *International Journal of Approximate Reasoning*, 66:119–138, 2015.

[8] B. De Baets. Analytical solution methods for fuzzy relation equations. In D. Dubois and H. Prade, editors, *The Handbooks of Fuzzy Sets Series*, volume 1, pages 291–340. Kluwer, Dordrecht, 1999.

[9] A. Di Nola, E. Sanchez, W. Pedrycz, and S. Sessa. *Fuzzy Relation Equations and Their Applications to Knowledge Engineering.* Kluwer Academic Publishers, Norwell, MA, USA, 1989.

[10] J. C. Díaz-Moreno and J. Medina. Multi-adjoint relation equations: Definition, properties and solutions using concept lattices. *Information Sciences*, 253:100–109, 2013.

[11] J. C. Díaz-Moreno and J. Medina. Solving systems of fuzzy relation equations by fuzzy property-oriented concepts. *Information Sciences*, 222:405–412, 2013.

[12] J. C. Díaz-Moreno and J. Medina. Using concept lattice theory to obtain the set of solutions of multi-adjoint relation equations. *Information Sciences*, 266(0):218–225, 2014.

[13] J. C. Díaz-Moreno, J. Medina, and E. Turunen. Minimal solutions of general fuzzy relation equations on linear carriers. an algebraic characterization. *Fuzzy Sets and Systems*, 311:112 – 123, 2017.

[14] J. Ignjatović, M. Ćirić, B. Šešelja, and A. Tepavčević. Fuzzy relational inequalities and equations, fuzzy quasi-orders, closures and openings of fuzzy sets. *Fuzzy Sets and Systems*, 260:1 – 24, 2015. Theme: Algebraic Structures.

[15] J. Ignjatović, M. Ćirić, and V. Simović. Fuzzy relation equations and subsystems of fuzzy transition systems. *Knowledge-Based Systems*, 38:48 – 61, 2013.

[16] J. Medina. Multi-adjoint property-oriented and object-oriented concept lattices. *Information Sciences*, 190:95–106, 2012.

[17] J. Medina. Minimal solutions of generalized fuzzy relational equations: Clarifications and corrections towards a more flexible setting. *International Journal of Approximate Reasoning*, pages –, 2017.

[18] J. Medina. Notes on 'solution sets of inf-α_T fuzzy relational equations on complete brouwerian lattice' and 'fuzzy relational equations on complete brouwerian lattices'. *Information Sciences*, 402:82 – 90, 2017.

 ISBN: 978-84-617-8694-7

[19] J. Medina, E. Turunen, E. Bartl, and J. C. Díaz-Moreno. Minimal solutions of fuzzy relation equations with general operators on the unit interval. In A. Laurent, O. Strauss, B. Bouchon-Meunier, and R. R. Yager, editors, *IPMU (3)*, volume 444 of *Communications in Computer and Information Science*, pages 81–90. Springer, 2014.

[20] K. Peeva. *Imprecision and Uncertainty in Information Representation and Processing: New Tools Based on Intuitionistic Fuzzy Sets and Generalized Nets*, chapter Intuitionistic Fuzzy Relational Equations in *BL* Algebras, pages 73–85. Springer International Publishing, Cham, 2016.

[21] I. Perfilieva. Fuzzy function as an approximate solution to a system of fuzzy relation equations. *Fuzzy Sets and Systems*, 147(3):363–383, 2004.

[22] I. Perfilieva and L. Nosková. System of fuzzy relation equations with inf-$\rightarrow$ composition: Complete set of solutions. *Fuzzy Sets and Systems*, 159(17):2256–2271, 2008.

[23] E. Sanchez. Resolution of composite fuzzy relation equations. *Information and Control*, 30(1):38–48, 1976.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

LPG Demand Forecast using Time Series

Aldina Correia[1], Eliana Costa e Silva[1], Cristina Lopes[2], Cláudio Henriques[3], Fábio Henriques[3], Mariana Pinto[3], Magda Monteiro[4], Rui Borges Lopes[5] and Ana Sapata[6]

[1] *CIICESI - Center for Research and Innovation in Business Sciences and Information Systems, ESTG - School of Management and Technology, P.PORTO - Polytechnic of Porto*

[2] *LEMA – Mathematical Engineering Lab, CEOS.PP – Centre for Organisational and Social Studies, ISCAP – Accounting and Business School, P.PORTO*

[3] *Mathematics Department, University of Aveiro*

[4] *ESTGA / CIDMA, University of Aveiro*

[5] *DEGEIT / CIDMA, University of Aveiro*

[6] *Mathematics Department, University of Évora, Portugal*

emails: aic@estg.ipp.pt, eos@estg.ipp.pt, cristinalopes@iscap.ipp.pt, c.henriques@ua.pt, fabiomchenriques@ua.pt, marianaapinto@ua.pt, msvm@ua.pt, rui.borges@ua.pt, ana_sapata@sapo.pt

Abstract

At the 127^{th} European Study Group with Industry an energy sector company proposed an industrial challenge that consisted on the asset acquisition planning for its liquefied petroleum gas (LPG) cylinder business, one of the most recent business areas in the company. This business area is still in a growing phase and to satisfy the market needs and assure a sustainable growth a very tight control of the main assets, the LPG cylinders, is of paramount importance. Therefore, a detailed planning of all the assets acquisition is required, taking into consideration several variables: sales growth rate, seasonality, cylinder rotation and corresponding return rate to the filling plant. The challenge was to develop a model for the assets acquisition planning. In order to tackle this challenge, it was necessary to forecast the demand. For that purpose, time

 ISBN: 978-84-617-8694-7

series techniques were used, in particular, moving averages and exponential smoothing. The results show that the seasonality does not explain all the variation of the demand, therefore it is necessary to use a model that would consider other possible explanatory variables. According to several authors, gas consumption may be influenced by several aspects, such as, atmospheric temperatures, heliophany (a measure of the day luminosity), wind, relative humidity, rains, minimum and maximum temperatures, demand in previous periods, and prices. Therefore, in an extended version of this article a multivariate linear regression model will be used.

Key words: Industrial Mathematics, LPG bottles demand, time series, exponential smoothing forecast, moving averages.

MSC 2000: 62P30 Applications in engineering and industry; 62-07 Data analysis; 62Hxx Multivariate analysis; 62M10 Time series, auto-correlation, regression.

1 Introduction

A Portuguese company in the energy sector, that will be named ALPHA for confidentiality reasons, started its activity in 2006, focusing in the production and distribution of biofuel. Since then, the company has been continuously growing, extending its business areas to other fuels, and is now operating at a national level. The company started the liquefied petroleum gas (LPG) cylinder business activities in 2012, and since then it has experienced a continuous growth. In this business, the LPG cylinder is the main asset and a correct planning of its needs is critical. The company currently commercializes propane gas and has two types of cylinders with different capacity, henceforth named type A and type B cylinders.

ALPHA wants to find a model to forecast the demand of each type of LPG cylinders. These forecasts are crucial for the company to define an assets acquisition plan, i.e., to determine the amount of LPG cylinders to acquire, and when to acquire them.

In [9], Virtullo et. al. look at the financial implication of forecasting natural gas, the nature of natural gas forecasting, the factors that impact natural gas consumption, and present a survey of the mathematical techniques and practices used to model natural gas demand. These authors argue that the most common mathematical modeling techniques used to forecast daily demand are multiple linear regression (MLR) and artificial neural networks (ANN) (for details see [8]). Paggi and Robledo presented, in [7], an application of ANN in the prediction of time series of the weekly wholesales of bottled propane gas.

In the present work, in order to tackle the challenge, an approach similar to these authors is followed and several statistical techniques are used.

This paper is organized as follows: In Section 2 an exploratory analysis of the data is made. Section 3 is devoted to determining the sales forecast of bottled propane gas for the Portuguese market and for ALPHA company. In the last section some conclusions are presented and future work is sketched.

 ISBN: 978-84-617-8694-7

2 Dataset and exploratory analysis

The dataset provided by the company included information about sales, return number of cylinders, operational stock of assets, total number of assets in the market, of the two types of propane cylinders (A – small bottles and B – large bottles), between January Year 0 (Jan/Y0) and December Year 1 (Dec/Y1). The forecast for sales during Year 2 are also provided by the company. Note that, for confidentiality reasons, the data used in this work was masked. Additional data about consumption of propane and temperature in Portugal were also collected from the Portuguese Association of Petroleum Companies – APETRO[1] and from the Portuguese Institute for Sea and Atmosphere – IPMA[2], respectively. Figure 1 and Figure 2 show this data.

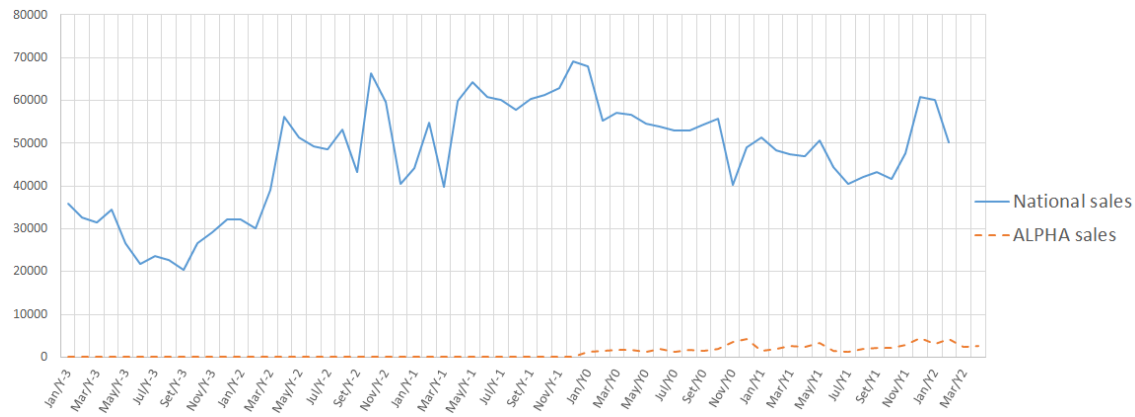


Figure 1: Comparison between national sales and ALPHA sales.

Figure 1 shows a general increase in the sales of ALPHA, however, it still represents a small percentage of the national market. Furthermore, the national and ALPHA sales of propane have different behaviors, thus in principle this is not a good indicator to forecast the company's sales when based on the national ones.

Figure 2 depicts the values of propane national sales and, simultaneously, the average temperature[3] in Portugal in the same period is shown. Since propane gas is used mostly for cooking and water heating, it is expected that whenever temperature decreases there is an increase in gas consumption. However, from Figure 2, this is not always true. In fact, for January Year -1 there is a decrease of the temperature and also a decrease of the consumption of gas.

[1] `http://www.apetro.pt`
[2] `https://www.ipma.pt`
[3] The air temperature in ($^{\circ}C$) was multiplied by a constant factor for a better comparison with the sales.

 ISBN: 978-84-617-8694-7

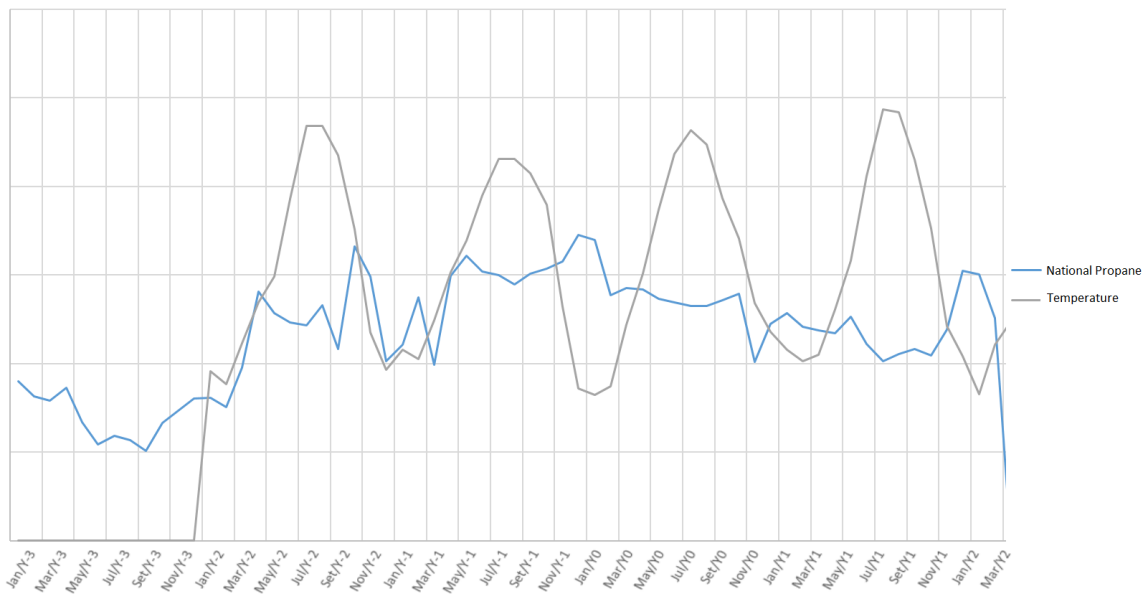


Figure 2: Average temperature and national sales.

In order to study the existence of seasonality in sales of type A and B cylinders, Figure 3 presents the sales for the three years. In this figure an increasing linear trend of ALPHA's sales, due to the company's market expansion, is observed. There is also some indicators of seasonality, because over the years the variations appear to be similar. In order to analyse the seasonality, moving averages seasonality coefficients (e.g. [4,10]) and exponential smoothing forecasts (for review see e.g. [2, 3, 5]), are calculated in the next Section.

3 Bottled propane gas sales forecast

3.1 Estimating seasonality coefficients

To study the seasonality of the data, the seasonality coefficients for total propane sales of ALPHA (in tons); sales of type A and B ALPHA assets; and sales of butane, propane and total in Portugal were calculated.

Let us consider, without loss of generality, ALPHA sales of propane between January Year 0 and December Year 2. The number of observations is given by $N = k \times s$, where k is the number of years and s the number of periods in the year, i.e. months. In this case $N = 3 \times 12 = 36$. Using the centered moving averages method, it is possible to calculate 12 seasonal indexes (one for each month) that express the amount of sales in each month

 ISBN: 978-84-617-8694-7

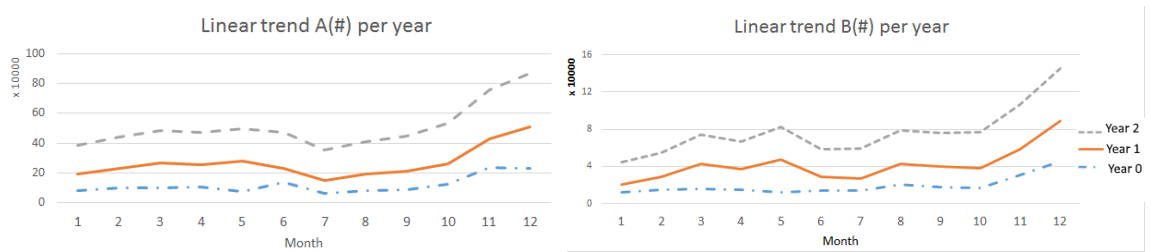


Figure 3: ALPHA sales of type A and B cylinders.

that are superior (or inferior) to the global mean sales. Using the multiplicative method, the seasonal indexes work as a percentage. The non-normalized estimates of the seasonal component at time i of each year are:

$$\bar{S}_i = \frac{1}{k-1}\sum_{j=1}^{k} S^*_{i+s(j-1)}\ , \qquad i = 1, 2, \ldots, s,$$

with

$$S^*_t = \frac{X_t}{M_t},$$

where M_t are the centered moving averages of the sales series X_t:

$$M_t = \frac{1}{s}\left(\frac{1}{2}X_{t-\frac{s}{2}} + X_{t-\frac{s}{2}+1} + \ldots + X_{t+\frac{s}{2}-1} + \frac{1}{2}X_{t+\frac{s}{2}}\right), \quad t = \frac{s}{2}+1, \ldots, N-\frac{s}{2}.$$

Finally, the standardized estimates of the seasonal components are:

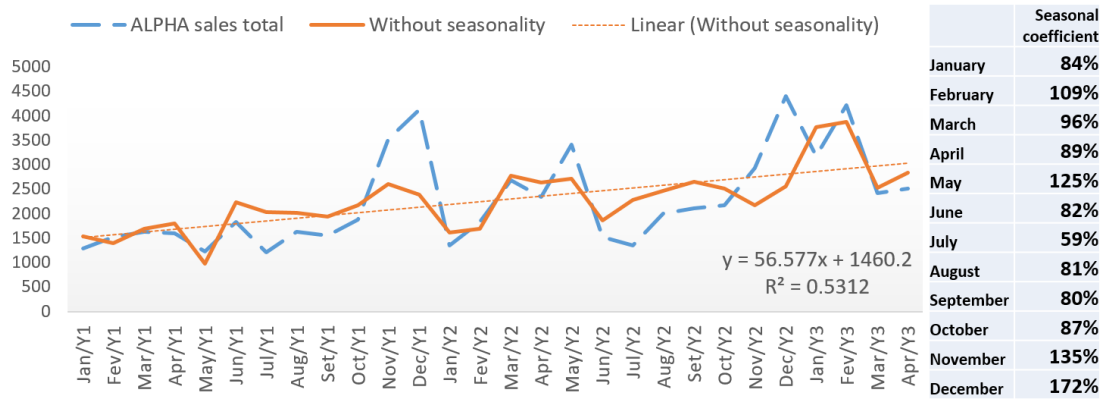
$$\hat{S}_i = \bar{S}_i \cdot \frac{s}{\sum_{j=1}^{s}\bar{S}_j}\ , \qquad 1, 2, \ldots, s.$$

Using these formulas it is possible to estimate the seasonal indexes of the demand and sales (see Figure 4).

Similarly, the seasonal coefficients presented in Table 1 were obtained. The national coefficients (PT butane, PT propane, PT total) do not have significant variations from month to month. In fact, for the national sales, the PT butane seasonal coefficients present a minimum of 84% in October and a maximum of 111% in May, while for PT propane and PT total the minimum is attained in September with 92% and 91%, respectively, and the maximum occurs in April with 111% and 110% 92%.

On the other hand, ALPHA sales show more significant variations. The seasonal coefficients of the total sales vary from 59% in July to 172% in December, while for type A

 ISBN: 978-84-617-8694-7



	Seasonal coefficient
January	84%
February	109%
March	96%
April	89%
May	125%
June	82%
July	59%
August	81%
September	80%
October	87%
November	135%
December	172%

Figure 4: ALPHA sales (in tons).

cylinder from 55% in July until 163% in December, and finally, type B cylinder sales vary from 65% in July to 183% in December (see Table 1). The months with high coefficients (above 100%) are February, May, November and December, and the ones with smallest coefficients (below 100%) are June, July, August and September.

Table 1: Seasonal coefficients.

	PT			ALPHA		
Month	butane	propane	total	total	type A	type B
January	100%	99%	99%	84%	96%	69%
February	102%	95%	97%	109%	110%	108%
March	103%	93%	96%	96%	95%	98%
April	107%	111%	110%	89%	92%	84%
May	111%	110%	110%	125%	121%	132%
June	96%	98%	98%	82%	91%	70%
July	109%	98%	101%	59%	55%	65%
August	101%	98%	99%	81%	69%	97%
September	90%	92%	91%	80%	74%	87%
October	84%	109%	102%	87%	91%	81%
November	88%	99%	96%	135%	144%	124%
December	108%	98%	101%	172%	163%	183%

Therefore, forecast for sales can be done using these seasonal coefficients.

 ISBN: 978-84-617-8694-7

3.2 Exponential smoothing forecast

Exponential smoothing forecast is a widely used method for time series forecast, including sales forecasting [1, 5, 6, 12].

Using the seasonal coefficients obtained in the previous subsection to desseasonalize the data, then Holt's method was used to forecast type A bottles sales. The model obtained using Holt's method [11] was such that: AIC = 671.9332, BIC = 678.5942, RMSE = 25697.11 for the training set, level coefficient $\alpha = 0.0253$ and trend coefficient $\beta = 0.0253$. Further details of this model are depicted in Table 2). In Figure 5, the observed values of the type A bottles sales (from Jan/Y0 to Apr/Y2) are presented alongside with the estimated values (from Jan/Y0 to Apr/Y2), and the forecast sales for May/Y2 until Dec/Y3.

sigma:	25697.11						
	AIC	AICc	BIC				
	671.9332	674.6605	678.5942				
Error measures:							
	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
Training set	-4739.01	25697.11	18961.96	-6.6708	14.9323	0.71456	0.05174

Table 2: Details of the model for type A bottles sales obtained using Holt's method.

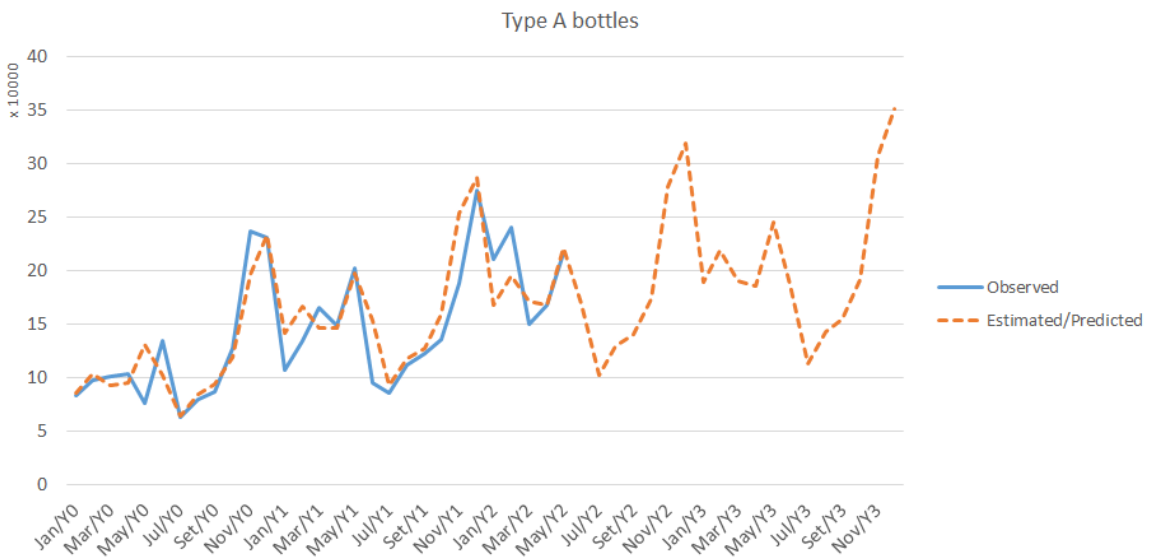


Figure 5: Forecast of type A bottle sales using Holt's method.

 ISBN: 978-84-617-8694-7

4 Conclusions and future work

The goal of this work was to forecast the sales of propane gas cylinders in order to plan the assets acquisition necessity of a Portuguese company of the energy sector. In this work the time series techniques exponential smoothing and moving averages were used to forecast the demand.

This approach allowed to see that the national seasonal coefficients are quite distinct from the ones observed for the sales of the company. ALPHA's sales present a larger variability in the seasonal coefficients than the total national sales. For the company the higher coefficients were observed in May, November and December, while the smaller were in July.

Since national and ALPHA seasonality coefficients are different, some other possible explanatory variables should be considered in order to forecast the demand with better accuracy. For that reason, several data has been collected, such as atmospheric temperatures, demand in previous periods, objectives of sales, expectation of price increase, and among others variables. In an extended version of this paper the focus will be in using approaches to forecast sales that take into account this data.

Acknowledgements

This article is based upon work from COST Action TD1409, Mathematics for Industry Network (MI-NET), supported by COST (European Cooperation in Science and Technology).

References

[1] Lucia Cassettari, Lucia Cassettari, Ilaria Bendato, Ilaria Bendato, Marco Mosca, Marco Mosca, Roberto Mosca, and Roberto Mosca. A new stochastic multi source approach to improve the accuracy of the sales forecasts. *foresight*, 19(1):48–64, 2017.

[2] Everette S Gardner. Exponential smoothing: The state of the art. *Journal of forecasting*, 4(1):1–28, 1985.

[3] PJ Harrison. Exponential smoothing and short-term sales forecasting. *Management Science*, 13(11):821–842, 1967.

[4] Charles C Holt. Forecasting seasonals and trends by exponentially weighted moving averages. *International Journal of Forecasting*, 20(1):5–10, 2004.

[5] Rob Hyndman, Anne B Koehler, J Keith Ord, and Ralph D Snyder. *Forecasting with exponential smoothing: the state space approach*. Springer Science & Business Media, 2008.

 ISBN: 978-84-617-8694-7

[6] John T Mentzer and James E Cox. Familiarity, application, and performance of sales forecasting techniques. *Journal of Forecasting*, 3(1):27–36, 1984.

[7] Horacio Paggi and Franco Robledo. A neural networks based model for the prediction of the bottled propane gas sales. In *Mathematics and Computers in Sciences and in Industry (MCSI), 2014 International Conference on*, pages 69–74. IEEE, 2014.

[8] Steven Vitullo. Disaggregating time series data for energy consumption by aggregate and individual customer. 2011.

[9] Steven R Vitullo, Ronald H Brown, George F Corliss, and Brian M Marx. Mathematical models for natural gas forecasting. *Canadian applied mathematics quarterly*, 17(7):807–827, 2009.

[10] Peter R Winters. Forecasting sales by exponentially weighted moving averages. *Management science*, 6(3):324–342, 1960.

[11] David J Wright. Forecasting data published at irregular time intervals using an extension of holt's method. *Management science*, 32(4):499–510, 1986.

[12] Lifeng Wu, Sifeng Liu, and Yingjie Yang. Grey double exponential smoothing model and its application on pig price forecasting in china. *Applied Soft Computing*, 39:117–123, 2016.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

Some results about randomized binary Markov chains: Theory and computing

J.-C. Cortés[1], A. Navarro-Quiles[1], J.-V. Romero[1] and M.-D. Roselló[1]

[1] *Instituto Universitario de Matemática Multidisciplinar,*
Universitat Politècnica de València
Camino de Vera s/n, 46022, Valencia, Spain

emails: jccortes@imm.upv.es, annaqui@doctor.upv.es, jvromero@imm.upv.es, drosello@imm.upv.es

Abstract

A Markov model is defined through its associated stochastic transition matrix whose entries are deterministic values representing the so-called transition probabilities between two possible states. In practice, these quantities are not known in a deterministic manner since they need to be fixed from sampling, therefore, it is more realistic to consider them as random variables (RV's) rather than deterministic quantities. This paper is addressed to give a generalization of the classical Markov methodology allowing the treatment of the entries of the transition matrix as RV's. The randomization of the classical Markov model permits the computation of the first probability density function (1-PDF) of the solution stochastic process taking advantage of the so-called Random Variable Transformation (RVT) technique. From the 1-PDF relevant probabilistic information about the evolution of the model described by Markov models can be calculated including all one-dimensional moments (mean, variance, symmetry, kurtosis, etc). As it is shown, RVT method also allows us the computation of the PDF's of the time instant until a certain proportion of the total population reaches a specific value of interest and of the steady state. All theoretical results are established under fairly general assumption since the PDF's of the randomized transition probabilities are assumed to be arbitrary. Finally, our theoretical findings are illustrated by means of a numerical example.

Key words: Binary Markov chain, random variable transformation technique, first probability density function

 ISBN: 978-84-617-8694-7

1 Introduction

A stochastic process (SP) is a mathematical representation that permits to describe how evolves a phenomenon over time in a probabilistic manner. Discrete Markov models, also referred to as Markov chains, are particular SP's where the outcome of an experiment depends only on the outcome of the previous experiment [5, 1]. Markov chains are applied in many different realms including Economy, Sociology, Reliability, Medical Decision Making, for instance. Markov chains are often chosen as a suitable tools for modelling very different phenomena because they are fairly general and adaptable to many contexts [4]. Moreover, excellent numerical techniques exist for computing its statistics.

This contribution is addressed to give a generalization of classical Markov chains by randomizing the entries of the transition matrix as random variables (RV's). To the best of our knowledge, this problem has not been considered yet in the extant literature. As a first step, we here will concentrate on the simplest type of Markov chains, usually referred to as binary Markov chains which are those having just two possible states.

Let $\{x_n,\, n = 0, 1, \dots\}$ be a Markov chain, where $n = 0, 1, 2, \dots$, denotes the cycle or period. As early indicated, we will consider binary Markov chains whose two possible states will be denoted by x_n^1 and x_n^2, being n the period or cycle. It is assumed that $0 < x_n^1, x_n^2 < 1$, thus x_n^1 and x_n^2 can be interpreted as percentages. As a key feature of Markov chains, it is assumed that $x_n^1 + x_n^2 = 1$ for every n. This means that the system is closed, that is, if for example, x_n^1 and x_n^2 represent the percentage of susceptibles and infected persons of a population, then it is implicitly assumed that any person can neither leave nor join the population. The evolution of the number of individuals in each cycle n is determined by the initial state $(x_0^1, x_0^2)^\top$ and the transition matrix while the long-term behaviour of the Markov chain only depends on the transition matrix. This matrix is a constant matrix whose entries represent the probabilities to change either from one state to another or to remain in the same state between to consecutive cycles. In practice, this entries are assumed to be deterministic quantities. In this contribution we generalize this feature by considering that the entries of the transition matrix are RV's rather than deterministic constants. Naturally, these RV's are assumed to take values in the interval $[0, 1]$, thus representing probabilities. In Figure 1, we show the flow diagram with the transitions among states.

In the classical context, a Markov chain is described as follows

$$x_{n+1} \;=\; a\, x_n, \quad n = 0, 1, 2, \dots, \quad a = \begin{pmatrix} p & 1-q \\ 1-p & q \end{pmatrix}, \tag{1}$$

where a is termed the (deterministic) transition matrix and $x_0 = (x_0^1, x_0^2)^\top = (x_0^1, 1 - x_0^1)^\top$ is the initial condition, i.e. the initial percentage of individuals in each group.

As indicated above, we will consider the entries of the transition matrix, p and q, as well as the initial condition, x_0, as RV's. For sake of clarity, in the following these quantities will

 ISBN: 978-84-617-8694-7

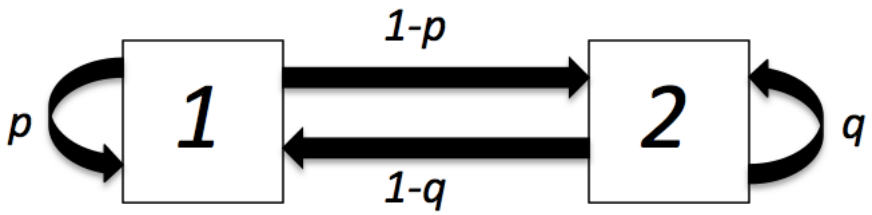


Figure 1: Flow diagram to a binary Markov chain.

be written using capital letters. The randomized Markov chain is then written as follows

$$\begin{aligned} X_{n+1} &= A\,X_n, \quad n=0,1,2,\ldots, \quad A=\begin{pmatrix} P & 1-Q \\ 1-P & Q \end{pmatrix}, \\ X_0 &= (X_0^1, 1-X_0^1)^\top, \end{aligned} \tag{2}$$

where X_0^1, P and Q are assumed to be absolutely continuous RV's defined on a common probability space $(\Omega, \mathcal{F}, \mathbb{P})$.

Solving a randomized Markov chain means not only to obtain its solution discrete SP, $X_n = (X_n^1, X_n^2)^\top = (X_n^1, 1-X_n^1)^\top$, but also to compute its mean, $\mathbb{E}[X_n]$ and its variance, $\mathbb{V}[X_n]$, for each cycle n. A more general purpose is the computation of its first probability density function (1-PDF), $f_1(x;n)$. This function provides a full probabilistic description of the solution SP in every cycle n. The aim of this paper is to obtain the 1-PDF of the solution SP for randomized binary Markov chains under very general hypotheses. To reach this objective, we will apply the Random Variable Transformation (RVT) method. This is a powerful technique that has been recently used by the authors to construct random phase portrait for planar systems [2] and to model the stroke disease [3]. The RVT technique permits to compute the PDF of a RV which results from mapping of another RV whose PDF is known. The multidimensional version of the RVT technique is stated in Theorem 1.

Theorem 1 *(Multidimensional version, [6, pp. 24–25]). Let* $\mathbf{U} = (U_1, \ldots, U_n)^\top$ *and* $\mathbf{V} = (V_1, \ldots, V_n)^\top$ *be two n-dimensional absolutely continuous random vectors. Let* $\mathbf{r} : \mathbb{R}^n \to \mathbb{R}^n$ *be a one-to-one deterministic transformation of* $\mathbf{U}$ *into* $\mathbf{V}$*, i.e.,* $\mathbf{V} = \mathbf{r}(\mathbf{U})$*. Assume that* $\mathbf{r}$ *is continuous in* $\mathbf{U}$ *and has continuous partial derivatives with respect to* $\mathbf{U}$*. Then, if* $f_{\mathbf{U}}(\mathbf{u})$ *denotes the joint probability density function of vector* $\mathbf{U}$*, and* $\mathbf{s} = \mathbf{r}^{-1} = (s_1(v_1, \ldots, v_n), \ldots, s_n(v_1, \ldots, v_n))^\top$ *represents the inverse mapping of* $\mathbf{r} = (r_1(u_1, \ldots, u_n), \ldots, r_n(u_1, \ldots, u_n))^\top$*, the joint probability density function of vector* $\mathbf{V}$ *is given by*

$$f_{\mathbf{V}}(\mathbf{v}) = f_{\mathbf{U}}\left(\mathbf{s}(\mathbf{v})\right) |J|, \tag{3}$$

 ISBN: 978-84-617-8694-7

where $|J|$ is the absolute value of the Jacobian, which is defined by

$$J = \det\left(\frac{\partial \mathbf{s}^\top}{\partial \mathbf{v}}\right) = \det\begin{pmatrix} \dfrac{\partial s_1(v_1,\ldots,v_n)}{\partial v_1} & \cdots & \dfrac{\partial s_n(v_1,\ldots,v_n)}{\partial v_1} \\ \vdots & \ddots & \vdots \\ \dfrac{\partial s_1(v_1,\ldots,v_n)}{\partial v_n} & \cdots & \dfrac{\partial s_n(v_1,\ldots,v_n)}{\partial v_n} \end{pmatrix}. \tag{4}$$

As the two states of a binary Markov chain make up a closed system, $X_n^1 + X_n^2 = 1$, we shall see that once the 1-PDF of one of the two states has been computed, the 1-PDF of the other state can be straightforwardly determined taking advantage of the following key lemma. This result can be derived as direct application of Theorem 1.

Lemma 2 *Let X and Y be two absolutely continuous random variables, such as $Y = 1 - X$. Let $f_X(x)$ denote the probability density function of the random variable X, then the probability density function of the random variable Y is given by*

$$f_Y(y) = f_X\left(1 - y\right). \tag{5}$$

2 Solving the randomized binary Markov chain

This section is divided in two parts. In the first subsection we will compute the 1-PDF of the solution to the randomized binary Markov chain (2) under very general assumptions and the PDF of its steady state as well. The second subsection is addressed to determine the PDF of the RV representing the time until a given proportion of the population is reached. These goals will be achieved by applying the RVT technique.

Let us recall that the solution of problem (2) is given by

$$X_n = A^n X_0 = \begin{pmatrix} \dfrac{-1+Q+(-1+P+Q)^n\left(1-Q+(-2+P+Q)X_0^1\right)}{-2+P+Q} \\ \dfrac{-1+P+(-1+P+Q)^n\left(-1+Q-(-2+P+Q)X_0^1\right)}{-2+P+Q} \end{pmatrix},\ n = 0, 1, \ldots \tag{6}$$

As P and Q are absolutely continuous RV's, then $\mathbb{P}\left[\{\omega \in \Omega :\ P(\omega) + Q(\omega) - 2\}\right] = 0$, for all event $\omega \in \Omega$. As a consequence, the denominator of both components of (6) is well-defined.

 ISBN: 978-84-617-8694-7

2.1 First Probability Density Function

As previously indicated, in this subsection we will obtain the 1-PDF of (6) using the RVT method. To do this, we will apply Theorem 1 for the following choice of mapping $\mathbf{r}$

$$\begin{aligned} y_1 &= r_1\left(x_0^1,p,q\right) = \frac{-1+q+(-1+p+q)^n\left(1-q+(-2+p+q)x_0^1\right)}{-2+p+q},\\ y_2 &= r_2\left(x_0^1,p,q\right) = p,\\ y_3 &= r_3\left(x_0^1,p,q\right) = q. \end{aligned}$$

Then, we will obtain the PDF of X_n^1 for n fixed. This can be achieved by applying Theorem 1, but we first need to compute the inverse mapping $\mathbf{s}$ of mapping $\mathbf{r}$

$$\begin{aligned} x_0^1 &= s_1\left(y_1,y_2,y_3\right) = \frac{y_1\left(-2+y_2+y_3\right)+\left(-1+y_3\right)\left(-1+\left(-1+y_2+y_3\right)^n\right)}{\left(-1+y_2+y_3\right)^n\left(-2+y_2+y_3\right)},\\ p &= s_2\left(y_1,y_2,y_3\right) = y_2,\\ q &= s_3\left(y_1,y_2,y_3\right) = y_3, \end{aligned}$$

and its jacobian

$$|J| = \left|\frac{\partial s_1}{\partial y_1}\right| = \left|\frac{1}{\left(-1+y_2+y_3\right)^n}\right|.$$

Therefore, the PDF of the random vector (Y_1,Y_2,Y_3) defined through mapping $\mathbf{r}$ is

$$\begin{aligned} f_{y_1,y_2,y_3}\left(y_1,y_2,y_3\right) &= f_{x_0^1,P,Q}\left(\tfrac{y_1(-2+y_2+y_3)+(-1+y_3)(-1+(-1+y_2+y_3)^n)}{(-1+y_2+y_3)^n(-2+y_2+y_3)},y_2,y_3\right)\\ &\times \left|\tfrac{1}{(-1+y_2+y_3)^n}\right|. \end{aligned}$$

Finally, marginalizing this expression with respect to P and Q and letting n arbitrary, we obtain the 1-PDF of X_n^1

$$\begin{aligned} f_1^{x^1}(x;n) &= \iint_{\mathcal{D}(P,Q)} f_{x_0^1,P,Q}\left(\tfrac{x(-2+p+q)+(-1+q)(-1+(-1+p+q)^n)}{(-1+p+q)^n(-2+p+q)},p,q\right)\\ &\times \left|\tfrac{1}{(-1+p+q)^n}\right| \mathrm{d}q\,\mathrm{d}p, \end{aligned} \tag{7}$$

where $\mathcal{D}(P,Q)$ stands for the domain of the random vector (P,Q).

Now, taking into account that $X_n^2 = 1 - X_n^1$ for every n, and applying Lemma 2, the 1-PDF of X_n^2 is given by

$$\begin{aligned} &f_1^{x^2}(x;n) = f_1^{x^1}(1-x;n) =\\ &= \iint_{\mathcal{D}(P,Q)} f_{x_0^1,P,Q}\left(\tfrac{(1-x)(-2+p+q)+(-1+q)(-1+(-1+p+q)^n)}{(-1+p+q)^n(-2+p+q)},p,q\right)\left|\tfrac{1}{(-1+p+q)^n}\right| \mathrm{d}q\,\mathrm{d}p. \end{aligned} \tag{8}$$

 ISBN: 978-84-617-8694-7

An important issue in dealing with Markov chains is to determine the steady state. From the deterministic theory it is known that the steady state to the Markov chain (1) is

$$x_\infty = \begin{pmatrix} \dfrac{1-q}{2-p-q} \\ \dfrac{1-p}{2-p-q} \end{pmatrix}. \tag{9}$$

By defining an appropriate mapping based on the expression (9) and using RVT technique, it can be shown that the PDF corresponding to the first component of the steady state is

$$f_{x_\infty^1}(x) = \int_{\mathcal{D}(Q)} f_{P,Q}\left(\frac{-1-x(2+q)+q}{x}, q\right)\left|\frac{1-q}{x^2}\right| \mathrm{d}q. \tag{10}$$

2.2 Distribution of time until a given proportion of the subpopulation is reached

It is useful to know when the percentage of a group in the population will attain a certain level. This motivates the computation of the distribution of the time, N_i, $i = 1, 2$, until a given proportion, ρ_i, of the population of state i is reached. Now, we concentrate on the computation of N_1 corresponding to the first subpopulation. Then, let us consider the following relation obtained from the first component of equation (6) one gets

$$\rho_1 = \frac{-1+q+(-1+p+q)^{n_1}(1-q+(-2+p+q)x_0^1)}{-2+p+q}. \tag{11}$$

In order to obtain the 1-PDF, $f_{N_1}(n)$, we first isolate n_1 from equation (11) and use the capital letter notation for random inputs X_0^1, P and Q. This leads to

$$N_1 = \frac{\log\left(\frac{1-Q+(-2+P+Q)\rho_1}{1-Q+(-2+P+Q)x_0^1}\right)}{\log(-1+P+Q)}. \tag{12}$$

This RV represents the time until a percentage ρ_1 of the subpopulation 1 has been reached. Using the RVT technique with an appropriate mapping inspired in (11), it can be proved that

$$\begin{aligned} f_{N_1}(n) &= \iint_{\mathcal{D}(P,Q)} f_{x_0^1,P,Q}\left(\tfrac{(-1+p+q)^{-n}(1-q+(-1+q)(-1+p+q)^n+\rho_1(-2+p+q))}{-2+p+q}, p, q\right) \\ &\times \left|\tfrac{(-1+p+q)^{-n}(-1+q-(-2+p+q)\rho_1)\log(-1+p+q)}{-2+p+q}\right| \mathrm{d}q\,\mathrm{d}q. \end{aligned} \tag{13}$$

 ISBN: 978-84-617-8694-7

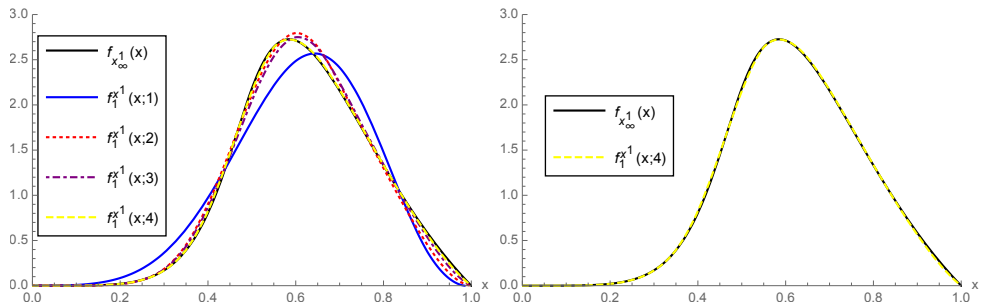


Figure 2: Left: 1-PDF of the solution SP, X_n^1, with $n \in \{1,2,3,4\}$ and the PDF of the equilibrium RV X_∞^1. Right: 1-PDF of the solution SP, X_n^1, with $n = 4$ and the PDF of the equilibrium RV X_∞^1.

3 Example

This section is addressed to show an example where the previous theoretical results are illustrated. With this aim, we need to choose the distributions of RV's X_0^1, P and Q. For the sake of clarity in the presentation, hereinafter we will assume that these RV's are independent. Therefore, its joint PDF can be factorized as the product of each individual PDF, $f_{X_0^1,P,Q}(x_0^1,p,q) = f_{X_0^1}(x_0^1)f_P(p)f_Q(q)$. On the one hand, as X_0^1 is the initial percentage of population in the state 1 (subpopulation 1), it lies between 0 and 1, then it is plausible to assume that X_0^1 has a Uniform distribution on the interval $[0,1]$. On the other hand, as P and Q represent the probabilities to remain in the states 1 (subpopulation 1) and 2 (subpopulation 2), respectively, then interval $[0,1]$ can represent their domains. Therefore, Beta is a flexible biparametric probability distribution for these two RV's, and this motivates that in the sequel we take $P \sim \text{Be}(3;2)$ and $Q \sim \text{Be}(3;5)$.

In Figure 2 the 1-PDF, $f_1^{X^1}(x;n)$, of the solution SP, X_n^1, for different values of $n \in \{1,2,3,4\}$ and the PDF, $f_{X_\infty^1}(x)$, of the equilibrium point, X_∞^1, are shown. As it has been previously pointed out, in this graphical representation one can observe that $f_1^{X^1}(x;n)$ tends to $f_{X_\infty^1}(x)$ when n increases. In Figure 2, we have plotted separately the PDF's $f_1^{X^1}(x;4)$ and $f_{X_\infty^1}(x)$ to better highlight this behaviour. On the right side of Figure 2, we can observe that both PDF's match. In Table 1, we report the total error of approximation $f_1^{X^1}(x;n)$ to $f_{X_\infty^1}(x)$ as n increases. This error is calculated as

$$e_n = \int_0^1 \left| f_1^{X^1}(x;n) - f_{X_\infty^1}(x) \right| \mathrm{d}x, \quad n = 1,2,\ldots \tag{14}$$

In Figure 3, the mean of the solution SP, X_n^1, and the threshold computed by the mean of the equilibrium have been represented. As it occurs with the 1-PDF, now we can observe

	$n=1$	$n=2$	$n=3$	$n=4$
e_n	0.155419	0.0508478	0.0405023	0.00884491

Table 1: Values of error e_n given by (14) for different cycles, $n \in \{1,2,3,4\}$.

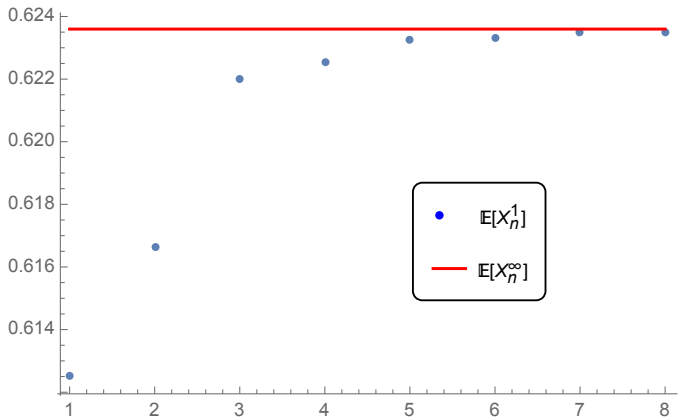


Figure 3: Blue points: $\mathbb{E}\left[X_n^1\right]$ for different $n \in \{1,2,\ldots,8\}$. Red line: $\mathbb{E}\left[X_\infty^1\right]$.

that $\mathbb{E}\left[X_n^1\right]$ tends to $\mathbb{E}\left[X_\infty^1\right]$ as n increases. This fact can be numerically observed in Table 2, where the error measured as the absolute value of the difference between $\mathbb{E}\left[X_n^1\right]$ and $\mathbb{E}\left[X_\infty^1\right]$, has been calculated for $n \in \{1,2,\ldots,8\}$.

	$n=1$	$n=2$	$n=3$	$n=4$
Error	0.0110972	0.00693051	0.00157336	0.00108635
	$n=5$	$n=6$	$n=7$	$n=8$
Error	0.00035583	0.000262173	0.000106079	0.0000817976

Table 2: Error between $\mathbb{E}\left[X_n^1\right]$ and $\mathbb{E}\left[X_\infty^1\right]$ for different cycles, $n \in \{1,2,\ldots,8\}$.

4 Conclusions

In this paper we have provided a full probabilistic description of random binary Markov chains under very general assumptions on the random inputs. These input are the probabilities of the transition matrix and the initial conditions. That description has been made

 ISBN: 978-84-617-8694-7

through the first probability density function of the discrete solution stochastic process and the probability density function associated to the steady state. Furthermore, the probability density function of a key time having specific interpretation in practice has been determined. Other quantities of great interest in the deterministic context of Markov chains, like first passage time, can be randomized using our approach. This permits the applications of the theoretical results established in this contribution when dealing with real models.

Acknowledgements

This work has been partially supported by the Ministerio de Economía y Competitividad grant MTM2013-41765-P. Ana Navarro Quiles acknowledges the doctorate scholarship granted by Programa de Ayudas de Investigación y Desarrollo (PAID), Universitat Politècnica de València.

References

[1] E. Behrends. *Introduction to Markov Chains: With Special Emphasis on Rapid Mixing*. Advanced Lectures in Mathematics. Vieweg+Teubner Verlag, 1999.

[2] J.-C. Cortés, A. Navarro-Quiles, J.-V. Romero, and M.-D. Roselló. Full solution of random autonomous first-order linear systems of difference equations. application to construct random phase portrait for planar systems. *Applied Mathematics Letters*, 68:150 – 156, 2017.

[3] J.-C. Cortés, A. Navarro-Quiles, J.-V. Romero, and M.-D. Roselló. Randomizing the parameters of a markov chain to model the stroke disease: A technical generalization of established computational methodologies towards improving real applications. *Journal of Computational and Applied Mathematics*, pages –, 2017,in press.

[4] J. Mar, F. Antoñanzas, R. Pradas, and A. Arrospide. Los modelos de Markov probabilísticos en la evaluación económica de tecnologías sanitarias: una guía práctica. *Gaceta Sanitaria*, 24(3):209–214, 2010.

[5] B. Sericola. *Markov Chains: Theory and Applications*. Applied stochastic methods series. Wiley, 2013.

[6] T. T. Soong. *Random Differential Equations in Science and Engineering*. Academic Press, New York, 1973.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

Computing the first probability density function of non-autonomous linear random differential equations by Karhunen-Loêve expansion

J.-C. Cortés[1], A. Navarro-Quiles[1], J.-V. Romero[1] and M.-D. Roselló[1]

[1] *Instituto Universitario de Matemática Multidisciplinar,*
Universitat Politècnica de València
Camino de Vera s/n, 46022, Valencia, Spain

emails: jccortes@imm.upv.es, annaqui@doctor.upv.es, jvromero@imm.upv.es, drosello@imm.upv.es

Abstract

Linear differential equations are important to model problems in many disciplines. When its input parameters are treated as random variables, the solution of a random differential equation is a stochastic process. Apart from obtaining the solution stochastic process, it is also important to determine its first probability density function. The first probability density function is important because it provides a comprehensive probabilistic description of the solution and from it one can calculate the mean and the variance functions and other important moments as well. In this paper we deal with the computation of approximations of the first-order probability density function to homogeneous variable-coefficient linear differential equations. This problem is studied under very general assumptions on both the diffusion coefficient and the initial conditions, which are assumed to be a stochastic process and a random variable, respectively. To achieve this goal, the Karhunen-Loève expansion and Random Variable Transformation technique will be applied. Our findings are illustrated by means of a numerical example.

Key words: Random non-autonomous linear differential equations, Random Variable Transformation, Karhunen-Loève expansion.

1 Introduction

Linear differential equations, which depend on a number parameters, are important to model some problems in many disciplines as Physics or Engineering, for example [1]. Although

most of differential equations that appear in practice are nonlinear, linearization can be used to study the local stability of an equilibrium point of a system of nonlinear differential equations or discrete dynamical systems [2]. This method is commonly used in fields such as engineering, Physics, Economics, Ecology, etc.

Coefficients and initial conditions of differential equations are treated as deterministic constants in spite of being usually obtained by measurements and experiments. Then, they contain errors. For this reason, it is more realist to treat these quantities as random variables (RVs) or, when depending on other variables as time and/or space, as stochastic processes (SPs).

A great part of the extant literature dealing with differential equations with uncertainty focuses on the study of Stochastic Differential Equations (SDEs). In that case, randomness is considered by means of special classes of SPs like markovian processes or, even more specific as the Wiener process (also designed as brownian motion). This latter case restricts the uncertainty to gaussian processes with irregular sample behaviour. This class of SDEs are usually referred to as Itô-type SDEs. Solving SDEs is based on Itô calculus [3, 4]. Apart from obtaining the solution SP, solving a SDE also means to determine the mean and the variance of the solution. Complementary to SDEs, uncertainty can be considered through a wider type of probabilistic distributions as binomial, Poisson, beta, gamma, etc. including gaussian, but having milder sample behaviour. This class of equations are referred to as Random Differential Equations (RDEs) and their rigorous analysis is usually based on a he so-called Mean Square calculus [5, 6] whose convergence is termed mean square convergence.

It is worthy to point out that the extension of results of deterministic differential equations to RDEs is not immediate at all. With the aim of motivating this assertion, and to justify our subsequent study for the linear non-autonomous RDE, next we state some well-know in the deterministic framework that do not fulfil in the random context unless additional assumptions are included. Indeed, let us consider the first-order linear RDE

$$x'(t;\omega) = a(\omega)x(t;\omega),\ t \geq 0; \quad x(0;\omega) = 1,$$

where $a = a(\omega)$ is a second-order RV, i.e. having finite variance defined on a complete probability space $(\Omega, \mathcal{F}, \mathbb{P})$. In [5, 7], it has been shown that the extension of the classical Lipschitz condition to mean square calculus in order to prove the uniqueness of the solution to random initial problem (IVP) is satisfied if, and only if, $a = a(\omega)$ is bounded almost surely. As a consequence, this important result is not applicable when $a = a(\omega)$ assumes a Gaussian or a Poisson probability distribution. As another example that highlights deep differences between deterministic and random differential equations, let us consider the following non-homogeneous linear RDE

$$x'(t;\omega) = a(t;\omega)x(t;\omega) + b(t;\omega), \quad x(t_0;\omega) = x_0(\omega). \tag{1}$$

Then, the mean of the solution, $\mu_x(t) = \mathbb{E}[x(t;\omega)]$, does not satisfy the averaged differential

 ISBN: 978-84-617-8694-7

equation (see [5, 8])

$$\frac{\mathrm{d}\,\mu_x(t;\omega)}{\mathrm{d}\,t} = \mathbb{E}[a(t;\omega)]\mu_x(t) + \mathbb{E}[b(t;\omega)], \quad \mu_x(t_0;\omega) = \mathbb{E}[x_0(\omega)].$$

In [9] a random power series solution to random IVP (1) is constructed and the main statistical functions (mean and variance) are calculated using mean square calculus as well.

Besides the calculation of first statistical moments, such as the mean and the variance, the computation of the first probability density function (1-PDF) of the solution SP is much more desirable since, from it, one can compute all the statistical moments. Furhtemore, the 1-PDF provides a comprehensive probabilistic description of the solution SP for each time instant t and it permits compute the probability that the solution lies on a certain set of specific interest.

In a previous contribution [10], some of the authors have provided a closed-form expression for the 1-PDF of the solution SP to first-order autonomous linear RDE whose input parameters (coefficients and initial condition) are assumed to be RVs rather than deterministic constants. This contribution is based on the application of the so-called Random Variable Transformation (RVT) technique. Now, we will reach the next level by considering that the diffusion coefficient of an homogeneous linear RDE is a SP instead of a RV. Therefore, the problem is formulated as follows

$$\left.\begin{array}{rcl} x'(t,\omega) & = & a(t,\omega)x(t,\omega), \quad t \in \mathcal{T} \subset \mathbb{R}^+ \\ x(t_0,\omega) & = & x_0(\omega), \end{array}\right\} \tag{2}$$

where $a : \mathcal{T} \times \Omega \to \mathbb{R}$ is a real-valued SP and $x_0 : \Omega \to \mathbb{R}$ is a real-valued RV. Hereinafter, $\mathrm{L}^2(\Omega, \mathrm{L}^2(\mathcal{T}))$ will denote the Hilbert space of square integrable real-valued SPs [11]. Our main goal is to obtain the 1-PDF, $f_1(x,t)$, of the solution SP, $x(t;\omega)$, to the random IVP (2). This objective will be achieved by applying RVT method together with Karhunen-Loève expansion. Despite of the simplicity in the statement of random IVP (2), as it shall be seen later the answer to this question is far from trivial. Moreover, as far as we know this problem has not be treated yet.

Karhunen-Loève expansion (KLE) is a Fourier type series that permits to represent a SP in $\mathrm{L}^2(\Omega, \mathrm{L}^2(\mathcal{T}))$ by a denumerable set of uncorrelated RVs, $\{\xi_i\}_{i=1}^{\infty}$, with mean zero and unit variance. KLE is a generalization of the spectral decomposition of real-valued symmetric matrices [11].

Theorem 1 (L^2 convergence of Karhunen-Loève) *([11, p.202]) Consider a stochastic process $\{y(t), t \in \mathcal{T}\}$ and suppose that $y(t) \in \mathrm{L}^2(\Omega, \mathrm{L}^2(\mathcal{T}))$. Then,*

$$y(t,\omega) = \mu_y(t) + \sum_{j=1}^{\infty} \sqrt{\nu_j}\phi_j(t)\xi_j(\omega), \quad \omega \in \Omega, \tag{3}$$

 ISBN: 978-84-617-8694-7

where the sum converges in $\mathrm{L}^2(\Omega, \mathrm{L}^2(\mathcal{T}))$,

$$\xi_j(\omega) := \frac{1}{\sqrt{\nu_j}} \left\langle y(t,\omega) - \mu_y(t), \phi_j(t) \right\rangle_{\mathrm{L}^2(\mathcal{T})},$$

being $\mu_y(t) = \mathbb{E}[y(t)]$ *the mean of* $y(t)$ *and* $\{\nu_j, \phi_j\}$ *denote, respectively, the eigenvalues with* $\nu_1 \geq \nu_2 \geq \cdots \geq 0$ *and eigenfunctions of the covariance function,* $C_y(s,t)$ *in*

$$(\mathfrak{C}f)(t) := \int_{\mathcal{T}} C_y(s,t) f(s) \mathrm{d}s, \quad \textit{for } f \in L^2(\mathcal{T}),$$

where $\mathfrak{C}$ *is the integral operator.*

Random variables $\xi_j(\omega)$ *have mean zero, unit variance and are pairwise uncorrelated. If the process is Gaussian, then* $\xi_j \sim N(0,1)$ *i.i.d.*

To apply the RVT technique, we will need to consider the truncation of order N of the infinite series (3), i.e.

$$y_N(t,\omega) = \mu_y(t) + \sum_{j=1}^{N} \sqrt{\nu_j} \phi_j(t) \xi_j(\omega), \quad \omega \in \Omega. \tag{4}$$

In this manner, only the following $N+1$ RVs, $x_0(\omega)$ and $\{\xi_i(\omega) : 1 \leq i \leq N\}$, will be involved in dealing with our approximate problem thas results after substituting the diffusion coefficient $a(t,\omega)$ by the truncated KLE of order N in the original random IVP (2). Then, we will obtain the 1-PDF, $f_1^N(x,t)$, of the solution to the approximate random IVP. The last step will be to prove that $f_1^N(x,t)$ converges to the exact 1-PDF $f_1(x,t)$ under mild conditions. To do this, the following result will be required.

Theorem 2 (Uniform convergence of Karhunen-Loève) *[11, p.203] Consider a stochastic process* $y(t,\omega) \in \mathrm{L}^2(\Omega, \mathrm{L}^2(\mathcal{T}))$, *and let* $y_J(t,\omega)$ *be the stochastic process defined in* (4). *If* $\mathcal{T} \subset \mathbb{R}$ *is a closed and bounded set and the covariance function of* $y(t,\omega))$, $C_y(s,t)$ *is continuous,* $C_y \in \mathbf{C}(\mathcal{T} \times \mathcal{T})$, *then* $\phi_j \in \mathbf{C}(\mathcal{T})$ *and the series expansion of* C_y *converges uniformly. In particular,*

$$\sup_{s,t \in \mathcal{T}} |C_y(s,t) - C_{y,J}(s,t)| \leq \sup_{t \in \mathcal{T}} \sum_{j=J+1}^{\infty} \nu_j \phi_j(t)^2 \to 0, \quad \textit{as } J \to \infty,$$

where $C_{y,J}$ *is the covariance function of the SP* $y_J(t,\omega)$ *defined by*

$$C_{y,J}(s,t) = \sum_{j=1}^{J} \nu_j \phi_j(s) \phi_j(t).$$

Moreover,

$$\sup_{t \in \mathcal{T}} \mathbb{E}\left[(y(t,\omega) - y_J(t,\omega))^2 \right] \to 0, \quad \textit{as } J \to \infty.$$

 ISBN: 978-84-617-8694-7

For the sake of completeness, now we state the RVT technique that will be used throughout our analysis. As previously indicated this result will allow us to determine the 1-PDF $f_1^N(x,t)$ of the truncated solution SP in terms of the joint PDF of the random vector $(x_0(\omega), \xi_1(\omega), \ldots, \xi_N(\omega)$, which will be assumed known.

Theorem 3 (Random Variable Transformation method) *([5, 12]) Let us consider* $\mathbf{x}(\omega) = [x_1(\omega), \ldots, x_m(\omega)]^{\mathsf{T}}$ *and* $\mathbf{y}(\omega) = [y_1(\omega), \ldots, y_m(\omega)]^{\mathsf{T}}$ *be two* m*-dimensional absolutely continuous random vectors defined on a complete probability space* $(\Omega, \mathfrak{F}, \mathbb{P})$*. Let* $\mathbf{r} : \mathbb{R}^m \to \mathbb{R}^m$ *be a one-to-one deterministic transformation of* $\mathbf{x}(\omega)$ *into* $\mathbf{y}(\omega)$*, i.e.,* $\mathbf{y}(\omega) = \mathbf{r}(\mathbf{x}(\omega))$ *for each* $\omega \in \Omega$*. Assume that* $\mathbf{r}$ *is continuous in* $\mathbf{x}(\omega)$ *for every for each* $\omega \in \Omega$ *fixed, and has continuous partial derivatives with respect to each* $x_i \equiv x_i(\omega)$*,* $1 \leq i \leq m$*, for every* $\omega \in \Omega$*. Then, if* $f_{\mathbf{x}(\omega)}(\mathbf{x})$ *denotes the joint probability density function of random vector* $\mathbf{x}(\omega)$*, and* $\mathbf{s} = \mathbf{r}^{-1} = (s_1(y_1, \ldots, y_m), \ldots, s_m(y_1, \ldots, y_m))$ *represents the inverse mapping of* $\mathbf{r} = (r_1(x_1, \ldots, x_m), \ldots, r_m(x_1, \ldots, x_m))$*, the joint probability density function of vector* $\mathbf{y}(\omega)$ *is given by*

$$f_{\mathbf{y}(\omega)}(\mathbf{y}) = f_{\mathbf{x}(\omega)}\left(\mathbf{s}(\mathbf{y})\right) |J_m| \,, \tag{5}$$

where $|J_m|$*, which is assumed to be different from zero, denotes the absolute value of the Jacobian defined by the determinant*

$$J_m = \det \begin{bmatrix} \dfrac{\partial s_1(y_1, \ldots, y_m)}{\partial y_1} & \cdots & \dfrac{\partial s_m(y_1, \ldots, y_m)}{\partial y_1} \\ \vdots & \ddots & \vdots \\ \dfrac{\partial s_1(y_1, \ldots, y_m)}{\partial y_m} & \cdots & \dfrac{\partial s_m(y_1, \ldots, y_m)}{\partial y_m} \end{bmatrix} . \tag{6}$$

The paper is organized as follows. Section 2 is addressed to compute the 1-PDF, $f_1^N(x,t)$, of the truncated solution SP and also to study under what conditions there is convergence in distribution to the 1-PDF of the exact solution SP, $f_1(x,t)$. Our findings are illustrated with a numerical in Section 3. Conclusions are drawn in Section 4.

2 Computing the 1-PDF of the truncated solution stochastic process

It is known that the exact closed solution to the IVP (2) is

$$x(t,\omega) = x_0(\omega) \mathrm{Exp}\left[\int_{t_0}^{t} a(s,\omega) \mathrm{d}s\right], \quad \omega \in \Omega. \tag{7}$$

It is important to note that given a SP $a(t)$, in general nothing is known about the distribution of the random integral $\int_{t_0}^{t} a(s)\,\mathrm{d}s$. An exception is the integrand is a Gaussian

 ISBN: 978-84-617-8694-7

SP. In this case, the integral is also a Gaussian SP, see [5]. This is the reason why Karhunen-Loève expansion is used here.

Let us first consider the truncated Karhunen-Loève expansion of SP $a(t,\omega)$ (see (1))

$$a_N(t,\omega) = \mu_a(t) + \sum_{j=1}^{N} \sqrt{\nu_j}\phi_j(t)\xi_j(\omega), \quad \omega \in \Omega, \tag{8}$$

where $\mu_a(t)$ represents the mean of the SP $a(t,\omega)$.

Secondly, we substitute (8) in (7). Then, the approximated solution SP of the random IVP (2) is

$$\begin{aligned} x_N(t,\omega) &= x_0(\omega)\mathrm{Exp}\left[\int_{t_0}^{t} a_N(s,\omega)\mathrm{d}s\right] \\ &= x_0(\omega)\mathrm{Exp}\left[\int_{t_0}^{t}\left(\mu_a(s) + \sum_{j=1}^{N}\sqrt{\nu_j}\phi_j(s)\xi_j(\omega)\right)\mathrm{d}s\right]. \end{aligned} \tag{9}$$

Now, fixed $t \in \mathcal{T}$, we will apply the RVT method (see Theorem 3), to obtain the PDF of the approximate solution SP (9) in function of the joint PDF of the random vector $\boldsymbol{\xi}_{N+1}(\omega) = (x_0(\omega), \xi_1(\omega), \ldots, \xi_N(\omega))$, $f_{\boldsymbol{\xi}_{N+1}(\omega)}(\boldsymbol{\xi}_{N+1})$, which is assumed to be known. To this end, we consider the following mapping $\mathbf{r} : \mathbb{R}^{N+1} \to \mathbb{R}^{N+1}$, whose components are r_i, $1 \le i \le N+1$,

$$\begin{aligned} y_1 &= r_1(x_0,\xi_1,\ldots,\xi_N) &&= x_0\ \mathrm{Exp}\left[\int_{t_0}^{t}\left(\mu_a(s) + \sum_{j=1}^{N}\sqrt{\nu_j}\phi_j(s)\xi_j\right)\mathrm{d}s\right], \\ y_2 &= r_2(x_0,\xi_1,\ldots,\xi_N) &&= \xi_1, \\ &\vdots \\ y_{N+1} &= r_{N+1}(x_0,\xi_1,\ldots,\xi_N) &&= \xi_N. \end{aligned}$$

Its inverse transformation, $\mathbf{s} = \mathbf{r}^{-1}$, is

$$\begin{aligned} x_0 &= s_1(y_1,y_2,\ldots,y_{N+1}) &&= y_1\ \mathrm{Exp}\left[-\int_{t_0}^{t}\left(\mu_a(s) + \sum_{j=1}^{N}\sqrt{\nu_j}\phi_j(s)y_{j+1}\right)\mathrm{d}s\right], \\ \xi_1 &= s_2(y_1,y_2,\ldots,y_{N+1}) &&= y_2, \\ &\vdots \\ \xi_N &= s_{N+1}(y_1,y_2,\ldots,y_{N+1}) &&= y_{N+1}. \end{aligned}$$

It can be checked that the absolute value of the Jacobian is

$$|J_{N+1}| = \mathrm{Exp}\left[-\int_{t_0}^{t}\left(\mu_a(s) + \sum_{j=1}^{N}\sqrt{\nu_j}\phi_j(s)y_{j+1}\right)\mathrm{d}s\right] \neq 0.$$

 ISBN: 978-84-617-8694-7

Applying the RVT method (see Theorem 3), we obtain the joint PDF of the random vector $\mathbf{y}_{N+1} = (y_1(\omega), y_2(\omega), \ldots, y_{N+1}(\omega))$ in function of the joint PDF of the random vector $\boldsymbol{\xi}_{N+1}(\omega)$

$$\begin{aligned} f_{\mathbf{y}_{N+1}(\omega)}(\mathbf{y}_{N+1}) &= f_{\boldsymbol{\xi}_{N+1}(\omega)}\left(y_1 \ \mathrm{Exp}\left[-\int_{t_0}^{t}\left(\mu_a(s)+\sum_{j=1}^{N}\sqrt{\nu_j}\phi_j(s)y_{j+1}\right)\mathrm{d}s\right], y_2, \ldots, y_{N+1}\right) \\ &\quad \times \mathrm{Exp}\left[-\int_{t_0}^{t}\left(\mu_a(s)+\sum_{j=1}^{N}\sqrt{\nu_j}\phi_j(s)y_{j+1}\right)\mathrm{d}s\right]. \end{aligned} \tag{10}$$

Finally, taking $t \in \mathcal{T}$ arbitrary and marginalizing expression (10), we obtain the 1-PDF of the truncated solution SP

$$\begin{aligned} f_1^N(x,t) &= \int_{\mathcal{D}(\boldsymbol{\xi}_N)} f_{\boldsymbol{\xi}_{N+1}}\left(x \ \mathrm{Exp}\left[-\int_{t_0}^{t}\left(\mu_a(s)+\sum_{j=1}^{N}\sqrt{\nu_j}\phi_j(s)\xi_j\right)\mathrm{d}s\right], \xi_1, \ldots, \xi_N\right) \\ &\quad \times \mathrm{Exp}\left[-\int_{t_0}^{t}\left(\mu_a(s)+\sum_{j=1}^{N}\sqrt{\nu_j}\phi_j(s)\xi_j\right)\mathrm{d}s\right] \mathrm{d}\xi_N \cdots \mathrm{d}\xi_1, \end{aligned} \tag{11}$$

where $\mathcal{D}(\boldsymbol{\xi}_N)$ denotes the domain of random vector $\boldsymbol{\xi}_N(\omega) = (\xi_1(\omega), \ldots, \xi_N(\omega))$.

The uniformly convergence of the sequence $\{f_1^N(x,t)\}$ in $\mathbb{R} \times \mathcal{T}$, i.e.

$$\lim_{N\to\infty} f_1^N(x,t) = f_1(x,t), \quad \forall (x,t) \in \mathbb{R} \times \mathcal{T}.$$

can be established.

3 Example

In this section we show a numerical example where the main results previously established are illustrated. In this example we assume that $a(t,\omega)$ is the standard Brownian motion, $B(t)$, and $t_0 = 0$, [11, 13]. Then, it is known that is mean, $\mu_a(t) = 0$, and its variance, $\mathbb{V}[a(t)] = 1$, $\forall t \in \mathcal{T} = [0,T]$. Also, the covariance function is given by

$$C_a(s,t) = \min(s,t), \quad (s,t) \in \mathcal{T} \times \mathcal{T},$$

 ISBN: 978-84-617-8694-7

which has the following eigenvalues and normalized eigenfunctions

$$\nu_j = \frac{4T^2}{\pi^2(2j-1)^2}, \quad \phi_j(t) = \sqrt{\frac{2}{T}} \sin\left(\frac{t\pi(2j-1)}{2T}\right), \quad j = 1, 2, \ldots \tag{12}$$

Then according to (11), the 1-PDF of the truncated solution SP, $x_N(t,\omega)$, is

$$f_1^N(x,t) = \int_{\mathcal{D}(\boldsymbol{\xi}_N)} f_{\boldsymbol{\xi}_{N+1}(\omega)} \left(x \prod_{j=1}^{N} \mathrm{e}^{-K_j(t)\xi_j}, \xi_1, \ldots, \xi_N \right) \prod_{j=1}^{N} \mathrm{e}^{-K_j(t)\xi_j} \, \mathrm{d}\xi_N \cdots \mathrm{d}\xi_1, \tag{13}$$

where

$$K_j(t) = \frac{4\sqrt{2}T^2}{\sqrt{T}\pi^2(2j-1)^2} \left(1 - \cos\left(\frac{t\pi(2j-1)}{2T}\right)\right).$$

In this example the 1-PDF of the exact solution SP, $x(t,\omega)$, can be computed. This allows us to check the accuracy of approximations constructed using our approach. Applying RVT technique and as $Z(t) = \int_0^t B(s)\mathrm{d}s \sim \mathrm{N}(0, t^3/3)$, the 1-PDF is given by

$$f_1(x,t) = \int_{-\infty}^{\infty} f_{X_0,Z}(x\,\mathrm{e}^{-z}, z)\,\mathrm{e}^{-z}\,\mathrm{d}z. \tag{14}$$

Hereinafter we will take $\xi_j(\omega)$, $j = 1, 2$, truncated standard Gaussian RVs in the interval $[-10, 10]$, $x_0(\omega)$ a uniform RV on the interval $[0, 1]$, $x_0(\omega) \sim \mathrm{Un}[0, 1]$ and we will assume that $x_0(\omega)$, $\xi_1(\omega)$ and $\xi_2(\omega)$ are independent RVs. In Table 1 the error given by the expression (15) is calculated for different times levels, $t \in \{0.1, 1, 2\}$ and truncation orders, $N = 1, 2$. We observe that the 1-PDF of the first truncation is close to the 1-PDF of the exact solution. Moreover, if the order of truncation increases this error decreases.

$$\mathrm{e}_N(t) = \int_{-\infty}^{\infty} \left| f_1(x,t) - f_1^N(x,t) \right| \mathrm{d}x. \tag{15}$$

$\mathrm{e}_N(t)$	$N=1$	$N=2$
$t=0.1$	0.019319	0.016788
$t=1$	0.077919	0.008663
$t=2$	0.005310	0.000832

Table 1: Error measure $\mathrm{e}_N(t)$ defined by (15) for different time instants, $t \in \{0.1, 1, 2\}$, and truncation orders, $N \in \{1, 2\}$.

 ISBN: 978-84-617-8694-7

4 Conclusions

In this paper we have proposed a method in order to compute the first probability density function of the solution stochastic process to the non-autonomous linear random differential equation with a random initial condition. The method is based on the application of both the Random Variable Transformation technique and the Karhunen-Loève expansion. To the best of our knowledge, thisis the first time that this approach has been considered. The method has been successfully tested through an example where the exact expression of the first probability density function of the solution is available. The proposed technique is promising, and we will study its application to complex random differential equations in our forthcoming research.

Acknowledgements

This work has been partially supported by the Ministerio de Economía y Competitividad grant MTM2013-41765-P. Ana Navarro Quiles acknowledges the doctorate scholarship granted by Programa de Ayudas de Investigación y Desarrollo (PAID), Universitat Politècnica de València.

Conflict of Interest Statement

The authors declare that there is no conflict of interests regarding the publication of this article.

References

[1] W. Xie, Differential Equations for Engineers, Cambridge University Press, 2010. URL `https://books.google.co.bw/books?id=II2QT2yXsucC`

[2] D. Cheng, X. Hu, T. Shen, Analysis and Design of Nonlinear Control Systems, Springer-Verlag, 2010.

[3] H. Holden, B. Øksendal, J. Ubøe, T. Zhang, Stochastic Partial Differential Equations A Modeling, White Noise Functional Approach, Springer-Verlag, New York, 2010. doi:10.1007/978-0-387-89488-1.

[4] P. E. Kloeden, E. Platen, Numerical Solution of Stochastic Differential Equations, 3rd Edition, Vol. 23 of Applications of Mathematics: Stochastic Modelling and Applied Probability, Springer, New York, 1999.

[5] T. T. Soong, Random Differential Equations in Science and Engineering, Academic Press, New York, 1973.

[6] T. Neckel, F. Rupp, Random Differential Equations in Scientific Computing, Versita, London, 2013.

[7] J. Strand, Random ordinary differential equations, Journal of Differential Equations 7 (3) (1970) 538 – 553. doi:http://dx.doi.org/10.1016/0022-0396(70)90100-2.
URL http://www.sciencedirect.com/science/article/pii/0022039670901002

[8] T. C. Gard, Introduction to Stochastic Differential Equations, Marcel Dekker, New York, 1988.

[9] G. Calbo, J. C. Cortés, L. Jódar, Mean square power series solution of random linear differential equations, Applied Mathematics and Computation 59 (1) (2010) 559–572. doi:10.1016/j.camwa.2009.06.007.

[10] M. C. Casabán, J. C. Cortés, J. V. Romero, M. D. Roselló, Determining the first probability density function of linear random initial value problems by the Random Variable Transformation (RVT) technique: A comprehensive study, Abstract and Applied Analysis 2014–ID248512 (2014) 1–25. doi:10.1155/2013/248512.

[11] G. Lord, C. Powell, T. Shardlow, An Introduction to Computational Stochastic PDEs, Cambridge Texts in Applied Mathematics, Cambridge University Press, 2014.

[12] A. Papoulis, S. U. Pillai, Probability, Random Variables and Stochastic Processes, 4th Edition, McGraw-Hill, New York, 2002.

[13] R. Ghanem, P. Spanos, Stochastic Finite Elements: A Spectral Approach, Dover Civil and Mechanical Engineering, Dover Publications Inc., 2012.

 ISBN: 978-84-617-8694-7

Proceedings of the 16th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

A parallel genetic algorithm for continuous and pattern-free heliostat field optimization

N.C. Cruz[1], S. Salhi[2], J.L. Redondo[1], J.D. Álvarez[1], M. Berenguel[1] and P.M. Ortigosa[1]

[1] *Dpt. of Informatics, ceiA3-CIESOL, University of Almería, Spain*

[2] *Centre for Logistics and Heuristic Optimisation (CLHO), University of Kent, Canterbury, UK*

emails: ncalvocruz@ual.es, s.salhi@kent.ac.uk, jlredondo@ual.es, jhervas@ual.es, beren@ual.es, ortigosa@ual.es

Abstract

The heliostat field of a solar power tower system, considering both its deployment cost and potential energy loss at operation, must be carefully designed. This procedure implies facing a complex continuous, constrained and large-scale optimization problem. Hence, its resolution is generally wrapped by extra distribution patterns or layouts with a reduced set of parameters. Griding the available surface is also an useful strategy. However, those approaches limit the degrees of freedom at optimization. In this context, the authors of this work are working on a new meta-heuristic for heliostat field optimization by directly addressing the underlying problem. Attention is also given to the benefits of modern High-Performance Computing (HPC) to allow a wider exploration of the search-space. Thus, a parallel genetic optimizer has been designed for direct heliostat field optimization. It relies on elitism, uniform crossover, static penalization of infeasible solutions and tournament selection.

Key words: heliostat field optimization, genetic algorithm, parallelization

1 Introduction

Solar central receiver systems, SCRS in what follows, are one the most promising flagships in the field of solar energy for large-scale electricity production. This is mainly due to the high thermodynamic efficiency and power output stability that can be achieved by these

 ISBN: 978-84-608-6082-2

systems [2, 6]. For the scope of this work, SCRS can be defined as a large set of high reflectance mirrors, called 'heliostats', and a radiation receiver on top of a tower. Heliostats feature an orientable structure and a tracking system that make them follow the apparent solar movement throughout days while also reflecting the incident radiation on the receiver. This energy is then progressively transferred to a working fluid which circulates inside the receiver. Finally, once the temperature of the fluid is high enough, it can be ultimately used in a turbine cycle to generate electricity. In Fig. 1, an illustrative depiction of this kind of systems is shown. Further information of SCRS can be found in [1, 16].

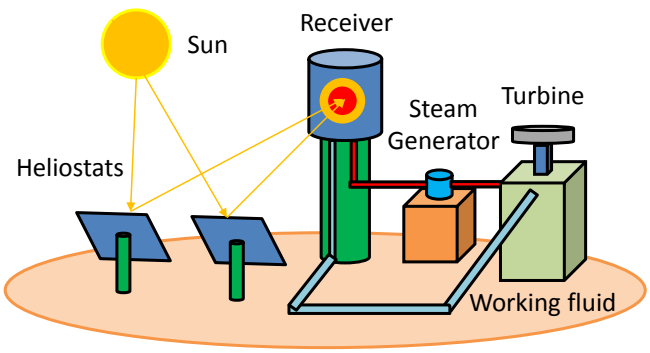


Figure 1: Scheme of a solar tower power plant.

The heliostat field supposes up to 50% of the initial investment costs and can cause up to 40% of energetic loss at operation [3, 6]. Consequently, its design must be carefully optimized. Some valid criteria to do so are the total intercepted power [8] (which is the objective chosen in this work), optical efficiency [3, 10] (which can be seen as a ratio between the really intercepted power and the theoretical maximum), land use [10, 15] and price of production [12]. Heliostat field optimization is a complex problem in which [5]: i) the coordinates of every heliostat should be directly considered (and commercial fields have, at least, several hundreds of heliostats) with placement constraints and ii) the objective function is computationally expensive and multi-modal without a direct solid approach. According to the concise and clear analysis done in [5, 8], most design methodologies relies on a) well-known parametric distribution patterns and b) selection of positions. Besides, these approaches could also be combined and expanded with what could be called 'special strategies'. Regarding known distribution patterns, there are some classic ones such as the popular radial-staggered scheme (used in [6, 10, 11]). There are also some recent and interesting proposals such as the parametric layout described in [12], which mainly expands the concept of radial staggering with more degrees of freedom. In [10], apart from trying to optimize a radial-stagger distribution, an innovative pattern is also proposed, a bio-inspired spiral later used in [3]. In relation to greedy selection of positions, the method of [15], in

 ISBN: 978-84-608-6082-2

which a grid is formed on the available surface, is widely known. The approach proposed in [5] is similar to that strategy. In fact, the concept of selection is also generally applied to pattern-based fields by generating larger layouts than needed and ultimately selecting the best heliostats as investigated in [3, 10, 11, 12]. Hence, layouts could be even seen as 'smart' ways of generating available positions. This is a good way to partially overcome the constrained perspective of parametric layouts as commented in [12]. Additionally, what has been tagged as 'special strategies', without considering some of the previous ones which also combine several ideas/steps, could contain the methods proposed in [4, 8] for their enhanced flexibility. In [4], an interesting procedure in which the refinement of a preliminary design is presented. In [8], based on some works exposing the sub-optimality of parametric approaches, the whole complexity of the problem is directly addressed with a classical optimization approach and a gradient-based method.

The authors of this work have defined a new meta-heuristic for heliostat field design. It aims to address the continuous and constrained large-scale problem of adjusting the coordinates of very heliostat through gradually altering the search-space shown to any selected optimizer. In this paper, the design of an optimizer specially designed for the problem at hand is described and commented. It is a genetic algorithm that has been designed with the main aim of keeping a simple but robust and parallel structure. We are aware that this type of algorithms have already been used for heliostat field optimization but i) their interesting theoretical principles worth its application in our specific context and, ii) their use is commonly focused on relatively reduced set of parameters as in [3, 4, 11, 12]. The underlying premise of design is: 'The more solutions can be explored per unit of time, the better results can be obtained'. In fact, the objective function is complex and requires simulating several candidate fields, which can be computationally expensive. Furthermore, as directly working with coordinates, it is expected that numerous cycles will be needed to achieve good solutions. Nevertheless, the procedure also relies on known principles to properly converge. Specifically, elitism, tournament selection and uniform crossover. Besides, points of potential knowledge-injection are also highlighted. This paper is organized as follows: In Section 2, the optimization problem is formally described in terms of maximizing the total power reflected on the receiver. Then, in Section 3, the genetic optimizer is described in detail. Finally, in Section 4, conclusions are drawn and future work is planned.

2 Problem statement

As introduced, the problem at hand consist in placing a certain number of heliostats on a flat ground so that, they reflect the maximum power on a known receiver. Let H be the total number of heliostats to deploy. All are assumed to be of the same size and specifications (as usual, to benefit from large-scale production). Specifically, their reflective surface is

 ISBN: 978-84-608-6082-2

rectangular and has a size of $l \times w$ (from length and width, respectively). Every heliostat h can be identified on the field by its central point, $C_h = (x_h, y_h)$ (assuming Cartesian coordinates). Hence, a field of H heliostats can be defined as a vector $F = (C_1, \cdots, C_H) = (x_1, y_1, \cdots, x_H, y_H)$ in $\mathbb{R}^{2H}$.

Considering the previous definitions, the problem to solve will have $2H$ dimensions, i.e., two coordinates per heliostat. Let $P_T(F)$ be the total power effectively reflected by a certain field F on the receiver throughout a fixed set of T instants of interest (i.e., defined apparent solar positions), $T = \{t_1, \cdots, t_T\}$. Depending on the final application requirements, T can vary from a single one (for design-point optimization) to many ones encompassing, for instance, a whole year (which increase the complexity and, specially, the computational cost). $P_T(F)$, which will be the objective function, can be analytically defined as expressed in Eq. (1) [8, 9].

$$P_T(F) = A \sum_{t=t_1}^{T} I_t \left(\sum_{h=h_1}^{H} \eta_h(t) \right) \tag{1}$$

In relation to A, it is the reflective area of the heliostat model (approximately $l \times w$ (m^2)) and I_t is the incident radiation density at instant t (kW/m^2). Regarding $\eta_h(t)$, it is the instantaneous efficiency factor of heliostat h at instant t (from 0 to 1, minimum and maximum efficiencies respectively). This factor depends on i) the instant, ii) the position of heliostat h in relation to the receiver and iii) the other heliostats due to potential interactions. Its computation implies both simulating and analyzing the behavior of the candidate field. As clearly explained in [3, 10, 16], this factor models different sources of energy loss at operation. In fact, it is composed by different sub-factors (also in range [0,1]). The abstract definition selected in this work is shown in Eq. (2), and it is the same selected in [3, 9, 10].

$$\eta_h = \eta_{cos} \eta_{sb} \eta_{itc} \eta_{aa} \eta_{ref} \tag{2}$$

A brief summary of the components of Eq. (2) and the way in which they are computed is shown next:

- η_{cos} **(Cosine loss)**: The effective reflective area of a heliostat is reduced by the cosine of angle of incidence of solar radiation. It is computed as described in [16].
- η_{sb} **(Shading and blocking loss)**: Every heliostat can partially obstruct the radiation either incident (shading) or reflected (blocking) from any other one. It is computed as recently proposed in [13] with the method for candidate filtering used in [10]. This is the most computationally expensive part pf the function.
- η_{itc} **(Interception loss)**: The reflected flux map of heliostats might not perfectly fit the desired zone of the receiver. It is computed according to the model proposed in [9] (which avoids its temporal component for computational efficiency).

 ISBN: 978-84-608-6082-2

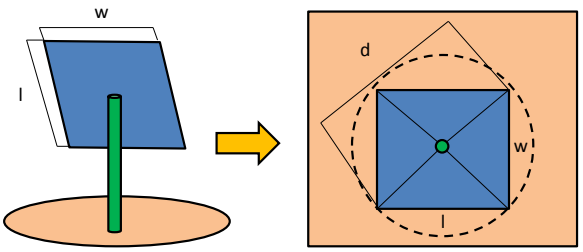


Figure 2: Characteristic safety distance for heliostats.

- η_{aa} **(Atmospheric attenuation loss)**: The atmosphere attenuates reflected radiation from heliostats along its trajectory. It is estimated by using the same model applied in [10] (which is also non-instant dependent).

- η_{ref} **(Reflectivity loss)**: Heliostats cannot grant a lossless reflection phenomenon. It is considered as a common fabrication constant as in [10].

Additionally, it must also be considered that: i) the receiver base will be at the origin of coordinates, ii) no heliostats can be neither nearer the receiver than R_{min} nor further than R_{max} and iii) heliostats should be able to freely move without colliding each other, where the safety distance d can be computed as $d = \sqrt{l^2 + w^2}$ (see Fig.2). At this point, the target optimization problem expressed in Eq. (3). Therefore, there are $H(H-1)/2 + 2H$ constrains to satisfy (as distance from any point a to b is the same that from b to a) and the potential problem dimensionality is really large.

$$\begin{aligned} \underset{F}{\text{maximize}} \quad & P_T(F) \\ \text{subject to} \quad & \sqrt{x^2+y^2} \geq R_{min} + \frac{d}{2}, \forall C = (x,y) \\ & \sqrt{x^2+y^2} \leq R_{max} - \frac{d}{2}, \forall C = (x,y) \\ & \sqrt{(x_i-x_j)^2+(y_i-y_j)^2} \geq d, \forall C_i = (x_i,y_i), C_j = (x_j,y_j) : (i \neq j) \end{aligned} \tag{3}$$

3 Method description

Genetic algorithms (GA), proposed by Holland in late seventies [7], are commonly used for complex global optimization problems. This is because their underlying theoretical principle is not linked to any particular problem but to the abstract evolution of species. Specifically,

 ISBN: 978-84-608-6082-2

a population of candidate solutions ('individuals') is generated and simulated to evolve (including interaction) until a certain halt condition met. In fact, as commented in Sec.1, they are usually considered for heliostat field optimization. In any case, the concept of GA (and even evolutionary computing in general) is quite wide and abstract. Hence, this kind of methods is ultimately adapted to the target problem, which usually determines both the selection of its base operators and their scope. Further information on population-based heuristic including GA, see [14].

In this work, a GA has been designed for the problem defined at Sec.2. Besides, as large populations are expected to be used it has the underlying aim of being able to run in parallel too. This strategy adapts the whole procedure to exploit modern high-performance environments, which is specially valuable to attenuate the potential cost of the objective function (specially for large fields and/or sets of instants).

The structure defined for the individuals is quite simple: They are vectors of length $2H$ with an additional field to record the fitness of that field design, i.e., the total power that it reflected on the receiver after simulation (evaluation of Eq. (1)). In Fig.3, the structure of individuals is depicted. Besides, it is important to remember, as included in that figure, that every pair of coordinates is linked to a certain heliostat. However, GA are mainly suited to unconstrained optimization [17] and as the problem at hand is constrained, some adaptations must be done. Specifically, the problem is treated as an unconstrained one and any unfeasible solution, i.e., those that does not respect all constraints, will be penalized with very low fitness. Penalization will only depend on the degree of violation, i.e., the more constraints are not respected (number and amount), the worse fitness is associated. This approach, which is quite common to handle constrained optimization problems with GA, is called 'static penalization' [17]. Thus, the designed method ignores the constraints shown in Eq. (3) but alter the evaluation of any candidate solution, F, as defined in Eq. 4. In that expression, m_c is the distance between heliostat c and the tower base, i.e., $m = \sqrt{x_c^2 + y_c^2}$, and I_T is the summation of solar radiation density at every instant in T. Finally, V is an abstract set which contains a record for every heliostat c in F and the constraints it violates (to compute only those factors). Every heliostat in V has also a special set linked, V_c listing any other heliostat z that is too near.

$$eval(F) = \begin{cases} P_T(F), & \text{if } V = \emptyset \\ 0 - AI_T \left(\frac{(R_{min}+d/2)-m_c}{(R_{min}+d/2)} + \frac{m_c-(R_{max}-d/2)}{m_c} \right. & \\ \left. + \frac{d-dist(C_c,C_z)}{d} \right) ; \forall c \in V, \forall z \in V_c & \text{otherwise} \end{cases} \quad (4)$$

Once the definition of candidate solutions and how constraints are handled, the genetic procedure can be described. It takes the input parameters listed below:

- pop_{size}: the population size, which will be kept constant during the search.
- num_{pairs}: the number of pairs to form for crossover.

 ISBN: 978-84-608-6082-2

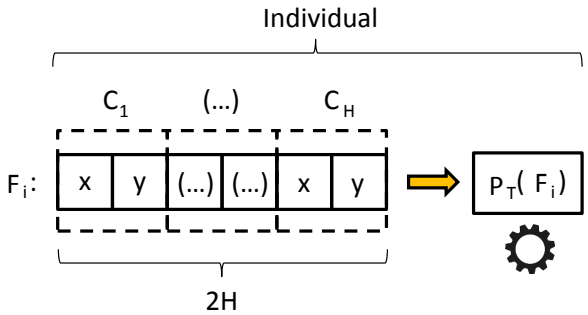


Figure 3: Diagram of an individual.

- $tourn_{size}$: the tournament size at any selection (both for crossover and replacement).
- ov_{mut}: the overall probability mutating a descendant.
- per_{mut}: the probability of altering every heliostat of a descendant once mutation started.
- *cycles*: the number of cycles, i.e., generations to run.

It must be noted that contextual information such as the description of the field, which is necessary to evaluate candidate solutions, is assumed implicitly available. With all this information, the algorithm described in Alg.1 is executed. As can be seen, it is a common evolutionary loop in which the parts that involve evaluating the objective function is distributed among concurrent threads (in a shared-memory environment). In fact, they all are forced to wait for the master to define the population of the next cycle before starting at line 12. By proceeding this way, threads can share the cost of evaluating the objective function (which is required at initialization (line 5), reproduction (line 9) and mutation (line 10)) while a consistent common population is maintained. Functions *createThreads*, *runInParallel* and *getChunkSize*, they are referred to the way in which a thread pool can be created and launched to work on different ranges of the population matrix. Similarly, tag *synchr* simple indicates that the update of that variable must be consistent. The tag *barrier_master_do* declares that the operation must be executed by the master while any other thread is forced to wait for him. Regarding the algorithmic behavior of the proposed GA, it is described next.

Function *GenerateInitialPop* (line 5) is expected to create as many candidate solutions, i.e., F vectors, as required by the population size. However, their evaluation according to Eq.4 is also required to form complete individuals. In fact, heliostats at this point are only forced to respect the constraints involving R_{min} and R_{max}, just to limit their position, but

Algorithm 1: Genetic algorithm for the first layer of the problem.

Input: Int $pop_{size}, num_{pairs}, tourn_{size}, cycles$, Real ov_{mut}, per_{mut}
Output: Vector F in $\mathbb{R}^{2H}$

1 IndividualSet *pop*, Individual *bestIndividual*; /* Shared among threads */
2 ThreadTeam $threads = createThreads()$; /* Create a team */
3 $threads.runInParallel()$; /* Thread-local below: */
4 Integers $range = getChunkSize()$; /* Get my zone of work as a thread */
5 $pop = GenerateInitialPop(pop_{size}, range, INJECT?)$;
6 $bestIndividual < synchr >= UpdateBest(pop)$;
7 **for** $i = 1$ ***to*** *cycles* **do**
8 IndividualSet $progs = SelectProgenitors(pop, range < num_{pairs} >, tourn_{size})$;
9 IndividualSet $desc = Reproduce(progs)$;
10 IndividualSet $descMut = Mutate(desc, ov_{mut}, per_{mut})$;
11 $bestIndividual < synchr >= UpdateBest(desc, descMut)$;
12 $pop < barrier\_master\_do >= Replace(pop, descMut, KEEP\_BEST)$;
13 **end**
14 **return** $bestIndividual.F$

collisions are not avoided. Thus, partially unfeasible solutions are possible from the very beginning. Nevertheless, its special label '$INJECT?$' declares the possibility of including some 'special individuals'. Specifically, it is referred to adding some fields obtained from any robust distribution pattern such as the biomimetic spiral proposed in [10]. This is a two-bladed option because it injects solid knowledge to the population from its origin, but also induces a serious influence in it because the fitness of those individuals will be much more higher than the other ones. Hence, it could lead to an important genetic drift and premature convergence. This option should be avoided or, at least, minimized when possible.

Function *SelectProgenitors* (line 8) simply looks for two different progenitors to form every pair. To do so, tournament selection is performed. This method is one of the most popular for GA as it can easily combine uniformity of exploration with adjustable selection pressure. Thus, every progenitor of each pair is selected out of a sample of $tourn_{size}$ participants. This is also the procedure applied in function *Replace* (line 12), in which the master selects every surviving individual out of $pop_{size} - 1$ tournaments. The previous -1 is caused because the best solution known so far is always left as part of the population (at it would not be guaranteed to participate at any tournament otherwise). This is called 'elitism' in the field of GA, and it is based on the idea that the structure of a very good solution could orientate the other ones to better zones of the search-space.

Function *Reproduce* (line 9) takes every pair and gets two descendant from each one.

 ISBN: 978-84-608-6082-2

To do so, uniform crossover is applied. This method is very popular because it features a high rate of mixing that tends to more complete explorations of the search-space. Its procedure consists in these steps: First, an auxiliary crossover mask is randomly defined. It is a binary string of length H (one per heliostat, i.e., pair of coordinates C) in which every bit had the same probability to be either a 0 or a 1. Second, a first descendant is formed by taking the heliostats (coordinates x and y) of its progenitor i for every position in which the auxiliary mask has a 1 while they taken from the progenitor $i+1$ otherwise. Third, the mask is inverted and a second descendant is obtained by applying the same rules. Any new individual must be ultimately evaluated according to Eq. 4.

Function $Mutate$ (line 10) is expected to allow the population to reach completely new zones of the search-space. To do so, every descendant has a probability of ov_{mut} of suffering any kind of mutation. Specifically, their set of heliostats is crossed and any of them has a probability of per_{mut} of being randomly repositioned. It must be noted that altered individuals must be evaluated as new ones. Besides, when applied, mutation can override promising solutions and make them worse. This is why a copy of the non-mutated ones is maintained to update the global reference of the best solution known so far.

Finally, the method returns to the best vector of coordinates defining a field that has been found during the search. That vector would be the solution that our meta-heuristic would receive for further processing.

4 Conclusions and future work

In this work, the problem of heliostat field optimization has been presented and formally stated. The field is not tried to be described by a reduced set of design variables, but the complete set of continuous coordinates is directly addressed. This approach maximizes the degrees of freedom at designing the field with respect to the use of parametric patterns. Increasing the mobility at search has been proven to lead to better designs. In fact, any final pattern can be seen just as a special case of a continuous pattern-free optimization. In this context, the authors of this paper are working on a meta-heuristic that can reduce the complexity of a complete resolution without significantly losing the achieved flexibility. However, as a meta-heuristic, it must be coupled with an optimizer to effectively solve the problem. Hence, a minimalist and parallel genetic algorithm has been designed for that purpose. Its aim is to perform a wide exploration of the search-space by using high-performance computing environments while also guiding the search. It relies on i) elitism, ii) uniform crossover, iii) static penalization and v) tournament selection. Additionally, the possibility of injecting distribution patterns as initial individuals is also allowed. It should be used with extreme care to avoid strong genetic drift and premature convergence.

Regarding future work, the most immediate one is to enhance the current preliminary implementation of the method. After that, it will included in our meta-heuristic and its real

 ISBN: 978-84-608-6082-2

performance for the problem at hand will be analyzed in depth. Hence, that some aspects might have to be further adapted.

Acknowledgements

This work has been funded by grants from the Spanish Ministry of Economy and Competitiveness (TIN2015-66680-C2-1-R and ENERPRO DPI 2014-56364-C2-1-R), Junta de Andalucía (P11-TIC7176 and P12-TIC301). Nicolás Calvo Cruz (FPU14/01728) is supported by an FPU Fellowship from the Spanish Ministry of Education. Juana López Redondo (RYC-2013-14174) and José Domingo Álvarez (RYC-2013-14107) are fellows of the Spanish 'Ramón y Cajal' contract program, co-financed by the European Social Fund.

References

[1] S. Alexopoulos and B. Hoffschmidt, *Advances in solar tower technology*, WIREs Energy Environ. **6** (2017) 1–19.

[2] A. L. Avila-Marín, J. Fernández-Reche and F. M. Tellez, *Evaluation of the potential of central receiver solar power plants: Configuration, optimization and trends*, Appl. Energ. **112** (2013) 274–288.

[3] S. M. Besarati and D. Y. Goswami, *A computationally efficient method for the design of the heliostat field for solar power tower plant*, Renew. Energ. **69** (2014) 226–232.

[4] R. Buck, A. Pfahl, and T. H. Roos, *Target aligned heliostat field layout for non-linear flat terrain*, In Proceedings of the First Southern African Solar Energy Conference (SASEC) (2012).

[5] E. Carrizosa, C. Domínguez-Bravo, E. Fernánded.Cara and M. Quero, *An optimization approach to the design of multi-size heliostat fields.*, Tech. rep., Institute of Mathematics of University of Seville (IMUS). (2014).

[6] F. J. Collado and J. Guallar, *Campo: Generation of regular heliostat fields*, Renew. Energ. **46** (2012) 49–59.

[7] J. H. Holland, *Adaptation in natural and artificial systems: An introductory analysis with applications to biology, control, and artificial intelligence*, Michigan Press (1975).

[8] S. L. Lutchman, A. A. Groenwold, P. Gauché and S. Bode, *On Using a Gradient-Based Method for Heliostat Field Layout Optimization*, Energ. Proc. **49** (2014) 1429–1438.

 ISBN: 978-84-608-6082-2

[9] S. L. Lutchman, *Heliostat field layout optimization for a central receiver*, Master Thesis at the University of Stellenbosch, 2014.

[10] C. J. Noone, M. Torrilhon and A. Mitsos, *Heliostat field optimization: A new computationally efficient model and biomimetic layout*, Sol. Energy. **86(2)** (2012) 792–803.

[11] R. Pitz-Paal, N. B. Botero and A. Steinfeld, *Heliostat field layout optimization for high-temperature solar thermochemical processing*, Sol. Energy. **85(2)** (2011) 334–343.

[12] A. Ramos and F. Ramos, *Strategies in tower solar power plant optimization*, Sol. Energy. **86(9)** (2012) 2536–2548.

[13] A. Ramos and F. Ramos, *Heliostat blocking and shadowing efficiency in the video-game era*, arXiv **86(1402.1690)** (2014).

[14] S. Salhi, *Heuristic Search: The emerging science of problem solving*, Springer, Switzerland, 2017.

[15] M. Sanchez and M. Romero, *Methodology for generation of heliostat field layout in central receiver systems based on yearly normalized energy surfaces*, Sol. Energy. **80(7)** (2006) 861–874.

[16] W. B. Stine and M. Geyer, *Power from the Sun*, On-line book available on http://powerfromthesun.net/, 2001. (Last accessed: 13^{th} May, 2017)

[17] Ö. Yeniay, *Penalty function methods for constrained optimization with genetic algorithms*, Math. and Comp. App. **10(1)** (2005) 45–56.

 ISBN: 978-84-608-6082-2

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

Numerical solution of surface integral equations based on spline quasi-interpolation

Catterina Dagnino[1] and Sara Remogna[1]

[1] *Department of Mathematics, University of Torino, via C. Alberto, 10 - 10123 Torino, Italy*

emails: catterina.dagnino@unito.it, sara.remogna@unito.it

Abstract

In this paper we propose a modified version of the classical collocation method and two spline collocation methods with high order of convergence, for the solution of integral equations on surfaces of $\mathbb{R}^3$. Such methods are based on optimal superconvergent quasi-interpolants defined on type-2 triangulations and based on the Zwart-Powell quadratic box spline.

Key words: surface integral equation, spline quasi-interpolation
MSC 2000: 65R20, 65D07

1 Introduction

In this paper we consider the surface integral equation

$$\rho(\mathbf{P}_1) - \int_S K(\mathbf{P}_1, \mathbf{P}_2)\rho(\mathbf{P}_2)dS_{\mathbf{P}_2} = \psi(\mathbf{P}_1), \quad \mathbf{P}_1 \in S, \tag{1}$$

where S is a connected surface in $\mathbb{R}^3$, described by a sufficiently smooth map $\mathbf{F} : \Omega \to S$, with Ω a polygonal domain in $\mathbb{R}^2$, and the kernel $K(\mathbf{P}_1, \mathbf{P}_2)$ is continuous for $\mathbf{P}_1$, $\mathbf{P}_2 \in S$.

Therefore, (1) can be written as

$$\rho(\mathbf{F}(u,v)) - \int_\Omega K(\mathbf{F}(u,v), \mathbf{F}(s,t))\rho(\mathbf{F}(s,t)) \, |(D_s\mathbf{F} \times D_t\mathbf{F})(s,t)| \, ds \, dt = \psi(\mathbf{F}(u,v)), \ (u,v) \in \Omega,$$

where $|(D_s\mathbf{F} \times D_t\mathbf{F})(s,t)|$ is the Jacobian of the map $\mathbf{F}(s,t)$.

If we denote by $\mathcal{K} : C(S) \longrightarrow C(S)$ the integral operator defined by

$$\mathcal{K}\rho(\mathbf{F}(u,v)) := \int_{\Omega} K(\mathbf{F}(u,v), \mathbf{F}(s,t))\rho(\mathbf{F}(s,t)) \, |(D_s\mathbf{F} \times D_t\mathbf{F})(s,t)| \; ds \; dt,$$

for $(u,v) \in \Omega$, then we can write (1) in the following operator form

$$(\mathcal{I} - \mathcal{K})\rho = \psi. \tag{2}$$

We remark that (2) has a unique solution $\rho \in C(S)$ for any given $\psi \in C(S)$ [3].

In the literature, standard methods for solving (2) consist in Nyström, Galerkin and collocation methods. For instance, we recall the collocation ones based on a sequence of linear interpolatory projection operators onto finite dimensional subspaces $\mathcal{X}_{mn}$ of $C(S)$, converging to the identity operator pointwise. A classical choice of $\mathcal{X}_{mn}$ is the space of C^0 piecewise polynomials of a given degree d (usually $d = 2$) on a triangulation of Ω (see [3, 5]).

In this paper we propose three collocation methods for (2), based on a sequence of optimal superconvergent spline quasi-interpolating operators $\{Q_{mn}\}$, that are not projectors and are defined on the space $\mathcal{X}_{mn} = S_2^1(\Omega, T_{mn})$ of the C^1 quadratic splines on a uniform type-2 triangulation T_{mn} of Ω, with Ω a rectangular domain. We recall [12] that the above quasi-interpolating splines are expressed by means of the scaled/translates of the Zwart-Powell quadratic box spline (ZP-element) (see e.g. [4, Chap. 1], [14, Chap. 2]). From a computational point of view, this is more convenient than the use of other spanning sets, for instance formed by bivariate B-splines with support completely included in Ω [1, 7, 9, 13], that, having different supports, have different expressions in the domain, while the ZP-element is always the same.

Given a rectangular domain $\Omega = [a,b] \times [c,d]$, by dividing it into mn equal squares $\{\Omega_{ij}\}_{i=1,\ j=1}^{m\ \ n}$ with a given edge h, $m, n \geq 4$, each of them being subdivided into 4 triangles by its diagonals, we obtain a uniform type-2 triangulation T_{mn} of Ω. We denote by $S_2^1(\Omega, T_{mn})$ the space of C^1 quadratic splines on T_{mn}, whose dimension is $(m+2)(n+2) - 1$ ([14] and the reference therein).

This space is generated by the $(m+2)(n+2)$ B-spline functions $\{B_{i,j}, (i,j) \in A_{mn}\}$, where $A_{mn} = \{(i,j),\ 0 \leq i \leq m+1,\ 0 \leq j \leq n+1\}$, obtained by dilation/translation of the ZP-element. Moreover, in order to obtain a B-spline basis for $S_2^1(\Omega, T_{mn})$ we have to neglect one B-spline from the spanning set ([14] and the reference therein).

In the space $S_2^1(\Omega, T_{mn})$ we consider special optimal quasi-interpolants (abbr. QIs) of the form

$$Q_{mn}f := \sum_{(i,j)\in A_{mn}} \lambda_{i,j}(f) B_{i,j}, \tag{3}$$

with $\{\lambda_{i,j},\ (i,j) \in A_{mn}\}$ a family of local linear functionals defined in this way

$$\lambda_{i,j}(f) := \sum_{(k,l)\in F_{i,j}} \sigma_{i,j}(k,l) f(M_{k,l}), \tag{4}$$

 ISBN: 978-84-617-8694-7

where the finite set of points $\{M_{k,l},\ (k,l) \in F_{i,j}\}$, $F_{i,j} \subset A_{mn}$, lies in some neighbourhood of $\text{supp}B_{i,j} \cap \Omega$ and the $\sigma_{i,j}(k,l)$'s are chosen such that $Q_{mn}f \equiv f$ for all f in $\mathbb{P}_2$ (the space of bivariate polynomials of total degree two) and superconvergence is induced at some specific points, i.e. the vertices, the centers, the midpoints of horizontal and vertical edges of each subsquare of the partition. The coefficient functional expression (4) is given in [12] and we recall that $\|Q_{mn}\|_\infty \leq 2$. The points $M_{k,l}$ in (4) are the mn centers of the squares, the $2(m+n)$ midpoints of boundary segments and the four vertices of Ω.

We remark that the QIs (3) can also be written in quasi-Lagrange form

$$Q_{mn}f := \sum_{(i,j)\in A_{mn}} f(M_{i,j})L_{i,j},$$

by means of the fundamental functions $L_{i,j}$, obtained as linear combination of the $B_{i,j}$'s.

Standard results in approximation theory and other specific ones given in [6] allow us to deduce the following theorem, where $D^\beta = D^{\beta_1\beta_2} = \frac{\partial^{|\beta|}}{\partial x^{\beta_1}\partial y^{\beta_2}}$, with $|\beta| = \beta_1 + \beta_2$, $\|D^\nu f\|_\infty = \max\limits_{|\beta|=\nu} \left\|D^\beta f\right\|_\infty$, $\omega\left(D^\nu f, h\right) = \max\left\{\omega\left(D^\alpha f, h\right), |\alpha| = \nu\right\}$, where $\omega(f,h) = \max\left\{|f(P_1) - f(P_2)|\,; P_1, P_2 \in \Omega, \|P_1 - P_2\| \leq h\right\}$ is the modulus of continuity of $f \in C(\Omega)$, and $\|\cdot\|$ is the Euclidean norm.

Theorem 1 *Let $f \in C^\nu(\Omega)$, $0 \leq |\alpha| \leq \nu \leq 2$, $|\alpha| = 0, 1$ then*

$$\|D^\alpha(f - Q_{mn}f)\|_\infty \leq K_{\alpha,\nu}h^{\nu-|\alpha|}\omega\left(D^\nu f, h\right),$$

where the error constant $K_{\alpha,\nu}$ is independent of h and depends only on α and ν.

If, in addition, $f \in C^3(\Omega)$, then

$$\|D^\alpha(f - Q_{mn}f)\|_\infty \leq K_{\alpha,3}h^{3-|\alpha|}\left\|D^3 f\right\|_\infty.$$

We underline that Q_{mn} has superconvergence properties. In particular, for $f \in C^4(\Omega)$, we have that $|(f - Q_{mn}f)(P)| = O(h^4)$ at specific points P in Ω, that are the vertices, the centers, the midpoints of horizontal and vertical edges of each subsquare of Ω partition.

Finally, the above superconvergent QIs can be applied to numerical integration, getting cubature rules that we will use in Section 2.2.

For any function $f \in C(\Omega)$, we can numerically evaluate the integral

$$I(f) = \int_\Omega f(s,t)\, ds\, dt$$

by the cubature rule defined by

$$I(Q_{mn}f) = \sum_{(i,j)\in A_{mn}} w_{i,j}f(M_{i,j}), \tag{5}$$

 ISBN: 978-84-617-8694-7

where the weights

$$w_{i,j} = \int_\Omega L_{i,j}(s,t)\ ds\ dt.$$

are reported in [8].

From Theorem 1, we can easily deduce the following result.

Theorem 2 *Let $f \in C(\Omega)$ and $E(f) = I(f) - I(Q_{mn}f)$.*
Then, $|E(f)| \leq \bar{C}\omega(f,h)$, where $\bar{C}$ is a positive constant independent of m and n.
Moreover if $f \in C^\nu(\Omega)$, $\nu = 1,2,3$, then $E(f) = O(h^\nu)$.

We remark that the above cubature has precision degree at least 2, because Q_{mn} is exact on $\mathbb{P}_2$. However, since uniform partitions are special cases of the ones with symmetric knots with respect to the center of Ω, Corollary 1 of [10] can be generalized to our case, getting

$$I(f) = I(Q_{mn}f) \text{ for } f(s,t) = s^{r_1}t^{r_2},$$

with $0 \leq r_1, r_2 \leq 3$, $r_1 + r_2 = 3$ and $r_1, r_2 = 1,3$, with $r_1 + r_2 = 4$. Therefore the precision degree of the cubature (5) is 3 and, if $f \in C^4(\Omega)$, then $E(f) = O(h^4)$.

2 Collocation methods for surface integral equations

In this section we present and analyse three collocation methods (see [8] for details) based on the sequence $\{Q_{mn}\}$ of spline QI operators defined in Section 1.

2.1 Modified collocation method

In this method, that we call *modified collocation method*, in (2) we replace the operator $\mathcal{K}$ by $Q_{mn}\mathcal{K}$ and the right hand side ψ by $Q_{mn}\psi$. We remark that the idea of defining a collocation method by operators that are not projectors has been proposed in [2] for univariate integral equations.

Therefore, we approximate the integral equation (2) by

$$(\mathcal{I} - Q_{mn}\mathcal{K})\rho_{mn} = Q_{mn}\psi. \tag{6}$$

We write the approximated solution ρ_{mn}, belonging to $S_2^1(\Omega, T_{mn})$, as

$$\rho_{mn}(\mathbf{F}(u,v)) = \sum_{\alpha \in A_{mn}} X_\alpha L_\alpha(u,v), \quad \text{with} \quad \alpha = (i,j).$$

Substituting the expressions of Q_{mn} and ρ_{mn} into (6), by identifying the coefficients of L_α, we obtain

$$X_\alpha - \sum_{\beta \in A_{mn}} X_\beta \bar{L}_\beta(M_\alpha) = \psi(\mathbf{F}(M_\alpha)), \quad \alpha \in A_{mn},$$

 ISBN: 978-84-617-8694-7

with $\bar{L}_\beta = \mathcal{K} L_\beta$. This is a linear system of $(m+2)(n+2)$ equations, that can be written in the form

$$(I - A)\mathbf{X} = \mathbf{a} \tag{7}$$

where A is the matrix with entries

$$A_{\alpha\beta} := \bar{L}_\beta(M_\alpha) = \int_\Omega K(\mathbf{F}(M_\alpha), \mathbf{F}(s,t)) \, |(D_s\mathbf{F} \times D_t\mathbf{F})(s,t)| \, L_\beta(s,t) \; ds \; dt \tag{8}$$

and $\mathbf{a}$ is the vector with elements $\mathbf{a}_\alpha := \psi(\mathbf{F}(M_\alpha))$.

Concerning the convergence, we can state the following theorem.

Theorem 3 *Let $\rho \in C^3(\Omega)$, then $\|\rho - \rho_{mn}\|_\infty = O\left(h^3\right)$.*

2.2 Collocation methods with high order of convergence

In these methods, that we call *collocation methods with high order of convergence*, in (2) we replace $\mathcal{K}$ by one of the two following finite rank operators

$$\mathcal{K}_{mn,i} := Q_{mn}\mathcal{K} + \mathcal{K}^*_{mn,i} - Q_{mn}\mathcal{K}^*_{mn,i}, \qquad i = 1, 2,$$

where

1. $\mathcal{K}^*_{mn,1}$ is the degenerate kernel operator defined by

$$\begin{aligned} &\mathcal{K}^*_{mn,1}\rho(\mathbf{F}(u,v)) \\ &:= \int_\Omega Q_{mn}\left(K(\mathbf{F}(u,v), \mathbf{F}(s,t)) \, |(D_s\mathbf{F} \times D_t\mathbf{F})(s,t)|\right) \rho(\mathbf{F}(s,t)) \; ds \; dt \\ &= \sum_{\alpha \in A_{mn}} K(\mathbf{F}(u,v), \mathbf{F}(M_\alpha)) \, |(D_s\mathbf{F} \times D_t\mathbf{F})(M_\alpha)| \cdot \int_\Omega L_\alpha(s,t)\rho(\mathbf{F}(s,t)) \; ds \; dt, \end{aligned} \tag{9}$$

2. $\mathcal{K}^*_{mn,2}$ is the Nyström operator based on Q_{mn} and defined by

$$\mathcal{K}^*_{mn,2}\rho(\mathbf{F}(u,v)) := \sum_{\alpha \in A_{mn}} w_\alpha K(\mathbf{F}(u,v), \mathbf{F}(M_\alpha)) \, |(D_s\mathbf{F} \times D_t\mathbf{F})(M_\alpha)| \, \rho(\mathbf{F}(M_\alpha)), \tag{10}$$

according to (5).

We remark that such methods are defined by a logical scheme similar to that one used in [1] to construct methods for 2D integral equations, based on other quasi-interpolants.

Therefore, we approximate (2) by

$$\rho_{mn,i} - (Q_{mn}\mathcal{K} + \mathcal{K}^*_{mn,i} - Q_{mn}\mathcal{K}^*_{mn,i})\rho_{mn,i} = \psi, \qquad i = 1, 2. \tag{11}$$

 ISBN: 978-84-617-8694-7

that can be reduced to two systems of $2(m+2)(n+2)$ linear equations.

After some algebra, from (9) and (11), we can write the approximate solution $\rho_{mn,1}$ as:

$$\begin{aligned}\rho_{mn,1}(\mathbf{F}(u,v)) = \; & \psi(\mathbf{F}(u,v)) + \sum_{\alpha\in A_{mn}} X_\alpha L_\alpha(u,v) \\ & + \sum_{\alpha\in A_{mn}} Y_\alpha K(\mathbf{F}(u,v),\mathbf{F}(M_\alpha))\,|(D_s\mathbf{F}\times D_t\mathbf{F})(M_\alpha)|\,,\end{aligned}$$

where the unknowns $\{X_\alpha\}$ and $\{Y_\alpha\}$, $\alpha\in A_{mn}$, are obtained by solving the linear system $(I-R)\mathbf{Z}=\mathbf{d}$, with

$$R := \begin{bmatrix} A & D-B \\ C & E \end{bmatrix}, \quad \mathbf{Z} := \begin{bmatrix} \mathbf{X} \\ \mathbf{Y} \end{bmatrix}, \quad \mathbf{d} := \begin{bmatrix} \mathbf{b} \\ \mathbf{c} \end{bmatrix} \tag{12}$$

and A, B, C, D, $E\in\mathbb{R}^{(m+2)(n+2)\times(m+2)(n+2)}$, $\mathbf{b}$, $\mathbf{c}\in\mathbb{R}^{(m+2)(n+2)}$, whose entries are given by

- $A_{\alpha,\beta} := \bar{L}_\beta(M_\alpha)$, see (8),
- $B_{\alpha,\beta} := K(\mathbf{F}(M_\alpha),\mathbf{F}(M_\beta))\,|(D_s\mathbf{F}\times D_t\mathbf{F})(M_\beta)|$,
- $C_{\alpha,\beta} := \displaystyle\int_\Omega L_\alpha(s,t)L_\beta(s,t)\;ds\;dt$,
- $D_{\alpha,\beta} := \displaystyle\int_\Omega K(\mathbf{F}(M_\alpha),\mathbf{F}(s,t))|(D_s\mathbf{F}\times D_t\mathbf{F})(s,t)|K(\mathbf{F}(s,t),\mathbf{F}(M_\beta))|(D_s\mathbf{F}\times D_t\mathbf{F})(M_\beta)|\;ds\;dt$,
- $E_{\alpha,\beta} := \displaystyle\int_\Omega K(\mathbf{F}(s,t),\mathbf{F}(M_\beta))\,|(D_s\mathbf{F}\times D_t\mathbf{F})(M_\beta)|\,L_\alpha(s,t)\;ds\;dt$,
- $\mathbf{b}_\alpha := \mathcal{K}\psi(\mathbf{F}(M_\alpha)) = \displaystyle\int_\Omega K(\mathbf{F}(M_\alpha),\mathbf{F}(s,t))\,|(D_s\mathbf{F}\times D_t\mathbf{F})(s,t)|\,\psi(\mathbf{F}(s,t))\;ds\;dt$,
- $\mathbf{c}_\alpha := \displaystyle\int_\Omega \psi(\mathbf{F}(s,t))L_\alpha(s,t)\;ds\;dt$.

Similarly, from (10) and (11), we can get that the solution $\rho_{mn,2}$ is

$$\begin{aligned}\rho_{mn,2}(\mathbf{F}(u,v)) = \; & \psi(\mathbf{F}(u,v)) + \sum_{\alpha\in A_{mn}} X_\alpha L_\alpha(u,v) \\ & + \sum_{\alpha\in A_{mn}} w_\alpha Y_\alpha K(\mathbf{F}(u,v),\mathbf{F}(M_\alpha))\,|(D_s\mathbf{F}\times D_t\mathbf{F})(M_\alpha)|\,,\end{aligned}$$

where the unknowns $\{X_\alpha\}$ and $\{Y_\alpha\}$, $\alpha\in A_{mn}$, are obtained by solving the linear system $(I-T)\mathbf{Z}=\mathbf{f}$, with

$$T := \begin{bmatrix} A & F-G \\ H & G \end{bmatrix}, \quad \mathbf{Z} := \begin{bmatrix} \mathbf{X} \\ \mathbf{Y} \end{bmatrix}, \quad \mathbf{f} := \begin{bmatrix} \mathbf{a} \\ \mathbf{e} \end{bmatrix} \tag{13}$$

and F, G, $H\in\mathbb{R}^{(m+2)(n+2)\times(m+2)(n+2)}$, $\mathbf{e}\in\mathbb{R}^{(m+2)(n+2)}$, whose entries are given by

- $F_{\alpha,\beta} := w_\beta \displaystyle\int_\Omega K(\mathbf{F}(M_\alpha), \mathbf{F}(s,t))|(D_s\mathbf{F}\times D_t\mathbf{F})(s,t)|K(\mathbf{F}(s,t), \mathbf{F}(M_\beta))|(D_s\mathbf{F}\times D_t\mathbf{F})(M_\beta)|\, ds\, dt,$
- $G_{\alpha,\beta} := w_\beta K(\mathbf{F}(M_\alpha), \mathbf{F}(M_\beta))\, |(D_s\mathbf{F} \times D_t\mathbf{F})(M_\beta)|,$
- $H_{\alpha,\beta} := L_\beta(M_\alpha),$
- $\mathbf{e}_\alpha := \psi(\mathbf{F}(M_\alpha).$

Concerning the convergence, we can state the following theorem.

Theorem 4 *If ρ is differentiable with bounded derivatives, $K(\cdot,\cdot) \in C^4(S \times S)$ and $\mathbf{F} \in C^5(\Omega)$, then $\|\rho - \rho_{mn,1}\|_\infty = O(h^7)$.*

If $\rho \in C^4(S)$, $K(\cdot,\cdot) \in C^4(S \times S)$ and $\mathbf{F} \in C^5(\Omega)$, then $\|\rho - \rho_{mn,2}\|_\infty = O(h^7)$.

3 Numerical results

By using the collocation methods (6) and (11), we have to evaluate many integrals and usually it must be done by suitable numerical integration formulas. Therefore, we have to discretize the proposed methods by introducing convenient cubatures and we denote by ρ^D_{mn}, $\rho^D_{mn,i}$, $i = 1, 2$, the corresponding solutions.

Here, we decide to compute the entries of the matrices and vectors appearing in (7), (12), (13), by using a composite Gaussian cubature on triangular domains (see [11]), implemented by the Matlab function `triquad` (see [15]), with N^2 nodes in each triangle of T_{mn} and with precision degree $2N - 1$. The number of nodes is chosen to preserve the approximation order of the method. Therefore, we choose $N = 2$ for the modified collocation method (6) and $N = 4$ for the two collocation methods with high order of convergence (11).

We test the performances of the proposed methods in the numerical solution of the surface integral equation from [3]

$$\rho(\mathbf{P}_1) - \frac{1}{30}\int_S \rho(\mathbf{P}_2)\frac{\partial}{\partial \mathbf{n}_{\mathbf{P}_2}}\left(\|\mathbf{P}_1 - \mathbf{P}_2\|^2\right) dS_{\mathbf{P}_2} = \frac{1}{30}\psi(\mathbf{P}_1),\ \mathbf{P}_1 \in S,$$

where S is the ellipsoidal surface given by $x^2 + \left(\frac{4y}{3}\right)^2 + (2z)^2 = 1$, $\mathbf{n}_{\mathbf{P}_2}$ is the inner normal to S at $\mathbf{P}_2$ and

$$\mathbf{F}(s,t) = \begin{bmatrix} \sin(s)\cos(t) \\ \frac{3}{4}\sin(s)\sin(t) \\ \frac{1}{2}\cos(s) \end{bmatrix}, \qquad (s,t) \in \Omega = [0,\pi] \times [0, 2\pi].$$

We choose $\rho(\mathbf{P}) = e^{\frac{1}{2}\cos(s)}$ and define ψ accordingly.

 ISBN: 978-84-617-8694-7

For each method we compute the maximum absolute errors

$$E_{mn} = \max_{(u,v)\in G} |\rho(u,v) - \rho^D_{mn}(u,v)|, \quad E_{mn,i} = \max_{(u,v)\in G} |\rho(u,v) - \rho^D_{mn,i}(u,v)|, \quad i = 1,2,$$

for increasing values of m and n, where G is a uniform grid of 100×100 points in Ω. We also compute the corresponding numerical convergence orders o_{mn}, $o_{mn,i}$, $i = 1,2$.

The results are shown in Table 1 and we can notice that they agree with the theoretical ones.

Table 1: Maximum absolute errors and numerical convergence orders.

m	n	E_{mn}	o_{mn}	$E_{mn,1}$	$o_{mn,1}$	$E_{mn,2}$	$o_{mn,2}$
4	8	7.56e-03	-	2.51e-05	-	3.17e-05	-
8	16	8.11e-04	3.22	2.09e-07	6.91	1.29e-07	7.94
16	32	8.21e-05	3.30	1.48e-09	7.14	1.71e-09	6.24
32	64	8.34e-06	3.30	1.12e-11	7.04	1.46e-11	6.87

Acknowledgements

This work has been supported by the program "Progetti di Ricerca 2016" of the Gruppo Nazionale per il Calcolo Scientifico (GNCS) - INdAM. Moreover, the authors thank the University of Torino for its support to their research.

References

[1] C. Allouch, P. Sablonnière, D. Sbibih *A collocation method for the numerical solution of a two dimensional integral equation using a quadratic spline quasi-interpolant*, Numer. Algorithms **62** (2013) 445–468.

[2] C. Allouch, P. Sablonnière, D. Sbibih *A modified Kulkarni's method based on a discrete spline quasi-interpolant*, Math. Comput. Simul. **81** (2011) 1991–2000.

[3] K.E. Atkinson, *The numerical solution of integral equations of the second kind*, Cambridge University Press, 1997.

[4] C. de Boor, K. Höllig, S. Riemenschneider, *Box Splines*, Springer-Verlag, New York, 1993.

 ISBN: 978-84-617-8694-7

[5] D. Chien, *Piecewise polynomial collocation for integral equations with a smooth kernel on surfaces in three dimensions*, J. Integral Equations Appl. **5** (1993) 315–344.

[6] C. Dagnino, P. Lamberti, *On the approximation power of bivariate quadratic C^1 splines*, J. Comput. Appl. Math. **131** (2001) 321–332.

[7] C. Dagnino, P. Lamberti, *On the construction of local quadratic spline quasi-interpolants on bounded rectangular domains*, J. Comput. Appl. Math. **221** (2008) 367–375.

[8] C. Dagnino, S. Remogna, *Quasi-interpolation based on the ZP-element for the numerical solution of integral equations on surfaces in* $\mathbb{R}^3$, BIT Numer. Math. DOI 10.1007/s10543-016-0633-x (2016).

[9] C. Dagnino, S. Remogna, P. Sablonnière, *Error bounds on the approximation of functions and partial derivatives by quadratic spline quasi-interpolants on non-uniform criss-cross triangulations of a rectangular domain*, BIT Numer. Math. **53** (2013) 87–109.

[10] P. Lamberti, *Numerical integration based on bivariate quadratic spline quasi-interpolants on bounded domains*, BIT Numer. Math. **49** (2009) 565–588.

[11] J. N. Lyness, R. Cools, *A Survey of Numerical Cubature over Triangles*, Mathematics and Computer Science Division, Argonne National Laboratory **III** 1994.

[12] S. Remogna, *Constructing good coefficient functionals for bivariate C^1 quadratic spline quasi-interpolants*, In: Daehlen, M., et al. (Eds.), Mathematical Methods for Curves and Surfaces. Lecture Notes in Computational Science, vol. **5862**. Springer-Verlag, Berlin, Heidelberg, (2010) 329–346.

[13] P. Sablonnière, *Quadratic spline quasi-interpolants on bounded domains of* $\mathbb{R}^d, d = 1, 2, 3$, Rend. Sem. Mat. Univ. Pol. Torino **61** (2003) 229–238.

[14] R. H. Wang, *Multivariate Spline Functions and Their Application*, Science Press, Beijing/New York, Kluwer Academic Publishers, Dordrecht/Boston/London, 2001.

[15] G. von Winckel, *Matlab procedure* `triquad`, http://www.mathworks.com/matlabcentral/fileexchange/9230-gaussian-quadrature-for-triangles.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

Parameter Uniform Numerical Approximation of the Solution of A System of Reaction Diffusion Problems involving A Small Perturbation Parameter

P. Das[1] and J. Vigo-Aguiar[2]

[1] *Department of Mathematics, Indian Institute of Technology, Patna, India*

[2] *Department of Applied Mathematics, University of Salamanca, Salamanca, Spain*

emails: pratibhamoy@iitp.ac.in, jvigo@usal.es

Abstract

We consider an optimal order numerical approximation for the boundary layer adaptive solution of a system of reaction-diffusion problems. To generate the layer adaptive mesh, we use the equidistribution of a positive monitor function. This technique enables the mesh movement to follow the rapidly changing behavior of the solution. It is observed that the discrete solution is second-order uniformly convergent on this mesh and the convergence is optimal with respect to the discretization of the continuous problem. Numerical experiments validate the performance of the present method.

Key words: Boundary layer, Singular perturbation, Adaptive mesh, Mesh equidistribution, Uniform convergence.
MSC 2000: 65L10

1 Introduction

In this work, we consider a system of reaction diffusion model which frequently rises in control theory [8]. For the numerical analysis, we consider the following singularly perturbed system of reaction-diffusion problems on $x \in \Omega = (0,1)$:

$$\boldsymbol{Lu}(x) = \boldsymbol{f}(x) \Leftrightarrow \begin{cases} L_1\boldsymbol{u}(x) \equiv -\varepsilon u_1''(x) + b_{11}(x)u_1(x) + b_{12}(x)u_2(x) = f_1(x), \\ L_2\boldsymbol{u}(x) \equiv -\varepsilon u_2''(x) + b_{21}(x)u_1(x) + b_{22}(x)u_2(x) = f_2(x), \\ u_1(0) = 0, \quad u_1(1) = 0, \quad u_2(0) = 0, \quad u_2(1) = 0, \end{cases} \tag{1}$$

where $0 < \varepsilon \ll 1$ is the singular perturbation parameter and $\boldsymbol{u}(x) = (u_1(x), u_2(x))^T$, $\boldsymbol{L} = (L_1,\ L_2)^T$ and $\boldsymbol{f}(x) = (f_1(x), f_2(x))^T$ which are assumed to be sufficiently smooth. The matrix $(b_{ij})_{i,j=1}^2$ is considered an L_0-matrix (*i.e.,* off-diagonals are nonpositive and diagonals are positive) with

$$\min_{x\in\overline{\Omega}=[0,1]} \left\{ \sum_{j=1}^{2} b_{mj}(x) \right\} \geq \beta > 0, \quad \text{for} \quad m = 1, 2, \tag{2}$$

and $\overline{\beta} > b_{ii}(x) > \beta > 0,\ i, j = 1, 2$, for some real number β, $\overline{\beta}$. Under these assumptions the equation (1) has a unique solution $\boldsymbol{u}(x)$ [3] with two boundary layers at $x = 0, 1$.

Numerical methods on a fixed uniform mesh do not capture boundary layer for an arbitrary value of ε, unless one uses the fitted operator technique. This technique is well developed in several works of Vigo-Aguiar et al. [9, 10]. In recent days, fitted mesh methods have shown its interest, where one needs to find an adaptive mesh which is dense inside the boundary layers. In [6], an almost second order convergence upto logarithmic term is observed on Shishkin's fitted mesh for (1). Here, our aim is to develop a numerical method for (1) which is exactly second order convergent. For this, we use the mesh equidistribution technique. A mesh $\{a = x_0 < x_1 < \cdots < x_N = b\}$ is said to be equidistributed [2, 4, 7], if

$$\int_{x_{i-1}}^{x_i} M\big(s, \boldsymbol{u}(s)\big) ds = \frac{1}{N} \int_a^b M\big(s, \boldsymbol{u}(s)\big) ds, \quad i = 1, \cdots, N, \tag{3}$$

where $M\big(x, \boldsymbol{u}(x)\big) (> 0)$ is called the error monitor function. This technique automatically adopts the mesh by equidistributing $M\big(s, \boldsymbol{u}(s)\big)$. Nowadays, several researchers have shown their interest on the moving mesh methods for scalar singularly perturbed reaction-diffusion [1, 3]. Here, our main goal is to obtain a parameter uniform optimal order convergent solution for a system of singularly perturbed reaction-diffusion problems using mesh equidistribution.

The paper is divided as follows. In Section 2, we propose an error monitor function whose equidistribution leads to an layer adaptive mesh. The finite difference discretization of (1) and its stability analysis is considered in Section 3. In Section 4, it is shown that the discrete solution is second-order parameter uniformly convergent on the equidistributed mesh. This order of convergence is optimal with respect to the discretization of the continuous problem. The numerical examples in Section 5 validates the theoretical prediction.

Throughout this paper, C denotes a generic positive constant, independent of ε, x_i (mesh points) and N (number of partitions of the domain $\overline{\Omega}$). We set $\phi_i = \phi(x_i)$ for any scalar function ϕ, while ϕ_i^N (or Φ_i^N for vector valued function) denote a numerical approximation of ϕ (analogously $\Phi = (\phi_1, \phi_2)^T \in \mathbb{R}^2$) at x_i. We denote $||\cdot||_\infty$ as $||\phi|| = ||\phi(x)||_D = \max_{\xi\in D} |\phi(\xi)|$ for any function ϕ defined on some domain D. Analogously for $\boldsymbol{\phi} \in \mathbb{R}^2$, $||\boldsymbol{\phi}||$ will be denoted as $||\boldsymbol{\phi}||_\infty = ||\boldsymbol{\phi}|| = \max_{x\in D} \{|\phi_1|, |\phi_2|\}$. In the analysis, we assume that ε satisfies $0 < \sqrt{\varepsilon} < N^{-1}$, as otherwise, the adaptive mesh is not required.

 ISBN: 978-84-617-8694-7

2 Derivative Bounds of the Solution

We decompose the analytical solution $\boldsymbol{u}$ into the smooth and singular components $\boldsymbol{v}$ and $\boldsymbol{w}$ resp. such that $\boldsymbol{u} = \boldsymbol{v} + \boldsymbol{w}$. The following lemma provides the derivative bounds of $\boldsymbol{v}$, $\boldsymbol{w}$.

Lemma 2.1 *The smooth component* $\boldsymbol{v} = (v_1, v_2)^T$, *where*

$$\boldsymbol{L}\boldsymbol{v}(x) = \boldsymbol{f}(x), \quad x \in \Omega, \quad \textit{where } \boldsymbol{v}(0) \textit{ and } \boldsymbol{v}(1) \textit{ are suitably chosen},$$

and the singular component $\boldsymbol{w} = (w_1, w_2)^T$, *where*

$$\boldsymbol{L}\boldsymbol{w}(x) = 0, \quad x \in \Omega, \quad \boldsymbol{w}(0) = \boldsymbol{u}(0) - \boldsymbol{v}(0), \quad \boldsymbol{w}(1) = \boldsymbol{u}(1) - \boldsymbol{v}(1),$$

satisfy

$$|\boldsymbol{v}^{(k)}(x)| \leq C(1 + \varepsilon^{(2-k)/2}), \textit{ for } k = 0, \cdots, 4, \tag{4}$$

and

$$|\boldsymbol{w}^{(k)}(x)| \leq C\varepsilon^{-k/2}\big(\exp(-x\sqrt{\beta/\varepsilon}) + \exp(-(1-x)\sqrt{\beta/\varepsilon})\big), \textit{ for } k = 0, \cdots, 4. \tag{5}$$

Proof. The proof of this lemma is given in [5]. For our error analysis, we shall use the leading order asymptotic expansion of the solution $\boldsymbol{u}(x)$ of (1) and its derivative bound. To find this, we use the decomposition of the solution $\boldsymbol{u}(x)$ into smooth component $\boldsymbol{v}(x)$ and the singular component $\boldsymbol{w}(x)$. To find this, we use the asymptotic expansion technique. For the smooth component $\boldsymbol{v}(x)$, we use the asymptotic expansion $\boldsymbol{v}(x) = \boldsymbol{v}_0(x) + \varepsilon\boldsymbol{v}_1(x) = (v_{01}(x) + \varepsilon v_{11}(x), v_{02}(x) + \varepsilon v_{12}(x))^T$ where $\boldsymbol{v}_1(0) = 0$ and $\boldsymbol{v}_1(1) = 0$. Hence the leading order term $\boldsymbol{v}_0(x)$ of $\boldsymbol{v}(x)$ satisfies (by putting $\varepsilon = 0$ in (1))

$$b_{11}(x)v_{01}(x) + b_{12}(x)v_{02}(x) = f_1(x), \quad b_{21}(x)v_{01}(x) + b_{22}(x)v_{02}(x) = f_2(x),$$

which ensures the ε uniform bound of $v_{01}(x)$ and $v_{02}(x)$. In a similar way, following the technique given in [6], we can obtain the derivative bounds in (4). Now, we decompose the singular/layer component $\boldsymbol{w}(x)$ as $\boldsymbol{w}(x) = \boldsymbol{w}^l(x) + \boldsymbol{w}^r(x)$ such that

$$\begin{cases} \boldsymbol{L}\boldsymbol{w}^l(x) = 0, \quad x \in \Omega, \\ \boldsymbol{w}^l(0) = \boldsymbol{w}(0), \quad \boldsymbol{w}^l(1) = 0, \end{cases} \quad \text{and} \quad \begin{cases} \boldsymbol{L}\boldsymbol{w}^r(x) = 0, \quad x \in \Omega, \\ \boldsymbol{w}^r(0) = 0, \quad \boldsymbol{w}^r(1) = \boldsymbol{w}(1). \end{cases}$$

Let us introduce the asymptotic expansions for the left-hand singular component $\boldsymbol{w}^l = (w_1^l, w_2^l)^T$ and the right-hand singular component $\boldsymbol{w}^r = (w_1^r, w_2^r)^T$ as

$$w_s^l(x) = w_{0s}^l(x) + \sqrt{\varepsilon}w_{1s}^l(x) + \varepsilon w_{2s}^l(x) + \cdots, \quad w_s^r(x) = w_{0s}^r(x) + \sqrt{\varepsilon}w_{1s}^r(x) + \varepsilon w_{2s}^r(x) + \cdots,$$

 ISBN: 978-84-617-8694-7

where $s = 1, 2$. To find the derivative bounds of $\boldsymbol{w}^l$, we shall use the transformation $\xi = x/\sqrt{\varepsilon}$. Using the Taylor series expansions of $b_{mn}(\xi\sqrt{\varepsilon})$, $m, n = 1, 2$, comparing the terms of order ε, it can be checked that the leading order terms of $\boldsymbol{w}^l(\xi)$ satisfies

$$\hat{L}_1 w_{01}^l + b_{12}(0) w_{02}^l = 0, \quad \hat{L}_2 w_{02}^l + b_{21}(0) w_{01}^l = 0,$$
$$w_{0k}^l(0) = -v_{0k}(0) \text{ with } \lim_{\xi\to\infty} w_{0k}^l(\xi) = 0, \text{ for } k = 1, 2,$$

where $\hat{L}_k \equiv -\dfrac{d^2}{d\xi^2} + b_{kk}(0)I$ and so on. In a similar way, the leading order terms of $\boldsymbol{w}^r(\xi)$ satisfies

$$\hat{L}_1 w_{01}^r + b_{12}(1) w_{02}^r = 0, \quad \hat{L}_2 w_{02}^r + b_{21}(1) w_{01}^r = 0,$$
$$w_{0k}^r(1) = -v_{0k}(1) \text{ with } \lim_{\xi\to-\infty} w_{0k}^r(\xi) = 0, \text{ for } k = 1, 2,$$

and so on. Now we can use a similar approach provided in [6] to have $|w_k^l(1)| < C\exp(-\beta/\varepsilon)$ and $|w_k^{l\ (p)}(x)| \le C\varepsilon^{-p/2}\exp\left(-x\sqrt{\beta/\varepsilon}\right)$, where $k = 1, 2$. By a similar way, using the asymptotic expansion technique, we can obtain the derivative bounds of the right-hand side of the boundary layer. This completes the proof of this lemma. ∎

From now onwards, we shall concentrate only on the finer bounds of the derivatives of $\boldsymbol{v}(x)$ and $\boldsymbol{w}(x)$. The *a priori* derivative bounds of the solution suggest that the boundary layer phenomena occurs from the singular component $\boldsymbol{w}(x) = (w_1(x), w_2(x))^T$ of the solution. Mesh equidistribution in (3) can also be thought of as a mapping $x = x(\xi)$ from a computational coordinate $\xi \in [0, 1]$ to the physical coordinate $x \in \overline{\Omega}$, defined by

$$\int_0^{x(\xi)} M\big(s, \boldsymbol{u}(s)\big) ds = \xi \int_0^1 M\big(s, \boldsymbol{u}(s)\big) ds. \tag{6}$$

To simplify the error analysis, we consider a monitor function, which only involves the singular component, since the derivatives of the singular component are not uniformly bounded with respect to ε. Here, we consider

$$M(x, \boldsymbol{u}(x)) = \alpha + |w_1''(x)|^{1/2} + |w_2''(x)|^{1/2}, \tag{7}$$

which involves the singular components (w_1, w_2) of $\boldsymbol{u}$. This choice will make the analysis simpler. Here α is a positive constant which is defined in Lemma 2.2. To approximate $w_k''(x)$, $k = 1, 2$, we use Lemma 2.1 with $0 < \varepsilon \ll 1$. Hence, the leading order expansion of $\boldsymbol{w}(x)$ follows

$$|w_m''(x)| \approx \begin{cases} \kappa_{m1}\varepsilon^{-1}\exp\left(-x\sqrt{\dfrac{\beta}{\varepsilon}}\right), & x \in [0, 1/2], \\ \kappa_{m2}\varepsilon^{-1}\exp\left(-(1-x)\sqrt{\dfrac{\beta}{\varepsilon}}\right), & x \in (1/2, 1], \end{cases}$$

 ISBN: 978-84-617-8694-7

where κ_{m1} and κ_{m2} for $m = 1, 2$, are constants which are independent of ε and x. Hence, we have

$$\int_0^1 (|w_1''(x)|^{1/2} + |w_2''(x)|^{1/2})dx \equiv K \approx 2\big[\sum_{i,j=1}^{2} |\kappa_{ij}|^{1/2}\big]/\sqrt{\beta}.$$

Substituting the approximate value of $w_k''(x)$ in (6), we obtain for $x(\xi) \leq 1/2$,

$$\xi\left(\frac{\alpha}{K}+1\right) = \alpha\frac{x(\xi)}{K} + \frac{2\sum_{i=1}^{2}|\kappa_{i1}|^{1/2}}{K\sqrt{\beta}}\left[1-\exp\left(-\frac{x(\xi)}{2}\sqrt{\frac{\beta}{\varepsilon}}\right)\right]. \tag{8}$$

Similarly, for $x(\xi) > 1/2$, we have

$$(1-\xi)\left(\frac{\alpha}{K}+1\right) = \alpha\frac{(1-x(\xi))}{K} + \frac{2\sum_{i=1}^{2}|\kappa_{i2}|^{1/2}}{K\sqrt{\beta}}\left[1-\exp\left(-\frac{(1-x(\xi))}{2}\sqrt{\frac{\beta}{\varepsilon}}\right)\right]. \tag{9}$$

Note that the equidistribution principle (6) is a mapping from the physical nonuniform coordinates $\{x_i\}_{i=0}^N$ to the computational coordinates of uniform meshes $\{\xi_i = i/N\}_{i=0}^N$. Hence, we have from (8) and (9) that

$$\frac{\alpha x_i}{K} + K_1\left[1-\exp\left(-\frac{x_i}{2}\sqrt{\frac{\beta}{\varepsilon}}\right)\right] = \frac{i}{N}\left(\frac{\alpha}{K}+1\right), \quad \text{for } x_i \leq 1/2, \tag{10}$$

and for $x_i > 1/2$

$$\frac{\alpha(1-x_i)}{K} + K_2\left[1-\exp\left(-\frac{(1-x_i)}{2}\sqrt{\frac{\beta}{\varepsilon}}\right)\right] = \left(1-\frac{i}{N}\right)\left(\frac{\alpha}{K}+1\right), \tag{11}$$

where

$$K_1 = \frac{\sum_{i=1}^{2}|\kappa_{i1}|^{1/2}}{\sum_{i,j=1}^{2}|\kappa_{ij}|^{1/2}} \quad \text{and} \quad K_2 = 1-K_1.$$

Hence, the adaptively generated mesh x_i is the solution of the nonlinear algebraic equations (10) and (11). Now let us denote the step sizes on an arbitrary mesh $\Omega^N \equiv \{0 = x_0 < x_1 < \cdots < x_N = 1\}$ as $h_i = x_i - x_{i-1}$. The following lemma provides the structure of the mesh distribution and a choice for α.

Lemma 2.2 *Let us assume that $\alpha = K$ in (10) and (11). Then the mesh points satisfy*

$$x_{k_l} < 2\sqrt{\frac{\varepsilon}{\beta}}\ln(N) < x_{k_l+1}, \text{ and } x_{k_r-1} < 1-2\sqrt{\frac{\varepsilon}{\beta}}\ln(N) < x_{k_r},$$

 ISBN: 978-84-617-8694-7

with $k_l = \left[\frac{K_1}{2}(N-1) + \sqrt{\frac{\varepsilon}{\beta}} N \ln(N)\right]$, $\quad k_r = \left[N - \frac{K_2}{2}(N-1) - \sqrt{\frac{\varepsilon}{\beta}} N \ln(N)\right] + 1$, *where* $[\cdot]$ *denotes the integral part of the inside quantity. In particular, the mesh widths within the boundary layers i.e., for* $i = 1, \cdots, k_l, k_r + 1, \cdots, N$, *satisfy* $h_i < C\sqrt{\varepsilon/\beta}$, *with*

$|h_{i+1} - h_i| \leq Ch_i^2, \quad i = 1, \cdots, k_l - 1$, *and* $|h_{i+1} - h_i| \leq Ch_{i+1}^2, \quad i = k_r + 1, \cdots, N-1$, *and*

$\exp\left(\frac{-x_i}{2}\sqrt{\frac{\beta}{\varepsilon}}\right) \leq CN^{-1}$, $i \geq k_l - 1$, $x_i \leq 1/2$, *and* $\exp\left(\frac{-(1-x_i)}{2}\sqrt{\frac{\beta}{\varepsilon}}\right) \leq CN^{-1}$, $i \leq k_r$, $x_i > 1/2$.

Proof. The proof of the above results on mesh structure follows from [1]. ∎

The following lemma provides upper bounds for the mesh spacings generated by (10-11).

Lemma 2.3 *Mesh spaces generated by the equidistribution of monitor function (7) satisfy*

$$h_i \leq CN^{-1}, \text{ for } i = 1, \cdots, N.$$

Proof. Note that (7) satisfies $M(x, \boldsymbol{u}(x)) \geq \alpha = K$. The derivative bounds of Lemma 2.1 follows $\int_0^1 M(x, \boldsymbol{u}(x))dx \leq C_1$ where C_1 is independent of ε. Hence, (3) implies

$$\alpha h_i \leq \int_{x_{i-1}}^{x_i} M(x, \boldsymbol{u}(x))dx = \frac{1}{N}\int_0^1 M(x, \boldsymbol{u}(x))dx \leq C_1 N^{-1}.$$

Therefore $h_i \leq CN^{-1}$. ∎

3 Discrete Problem

In this section, we consider the discrete problem and its stability for (1). We use the following finite difference schemes to discretize (1) on $\Omega^N \equiv \{0 = x_0 < x_1 < \cdots < x_N = 1\}$. For a given discrete function $\eta(x_i) = \eta_i$, we use the central difference operator

$$\delta^2\eta_i = \frac{(D^+\eta_i - D^-\eta_i)}{\hbar_i} \quad \text{where} \quad D^+\eta_i = \frac{\eta_{i+1} - \eta_i}{h_{i+1}} \text{ and } D^-\eta_i = \frac{\eta_i - \eta_{i-1}}{h_i},$$

where $\hbar_i = (h_i + h_{i+1})/2$ to discretize the problem (1). Therefore, the discrete problem corresponding to the continuous version (1) becomes:
Find $\boldsymbol{U} = (U_1, U_2)^T$ such that

$$\boldsymbol{L}^N \boldsymbol{U}_i = \boldsymbol{f}_i \Leftrightarrow \begin{cases} L_1^N \boldsymbol{U}_i \equiv -\varepsilon\delta^2 U_{1,i} + b_{11,i} U_{1,i} + b_{12,i} U_{2,i} = f_{1,i}, \\ L_2^N \boldsymbol{U}_i \equiv -\varepsilon\delta^2 U_{2,i} + b_{21,i} U_{1,i} + b_{22,i} U_{2,i} = f_{2,i}, \quad i = 1, \cdots, N-1, \\ U_{1,0} = U_{1,N} = U_{2,0} = U_{2,N} = 0, \end{cases} \quad (12)$$

 ISBN: 978-84-617-8694-7

where $\boldsymbol{L}^N = (L_1^N,\ L_2^N)^T$ is the discrete operator corresponding to the continuous operator $\boldsymbol{L}$ and $\boldsymbol{f}_i = (f_{1,i}, f_{2,i})^T$. By solving the system of linear algebraic equations in (12) on the adaptive equidistributed mesh obtained by the equidistribution of (7), we obtain the numerical solution $\boldsymbol{U}$ on the adaptive mesh $x_0, x_1, \cdots, x_N$. Observe that the discrete operator $\boldsymbol{L}^N$ satisfies the following comparison principle.

Lemma 3.1 *If two mesh functions $\boldsymbol{V}$ and $\boldsymbol{W}$, where $\mathbf{V}_i = (V_{1,i}, V_{2,i})$ and $\mathbf{W}_i = (W_{1,i}, W_{2,i})$ satisfy $(\boldsymbol{L}^N \boldsymbol{V})_i \geq (\boldsymbol{L}^N \boldsymbol{W})_i$, for $1 \leq i \leq N-1$, with $\boldsymbol{V}_0 \geq \boldsymbol{W}_0$ and $\boldsymbol{V}_N \geq \boldsymbol{W}_N$, then $\boldsymbol{V}_i \geq \boldsymbol{W}_i$, for $1 \leq i \leq N-1$.*

Proof. Let us assume $\mathbf{Z}_i = \mathbf{V}_i - \mathbf{W}_i$. Therefore, it is enough to show that $\mathbf{Z}_i \geq 0$ under the conditions stated above. Now, let us assume that the above result is not true, i.e., there exists $x_p \in \Omega^N$ such that $Z_{q,p} < 0$ for some q. Let $\boldsymbol{S}_i = (S_{1,i}, S_{2,i})^T$ where $S_{k,i} = 1$ for all i and k and assume $\phi = \max_k\{\max_i(-Z_k/S_k)(x_i)\}$. This implies that there exists a mesh point $x_* \in \Omega^N$ such that $(Z_t + \phi S_t)(x_*) = 0$ for some t. Note that $(Z_{k,i} + \phi S_{k,i}) \geq 0$ for all $i = 0, \cdots, N$. Therefore, the minimum of $Z_k + \phi S_k$ attains at x_*.

Again $(L_k^N \boldsymbol{S})_i = \sum_{j=1}^2 b_{kj,i} > 0$ from (2). Therefore

$$\begin{aligned} 0 < (L_t^N(\boldsymbol{Z} + \phi \boldsymbol{S}))_* &= -\varepsilon\delta^2(Z_t + \phi S_t)_* + \textstyle\sum_{j=1}^2 b_{tj,*}(Z_{j,*} + \phi S_{j,*}) \\ &= -\varepsilon\delta^2(Z_t + \phi S_t)_* + \textstyle\sum_{j=1, j\neq t}^2 b_{tj,*}(Z_{j,*} + \phi S_{j,*}) \leq 0 \end{aligned}$$

which is a contradiction. Therefore $\mathbf{Z}_i \geq 0$ for all i. ∎

An immediate consequence of the above lemma is the following stability result of (12).

Lemma 3.2 *If $\boldsymbol{U}_i = (U_{1,i}, U_{2,i})$, $0 \leq i \leq N$ is any discrete solution satisfying the problem (12), then*

$$\|\boldsymbol{U}\| \leq C \max\{\|\boldsymbol{U}_0\|, \|\boldsymbol{U}_N\|, \|\boldsymbol{L}^N \boldsymbol{U}_i\|\}. \tag{13}$$

Proof. Define the barrier function $\boldsymbol{U}_i^{\pm} = (U_{1,i}^{\pm}, U_{2,i}^{\pm})^T$ as $U_{k,i}^{\pm} = C_1 \pm U_{k,i}$, where $C_1 = C_2 \max\{\|\boldsymbol{U}_0\|, \|\boldsymbol{U}_N\|, \|\boldsymbol{L}^N \boldsymbol{U}_i\|\}$ where C_2 is a ε- independent positive user chosen constant. Then $\boldsymbol{U}_0^{\pm} \geq 0$ and $\boldsymbol{U}_N^{\pm} \geq 0$. Again,

$$(L_k^N \boldsymbol{U}^{\pm})_i = \pm L_k^N \boldsymbol{U}_i^N + C_1 \sum_{j=1}^2 b_{kj,i} \geq \pm L_k^N \boldsymbol{U}_i^N + C_1\beta > 0$$

for $C_2 \geq 1/\beta$. Therefore $\boldsymbol{U}_i^{\pm} \geq 0$ which proves the lemma. ∎

4 Convergence Analysis

Here, we describe the ε-uniform convergence analysis of the solution of (12). First, we decompose the numerical solution $\boldsymbol{U}$, into smooth and singular components $\boldsymbol{V}$ and $\boldsymbol{W}$

 ISBN: 978-84-617-8694-7

respectively, *i.e.,* $\boldsymbol{U} = \boldsymbol{V} + \boldsymbol{W}$, where $\boldsymbol{V}$ and $\boldsymbol{V}$ satisfy

$$\begin{cases} \boldsymbol{L}^N \boldsymbol{V}_i = \boldsymbol{f}_i, & i = 1, \cdots, N-1, \\ \boldsymbol{V}_0 = \boldsymbol{v}(0), \quad \boldsymbol{V}_N = \boldsymbol{v}(1), \end{cases} \text{ and } \begin{cases} \boldsymbol{L}^N \boldsymbol{W}_i = 0, & i = 1, \cdots, N-1, \\ \boldsymbol{W}_0 = \boldsymbol{w}(0), \quad \boldsymbol{W}_N = \boldsymbol{w}(1). \end{cases}$$

Therefore, using the triangle inequality, we can write that the error of the numerical solution $\boldsymbol{U}$ satisfies

$$||\boldsymbol{U}_i - \boldsymbol{u}(x_i)|| \leq ||\boldsymbol{V}_i - \boldsymbol{v}(x_i)|| + ||\boldsymbol{W}_i - \boldsymbol{w}(x_i)||, \tag{14}$$

in the region Ω^N. Now, consider the error of the smooth component $\boldsymbol{V}$. From (4) and Lemma 2.3, we have $|L_k^N(\boldsymbol{V}-\boldsymbol{v})(x_i)| = |(L_k - L_k^N)\boldsymbol{v}(x_i)| \leq C\varepsilon(h_i+h_{i+1})||v_k'''(x)||_{(x_{i-1},x_{i+1})} \leq CN^{-2}$, for $k = 1, 2$, since $\sqrt{\varepsilon} \ll N^{-1}$. Hence, we have

$$||\boldsymbol{L}^N(\boldsymbol{V} - \boldsymbol{v})(x_i)|| \leq CN^{-2}. \tag{15}$$

The consistency analysis of the singular component for $i = k_l, \cdots, k_r$, shows that the truncation error of the singular components W_1 and W_2 satisfy

$$|L_k^N(\boldsymbol{W} - \boldsymbol{w})(x_i)| \leq C\varepsilon|(\delta^2 w_k - w_k'')(x_i)| \leq C\varepsilon|w_k''(x)|_{[x_{i-1},x_{i+1}]}, \text{ for } k = 1, 2.$$

Now, Lemma 2.1 implies that $||\boldsymbol{L}^N(\boldsymbol{W} - \boldsymbol{w})(x_i)|| \leq C\exp(-x_{i-1}\sqrt{\beta/\varepsilon})$ for $x_i \leq 1/2$. Therefore, from Lemma 2.2 for $x_i \geq 1/2$ with $k_l \leq i$, we get

$$||\boldsymbol{L}^N(\boldsymbol{W} - \boldsymbol{w})(x_i)|| \leq C\exp\left(-x_{k_l-1}\sqrt{\beta/\varepsilon}\right) = C\left(\exp\left(-\frac{x_{k_l-1}}{2}\sqrt{\beta/\varepsilon}\right)\right)^2 \leq CN^{-2}.$$

In a similar manner, the case $x_i > 1/2$ with $k_r \geq i$ can be carried out. Hence, for $i = k_l, \cdots, k_r$,

$$||\boldsymbol{L}^N(\boldsymbol{W} - \boldsymbol{w})(x_i)|| \leq CN^{-2}. \tag{16}$$

Now, we consider the error estimate at the nodes corresponding to $i = 1, \cdots, k_l - 1$ and $i = k_r + 1, \cdots, N-1$. We only consider the left-hand boundary layer region as the analysis corresponding to right-hand boundary layer region can be derived by the same technique. Using Taylor's series expansion, we obtain

$$|L_k^N(\boldsymbol{W} - \boldsymbol{w})(x_i)| = \frac{\varepsilon|h_i^2 w_k'''(\eta_i^1) - h_{i+1}^2 w_k'''(\eta_i^2)|}{3(h_i + h_{i+1})}, \text{ where } \eta_i^1 \in (x_{i-1}, x_i) \text{ and } \eta_i^2 \in (x_i, x_{i+1}),$$

for $k = 1, 2$. Again $|h_i^2 w_k'''(\eta_i^1) - h_{i+1}^2 w_k'''(\eta_i^2)| \leq C(|h_i^2 - h_{i+1}^2||w_k'''(x_i)| + h_i^2(h_i+h_{i+1})|w_k^{(iv)}(x_i)|)$. Hence, Lemma 2.2 implies that

$$||\boldsymbol{L}^N(\boldsymbol{W} - \boldsymbol{w})(x_i)|| \leq C\varepsilon^{-1}h_i^2\exp\left(-x_i\sqrt{\frac{\beta}{\varepsilon}}\right) \leq C\varepsilon^{-1}\left(\int_{x_{i-1}}^{x_i}\exp\left(-\frac{s}{2}\sqrt{\frac{\beta}{\varepsilon}}\right)ds\right)^2.$$

 ISBN: 978-84-617-8694-7

It follows that

$$||\boldsymbol{L}^N(\boldsymbol{W}-\boldsymbol{w})(x_i)|| \leq C\varepsilon^{-1}\left(\sqrt{\varepsilon}\int_{x_{i-1}}^{x_i} M(t,\boldsymbol{u}(t))dt\right)^2 \leq CK^2N^{-2} \leq CN^{-2}.$$

where we have used the definition of the equidistribution principle (3) in the above inequality. Therefore, for $i = 1, \cdots, k_l - 1$ and $i = k_r + 1, \cdots, N-1$, we have

$$||\boldsymbol{L}^N(\boldsymbol{W}-\boldsymbol{w})(x_i)|| \leq CN^{-2}. \tag{17}$$

This completes the proof. ∎

Now, we write the main convergence result of the numerical solution $\boldsymbol{U}_i$.

4.1 The main convergence result

The following theorem shows that the discrete solution $\mathbf{U}$ of (12) is second-order uniformly convergent on the equidistributed mesh.

Theorem 4.1 *Let $\boldsymbol{u}$ be the solution of (1) and $\boldsymbol{U}$ be the solution of the discrete problem (12), on a mesh obtained by equidistributing the monitor function (7) with $\alpha = K$. Then, there exists a constant C, independent of ε, x_i and N, such that*

$$||\boldsymbol{u}-\boldsymbol{U}||_{\Omega^N} \leq CN^{-2}.$$

Proof. Lemma 3.2 implies that the inverse of the discrete operator $\boldsymbol{L}^N$ is ε-uniformly bounded. Therefore, using (15) and (17) in (14), we obtain that

$$|\boldsymbol{u}(x_i)-\boldsymbol{U}_i| \leq |\boldsymbol{V}_i - \boldsymbol{v}(x_i)| + |\boldsymbol{W}_i - \boldsymbol{w}(x_i)| \leq CN^{-2} \text{ for } i = 0, \cdots, N,$$

which is the required result. ∎

5 Computational Experiments

The layer adaptive meshes will be obtained by the equidistribution of the proposed monitor function (7) based on a moving mesh algorithm. We use the de Boor algorithm to move any user chosen mesh, whose convergence analysis is considered in [11]. For the convergence of this algorithm, it is sufficient to consider the following weak discrete version of (3):

$$M_i h_i \leq \frac{C_0}{N}\sum_{j=1}^{N} M_j h_j, \quad \text{for} \quad i = 1, \cdots, N$$

where M_i is the discrete approximation of the monitor function $M(x, \boldsymbol{u}(x))$ in the subinterval (x_{i-1}, x_i). In practice, we use the central difference approximation to discretize the second

 ISBN: 978-84-617-8694-7

order derivatives appeared in (7). Here $C_0 > 1$ is a user chosen constant. C_0 can be chosen large enough to get fewer iterations for the convergence of the algorithm.

Algorithm 1: (Mesh Generation and Adaptive Solution)

Step 1: Define the initial uniform mesh $\{x^{(0)} = i/N : 0 \le i \le N\}$ and go to Step 2 with $p = 0$.

Step 2: Solve the discrete problem $\boldsymbol{L}^N \boldsymbol{U}_i^{(p)} = \boldsymbol{f}_i^{(p)}$, defined in (12) for $(\boldsymbol{U}_0^{(p)}, \cdots, \boldsymbol{U}_N^{(p)})$ on the uniform mesh $\{x_i^{(p)} : i = 0, \cdots, N\}$. Define $h_i^{(p)} = x_i^{(p)} - x_{i-1}^{(p)}$ for $i = 1, \cdots, N$.

Step 3: Find the smooth component $\boldsymbol{V}_i^{(p)}$ by solving (1) with $\varepsilon = 0$. Denote the layer component of $\boldsymbol{U}_i^{(p)}$ as $\boldsymbol{W}_i^{(p)} = \boldsymbol{U}_i^{(p)} - \boldsymbol{V}_i^{(p)}$. Find the discretized monitor function $M_i^{(p)} = \left[\alpha^{(p)} + \sum_{k=1}^{2} |\overline{\delta}^2 W_{k,i}^{(p)}|^{1/2}\right]$ for $i = 1, \cdots, N$. by defining $\overline{\delta}^2 W_{k,i} = (\delta^2 W_{k,i} + \delta^2 W_{k,i-1})/2$ with $\overline{\delta}^2 W_{k,1} = \delta^2 W_{k,1}$ and $\overline{\delta}^2 W_{k,N} = \delta^2 W_{k,N-1}$, for $k = 1, 2$. where $\alpha^{(p)} = \sum_{i=1}^{N} h_i (\sum_{k=1}^{2} |\overline{\delta}^2 W_{k,i}|^{1/2})$. Compute the total length $\Phi_j^{(p)} = \sum_{i=1}^{j} h_i^{(p)} M_i^{(p)}$.

Step 4: Choose a constant $C_0 \ge 1$. The stopping criteria for iterative technique is $\max\limits_{i=1,\ldots,N} h_i^{(p)} M_i^{(p)} / \Phi_N^{(p)} \le C_0 N^{-1}$. If it holds, then go to Step 6, else continue with step 5.

Step 5: Generate a new mesh: Set $Y_i^{(p)} = i\Phi_N^{(p)}/N$ for $i = 0, \cdots, N$. Now interpolate $(Y_i^{(p)}, x_i^{(p+1)})$ to $(\Phi_i^{(p)}, x_i^{(p)})$ using piecewise linear interpolation. Generate a new mesh $x^{(p+1)} \equiv \{0 = x_0^{(p+1)} < x_1^{(p+1)} < \cdots < x_N^{(p+1)} = 1\}$ and return to Step 2.

Step 6: Set $x^* = \{0 = x_0^* < x_1^* < \cdots < x_N^* = 1\} = x^{(p+1)}$ and $\boldsymbol{U}^* = \boldsymbol{U}^{(p+1)}$, where x^* is the boundary layer adaptive mesh and $\boldsymbol{U}^*$ is our desired adaptive solution. Stop.

5.1 Numerical Examples

Now, we demonstrate two numerical examples to validate our monitor function (7).

Example 5.1 *Consider the following system of second-order reaction-diffusion problems:*

$$\begin{cases} -\varepsilon u_1''(x) + (10 + \exp(-x))u_1(x) - 6x^2 u_2(x) = 6 + 5x^2, \quad x \in \Omega = (0,1), \\ -\varepsilon u_2''(x) - x^4|x|u_1(x) + (7 + 2x^3)u_2(x) = 5 + x^3, \\ u_1(0) = u_1(1) = u_2(0) = u_2(1) = 0. \end{cases}$$

Example 5.2 *Consider the system of singularly perturbed second-order problems:*

$$\begin{cases} -\varepsilon u_1''(x) + 2(x+1)^2 u_1(x) - (1 + x^3)u_2(x) = 2\exp(x), \quad x \in \Omega = (0,1), \\ -\varepsilon u_2''(x) - 2\cos\left(\dfrac{\pi x}{4}\right) u_1(x) + 2.2\exp(1-x)u_2(x) = 10x + 1, \\ u_1(0) = u_1(1) = u_2(0) = u_2(1) = 0. \end{cases}$$

 ISBN: 978-84-617-8694-7

We compute the accuracy of the numerical solution by using the following double mesh principle: for any value of N, the maximum pointwise error $E^N_{m,\varepsilon}$, $m = 1, 2$, of the numerical solution is calculated by $E^N_{m,\varepsilon} = \max\limits_{0\le i\le N} |U^N_m - \overline{U}^{2N}_m|$, where $\boldsymbol{U}^N$ is the computed solution with N number of intervals and $\overline{\boldsymbol{U}}^{2N} = (\overline{U}^N_1, \overline{U}^N_2)$ is the numerical solution on a mesh, obtained by bisecting the original mesh such that the ith point of the original mesh coincides with the $2i$ th point of the newly obtained mesh. The uniform error, say E^N_m, for each fixed N is defined by $E^N_m = \max_{\varepsilon\in S} E^N_{m,\varepsilon}$, where the set $S = \{\varepsilon|\varepsilon = 1, 2^{-2}, \cdots, 2^{-30}\}$. The order of convergence is calculated by $r^N_m = \log_2\left(E^N_m/E^{2N}_m\right)$. We took $C_0 = 1.2$. in Algorithm 1.

Table 1: *Uniform errors and orders of convergence of U_1 for Example 5.1.*

$\varepsilon \in S$	Number of intervals N						
	64	128	256	512	1024	2048	4096
E^N_1	3.9638e-3	1.1440e-3	2.2835e-4	5.4579e-0	1.3420e-5	3.3254e-6	8.3215e-7
r^N_1	1.7928	2.3248	2.0648	2.0239	2.0128	1.9986	-

Table 2: *Uniform errors and orders of convergence of U_2 for Example 5.1.*

$\varepsilon \in S$	Number of intervals N						
	64	128	256	512	1024	2048	4096
E^N_2	2.2397e-3	6.1678e-4	1.2372e-4	2.9406e-5	7.2295e-6	1.7923e-6	4.4822e-7
r^N_2	1.8605	2.3177	2.0729	2.0241	2.0121	1.9996	-

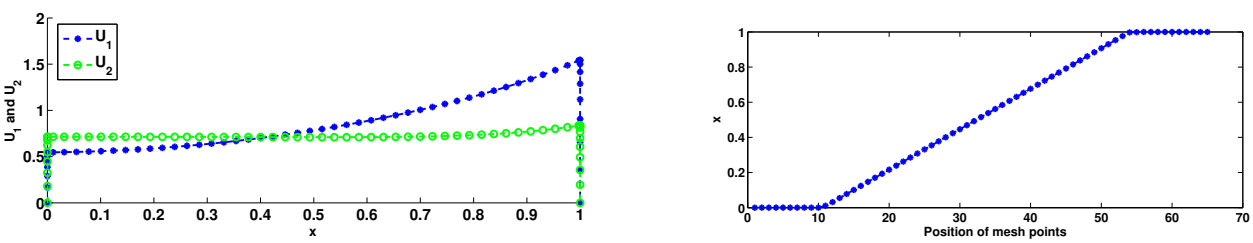


Figure 1: *Solution plot and the mesh density for $N = 64$, $\varepsilon = 2^{-30}$ resp. for Example 5.1.*

The maximum ε-uniform errors E^N_m and the orders of convergence r^N_m of the (U_1, U_2) displayed in Tables 1, 2,3, 4 shows that the second order convergence hold true in practice. This rate is better than the rate of convergence on uniform mesh (where numerical solution diverges) and Shishkin mesh (where almost second order accuracy observed in [3]) which appear in Tables 6, 7 resp. The Shishkin mesh is constructed by dividing the domain $[0, 1]$ into three subdomains $[0, \tau]$, $[\tau, 1-\tau]$ and $[1-\tau, 1]$ with $\tau = \min\{1/4, 2\sqrt{\varepsilon/\beta}\log(N)\}$. Here each subdomain contains $N/4$, $N/2$ and $N/4$ number of mesh intervals respectively with

Table 3: *Uniform errors and orders of convergence of U_1 for Example 5.2.*

$\varepsilon \in S$	Number of intervals N						
	64	128	256	512	1024	2048	4096
E_1^N	8.5058e-3	2.7059e-3	7.7061e-4	1.8278e-4	4.5170e-5	1.1221e-5	2.7994e-6
r_1^N	1.6523	1.8120	2.0759	2.0166	2.0092	2.0030	-

Table 4: *Uniform errors and orders of convergence of U_2 for Example 5.2.*

$\varepsilon \in S$	Number of intervals N						
	64	128	256	512	1024	2048	4096
E_2^N	6.1937e-3	1.4829e-3	3.5946e-4	8.3424e-5	2.0669e-5	5.1377e-6	1.2831e-6
r_2^N	2.0624	2.0445	2.1073	2.0130	2.0082	2.0014	-

uniform step size. We take $\beta = .5$ to generate Table 7. Note that the present method also works for nonsingularly perturbed problems as the set S contains $\varepsilon = 1$. Table 5 presents the number of iterations taken by Algorithm 1 which shows the effectiveness our preferred monitor function.

Figures 1, 2 show the boundary layer phenomena and mesh density towards the boundary points $x = 0, 1$ of Examples 5.1 and 5.2. The mesh movements in Figure 3 shows the effectiveness of Algorithm 1 to generate layer adapted mesh. In addition, Figure 4 shows that the depicted second order convergence holds true in logarithmic scale as N increases.

6 Conclusion

In this paper, we have proposed a computational technique for a system of singularly perturbed reaction-diffusion problems exhibiting boundary layers by moving mesh methods. An error monitor function is proposed whose equidistribution will lead to a boundary layer adapted mesh. Through theoretical and numerical results, it is shown that the proposed numerical method is second order ε-uniform convergent.

Table 5: *Number of iterations taken by the Algorithm 1 for Example 5.2.*

ε	Number of intervals N					
	64	128	256	512	1024	2048
1	0	0	0	0	0	0
2^{-6}	1	1	1	1	1	1
2^{-12}	2	2	1	1	1	1
2^{-18}	3	3	2	2	2	1
2^{-24}	8	5	4	3	3	2

 ISBN: 978-84-617-8694-7

Table 6: *Uniform errors and orders of convergence on uniform mesh for Example 5.2.*

$\varepsilon \in S$	Number of intervals N						
	64	128	256	512	1024	2048	4096
E_1^N	6.4085e-2	6.4311e-2	6.4429e-2	6.4489e-2	6.4519e-2	6.4535e-2	6.4542e-2
r_1^N	-5.0786e-3	-2.6466e-3	-1.3500e-3	-6.8165e-4	-3.4249e-4	-1.7166e-4	-
E_2^N	1.8680e-1	1.8768e-1	1.8813e-1	1.8836e-1	1.8847e-1	1.8853e-1	1.8856e-1
r_2^N	-6.8190e-3	-3.4479e-3	-1.7335e-3	-8.6913e-4	-4.3516e-4	-2.1773e-4	-

Table 7: *Uniform errors and orders of convergence on Shishkin mesh for Example 5.2.*

$\varepsilon \in S$	Number of intervals N						
	64	128	256	512	1024	2048	4096
E_1^N	2.6150e-2	9.6536e-3	3.3847e-3	1.0961e-3	3.3995e-4	1.0304e-4	3.0673e-5
r_1^N	1.4377	1.5120	1.6267	1.6889	1.7221	1.7482	-
E_2^N	6.2734e-2	2.2667e-2	7.5931e-3	2.4204e-3	7.5022e-4	2.2707e-4	6.7596e-5
r_2^N	1.4687	1.5778	1.6495	1.6898	1.7242	1.7481	-

References

[1] M.G. Beckett and J.A. Mackenzie. On a uniformly accurate finite difference approximation of a singularly perturbed reaction-diffusion problem using grid equidistribution. *J. Comput. Appl. Math.*, **131**:381–405, 2001.

[2] P. Das. Comparison of a priori and a posteriori meshes for singularly perturbed nonlinear parameterized problems. J. Comput. Appl. Math., **290**, 16-25, 2015.

[3] P. Das and S. Natesan. A uniformly convergent hybrid scheme for singularly perturbed system of reaction-diffusion Robin type boundary value problems. *J. Appl. Math. Comput.*, **41(1-2)**:441-471, 2013.

[4] W. Huang and R. D. Russell. Adaptive moving mesh methods. *Springer*, 2011.

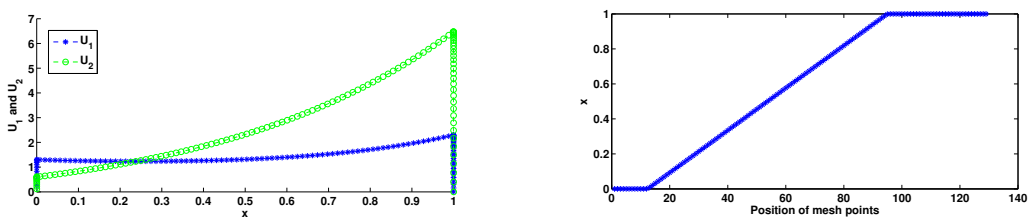
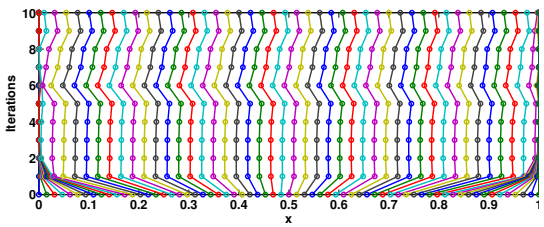
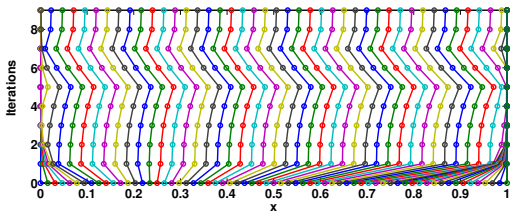


Figure 2: *Solution plot and the mesh density for* $N = 128$, $\varepsilon = 2^{-30}$ *resp. for Example 5.2.*

 ISBN: 978-84-617-8694-7

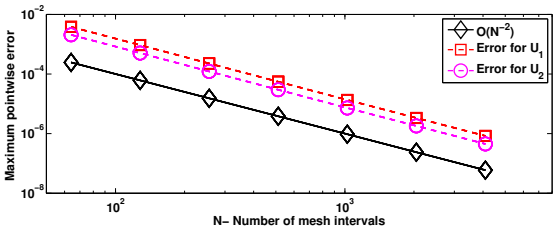


(a) Examples 5.1

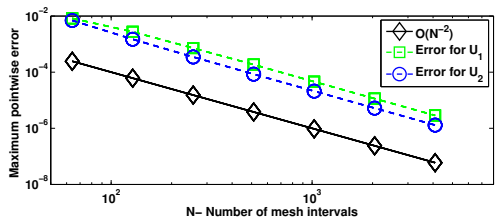


(b) Examples 5.2.

Figure 3: *Mesh trajectories towards boundary layers* $x = 0$ *and* $x = 1$ *for* $N = 64$, $\varepsilon = 2^{-30}$.



(a) $\varepsilon = 2^{-34}$ *for Example 5.1.*



(b) $\varepsilon = 2^{-40}$ *for Example 5.2.*

Figure 4: *Loglog plot of maximum point-wise errors.*

[5] N. Madden and M. Stynes. A uniformly convergent numerical method for a coupled system of two singularly perturbed linear reaction-diffusion problems. *IMA J. Numer. Anal.*, **23(4)**:627–644, 2003.

[6] S. Matthews, E. O'Riordan,G.I. Shishkin.A numerical method for a system of singularly perturbed reaction-diffusion equations. *J. Comput. Appl. Math.*,**145**:151-166, 2002.

[7] J. Mohapatra. Equidistribution grids for two parameter convection-diffusion boundary value problems, *J. Math. Model.*, **2**, (2014) 1-21.

[8] D.S. Naidu and A. Calise. Singular perturbations and time scales in guidance and control of aerospace systems:a survey. *J. Guidance Control Dynam.*,**24**,1057-1078, 2001.

[9] S. Natesan, J. Vigo-Aguiar, N. Ramanujam. A numerical algorithm for singular perturbation problems exhibiting weak boundary layers. *Comput. Math. Appl.*, **45**:46-79,2003.

[10] J. Vigo-Aguiar and S Natesan. An efficient numerical method for singular perturbation problems. *J. Comput. Appl. Math.*, **192 (1)**, 132-141, 2006.

[11] X. Xu, W. Huang, R. D. Russell, J. F. Williams. Convergence of de Boor's algorithm for the generation of equidistributing meshes. *IMA J. Numer. Anal.*, **31**, 580-596, 2011.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

On a generalized variable-coefficient Gardner equation with forcing term

R. de la Rosa[1], E. Recio[1], T.M. Garrido[1] and M.S. Bruzón[1]

[1] *Departamento de Matemáticas, Universidad de Cádiz*

emails: rafael.delarosa@uca.es, elena.recio@uca.es, tamara.garrido@uca.es, m.bruzon@uca.es

Abstract

Key words: Conservation laws, Equivalence transformations, Gardner equation, Nonlinear partial differential equations.

In the last decade, there has been an increasing interest in the research of several generalizations of some important dispersive wave equations such as the Korteweg-de Vries (KdV) equation, mKdV equations and the Burgers equation. Among these generalizations, special attention should be paid to variable-coefficient nonlinear equations because these describe a great number of nonlinear phenomena more realistically than their constant-coefficient counterparts. Specifically, the Gardner equation, also known as combined KdV-mKdV equation, is widely used in different branches of physics such as fluid dynamics, solid state physics or quantum field theory. Moreover, it describes interesting physical phenomena: the long wave propagation in an inhomogeneous two-layer shallow liquid, internal waves in a stratified ocean, ion acoustic waves in plasma with a negative ion...

In this paper, we consider a generalized variable-coefficient Gardner equation with nonlinear terms of any order and forcing term given by

$$u_t + a(t)u^n u_x + b(t)u^{2n}u_x + c(t)u_{xxx} + h(t)u_x + f(t)u = -r(t), \tag{1}$$

where n is a positive constant, $a(t)$ and $b(t)$ are not simultaneously equal to zero, $c(t) \neq 0$, $f(t)$, $h(t)$ and $r(t)$ are arbitrary functions. The Gardner equation has been recently studied by different authors [5, 6, 7, 8, 9]. In [9] a large number of solutions of equation (1)

ISBN: 978-84-617-8694-7

with $r(t) = 0$ were constructed. The authors obtained these solutions with the aid of two first-order nonlinear ordinary differential equations and a new generalized algebraic method. Later, in [6], exact solutions of equation (1) were obtained by using the general mapping deformation method which included soliton solutions, Jacobi elliptic wave solutions and Weierstrass elliptic function solution, among others. Finally, in [5] the authors performed an analysis of the classical and nonclassical symmetries admitted by the constant-coefficient counterpart of equation (1) and they constructed some exact travelling wave solutions by using the simplest equation method.

The problem lies in the fact that the study of variable-coefficient equations is often difficult. Equivalence transformations fit perfectly into the study of variable-coefficient partial differential equations (PDEs). Equivalence transformations allow us to determine those inessential elements of the class under consideration and to perform a transformation which maps these arbitrary elements to chosen simple values from the beginning. The use of these transformations often provides an alternative way to simplify a group classification problem and to show the results in a simple and clear manner.

An equivalence transformation of class (1) is a nondegenerate point transformation, (t, x, u) to $\left(\tilde{t}, \tilde{x}, \tilde{u}\right)$ in the augmented space $(t, x, u, a, b, c, h, f, r)$ with the property that it preserves the differential structure of the equation but with different arbitrary functions, $\tilde{a}(\tilde{t})$, $\tilde{b}(\tilde{t})$, $\tilde{c}(\tilde{t})$, $\tilde{h}(\tilde{t})$, $\tilde{f}(\tilde{t})$ and $\tilde{r}(\tilde{t})$.

The symmetry group of a PDE is the largest group of transformations acting on the space of independent and dependent variables which transforms solutions of the equation into other solutions. The method of Lie symmetry groups is one of the most powerful methods to analyse PDEs. Among its well-known applications, we highlight that symmetry groups can be used to obtain exact solutions or to construct conservation laws.

Given a PDE, a conservation law is a space-time divergence expression

$$D_t T + D_x X = 0, \tag{2}$$

that vanishes on all solutions of the PDE, where the conserved density T and the spatial flux X are functions of t, x, u and derivatives of u, whereas D_t and D_x denote the total derivative operators with respect to t and x respectively. This concept has its origin in physics, nevertheless it presents important applications in the study of differential equations or systems of differential equations. To begin with, the integrability of a differential equation is closely linked with the existence of a large number of conservation laws. Furthermore, these laws can be used to assess the accuracy and stability of numerical methods for the solutions of PDE.

 ISBN: 978-84-617-8694-7

Anco and Bluman [1, 2, 3] proved a general method to construct conservation laws for PDEs. This method makes use of the concept of multiplier. A multiplier is a non-singular function $Q(t, x, u, u_t, u_x, \cdots)$ on the set of solutions $u(t, x)$, which satisfies that $(u_t + a(t)u^n u_x + b(t)u^{2n}u_x + c(t)u_{xxx} + h(t)u_x + f(t)u + r(t))Q$ is a divergence expression not only for solutions of equation (1), but for any function $u(t, x)$.

A conservation law is called locally trivial if there is a function $\Theta(t, x, u, u_t, u_x, \ldots)$ so that the conserved vector $(T, X) = (D_x\Theta, -D_t\Theta)$ holds for every solution $u(t, x)$, therefore equation (2) becomes an identity. From this definition, two conservation laws are considered to be locally equivalent if they differ by a locally trivial conservation law. Any non-trivial conservation law can be expressed in a general form

$$\frac{d}{dt}\int_\Omega T dx = -X\Big|_{\partial\Omega},$$

where $\Omega \subseteq \mathbb{R}$ is a fixed spatial domain. Moving off of the set of solutions of equation (1), each conservation law can be stated by using the characteristic form

$$D_t\tilde{T} + D_x\tilde{X} = (u_t + a(t)u^n u_x + b(t)u^{2n}u_x + c(t)u_{xxx} + h(t)u_x + f(t)u + r(t))Q. \qquad (3)$$

In particular, from the characteristic form (3) it follows that each conserved vector given by (2) derives from a multiplier Q of equation (1). Multipliers Q are obtained by requiring that the divergence condition must be verified identically

$$\frac{\delta}{\delta u}\left((u_t + a(t)u^n u_x + b(t)u^{2n}u_x + c(t)u_{xxx} + h(t)u_x + f(t)u + r(t))Q\right) = 0,$$

where $\frac{\delta}{\delta u} = \partial_u - D_x\partial_{u_x} - D_t\partial_{u_t} + D_xD_t\partial_{u_{xt}} + D_x^2\partial_{u_{xx}} + ...$, represents the variational derivative. Divergence condition splits with respect to u_t, u_{xxx} and their differential consequences, yielding an overdetermined system in Q and the arbitrary functions which equation (1) involves. Once multipliers have been determined, the conserved vectors can be constructed by integrating the characteristic equation (3) [3, 4].

The aim of this work is to analyse the generalized Gardner equation (1) from the point of view of Lie symmetries and conservation laws. To achieve this objective, we will derive the continuous equivalence group of equation (1). The gauging of arbitrary functions by using equivalence transformations allows us to perform an exhaustive study of equation (1) and a clear presentation of the results. Moreover, we obtain a classification of the Lie symmetries of the reduced equation. Finally, we determine nontrivial conservation laws via the direct method of the multipliers [1, 2, 3].

 ISBN: 978-84-617-8694-7

Acknowledgements

The authors gratefully acknowledge the financial support from Junta de Andalucía group FQM-201, Universidad de Cádiz and they express their sincere gratitude to the Plan Propio de Investigación de la Universidad de Cádiz.

References

[1] S. C. Anco and G. W. Bluman, *Direct construction method for conservation laws of partial differential equations Part I: Examples of conservation law classifications*, Eur. J. Appl. Math. **13** (2002), 545–566.

[2] S. C. Anco and G. W. Bluman, *Direct construction method for conservation laws of partial differential equations Part II: General treatment*, Eur. J. Appl. Math. **13** (2002), 567–585.

[3] S. C. Anco, *Generalization of Noether's theorem in modern form to non-variational partial differential equations*, To appear in Fields Institute Communications: Recent progress and Modern Challenges in Applied Mathematics, Modeling and Computational Science.

[4] G. W. Bluman, A. Cheviakov and S. C. Anco, *Applications of symmetry methods to partial differential equations*, Springer, New York, 2009.

[5] R. de la Rosa and M. S. Bruzón, *On the classical and nonclassical symmetries of a generalized Gardner equation*, Applied Mathematics and Nonlinear Sciences **1(1)** (2016) 263–272.

[6] B. Hong and D. Lu, *New exact solutions for the generalized variable-coefficient Gardner equation with forcing term*, Applied Mathematics and Computation **219** (2012) 2732–2738.

[7] O. Vaneeva, O. Kuriksha and C. Sophocleous, *Enhanced group classification of Gardner equations with time-dependent coefficients*, Commun Nonlinear Sci Numer Simulat **22** (2015) 1243–1251.

[8] A. M. Wazwaz, *A study on KdV and Gardner equations with time-dependent coefficients and forcing terms*, Applied Mathematics and Computation **217** (2010) 2277–2281.

[9] L. H. Zhang, L. H. Dong and L. M. Yan, *Construction of non-travelling wave solutions for the generalized variable-coefficient Gardner equation*, Applied Mathematics and Computation **203** (2008) 784–791.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

An study on the distances of an extension of the SMOTE algorithm for Time Series

Enrique A.de la Cal[1], José R. Villar[1], Paula Vergara[1] and Javier Sedano[2]

[1] *Computer Science Dpt., University of Oviedo*

[2] *Electronics and Artificial Intelligence Area, Instituto Tecnológico de Castilla y León*

emails: delacal@uniovi.es, villarjose@uniovi.es, paulavg09@gmail.com, javier.sedano@itcl.es

Abstract

Big data field usually involves the use of Time Series (TS) datasets. In this scenario, the detection of complex TS events are rarely presented; thus, the learning algorithms need to tackle with the TS data balancing problem, which has been barely studied. This study addresses this issue, describing a very simple TS extension of the well-known SMOTE algorithm for balancing datasets where a preliminary study on the TS distances used in different parts of the extension of the SMOTE algorithm. Besides, as the study will carry out the experiments on a realistic TS dataset simulating epilepsy attacks, a proposal to force an erratic behavior is presented. A study on the characteristics of the dataset before and after the performance of this TS balancing algorithm is performed, showing evidence on the requirements for the research on this topic, the energy efficiency of the algorithm and the TS generation process among them.
Key words: Dataset balancing algorithms, SMOTE, Time Series, Time Series Distances

1 Introduction

Big data field usually involves the use of Time Series (TS) datasets. Such cases includes the management of the sensory systems located on wearable devices, like in the human activity recognition and the abnormal movement detection [1, 2]. Furthermore, the TS datasets have become into multivariate TS datasets, which makes the data analysis even more complex.

In this context, when leaning models for the detection of some complex events, the problem of lacking data balance arises: there are many more TS segments belonging to

normal class than to the abnormal class to detect. For instance, in the problem of epilepsy seizure detection [3, 4], the occurrence of a seizure might be once in a month or even less.

The main part of the literature concerning the dataset balancing problem is focused on classical datasets, where a sample includes an atomic value for each of the features. These balancing techniques can rely on oversampling the minority classes or undersampling the majority classes; however, as long as oversampling does not produce information losses, it is preferred over undersampling.

Some valid alternatives have also been published, coping with imbalanced problems specific algorithms [5], or proposing ensembles for the minority class together with a kind of undersampling of the majority classes [6]. Examples of oversampling techniques include well-known algorithms as SMOTE (Synthetic Minority Over-sampling Technique, [7, 8]), ADASYN (ADAptive SYNthetic Sampling, [9]), ADOMS (Adjusting the Direction Of the synthetic Minority clasS examples, [10]) or SPIDER (Selective Preprocessing of Imbalanced Data, [11]).

However, the problem of balancing TS datasets has not received much attention. In a TS datasets, each sample includes a TS for each feature. Moreover, the sample is assigned a class, but also a TS is attached as the labelling TS for that sample. From now on, we consider all the TS features from a sample with the same length and sampling frequency; however, the variability in these factors needs further study. Some approaches for TS datasets balancing have focused on univariate TS problems, either.

Several solutions studied how to classify the values in the incoming sequence [12, 13, 14, 15], where the known data sequence labels are clearly biased to the majority class. Therefore, the solutions rely on drawing new synthetic atomic values based on any of the above mentioned algorithms. On the other hand, Koknar et al proposed the balancing of univariate TS based on suggesting ghost points [16]. These ghosts points belong to the domain space of TS distances. With the distance matrix a SVM classifier is learned; allowing to generate a new TS and assigning a it a class. Different TS distance measurements were proposed, the Dynamic Time Warping (DTW) distance measurement along them.

Clearly, there is a need of tackling the multivariate TS datasets balancing problem. This study addresses this topic, extending the well-known SMOTE algorithm to cope with multivariate TS. Besides, our proposal use a simpler strategy to draw the synthetic TS than the one used by Koknar [16]. Current proposal includes the use of KNN algorithm to select the parents for new synthetic TS as well as the merging operator TS_AVERAGE presented in our previous work[17]. In addition, as the study will carry out the experiments on a realistic TS dataset simulating epilepsy attacks, a proposal to force an erratic behavior of the simulated TS is presented

The experimentation will analyze the distortion in the dataset due to the inclusion of the new TS samples. This study is structured as follows. Next section outlines the SMOTE algorithm, while the design issues are explained and possible solutions are given in Sect. 3.

 ISBN: 978-84-617-8694-7

Experimentation and the discussion on the results are coped in Sect. 4. Finally, the main conclusions are drawn.

2 The SMOTE algorithm

The SMOTE algorithm is an oversampling method [7], where each sample from the minority class is randomly combined with each of its nearest neighbors to balance the dataset. This method assumes a two-class problem, however, it can be easily extended to a multi-class problem.

Algorithm 1 reproduce the algorithm from the original paper for the sake of autocompletion. The parameters of this method include the number of nearest neighbor to consider (k, by default $k = 5$ has been proposed), the number of samples belonging to the minority class (T) and the number of synthetic samples to generate for each original sample from the minority class (N). This parameter N is given as a percentage; values smaller than 100% reduces the original minority subset and produces a new dataset of the same size as the original. Whenever $N > 100$ means that $N/100$ synthetic samples are to be generated for each sample from the minority class.

As can be seen, SMOTE takes a sample and searches for some neighbors; each synthetic sample is generated as random linear combination of the two considered samples. This method has been successfully tested on different domains; and plenty of different versions have been published [8]. One the most known version of SMOTE is*SMOTE-ENN*, a sort of de-noising version, that will not be considered in current work.

3 Tackling the TS balancing problem

Two main concerns, at least, have to be solved in order to allow the SMOTE algorithm to cope with TS datasets. The first concern is related to the method for choosing the parents TS to mate, the second focuses on the generation of the new TS sample.

Choosing the two TS samples that will be used in generating the new TS offspring should consider TS grouping according to some measurement. The original SMOTE randomly selects the parents for mating among those belonging to the minority class that are the nearest neighbours (KNN). However, different solutions can be considered; for instance, the solution proposed in ADASYN [9], where the parents are randomly chosen according to the distribution of the size of the neighborhood, is totally valid as well. It seems, according to the published results for the different SMOTE-based flavors, that problem-oriented heuristics might be the best solution for each problem. An example of such heuristic can be grouping the TS samples for the minority class using the mean value of the Phan et al distance[18]; afterwards, two different groups are randomly chosen; finally, one TS sample is chosen form

 ISBN: 978-84-617-8694-7

Algorithm 1 The SMOTE original algorithm. Three parameters (T, N, k) are needed, as stated above.

N, k)

```
1: if N < 100 then
2:     Randomize the T minority class samples
3:     T = (N / 100) * T
4:     N = 100
5: end if
6: N = int( N / 100 )
7: numattrs = Number of attributes
8: Sample[][]: array for original minority class samples
9: newindex: counts the number of generated synthetic samples
10: Synthetic[][]= array for synthetic samples
11: for i = 1 : T do
12:     Compute the k nearest neighbors of sample i, saving the indexes in nnarray
13:     Populate(N, i, nnarray)
14: end for
15: function Populate(N, i, nnarray)
16:     while N ≠ 0 do
17:         Choose a random number nn in {1, k}
18:         for  attr = 1 : numattrs do
19:             dif = Sample[nnarray[nn]][attr]-Sample[i][attr]
20:             gap = random number in {0, 1}
21:             Synthetic[newindex][attr] = Sample[i][attr] + gap * dif
22:         end for
23:         newindex ++
24:         N = N - 1
25:     end while
26: end function
```

 ISBN: 978-84-617-8694-7

each of the two candidate groups. Nevertheless, this distance measurement suites the best when the length of the TS is bounded to less than, say, 30 values.

On the other hand, the generation of a new TS is not a simple task: as long as multivariate TS are considered, the new TS sample will need a TS for each of the available features. For each feature to generate, a combination of the parents' feature should be performed. Further, the combination must be coherent for all the features considered as a single sample. Finally, the class TS needs to be generated as well, which is much of a compromise. Again, general algorithms can be provided, but it should be expected that specific heuristic would eventually be needed in order to obtain a better performance.

In this study, the selection of the parents is performed using the same strategy than in the original SMOTE, using KNN with euclidean distance among the minority class samples. The euclidean distance between two multivariate TS X and Y (See equation 1) will be calculated on the average of all the features. In case the parents have different length, the shortest one will be considered.

$$KnnDistance_{X,Y} = \sqrt{\sum_{i=min\{\#X,\#Y\}} (\frac{\sum_{f \in features} X_i^f - Y_i^f}{\#features})^2} \quad (1)$$

The generation of a new TS sample is performed as follows, from now on, this proposal is referred as AVG_TS_SMOTE:

- For each feature, the average of the corresponding TS from the parents is computed.
- The class TS is calculated with the maximum of the values from the two parents.
- The length of the new TS sample, for every feature and for class TS, is bounded to the shortest of the two parents.

From now on, this new release of SMOTE will be referred as TS_SMOTE.

4 Experiments and results

4.1 Experimental setup

For this experimentation, a real world TS dataset obtained from the simulation of epileptic seizures is used; this dataset is publicly available at [3, 19]. This TS dataset was gathered following a previously defined and very strict protocol, defining a set of activities, namely, the simulation of the epileptic convulsions and three activities: running, sawing and walking -either gesturing while walking slowly or normal walking at different paces. A wearable triaxial accelerometer sensor (3DACM) included in a bracelet placed on the affected wrist measured the participant movements.

 ISBN: 978-84-617-8694-7

Transformation	**Calculation**
$SMA_t(\vec{s})$	$\frac{1}{w}\sum_{i=1}^{w-1}(\sum_{c\in\{x,y,z\}}\|b_{c,t-1}\|)$
$AoM_t(\vec{s})$	$\sum_{i=0}^{i=w-1}\sum_{c\in\{x,y,z\}}\|max(b_{c,t-i})$ $-min(b_{c,t-i})\|$
$TbP_t(\vec{s})$	Computed with the following algorithm: 1.- Find the sequences with value higher than mean+K*std within the window ($K = 0.9$) 2.- Keep the rising points from each of these sequences 3.- Measure the mean time between them

Table 1: The transformations of the components of the acceleration, where $b_{c,i}$ stands for the body acceleration.

The bracelets have wireless data sampling capabilities at a rate of 16 Hz, the 3DACM have a range of $2 \times g$. Up to 6 healthy participants, all of whom remained anonymous, successfully completed this experiment, each running 10 trials of each activity. The ages of the participants ranged from 22 to 47, with four participants of around 40 years old. One participant out of six was female, and the eldest was left-handed. An identification number was given to each Time Series (TS), including information fields on participant ID, the number of trials, the activity, etc.

The acceleration has been filtered and processed, becoming into a three variable TS dataset: the features are depicted in Table 1: the Signal-Magnitude Area (SMA), the Amount of Movement (AoM) and the Time between Peaks (TbP). The complete pre-processing have been described in [3].

This TS dataset, consisting on TS samples of three TS each -SMA, AoM and TbP- $\{\overline{TS}_s\}$, with the label for each activity $\{c_s\}$ and with the TS for each timestamp label $\{\overline{C}_s\}$, has been used in this experimentation. We denote this TS dataset as ORIG, while the TS dataset after applying AVG_TS_SMOTE is denoted as SMT.

To select the number of TS samples to introduce in the dataset we used the following criteria. In an imbalanced dataset, there exists $R = 3$ times more examples belonging to the MC class than to the mC class for the s data source. So, to balance the number of samples for both classes means injecting $(R1) \times |mC_s|$ new TS samples.

Finally, the α parameter was allowed to drift in the interval [-1.0, 1.0]. Although this is a rather wide interval, it was used to evaluate the robustness of the algorithm when the generation of the synthetic TS samples generate disperse samples. As stated in [8], this scenario highly penalizes the performance of balancing dataset algorithms; therefore, the conclusions can be extracted on adverse scenarios. The next experimentation focuses on

 ISBN: 978-84-617-8694-7

analyzing the correlation between each feature and the class for the ORIG and the SMT datasets.

4.1.1 Shifting process

As it was explained in section 4.1 a very strict protocol, with a high control of the timing in each stage, was carried out to gather the TS dataset, thus all the TSs corresponding to EPILEPSY class have the same timing (see figure 1.top). This is a problem when facing the same problem of generating synthetic TS in different domains, or even in the same domain, in different contexts -such as in everyday life- where the TS are not so similar, even totally different. So, in order to inject an erratic behavior in the datasets as in the daily live, a shifting procedure to induce more dispersion in the timing of the EPILEPSY class TS is proposed. This procedure consists of the following steps:

1. With a probability of **Pshift** (Pshift=50%) each Ts belonging to EPILEPSY class will be shifted.
2. Each chosen Ts will be shifted circularly **Sshift tics**to the right or to the left with a probability of 50% .
3. **Sshift** is the 75% of the number of NO_EPILEPSY class tics at beginning or at the end of the TS depending of the shifting direction, left or right respectively.

For the shake of example, a right shifted TS is shown in Figure 1bottom.

Once we have defined the shifting procedure, it will be carried out on the whole EPILEPSY class TS obtaining the Shifted Dataset. Thus we have two input datasets for our TS_SMOTE algorithm: the original dataset (ORIG) and the shifted dataset(SHIFTED), besides TS_SMOTE algorithm will be run with these two inputs obtaining the balanced datasets called ORIGSTM and SHIFTEDSTM respectively.

4.2 Correlation study

Two different measurements have been applied in this study in order to assess the relationship between the distribution of the ORIG and the SMT datasets, namely: the Pearson Correlation ($\rho_{X,Y}$, Eq. 2) coefficient and the Mutual Information ($MI(X,Y)$, Eq. 3); where cov is the covariance, σ_X is the standard deviation of X, $p(x)$ is the probability of the event x and $p(x,y)$ is the conditional probability of x given y.

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y} \tag{2}$$

$$MI(X,Y) = \sum_{y \in Y} \sum_{x \in X} p(x,y) \log \left(\frac{p(x,y)}{p(x)\, p(y)} \right) \tag{3}$$

 ISBN: 978-84-617-8694-7

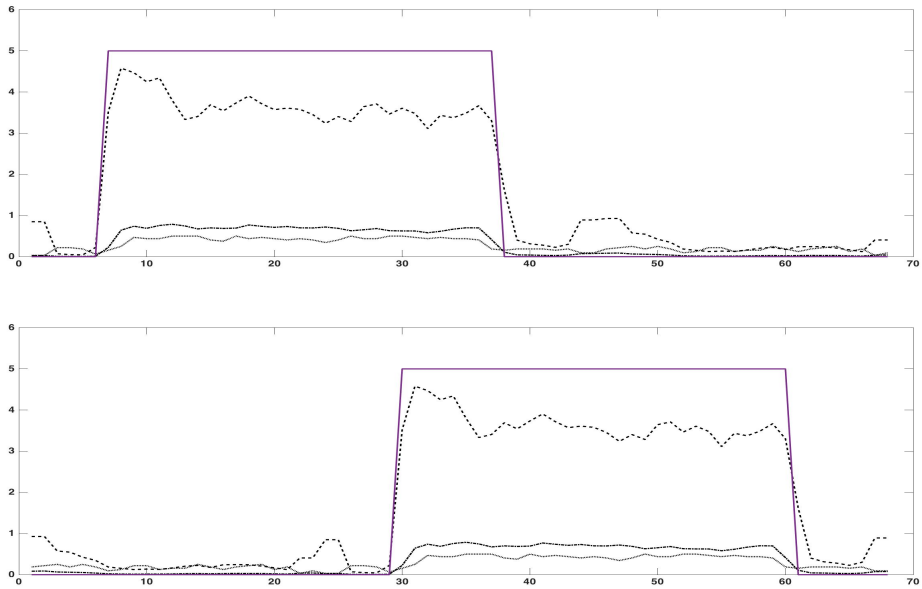


Figure 1: top) Case of EPILEPSY class TS where the solid line represents the class and the dashed, dotted and dashed-dotted lines represent the features SMA, AoM and TbP respectively, bottom) Right shifting of the top TS.

In order to analyze the robustness of the TS_SMOTE proposal, it's included a comparative study of correlation between the features and the class of each TS for the four datasets ORIG, SHIFTED, ORIGSMT and SHIFTEDSMT. For the sake of space restrictions only the detailed boxplot data for participant number 1 is depicted in Fig. 2. It can be noted that although the results for both flavors (shifted and not shifted datasets) are close similar it doesn't show the real differences between both results.

Let's see one the synthetic TS (Figure 3) from the SHIFTEDSMT dataset. It can be see that the synthetic TS (solid line) perverts totally at least one of the characteristics of the parent TSs (the dashed and the dotted lines), the length, which an important characteristic of a epilepsy attack.

Therefore, the table 2 includes a second correlation analysis, where every TS of not balanced dataset (ORIG or SHIFTED) will be compared with every TS of the corresponding balanced dataset (ORIGSMT or SHIFTEDSMT). It can be stated clearly that the balanced dataset obtained (SHIFTEDSMT) from the shifted dataset (SHIFTED) has a very low correlation with the respective not balanced dataset, with values of Pearson Correlation under 0.3-0.4 while the not shifted results are above 0.7-0.9.

Thus, we can conclude that the timing factor is a critical factor in the balancing process

 ISBN: 978-84-617-8694-7

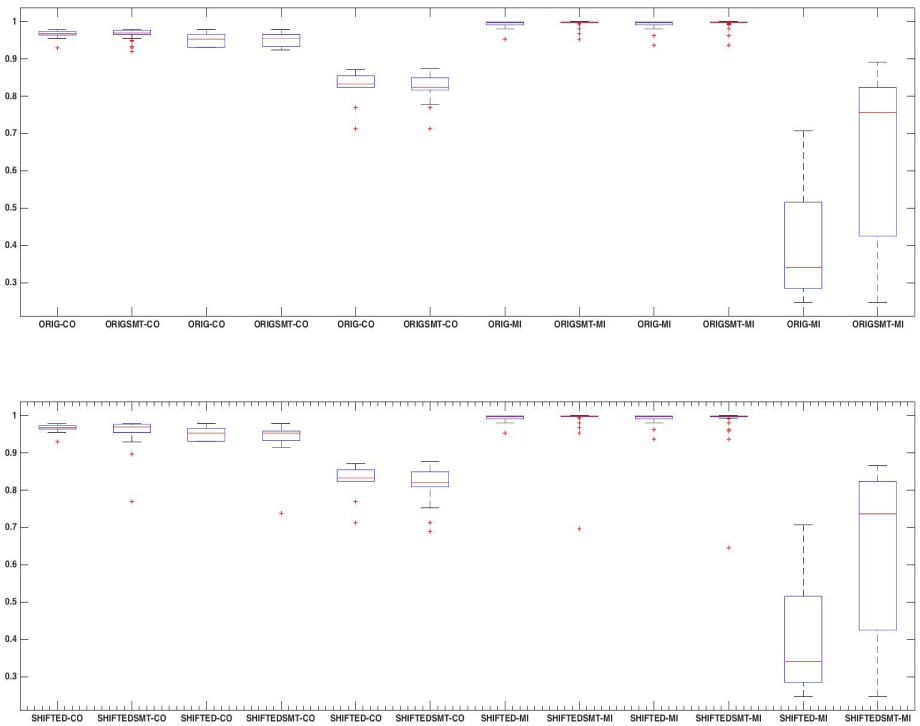


Figure 2: Boxplots of the relationship measurements between each feature and the class TS for the output of TS_SMOTE: ORIGSMT (top figure) and SHIFTEDSMT datasets (bottom figure). The 6 left-most boxes correspond to the Pearson Correlation, while the 6 right-most correspond to the Mutual Information.

of TS, so it's necessary to consider a kind of TS distances independent of shifting like DTW or OSB.

5 Conclusions

This research, presents a simple extension of the datasets balancing SMOTE algorithm but adapted to TS (TS_SMOTE). Our proposal is based on the following points: a) the inclusion of KNN algorithm for multivariate TS to select the parents of the synthetic TSs, where the distance between TS is based on a combination of the features; b) the definition of the operator AVG_TS_SMOTE to obtain the synthetic TSs.

Besides a shifting process is presented in order to inject an erratic behavior in the

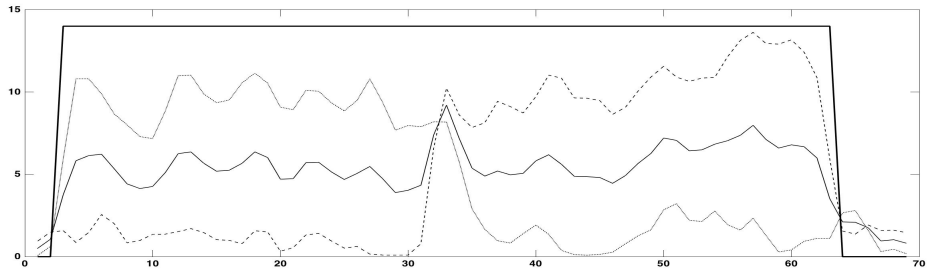


Figure 3: Example of SHIFTEDSMT Synthetic TS. From top to bottom, the two parents TS -the dashed and the dotted lines- and the synthetic TS -the solid line-.

datasets as in the daily live. So, the results of TS_SMOTE have been run on both unbalanced datasets (the original TS dataset and the shifted TS dataset) and we can conclude that timing is a critical factor in the balancing process of TS, so it's necessary to consider a kind of TS distances independent of shifting like DTW or OSB [20, 21].

6 Acknowledgment

This research has been funded by the Spanish Ministry of Science and Innovation, under project MINECO-TIN2014-56967-R.

References

[1] Beniczky, S., Polster, T., Kjaer, T., Hjalgrim, H.: Detection of generalized tonic-clonic seizures by a wireless wrist accelerometer: a prospective, multicenter study. Epilepsia **4**(54) (2013) e58–61

[2] Villar, J.R., González, S., Sedano, J., Chira, C., Trejo-Gabriel-Galán, J.M.: Improving human activity recognition and its application in early stroke diagnosis. International Journal of Neural Systems **25**(4) (2015) 1450036–1450055

[3] Villar, J.R., Vergara, P., Menéndez, M., de la Cal, E., González, V.M., Sedano, J.: Generalized models for the classification of abnormal movements in daily life and its applicability to epilepsy convulsion recognition. accepted for publication, International Journal of Neural Systems (2016)

[4] Villar, J.R., Menéndez, M., de la Cal, E., González, V.M., Sedano, J.: Identification of abnormal movements with 3d accelerometer sensors for its application to seizure recognition. accepted for publication, International Journal of Applied Logic (2016)

 ISBN: 978-84-617-8694-7

	Mean/std					
	$\rho_{X,Y}$					
	SMA		AoM		TbP	
s	**ORIG**	**SMT**	**ORIG**	**SMT**	**ORIG**	**SMT**
1	0.92/0.06	0.95/0.05	0.92/0.07	0.92/0.07	0.84/0.11	0.84/0.11
2	0.85/0.11	0.92/0.08	0.90/0.09	0.90/0.09	0.84/0.10	0.84/0.10
3	0.87/0.08	0.91/0.07	0.84/0.11	0.84/0.11	0.72/0.13	0.72/0.13
4	0.98/0.01	0.99/0.01	0.98/0.02	0.98/0.02	0.86/0.07	0.86/0.07
5	0.77/0.24	0.84/0.18	0.82/0.17	0.82/0.17	0.74/0.18	0.74/0.18
6	0.95/0.06	0.98/0.04	0.97/0.05	0.97/0.05	0.92/0.05	0.92/0.05
	$\rho_{X,Y}$					
	SMA		AoM		TbP	
s	**SHIFTED**	**SHIFTEDSMT**	**SHIFTED**	**SHIFTEDSMT**	**SHIFTED**	**SHIFTEDSMT**
1	0.44/0.66	0.60/0.60	0.58/0.57	0.58/0.57	0.53/0.53	0.53/0.53
2	0.16/0.68	0.42/0.61	0.41/0.61	0.41/0.61	0.38/0.53	0.38/0.53
3	0.21/0.66	0.38/0.63	0.35/0.59	0.35/0.59	0.32/0.49	0.32/0.49
4	0.20/0.67	0.38/0.61	0.37/0.60	0.37/0.60	0.34/0.54	0.34/0.54
5	0.17/0.59	0.33/0.58	0.34/0.55	0.34/0.55	0.25/0.51	0.25/0.51
6	0.21/0.70	0.40/0.64	0.39/0.63	0.39/0.63	0.40/0.55	0.40/0.55

Table 2: Upper part shows the Pearson Correlation results between all the pairs of time series from ORIG and ORIGSMT datasets. The lower part shows the correlation between the SHIFTED and SHIFTEDSMT datasets. The left-most column s refers to the participant id. Each cell contains the mean and the standard statistics for the calculated values on the completed TS dataset.

[5] López, V., Fernández, A., del Jesus, M., Herrera, F.: A hierarchical genetic fuzzy system based on genetic programming for addressing classification with highly imbalanced and borderline data-sets. Knowledge-Based Systems **38** (2013) 85–104

[6] Galar, M., Fernández, A., Barrenechea, E., Herrera, F.: Eusboost: Enhancing ensembles for highly imbalanced data-sets by evolutionary undersampling. Pattern Recognition **46**(12) (2013) 3460–3471

[7] Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: Smote: synthetic minority over-sampling technique. Journal of artificial intelligence research (2002) 321–357

[8] Batista, G., Prati, R., Monard, M.: A study of the behavior of several methods for balancing machine learning training data. SIGKDD Explorations (2004) 20–29

[9] He, H., Bai, Y., Garcia, E., Li, S., et al.: Adasyn: Adaptive synthetic sampling approach for imbalanced learning. In: Neural Networks, 2008. IJCNN 2008.(IEEE World Congress on Computational Intelligence). IEEE International Joint Conference on, IEEE (2008) 1322–1328

[10] Tang, S., Chen, S.: The generation mechanism of synthetic minority class examples. In: Proceedings of 5th International Conference on Information Technology and Applications in Biomedicine (ITAB 2008). (2008) 444–447

[11] Stefanowski, J., Wilk, S.: Selective pre-processing of imbalanced data for improving classification performance. In: Proceedings of the 10th International Conference

in Data Warehousing and Knowledge Discovery (DaWaK2008). Volume LNCS 5182., Springer (2008) 283–292

[12] Fu, T.c.: A review on time series data mining. Engineering Applications of Artificial Intelligence **24**(1) (2011) 164–181

[13] Mishra, S., Saravanan, C., Dwivedi, V., Pathak, K.: Discovering flood rising pattern in hydrological time series data mining during the pre monsoon period. Indian Journal of Marine Sciences **44**(3) (2015) 3

[14] Montgomery, D.C., Jennings, C.L., Kulahci, M.: Introduction to time series analysis and forecasting. John Wiley & Sons (2015)

[15] Moses, D., et al.: A survey of data mining algorithms used in cardiovascular disease diagnosis from multi-lead ecg data. Kuwait Journal of Science **42**(2) (2015)

[16] Köknar-Tezel, S., Latecki, L.J.: Improving svm classification on imbalanced time series data sets with ghost points. Knowledge and information systems **28**(1) (2011) 1–23

[17] de la Cal, E., Villar, J.R., Vergara, P., Sedano, J., Herrero, Á.: A smote extension for balancing multivariate time series datasets. In: accepted for Proceedings of the International Conference on Soft Computing Models in Industrial and Environmental Applications, Springer (2017)

[18] Phan, S., Famili, F., Tang, Z., Pan, Y., Liu, Z., Ouyang, J., Lenferink, A., Oconnor, M.M.C.: A novel pattern based clustering methodology for time-series microarray data. International Journal of Computer Mathematics **84**(5) (2007) 585–597

[19] Villar, J.R.: Jrv's research home page

[20] Berndt, D.J., Clifford, J.: Using dynamic time warping to find patterns in time series. In: Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining. AAAIWS'94, AAAI Press (1994) 359–370

[21] Latecki, L.J., Wang, Q., Koknar-Tezel, S., Megalooikonomou, V.: Optimal subsequence bijection. In: Data Mining, 2007. ICDM 2007. Seventh IEEE International Conference on, IEEE (2007) 565–570

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

Tricky Aspects of Kronecker Power Series in Constancy Adding Space Extention (CASE) Perspective

Metin Demiralp[1]

[1] *Computational Science and Engineering Department, Informatics Institute, İstanbul Technical University*

emails: metin.demiralp@gmail.com

Abstract

Kronecker power series is one of the fundamental elements of the Probabilistic Evolution Theory (PREVTH) developed in Demiralp group studies during the last decade. This is a concise representation of multivariate Taylor series where the independent variable is just a vector instead of many scalar independent variables and based on Kronecker powers. In contrast to Taylor series it is not unique because the greater than one Kronecker powers of a vector becomes orthogonal to certain constant vectors such that their population rapidly grows as the power increases. This property brings a lot of interesting flexibilities to change the structure of the series without changing the target function the Kronecker power series represents. These can be used at our favor to facilitate the analyses especially through the Constancy Adding Space Extension (CASE). This work focuses on these issues and basically on converting a Kronecker power multinomial to its highest power monomial which gains great importance to get a new extension to PREVTH.

Key words: Kronecker product, Kronecker power, Kronecker power series, Kronecker multinomial, Kronecker monomial, PREVTH.

1 Introduction: Kronecker Products, Kronecker Powers, Kronecker Power Series

Kronecker power series concept is somehow backbone of the Probabilistic Evolution Theory (PREVTH) which has been recently developed in Demiralp group studies, [1–16] for constructing analytic solutions to the first order ODEs under initial condition impositions. For better understanding of these series we need to start with Kronecker product of two

ordinary linear algebraic arrays (vectors, matrices). It is defined in a way such that each element of first (left) array is replaced by the product of second (right) array with that element. If we denote the factor arrays by $\boldsymbol{\mathcal{A}}_1$ $(m_1 \times n_1)$ and $\boldsymbol{\mathcal{A}}_2$ $(m_2 \times n_2)$ respectively then the Kronecker power can be symbolized as $\boldsymbol{\mathcal{A}}_1 \otimes \boldsymbol{\mathcal{A}}_2$ and its type is $m_1 m_2 \times n_1 n_2$. To be more explicit we can write the following equality for two given linear algebraic vectors, $\mathbf{a}$ and $\mathbf{b}$ with m_a and m_b elements respectively

$$\mathbf{a} \otimes \mathbf{b} \equiv \left[a_1 \mathbf{b}^T \;\; ... \;\; a_{m_a} \mathbf{b}^T \right]^T \tag{1}$$

where the product is a linear algebraic vector of $m_a m_b$ elements. This explicit definition can be extended to vector-matrix, matrix-vector, matrix-matrix products accordingly even though we are not going to get into further details. Kronecker power is not commutative generally for different factors. Throughout this presentation we use lower and upper letter bold symbols for ordinary linear algebraic vectors and matrices respectively. When we need to use just a single symbol for an entity which can be either vector or matrix depending on the situation we use bold calligraphic upper letters.

Now we can define the Kronecker power of an array (vector or matrix) as sufficient number of consecutive Kronecker products of that array. If we denote the array whose mth Kronecker power is under consideration by $\boldsymbol{\mathcal{A}}$ then we can write

$$\boldsymbol{\mathcal{A}}^{\otimes m} \equiv \underbrace{\boldsymbol{\mathcal{A}} \otimes \cdots \otimes \boldsymbol{\mathcal{A}}}_{m \text{ factors}}, \qquad m = 1, 2, ...; \qquad \boldsymbol{\mathcal{A}}^{\otimes 0} \equiv 1 \tag{2}$$

where we have used the algebraic convention dictating us that the zero power of something is just the constant 1.

Kronecker powers enable us to define the powers of ordinary algebraic vectors. This property can be used to much more concisely rewrite a multivariate Taylor series in a single infinite sum. If we consider a multivariate function f which depends on N number of independent variables then we can define the following vector

$$\mathbf{x} \equiv \left[x_1 - x_1^{(e)} \;\; ... \;\; x_N - x_N^{(e)} \right]^T \tag{3}$$

which can be called "System Vector" in system theoretical applications. Then the following infinite sum corresponds to multivariate Taylor series expanded at the position denoted by the vector whose elements $x_1^{(e)}$,..., $x_N^{(e)}$ respectively ((e) stands for recalling expansion) for the function $f(\mathbf{x})$.

$$f(\mathbf{x}) = \sum_{j=0}^{\infty} \mathbf{f}_j^T \mathbf{x}^{\otimes j} \tag{4}$$

where $\mathbf{f}_j$ stands for an N^j element constant vector whose elements can be evaluated from the f function's appropriate partial derivatives evaluated at $\mathbf{x} = \mathbf{0}_N$.

 ISBN: 978-84-617-8694-7

First two coefficient vectors, 1×1 type (scalar) f_0 and $N \times 1$ type f_1 vectors, are unique in this representation. However, all other f_j vectors are not unique because the greater than or equal to 2 Kronecker powers of the independent variable vector $\mathbf{x}$ is orthogonal to some finite number constant matrices. To explain this situation we can consider the Kronecker square of the vector $\mathbf{x}$. The vector $\mathbf{x}^{\otimes 2}$ has elements which are binary products like $a_i a_j$ $(i \neq j)$. The elements $a_i a_j$ and $a_j a_i$ are same even though their positions in element ordering are $(i-1)N + j$ and $(j-1)N + i$ respectively. So the vector whose elements are all zero except the ones at these locations have same magnitude but opposite signs is orthogonal to $\mathbf{x}^{\otimes 2}$. This implies that there are $N(N-1)/2$ constant vectors to which the vector $\mathbf{x}^{\otimes 2}$ is orthogonal. Hence the addition of the linear combination of these constant vector transposes with arbitrary linear combination coefficients does not change the contribution of the term proportional to f_2 to whole sum. Similar situations exist for other Kronecker powers greater than 2 such that the number of the constant vectors rapidly grows unboundedly as the power increases. All these discussions mean that Kronecker power series is not a unique entity and many appropriately defined constant vectors with arbitrary magnitudes can be added to the coefficients without changing the entire expansion.

2 Constancy Adding Space Extension (CASE)

The nonuniqueness of Kronecker power series in fact presents an important facilitation in the restructuring of Kronecker power series since certain arbitrary parameters can be introduced to the structure such that we can monitor the series coefficients by choosing appropriate values to these parameters. On the other hand, another important concept, "Space Extension" may take us to much better structures in Kronecker power series. Space extension means the definition of new independent variables in terms of the given ones to extend the space where the Kronecker power series basic vector $\mathbf{x}$ lies in such a way that the original Kronecker power series becomes having coefficients which are close to what we desire to have.

"Constancy Adding Space Extension" takes an important specific part amongst many other space extensions. It basically adds a constant function with an arbitrary value, as if a new independent variable. Let us now define an augmented independent variable vector from the independent variable of the Kronecker power series in (4) as follows

$$\mathbf{x}_{aug} \equiv \begin{bmatrix} \mathbf{x}^T & x_{n+1} - x_{n+1}^{(e)} \end{bmatrix}^T, \qquad x_{n+1} \equiv a \tag{5}$$

where we have assumed that the augmented expansion point component, $x_{n+1}^{(e)}$ vanishes and a is arbitrary at this moment. We can write the following Kronecker power series instead

 ISBN: 978-84-617-8694-7

of the one in (4)

$$f(\mathbf{x}) = \sum_{j=0}^{\infty} \mathbf{f}_{aug,j}^{T} \mathbf{x}_{aug}^{\otimes j} \tag{6}$$

where we need to express $\mathbf{f}_{aug,j}$ vectors in terms of $\mathbf{f}_j$ vectors. To accomplish this task we focus on the Kronecker powers $\mathbf{x}_{aug}^{\otimes j}$s. $\mathbf{x}_{aug}^{\otimes j}$ is a jth degree polynomial of a as can be noticed immediately even though this dependence may be considered quite complicated. A careful look at this vector reveals that each element is proportional to a nonengative integer power of a even though there seems to be existing rather a disordered ordering of these powers. However it is always possible that we can use permutations amongst the elements of this vector such that the new formed vector has vector blocks located as descending Kronecker powers of the vector $\mathbf{x}$ starting from the power j. Each block should be multiplied by a power of a such that the sum of this power with the Kronecker power of that term always remains equal to j. There is only one block containing jth Kronecker power of $\mathbf{x}$ which has no prefactor of an a power while the number of blocks proportional to the $(j-1)$th Kronecker power of $\mathbf{x}$ is j and they are multiplied by just a. The number of the other terms are relevant binomial coefficients and the degree of the a factors increases by one as the Kronecker power decreases by one. Therefore we can write

$$\mathbf{x}_{aug}^{\otimes j} \equiv \mathbf{\Pi}_j \left[\mathbf{x}^{\otimes j^T} \quad \underbrace{a\mathbf{x}^{\otimes (j-1)^T} \quad ... \quad a\mathbf{x}^{\otimes (j-1)^T}}_{j \text{ identical terms}} \quad \underbrace{a^2\mathbf{x}^{\otimes (j-2)^T} \quad ... \quad a^2\mathbf{x}^{\otimes (j-2)^T}}_{j(j-1)/2 \text{ identical terms}} \quad ... \quad a^j \right]^T \tag{7}$$

where $\mathbf{\Pi}_j$ is a permutation matrix of $(N+1)^j \times (N+1)^j$ type. Due to the specific nature of the permutation operation, this permutation matrix is unitary. We do not intend to explicitly give its analytic structure.

(7) implies that the relation $\mathbf{f}_{aug,j} = \mathbf{f}_{zpad,j}\mathbf{\Pi}^T$ can be constructed between the $\mathbf{f}_j$ and $\mathbf{f}_{aug,j}$ vectors where $\mathbf{f}_{zpad,j}$ stands for the $(N+1)^j$ element vector whose first N^j element block is equal to $\mathbf{f}_j$ while its remaining part is padded with zero.

3 Monomial Shifting In Constancy Adding Space Extension (CASE)

Abovementioned formulation shows that the use of CASE may not be considered so interesting unless certain extra actions are realized to facilitate the Kronecker Power Series utilization under consideration. To this end we can start with the following equalities: $1 \equiv a/a \equiv (1/a)\mathbf{e}_{N+1}^T \mathbf{x}_{aug},\ a \neq 0$ where $\mathbf{e}_{N+1}$ stands for the standard unit vector whose only nonzero element is 1 and located at the $(N+1)$th position, in $(N+1)$ dimensional

 ISBN: 978-84-617-8694-7

Cartesian space. The following extended form of this equality is much more facilitating for the term shifts in Kronecker power series.

$$1 \equiv \frac{a^m}{a^m} \equiv \frac{1}{a^m}\left(\mathbf{e}_{N+1}^T \mathbf{x}_{aug}\right)^m \equiv \frac{1}{a^m}\mathbf{e}_{N+1}^{\otimes m\ T}\mathbf{x}_{aug}^{\otimes m}, \qquad a \neq 0, \qquad m = 0, 1, 2, ... \tag{8}$$

This leads us to write

$$\begin{aligned}\mathbf{f}_{aug,j}^T \mathbf{x}_{aug}^{\otimes j} = 1\mathbf{f}_{aug,j}^T \mathbf{x}_{aug}^{\otimes j} 1 &= \frac{1}{a^m}\left(\mathbf{e}_{N+1}^{\otimes m_1\ T}\mathbf{x}_{aug}^{\otimes m_1}\right)\left(\mathbf{f}_{aug,j}^T \mathbf{x}_{aug}^{\otimes j}\right)\left(\mathbf{e}_{N+1}^{\otimes (m-m_1)^T}\mathbf{x}_{aug}^{\otimes (m-m_1)}\right) \\ &= \left(\mathbf{e}_{N+1}^{\otimes m_1} \otimes \mathbf{f}_{aug,j} \otimes \mathbf{e}_{N+1}^{\otimes (m-m_1)}\right)\mathbf{x}_{aug}^{\otimes (j+m)} = \mathbf{f}_{aug,j\to m_1+j+m-m_1}^T \mathbf{x}_{aug}^{\otimes (j+m)}\end{aligned} \tag{9}$$

where the $(N+1)^{j+m}$ type new coefficient vector can be given through the following identicality.

$$\mathbf{f}_{aug,j\to m_1+j+m-m_1} \equiv \mathbf{e}_{N+1}^{\otimes m_1} \otimes \mathbf{f}_{aug,j} \otimes \mathbf{e}_{N+1}^{\otimes (m-m_1)} \tag{10}$$

where the subscript component $j \to m_1 + j + m - m_1$ explains the shift from jth term to $(j+m)$th term in a way such that the total shift is partitioned to three elementary shifts: (i) first from constant to m_1 Kronecker power, (ii) then the next j step shift in power ascending direction, (iii) and finally, $(m - m_1)$ step in power ascending direction. This implies that there are $(m+1)$ number of three partitioned consecutive shifts from jth power to $(j+m)$th power.

Now, we can write the following equality from (9)

$$\begin{aligned}\mathbf{f}_{aug,j}^T \mathbf{x}_{aug}^{\otimes j} = \left(\sum_{m_1=0}^{m} \alpha_{m_1} \mathbf{f}_{aug,j\to m_1+j+m-m_1}\right)^T \mathbf{x}_{aug}^{\otimes (j+m)} &= \mathbf{f}_{aug,j\to j+m}\left(\boldsymbol{\alpha}\right)^T \mathbf{x}_{aug}^{\otimes j}, \\ \alpha_0 + \cdots + \alpha_m &= 1\end{aligned} \tag{11}$$

where $\boldsymbol{\alpha}$ stands for the set of the α parameters while the rightmost part coefficient is a concise notation not explicitly representing the partitioning and the coefficient of the $(j+m)$ Kronecker power of $\mathbf{x}_{aug}$ therein can be called the "Coefficient to m-Power-Shift". This means that m-power-shift introduces m arbitrary scalars as flexible parameters (if m vanishes then there is no inserted flexible parameter; however this means no shift and it is a trivial issue). These parameters can be determined to suppress certain array norms to get wider convergence domains.

The above analysis imposes no condition on the structure of the vector $\mathbf{f}_j$ and inserts a lot of arbitrary parameters to that vector during the m-power-shift such that the number of the parameters ascends up to infinity as j grows unboundedly. However, this number can be much more increased if the vector $\mathbf{f}_j$ has certain specific natures. To this end, first thing coming to mind, is the binary Kronecker product decomposition and we can write

$$\mathbf{f}_{aug,j} = \mathbf{f}_{aug,j_1}^{(1)} \otimes \mathbf{f}_{aug,j-j_1}^{(2)}, \qquad \mathbf{f}_{aug,j}\mathbf{x}^{\otimes j} = \left(\mathbf{f}_{aug,j_1}^{(1)}\mathbf{x}^{\otimes j_1}\right)\left(\mathbf{f}_{aug,j-j_1}^{(2)}\mathbf{x}^{\otimes (j-j_1)}\right) \tag{12}$$

 ISBN: 978-84-617-8694-7

where j_1 can take just a single or more-than-one values between 0 and j exclusive and each factor between parantheses can be parametrized by using m-power-shift formula given above. Since the total power shift will be considered as m, each paranthesed factor can be power shifted such that the total power shift remains as m-power-shift.

By using (11) and distributive property of matrix product over Kronecker product we can write

$$\begin{aligned}\mathbf{f}_j^T \mathbf{x}_{aug}^{\otimes j} &= \sum_{m_1=0}^{m} \gamma_{m_1} \left(\mathbf{f}_{aug,j_1\to j_1+m_1}\left(\boldsymbol{\alpha}_{m_1}\right) \otimes \mathbf{f}_{aug,(j-j_1)\to(j-j_1)+(m-m_1)}\left(\boldsymbol{\beta}_{m-m_1}\right)\right)^T \mathbf{x}_{aug}^{\otimes(j+m)} \\ &= \mathbf{f}_{j\to j+m;KPD}\left(\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\gamma}\right)\mathbf{x}^{\otimes j+m} \end{aligned} \tag{13}$$

where subscript component KPD after semicolon stands for implying the Kronecker product decomposition of the vector $\mathbf{f}_j$ at two factor (binary) product level while the symbols $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ have been used to denote the unions of the sets over $\boldsymbol{\alpha}_{m_1}$s and the unions of the sets over $\boldsymbol{\beta}_{m_1}$s respectively whereas $\boldsymbol{\gamma}$ denotes the set composed of γs.

The set $\boldsymbol{\alpha}_{m_1}$ has (m_1+1) elements denoted by $\alpha_{m_1,0}$,...,α_{m_1,m_1} such that the sum of these elements separately vanish for all m_1 values between 1 and m inclusive as long as m does not vanish (otherwise trivial). Same thing is also valid for the set $\boldsymbol{\beta}_{m-m_1}$ whose elements can be denoted by $\beta_{m-m_1,0}$, ..., $\beta_{m-m_1,m-m_1}$ such that the sum of these elements separately vanish for each m_1 values between 0 and $m-1$ inclusive as long as m does not vanish (otherwise trivial). Hence, for a specific m_1 value between 0 and m, $\boldsymbol{\alpha}_{m_1}$ brings m_1 flexible parameters while the number of flexible parameters brought by the set $\boldsymbol{\beta}_{m-m_1}$ is $(m-m_1)$. Thus the number of flexible parameters entering the shift for a single given m_1 value from αs and βs is always m. This makes the total number of flexible parameters coming from αs and βs is $m(m+1)$. Since we have also used γs with indices between 0 and m inclusive as flexible parameters, the total number of flexible parameters is $m(m+2)$ because of the vanishing sum over γs. This means that the number of flexible parameters in (13) can be represented by a second degree polynomial of m. This proves the rapid increase in the parametrization when the vector $\mathbf{f}_j$ can be expressed as a Kronecker product of two appropriate subvectors. Here all these are true for just a single binary Kronecker product case. If there are more than one cases each of which leads us to a separate binary Kronecker power for the vector $\mathbf{f}_j$ then the number of the flexible parameters grows much more rapidly.

The latest analysis here in this section can be extended to the case where the vector $\mathbf{f}_j$ can be expressed as more than two factor Kronecker products. In those cases the number of the flexible parameters increases from second degree polynomials of m to much more higher polynomials of m even though we do not intend to proceed for giving much more informations for these cases since they are not primary items for this work.

In many cases the vector $\mathbf{f}_j$ may not be uniquely (but approximately) represented as the Kronecker products of some number of factors. In those cases, certain more sophisticated schemes, for example singular value decomposition like or enhanced multivariate products

 ISBN: 978-84-617-8694-7

representation like algorithms, to represent this vector can be brought to the stage. However, these algorithms alone may necessitate quite rigorous studies and we find them too detailed for this proceeding paper even though we have current research projects on these items.

The higher number of flexible parameters which can be inserted to Kronecker power terms during the shifts in Kronecker powers is quite important since their optimisation can be used to suppress the norm of the coefficient vector of shifted Kronecker power term. This suppression generally leads us to get wider convergence domain in the resulting-after-shift Kronecker power series and this is very important issue for practical applications. Hence high level parametrization of the m-power-shifts in Kronecker power series is perhaps most desired issue. We find this discussion sufficient for our goal in this work.

4 Converting Multinomiality to Highest Degre Monomiality via Power Shifts

Kronecker power series are somehow is the backbone of the Probabilistic Evolution Theory (PREVTH) as we pronounced at the beginning of this paper. PREVTH extensively uses the concept of space extension as unfamiliar reader can find sufficient information in our publications on this issue [1–16] and many practical cases have possibilities to get multinomiality (multinomial=polynomial of more than one independent variables) at the right hand side functions of the explicit ODEs in PREVTH. Until now we have also attempted to further reduce multinomiality to conicality (the second degree multinomiality) to get the analytical solution for PREVTH. Our recent considerations have shown that the Highest Degree Monomiality (just a single Kronecker power with highest degree together with its coefficient matrix or vector) can also facilitate the PREVTH analysis and may yield analytical solutions as again appropriate Kronecker power series.

Hence to proceed we can focus on the Kronecker power multinomial defined as follows

$$f\left(\mathbf{x}\right) = \sum_{j=0}^{N} \mathbf{f}_{aug,j}^{T} \mathbf{x}_{aug}^{\otimes j} \tag{14}$$

which is a restricted form of the Kronecker power series given in (6). Our purpose is now to use m-power-shifts to reduce this multinomial to a single power multinomial such that only the highest Kronecker power appears. To this end we can start with the following new form of (13)

$$\mathbf{f}_{j}^{T} \mathbf{x}_{aug}^{\otimes j} = \mathbf{f}_{j \rightarrow j+m;Dec}\left(\boldsymbol{\kappa}_{m}\right) \mathbf{x}^{\otimes j+m} \tag{15}$$

where the separately symbolized sets $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, and, $\boldsymbol{\gamma}$ have been combined to a single set denoted by κ_m (m specifies the shift) for brevity here. Beyond that the subscript string

 ISBN: 978-84-617-8694-7

KPD, which has been used to mean a very specific situation, can be interchanged with another string Dec to mean a more general vector decomposition.

We can now change m with $N-j$ in (15) to get the following more specific shift

$$\mathbf{f}_j^T \mathbf{x}_{aug}^{\otimes j} = \mathbf{f}_{j\to N;Dec}\left(\boldsymbol{\kappa}_{N-j}\right) \mathbf{x}^{\otimes N} \tag{16}$$

whose utilization in (14) produces

$$f\left(\mathbf{x}\right) = \left(\sum_{j=0}^{N} \mathbf{f}_{j\to N;Dec}\left(\boldsymbol{\kappa}_{N-j}\right) \right) \mathbf{x}^{\otimes N} \tag{17}$$

whose right hand side is apparently highest degree monomial of the original multinomial. This is a quite important relation and stands for opening new horizons to PREVTH. Here the number of elements in the set $\boldsymbol{\kappa}_{N-j}$ may be quite nonlinear in comparison to the term $(N-j)$ and therefore the total level of parametrization may be quite high and nonlinear in $(N-j)$ depending on the decomposition used in the parametrization. Even though this is an important issue in the sense of combinatorial analysis, we can consider this term as a secondary issue in this conceptual work. However, for further practicality this issue is quite important and we have launched a new peoject to reveal important aspects of this topic.

5 Concluding Remarks

The main goal of this work has been the conversion of a Kronecker multinomial to its highest degree monomial with a different coefficient. To this end we have basically used the Constancy Adding Space Extension (CASE) which principally adds a constant function with an unknown value to the existing unknowns such that the added constant function is considered as if it is a new unknown. This addition of course affects the existing coefficients of a Kronecker power series. However the most important aspect of CASE is the possibility of moving certain Kronecker powers to a higher Kronecker power and add to its existing value. This can be successed in such a way that the resulting affected higher Kronecker power coefficient includes some number of arbitrary parameters which can be used to suppress the relevant coefficient norms and therefore to obtain much wider convergence domain in the relevant Kronecker power series. The parametrization level (the number of the flexible parameters to be optimized) depends on the nature of the coefficients entering the shifts from certain Kronecker powers to higher degree Kronecker powers.

Kronecker power series appear comprehensively in PREVTH and are not unique expansions. This nonuniqueness shows up as flexible parameter insertion in Constancy Adding Space Extension (CASE). However, in PREVTH, the infinite Kronecker power series do not help us so much to facilitate the analysis and to get an analytical solution. On the other hand, in many practical applications the right hand sides of the explicit ODEs to

 ISBN: 978-84-617-8694-7

which PREVTH will be applied there exist certain functions of the unknown temporally changing functions such that the use of those functions as new unknowns with or without existing ones produces new extended set of ODEs having right hand sides in multinomials of the new unknowns. Once the multinomiality is obtained at the right hand side of the considered ODEs the rest is quite straightforward and the multinomiality can be converted to conicality as certain theorems guarantees. Even though the present widely used form of PREVTH uses conicality our brainstorming actions implied that highest degree monomial structured right hand sides even facilitate the PREVTH analysis pretty much and the power shifts mentioned in this paper permits us to convert a given Kronecker power multinomial to its highest degree monomial with a new coefficient and this brings another analytical solution having PREVTH version to applications.

References

[1] M. Demiralp, *Quantum Expected Value Dynamics in Probabilistic Evolution Perspective*, Proceedings of the 12th International Conference on Computational and Mathematical Methods in Science and Engineering (CMMSE), ISBN: 978-84-615-5392-1, Murcia, Spain (2012), 449-459.

[2] M. Demiralp, Journal of Mathematical Chemistry, URL doi:http://dx.doi.org/10.1007/s10910-012-0079-6, **51**(4), (2012) 1170.

[3] M. Demiralp and B. Tunga, Journal of Mathematical Chemistry, URL doi:http://dx.doi.org/10.1007/s10910-012-0081-z, **51**(4), (2012), 1198.

[4] E. Demiralp, M. Demiralp and L. Hernandez-Garcia, *Journal of Mathematical Chemistry*, URL doi:http://dx.doi.org/10.1007/s10910011-9929-x, **50**, (2012) 850.

[5] M. Demiralp and E. Demiralp, Journal of Mathematical Chemistry, URL doi:http://dx.doi.org/10.1007/s10910-012-0070-2 **51**(1) (2012) 58.

[6] M. Demiralp and E. Demiralp, Journal of Mathematical Chemistry, URL doi:http://dx.doi.org/10.1007/s10910-012-0064-0 **51**(19) (2012) 38.

[7] S. Tuna and M. Demiralp, *Certain Validations of Probabilistic Evolution Approach for Initial Value Problems*, Proceedings of the 12th WSEAS International Conference on Systems Theory and Scientific Computation (ISTASC'12), ISBN: 978-1-61804-115-9, İstanbul, Türkiye (2012) 246-249.

[8] N. A. Baykara, E. Gürvit and M. Demiralp, *Univariate single quantum harmonic oscillator from probabilistic evolution perspective*, Proceedings of the 13th WSEAS

International Conference on Mathematics and Computers in Biology and Chemistry (MCBC'12), Wisconsin, ABD (2012) 27-32.

[9] M. Ayvaz and M. Demiralp, *Getting Triangularity and Conicality in the Probabilistic Evolutionary Expectation Dynamics of the Purely Quartic Quantum Anharmonic Oscillator*, Proceedings of the 12th WSEAS International Conference on Systems Theory and Scientific Computation (ISTASC'12), ISBN: 978-1-61804-115-9, İstanbul, Türkiye (2012) 268-271.

[10] E. Demiralp, M. Demiralp and L. Hernandez-Garcia, *Journal of Mathematical Chemistry*, URL doi:http://dx.doi.org/10.1007/s10910-011-9930-4, **50**, (2012) 870.

[11] B. Tunga and M. Demiralp, *Probabilistic Evolution of the State Variable Expected Values in Liouville Equation Perspective, for a Many Particle System Interacting Via Elastic Forces*, Proceedings of the 12th International Conference on Computational and Mathematical Methods in Science and Engineering (ICCMSE), ISBN: 978-84-615-5392-1, Murcia, Spain (2012) 1186-1197.

[12] M. Demiralp, *Squarificating the Telescope Matrix Images of Initial Value Vector in Probabilistic Evolution Theory (PET)*, Proceedings of the 19th International Conference on Applied Mathematics (AMATH'14), ISBN: 978-1-61804-258-3, İstanbul, Türkiye (2014) 99104.

[13] M. E. Kirkin and C. Gözükirmizi, *Probabilistic Evolution Theory for ODE Sets with Second Degree Multinomial Right Hand Side Functions: Certain Reductive Cases*, ICCMSE, Athens, Greece (2015).

[14] C. Gözükirmizi, *Probabilistic Evolution Theory for ODE Sets with Second Degree Multinomial Right Hand Side Functions: Implementation*, ICCMSE, Athens, Greece (2015).

[15] C. Gözükirmizi and M. E. Kirkin, *Classical Symmetric Fourth Degree Potential Systems in Probabilistic Evolution Theoretical Perspective: Most Facilitative Conicalization and Squarification of Telescope Matrices*, International Conference in Nonlineer Problems in Aviation and Aerospace (ICNPAA), La Rochelle, France (2016).

[16] C. Gözükirmizi, M. E. Kirkin and M. Demiralp, Journal of Mathematical Chemistry, URL doi:http://dx.doi.org/10.1007/s10910-016-0678-8 (2017) 1-20.

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

Binary Kronecker Product Based Orthogonal Decompositions of Linear Algebraic Vectors

Metin Demiralp[1]

[1] *Computational Science and Engineering Department, Informatics Institute, İstanbul Technical University*

emails: metin.demiralp@gmail.com

Abstract

An algebraic vector which is a one way array can be decomposed to a linear combination of elements in a complete basis set. This is a quite ordinary idea for vector decomposition. However it becomes quite interesting when we take a basis set of orthonormal vectors each of which is a binary Kronecker product of same type. This work focuses on such decompositions each of which in fact uses appropriate matrix forms obtained from the target vector. We focus on singular value decomposition (SVD) and tridiagonal matrix enhanced multivariance products representation (TMEMPR) only in this work.

Key words: Kronecker product, Singular value decomposition, Tridiagonal Matrix Enhanced Multivariance Products Representation.

1 Introduction: Needs for Linear Algebraic Vector Decomposition

Probabilistic Evolution Theory (PREVTH) has been developed to get analytic solutions to explicit ODE(s) under initial value impositions recently in our group (Group for Science and Methods of Computing (G4SMC)) studies [1–16]. PREVTH solutions are given in Kronecker power series as long as the explicit ODE set at the focus has a conical right hand side. Space extension concept is also used to reduce the conicality to the highest degree monomial. The definition of Kronecker product, Kronecker power and Kronecker power series have been given in another paper [17] of this author in this conference proceeding.

 ISBN: 978-84-617-8694-7

Kronecker power series representation of a multivariate function is defined by using its multivariate Taylor series and therefore the partial derivative values evaluated at a specific expansion point denoted by the components $x_1^{(e)}$, ..., $x_N^{(e)}$ in N dimensional Cartesian space

$$f(\mathbf{x}) = \sum_{j=0}^{\infty} \mathbf{f}_j^T \mathbf{x}^{\otimes j} \tag{1}$$

where

$$\mathbf{x} \equiv \begin{bmatrix} x_1 - x_1^{(e)} & \dots & x_N - x_N^{(e)} \end{bmatrix} \tag{2}$$

which can be named "System Vector" in system theoretical point of view. In (1) f stands for a multivariate function depending on system vector $\mathbf{x}$ while the coefficient $\mathbf{f}_j$s are constant vectors of $N^j \times 1$ type, composed of constant elements in terms of the partial derivatives of f evaluated at the expansion point $\mathbf{x}^{(e)}$ in N-dimensional Cartesian space. These varying types of the coefficients are balanced by the types of the relevant Kronecker powers. Thus the jth monomial of the series has $1 \times N^j$ type array coefficient is balanced to a scalar by$N^j \times 1$ type $\mathbf{x}^{\otimes j}$. First two coefficients of the Kronecker power series in (1) are 1×1 type (scalar) $\mathbf{f}_0^T$ and $1 \times N$ type $\mathbf{f}_1$ vector transpose are unique in this representation. However, all remaining $\mathbf{f}_j$ vectors are not unique because the greater than or equal to 2 Kronecker powers of the independent variable vector $\mathbf{x}$ are orthogonal to certain finite number constant vectors. A detailed discussion on this issue has been given in another paper of this papers author [17]. Therein, these nonuniquenesses have been used for parametrization of the shifted coefficients through certain flexible parameters. Beyond these, certain decomposition relation additions have also been investigated up to certain level of detailing. Hence the decomposition of a vector to a linear combination of binary Kronecker products gains a lot of importance and we have devoted the remaining part of this presentation to this issue. We will present basically two decompositions: (i) Kronecker product based singular value decomposition, (ii) Tridiagonal Matrix Enhanced Multivariance Products Representation (TMEMPR) Based Decomposition.

2 Kronecker Product Based Singular Value Decomposition of Linear Algebraic Vectors

Singular Value Decomposition (SVD) of a matrix which can be even rectangular is formulated generally by using two mappings from one Cartesian space whose dimension matches target matrix column space dimension to another Cartesian space whose dimension matches target matrix row space dimension and its reverse such that each mapping somehow define a semi-eigenvalue problem to get a true eigenvalue problem when they are combined. We do not intend to repeat the details of this well-known formulation here. Beyond the

 ISBN: 978-84-617-8694-7

above formulation we can generate the SVD formula by optimizing the Euclidean distance between the target matrix and an outer product of two unit normed vectors whose number of elements match the row and column number of the target matrix scaled by an arbitrary constant where everything is considered real-valued for brevity. This gives some number of possibilities for the outer product and the scaling factor such that all solutions match one of the singular value decompositions additive terms as can be proven by using consecutive constrained optimizations until the Euclidean distance of the remaining target vanishes. We do not intend to detail this formulation since the following parts will reveal the same type actions for our present Kronecker product based singular value decomposition of a linear algebraic vector. Now for our current decomposition we will denote the target vector to be decomposed by $\mathbf{a}$ and assume that its number of element is mn where m and n are two positive integer numbers for convenience. We are going to propose the following cost functional for optimisation

$$\begin{aligned} \mathcal{J}\left(\sigma, \mathbf{u}, \mathbf{v}, \lambda_u, \lambda_v\right) &\equiv \left\|\mathbf{a} - \sigma \mathbf{u} \otimes \mathbf{v}\right\|^2 + \lambda_u \left(\mathbf{u}^T \mathbf{u} - 1\right) + \lambda_v \left(\mathbf{v}^T \mathbf{v} - 1\right) \\ &= \mathbf{a}^T \mathbf{a} - 2\sigma \mathbf{a}^T \left(\mathbf{u} \otimes \mathbf{v}\right) + \sigma^2 \mathbf{u}^T \mathbf{u} \mathbf{v}^T \mathbf{v} + \lambda_u \left(\mathbf{u}^T \mathbf{u} - 1\right) \\ &\quad + \lambda_v \left(\mathbf{v}^T \mathbf{v} - 1\right) \end{aligned} \tag{3}$$

where we have used the unit norm constraints on the vectors $\mathbf{u}$ of m elements and $\mathbf{v}$ of n elements. This cost functional depends on three scalars, σ, λ_u, λ_v and two vectors, $\mathbf{u}$ and $\mathbf{v}$. Its optimization necessitates the setting of its partial derivatives with respect to σ, λ_u and λ_v equal to zero and setting its gradients with respect to the vectors $\mathbf{u}$ and $\mathbf{v}$ equal to $\mathbf{0}_m$ and $\mathbf{0}_n$, the zero vectors of m and n elements respectively. We can at first write

$$\frac{\partial \mathcal{J}}{\partial \sigma} = 2\sigma \mathbf{u}^T \mathbf{u} \mathbf{v}^T \mathbf{v} - 2\mathbf{a}\left(\mathbf{u} \otimes \mathbf{v}\right) \tag{4}$$

$$\frac{\partial \mathcal{J}}{\partial \lambda_u} = \mathbf{u}^T \mathbf{u} - 1 \tag{5}$$

$$\frac{\partial \mathcal{J}}{\partial \lambda_v} = \mathbf{v}^T \mathbf{v} - 1 \tag{6}$$

These form three scalar equations for solving unknowns. We need two additional vector equations. To this end we have to evaluate the gradients of additive terms in the expression of the cost functional. We can write first

$$\mathbf{a} \equiv \left[\, \mathbf{a}_1^T \;\; \ldots \;\; \mathbf{a}_N^T \,\right] \tag{7}$$

where $\mathbf{a}_i$s stand for n-element subvector components of $\mathbf{a}$. This takes us to the following equality

$$\mathbf{a}^T \left(\mathbf{u} \otimes \mathbf{v}\right) = \sum_{i=1}^{m} u_i \mathbf{a}_i^T \mathbf{v} = \mathbf{u}^T \mathbf{A} \mathbf{v} \tag{8}$$

 ISBN: 978-84-617-8694-7

where $\mathbf{A}$ stands for the $m \times n$ type matrix defined as follows

$$\mathbf{A} \equiv [\, \mathbf{a}_1 \;\; ... \;\; \mathbf{a}_m \,]^T \tag{9}$$

All these enable us to rewite the explicit expression of the cost functional as follows

$$\begin{aligned} \mathcal{J}\left(\sigma, \mathbf{u}, \mathbf{v}, \lambda_u, \lambda_v\right) &= \mathbf{a}^T\mathbf{a} - 2\sigma\mathbf{a}^T\left(\mathbf{u} \otimes \mathbf{v}\right) + \sigma^2\mathbf{u}^T\mathbf{u}\mathbf{v}^T\mathbf{v} + \lambda_u\left(\mathbf{u}^T\mathbf{u} - 1\right) \\ &\quad + \lambda_v\left(\mathbf{v}^T\mathbf{v} - 1\right) \end{aligned} \tag{10}$$

which permit us to write

$$\boldsymbol{\nabla}_{\mathbf{u}}\mathcal{J} = -2\sigma\mathbf{A}\mathbf{v} + 2\sigma^2\mathbf{v}^T\mathbf{v}\mathbf{u} + 2\lambda_u\mathbf{u} = \mathbf{0}_m \tag{11}$$

$$\boldsymbol{\nabla}_{\mathbf{v}}\mathcal{J} = -2\sigma\mathbf{A}^T\mathbf{u} + 2\sigma^2\mathbf{u}^T\mathbf{u}\mathbf{v} + 2\lambda_v\mathbf{v} = \mathbf{0}_n \tag{12}$$

Premultiplication of (11) and (12) with the transposes of vectors $\mathbf{u}$ and $\mathbf{v}$ results in the following scalar equations

$$-2\sigma\mathbf{u}^T\mathbf{A}\mathbf{v} + 2\sigma^2 + 2\lambda_u = 0, \qquad -2\sigma\mathbf{v}^T\mathbf{A}^T\mathbf{u} + 2\sigma^2 + 2\lambda_v = 0 \tag{13}$$

where we have used (5) and (6). On the other hand, the combination of (8) with (4), (5) and (6) makes the following equality valid.

$$\sigma = \mathbf{u}^T\mathbf{A}\mathbf{v} \tag{14}$$

The employment of this intermediate result in (13) takes us to the conclusion that the Lagrange parameters of optimization, λ_u and λ_v, vanish. This implies that optimisation could have been conducted without imposing the unit norm condition on $\mathbf{u}$ and $\mathbf{v}$ through the constraints. Instead we could straightforwardly impose to the construction of cost functional by taking the norms of these vectors directly equal to 1. This achievement is quite natural for spectral-decomposition-like or SVD-like formulations.

Now we can reorganize (11) and (12) by setting λ_u and λ_v equal to 0 while setting the norms of the vectors $\mathbf{u}$ and $\mathbf{v}$ equal to 1. This gives

$$\mathbf{A}\mathbf{v} = \sigma\mathbf{u}, \qquad \mathbf{A}^T\mathbf{u} = \sigma\mathbf{v} \tag{15}$$

which implies that the σ parameter can take the singular values of $\mathbf{A}$ while $\mathbf{u}$ and $\mathbf{v}$ correspond to the relevant singular vectors of same matrix. The matrix $\mathbf{A}$ is composed of rows each of which is equal to one of $\mathbf{a}_i$s. Hence it is rowwise folding of the original vector $\mathbf{a}$. For this folding the vectors $\mathbf{v}$ and $\mathbf{u}$ are the relevant right and left singular vectors of the matrix $\mathbf{A}$ to the singular value σ. All these mean that we can define the Kronecker product

 ISBN: 978-84-617-8694-7

based singular value decomposition as the singular value decomposition of the matrix $\mathbf{A}$. We can therefore write

$$\mathbf{a} = \sum_{i=1}^{\min(m,n)} \sigma_i \mathbf{u}_i \otimes \mathbf{v}_i \tag{16}$$

where σ_i, $\mathbf{u}_i$, $\mathbf{v}_i$ are the ith Kronecker Product Based SVD singular value, normalized right singular vector, normalized left singular vector respectively as long as the multiple singular states are counted separately.

3 Kronecker Product Based Tridiagonal Matrix Enhanced Products Representation for an Ordinary Linear Algebraic Vector

We have been able to develop many extensions to the High Dimensional Model Representation (HDMR) and Enhanced Multivariance Products Representation (EMPR) [18–21] takes a very important place in these extensions. We have even brought new extensions to EMPR such that the Bivariate EMPR has been used as the backbone of the new methodology and its utilization is repeated with different support functions which are the basic additional components of EMPR. Bivariate EMPR can be written as follows for a given matrix $\mathbf{A}$ of $m \times n$ type

$$\mathbf{A} = \alpha \mathbf{u}\mathbf{v}^T + \overline{\mathbf{a}}_1 \mathbf{v}^T + \mathbf{u}\overline{\mathbf{a}}_2^T + \overline{\mathbf{A}}_{1,2} \tag{17}$$

where m-element $\mathbf{u}$ and n-element $\mathbf{v}$ vectors are given and called "Left" and "Right" "Support Vectors". They have unit norms and can be optimised to increase the descending dominance of right hand side additive terms. The right hand side entities, α, $\overline{\mathbf{a}}_1$, $\overline{\mathbf{a}}_2$ and $\overline{\mathbf{A}}_{1,2}$, stand for a scalar, an m-element vector, an n-element vector and an $m \times n$ type matrix. They are called in a way such that "constant component" for α, "univariate components for $\overline{\mathbf{a}}_1$ and $\overline{\mathbf{a}}_2$, and finally, "remainder matrix component" for $\mathbf{A}$. These are determined in such a way that additive terms become mutually orthogonal in Frobenius matrix inner product. We do not give further details since the basic content of this section will reveal all these details parallel to the formulations of this section.

Now as being inspired by Bivariate EMPR for a given matrix we can propose the following expression for a given vector $\mathbf{a}$ with mn elements where m and n are given positive integers.

$$\mathbf{a} = \alpha \mathbf{u} \otimes \mathbf{v} + \overline{\mathbf{a}}_1 \otimes \mathbf{v} + \mathbf{u} \otimes \overline{\mathbf{a}}_2^T + \overline{\mathbf{a}}_{1,2} \tag{18}$$

where the remainder term can be considered as the agent to be suppressed in norm. We can define the following Euclidean Distance cost functional by considering support vectors

 ISBN: 978-84-617-8694-7

as given entities while EMPR components are considered as the items to be optimised.

$$\begin{aligned}\mathcal{J}_{cd}\left(\alpha,\overline{\mathbf{a}}_1,\overline{\mathbf{a}}_2\right) &\equiv \left\|\mathbf{a}-\alpha\mathbf{u}\otimes\mathbf{v}-\overline{\mathbf{a}}_1\otimes\mathbf{v}-\mathbf{v}\otimes\overline{\mathbf{a}}_2\right\|^2\\ &= \mathbf{a}^T\mathbf{a}+\alpha^2+\overline{\mathbf{a}}_1^T\overline{\mathbf{a}}_1+\overline{\mathbf{a}}_2^T\overline{\mathbf{a}}_2+2\alpha\mathbf{u}^T\overline{\mathbf{a}}_1+2\alpha\mathbf{v}^T\overline{\mathbf{a}}_2+2\mathbf{u}^T\overline{\mathbf{a}}_1\mathbf{v}^T\overline{\mathbf{a}}_2\\ &\quad -2\mathbf{a}^T\left(\mathbf{u}\otimes\mathbf{v}\right)-2\mathbf{a}^T\left(\overline{\mathbf{a}}_1\otimes\mathbf{v}\right)-2\mathbf{a}^T\left(\mathbf{v}\otimes\overline{\mathbf{a}}_2\right)\\ &= \mathbf{a}^T\mathbf{a}+\alpha^2+\overline{\mathbf{a}}_1^T\overline{\mathbf{a}}_1+\overline{\mathbf{a}}_2^T\overline{\mathbf{a}}_2+2\alpha\mathbf{u}^T\overline{\mathbf{a}}_1+2\alpha\mathbf{v}^T\overline{\mathbf{a}}_2+2\mathbf{u}^T\overline{\mathbf{a}}_1\mathbf{v}^T\overline{\mathbf{a}}_2\\ &\quad -2\mathbf{u}^T\mathbf{A}\mathbf{v}-2\overline{\mathbf{a}}_1^T\mathbf{A}\mathbf{v}-2\mathbf{u}^T\mathbf{A}\overline{\mathbf{a}}_2 \end{aligned} \tag{19}$$

where we have used the fact that each support function has unit norm by definition. This may not be considered as the cost functional to be used in optimisation because of certain constraints on the right hand side components.

(18) imposes just a single vector equation amongst one scalar, two vectors, and, one matrix components and is not sufficient to uniquely determine the right hand side components. To get uniqueness the extended forms of Sobol conditions in High Dimensional Model Representation are needed to be used. They are called vanishing product conditions and two of them dictates us that $\overline{\mathbf{a}}_1$ and $\overline{\mathbf{a}}_2$ need to be orthogonal to the support vectors, $\mathbf{u}$ and $\mathbf{v}$, respectively. Thus using these vanishing inner products as constraints we can define the following new cost functional instead of the one in (19) as follows

$$\begin{aligned}\mathcal{J}\left(\alpha,\overline{\mathbf{a}}_1,\overline{\mathbf{a}}_2,\lambda_1,\lambda_2\right) &\equiv \mathcal{J}_{cd}\left(\alpha,\overline{\mathbf{a}}_1,\overline{\mathbf{a}}_2\right)+\lambda_1\mathbf{u}^T\overline{\mathbf{a}}_1+\lambda_2\mathbf{v}^T\overline{\mathbf{a}}_2\\ &= \mathbf{a}^T\mathbf{a}+\alpha^2+\overline{\mathbf{a}}_1^T\overline{\mathbf{a}}_1+\overline{\mathbf{a}}_2^T\overline{\mathbf{a}}_2-2\mathbf{u}^T\mathbf{A}\mathbf{v}-2\overline{\mathbf{a}}_1^T\mathbf{A}\mathbf{v}-2\mathbf{u}^T\mathbf{A}\overline{\mathbf{a}}_2\\ &\quad +\lambda_1\mathbf{u}^T\overline{\mathbf{a}}_1+\lambda_2\mathbf{v}^T\overline{\mathbf{a}}_2\end{aligned} \tag{20}$$

By setting the first partial derivative of this cost functional with respect to α equal to zero and then solving the resulting equation we can obtain

$$\alpha=\mathbf{u}^T\mathbf{A}\mathbf{v} \tag{21}$$

On the other hand, by setting the first partial derivatives of this cost functional with respect to λ_1 and λ_2 equal to zero and then solving the resulting equations we can obtain

$$\mathbf{u}^T\overline{\mathbf{a}}_1=0 \qquad \mathbf{v}^T\overline{\mathbf{a}}_2=0 \tag{22}$$

The remaining equations can be obtained by setting the gradients of the cost functional with respect to the vectors $\overline{\mathbf{a}}_1$ and $\overline{\mathbf{a}}_2$ equal to m element and n element zero vectors respectively. We can write

$$2\overline{\mathbf{a}}_1-2\mathbf{A}\mathbf{v}-\lambda_1\mathbf{u}=\mathbf{0}_m, \qquad 2\overline{\mathbf{a}}_2 2\mathbf{A}^T\mathbf{u}-\lambda_2\mathbf{v}=\mathbf{0}_n \tag{23}$$

 ISBN: 978-84-617-8694-7

whose equations can be premultiplied by the transposes of the vectors, $\mathbf{u}$ and $\mathbf{v}$, respectively; and then the resulting scalar equations can be solved for λ_1 and λ_2. This action produces

$$\lambda_1 = \lambda_2 = 2\mathbf{u}^T\mathbf{A}\mathbf{v} = 2\alpha \tag{24}$$

This takes us from (23) to the following results

$$\overline{\mathbf{a}}_1 = \mathbf{A}\mathbf{v} + \alpha\mathbf{u}, \qquad \overline{\mathbf{a}}_2 = \mathbf{A}^T\mathbf{u} + \alpha\mathbf{v} \tag{25}$$

(21) and (25) shows that α, $\overline{\mathbf{a}}_1$, $\overline{\mathbf{a}}_2$ components of the vector $\mathbf{a}$ match the Bivariate EMPR components of the matrix $\mathbf{A}$ which is a specific folded form of the vector $\mathbf{a}$. This implies that the remainder vector $\overline{\mathbf{a}}_{1,2}$ should give the remainder matrix, $\mathbf{A}_{1,2}$ when it is folded in accordance with the folding of the target vector $\mathbf{a}$.

TMEMPR is based on consecutive Bivariate EMPR (BEMPR) application to the target matrix. The remainder matrix of first BEMPR can be taken as a new target and then the univariate components of the first BEMPR can used as the new support vectors after they are scaled such that their norms become 1. This produces new α and new univariate components together with a new remainder matrix. Then a new BEMPR is launched with this remainder at the focus by using the new univariate vectors as the new support vectors. As the number of consecutive BEMPRs increases the newly formed remainder matrix becomes having decreasing ranks such that after some number of consecutive steps the remainder becomes disappearing. This builds Tridiagonal Matrix Enhanced Multivariance Products Representation (TMEMPR) whose explicit structure is given below

$$\mathbf{A} = \sum_{i=1}^{n_\alpha} \alpha_i \mathbf{u}_i \mathbf{v}_i^T + \sum_{i=1}^{n_\beta} \beta_i \mathbf{u}_{i+1} \mathbf{v}_i^T + \sum_{i=1}^{n_\gamma} \mathbf{u}_i \mathbf{v}_{i+1}^T \tag{26}$$

where n_α stands for $\min(m, n)$, the minimum of the positive integer row and column numbers, while n_β characterizes $(m-1)$ when $m < n$ and m when $m > n$. Similarly, n_γ represents m when $m < n$ and $(m-1)$ when $m > n$. In the case where $m = n$ both n_β and n_γ take the common value, $(m-1)$. Curious readers can refer to our basic relevant publications [18–21] for much more detailed information on TMEMPR.

Above analysis and the most recently given formula can be parallelly adapted to the case of vector decomposition through TMEMPR as follows

$$\mathbf{a} = \sum_{i=1}^{n_\alpha} \alpha_i \mathbf{u}_i \otimes \mathbf{v}_i^T + \sum_{i=1}^{n_\beta} \beta_i \mathbf{u}_{i+1} \otimes \mathbf{v}_i^T + \sum_{i=1}^{n_\gamma} \mathbf{u}_i \otimes \mathbf{v}_{i+1}^T \tag{27}$$

where all entities, α, β, and, γ parameters together with the support vectors, $\mathbf{u}$s and $\mathbf{v}$s, are known in terms of the target vector $\mathbf{a}$ and the initially given support vectors. This completes the construction of our basic goal, the vector decomposition via TMEMPR.

4 Concluding Remarks

This work has been devoted to the linear algebraic vector decomposition by using binary Kronecker product terms. Our important remarks are enumerated below

1. We have focused on Singular Value Decomposition (SVD) of an ordinary linear algebraic vector such that each additive term is proportional to the Kronecker product of two vectors each of which has been taken from an orthonormal vector set, each of which is from a different Cartesian vector space.

2. We have also focused on Tridiagonal Matrix Enhanced Multivariance Products Representation (TMEMPR). We have immediately noticed that each additive term outer product should be changed with the Kronecker product of the same factors of the relevant outer product.

3. Kronecker Product Based SVD and TMEMPR are not the only possible ordinary linear algebraic vector decompositions. However, these are the most basic two of standing approaches based on powerful theoretical backgrounds.

4. These two decompositions can be used for shifting from a Kronecker power to a another higher Kronecker power. This inserts quite many arbitrary parameters to shifted Kronecker power coefficient and enable us to suppress the norms of coefficient vectors or matrices as much as possible. Even though the handling of these parameter evaluations stands rather comprehensive it has the power of monitoring the Kronecker power series convergences.

5. By having convergence monitoring powers of these decompositions we are now at an important point that we can extend the Probabilistice Evolution Theory (PREVTH) we have developed for conical systems in last decade to highest power monomiality. This will form the content of our very near future publication. We have been now equipped very much to this end.

References

[1] M. Demiralp, *Quantum Expected Value Dynamics in Probabilistic Evolution Perspective*, Proceedings of the 12th International Conference on Computational and Mathematical Methods in Science and Engineering (CMMSE), ISBN: 978-84-615-5392-1, Murcia, Spain (2012), 449-459.

[2] M. Demiralp, Journal of Mathematical Chemistry, URL doi:http://dx.doi.org/10.1007/s10910-012-0079-6, **51**(4), (2012) 1170.

[3] M. Demiralp and B. Tunga, Journal of Mathematical Chemistry, URL doi:http://dx.doi.org/10.1007/s10910-012-0081-z, **51**(4), (2012), 1198.

[4] E. Demiralp, M. Demiralp and L. Hernandez-Garcia, *Journal of Mathematical Chemistry*, URL doi:http://dx.doi.org/10.1007/s10910011-9929-x, **50**, (2012) 850.

[5] M. Demiralp and E. Demiralp, Journal of Mathematical Chemistry, URL doi:http://dx.doi.org/10.1007/s10910-012-0070-2 **51**(1) (2012) 58.

[6] M. Demiralp and E. Demiralp, Journal of Mathematical Chemistry, URL doi:http://dx.doi.org/10.1007/s10910-012-0064-0 **51**(19) (2012) 38.

[7] S. Tuna and M. Demiralp, *Certain Validations of Probabilistic Evolution Approach for Initial Value Problems*, Proceedings of the 12th WSEAS International Conference on Systems Theory and Scientific Computation (ISTASC'12), ISBN: 978-1-61804-115-9, İstanbul, Türkiye (2012) 246-249.

[8] N. A. Baykara, E. Gürvit and M. Demiralp, *Univariate single quantum harmonic oscillator from probabilistic evolution perspective*, Proceedings of the 13th WSEAS International Conference on Mathematics and Computers in Biology and Chemistry (MCBC'12), Wisconsin, ABD (2012) 27-32.

[9] M. Ayvaz and M. Demiralp, *Getting Triangularity and Conicality in the Probabilistic Evolutionary Expectation Dynamics of the Purely Quartic Quantum Anharmonic Oscillator*, Proceedings of the 12th WSEAS International Conference on Systems Theory and Scientific Computation (ISTASC'12), ISBN: 978-1-61804-115-9, İstanbul, Türkiye (2012) 268-271.

[10] E. Demiralp, M. Demiralp and L. Hernandez-Garcia, *Journal of Mathematical Chemistry*, URL doi:http://dx.doi.org/10.1007/s10910-011-9930-4, **50**, (2012) 870.

[11] B. Tunga and M. Demiralp, *Probabilistic Evolution of the State Variable Expected Values in Liouville Equation Perspective, for a Many Particle System Interacting Via Elastic Forces*, Proceedings of the 12th International Conference on Computational and Mathematical Methods in Science and Engineering (ICCMSE), ISBN: 978-84-615-5392-1, Murcia, Spain (2012) 1186-1197.

[12] M. Demiralp, *Squarificating the Telescope Matrix Images of Initial Value Vector in Probabilistic Evolution Theory (PET)*, Proceedings of the 19th International Conference on Applied Mathematics (AMATH'14), ISBN: 978-1-61804-258-3, İstanbul, Türkiye (2014) 99104.

[13] M. E. Kirkin and C. Gözükirmizi, *Probabilistic Evolution Theory for ODE Sets with Second Degree Multinomial Right Hand Side Functions: Certain Reductive Cases*, ICCMSE, Athens, Greece (2015).

[14] C. Gözükirmizi, *Probabilistic Evolution Theory for ODE Sets with Second Degree Multinomial Right Hand Side Functions: Implementation*, ICCMSE, Athens, Greece (2015).

[15] C. Gözükirmizi and M. E. Kirkin, *Classical Symmetric Fourth Degree Potential Systems in Probabilistic Evolution Theoretical Perspective: Most Facilitative Conicalization and Squarification of Telescope Matrices*, International Conference in Nonlineer Problems in Aviation and Aerospace (ICNPAA), La Rochelle, France (2016).

[16] C. Gözükirmizi, M. E. Kirkin and M. Demiralp, Journal of Mathematical Chemistry, URL doi:http://dx.doi.org/10.1007/s10910-016-0678-8 (2017) 1-20.

[17] M. Demiralp, *Tricky Aspects of Kronecker Power Series in Constancy Adding Space Extention (CASE) Perspective*, 17th International Conference on Computational Mathematical Methods in Science and Engineering (CMMSE), (to appear in Proceedings)

[18] E. K. Özay, M. Demiralp, *Weighted Tridiagonal Matrix Enhanced Multivariance Products Representation (WTMEMPR) for Decompositions of Multiway Arrays: Applications on Certain Chemical System Data Sets*, J. Math. Chem., **55**, (2014) 455-476, DOI:10.1007/s10910-014-0371-8

[19] E. K. Özay, M. Demiralp, *Reductive Enhanced Multivariance Product Representation for Multi-way Arrays*, J. Math. Chem., **52**, (2014) 2546-2558, DOI:10.1007/s10910-014-0371-8

[20] E. K. Özay, M. Demiralp, *Combined Small Scale High Dimensional Model Representation*, J. Math. Chem., 50, (2012) 2023-2042, DOI:10.1007/s10910-012-0018-6

[21] B. Kalay, M. Demiralp, *Fundamental Elements of Vector Enhanced Multivariance Product Representation*, International Conference of Numerical Analysis and Applied Mathematics, 15-20 September 2012, Kos Island, Greece, AIP Proceedings, **1479**, (2012) 1998-2001

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

Highest Monomiality Based Probabilistic Evolution Theoretical (PREVTH) Solutions to Explicit Ordinary Differential Equations

Metin Demiralp[1]

[1] *Computational Science and Engineering Department, Informatics Institute, İstanbul Technical University*

emails: metin.demiralp@gmail.com

Abstract

This work can be regarded as the most contemporary and important development realized in most recent Demiralp group studies, nine of which are reported in this conference's presentations. A well developed Probabilistic Evolution Theory (PREVTH) has been developed by us in the last decade through step by step, maybe small but quite important, developments. However, that theory basically constructed to solve the explicit ODE(s) whose autonomous right hand side functions are conical (second degree multinomial) in unknown temporal functions. The obtention of conicality is generally based on certain space extension procedures which mostly increase the dimensionality of the space of unknowns. Our recent efforts have revealed that the conicality is not the only important facilitation to get rather simple analytic series solutions. It is also almost equivalently possible to develop similar algorithms to one for conicality case for any multinomiality in the right hand sides since as we have caught the fact that any multinomial can be converted to highest degree monomial by using some shift operations. This work is designed to focus on these issues by also referring to other two important papers of the author in this conference contributions.

Key words: Kronecker product, Kronecker power series, PREVTH, PREVTH recursions.

1 Introduction: Probabilistic Evolution Theory (PREVTH) in its Contemporary Form

In last decade of our group (Group for Science and Methods of Computing (G4SMC)) studies [1–16], we have developed a productive structure we have called Probabilistic Evolution

Theory (PREVTH) to obtain analytic solutions to first order autonomous explicit ODE(s) under initial value impositions. As long as the explicit ODE set at the focus has a conical right hand side PREVTH solutions are given in Kronecker power series even though the coefficient matrices or vectors may not be analytically found. The "Space Extension" concept can also be used to convert the conicality to the highest (second) degree monomiality to get analytically evaluable coefficient arrays (matrices or vectors). Curious readers can refer to our relevant publications [1–16]. We do not intend to repeat the details of solution construction procedure since almost same thing will be realized for the PREVTH's present form solution in this work.

2 Probabilistic Evolution Theory (PREVTH) on Highest Power Multinomiality

Let us consider the following vector ODE

$$\dot{\mathbf{x}}(t) = \mathbf{F}\mathbf{x}(t)^{\otimes n}, \qquad \mathbf{x}(0) = \mathbf{a} \tag{1}$$

where $\mathbf{x}(t)$ stands for the N-element system vector while n stands for a positive integer whereas $\mathbf{F}$ is a constant rectangular matrix of $N \times N^n$ type. Even though we can construct a Kronecker power series solution to this vector ODE it opens a new formalism from scratch. Instead of that construction it seems to be better to use an ODE over not system vector but its $(n-1)$th Kronecker power. To this end, we can write the following equality via simple temporal differentiation.

$$\begin{aligned} \frac{d\mathbf{x}(t)^{\otimes (n-1)}}{dt} &= \sum_{j=0}^{n-2} \mathbf{x}(t)^{\otimes j} \otimes \dot{\mathbf{x}}(t) \otimes \mathbf{x}(t)^{\otimes (n-j-2)} = \mathbf{G}\mathbf{x}(t)^{\otimes (2n-2)}, \\ \mathbf{G} &\equiv \sum_{j=0}^{n-2} \mathbf{I}_N^{\otimes j} \otimes \mathbf{F} \otimes \mathbf{I}_N^{\otimes (n-j-2)}, \qquad \mathbf{x}(0)^{\otimes (n-1)} = \mathbf{a}^{\otimes (n-1)} \end{aligned} \tag{2}$$

where $\mathbf{G}$ matrix is of $N^{n-1} \times N^{2n-2}$ type. Now we can write

$$\dot{\mathbf{y}}(t) = \mathbf{G}\mathbf{y}(t)^{\otimes 2}, \qquad \mathbf{y}(0) = \mathbf{a}^{\otimes (n-1)} \equiv \mathbf{b}, \qquad \mathbf{y}(t) \equiv \mathbf{x}(t)^{\otimes (n-1)} \tag{3}$$

where $\mathbf{y}$ is an N^{n-1} element vector. The ordinary differential equation herein allows us to use conicality based Probabilistic Evolution Theory (PREVTH) to construct a Kronecker Power Series solution. We are going to give intermediate steps explicitly for better explanation.

 ISBN: 978-84-617-8694-7

We can write the following equality from this purely conical ODE set.

$$\begin{aligned}\frac{d\mathbf{y}(t)^{\otimes j}}{dt} &= \sum_{k=0}^{j-1}\mathbf{y}(t)^{\otimes k}\otimes\dot{\mathbf{y}}(t)\otimes\mathbf{y}(t)^{\otimes(j-k-1)} = \mathbf{M}_j\mathbf{y}(t)^{\otimes j+1},\\ &\mathbf{M}_j \equiv \sum_{k=0}^{j-1}\mathbf{I}_{\overline{\mathrm{N}}}^{\otimes k}\otimes\mathbf{G}\otimes\mathbf{I}_{\overline{\mathrm{N}}}^{\otimes(j-k-1)},\quad \overline{\mathrm{N}}\equiv N^{n-1}\qquad \mathbf{y}(0)^{\otimes j}=\mathbf{b}^{\otimes j}\end{aligned}\tag{4}$$

where $\overline{N}^j\times\overline{N}^{j+1}$ type array, $\mathbf{M}_j$, can be called "Monocular Matrix" as we have done in Conicality Based Probabilistic Evolution Theory (PREVTH). The temporal integration of both sides in (4) takes us to the following integral equation

$$\mathbf{y}(t)^{\otimes j}=\mathbf{b}^{\otimes j}+\int_0^t d\tau\mathbf{M}_j\mathbf{y}(\tau)^{\otimes(j+1)},\qquad j=1,2,3,...\tag{5}$$

First J equations of this recursion can be used to get the following equality by eliminating the integrals over the Kronecker powers (except the highest one) of the system vector and performing all remaining intermediate simple integrations

$$\begin{aligned}\mathbf{y}(t) &= \sum_{j=0}^{J-1}\frac{t^j}{j!}\mathbf{T}_j\mathbf{b}^{\otimes(j+1)}+\int_0^t dt_1...\int_0^{t_{J-1}}dt_J\mathbf{T}_J\mathbf{y}\left(t_J\right)^{\otimes(J+1)},\qquad J=1,2,3,...\\ &\mathbf{T}_k\equiv\prod_{\ell=1}^{k}\mathbf{M}_\ell,\qquad k=0,1,2,...\end{aligned}\tag{6}$$

where $t_0\equiv t$ and the finite product over Monocular Matrices, whose upper limit is less than the corresponding lower limit, is assumed to be $\overline{N}$ element identity matrix. The multifold integral term in the first part of this formula tends to vanish as long as its integrand's norm tends to decrease as J increases unboundedly. However, this is equivalent to the convergence of the preceding multinomial part in the first equation of this formula. If this convergence happens to exist then we can write the solution as follows

$$\mathbf{y}(t)=\sum_{j=0}^{\infty}\frac{t^j}{j!}\mathbf{T}_j\mathbf{b}^{\otimes(j+1)},\tag{7}$$

As have done in Conicality Based Probabilistic Evolution Theory, the matrix $\mathbf{T}_j$ is called "j-th Telescope Matrix" and it is somehow consistent cascaded form of the Monocular Matrices.

The solution given through (7) is a denumerable infinite linear combination of $\overline{N}$ element vectors, jth of which is in fact the image of the $(j+1)$th augmented initial vector ($\mathbf{b}$) Kronecker power under the jth Telescope matrix ($\mathbf{T}_j$). The conicality coefficient matrix $\mathbf{G}$

is defined as a sum of certain left and right Kronecker products of the matrix $\mathbf{F}$ with appropriate identity matrices. Hence it is rather sparse and the sparsity dramatically increases as the positive integer n increases. On the other hand, Monocular Matrices add more sparsity because of their structures connecting them to the matrix $\mathbf{G}$. A similar sparsity addition comes from the relations of the Telescope Matrices to Monocular matrices. All these imply that the sparsity is a quite undesired happening because it consumes much more memory and execution times in computer applications, and hence, must be avoided. To this end a new compaction method we call squarification has been proposed in Demiralp group studies. Basic definition of the squarification is as follows

$$\mathbf{T}_j \mathbf{b}^{\otimes (j+1)} \equiv \mathbf{S}_j (\mathbf{b}) \mathbf{b}, \qquad j = 0, 1, 2, ... \tag{8}$$

where $\mathbf{S}_j (\mathbf{b})$ stands for a matrix of $\overline{N} \times \overline{N}$ with elements depending on initial vector. The square matrices, $\mathbf{S}_j (\mathbf{b})$s can be evaluated by using specific procedure we call "Squarification". Since $\mathbf{T}_0 = \mathbf{I}_{\overline{N}}$ there is specific need for squarification and we take $\mathbf{S}_0 \equiv \mathbf{I}_{\overline{N}}$. However for $\mathbf{S}_1 (\mathbf{b})$ we can write first

$$\mathbf{T}_1 \mathbf{b}^{\otimes 2} = \mathbf{G}\mathbf{b}^{\otimes 2} \equiv \left[\mathbf{G}_1 \ \ ... \ \ \mathbf{G}_{\overline{\mathrm{N}}} \right] \mathbf{b}^{\otimes 2} = \left(\sum_{j=1}^{\overline{\mathrm{N}}} b_j \mathbf{G}_j \right) \mathbf{b} \tag{9}$$

which urges us to define the following "Squarification" procedure

$$\lfloor \mathbf{G}, \mathbf{b} \rceil = \left(\sum_{j=1}^{\overline{\mathrm{N}}} b_j \mathbf{G}_j \right) \mathbf{b}, \qquad \mathbf{S}_1 (\mathbf{b}) \equiv \lfloor \mathbf{G}, \mathbf{b} \rceil \tag{10}$$

where the leftmost symbol can be called "Squarification of $\mathbf{G}$ by $\mathbf{b}$", or briefly, "Squarification" while the matrix $\mathbf{G}$ and the vector $\mathbf{b}$ are called "Squarificant" and "Squarifier" respectively. The essential points in this squarification are the rules on the types of the squarificant and the squarifier. The matrix column number must be square of its row number which must also be equal to the number of elements in the squarifier. Beyond these, the matrices, $\mathbf{S}_j (\mathbf{b})$ are called "Squarified Telescope Matrices (SquTelMats)". These are vald only when the rectangular matrix to be squarified acts on a Kronecker square. If the rectangular matrix acts on a rather general entity, a Kronecker product of two different vectors then the rules may change accordingly. We are going to encounter such cases at the end of this section.

Even though the first SquTelMat, $\mathbf{S}_1 (\mathbf{b})$ is coincidentally a sole squarification, the other higher indexed SquTelMats generally have nested and powered squarifications. The expressions of these entities rapidly become quite complicated as the index of the SquTelMat tends to increase unboundedly. And, as a matter of fact, the determination of these expressions become formidable tasks even computer facilites are appropriately used. These facts have

 ISBN: 978-84-617-8694-7

urged us to seek a recursion amongst the SquTelMats. After comprehensive studies we have been able to get

$$\mathbf{S}_j(\mathbf{b}) = \sum_{k=0}^{j-1} \binom{j-1}{k} \lfloor \mathbf{G}, \mathbf{S}_k(\mathbf{b})\,\mathbf{b} \rceil\, \mathbf{S}_{j-k-1}(\mathbf{b}), \qquad j = 1, 2, ... \qquad \mathbf{S}_0(\mathbf{b}) = \mathbf{I}_{\overline{\mathbb{N}}} \tag{11}$$

whose validity had been shown by using some number of first SquTelMat expressions even though the general induction stage of the mathematical induction proof method has not been realized. Quite recently we could have been able to prove the validity of this recursion for all positive integer j values [21].

Although this recursion is constructed on square matrices, it is much better to convert it to a vector recursion for facilitating the computer algebra. To this end we can write

$$\mathbf{v}_j(\mathbf{b}) = \sum_{k=0}^{j-1} \binom{j-1}{k} \lfloor \mathbf{G}, \mathbf{v}_k(\mathbf{b}) \rceil\, \mathbf{v}_{j-k-1}(\mathbf{b}), \quad j = 1, 2, ... \quad \mathbf{v}_0(\mathbf{b}) = \mathbf{b} \tag{12}$$

where

$$\mathbf{v}_j(\mathbf{b}) \equiv \mathbf{S}_j(\mathbf{b})\,\mathbf{b} \tag{13}$$

Now the utilization of the last definition in (7) gives the ultimate PREVTH solution of (3) as follows

$$\mathbf{y}(t) = \sum_{j=0}^{\infty} \frac{t^j}{j!} \mathbf{v}_j(\mathbf{b}) \tag{14}$$

Now we can rewrite (1) as follows in the light of our most recent findings

$$\dot{\mathbf{x}}(t) = \mathbf{F}\mathbf{x}(t)^{\otimes n} = \mathbf{F}\left(\mathbf{y}(t) \otimes \mathbf{x}(t)\right) \qquad \mathbf{x}(0) = \mathbf{a} \tag{15}$$

which urges us to write the following much more general squarification where two vector Kronecker product appears instead of a single vector's Kronecker square.

$$\mathbf{F} \equiv [\,\mathbf{F}_1 \;\; ... \;\; \mathbf{F}_{N^{n-1}}\,], \qquad \lfloor \mathbf{F}, \mathbf{y}(t) \rceil \equiv \sum_{j=1}^{N^{n-1}} y_j(t) \mathbf{F}_j \tag{16}$$

The first vector factor of the Kronecker product must have same number of elements as the number of square blocks in the rectangular matrix squarificant while the second vector factor of the Kronecker product must have same number elements as the number of the rows of the squarificant. We do not distinguishly symbolize this squarifier for brevity and this does not create any confusion as long as the Kronecker product factors' and squarificant's

 ISBN: 978-84-617-8694-7

row and column numbers are incompatible for multiplication. This squarification enables us to finally write

$$\dot{\mathbf{x}}(t) = \lfloor \mathbf{F}, \mathbf{y}(t) \rceil \, \mathbf{x}(t) \qquad \mathbf{x}(0) = \mathbf{a} \tag{17}$$

This is a linear vector ODE with variable known coefficient and can be at least numerically solved since we have assumed that $\mathbf{y}(t)$ has been found from the PREVTH solution of the auxiliary vector ODE. On the other hand, this equation can also be handled by using a PREVTHwise recursion.

3 Solving Previous Section's Linear Vector ODE via Recursion Construction

We can now proceed to get the solution of (17) by proposing the following expansions

$$\mathbf{x}(t) \equiv \sum_{j=0}^{\infty} \frac{t^j}{j!} \mathbf{x}_j, \qquad \lfloor \mathbf{F}, \mathbf{y}(t) \rceil = \sum_{j=1}^{\infty} \frac{t^j}{j!} \lfloor \mathbf{F}, \mathbf{y}_j \rceil \tag{18}$$

which allows us to write

$$\dot{\mathbf{x}}(t) \equiv \sum_{j=0}^{\infty} \frac{t^j}{j!} \mathbf{x}_{j+1}, \qquad \lfloor \mathbf{F}, \mathbf{y}(t) \rceil \, \mathbf{x}(t) = \sum_{j=1}^{\infty} \frac{t^j}{j!} \sum_{k=0}^{j} \binom{j}{k} \lfloor \mathbf{F}, \mathbf{y}_k \rceil \, \mathbf{x}_{j-k} \tag{19}$$

as long as the convergences permit us. The plugging into (17) produces

$$\mathbf{x}_{j+1} = \sum_{k=0}^{j} \binom{j}{k} \lfloor \mathbf{F}, \mathbf{y}_k \rceil \, \mathbf{x}_{j-k}, \qquad j = 0, 1, 2, ...; \qquad \mathbf{x}_0 \equiv \mathbf{a}. \tag{20}$$

This is an infinite linear recursion whose solution's convergence characteristics can be investigated by using, for example, majorant series.

Let us consider the linear vector space spanned by the matrices of $N \times N^{n-1}$ type and denote it by $\mathcal{M}_{N \times N^{n-1}}$. We are going to consider the Frobenius inner product and relevant induced norm which can be explicitly defined over this space as follows

$$(\mathbf{M}_1, \mathbf{M}_2) \equiv Tr\left(\mathbf{M}_1^T \mathbf{M}_2\right), \qquad \|\mathbf{M}\| \equiv (\mathbf{M}, \mathbf{M}), \qquad \mathbf{M}_1, \mathbf{M}_2, \mathbf{M} \in \mathcal{M}_{N \times N^{n-1}} \tag{21}$$

Now we can write the following norm equality from the second equality in (16)

$$\|\lfloor \mathbf{F}, \mathbf{y}(t) \rceil\| \leq \sum_{j=1}^{N^{n-1}} |y_j(t)| \, \|\mathbf{F}_j\| \leq \left(\sum_{j=1}^{N^{n-1}} y_j(t)^2 \right)^{\frac{1}{2}} \left(\sum_{j=1}^{N^{n-1}} \|\mathbf{F}_j\|^2 \right)^{\frac{1}{2}} \equiv \|\mathbf{y}(t)\| \, \|\mathbf{F}\| \tag{22}$$

 ISBN: 978-84-617-8694-7

where we have used the Cauchy-Schawarz inequality when we write the rightmost inequality. On the other hand, we can also write

$$\|\mathbf{y}(t)\| \le \sum_{j=0}^{\infty} \frac{|t|^j}{j!} \|\mathbf{y}_k\| \equiv Y(t), \qquad \|\lfloor \mathbf{F}, \mathbf{y}(t) \rceil\| \le \|\mathbf{F}\| \, Y(t) \tag{23}$$

$$\|\mathbf{x}(t)\| \le \sum_{j=0}^{\infty} \frac{|t|^j}{j!} \|\mathbf{x}_k\| \equiv X(t), \qquad \|\lfloor \mathbf{F}, \mathbf{y}(t) \rceil \mathbf{x}(t)\| \le \|\mathbf{F}\| \, Y(t) X(t) \tag{24}$$

The first one of this couple of equation imply the following formulae when it is considered together with the second one of the same couple of equations

$$\|\dot{\mathbf{x}}(t)\| \le \sum_{j=0}^{\infty} \frac{|t|^j}{j!} \|\mathbf{x}_{k+1}\|, \qquad \dot{X}(t) = \|\mathbf{F}\| \, Y(t) X(t) \tag{25}$$

where we have converted the inequality to equality at the second formula since $X(t)$ has been considered as a majorant function. It should be accompanied with the initial imposition $X(0) = \mathbf{a}$. Then the resulting simple scalar ODE can be solved as follows

$$X(t) = \mathrm{e}^{\|\mathbf{F}\| \int_0^t d\tau Y(\tau)} \mathbf{a}, \qquad t \ge 0 \tag{26}$$

In accordance with this majorant function $\mathbf{x}(t)$ should converge in the same temporal interval as $Y(t)$ converges. This means that the essential point is the temporal convergence of the vector $\mathbf{y}(t)$. This convergence has been investigated under sufficient rigor in our relevant publications. [1–16]. Curios readers can refer to them.

4 Conicality Matrix and Convergence of Auxiliary Solution

We have used an auxiliary conicality based PREVTH solution in the previous sections. Therein the auxiliary vector $\mathbf{y}(t)$ has played an important role in the construction of the main solution. Hence its temporal series convergence is also a very important issue in the quality of the PREVTH solution for the highest monomiality based ODEs. On the other hand, our conicality based PREVTH solution convergence is basically determined by the matrix $\mathbf{G}$ we may call "Conicality Matrix" and also initial vector $\mathbf{b}$. We are not going to repeat the details of this issue. However, we may consider to investigate the specific structure of conicality in the sense of coefficient nonuniquenesses and the relevant flexibilities.

Second equality of (2) defines the conicality matrix, $\mathbf{G}$ explicitly. The structure of this matrix includes many different types identity matrices each of which is multiplied by a (generally) different Kronecker powers of the system vector such that these powers are

 ISBN: 978-84-617-8694-7

orthogonal to certain number of constant vectors. This enables us to add certain number of specific constant matrices to these identity matrices with arbitrary coefficients. So it is quite possible to insert many flexible parameters to the conicality matrix, **G**. Same thing happens not only to identity matrices but also to the core matrix (highest monomiality matrix), **F**. All these flexible parameters can be determined in such a way that the norm of the Conicality Matrix, **G** can be suppressed as much as possible to get a close quality to desired convergence rate. We do not intend to go beyond this point in this issue.

5 Concluding Remarks

We have presented how to construct a Kronecker power series solution to an explicit autonomous first order ODE set with a monomial at the right hand side and to this end we have used the fact that the right hand side multinomial's highest degree monomial can be considered as the only Kronecker power at the right hand side of the focused ODE set without any loss of generality due to our findings [27, 28]. Since we believe that our basic task has been completed in this work, we are not willing to repeat all details here.

Beside this work we have some other papers reporting most recent developments about PREVTH in this conference. One paper [21] focuses on the proof of the conicality based PREVTH and that development is a milestone as we believe.

Another paper [22] includes the solution of sensitivity coefficient evaluation problem and has been one of the basic motivation agent to put this work on the loom if the statement holds.

The quantum expectation value dynamics is a quite important issue in the solution of the quantum dynamical problems and until now we have used the Ehrenfest or Ehrenfest-like analysis. This was presenting a lot of undesired negativities. Now in three different papers [23–25], Heisenberg picture related concepts have been used first time and a set of operator equatons we call "Evolver Dynamical Equations" have been constructed. We believe that this development opens a new horizon in PREVTH use for quantum dynamical problems.

Finally, in our another paper [26] we have attempted to get analytic continuations to PREVTH solutions by using Padé approximants.

References

[1] M. Demiralp, *Quantum Expected Value Dynamics in Probabilistic Evolution Perspective*, Proceedings of the 12th International Conference on Computational and Mathematical Methods in Science and Engineering (CMMSE), ISBN: 978-84-615-5392-1, Murcia, Spain (2012), 449-459.

[2] M. Demiralp, Journal of Mathematical Chemistry, URL doi:http://dx.doi.org/10.1007/s10910-012-0079-6, **51**(4), (2012) 1170.

[3] M. Demiralp and B. Tunga, Journal of Mathematical Chemistry, URL doi:http://dx.doi.org/10.1007/s10910-012-0081-z, **51**(4), (2012), 1198.

[4] E. Demiralp, M. Demiralp and L. Hernandez-Garcia, *Journal of Mathematical Chemistry*, URL doi:http://dx.doi.org/10.1007/s10910011-9929-x, **50**, (2012) 850.

[5] M. Demiralp and E. Demiralp, Journal of Mathematical Chemistry, URL doi:http://dx.doi.org/10.1007/s10910-012-0070-2 **51**(1) (2012) 58.

[6] M. Demiralp and E. Demiralp, Journal of Mathematical Chemistry, URL doi:http://dx.doi.org/10.1007/s10910-012-0064-0 **51**(19) (2012) 38.

[7] S. Tuna and M. Demiralp, *Certain Validations of Probabilistic Evolution Approach for Initial Value Problems*, Proceedings of the 12th WSEAS International Conference on Systems Theory and Scientific Computation (ISTASC'12), ISBN: 978-1-61804-115-9, İstanbul, Türkiye (2012) 246-249.

[8] N. A. Baykara, E. Gürvit and M. Demiralp, *Univariate single quantum harmonic oscillator from probabilistic evolution perspective*, Proceedings of the 13th WSEAS International Conference on Mathematics and Computers in Biology and Chemistry (MCBC'12), Wisconsin, ABD (2012) 27-32.

[9] M. Ayvaz and M. Demiralp, *Getting Triangularity and Conicality in the Probabilistic Evolutionary Expectation Dynamics of the Purely Quartic Quantum Anharmonic Oscillator*, Proceedings of the 12th WSEAS International Conference on Systems Theory and Scientific Computation (ISTASC'12), ISBN: 978-1-61804-115-9, İstanbul, Türkiye (2012) 268-271.

[10] E. Demiralp, M. Demiralp and L. Hernandez-Garcia, *Journal of Mathematical Chemistry*, URL doi:http://dx.doi.org/10.1007/s10910-011-9930-4, **50**, (2012) 870.

[11] B. Tunga and M. Demiralp, *Probabilistic Evolution of the State Variable Expected Values in Liouville Equation Perspective, for a Many Particle System Interacting Via Elastic Forces*, Proceedings of the 12th International Conference on Computational and Mathematical Methods in Science and Engineering (ICCMSE), ISBN: 978-84-615-5392-1, Murcia, Spain (2012) 1186-1197.

[12] M. Demiralp, *Squarificating the Telescope Matrix Images of Initial Value Vector in Probabilistic Evolution Theory (PET)*, Proceedings of the 19th International Conference on Applied Mathematics (AMATH'14), ISBN: 978-1-61804-258-3, İstanbul, Türkiye (2014) 99104.

[13] M. E. Kirkin and C. Gözükirmizi, *Probabilistic Evolution Theory for ODE Sets with Second Degree Multinomial Right Hand Side Functions: Certain Reductive Cases*, ICCMSE, Athens, Greece (2015).

[14] C. Gözükirmizi, *Probabilistic Evolution Theory for ODE Sets with Second Degree Multinomial Right Hand Side Functions: Implementation*, ICCMSE, Athens, Greece (2015).

[15] C. Gözükirmizi and M. E. Kirkin, *Classical Symmetric Fourth Degree Potential Systems in Probabilistic Evolution Theoretical Perspective: Most Facilitative Conicalization and Squarification of Telescope Matrices*, International Conference in Nonlineer Problems in Aviation and Aerospace (ICNPAA), La Rochelle, France (2016).

[16] C. Gözükirmizi, M. E. Kirkin and M. Demiralp, Journal of Mathematical Chemistry, URL doi:http://dx.doi.org/10.1007/s10910-016-0678-8 (2017) 1-20.

[17] E. K. Özay, M. Demiralp, *Weighted Tridiagonal Matrix Enhanced Multivariance Products Representation (WTMEMPR) for Decompositions of Multiway Arrays: Applications on Certain Chemical System Data Sets*, J. Math. Chem., **55**, (2014) 455-476, DOI:10.1007/s10910-014-0371-8

[18] E. K. Özay, M. Demiralp, *Reductive Enhanced Multivariance Product Representation for Multi-way Arrays*, J. Math. Chem., **52**, (2014) 2546-2558, DOI:10.1007/s10910-014-0371-8

[19] E. K. Özay, M. Demiralp, *Combined Small Scale High Dimensional Model Representation*, J. Math. Chem., 50, (2012) 2023-2042, DOI:10.1007/s10910-012-0018-6

[20] B. Kalay, M. Demiralp, *Fundamental Elements of Vector Enhanced Multivariance Product Representation*, International Conference of Numerical Analysis and Applied Mathematics, 15-20 September 2012, Kos Island, Greece, AIP Proceedings, **1479**, (2012) 1998-2001

[21] C. Gözükirmizi, M. Demiralp, *Probabilistic Evolution Theory for Explicit Autonomous Ordinary Differential Equations: Recursion of Squarified Telescope Matrices and Optimal Space Extension*,International Conference on Computational and Mathematical Models in Science and Engineering (CMMSE 2017), 4-8 July 2017, Rota, Cadiz, Spain, CMMSE 2017 Proceedings, (2017), another contribution in the same minisymposium

[22] M. E. Kirkin, M. Demiralp, *Recursion Based Sensitivity Coefficient Determination for Probabilistic Evolution Theoretical (PREVTH) Solutions to Explicit Autonomous Ordinary Differential Equations*, International Conference on Computational

and Mathematical Models in Science and Engineering (CMMSE 2017), 4-8 July 2017, Rota, Cadiz, Spain, CMMSE 2017 Proceedings, (2017), another contribution in the same minisymposium

[23] B. Kalay, M. Demiralp, *A Probabilistic Evolution Theoretical (PREVTH) Approach to Quantum Evolver Dynamical Equations for Singular Hamiltonians: Fluctuationlessness Approximation*, International Conference on Computational and Mathematical Models in Science and Engineering (CMMSE 2017), 4-8 July 2017, Rota, Cadiz, Spain, CMMSE 2017 Proceedings, (2017), another contribution in the same minisymposium

[24] B. Kalay, M. Demiralp, *Initial Wavefunction Construction for Probabilistic Evolution Theoretical (PREVTH) Evolver Dynamics via PREVTH Parameters and Initial Wave Function Optimization*, International Conference on Computational and Mathematical Models in Science and Engineering (CMMSE 2017), 4-8 July 2017, Rota, Cadiz, Spain, CMMSE 2017 Proceedings, (2017), another contribution in the same minisymposium

[25] S. B. Özdemir, M. Demiralp, *Probabilistic Evolution Theoretical Formulation of Anharmonic Symmetric Quantum Oscillator by Using Quantum Evolver Dynamics*, International Conference on Computational and Mathematical Models in Science and Engineering (CMMSE 2017), 4-8 July 2017, Rota, Cadiz, Spain, CMMSE 2017 Proceedings, (2017), another contribution in the same minisymposium

[26] E. Tataroğlu. M. Demiralp, *Padé Approximants to Conicality Based Probabilistic Evolution Theory (PREVTH) Solutions: Two Classical Particles Systems Interacting via Central Forces*, International Conference on Computational and Mathematical Models in Science and Engineering (CMMSE 2017), 4-8 July 2017, Rota, Cadiz, Spain, CMMSE 2017 Proceedings, (2017), another contribution in the same minisymposium

[27] M. Demiralp, *Tricky Aspects of Kronecker Power Series in Constancy Adding Space Extention (CASE) Perspective*, 17th International Conference on Computational and Mathematical Methods in Science and Engineering (CMMSE 2017), 4-8 July 2017, Rota, Cadiz, Spain, CMMSE 2017 Proceedings, (2017), another contribution in the same minisymposium

[28] M. Demiralp, *Binary Kronecker Product Based Orthogonal Decompositions of Linear Algebraic Vectors*, International Conference on Computational and Mathematical Models in Science and Engineering (CMMSE 2017), 4-8 July 2017, Rota, Cadiz, Spain, CMMSE 2017 Proceedings, (2017), another contribution in the same minisymposium

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

A class of predator-prey models with a non-differentiable functional response

Josué Díaz-Avalos[1] and Eduardo González-Olivares[2]

[1] *Facultad de Ciencias Matemáticas, Universidad Nacional Mayor de San Marcos*

[2] *Pontifícia Universidad Católica de Valparaíso,*

emails: josuediazavalos@gmail.com, ejgonzal@ucv.cl

Abstract

In this work, a predator-prey model considering a non-differentiable functional response similar to the Cobb-Douglas type production function. We show that this function has a strong influence on the locally and globally dynamics of the model.

The non-uniqueness of solutions on the coordinate axis for any initial conditions is proved, i.e. the system is non-lipchitzian. Moreover, the model has at least one small amplitude limit cycles generated by Hopf bifurcation of a fine focus.

Tthe existence of a homocinic curve on the phase plane (threshold curve), which divides the behavior of the trajectories, iimplying that two solutions, to distinct side of this curve have different $\omega-limit$; so, the solutions highly sensitive to initial conditions.

Key words: predator-prey model, stabiity, separatrix curve, limit cycles,
MSC 2000: AMS codes 92D25; 34C23; 58F14; 58F21

1 Introduction

We analyzed a predator-prey model, in which the following aspects are considered: i) the prey natural growth is the logistic equation, [17],

ii) the functional response is a non-differentiable [4, 7, 11], and

iii) the equation for predator growth is the logistic type [1, 15, 17].

This last assumption characterize to the Leslie-Gower type models, in which the conventional environmental carrying capacity for predators K_y is a function of the available prey quantity [1, 10, 13].

We will consider that the predator consumption function or functional response is expressed by the $H(x,y) = qx^{\alpha}y^{\beta}$, known in Economy Sciences as a Cobb-Douglas type function [4]. This function is more realistic than the bilinear functional response $H(x,y) = qxy$, proposed in the Lesie-Gower predator-prey model [8, 9, 13].

To establish the local stability of the equilibrium points, the standard methodoogy cannot be used, [2, 3, 12], due the non-uniqueness of solutions over the coordinates axis. Thus, new methods must be proposed for the analysis of the singuarities over the axis.

2 The model

The predator-prey model is described by a family of vector fields or by autonomous bidimensional differential equations system of the Kolmogorov type [6] given by

$$X_{\mu}: \left\{ \begin{array}{ll} \frac{dx}{dt} & = r\left(1-\frac{x}{K}\right)x - qx^{\alpha}y^{\beta} \\ \frac{dy}{dt} & = s\left(1-\frac{y}{nx+c}\right)y \end{array} \right. \tag{1}$$

where $x(t)$ and $y(t)$ denote the prey and predator population size, respectively, of densities as functions of time, and the vector of parameters

$$\mu = (r, K, q, s, n, c, \alpha, \beta) \in \mathbb{R}^6_+ \times]0,1]^2$$

having different ecological meanings.

In order to make an adequate description of the behaviour of system (1) and to simplify calculus, we follow the methodology used in [8, 9, 10], making a change in variables and a time rescaling given by

$$\varphi : \mathbb{R}^2_+ \times \mathbb{R} \to \mathbb{R}^2_+ \times \mathbb{R},$$

such that

$$\varphi(u,v,\tau) = \left(ku, nkv, \frac{(u+C)}{r}\tau\right) = (x,y,t)$$

with

$$detD\varphi(u,v,\tau) = \left| \begin{pmatrix} k & 0 & 0 \\ 0 & nk & 0 \\ \frac{\tau}{r} & 0 & \frac{(u+C)}{r} \end{pmatrix} \right| = \frac{nk^2(u+C)}{r} > 0.$$

Then, φ is a diffeomorphism preserving time orientation; hence, in the new coordinates, the topologically equivalent vector field [5, 3] to X_{μ} is

$$Y_{\eta}: \left\{ \begin{array}{ll} \frac{du}{d\tau} & = \left((1-u)u - Qu^{\alpha}v^{\beta}\right)(u+C) \\ \frac{dv}{d\tau} & = S(u+C-v)v \end{array} \right. , \tag{2}$$

where $\eta = (Q, C, S, \alpha, \beta) \in \mathbb{R}^3 \times]0,1]^2$ with $Q = \frac{q}{r}k^{\alpha+\beta-1}n^{\beta}$, $C = \frac{c}{nk}$ and $S = \frac{s}{r}$.

 ISBN: 978-84-617-8694-7

The equilibrium points of system (2) or singularities of vector field Y_η are: $(0,0)$, $(1,0)$, $(0,C)$ and $(u_e,v_e)\in\mathbb{R}^2_+$, satisfying the equations of the isoclines

$$\begin{aligned}(1-u)\,u-Qu^\alpha v^\beta &= 0\\ u+C-v &= 0\end{aligned}\tag{3}$$

Since $(1-u)\,u=Qu^\alpha v^\beta>0$ then $0<u<1$. Thus, of the second isocline, we obtain $C<v<1+C$. The abscissa u is a solution of the equation

$$f(u)=Q$$

where the function $f:[0,1]\to\mathbb{R}_{+,0}$ is given by

$$f(u)=\tfrac{(1-u)u}{u^\alpha(u+C)^\beta},\qquad (C,\alpha,\beta)\in\mathbb{R}_+\times\,]0,1]^2.$$

We consider two cases: $\alpha<1$ and $\alpha=1$.

If $\alpha<1$, $f(0)=f(1)=0$ and we obtain

$$\tfrac{df}{du}(u)=-\tfrac{q(u)}{u^\alpha(u+C)^{\beta+1}},\qquad q(u)=au^2+bu-C(1-\alpha),$$

with $a=2-\beta-\alpha>0$ and $b=\alpha+\beta-1+(2-\alpha)C$. q is a parabola that opens upward with only positive root $\bar{u}$, then

$$\tfrac{df}{du}(\bar{u})=0\Leftrightarrow q(\bar{u})=0\Leftrightarrow \bar{u}=\tfrac{-b+\sqrt{b^2+4aC(1-\alpha)}}{2a}$$

Moreover, since $q(\bar{u})=0<1+C=q(1)$, $\bar{u}<1$. Then $\bar{u}\in\,]0,1[$. Deriving again and evaluating in $\bar{u}$ we have

$$\tfrac{d^2f}{du^2}(u)=\tfrac{q(u)((1+\alpha+\beta)u+\alpha C)}{u^{\alpha+1}(u+C)^{\beta+2}}-\tfrac{b+2au}{u^\alpha(u+C)^{\beta+1}},\qquad \tfrac{d^2f}{du^2}(\bar{u})=-\tfrac{\sqrt{b^2+4aC(1-\alpha)}}{\bar{u}^\alpha(\bar{u}+C)^{\beta+1}}<0.$$

Therefore, f is increasing in $[0,\bar{u}]$, decreasing in $[\bar{u},1]$ and with maximun value in $\bar{u}$.

For $\alpha=1$ we have

$$f(u)=\tfrac{(1-u)}{(u+C)^\beta},\qquad f(0)=\tfrac{1}{C^\beta},\qquad f(1)=0,\qquad \tfrac{df}{du}(u)=-\tfrac{\beta+(1-\beta)u}{(u+C)^{\beta+1}}<0.$$

then f is decreasing.

3 Main Results

For the system (2) or vector field Y_η, we have the following results.

Lemma 1 *The set* $\Gamma=\left\{(u,v)\in\mathbb{R}^2\mid 0<u<1, C<v<1+C\right\}$ *is an invariant region.*

 ISBN: 978-84-617-8694-7

Proof. On the upper boundary of Γ,

$$\frac{dv}{d\tau}(u, 1+C) = S(u-1)(1+C) \leq 0, \qquad u \in [0,1].$$

On the lower boundary of Γ,

$$\frac{dv}{d\tau}(u, C) = SuC \geq 0, \qquad u \in [0,1].$$

On the right boundary of Γ,

$$\frac{du}{d\tau}(1, v) = -Qv^{\beta}(1+C) \leq 0, \qquad v \in [0, 1+C].$$

Therefore, all orbits starting from the interior of Γ cannot cross the boundary. ■

We define the functions $F : [0,1] \times \mathbb{R}_+ \to \mathbb{R}_{+,0}$ and $g : [0,1] \to \mathbb{R}_{+,0}$ such as

$$F(u, C) = \frac{(1-u)u}{u^{\alpha}(u+C)^{\beta}}, \qquad g(u) = (1-u)u^{1-\alpha-\beta}, \ with\,(\alpha, \beta) \in]0,1[\times]0,1].$$

Also, we define $\bar{u}$ as function of C, this is, $\bar{u} : \mathbb{R}_+ \to [0,1]$ such that

$$\bar{u}(C) = \frac{-(\alpha+\beta-1+(2-\alpha)C)+\sqrt{(\alpha+\beta-1+(2-\alpha)C)^2+4(2-\beta-\alpha)C(1-\alpha)}}{2(2-\beta-\alpha)},$$

with $(\alpha, \beta) \in]0,1[\times]0,1]$.

For $(\alpha, \beta) \in]0,1[\times]0,1]$, the following statements are true:

i) For $u \in]0,1]$ fixed, $lim_{C\to 0} F(u, C) = g(u)$.

ii) $\bar{u}(0) = 0$ if $1-\alpha-\beta \leq 0$ and $\bar{u}(0) = \frac{(1-\alpha-\beta)}{(2-\alpha-\beta)}$ if $1-\alpha-\beta > 0$.

iii) $lim_{C\to\infty} F(u, C) = 0$ uniformly in u.

iv) $\frac{dF(\bar{u}(C),C)}{dC} = \frac{\partial F}{\partial u}(\bar{u}(C), C)\frac{d\bar{u}}{dC}(C) + \frac{\partial F}{\partial C}(\bar{u}(C), C) = (0)\frac{d\bar{u}}{dC}(C) - \beta\frac{(1-u)u}{u^{\alpha}(u+C)^{\beta+1}} < 0$, namely, $F(\bar{u}(C), C)$ is decreasing in C.

Ultimately, we defined the following sets of parameters

$$\Delta^{sg(\varepsilon)} = \left\{\eta \in \mathbb{R}^5_+ \mid 1-\alpha = sg(\varepsilon), \beta \leq 1\right\},$$
$$\Upsilon^{sg(\varepsilon)} = \left\{\eta \in \Delta^+ \mid 1-\alpha-\beta = sg(e)\right\},$$
$$\Lambda^{sg(\varepsilon)} = \left\{\eta \in \Delta^+ \mid Q - (1-\alpha-\beta)^{1-\alpha-\beta}(2-\alpha-\beta)^{\alpha+\beta-2} = sg(\varepsilon)\right\},$$

where $\eta = (Q, C, S, \alpha, \beta)$.

Lemma 2 *For $\eta = (Q, C, S, \alpha, \beta) \in \Delta^+$:*

1. *If $\eta \in \Upsilon^-$, $\exists!$ $C_1 > 0$ such that $F(\bar{u}(C_1), C_1) = Q$.*

 ISBN: 978-84-617-8694-7

2. *If* $\eta \in \Upsilon^0 \cap \Lambda^-$, $\exists!\ C_2 > 0$ *such that* $F\left(\bar{u}\left(C_2\right), C_2\right) = Q$;

3. *If* $\eta \in \Upsilon^+ \cap \Lambda^-$, $\exists!\ C_3 > 0$ *such that* $F\left(\bar{u}\left(C_3\right), C_3\right) = Q$.

Proof. Proof of 1. Let $K > 0$. If $\eta \in \Upsilon^-$, for the function g we have

$$g\left(1\right) = 0, \qquad lim_{u \to 0} g\left(u\right) = \infty, \qquad \frac{dg}{du} = \frac{(1-\alpha-\beta)-(2-\alpha-\beta)u}{u^{\alpha+\beta}} < 0,$$

then, there is $u_1 < 1$ such that $K = g\left(u_1\right)$. Let $\varepsilon < u_1 = g^{-1}\left(K\right)$. As g is decreasing, we can choose some $\delta > 0$ such that $\delta < g\left(\varepsilon\right) - K$. Since i), for u fixed, there is $\tilde{C}$ such that

$$g\left(u\right) - F\left(u, C\right) < \delta < g\left(\varepsilon\right) - K, \qquad C < \tilde{C}.$$

We choose $u = \varepsilon$, then $K < F\left(\varepsilon, C\right)$ for $C < \tilde{C}$. By ii), there is $\bar{C}$ such that

$$\bar{u}\left(C\right) < \varepsilon, \qquad C < \bar{C},$$

and since $\bar{u}$ is a maximun, $F\left(\bar{u}\left(C\right), C\right) > F\left(\varepsilon, C\right)$ for all $C < \bar{C}$. Then, we choose $\check{C} = min\left\{\tilde{C}, \bar{C}\right\}$. Thus

$$F\left(\bar{u}\left(C\right), C\right) > F\left(\varepsilon, C\right) > K, \qquad C < \check{C}.$$

Therefore, $lim_{C \to 0} F\left(\bar{u}\left(C\right), C\right) = \infty$. By iii), there is $\hat{C}$ such that $F\left(\bar{u}\left(\hat{C}\right), \hat{C}\right) < Q$. Then, by intermediate value theorem, there is $C_1 \in \left]0, \hat{C}\right[$ such that $F\left(\bar{u}\left(C_1\right), C_1\right) = Q$. By vi), $F\left(\bar{u}\left(C\right), C\right)$ is decreasing, then C_1 is unique.

Proof of 2. Let $\delta > 0$. If $\eta \in \Upsilon^0$, $g\left(u\right) = 1 - u$. We choose $\varepsilon < min\left\{1, \delta\right\}$ and $\theta = \delta - \varepsilon = \delta - 1 + g\left(\varepsilon\right) > 0$. Since i), for u fixed, there is $\tilde{C}$ such that

$$g\left(u\right) - F\left(u, C\right) < \theta = \delta - 1 + g\left(\varepsilon\right) \qquad C < \tilde{C}.$$

We choose $u = \varepsilon$, then $1 - F\left(\varepsilon, C\right) < \delta$ for all $C < \tilde{C}$. By ii), there is $\bar{C}$ such that

$$\bar{u}\left(C\right) < \varepsilon, \qquad C < \bar{C},$$

and since $\bar{u}$ is a maximun, $F\left(\bar{u}\left(C\right), C\right) > F\left(\varepsilon, C\right)$ for all $C < \bar{C}$. Then we choose $\check{C} = min\left\{\tilde{C}, \bar{C}\right\}$. Thus

$$1 - F\left(\bar{u}\left(C\right), C\right) < 1 - F\left(\varepsilon, C\right) < \delta, \qquad C < \check{C}.$$

Therefore, $lim_{C \to 0} F\left(\bar{u}\left(C\right), C\right) = 1$. By iii) there is $\hat{C}$ such that $F\left(\bar{u}\left(\hat{C}\right), \hat{C}\right) < Q$. If $\eta \in \Lambda^-$, $Q < 1$ and by intermediate value theorem, there is $C_2 \in \left]0, \hat{C}\right[$ such that $F\left(\bar{u}\left(C_2\right), C_2\right) = Q$. By vi) $F\left(\bar{u}\left(C\right), C\right)$ is decreasing, then C_2 is unique.

 ISBN: 978-84-617-8694-7

Proof of 3. Let $\delta > 0$. If $\eta \in \Upsilon^+$, for the function g we have

$$g(0) = g(1) = 0, \qquad \frac{dg}{du} = \frac{(1-\alpha-\beta)-(2-\alpha-\beta)u}{u^{\alpha+\beta}}, \qquad \frac{dg}{du}(\hat{u}) = 0 \Leftrightarrow \hat{u} = \frac{(1-\alpha-\beta)}{(2-\alpha-\beta)}, \qquad \frac{d^2g}{du^2} < 0,$$

then g has a maximun at $\hat{u}$ and $g(\hat{u}) = \frac{(1-\alpha-\beta)^{1-\alpha-\beta}}{(2-\alpha-\beta)^{2-\alpha-\beta}}$. $F(\bar{u}(C), C)$ is $C^0([0,\infty[)$ and by ii), $F(\bar{u}(0), 0) = g(\hat{u})$. Since iii), there is $\hat{C}$ such that $F\left(\bar{u}\left(\hat{C}\right), \hat{C}\right) < Q$. Thus, if $\eta \in \Lambda^-$, $Q < g(\hat{u})$ by intermediate value theorem, there is $C_3 \in \left]0, \hat{C}\right[$ such that $F(\bar{u}(C_3), C_3) = Q$. By vi), $F(\bar{u}(C), C)$ is decreasing, then C_3 is unique. ■

Theorem 3 *The existence of singularities in* Γ *is given by:*

1. *For* $\eta = (Q, C, S, \alpha, \beta) \in \Delta^0$,
   (a) *if* $QC^\beta \geq 1$, *there are not singularities;*
   (b) *if* $QC^\beta < 1$, $Sing(Y_\eta(\Gamma)) = \{(u_0, u_0 + C)\}$.
2. *For* $\eta = (Q, C, S, \alpha, \beta) \in \Delta^+$,
   (a) *either if* $\eta \in \Upsilon^-$ *with* $C > C_1$ *or if* $\eta \in \Upsilon^0 \cap \Lambda^-$ *with* $C > C_2$ *or if* $\eta \in \Upsilon^+ \cap \Lambda^-$ *with* $C > C_3$ *or if* $\eta \in \Upsilon^{0,+} \cap \Lambda^{0,+}$, *there are not singularities;*
   (b) *either if* $\eta \in \Upsilon^-$ *with* $C = C_1$ *or if* $\eta \in \Upsilon^0 \cap \Lambda^-$ *with* $C = C_2$ *or if* $\eta \in \Upsilon^+ \cap \Lambda^-$ *with* $C = C_3$, $Sing(Y_\eta(\Gamma)) = \{(u_1, u_1 + C)\}$;
   (c) *either if* $\eta \in \Upsilon^-$ *with* $C < C_1$ *or if* $\eta \in \Upsilon^0 \cap \Lambda^-$ *with* $C < C_2$ *or if* $\eta \in \Upsilon^+ \cap \Lambda^-$ *with* $C < C_3$, $Sing(Y_\eta(\Gamma)) = \{(u_2, u_2 + C), (u_3, u_3 + C)\}$.

Proof.

Proof of 1. Since the function f is decreasing if $\alpha = 1$, there is u_0 satisfying $f(u) = Q$ iff $f(0) = \frac{1}{C^\beta} > Q$. Moreover, u_0 is unique. This proves a) and b).

Proof of 2. (u_e, v_e) is solution of (3) iff is solution of

$$f(u) = Q, \qquad v = u + C. \tag{4}$$

and by Lemma 2, if $\eta \in \Upsilon^- \cup \left(\Upsilon^0 \cap \Lambda^-\right) \cup (\Upsilon^+ \cap \Lambda^-)$ there is C_i, with $i \in \{1, 2, 3\}$ depending on the case, such that $F(\bar{u}(C_i), C_i) = Q$. Moreover, f has a maximum value in $\bar{u}$.

a. If $C > C_i$ and as by the vi) property, $F(\bar{u}(C), C)$ is decreasing, we obtain

$$f(u) \leq f(\bar{u}) = F(\bar{u}(C), C) < F(\bar{u}(C_i), C_i) = Q, \qquad u \in [0, 1].$$

Thus, any u can verify (4). If $\eta \in \Upsilon^0 \cap \Lambda^{0,+}$, then $g(u) = 1 - u$ and $Q \geq 1$. Then

$$f(u) < g(u) \leq Q, \qquad u \in]0, 1[,$$

 ISBN: 978-84-617-8694-7

y thus, any u can verify (4). If $\eta \in \Upsilon^{+} \cap \Lambda^{0,+}$, then

$$f(u) < g(u) \leq g(\hat{u}) = \frac{(1-\alpha-\beta)^{1-\alpha-\beta}}{(2-\alpha-\beta)^{2-\alpha-\beta}} \leq Q, \qquad u \in]0,1[,$$

Thus, any u can verify (4).

b. If $C = C_i$ then

$$f(u) \leq f(\bar{u}) = F(\bar{u}(C), C) = Q, \qquad u \in [0,1].$$

Thus, only $u_1 = \bar{u}$ can verify (4).

c. If $C < C_i$ and as by vi), $F(\bar{u}(C), C)$ is decreasing, then

$$Q = F(\bar{u}(C_i), C_i) < F(\bar{u}(C), C) = f(\bar{u}).$$

As $f(0) = f(1) = 0$ and f has a maximun value in $\bar{u}$, there are $u_2, u_3 \in]0,1[$ such that $u_2 < \bar{u} < u_3$ and $f(u_2) = f(u_3) = Q$. ■

Proposition 4 *The singularity* $(0,0)$ *is an unstable node.*

Proof. The Jacobian matrix at $(0,0)$ is

$$DY_{\eta}(0,0) = \begin{pmatrix} C & 0 \\ 0 & SC \end{pmatrix}$$

with $detDY_{\eta}(0,0) = SC^2 > 0$ and $traDY_{\eta}(0,0) = C(1+S) > 0$, then the point is an unstable node. ■

Proposition 5 *The singularity* $(1,0)$ *is a saddle point.*

Proof. The Jacobian matrix at $(1,0)$ is

$$DZ_{\eta}(1,0) = \begin{pmatrix} -1-C & -Q(1+C) \\ 0 & \beta S(1+C) \end{pmatrix}$$

with $detDZ_{\eta}(1,0) = -(1+C)\beta S(1+C) < 0$ and therefore, $(1,0)$ is a saddle point. ■

Proposition 6 *If* $\eta = (Q, C, S, \alpha, \beta) \in \Delta^{0}$, *i. e.* $\alpha = 1$ *for* $(0,C)$ *we have:*

1. *if* $QC^{\beta} < 1$, $(0,C)$ *is a saddle point.*
2. *if* $QC^{\beta} > 1$, $(0,C)$ *is stable.*

 ISBN: 978-84-617-8694-7

Proof. The Jacobian matrix at $(0,C)$ is

$$DY_\eta(0,C)=\begin{pmatrix}\left(1-QC^\beta\right)C & 0\\ SC & -SC\end{pmatrix}$$

with $detDY_\eta(0,C)=\left(QC^\beta-1\right)SC^2$ and $traDY_\eta(0,C)=C\left(1-QC^\beta-S\right)$.

(1) If $QC^\beta<1$, $detDY_\eta(0,C)<0$ and $(0,C)$ is a saddle point.

(2) If $QC^\beta>1$, $traDY_\eta(0,C)<0<detDY_\eta(0,C)$ and $(0,C)$ is stable. ■

Lemma 7 *If $\eta\in\Delta^+$, there is not uniqueness of solutions at points in the positive v-axis, different from the singularity $(0,C)$.*

Proof. If $\eta\in\Delta^+$, $\alpha<1$. We consider the change of variables and the time rescaling given by

$$\phi:\mathbb{R}^2_+\times\mathbb{R}\to\mathbb{R}^2_+\times\mathbb{R},\qquad \phi(z,w,\kappa)=\left(z^{\frac{2}{1-\alpha}},w,z\kappa\right)=(u,v,\tau),$$

with

$$detD\phi(z,w,\kappa)=\left|\begin{pmatrix}\frac{2}{1-\alpha}z^{\frac{2}{1-\alpha}-1} & 0 & 0\\ 0 & 1 & 0\\ \kappa & 0 & z\end{pmatrix}\right|=\frac{2}{1-\alpha}z^{\frac{2}{1-\alpha}}>0\ .$$

and $\phi(0,C)=(0,C)$. By the rule chain $\frac{dz}{d\kappa}=\frac{1-\alpha}{2}z^{1-\frac{1-\alpha}{2}}\frac{du}{d\tau}\frac{d\tau}{d\kappa}$ and $\frac{dw}{d\kappa}=\frac{dv}{d\tau}\frac{d\tau}{d\kappa}$, in the new coordinates, the new vector field is

$$Z_\eta:\begin{cases}\frac{dz}{d\kappa} &= \frac{1-\alpha}{2}\left(\left(1-z^{\frac{2}{1-\alpha}}\right)z^2-Qw^\beta\right)\left(z^{\frac{2}{1-\alpha}}+C\right)\\ \frac{dw}{d\kappa} &= S\left(z^{\frac{2}{1-\alpha}}+C-w\right)wz\end{cases}. \tag{5}$$

Clearly $Z_\eta(0,w)=-\frac{1-\alpha}{2}QCw^\beta\frac{\partial}{\partial z}$. Then, for $w>0$, the vector field Z_η is orthogonal to the w-axis. If $0<w_0\neq C$ and γ is the orbit of vector field (5) with initial condition in the point $(0,w_0)$, then $\gamma^*=\gamma-\{(0,w_0)\}$ is also an orbit to the vector field (5).

As ϕ is a homeomorphism, systems (2) and (5) are C^0-equivalent in the first quadrant $\mathbb{R}^2_+$; hence if $\phi(\gamma^*)$ is an orbit of (2), by continuity $\phi(\gamma)$ is an orbit of (2) that is tangent to the vector field Y_η a the point $(0,w_0)$. But the v-axis, $u=0$ is clearly an invariant set and $Y_\eta(0,w_0)=S(C-w_0)w_0\frac{\partial}{\partial v}\neq 0$. Thus, for the point $(0,w_0)$, at least two orbits exist. ■

Evaluating the Jacobian matrix of vector field Y_η at the equilibrium point (u_e,v_e), we have

$$DY_\eta(u_e,v_e)=\begin{pmatrix}\left(1-2u_e-\alpha(1-u_e)\right)(u_e+C) & -\beta(1-u_e)u_e\\ S(u_e+C) & -S(u_e+C)\end{pmatrix},$$

with trace and determinant given by

$$\begin{aligned} detDY_\eta(u_e,v_e) &= \left((2-\alpha-\beta)u_e^2+(\alpha+\beta-1+(2-\alpha)C)u_e-C(1-\alpha)\right)S(u_e+C)\\ &= q(u_e)S(u_e+C),\\ traDY_\eta(u_e,v_e) &= (1-\alpha+(\alpha-2)u_e-S)(u_e+C).\end{aligned}$$

 ISBN: 978-84-617-8694-7

In the parameters space, in order to obtain a simpler description of the bifurcation diagram of (2), for $\eta = (Q, C, S, \alpha, \beta)$, we define the sets

$$\Theta^{sg(\varepsilon)} = \{\eta \in \Delta^+ \mid S - 1 + \alpha + (2-\alpha)\,\bar{u} = sg(\varepsilon), \bar{u} = \bar{u}(C, \alpha, \beta)\},$$
$$\Pi^{sg(\varepsilon)} = \left\{\eta \in \Delta^+ \mid Q - (1+S)(1-\alpha-S)^{1-\alpha}(2-\alpha)^{\alpha+\beta-2}(C(2-\alpha)+1-\alpha-S)^{-\beta} = sg(\varepsilon)\right\}.$$

Theorem 8 *The nature of singularities in* Γ *is given by:*

1. *The singularity* $p_0 = (u_0, u_0 + C)$ *is stable.*
2. *For the singularity* $p_1 = (u_1, u_1 + C)$ *we have*
   (a) *if* $\eta \in \Theta^-$, p_1 *is a unstable saddle-node;*
   (b) *if* $\eta \in \Theta^0$, p_1 *is a cusp point;*
   (c) *if* $\eta \in \Theta^+$, p_1 *is a stable saddle-node.*
3. *The singularity* $p_2 = (u_2, u_2 + C)$ *is a saddle-point*
4. *For the singularity* $p_3 = (u_3, u_3 + C)$ *we have*
   (a) *either if* $\eta \in \Theta^- \cap \Pi^-$ *or if* $\eta \in \Theta^0 \cup \Theta^+$, p_3 *is stable;*
   (b) *if* $\eta \in \Theta^- \cap \Pi^0$, *a Hopf bifurcation occurs at* p_3*;*
   (c) *if* $\eta \in \Theta^- \cap \Pi^+$, p_3 *is unstable.*

Proof. Proof of 1. The trace and the determinat of the Jacobian matrix $DY_\eta(p_0)$ are

$$detDY_\eta(p_0) = \left((1-\beta)u_0^2 + (\beta + C)u_0\right) S(u_0 + C) > 0,$$
$$traDY_\eta(p_0) = -(u_0 + S)(u_0 + C) < 0,$$

thus, p_1 is stable.

Proof of 2. Since $u_1 = \bar{u}$ y $q(\bar{u}) = 0$, the trace and the determinant of $DY_\eta(p_1)$ are

$$detDY_\eta(p_1) = p(\bar{u})\, S(\bar{u} + C) = 0,$$
$$traDY_\eta(p_1) = (1 - \alpha + (\alpha - 2)\,\bar{u} - S)(\bar{u} + C),$$

then, the sign of $traDY_\eta(p_1)$ depends of the factor $\bar{S} = 1 - \alpha + (\alpha - 2)\,\bar{u} - S$. If $\eta \in \Theta^-$ or Θ^+, $\bar{S}$ is positive o negative, respectively, therefore, we obtain (a) and (c). In the particular case $\eta \in \Theta^0$, $\bar{S} = 0$, and the Jacobian matrix at p_1 is

$$DY_\eta(p_1) = S(\bar{u} + C)\begin{pmatrix} 1 & -1 \\ 1 & -1 \end{pmatrix}.$$

 ISBN: 978-84-617-8694-7

The Jordan form matrix corresponding to this Jacobian matrix is

$$\begin{pmatrix} 0 & S(\bar{u}+C) \\ 0 & 0 \end{pmatrix}.$$

and we have the Bogdanov–Takens bifurcation [16], that is, the point p_1 is a cusp point.

Proof of 3. Since $q(u_2) < 0$, $detDY_\eta(p_2) = q(u_2)S(u_2+C) < 0$ and p_2 is a saddle point.

Proof of 4. Since $q(u_3) > 0$, $detDY_\eta(p_3) = q(u_3)S(u_3+C) > 0$ and the nature of p_3 depends on the trace, given by

$$traDY_\eta(p_3) = (1-\alpha+(\alpha-2)u_3 - S)(u_3+C),$$

i.e., depends on the factor $\bar{S} = 1-\alpha+(\alpha-2)u_3 - S$. If $\eta \in \Theta^0 \cup \Theta^+$ and since $\bar{u} < u_3$, is true that

$$1-\alpha-(2-\alpha)u_3 < 1-\alpha-(2-\alpha)\bar{u} \leq S,$$

hence, $\bar{S} < 0$ and then p_3 is stable. If $\eta \in \Theta^-$, $\bar{u} < \frac{1-\alpha-S}{2-\alpha}$ and since f is decreasing in $]\bar{u}, 1[$, we have that

$$\eta \in \Pi^- \Rightarrow f(u_3) = Q < \frac{(1+S)(1-\alpha-S)^{1-\alpha}(2-\alpha)^{\alpha+\beta-2}}{(C(2-\alpha)+1-\alpha-S)^\beta} = f\left(\frac{1-\alpha-S}{2-\alpha}\right) \Rightarrow \frac{1-\alpha-S}{2-\alpha} < u_3 \Rightarrow \bar{S} < 0,$$

$$\eta \in \Pi^0 \Rightarrow f(u_3) = Q = \frac{(1+S)(1-\alpha-S)^{1-\alpha}(2-\alpha)^{\alpha+\beta-2}}{(C(2-\alpha)+1-\alpha-S)^\beta} = f\left(\frac{1-\alpha-S}{2-\alpha}\right) \Rightarrow \frac{1-\alpha-S}{2-\alpha} = u_3 \Rightarrow \bar{S} = 0,$$

$$\eta \in \Pi^+ \Rightarrow f(u_3) = Q > \frac{(1+S)(1-\alpha-S)^{1-\alpha}(2-\alpha)^{\alpha+\beta-2}}{(C(2-\alpha)+1-\alpha-S)^\beta} = f\left(\frac{1-\alpha-S}{2-\alpha}\right) \Rightarrow \frac{1-\alpha-S}{2-\alpha} > u_3 \Rightarrow \bar{S} > 0.$$

Thus, we obtain (a) and (c). The proof of (b) follows from that $traDY_\eta(p_3)$ changes sign and the $detDY_\eta(p_3)$ is always positive. Moreover, verifying the transversality condition [12], we have

$$\frac{\partial(traDY_\eta(p_3))}{\partial S} = -(u_2+C) < 0. \ .$$

Therefore, at the point $p_3 = \left(\frac{1-\alpha-S}{2-\alpha}, \frac{1-\alpha-S}{2-\alpha}+C\right)$ occurs a Hopf bifurcation [16] for the bifurcation value $Q = \frac{(1+S)(1-\alpha-S)^{1-\alpha}(2-\alpha)^{\alpha+\beta-2}}{(C(2-\alpha)+1-\alpha-S)^\beta}$. ■

Remark. The existence of homoclinic curve can be proved, which is generated by the intersection of the stable and unstable manifold of the saddle point p_2.

Proposition 9 *If $\eta \in \Delta^0$, there are not limit cycle in $\mathbb{R}^2_+$.*

Proof. Let $g(u,v) = \frac{1}{u(u+C)v}$. Clearly $g \in C^1\left(\mathbb{R}^2_+\right)$. Then

$$\begin{aligned} div(g(u,v)Y_\eta(u,v)) &= \frac{d}{du}\left(\frac{\left((1-u)u - Qu^\alpha v^\beta\right)(u+C)}{u(u+C)v}\right) + \frac{d}{dv}\left(\frac{S(u+C-v)v}{u(u+C)v}\right), \qquad \forall\,(u,v) \in \mathbb{R}^2_+. \\ &= -\frac{1}{v} - \frac{Q(\alpha-1)}{u^{2-\alpha}v^{1-\beta}} - \frac{S}{u(u+C)} \end{aligned}$$

 ISBN: 978-84-617-8694-7

If $\alpha = 1$, $div\left(g\left(u,v\right)Y_{\eta}\left(u,v\right)\right) = -\frac{1}{v} - \frac{S}{u(u+C)} < 0$. Therefore, by the Dulac criteria [16], there are not limit cycles in $\mathbb{R}^2_+$. ■

If $\beta < 1$, the vector fied Y_{η} is non-lipchitzian in the u-axis. However, we have the following property:

Proposition 10 *If $\beta < 1$, Y_{η} is topologically equivalent to a vector field of Kolmogorov type and lipchitzian in the u-axis.*

Proof. Making a change in variables $\varphi : \mathbb{R}^2_+ \times \mathbb{R} \to \mathbb{R}^2_+ \times \mathbb{R}$, such as

$\varphi\left(z,w\right) = \left(z, w^{\frac{1}{\beta}}\right) = (u,v)$ with

$\mathrm{detD}\varphi\left(z,w\right) = \left|\begin{pmatrix} 1 & 0 \\ 0 & \frac{1}{\beta}w^{\frac{1-\beta}{\beta}} \end{pmatrix}\right| = \frac{1}{\beta}w^{\frac{1-\beta}{\beta}} > 0.$

Then, φ is a diffeomorphism preserving time orientation; hence, in the new coordinates, the topologically equivalent vector field to Y_{η} is

$$Z_{\eta} : \begin{cases} \frac{dz}{d\tau} &= \left(\left(1-z\right)z - Qz^{\alpha}w\right)\left(z+C\right) \\ \frac{dw}{d\tau} &= \beta S\left(z + C - w^{\frac{1}{\beta}}\right)w \end{cases},$$

which it is of Kolmogorov type and lipchitzian on the z-axis (u-axis). ■

Acknowledgements

This work has been partially supported by by DIEA-PUCV 124.730/2012 project

References

[1] P. Aguirre, E. González-Olivares, E. Sáez, *Three limit cycles in a Leslie-Gower predator-prey model with additive Allee effect*, SIAM Journal on Applied Mathematics 69(5) (2009) 1244-1269.

[2] A. D. Bazykin, *Nonlinear Dynamics of interacting populations*, World Scientific Publishing Co. Pte. Ltd., 1998.

[3] C. Chicone, *Ordinary differential equations with applications* (2nd edition), Texts in Applied Mathematics 34, Springer, 2006.

[4] C. W. Clark, *Mathematical Bioeconomics: The optimal management of renewable resources* (2nd edition), John Wiley and Sons, New York, 1990.

[5] F. Dumortier. J. Llibre and J. C. Artés,, *Qualitative Theory of Planar Differential Systems*, Springer, Berlin, 2006.

 ISBN: 978-84-617-8694-7

[6] H. I. Freedman, *Deterministic Mathematical Model in Population Ecology*, Marcel Dekker, New York, 1980.

[7] E. González-Olivares, E. Sáez, E. Stange and I. Szantó, *Topological description of a non-differentiable bio-economics model*, Rocky Mountain Journal of Mathematics 35(4) 1133-1155.

[8] E. González-Olivares, P. Tintinago-Ruiz, and A.Rojas-Palma, *A Leslie-Gower-type predator-prey model with sigmoid functional response*, International Journal of Computer Mathematics 92 (2014)1895-1909.

[9] E. González-Olivares, L. M. Gallego-Berrio, B. González-Yañez, and A. Rojas-Palma, *Consequences of weak Allee effect on prey in the May-Holling-Tanner predator-prey model*, Mathematical Methods in the Applied Sciences 38 (2015) 5183-5196.

[10] E. González-Olivares, B. González-Yañez and A. Rojas-Palma, *Multiple limit cycles in a Leslie-Gower type predator-prey model considering weak Allee effect on prey*, Nonlinear Analysis: Modelling and Control 22(3) (2017) 347-365.

[11] E. González-Olivares, K Vilches-Ponce and A. Rojas-Palma, *A class of predator-prey models with a non differentiable functional response*, submitted (2015).

[12] F. Guckenheimer and P. Holmes, *Nonlinear oscillations, dynamical systems, and bifurcations of vector fields*, Springer, New York,1983.

[13] P. H. Leslie and J. C. Gower, *The properties of astochastic model for the predator-prey type of interaction between two species*, Biometrica 47 (1960) 219-234.

[14] R. M. May, *Stability and complexity in model ecosystems* (2nd edition), Princeton University Press (2001).

[15] J. Mena-Lorca, E. González-Olivares, B. González-Yañez, *The Leslie-Gower predator-prey model with Allee effect on prey: A simple model with a rich and interesting dynamics*, in: R. Mondaini (Ed.), Proceedings of the 2006 International Symposium on Mathematical and Computational Biology BIOMAT 2006, E-papers Serviços Editoriais Ltda., Rio de Janeiro, 2007 105-132.

[16] L. Perko, *Differential Equations and Dynamical Systems* (3rd ed), Springer, New York, 2001.

[17] P. Turchin, *Complex population dynamics. A theoretical/empirical synthesis*, Monographs in Population Biology 35, Princeton University Press (2003).

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

Difference method of fourth order accuracy for the Laplace equation with multilevel nonlocal condition

Adiguzel A. Dosiyev

Department of Mathematics, Near East University
PO Box 99138, Nicosia, TRNC, Mersin 10, Turkey

emails: adiguzel.dosiyev@emu.edu.tr

Abstract

We consider the multipoint nonlocal boundary value problem for the two-dimensional Laplace equation in a rectangular domainThe solution of this problem is defined as a 9-point finite difference solution, with the fourth order gluing operator of the local Dirichlet boundary value problem, by constructing a special method to find a function as the boundary value on the side of the rectangle, where the nonlocal condition is given. Numerical experiments are illustrated to support the analysis made.
Key words: Bitsadze-Smarskii problem, Elliptic equation, Nonlocal boundary value problems, Difference scheme
MSC 2000: AMS codes: 35A35, 65N06, 65N15

1 Introduction

The paper of A.V. Bitsadze and A.A. Samarskii [1] stated a nonlocal boundary value problem for finding a function on the rectangle, for the given continuous boundary values on three sides of the rectangle. This function has the following properties: it is a solution of Poisson's equation on the open rectangle, it is continuous on the closed rectangle, also on the fourth side of the rectangle and on the midline of the rectangle, which is parallel to this side the function takes coinciding values.

Different generalizations of the nonlocal conditions and their approximate solutions were investigated by numerous authors (see [2], [3], [4], and references therein). As follows from the existing papers, difficulties arise in the analysis of both the exact and numerical solutions due to the existence of the nonlocal conditions.

In [5]-[7] a new constructive method for the solution of the Poisson equation with nonlocal boundary condition have been proposed and justified.

In this paper by the 9-point approximation of the Laplace equation with the fourth order interpolation operator, the approximate method used in [5]-[7] is developed. The 9-point solution of the multilevel nonlocal problem is defined as the 9-point solution of the Dirichlet problem by finding a function given as the boundary value on the side of the rectangle where the nonlocal condition was given. Finally, the numerical experiments are illustrated to support the obtained theoretical results.

2 Multipoint Nonlocal Boundary Value Problem

Let

$$R = \{(x, y) : 0 < x < 1, 0 < y < 2\} \tag{1}$$

be an open rectangle, γ^p, $p = 1, 2, 3, 4$ be its sides including the ends, numerated in the clockwise direction, starting with the side which lies on the y-axis, and let $\gamma = \cup_{p=1}^{4}\gamma^p$ be the boundary of R, $\overline{R} = R \cup \gamma$.

Let $\eta_1, \eta_2, ..., \eta_m$, and $\alpha_1, \alpha_2, ..., \alpha_m$ be given numbers satisfying for some fixed number $\delta > 0$ the inequalities

$$0 < \delta \le \eta_1 < \eta_2 < ... < \eta_m \le 2 - \delta < 2, \tag{2}$$

$$\Big(1 - \frac{\eta_1}{2}\Big) \sum_{k=1}^{m} |\alpha_k| < 1. \tag{3}$$

We consider the following nonlocal boundary value problem on R :

$$\Delta u = 0 \ \text{ on } \ R, \ u = 0 \ \text{ on } \gamma^1 \cup \gamma^3, \ u = \tau(x) \ \text{ on } \gamma^2, \tag{4}$$

$$\sum_{k=1}^{m} \alpha_k u(x, \eta_k) = u(x, 0), \quad 0 \le x \le 1, \tag{5}$$

where $\Delta \equiv \partial^2/\partial x^2 + \partial^2/\partial y^2$ is the Laplace operator, $\tau(x)$ is the given continuous function on $[0, 1]$ and $\tau(0) = \tau(1) = 0$.

The existence and uniqueness of the classical solution $u \in C(\overline{R}) \cap C^2(R)$ of the problem (4), (5) is given in [8], [3].

3 Approximate solution of the nonlocal problem by the finite difference method

We say that $\phi \in C^{k,\lambda}(D)$, if ϕ has k-th derivatives on D satisfying the Hölder condition with exponent λ. Let $\tau(x) \in C^{4,\lambda}(\gamma^2)$, $0 < \lambda < 1$, and $\tau^{(2q)}(0) = \tau^{(2q)}(1)$, $q = 0, 1, 2$.

 ISBN: 978-84-617-8694-7

We assign a square mesh D_h obtained with the lines $x, y = 0, h, 2h, \ldots$, where $h = \frac{1}{N}$ is the step size, $N > 2$ is an integer, such that h is less than half of the minimum length of the intervals $[0, \eta_1]$, $[\eta_1, \eta_2], ..., [\eta_m, 2]$, and we denote by j_l the number for which

$$j_l h \le \eta_l < (j_l + 1)h.$$

We denote by $R_h = D_h \cap R$, γ_h^p is the set of grids on γ^p, $p = 1, ..., 4$, $\gamma_h = \cup_{p=1}^4 \gamma_h^p$, $\overline{R}_h = R_h \cup \gamma_h$, and for each integer l, $1 \le l \le m$, the set of intersection points of the line $y = \eta_l$ with the grid lines $x = ih$, $i = 1, ..., N-1$, is denoted by Y_l^h; $\widetilde{R}_h = \overline{R}_h \cup (\cup_l^m Y_l^h)$.

On the set $Y^h = \cup_{l=1}^m Y_l^h$, we introduce the fourth order liear matching operator S^4 from the condition that the expression $S^4(F_3)$ gives the exact value of any third order harmonic polynomial $F_3(x, y)$ at each point $P \in Y^h$. Let P_0 be the grid node closest to P. We place the origin of the rectangular system of coordinates at the node P_0 and direct the positive axis of x along the grid line so that $P = P(\sigma h, 0)$, $0 \le \sigma < 1$. In the neighborhood $|z| < 2h$, where $z = x + iy$ in the new coordinate system, for the harmonic function $u \in C^{4,0}$ on the basis of Taylor's formula, we have

$$u(x, y) = F_3(x, y) + O(h^4), \tag{6}$$

where

$$F_3(x, y) = \sum_{k=0}^{3} a_k \,\mathrm{Re}\ z^k + \sum_{k=1}^{3} b_k \,\mathrm{Im}\ z^k \tag{7}$$

$$a_0 = u(0,0), a_1 = \frac{\partial u(0,0)}{\partial x}, a_2 = \frac{1}{2}\frac{\partial^2 u(0,0)}{\partial x^2}, a_3 = \frac{1}{3!}\frac{\partial^3 u(0,0)}{\partial x^3};$$
$$b_1 = \frac{\partial u(0,0)}{\partial y}, b_2 = \frac{1}{2}\frac{\partial^2 u(0,0)}{\partial x \partial y}, b_3 = \frac{1}{3!}\frac{\partial^3 u(0,0)}{\partial x^2 \partial y}.$$

We take the points $P_1(h, 0)$, $P_2(h, h)$, and $P_3(0, h)$ to find numerical coefficients μ_0', μ_1', μ_2' and μ_3' such that the representation

$$u_0 = \mu_0' u + \mu_1' u_1 + \mu_2' u_2 + \mu_3' u_3 \tag{8}$$

is satisfied for the harmonic polynomials $\mathrm{Re}\, z^n$, $z = x + iy$, $n = 0, 1, 2, 3$, where $u = u(P)$, $u_k = u(P_k)$, $k = 0, 1, 2, 3$. We then have

$$\mu_0' + \mu_1' + \mu_2' + \mu_3' = 1, \quad \sigma\mu_0' + \mu_1' + \mu_2' = 0, \tag{9}$$
$$\sigma^2\mu_0' + \mu_1' - \mu_3' = 0, \quad \sigma^3\mu_0' + \mu_1' - 2\mu_2' = 0. \tag{10}$$

Solving system (9), (10), we obtain $\mu_0' = 3\lambda_0$, $\mu_1' = -\sigma(2 + \sigma^2)\lambda_0$, $\mu_2' = -\sigma(1 - \sigma^2)\lambda_0$, $\mu_3' = -\sigma(2 - \sigma)(1 - \sigma)\lambda_0$, where $\lambda_0 = 1/(3 - 5\sigma + 3\sigma^2 - \sigma^3)$.

Now, we take the nodal points $P_4(h,-h)$ and $P_5(0,-h)$ respectively symmetric to the points P_2 and P_3 with respect to the x axis. Since Im $z^k = 0$, $k = 1,2,3$ for $y = 0$ and is odd with respect to y, and Re z^k, $k = 0,1,2,3$ is even with respect to y, from (7) and (8) we obtain the expression

$$S^4 u \equiv \sum_{k=0}^{5} \mu_k u_k, \tag{11}$$

which, on the basis of $(6)-(7)$, gives the exact value of the harmonic polynomial $F_3(x,y)$ at the point P, where

$$\mu_0 = 1/\mu_0',\ \mu_1 = -\mu_1'/\mu_0',\ \mu_q = \mu_{q+2} = -\mu_q'/2\mu_0',\ q = 2,3. \tag{12}$$

It is easy to check that

$$\mu_0 > 0,\ \mu_q \geq 0, q \neq 0;\quad \sum_{k=0}^{5} \mu_k = 1. \tag{13}$$

More general construction of the fourth order matchinh operator S^4 see [9].

Let $[0,1]_h = \left\{x = x_i,\ x_i = ih,\ i = 0,1,...,N;\ h = \frac{1}{N}\right\}$, be the set of nodes on the interval $[0,1]$ with the step size h. Let C_h^0 be the set of grid functions f_h on $[0,1]_h$ for which $f_h(0) = f_h(1) = 0$. We define the norm $\|f_h\|_{C_h^0} = \max_{x\in[0,1]_h} |f_h|$. It is obvious that the space C_h^0 is complete equipped with this norm.

Let $\widetilde{v}_h$ be a solution of the following system of equations defined on $\widetilde{R}_h$:

$$\widetilde{v}_h = B\widetilde{v}_h \ \text{ on } R_h, \tag{14}$$

$$\widetilde{v}_h = S^4\widetilde{v}_h \ \text{ on } Y_l^h,\ l = 1,2,...,m, \tag{15}$$

$$\widetilde{v}_h = \tau_h \ \text{ on } \gamma_h^2,\ \ \widetilde{v}_h = 0 \ \text{ on } \gamma_h \setminus \gamma_h^2, \tag{16}$$

where τ_h is the trace of τ on γ_h^2, S^4 be the operator defined by (11), and

$$\begin{aligned} Bu(x,y) = {} & (u(x+h,y) + u(x-h,y) + u(x,y+h) + u(x,y-h))/5 \\ & + (u(x+h,y+h) + u(x+h,y-h) + u(x-h,y+h) + u(x-h,y-h))/20. \end{aligned}$$

On the basis of maximum principle, the problem (14)-(16) has a unique solution.

We define

$$\widetilde{\varphi}_h = \widetilde{\varphi}_h(x) = \sum_{l=1}^{m} \alpha_l \widetilde{v}_h(x,\eta_l) \in C_h^0,\ \ x \in [0,1]_h\,, \tag{17}$$

where $\widetilde{v}_h(x,\eta_l)$ are the solution values of the system $(14)-(16)$ on Y_l^h, $l = 1,2,...,m$.

We consider the following problem on $\widetilde{R}_h$

$$w_h = Aw_h \ \text{ on } R_h, \tag{18}$$

$$w_h = S^4 w_h \ \text{ on } Y_l^h,\ l = 1,2,...,m, \tag{19}$$

$$w_h = 0 \ \text{ on } \gamma_h^m,\ m = 1,2,3,\ \ w_h = f_h \ \text{ on } \gamma_h^4, \tag{20}$$

 ISBN: 978-84-617-8694-7

where $f_h \in C_h^0$, is an arbitrary function. On the basis of (13) and the maximum principle, for any f_h problem $(18)-(20)$ has a unique solution.

We introduce a linear operator $B_l^h : C_h^0 \to C_h^0$, and let for any grid function $f_h = f_h(x) \in C_h^0$

$$B_l^h f_h \equiv w_h(x, \eta_l) \in C_h^0,\ l = 1, 2, ..., m, \tag{21}$$

where w_h is the solution of the problem (18)-(20).

On the basis of properties (13) of the matching operator S^4 and maximum principle, we obtain

$$\begin{aligned}\left\|B_l^h f_h\right\|_{C_h^0} &\le \frac{1}{2}\left\|f_h\right\|_{C_h^0}(\mu_1+\mu_2+\mu_4)(2-(j_l+1)h)\\ &+\frac{1}{2}\left\|f_h\right\|_{C_h^0}(\mu_0+\mu_3+\mu_5)(2-j_l h)\\ &=\frac{1}{2}\left\|f_h\right\|(2-\eta_l),\ l=1,2,...,m,\end{aligned} \tag{22}$$

i.e., the norm of operator B_l^h does not exceed $q_l = 1 - \frac{\eta_l}{2}$, $l = 1, 2, ..., m$.

Let $\widetilde{\psi}_{l,h}$, $l = 1, 2, ..., m$ be the solution of the system of equations

$$\widetilde{\psi}_{l,h} = B_l^h\left(\widetilde{\varphi}_h + \sum_{k=1}^{m}\alpha_k\widetilde{\psi}_{k,h}\right),\ l = 1, 2, ..., m, \tag{23}$$

where $\widetilde{\varphi}_h$ is defined by (17).

We seek the solution of system (23) by the following fixed-point iteration´

$$\widetilde{\psi}_{l,h}^0 = 0,\quad \widetilde{\psi}_{l,h}^n = B_l^h\left(\widetilde{\varphi}_h + \sum_{k=1}^{m}\alpha_k\widetilde{\psi}_{k,h}^{n-1}\right),\ l = 1, 2, ..., m;\ n = 1, 2, ... \tag{24}$$

On the basis of inequality (22) follows that the iteration converges to the unique solution of the equation (23).

We define the function

$$\widetilde{f}_h^n = \widetilde{\varphi}_h + \sum_{l=1}^{m}\alpha_l\widetilde{\psi}_{l,h}^n,\quad \text{on } \gamma_h^4, \tag{25}$$

where $\widetilde{\varphi}_h$ is defined by (17), $\widetilde{\psi}_{l,h}^n$ is the n-th element of the sequence (24)

For the approximate solution of problem (4), (5), we take the solution of the following finite difference problem

$$\widetilde{u}_h^n = A\widetilde{u}_h^n \ \text{ on } R_h,\quad \widetilde{u}_h^n = 0 \ \text{ on } \gamma_h^m,\ m = 1, 3\ \ \widetilde{u}_h = \tau_h \ \text{ on } \gamma_h^2, \tag{26}$$

$$\widetilde{u}_h^n = \widetilde{f}_h^n \ \text{ on } \gamma_h^4, \tag{27}$$

where $\widetilde{f}_h^n$ is defined by (25).

The following Theorem is proved:

 ISBN: 978-84-617-8694-7

Theorem 1 *Let the boundary function $\tau(x)$ in the problem (4), (5) be from the Hölder classes $C^{4,\lambda}(\gamma^2)$ and $\tau^{(2q)}(0) = \tau^{(2q)}(1) = 0$, $q = 0, 1, 2$. The next estimation holds*

$$\max_{(x,y)\in\overline{R}_h} |\widetilde{u}_h^n - u_h| \le ch^4 + c_1 \frac{q^{n+1}}{1-q}, \ n \ge 2, \tag{28}$$

where $\widetilde{u}_h^n$ is a solution of the problem (26), (27), u_h is the trace of the exact solution of the nonlocal problem (4), (5), c and c_1 are constants independent of n and h,
$q = \max\left\{1 - \frac{\eta_1}{2}, \left(1 - \frac{\eta_1}{2}\right)\sum_{k=1}^{m} |\alpha_k|\right\} < 1.$

4 Numerical results

Let

$$R = \{(x, y) : 0 < x < 1, 0 < y < 2\}.$$

Problem 1:

$$\begin{aligned} \Delta u = 0 \ \text{ on } R, \ u(0, y) = u(1, y) = 0, \ 0 \le y \le 2, \\ u(x, 2) = 100e^{-\pi} \sin \pi x, \ 0 \le x \le 1, \\ u(x, 0) = \frac{1}{2}u\left(x, \frac{3}{5}\right) + \frac{1}{4}u\left(x, \frac{6}{5}\right) + \frac{1}{4}u\left(x, \frac{9}{5}\right), \ 0 \le x \le 1. \end{aligned}$$

Problem 2:

$$\begin{aligned} \Delta u = 0 \ \text{ on } R, \ u(0, y) = u(1, y) = 0, \ 0 \le y \le 2, \\ u(x, 2) = 100e^{-\pi} \sin \pi x, \ 0 \le x \le 1, \\ u(x, 0) = \frac{1}{4}u\left(x, \frac{3}{2}\right) + \frac{1}{4}u\left(x, \frac{9}{5}\right), \ 0 \le x \le 1. \end{aligned}$$

The exact solutions of this problems are unknown. The approximate solutions obtained by the proposed method are demonstrated on line $y = 0$ (Table 1 and Table 2). According to the repeated digits, for the decreasing mesh steps $h = \frac{1}{16}, \frac{1}{32}, \frac{1}{64}, \frac{1}{128}$ it follows that the maximum error on these line decreases as $O(h^4)$. To achieve this accuracy just 11 iterations in (25) are needed. These problems were solved in [7] with the $O(h^2)$ order of accuracy.

References

[1] A.V. Bitsadze, A.A. Samarskii, On some simplest generalizations of linear elliptic problems, Dokl.Akad. Nauk SSSR, 185(4), (1969).739-740.

 ISBN: 978-84-617-8694-7

$h = 1/16$	$h = 1/32$	$h = 1/64$	$h = 1/128$
0.1466296233	0.1466298514	0.1466298831	0.1466298841
0.2876243524	0.2876247999	0.2876248621	0.2876248641
0.4175658389	0.4175664886	0.4175665789	0.4175665818
0.5314605045	0.5314613313	0.5314614462	0.5314614499
0.6249314409	0.6249324132	0.6249325483	0.6249325526
0.6943866124	0.6943876928	0.6943878429	0.6943878477
0.7371568959	0.7371580428	0.7371582022	0.7371582073
0.7515986533	0.7515998227	0.7515999852	0.7515999904
0.7371568959	0.7371580428	0.7371582022	0.7371582073
0.6943866124	0.6943876928	0.6943878429	0.6943878477
0.6249314409	0.6249324132	0.6249325483	0.6249325526
0.5314605045	0.5314613313	0.5314614462	0.5314614499
0.4175658389	0.4175664886	0.4175665789	0.4175665818
0.2876243524	0.2876247999	0.2876248621	0.2876248641
0.1466296233	0.1466298514	0.1466298831	0.1466298841

Table 1: Solutions on the line y=0 of Problem1

$h = 1/16$	$h = 1/32$	$h = 1/64$	$h = 1/128$
1.40667736245870	1.40667890400060	1.40667919485265	1.40667920245065
2.75929690275187	2.75929992659508	2.75930049712190	2.75930051202591
4.00587821050381	4.00588260044373	4.00588342872030	4.00588345035757
5.09851586514848	5.09852145248217	5.09852250667828	5.09852253421731
5.99522041437613	5.99522698438550	5.99522822398898	5.99522825637145
6.66153200523764	6.66153930544092	6.66154068281451	6.66154071879598
7.07184465696806	7.07185240682252	7.07185386903453	7.07185390723225
7.21039028446737	7.21039818615046	7.21039967700889	7.21039971595495
7.07184465696806	7.07185240682252	7.07185386903453	7.07185390723225
6.66153200523764	6.66153930544092	6.66154068281451	6.66154071879598
5.99522041437613	5.99522698438550	5.99522822398898	5.99522825637145
5.09851586514848	5.09852145248217	5.09852250667828	5.09852253421731
4.00587821050381	4.00588260044373	4.00588342872030	4.00588345035757
2.75929690275187	2.75929992659508	2.75930049712190	2.75930051202591
1.40667736245870	1.40667890400060	1.40667919485265	1.40667920245065

Table 2: Solutions on the line y=0 of Problem 2

 ISBN: 978-84-617-8694-7

[2] A. Ashyralyev and E. Ozturk, On Bitsadze-Samarskii type nonlocal boundary value problems for elliptic differential and difference equations:Well-posedness, Applied Mathematics and Computation, 219 (2012) 1093-1107.

[3] E.A. Volkov, Solvability analysis of a nonlocal boundary value problem by applying the contraction mapping priciple. Comput. Math. Math. Phys., 53(10), (2013) 1494-1498.

[4] S. Sajavicius, Radial basis function method for a multidimensional linear elliptic equation with nonlocal boundary conditions, Comput. Math. Appl. 67 (2014) 1407-1420.

[5] E.A. Volkov, Approximate grid solution of a nonlocal boundary value problem for Laplace's equation on a rectangle Comput.Math. Math. Phys., 53(8), (2013) 1128-1138.

[6] E.A. Volkov, A.A. Dosiyev, S.C. Buranay, On the solution of a nonlocal problem, Comput. Math.Appl.66 (2013) 330-338.

[7] E.A. Volkov and A.A. Dosiyev, On the numerical solution of a multilevel nonlocal problem, Mediterr.J. Math.13 (2016) 3589-3604.

[8] V.A Il'in, E.I. Moiseev, Two-dimensional nonlocal boundary value problems for Poisson's operator in differential and difference variants, Mat. Model. 2(8), (1990) 139-150.

[9] A.A. Dosiyev, A fourth-order accurate composite grid method for solving Laplace's boundary value problems with singularities, Comput. Math. Math. Phys.,42(6), (2002) 832-849.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

Boundary layer flow control using synthetic jets on the flow over a NACA 0012 airfoil

David Duran-Perez[1], Ivette Rodriguez[1], Manel Soria[1] and Oriol Lehmkuhl[2]

[1] *TUAREG-Turbulence and Aerodynamics Research Group, Universitat Politècnica de Catalunya*

[2] *Barcelona Supercomputing Center (BSC),*

emails: davidduranperez@gmail.com, ivette.rodriguez@upc.edu, manel.soria@upc.edu, oriol.lehmkuhl@bsc.es

Abstract

Numerical simulations of the flow control using a synthetic jet are presented. The flow considered is around a NACA 0012 airfoil at $Re = 5000$ and the actuator is located at 6% of the chord from the leading edge. This work will be focused on the study of the effect of the frequency of the actuator on the flow parameters in order to find in which range of frequencies the actuation is more efficient.

Key words: flow control, synthetic jets, numerical simulations

1 Introduction

The field of active flow control (AFC) has experimented a dramatic growth in the recent years, specially in the aeronautic field (despite being a multidisciplinary field). AFC consists on techniques in which energy is actively expended to modify the flow around a specific surface. Periodic excitation introduced at the surface has shown as an efficient and practical means of flow control [1]. It has, among other advantages, the potential to significantly change the lift and drag of an airfoil and the separation of the boundary layer [2].

Regarding periodic actuation, it has been reported two different ranges of frequency in which actuation has been considered optimal i.e. at $F^+O(1)$ and at $F^+O(10)$, where $F^+ = f_{act}L/U_{ref}$ is defined in terms of the position of the actuator from the trailing edge

ISBN: 978-84-617-8694-7

L [3]. Another parameter that should also be considered when it comes to synthetic jets actuators (SJA) is the location and number of jets. This work is a preliminary study regarding the influence of the frequency of the actuation on the flow past a NACA 0012. Most of the studies conducted so far have been performed on thick airfoils with a trailing edge stall and at relatively large Reynolds numbers $Re > 1e5$ (see for instance [3, 4, 5] and citations therein). However, in spite of the studies done so far, there is no consensus on whether the optimum frequency should be on the order of the vortex shedding frequency or an order of magnitude larger, nor if these frequencies have the same effects on thinner airfoils at lower Reynolds numbers (typical of UAVs) where a combination of leading-edge/trailing-edge stall might occur [6]. In order to analyse the influence of the frequency of the actuator, several numerical simulations of a NACA 0012 at $Re = 5000$ and $AoA = 6°$ and $10°$ are performed.

2 Mathematical and numerical model

The incompressible Navier-Stokes equations can be written as

$$\frac{\partial u_i}{\partial x_i} = 0 \tag{1}$$

$$\frac{\partial u_i}{\partial t} + \frac{\partial u_i u_j}{\partial x_j} = -\rho^{-1}\frac{\partial p}{\partial x_i} + \nu\frac{\partial^2 u_i}{\partial x_j \partial x_j} \tag{2}$$

where x_i are the spatial coordinates (or x, y, and z) in the stream-wise, cross-stream and span-wise directions. u_i (or u, v, and w) stand for the velocity components and p is the pressure. ν is the kinematic viscosity and ρ the density of the fluid.

The discretisation of the governing equations has been performed using a low dissipation finite element (FE) scheme, which is based on the same principles followed by Verstappen and Veldman [7], generalised for unstructured finite volumes by Jofre et al. [8] and Trias et al. [9] and extended to finite element (FE) schemes by Lehmkuhl et al. [10]. The basic idea behind this approach remains the same: to mimic the fundamental symmetry properties of the underlying differential operators, i.e., the convective operator is approximated by a skew-symmetric matrix and the diffusive operator by a symmetric, positive-definite matrix. The final set of equations is time integrated using an explicit third order conservative Runge-Kutta method. The pressure stabilisation is achieved by means of a non-incremental fractional step. The chosen low dissipation FE scheme presents good accuracy compared to other low dissipation finite volume and finite difference methods with the advantage of being able to increase the order of accuracy at will without breaking the fundamental symmetry properties of the discrete operators. This methodology is implemented into Alya code, which is a multi-physics parallel code organized in a modular way: kernel, services and modules, which can be separately compiled and linked. Each module represents a single

 ISBN: 978-84-617-8694-7

set of Partial Differential Equations (PDE) for a given physical model. To solve a coupled multi-physics problem, all the required modules must be active and interacting following a well-defined workflow. For more details, the reader is referred to [11].

2.1 Definition of the cases and boundary conditions

All computed flows around a NACA 0012 airfoil extended to include a sharp trailing edge are at Reynolds number $Re = U_{ref}C/\nu = 5000$. The Reynolds number is defined in terms of the free-stream velocity U_{ref} and the airfoil chord C. Two different angles of attack, $AoA = 6^\circ$ and 10° have been considered. As separation at $AoA = 10^\circ$ occurs close to the leading edge of the airfoil, a synthetic jet actuator has been placed at a distance from the leading edge of $x/C = 0.057$. The width of the actuator is $h/C = 0.00748$, similar to that used in the experimental work of Gilarranz [1]. The actuator results in a periodic injection and suction of mass flow with a constant frequency. The velocity at the neck of the actuator is given by,

$$(u, v) = (cos(AoA), -sin(AoA))A_p \sin(2\pi f_{act} t)U_{ref} \tag{3}$$

Considering that there is no consensus regarding the most effective frequency of the sinusoidal oscillation f_{act}, in this work different frequencies are varied from $F^+ = f_{act}U_{ref}/C = 0.85$ to 10 [3]. A_p is the amplitude of the oscillation and is characterised by the momentum coefficient C_μ,

$$C_\mu = \frac{h(\rho U_{max}^2)sin\theta_j}{C(\rho U_{ref}^2)} \tag{4}$$

being θ_j the jet angle with respect to the airfoil surface. In this work it has been set to $\theta_j = 30^\circ$ as in [1].

Due to the large computational resources that might be required to perform a thorough study on three-dimensional meshes, in this work and as a first step in order to analyse the influence of the frequency on the performance of the actuator, two-dimensional simulations are here considered. It is of course known that a final study with a three-dimensional domain will be necessary, but in order to find the range and order of magnitude of the most effective frequencies, it is expected that the results will be acceptable without loss of generality. Thus, the domain to be computed has dimensions of $25C \times 30C$ with the airfoil leading edge placed at $(0, 0)$, see figure 1. In all cases, the meshes used are unstructured and have been constructed clustering the points around the airfoil surface and in the near wake behind the airfoil. In the regions far from the airfoil and the wake, the flow complexity diminishes as the distance increases and so does the computational mesh. Moreover, at the outlet of the actuator, the mesh has been refined so as to carefully capture the actuation of the jet on the boundary layer.

 ISBN: 978-84-617-8694-7

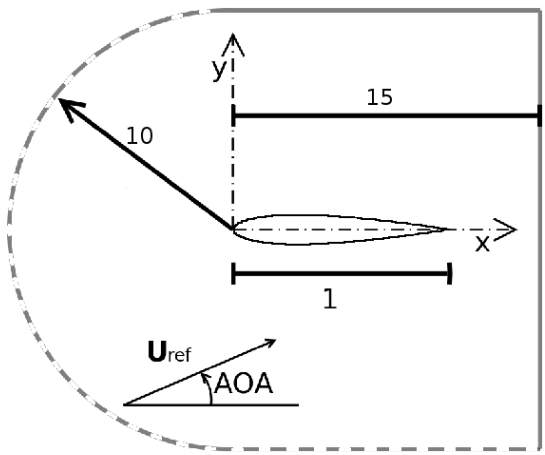


Figure 1: Computational domain (not to scale)

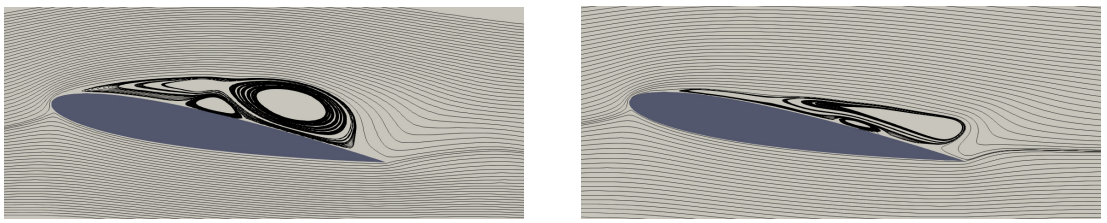
3 Preliminary results

Preliminary results are obtained for $AoA = 10°$ and actuation frequency $F^+ = 1$. In figure 2 the streamlines for both the baseline flow and the actuated one are plotted. As can be seen from the figure, at this AoA a large recirculation bubble can be observed (figure 2a) and the flow reattaches to the airfoil surface close to the trailing edge. For the actuated case, separation occurs at a position downstream of the baseline case and even though a large recirculation bubble is formed, the height of this bubble is smaller than for the baseline case. Indeed, if the streamwise velocity profiles at different streamwise locations are inspected (see figure 3) it can be seen that the size of the separated boundary layer is smaller for the actuated case than for the baseline flow. However, in the case of the baseline case, reattachment occurs earlier than for the actuated case.

At the time this work is being written, several actuation frequencies are being simulated. It is expected to present in the conference a complete set of results with the optimal range of frequencies for which the efficiency of the airfoil is higher.

Acknowledgements

This work has been partially supported the Spanish Ministry project (MEC) FIS2016-77849-R.



(a) (b)

Figure 2: Average flow streamlines at AoA=10. (a) Flow without actuation, (b) Actuated flow $F^+ = 1$

References

[1] J. L. Gilarranz, L. W. Traub, and O. K. Rediniotis. A New Class of Synthetic Jet Actuators-Part I: Design, Fabrication and Bench Top Characterization. *Journal of Fluids Engineering*, 127(2):367, 2005.

[2] Jf Donovan, Ld Kral, and Aw Cary. Active flow control applied to an airfoil. *AIAA paper*, 1998.

[3] M. Amitay and a. Glezer. Role of actuation frequency in controlled flow reattachment over a stalled airfoil. *AIAA Journal*, 40(2):209–216, 2002.

[4] Sebastian D. Goodfellow, Serhiy Yarusevych, and Pierre E. Sullivan. Momentum coefficient as a parameter for aerodynamic flow control with synthetic jets. *AIAA Journal*, 51(3):623–631, 2013.

[5] Régis Duvigneau and Michel Visonneau. Optimization of a synthetic jet actuator for aerodynamic stall control. *Computers and Fluids*, 35(6):624–638, 2006.

[6] I. Rodríguez, O. Lehmkuhl, R. Borrell, and a. Oliva. Direct numerical simulation of a NACA0012 in full stall. *International Journal of Heat and Fluid Flow*, 43:194–203, 2013.

[7] R. W. C. P. Verstappen and A. E. P. Veldman. Symmetry-preserving discretization of turbulent flow. *Journal of Computational Physics*, 187:343–368, 2003.

[8] L. Jofre, O. Lehmkuhl, J. Ventosa, F. X. Trias, and A. Oliva. Conservation properties of unstructured finite-volume mesh schemes for the Navier-Stokes equations. *Numerical Heat Transfer, Part B: Fundamentals*, 54(1):53–79, 2014.

 ISBN: 978-84-617-8694-7

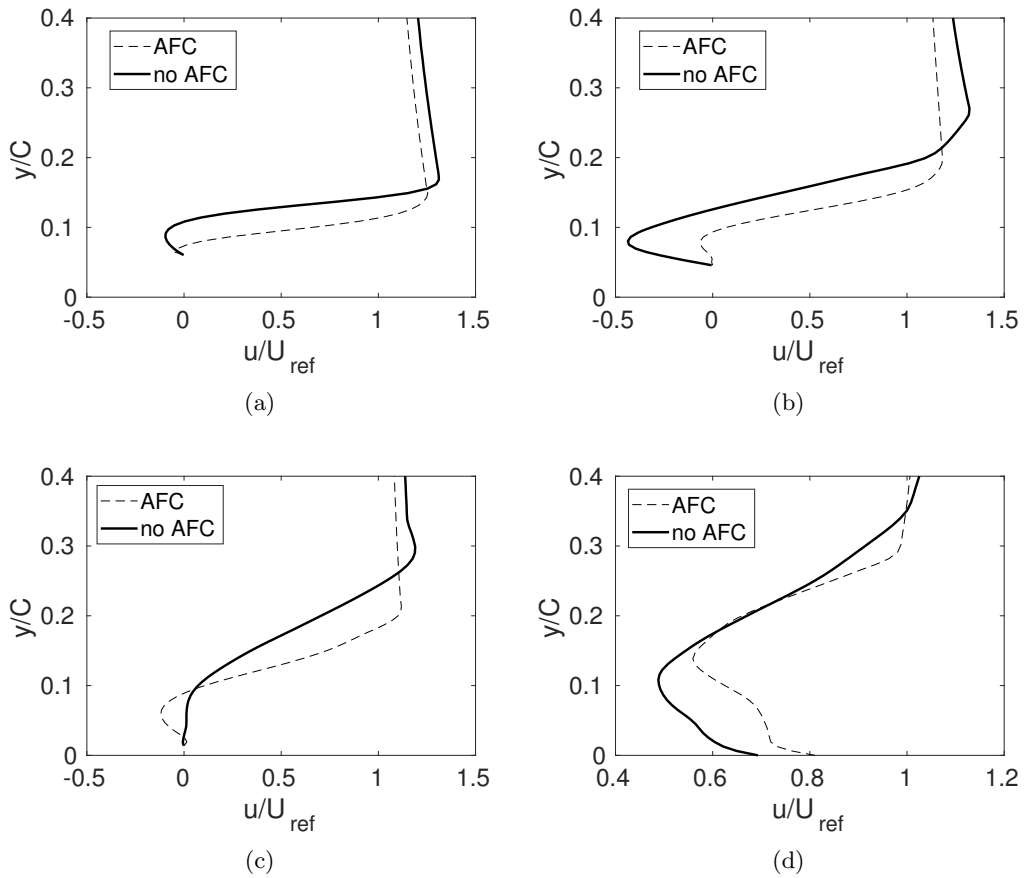


Figure 3: Average flow. Comparison of the actuated (F^+1 and base flows. Streamwise velocity profiles close to the airfoil surface (a,b,c) and in the near wake (d). (a) at $x/C = 0.3$, (b) at $x/C = 0.6$, (c) $x/C = 0.9$, (d) at $x/C = 1.2$.

 ISBN: 978-84-617-8694-7

[9] F. X. Trias, O. Lehmkuhl, A. Oliva, C. D. Pérez-Segarra, and R. W. C. P. Verstappen. Symmetry-preserving discretization of Navier-Stokes equations on collocated unstructured grids. *Journal of Computational Physics*, 258:246–267, 2014.

[10] O. Lehmkuhl, G. Houzeaux, M. Avila, H. Owen, M. Vazquez, and D. Mira. A low dissipation finite element scheme for the large eddy simulation on complex geometries. In *19h International Conference on Finite Elements in Flow Problems - FEF 2017*, 2017.

[11] Mariano Vzquez, Guillaume Houzeaux, Seid Koric, Antoni Artigues, Jazmin Aguado-Sii erra, Ruth Ars, Daniel Mira, Hadrien Calmet, Fernando Cucchietti, Herbert Owen, Ahmedd Taha, Evan Dering Burness, Jos Mara Cela, and Mateo Valero. Alya: Multiphysics engineering simulation toward exascale. *Journal of Computational Science*, 14:15 – 27, 2016.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

A comparison in numerical solution of Richards equation

N. Egidi[1], E. Gioia[2], P. Maponi[1] and L. Spadoni[1]

[1] *School of Science and Tecnology, University of Camerino*

[2] *Department of Life Sciences and the Environment, Marche Polytechnic University*

emails: nadaniela.egidi@unicam.it, e.gioia@staff.univpm.it,
pierluigi.maponi@unicam.it, lorenza.spadoni@unicam.it

Abstract

We consider the Richards equation for the fluid flow in variably saturated porous media, which is a crucial problem in several application fields like water resource management, assessment of water-related disaster and agriculture. So, efficient methods for the numerical solution of the Richards equation can provide great benefits for society. We define four numerical procedures for the solution of such an equation. The computational performances of these procedures are tested and compared in two experiments: a water infiltration problem for three types of soil, a landslide hazard evaluation on a geographical area.

Key words: soil moisture, Richards equation, hydrological modelling, landslide hazard.

MSC 2000: 76S05

1 Introduction

The prediction of fluid movement in porous media is an important problem in many branches of sciences and engineering [6]. The most important application field is probably the soil moisture dynamics, which is a fundamental component of the hydrological cycle. In fact, it is a relevant phenomenon in several natural man-induced processes, such as the soil pollution, which consists of altering the chemical and geological balance caused by the transmission and transport of pollutants into groundwater, compromising quality of freshwater [1]. The soil moisture dynamics can give also important inputs to the efficient use of water resources for agriculture and silviculture activities, avoiding unnecessary irrigations and providing

information for the optimal management operations [2]. It is also strictly connected to the water-related risk analysis: heavy rains combined with the particular conditions characterizing a region, can cause natural disasters such as landslides or floods. An efficient landslide warning system should combine the dynamics of soil moisture, with a model for a quantitative evaluation of the slope stability [3].
The dynamics of the soil moisture can be formally described by the Darcy's law and mass continuity law, in fact the combination of these two principles gives the Richards equation [4], that is a non linear partial differential equation, defining the water flow both in the saturated and unsaturated porous media.
Richards equation cannot be solved analytically under realistic situations, such as three-dimensional heterogeneous media, therefore a numerical solution of this equation has to be computed by discretization methods. The most usual numerical approaches, i.e. the finite difference method [11] and the finite element method [12], can also be applied to Richards equation producing in this way non-linear systems of algebraic equations. However, these general approaches are usually coupled with special linearization techniques to deal with the strong non-linearity of the soil hydraulic functions. In particular, the main objective of these techniques is the achievement of the accurate numerical solutions of the Richards equation, also providing a good approximation of the mass-conservation property [5].
In this paper, four numerical procedures are considered for the solution of Richards equation: such procedures differ for the discretization scheme and/or for the linearization approach as well as for the solution of the linearized equations. These procedures are tested in two numerical experiments concerning the water infiltration on three homogeneous types of soil, and the landslide hazard evaluation on a geographical area having a heterogeneous soil.

The remaining part of the paper is organized as follows. Section 2 defines the Richards equation. In Section 3, we describe two numerical methods for the discretization of the space domain, i.e. the finite difference method and the finite element method, then we define four different implementation procedures. Section 4 provides the first numerical experiment comparing the performance of the procedures on a water infiltration problem in a three types of soil. In Section 5, an application of the Richards equation relative to the landslide hazard problem is studied and the corresponding results are provided for a test area in the Italian territory. In section 6 we give some conclusions and future development of this work.

2 The dynamics of soil humidity

The water movement into the soil can be formally described by Darcy's law and mass continuity law; in fact, these two principles can be combined to obtain the so called Richards

equation [4], that is

$$\Big(C(\psi) + S_s \frac{\theta(\psi)}{n_\epsilon}\Big)\frac{\partial h}{\partial t} = div(K\nabla h) + W - ET, \tag{1}$$

where $div(\cdot) = \frac{\partial(\cdot)}{\partial x}\hat{i} + \frac{\partial(\cdot)}{\partial y}\hat{j} + \frac{\partial(\cdot)}{\partial z}\hat{k}$, $\nabla h = \Big(\frac{\partial h}{\partial x}, \frac{\partial h}{\partial y}, \frac{\partial h}{\partial z}\Big)$, $C(\psi) = \frac{d\theta}{d\psi}$ is the specific capillary capacity, ψ is the pressure head, S_s is the storage coefficient, $\theta(\psi)$ is the water content, n_ϵ is the porosity, $h = \psi + z$ is the hydraulic head, $K(\psi)$ is the hydraulic conductivity, W is the recharge and it is related to the rate of precipitation, ET is the evapotranspiration rate (usually estimated by $Penman - Montieth\ equation$ [9]).
It is worth noticing that equation (1) defines the water flow both in the saturated and in the unsaturated porous media.
The Van Genuchten model is probably the most used formulation of functions $\theta(\psi)$ and $K(\psi)$, see [10] for a detailed definition of these relations. We note that this model depends on various hydraulic parameters, e.g., α that is the inverse of the value of ψ at the air entry point, the residual water content θ_r, the saturated water content θ_s, n that is an empirical parameter, K_s that is the value of the hydraulic conductivity when the soil is saturated. These parameters are all related to the pore size distribution and geometry and so they ultimately depend on the soil type.
The solution of equation (1) requires knowledge of the initial distribution of the hydraulic head h inside the space domain Ω. Moreover, it requires knowledge of appropriate boundary conditions to describe the interactions along the domain boundary $\partial\Omega$: specified hydraulic head (Dirichlet type) and specified flux (Neumann type) are the most commonly used for Richards equation [6].

3 The discretization schemes and their implementation

We describe two numerical methods for the discretization of the Richards equation in the space domain: the finite difference method and the finite element method. Both methods yield to a first order non-linear initial-value problem in the time variable whose solution can be found by applying an iterative procedure. Moreover, linearization strategy is taken into account to deal with such a non linear problem. On the base of these numerical approximations, we define four different procedures for the solution of the problem.

3.1 The Finite difference approximation

The finite difference scheme is obtained from the equation (1) by using central difference quotients for the space derivatives:

$$\Big(C(\psi_{i,j,k}) + S_s \frac{\theta(\psi_{i,j,k})}{n_\epsilon}\Big)\frac{\partial h}{\partial t}(x_i, y_j, z_k) =$$

 ISBN: 978-84-617-8694-7

$$= \frac{1}{(\Delta x)^2}\Big[K(\psi_{i+1/2,j,k})(h_{i+1,j,k} - h_{i,j,k}) - K(\psi_{i-1/2,j,k})(h_{i,j,k} - h_{i-1,j,k})\Big] +$$

$$+\frac{1}{(\Delta y)^2}\Big[K(\psi_{i,j+1/2,k})(h_{i,j+1,k} - h_{i,j,k}) - K(\psi_{i,j-1/2,k})(h_{i,j,k} - h_{i,j-1,k})\Big] +$$

$$+\frac{1}{(\Delta z)^2}\Big[K(\psi_{i,j,k+1/2})(h_{i,j,k+1} - h_{i,j,k}) - K(\psi_{i,j,k-1/2})(h_{i,j,k} - h_{i,j,k-1})\Big] + W_{i,j,k} - ET_{i,j,k}, \quad (2)$$

where Δx, Δy, Δz are the discretization steps in the x, y and z direction, respectively, $h_{i,j,k}$ and $\psi_{i,j,k}$ denote the approximate value of h and ψ at in the generic grid point (x_i, y_j, z_k); for brevity, in equation (2) the time-dependence of these functions is not explicitly denoted.

3.2 Finite element approximation

In the following we suppose $\partial\Omega = \partial\Omega^D \cup \partial\Omega^N$, where we impose a Dirichlet boundary condition in $\partial\Omega^D$ and a Neumann boundary condition in $\partial\Omega^N$. We consider the Galerkin method [12], that is based on the weak formulation of the Richards equation (1):

$$\int_\Omega w\Big(C(\psi) + S_s\frac{\theta(\psi)}{n_\epsilon}\Big)\frac{\partial h}{\partial t}\; d\underline{x} + \int_\Omega \nabla w \cdot [K(\psi)\cdot\nabla h]\; d\underline{x}$$

$$-\int_\Omega w(W - ET)\; d\underline{x} + \int_{\partial\Omega^N} wq_h\; ds(\underline{x}) = 0; \quad \forall w \in {H^1}_0(\Omega) \quad (3)$$

where $q_h = [K(\psi)\cdot\nabla h]\cdot n$, and $H_0^1(\Omega)$ is the space of square integrable functions w having square integrable derivative defined almost everywhere, an such that $w(x) = 0$, $x \in \partial\Omega^D$. Solution h is approximated by a function $\tilde{h}(x,t) \approx \sum_{j=1}^{N_p} N_j(x)h_j(t)$ where N_j, $j = 1, 2, \ldots, N_p$ are basis functions and $h_j(t)$ are unknown coefficients to be determined; for semplicity we suppose that an homogeneous Dirichilet boundary condition is prescribed on $\partial\Omega^D$. In particular, this basis is usually defined by piece-wise polynomial functions N_i having a small support Ω_i with respect to the domain. The Galerkin method considers the weighting function w equal to the representation functions, i.e. N_i, so formula (3) becomes

$$\sum_e \int_{\Omega_e} N_i\Big(C(\psi^e) + S_s^e\frac{\theta(\psi^e)}{n_\epsilon^e}\Big)\frac{\partial}{\partial t}\Big(\sum_j N_j h_j\Big)\; d\underline{x} + \sum_e \int_{\Omega_e} \nabla N_i \cdot [K(\psi^e)\cdot\nabla(\sum_j N_j h_j)]\; d\underline{x}$$

$$-\sum_e \int_{\Omega_e} N_i(W^e - ET^e)d\underline{x} + \sum_e \int_{\partial\Omega_e{}^N} N_i {q_h}^e ds(d\underline{x}) = 0$$

$$1 \le i \le N_p. \quad (4)$$

 ISBN: 978-84-617-8694-7

3.3 Time discretization

Both equation (2) and equation (4) represent a first-order initial-value problem in the time variable. This problem is numerically solved by an iterative process. Let $\Delta t > 0$ be a time step which defines a partion of the time domain $[0, T_0]$ such that $t_n = n\Delta t$, $n = 0, \dots N$. On the base of this discretization, both (2) and (4) can be formally rewritten as

$$\frac{\partial \underline{h}^n}{\partial t} = F(\underline{h}^n, t_n), \tag{5}$$

where $\underline{h}^n$ is the vector of the unknowns at time $t = t_n$, $F(\underline{h}^n, t_n)$ is a non linear vectorial function. We replace the time derivative by the finite difference approximation

$$\frac{\partial \underline{h}^n}{\partial t} = \frac{\underline{h}^{n+1} - \underline{h}^n}{\Delta t}. \tag{6}$$

So, in order to define the iterative process for the numerical solution of (5), we take into account the single step methods which are expressed in their generalized form as

$$\underline{h}^{n+1} = \underline{h}^n + \Delta t[(1-\lambda)F(\underline{h}^n, t_n) + \lambda F(\underline{h}^{n+1}, t_{n+1})], \tag{7}$$

where λ is a time-weighting factor, $0 \leq \lambda \leq 1$.
In particular, the case $\lambda = 0$ is called explicit Euler method, the case $\lambda = \frac{1}{2}$ is called Crank Nicolson method while the case $\lambda = 1$ is called implicit Euler method.

3.4 Linearization tecniques

At each time t_n, the numerical solution of nonlinear system (7) requires an iterative procedure, where, roughly speaking, several local linearized models are subsequently solved. We consider the following iterative procedure: we define $\underline{h}^{n+1,r}$ as the vector solution at time level $n+1$ at the r-th refinement iterate and for $r = 1, \cdots R$ we compute

$$\underline{h}^{n+1,r} = \underline{h}^n + \Delta t[(1-\lambda)F(\underline{h}^n, t_n) + \lambda F(\underline{h}^{n+1,r-1}, t_{n+1})], \tag{8}$$

where R is the first iterate that, given an appropriate tolerance *toll*, satisfies

$$\left|\left|\underline{h}^{n+1,r} - \underline{h}^{n+1,r-1}\right|\right|_\infty < toll. \tag{9}$$

At the end of the cycle we define

$$h^{n+1} = h^{n+1,R}. \tag{10}$$

Other linearization methods used to solve (7) are *Picard linearization scheme* and *Newton scheme* [7].

 ISBN: 978-84-617-8694-7

3.5 The procedures

We consider four numerical procedures for the solution of Richards equation. These procedures are obtained by combining the previously described approximation approaches in order to compare their efficiency.

Procedure 1 *The finite difference method is used for space discretization; the solution of the system (7) is carried out by Crank-Nicolson method togheter with the linearization scheme (8) and the solution at the iterate* $r = 1$ *is given by the explicit Euler method.*

Procedure 2 *The finite difference method is used for space discretization; the solution of the system (7) is carried out by implicit Euler method togheter with the Picard linearization scheme and the solution of the linearized system is obtainded by means of a direct method based on a sparse variant of Gaussian elimination, that is called multifrontal method [13].*

Procedure 3 *The finite difference method is used for space discretization; the solution of the system (7) is carried out by implicit Euler method togheter with the Picard linearization scheme and the solution of the linearized system is obtainded by means of the preconditioned conjugate gradient method.*

Procedure 4 *The finite element method is used for space discretization; the solution of the system (7) is carried out by implicit Euler method togheter with the Picard linearization scheme and the solution of the linearized system is obtainded by means of the preconditioned conjugate gradient method.*

4 Numerical experiments

We tested the four numerical procedures presented in the previous section with three examples. All these examples consider the same spatial domain $\Omega_1 = [0, L_x] \times [0, L_y] \times [0, L_z]$, where $L_x = 90$ m, $L_y = 50$ m and $L_z = 4$ m and the same time period of 30 days, but with three different soil types. Table 1 shows the values of α, θ_r, θ_s, n, K_s of these considered soil types in the Van Genucthen model (see [8] for details). In particular, the first example considers a clay soil, at time $t = 0$ the simulation starts from a saturation equal to 50% and the time period is characterized by a precipitation rate of 100 mm/day. The second and third examples consider sand and coarse sand soils, respectively, at time $t = 0$ the saturation is equal to 30% and the precipitation rate considered is 50 mm/day. At the boundary of Ω_1 we impose the zero normal flow condition as we supposed that these boundaries do not interact with the outside. Procedure 1 is implemented inside a program written in Fortran language, while, in order to implement the multifrontal method in Procedure 2, we used the package MA57 [14], taken from the HSL software library. Procedure 3 is implemented inside the computer model MODFLOW-2005 [11] which numerically solves the three-dimensional

 ISBN: 978-84-617-8694-7

ground-water flow equation by using a finite-difference method. Procedure 4 is implemented inside the demo version of FEFLOW software [12], that is an advanced Finite-Element subsurface flow and transport modeling system.

Grainsize	Description	θ_r	θ_s	$\alpha[1/m]$	n	$K_s[m/s]$
Clay	Cl> 75%	0.11	0.48	1.33	1.31	1.00E-08
Sand	Sa> 75%	0.05	0.39	3.36	2.11	5.83E-05
Coarse Sand	CSa> 75%	0.02	0.36	15.85	2.91	1.00E-03

Table 1: Different types of grain sizes and the corresponding parameters θ_r, θ_s, α, n, K_{sat} in Van Genucthen model.

Table 2 shows a comparison between the results obtained by the four numerical procedures on the basis of the time step Δt, the elapsed CPU time of the computation and the $RE(i,j)$, i.e. the relative error in the 2-norm between the solution computed by the procedure i and the one computed by procedure j at the end of the simulation period. We observe that, despite its simple structure, Procedure 1 performs quite well with the exception of the coarse sand type where, in order to achieve the convergence of the method, we had to decrease the time step by a factor 100, compromising the efficiency of the procedure. Figure 1 shows the saturation in the domain Ω_1 at the end of the simulation period relative to the clay case: from this figure, we observe a similar behaviour in the solutions computed by the four procedures.
We tested our simulations using Window PCs with 128 GB of RAM and 8 Core of CPU.

5 A landslide hazard application

We consider the application of the dynamics of the soil moisture for a quantitative evaluation of landslide hazard.
The landslide hazard problem is usually formulated in terms of the $Safety\ Factor$, which is given by the ratio between the forces that prevent the slope from failing and those that bring the slope to collapse; in particular, *the last forces depend on the soil saturation, given by the solution of Richards equation* (1). Therefore, the safety factor is a hazard index: a value larger than 1 indicates stable conditions, a value smaller than 1 indicates unstable conditions.
The $Infinite\ Slope\ Model$ [15] is probably the easiest evaluation method for the safety factor F, that is given by

$$F = \frac{C + (z\gamma - z_w\gamma_w)\cos^2\beta\tan\phi}{z\gamma\sin\beta\cos\beta}, \tag{11}$$

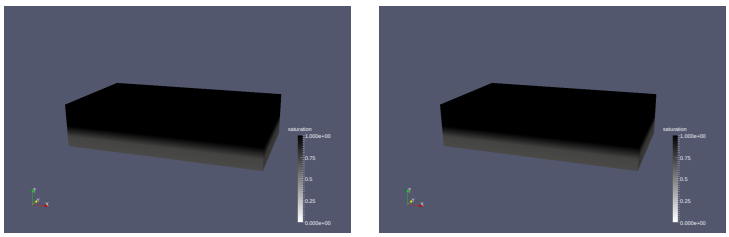
 ISBN: 978-84-617-8694-7

Procedure	Δt (min)	Computation time (min)	RE with Procedure 2	RE with Procedure 3	RE with Procedure 4
			CLAY		
Procedure 1	15	0.25	7.96E-5	0.00553	0.0555
Procedure 2	15	0.3	/	0.00552	0.0554
Procedure 3	15	0.4	/	/	0.0664
Procedure 4	15	8	/	/	/
			SAND		
Procedure 1	15	0.25	0.000757	0.00659	0.188
Procedure 2	15	0.3	/	0.0664	0.187
Procedure 3	15	0.4	/	/	0.230
Procedure 4	15	8	/	/	/
			COARSE SAND		
Procedure 1	0.15	30	0.0905	0.0022	0.118
Procedure 2	15	0.3	/	0.0889	0.145
Procedure 3	15	0.4	/	/	0.0053
Procedure 4	15	8	/	/	/

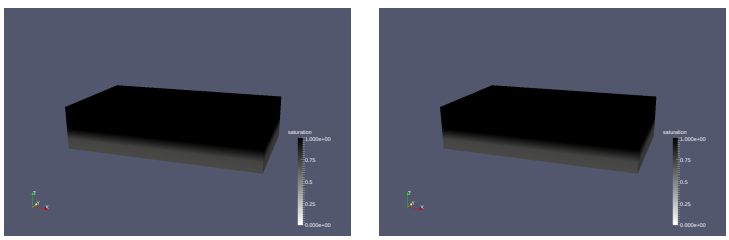
Table 2: Comparison between the solution computed by the four different procedures relative to the three soil types described in Table 1.

where C is the effective cohesion; γ is the unit weight of the soil; γ_w is the unit weight of water; z is the depth of the failure surface; z_w is the height of the watertable above failure surface; β is the slope of the inclined surface; ϕ is the angle of internal friction.
The Infinite Slope Model has been used to analyse a geographical area of 11,69 km^2 located in the hydrographic basin of the Esino River in the province of Ancona (Italy), where a landslide occurred on March 2015. This area is representative due to its landslide susceptibility, moreover several meteorological stations are located in it. Previous geological and geotechnical studies provided the geomorphological information: in particular, we know that the soil is mostly characterized by clay and sand.
The proposed experiment considers the test area during the three months before the landslide event: Figure 2 shows the millimeters of rain fell every days during the observation period (12/6/2014 to 3/6/2015). The four procedures described in section 3, are combined with the infinite slope model (11): in particular, the solution of Richards equation (1) is used in order to obtained information about z_w in formula (11).
Figure 2 shows the trend of the minimum value of F computed during the observation period: this diagram shows that during the three months before the landslide initiation, the safety factor values gradually decreased, passing from values larger than 1 (indicating stable region) to values smaller than 1 (indicating unstable region).
Figure 3 shows a graphical representation of the safety factor values concerning the last day of the simulation: a nearly black zone is an unstable region ($F < 1$); a nearly white zone

 ISBN: 978-84-617-8694-7



(a) *Procedure 1* (b) *Procedure 2*



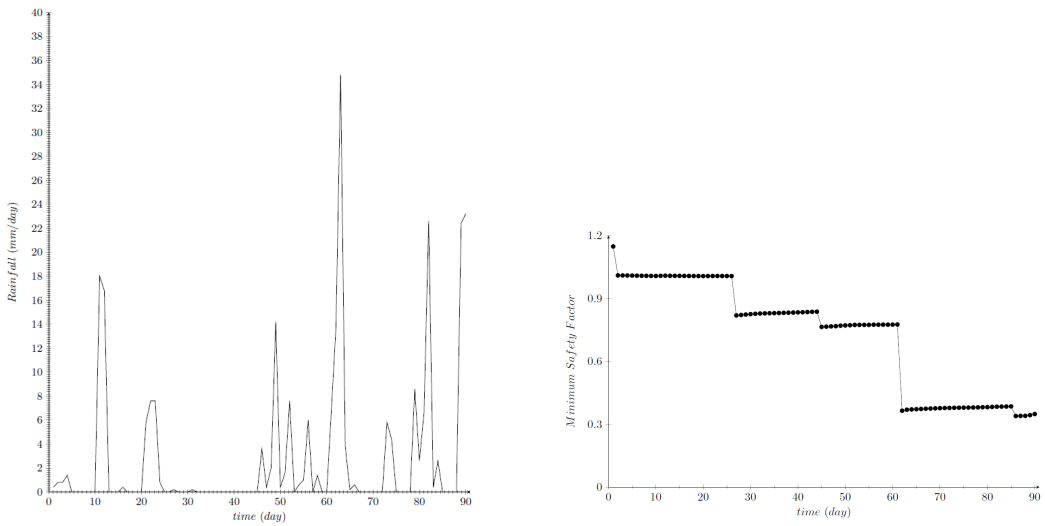
(c) *Procedure 3* (d) *Procedure 4*

Figure 1: The saturation in Ω_1 at the time $t = 30$ dayes related to clay.

is a stable region ($F > 1$). The four images in this figure are relative to the four numerical procedures for the soil moisture dynamics. So these results show that the four procedures give the same solution for the computation of the safety factor.

6 Conclusions

The water movement in the unsaturated porous media is described by Richards equation that is a non-linear partial differential equation. Four numerical procedures are considered for the solution of this equation: such procedures differ for the discretization scheme and/or for the linearization approach as well as for the solution of the linearized equations.
We tested and compared the four procedures in two numerical experiments: the soil moisture dynamics in three soil types (clay, sand, coarse sand) and the landslide hazard evaluation on a geographical area with mixed soil (clay, sand). Results show that, Procedure 1 performs quite well, even if it breaks down when we consider the coarse sand type with a discretization time of 15 minutes, however the convergence can be achieved if we decrease the time step. In the other cases, Procedure 1 seems to provide acceptable results in a very efficient way.
This preliminary study should be extended by comparing the four procedures on the basis of a larger range of soil types. This would allow a more detailed evaluation of the efficiency of Procedure 1 and the possible remedies to overcome its inefficiencies. The results of such



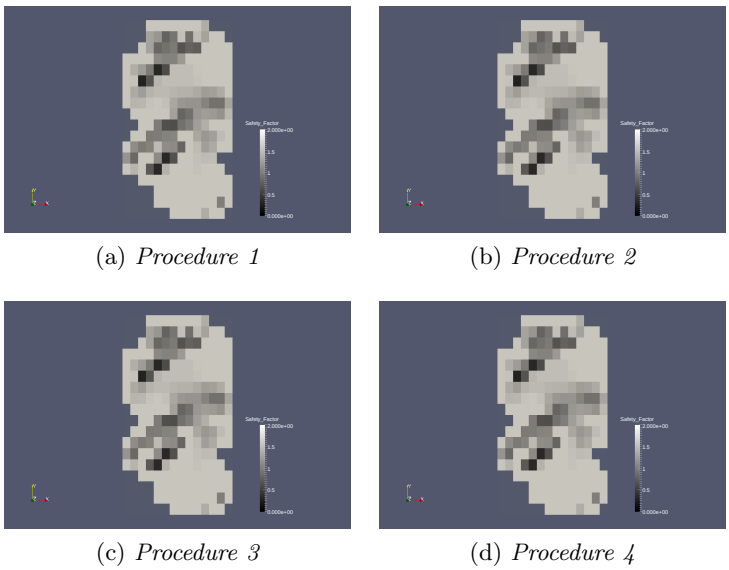
(a) (b)

Figure 2: (a) the millimeters of rain fell every dayes during the three months before the landslide event; (b) trend of the minimum safety factor during the observation period.

an analysis will be useful in several application fields like natural hazard evaluation, water resource analysis and weather forecast.
Another further difficulty in Richards equation is the efficient acquisition of model parameters. Actually, this is a quite common situation to all mathematical models for complex phenomena; however, in this particular case, it is a very crucial issue since it depends on the weather data and on the geotechnical features of the soil. Weather is usually acquired by proper stations on the territory or by satellite measurements. Soil features are mainly characterized by the granulometry, which can be used to identify the textural class of the soil studied among a number of possible classes and it can be obtained by direct measurements and/or general geomorphological features of the territory. So, both sets of these data are difficult to get with high spatial resolution and accuracy. Unfortunately the lack of information can limit the successfull application of Richards equation to real-world problems, so a precise sensitivity analysis for such a model should be provided in order to obtain reliable results.

7 Acknowledgment

This activity was partially supported by LANDSLIDE (ECO/SUB/2014/693902), an European project co-financed by the Directorate General Humanitarian Aid and Civil Protection



(a) *Procedure 1* (b) *Procedure 2*

(c) *Procedure 3* (d) *Procedure 4*

Figure 3: Graphical representation of the safety factor values concerning the last day of the simulation.

of the European Commission, whose aim is the development, the implementation and the testing of an automatic software for the dynamic evaluation of the soil-moisture content and the corresponding prediction of the daily hazard of landslides in the European territories. LANDSLIDE is made up of 6 organisation partners coming from Italy, Bulgaria, Greece and Poland.

References

[1] Pan, Tongyan and Tao Miao, *Contamination of roadside soils by runoff pollutants: a numerical study*, Transportation Geotechnics **2** (2015) 1–9.

[2] Mailhol, Jean Claude, Pierre Ruelle, and Zornitsa Popova, *Simulation of furrow irrigation practices (SOFIP): a field-scale modelling of water management and crop yield for furrow irrigation*, Irrigation science **24** (2005) 37–48.

[3] Baum, Rex L. and Godt, Jonathan W., *Early warning of rainfall-induced shallow landslides and debris flows in the USA*, Landslides **7** (2005) 259–272.

[4] Richards, Lorenzo Adolph, *Capillary conduction of liquids through porous mediums*, Physics **1** (1931) 318–333.

 ISBN: 978-84-617-8694-7

[5] Zarba, Raaecca L., E. T. Bouloutas, and M. Celia, *General mass-conservative numerical solution for the unsaturated flow equation*, Water Resources Research WRERAQ **26** (1990) 1483–1496.

[6] John Wiley & Sons, *Subsurface hydrology*, Pinder, George F and Celia, Michael A, 2006.

[7] Lehmann, F., and P. H. Ackerer, *Comparison of iterative methods for improved solutions of the fluid flow equation in partially saturated porous media*, Transport in Porous Media **31** (1998) 275–292.

[8] Gioia Eleonora et al, *Application of a process-based shallow landslide hazard model over a broad area in Central Italy*, Landslides **13** (2016) 1197–1214.

[9] Allen, Richard G., et al, *Crop evapotranspiration-Guidelines for computing crop water requirements-FAO Irrigation and drainage paper 56*, FAO, Rome **300** (1998) D05109.

[10] Van Genuchten, M. Th., *A closed-form equation for predicting the hydraulic conductivity of unsaturated soils*, Soil Science Society of Americae **44** (1980) 892–898.

[11] Harbaugh, Arlen W, *MODFLOW-2005, the US Geological Survey modular groundwater model: the ground-water flow process*, Reston, VA, USA: US Department of the Interior, US Geological Survey, 2005.

[12] Diersch, Hans-Joerg, *FEFLOW: finite element modeling of flow, mass and heat transport in porous and fractured media*, Springer Science & Business Media, 2013.

[13] Duff, Iain S., *MA57—a code for the solution of sparse symmetric definite and indefinite systems*, ACM Transactions on Mathematical Software (TOMS) **30** (2004) 118–144.

[14] HSL lybrary, `http://en.wikipedia.org/wiki/Business_logic_layer`, 2 9 2016.

[15] Duncan, J Michael and Wright, Stephen G and Brandon, Thomas L, *Soil strength and slope stability*, John Wiley & Sons, 2014.

 ISBN: 978-84-617-8694-7

Volume III

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

Wildland fire propagation modeling: fire-spotting parametrisation and energy balance

Vera N. Egorova[1], **Gianni Pagnini**[1,2] **and Andrea Trucchia**[1,3]

[1] *BCAM - Basque Center for Applied Mathematics, Alameda de Mazarredo 14, E-48009 Bilbao, Basque Country, Spain*

[2] *Ikerbasque Basque Foundation for Science, Calle de María Díaz de Haro 3, E-48013 Bilbao, Basque Country, Spain*

[3] *University of the Basque Country UPV/EHU, Barrio Sarriena s/n, E-48940 Leioa, Basque Country, Spain*

emails: vegorova@bcamath.org, gpagnini@bcamath.org, atrucchia@bcamath.org

Abstract

Present research concerns the physical background of a wild-fire propagation model based on the split of the front motion into two parts - drifting and fluctuating. The drifting part is solved by the level set method and the fluctuating part describes turbulence and fire-spotting. These phenomena have a random nature and can be modeled as a stochastic process with the appropriate probability density function. Thus, wildland fire propagation results to be described by a nonlinear partial differential equation (PDE) of the reaction-diffusion type. A numerical study of the effects of the atmospheric stability on wildfire propagation is performed through its effects on fire-spotting. Moreover, it is shown that the solution of the PDE as an indicator function allows to construct the energy balance equation in terms of the temperature.

Key words: fire propagation, fire-spotting, level set method, energy balance
MSC 2000: 00A59, 35K57, 65C20, 70H20

1 Introduction

In wildland fire propagation, fire-spotting phenomena cause isolated fire from the main fire. It is an important aspect because it affects the rate of spread of the fire and may cause

dangerous effects. The present study deals with fire propagation modelling and its physical background.

In the proposed model, the front motion is splitted into two parts - drifting and fluctuating. Each of them can be solved by using an appropriate method. In the present study, the Eulerian Level Set Method (LSM) is chosen for the drifting part while the fluctuating part is the result of a comprehensive statistical description of the physics of the system and takes into account the randomness of the hot air turbulent transport and fire-spotting. Thus, in order to treat the fluctuating part, specific probability density function has to be taken into account [1].

Fire-spotting is a complicated physical process due to many factors, such as wind, fire intensity, fuel characteristics, atmospheric conditions, etc. The statistical formulation of fire-spotting has been proposed in [1] and completed by the physical parametrisation in [2]. This formulation does not depend on the method used for fire propagation. In the present study we extend this approach and include into account the change of wind direction and possibility of the several secondary fires appearance.

The main aim of the present study is to explain the physical background of the proposed model. The solution of the underlying PDE is connected to a temperature field. The transfer of the temperature due to the turbulent flows is then described by the energy balance equation. The rest of the paper is organized as follows. The wildland fire propagation model is proposed in the next section, including a brief description of the level-set method and physical parametrisation of the fire-spotting. Section 3 deals with the energy balance equation for the proposed model. Numerical examples for several test cases, such as different wind conditions, and merging of the secondary fires, are provided in Section 4.

2 Fire propagation model

2.1 Level-set method

LSM is widely used as effective method for the front-tracking. For some computational domain S the fire front contour is represented by a closed curve Γ. The region bounded by Γ is denoted by Ω and represents the burnt area. Let us introduce an indicator function $\phi(\mathbf{x}, t)$:

$$\phi(\mathbf{x}, t) = \begin{cases} 1, & \mathbf{x} \in \Omega(t), \\ 0, & \text{othervise.} \end{cases} \tag{1}$$

Let $\gamma : S \times [0, \infty) \to \mathbb{R}$ be a level-set function, such that for some fixed γ^* at the moment t the fire front can be described by $\Gamma(t) = \{\mathbf{x} \in S | \gamma(\mathbf{x}, t) = \gamma^*\}$. If $\gamma(\mathbf{x}, t) > \gamma^*$, then the ignition is observed at the point $\mathbf{x}$. The level-set function $\gamma(\mathbf{x}, t)$ evolves according to the following ordinary level-set equation

 ISBN: 978-84-617-8694-7

$$\frac{\partial \gamma}{\partial t} = V(\mathbf{x}, t)||\nabla \gamma||, \tag{2}$$

where $V(\mathbf{x}, t)$ is a rate of spread (ROS) of the fire front. The ROS value depends on many elements, such as the intensity and direction of the wind, fuel conditions, etc.

2.2 Turbulence and fire-spotting modelling

Together with other factors, turbulence and fire-spotting cause the random motion of the fire-front. Thus, the random front contour can be defined by the effective indicator function [1]:

$$\phi_e(\mathbf{x}, t) = \int_{\Omega(t)} f(\mathbf{x}; t|\bar{\mathbf{x}}) d\bar{\mathbf{x}}, \tag{3}$$

where $f(\mathbf{x}; t|\bar{\mathbf{x}})$ is the probability density function (PDF) that accounts for turbulence and fire-spotting effects. Note that the point is labelled as burnt, if $\phi_e(\mathbf{x}, t)$ exceeds some threshold value ϕ_e^{th}.

In accordance with the Reynold transport theorem, the evolution of the effective indicator function $\phi_e(\mathbf{x}, t)$ takes the form

$$\frac{\partial \phi_e}{\partial t} = \int_{\Omega(t)} \frac{\partial f}{\partial t} d\bar{\mathbf{x}} + \int_{\Omega(t)} \nabla_{\bar{\mathbf{x}}} \left[V(\bar{\mathbf{x}}, t) f(\mathbf{x}; t|\bar{\mathbf{x}}) \right] d\bar{\mathbf{x}}. \tag{4}$$

Evolution equation for the PDF $f(\mathbf{x}; t|\bar{\mathbf{x}})$ takes the form

$$\frac{\partial f}{\partial t} = \epsilon f, \tag{5}$$

where $\epsilon = \epsilon(\mathbf{x})$ is a generic evolution operator. Hence, (4) can be rewritten in the following form

$$\frac{\partial \phi_e}{\partial t} = \epsilon \phi_e + \int_{\Omega(t)} \nabla_{\bar{\mathbf{x}}} \left[V(\bar{\mathbf{x}}, t) f(\mathbf{x}; t|\bar{\mathbf{x}}) \right] d\bar{\mathbf{x}}. \tag{6}$$

In (6) the front-line velocity is controlled by the ROS, while random process, such as turbulence and fire-spotting, are modelled by modifying PDF. Thus, if $f(\mathbf{x}; t|\bar{\mathbf{x}}) = \delta(\mathbf{x} - \bar{\mathbf{x}})$, equation (6) reduces to the deterministic case described by (2).

Assuming that the downwind phenomenon of fire-spotting is independent of turbulence, the random process handled by $f(\mathbf{x}; t|\bar{\mathbf{x}})$ in (6) can be defined as follows

$$f(\mathbf{x}; t|\bar{\mathbf{x}}) = \begin{cases} \int_0^\infty G(\mathbf{x} - \bar{\mathbf{x}} - l\hat{\mathbf{n}}; t) q(l) dl, & \text{if downwind,} \\ G(\mathbf{x} - \bar{\mathbf{x}}; t), & \text{otherwise.} \end{cases} \tag{7}$$

 ISBN: 978-84-617-8694-7

The shape of the PDF is defined by the isotropic bi-variate Gaussian function (considering turbulence effects) $G(\mathbf{x}-\bar{\mathbf{x}};t)$ and the firebrand landing distribution $q(l)$ is defined by a lognormal distribution as follows

$$q(l)=\frac{1}{\sqrt{2\pi}\sigma l}\exp\frac{-(\ln l/\mu)^2}{2\sigma^2}, \tag{8}$$

where μ is the ratio between the square of the mean of landing distance l and its standard deviation, σ is the standard deviation of $\ln l/\mu$.

In [3] authors complete the study of the fire-spotting proposed in [1] by describing this phenomenon in terms of the fire intensity, wind velocity and fuel characteristics. It includes only the vital ingredients, each firebrand is assumed to be spherical of the constant size. Then lognormal parameters μ and σ in (8) take the following form

$$\mu=H\left(\frac{3\rho_a C_d}{2\rho_f r g}\right)^{1/2},\quad \sigma=\frac{1}{2z_p}\ln\left(\frac{U^2}{rg}\right), \tag{9}$$

where U is the wind velocity and H is the maximum loftable height, that is according to [4],

$$H=\alpha H_{ABL}+\beta\left(\frac{I}{dP_{f_0}}\right)^{\gamma}\exp\left(-\frac{\delta N_{FT}^2}{N_0^2}\right), \tag{10}$$

where all the parameters are defined in Table 1.

3 Energy balance equation

Wildfire model can be formulated based on balance equations for energy and fuel, as it is proposed in [5,6]. In the present study we follow the level-set formulation with the stochastic process. However, there is a connection between these two formulations. In order to derive the energy balance equation, the physical laws are used, mainly, conservation of energy and fuel reaction. In present study we focus on the solution in the form of indicator function. Thus, it is important to show, that the found solution is connected to the temperature and the energy balance equation can be derived by using the indicator function.

An important part of study deals with the temperature field. Temperature is transferred due to the turbulent flows, and it can be modelled by the diffusion process. The heating-before-burning mechanism, that is accumulation in time of potential fire, can be associated with an amount of heat:

$$\psi(\mathbf{x},t)=\int_0^t \phi_e(\mathbf{x},\eta)\frac{d\eta}{\tau}, \tag{11}$$

where τ is the ignition delay, that can be understood as a resistance to the hot-air heating and firebrand landing in parallel.

 ISBN: 978-84-617-8694-7

Notation	Description	
α	Part of ABL passed freely, $\alpha < 1$	0.24 [4]
β	$[m]$ Contribution of the fire intensity, $\beta > 0$	170 [4]
γ	Power-law dependence on FRP, $\gamma < 0.5$	0.35 [4]
δ	Dependence on stability of the FT, $\delta \geq 0$	0 [4]
H_{abl}	$[m]$ Height of the atmospheric boundary layer (ABL)	1200
d	$[m\ ]$ Unit depth of the combustion zone	1
P_{f_0}	$[MWm^{-2}]$ Ratio of reference fire power	1 [4]
N^2_{FT}	$[s^{-2}]$ Brunt-Väisälä frequency in the FT	$2.789 \cdot 1e-4$
N^2_0	$[s^{-2}]$ Brunt-Väisälä frequency	$2.5 \cdot 1e-4$
H	$[m]$ The maximum loftable height	
ρ_a	$[kg/m^3]$ Density of the ambient air	1.1
ρ_f	$[kg/m^3]$ Density of the wild-land fuels	542
C_d	Drag coefficient	0.45
z_p	p-th percentile	0.45
r	$[m]$ Brand radius	0.015
g	$[ms^{-2}]$ Acceleration due to gravity	9.81

Table 1: Physical parameters of the atmospheric boundary layer and fire-spotting.

The amount of heat is proportional to the increasing of the temperature $T(\mathbf{x}, t)$. For the sake of simplicity, we can assume that

$$\psi(\mathbf{x}, t) = \frac{T(\mathbf{x}, t) - T_a(\mathbf{x})}{T_{ign} - T_a(\mathbf{x})}, \quad T < T_{ign}, \tag{12}$$

where ambient temperature is denoted by T_a, and T_{ign} stands for the ignition temperature. From (12) one can see, that $\psi(\mathbf{x}, t) = 1$ entails that $T(\mathbf{x}, t) = T_{ign}$ and the spacial point $\mathbf{x}$ at the moment t belongs to the burning area.

Temperature $T(\mathbf{x}, t)$ can be found from (12) as follows

$$T(\mathbf{x}, t) = T_a(\mathbf{x}) + \psi(\mathbf{x}, t)\left(T_{ign} - T_a(\mathbf{x})\right), \tag{13}$$

that by the differentiation leads to

$$\frac{\partial T}{\partial t} = \frac{(T_{ign} - T_a(\mathbf{x}))}{\tau}\phi(\mathbf{x}, t). \tag{14}$$

According to the heat-before-burning mechanism, the temperature field is described by the following reaction-diffusion type equation [1]

$$\frac{\partial T}{\partial t} = \epsilon T + \frac{T_{ign} - T_a}{\tau}\left(I_{\Omega_0}(\mathbf{x}) + W(\mathbf{x}, t)\right), \tag{15}$$

 ISBN: 978-84-617-8694-7

where T_{ign} is the ignition temperature, $T_a(\mathbf{x})$ is the ambient temperature, $I_{\Omega_0}(\mathbf{x}) = \phi_e(\mathbf{x}, 0)$ and

$$W(\mathbf{x}, t) = \int_0^t \left(\int_{\Omega(\theta)} \nabla_{\bar{x}} \cdot [\mathbf{V}(\bar{\mathbf{x}}, \theta) f(\mathbf{x}; \theta | \bar{\mathbf{x}})] \, d\bar{\mathbf{x}} \right) d\theta. \tag{16}$$

Equation (15) can be understood as the energy balance equation associated to the model. In (15) $\epsilon = \epsilon(\mathbf{x})$ is a generic evolution operator, associated to the evolution of $f(\mathbf{x}; t|\bar{\mathbf{x}})$, and it models the turbulent heat transfer due to radiation. Moreover, the second term on the RHS corresponds to the convective heat lost to the atmosphere and the third to the rate of fuel consumed by the fire with Rate of Spread $\mathbf{V}(\bar{\mathbf{x}}, t)$ in the outward direction.

4 Numerical results

In this section some numerical examples are considered in order to study effects of the wind and the choice of underlying PDF on the rate of spread. It is natural that if there is no wind, for the homogeneous moisture the fire front would grow slower and in all directions, as it is shown in Figure 1. If there is wind (we consider $U = 6.7ms^{-1}$, the fire intensity $I = 20MW$), then the fire propagates in the downwind direction (see Figure 2).

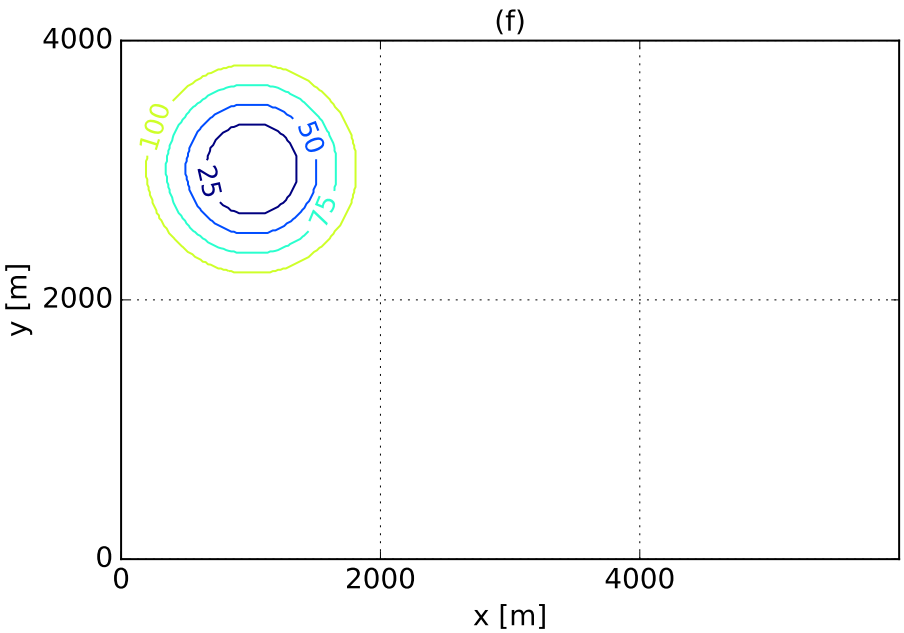


Figure 1: Fire propagation with zero wind.

 ISBN: 978-84-617-8694-7

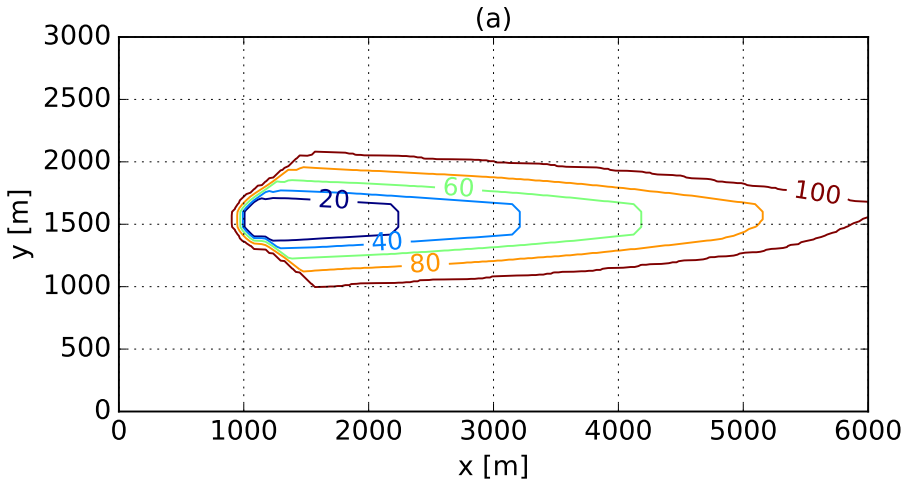


Figure 2: Fire propagation with mean wind $U = 6.7ms^{-1}$.

The model treats the fire-spotting. In order to model this phenomenon the lognormal PDF (8), as it is shown in Figure 3. The point of ignition of the secondary fire also depends on the wind direction. If at some moment $t = t^*$ the wind changes the direction, then new secondary fires appear in new direction, as it is shown in Figure 4. For the sake of simplicity, the initial conditions for new ignitions are chosen the same as for the main fire zone, that causes the same form of the secondary fires.

However, not for any set of the parameters the sire-spotting can be observable. In some cases the fire intensity is not high enough to let the firebrands jump from the main column and produce new independent fire. Moreover, it can occur the situation when the firebrand jumps not that far, so it merges to the main fire changing the curvature of the fire front (see Figure 5).

Acknowledgements

This research is supported by the Basque Government through the BERC 2014-2017 program and by Spanish Ministry of Economy and Competitiveness MINECO through BCAM Severo Ochoa excellence accreditation SEV-2013-0323 and through project MTM2016-76016-R MIP and by the PhD grant "La Caixa 2014".

References

[1] G. Pagnini and A. Mentrelli, "Modelling wildland fire propagation by tracking

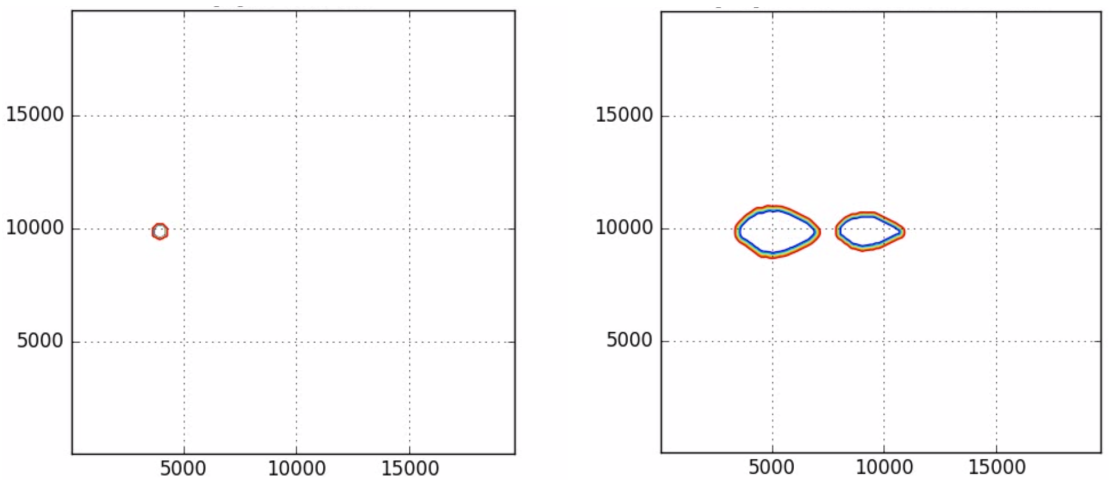


Figure 3: Fire-spotting effect (initial moment $t = 0$ and $t = 74$ min).

random fronts," *Natural Hazards and Earth System Sciences*, vol. 14, no. 8, pp. 2249–2263, 2014.

[2] I. Kaur, A. Mentrelli, F. Bosseur, J.-B. Filippi, and G. Pagnini, "Turbulence and fire-spotting effects into wild-land fire simulators," *Communications in Nonlinear Science and Numerical Simulation*, vol. 39, pp. 300 – 320, 2016.

[3] I. Kaur and G. Pagnini, "Fire-spotting modelling and parametrisation for wild-land fires," *International Congress on Environmental Modelling and Software*, Paper 55, 2016.

[4] M. Sofiev, T. Ermakova, and R. Vankevich, "Evaluation of the smoke-injection height from wild-land fires using remote-sensing data," *Atmospheric Chemistry and Physics*, vol. 12, no. 4, pp. 1995–2006, 2012.

[5] M. I. Asensio and L. Ferragut, "On a wildland fire model with radiation," *InternationalL Journal for Nuerical Methods in Engineering*, vol. 54, pp. 137-157,2002.

[6] J. Mandel, L. S. Bennethuma, J. D. Beezley, J. L. Coenb, C. C. Douglas, M. Kima, and A. Vodacek, "A wildland fire model with data assimilation," *Mathematics and Computers in Simulation*, vol. 79, pp. 584-606,2008.

 ISBN: 978-84-617-8694-7

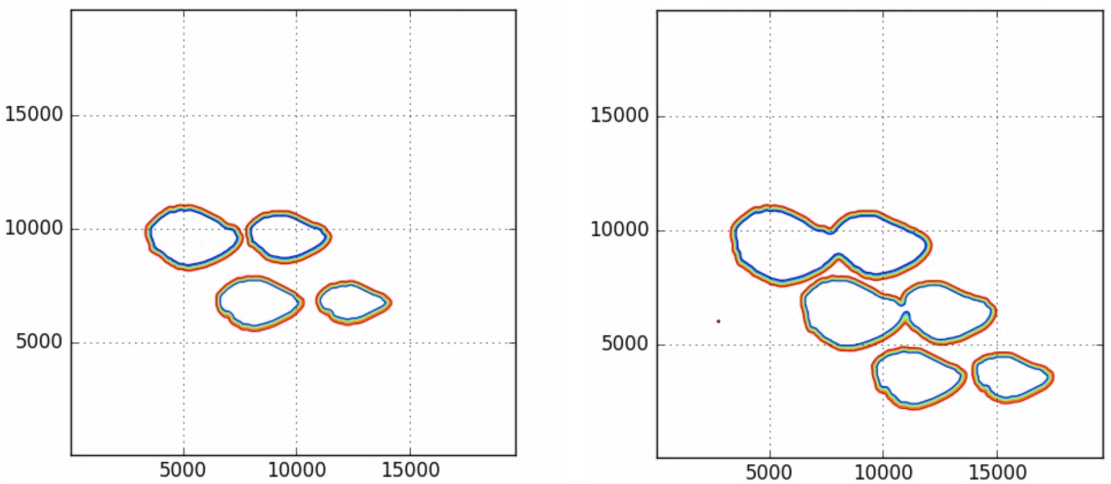


Figure 4: Fire-spotting effect with wind changing ($t = 90$ and $t = 110$ min).

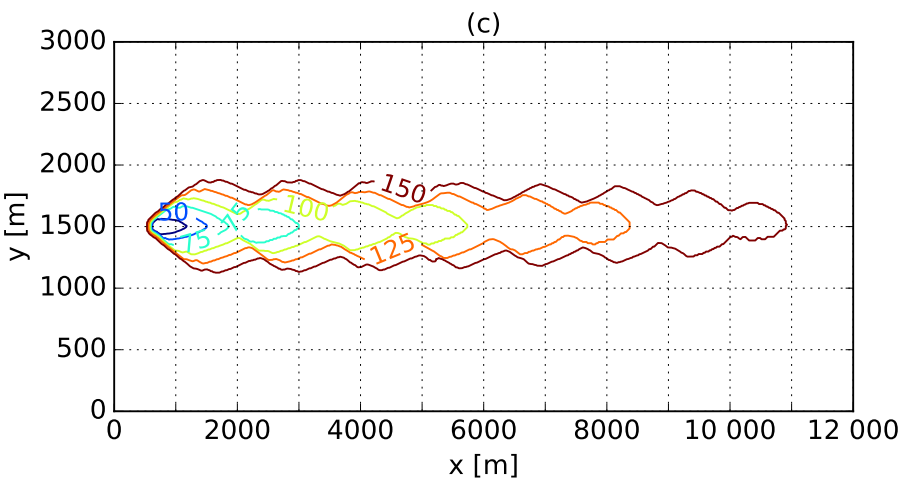


Figure 5: Secondary fire merges to the main fire.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

Improved bisection eigenvalue method for band symmetric Teoplitz matrices

Yuli Eidelman[1] **and Iulian Haimovici**[1]

[1] *Department of Mathematics, Tel-Aviv University*

emails: eideyu@post.tau.ac.il, iulianh@zahav.net.il

Abstract

We specify a new bisection eigenstructure algorithm developed for symmetric quasiseparable of any order matrices to band Toeplitz matrices. It is shown that there is a small perturbation of a band symmetric Toeplitz matrix with the eigendata obtained explicitly and with interlacing property between eigenvalues of the original matrix and of the perturbed one. This allows to improve the performance of the eigenvalue algorithm essentially.

Key words: Toeplitz, quasiseparable, band, eigenstructure, bisection
MSC 2000: 15A18, 15A42, 65N25

Given an $N \times N$ Toeplitz symmetric $2q+1$-band matrix T_q with the only nonzero entries $T_q(i,j) = t_{|i-j|},\ |i-j| \le q,\ i,j = 1,\dots,N$. We perform on the corner of its first entries a small perturbation, such that

$$T_q(i,j) = t_{|i-j|} - t_{i+j}, \quad 2 \le i+j \le q$$

and a centrosymmetric perturbation on the corner of its last entries, no matter how large N is. We prove that if $t_q > 0$, which we can presume without loss of generality, the eigenvalues of the such obtained matrix A_q are

$$\lambda_k^A = t_0 + 2\sum_{j=1}^{q} t_j \cos\left(\frac{jk\pi}{N+1}\right), \quad k = 1,\dots,N, \tag{1}$$

which in the sequel will be sorted in ascending order to almost interlace with the eigenvalues of T_q. More precisely, if the corner matrix $T_q(1:q-1,1:q-1) - A_q(1:q-1,1:q-1)$

 ISBN: 978-84-617-8694-7

is strongly regular and we denote by n the number of its negative eigenvalues and by $p = q - 1 - n$ the number of its positive eigenvalues (including multiplicities), then

$$\lambda^A_{k-2n} \leq \lambda^T_k \leq \lambda^A_{k+2p}, \quad k = 2n+1, \ldots, N-2p.$$

For $q = 2$, if $t_q > 0$, we show that the eigenvalues of A_2 and T_2 almost interlace, more exactly

$$\lambda^A_k \leq \lambda^T_k \leq \lambda^A_{k+2}, \quad k = 1, \ldots, N-2.$$

For 7-band matrices, i.e. $q = 3$, we show that the eigenvalues of A_3 and T_3 satisfy

$$\lambda^A_{k-2} \leq \lambda^T_k \leq \lambda^A_{k+2}, \quad k = 3, \ldots, N-2.$$

As the perturbing matrix is null on $N-2q+2$ rows, i.e. except of two diagonal blocks of size $(q-1)\times(q-1)$, which are a Hankel matrix, we use known results about small symmetric matrix perturbations of a symmetric matrix in order to approximate the eigenvalues of T_q. For instance, we show that if $m_k,\ k = 1, \ldots, N$ are the unwanted differences between the eigenvalues of T_q and A_q, then $\sum_{k=1}^N m_k^2 \leq 2\sum_{j=1}^{q-1} j t_{j+1}^2$ independently on how large is the size N of the matrices T_q and A_q. The approximation is better for larger matrices and/or tridiagonally dominated matrices. For instance for 11-band $10,000 \times 10,000$ matrices with $t_j = t_0/2^j,\ j = 1, \ldots, 5$ we obtain five exact digits. This permits us to find lower and upper bounds when we look only for selected eigenvalues of T_q.

We then specify the bisection algorithm developed for quasiseparable Hermitian matrices to the particular case of symmetric band Toeplitz matrices. We obtain a fast and accurate complete algorithm to solve the eigenproblem in this case. To this end, we use interlacing properties between the eigenvalues of T_q and A_q.

The basic part of the bisection algorithm is the computation of the number of sign changes in the Sturm sequence

$$D_k(\lambda) = \frac{\gamma_k(\lambda)}{\gamma_{k-1}(\lambda)}, \quad k = 1, 2, \ldots, N, \tag{2}$$

where

$$\gamma_0(\lambda) \equiv 1, \gamma_1(\lambda), \gamma_2(\lambda), \ldots, \gamma_N(\lambda)$$

are characteristic polynomials of the principal leading submatrices of a matrix. In our case this number is obtained as follows.

Algorithm. *The number ν of sign changes in the Sturm sequence (2) for a given real value λ and for an $N \times N$ Toeplitz symmetric $2m+1$-band matrix T_m with the given diagonal entry $d \equiv t_0$ and the given row vector $q = (t_1, \cdots, t_m)^T$ of Toeplitz coefficients is computed by the following steps.*

1. Compute $r = m-1, \delta = d - \lambda, v = q^T/\delta; f = vq$. Set $v(m) = q(m)$ and set $\nu = 1$ if $\delta < 0$ and $\nu = 0$ otherwise.

 ISBN: 978-84-617-8694-7

2. Compute $N-2$ *times* $D=\delta-f(1,1)$, $v(1:r)=q(1:r)^T-f(2:m,1)$ *and*

$$\phi=vv^T/D,\quad \phi(1:r,1:r)=\phi(1:r,1:r)+f(2:m,2:m),\quad f=\phi.$$

Set $\nu=\nu+1$ *if* $D<0$.

3. Set $\nu=\nu+1$ *if* $\delta-f(1,1)<0$.

For a given simple eigenvalue λ the normalized eigenvector is obtained as follows.

Algorithm. *The normalized eigenvector for a given simple eigenvalue* λ *and for an* $N\times N$ *Toeplitz symmetric* $2m+1$*-band matrix* T_m *with the given diagonal entry* $d\equiv t_0$ *and the given row vector* $q=(t_1,\cdots,t_m)^T$ *of Toeplitz coefficients is computed by the following steps.*

1.1. Setr $=m-1$, $h_m=(N-\rho)/2$ *where* ρ *is the remainder of the integer division of* N *to* 2 *and* $h_p=h_m+1$, $u(m,N-1)=1$. *Set also* $\delta=d-\lambda$, $u(:,1)=q(:)^T/\delta$, $f=u(:,1)q(:)$, *where* f *is a* $m\times m$ *auxiliary matrix.*

1.2. For $k=2,\ldots,N-1$ *perform:*

$$D=\delta-f(1,1),\quad u(m,k)=q(m)/D,\quad u(1:r,k)=(q(1:r))^T-f(2:m,1:1))/D,$$

$$F=u(:,k)(u(:,k))^TD,\quad F(1:r,1:r)=F(1:r,1:r)+f(2:m,2:m),\quad f=F.$$

2.1. Set $x(N)=1$, $s(2:m)=0$, $s(1)=1$.

2.2. For $k=N-1,N-2,\ldots,h_m$ *set* $n=1$ *and perform:*

$$x(k)=-u(:,k)'s(:),\quad s(2:m)=s(1:r),\quad s(1)=x(k),\quad n=n+(x(k))^2.$$

2.3. Compute $n=\sqrt{2(n-(x(h_m))^2)-\rho(x(h_p))^2}$.

3.1. Set $\sigma=1$ *if* $x(h_m)x(h_p+\rho)>0$ *else* $\sigma=-1$.

3.2. For $j=h_p+1,\ldots,N$ *set* $x(N+1-j)=\sigma x(j)$.

3.3. Normalize $x(1:N)=x(1:N)/n$.

The results of numerical tests demonstrate a good behavior of presented algorithms.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

Positive solutions for second order boundary value problems with sign changing Green's functions

Ricardo Enguiça[1]

[1] *Área Departamental de Matemática, Instituto Superior de Engenharia de Lisboa*

emails: rroque@adm.isel.pt

Abstract

Key words: second order differential equations; Dirichlet boundary conditions, periodic boundary conditions, change sign Green's function.
MSC 2000: 34B15, 34A40

In the literature it has been widely studied the existence of positive solutions for boundary value problems (BVP), namely second order BVP with Periodic and Dirichlet boundary conditions. A standard technique consists on obtaining the existence of positive solutions through Krasnoselskii's fixed point theorem on cones, or to use fixed point index theory. In these cases, the positivity of the associated Green's functions is usually fundamental to prove such results. In this paper we are able to prove existence of solutions for several problems where the associated Green's function changes sign.

Hill's operator properties have been described in several papers, where existence and multiplicity results, comparison principles, Green's functions and spectral analysis were studied.

Positivity results for BVP where the Green's function can vanish are treated for example in [4, 8]. In [4], Graef, Kong and Wang studied the periodic BVP (with $T = 1$ in the paper)

$$u''(t) + a(t)\,u(t) = g(t)\,f(u(t)), \quad u(0) = u(T),\ u'(0) = u'(T), \tag{1}$$

with f and g nonnegative continuous functions and g satisfying $\min_{t\in[0,1]} g(t) > 0$. They assumed the Green's function to be nonnegative and to satisfy the following condition

$$\min_{0\le s\le T} \int_0^T G(t,s)\,dt > 0. \tag{2}$$

Webb ([8]) considered weaker assumptions to prove the existence of positive solutions of the previous problem, but he still assumed the Green's function to be nonnegative.

 ISBN: 978-84-617-8694-7

Despite our results do not require the Green's function to be nonnegative, they could be applied to this particular case, obtaining positive solutions assuming an integral condition weaker than the one above.

On the other hand, some existence results for BVP with sign-changing Green's function have been considered in [3, 6], where the authors asked for the existence of a subinterval $[c,d] \subset [0,T]$, a function $\phi \in L^1([0,T])$ and a constant $c \in (0,1]$ such that the Green's function G satisfies the following condition:

$$|G(t,s)| \leq \phi(s),\ t \in [0,T],\ \ pps \in [0,T], \tag{3}$$

$$G(t,s) \geq c\,\phi(s)\ t \in [c,d],\ \ pps \in [0,T]. \tag{4}$$

It must be pointed out that, if we consider a periodic problem with constant potential $a(t) = \rho^2$ for which the related Green's function changes it's sign (i.e. $\rho > \pi/T$, $\rho \neq 2k\pi/T$, $k = 1,2,...$), the condition above is never fulfilled for any strictly positive function ϕ. This is due to the fact that in such situation the Green's function is constant along the straight lines of slope equals to one (see [2] for details). Meanwhile, our results can be applied without further complications for this case.

Moreover, for Dirichlet BVP with constant potential $a(t) = \rho^2$ with sign change Green's function (i.e. $\rho > \pi/T$, $\rho \neq k\pi/T$, $k = 1,2,...$) it is easy to verify that the condition above holds if and only if ρ^2 lies between the first and the second eigenvalues of the problem ($\frac{\pi}{T} < \rho < \frac{2\,\pi}{T}$) but it is never satisfied for $\rho > \frac{2\,\pi}{T}$. However, our results can be applied for any nonresonant value of $\rho > \pi/T$. Despite this, we must note that the imposed restrictions increase with ρ.

Furthermore, in [3, 6] the authors proved the existence of solutions in the cone

$$K_0 = \left\{ u \in \mathcal{C}[0,T],\ \min_{t\in[c,d]} u(t) \geq c\|u\| \right\},$$

that is, they ensured the positivity of the solutions on the subinterval $[c,d]$ but such solutions were allowed to change sign when considering the whole interval $[0,T]$.

As far as we know, positive solutions for BVP with sign-changing Green's function can be tracked only as back as 2011 in the papers [7, 10]. In the first of these papers, R. Ma considers the following one parameter family of problems,

$$u''(t) + a(t)\,u(t) = \lambda\, g(t)\, f(u(t)),\ t \in (0,T), \quad u(0) = u(T),\ u'(0) = u'(T). \tag{5}$$

By using the Schauder's fixed point Theorem, the author obtains the existence of a positive solution for sufficiently small values of λ. These existence results are not comparable with the ones we will obtain in this paper. On the second paper [10], S. Zhong and Y. An study the following autonomous periodic BVP, with constant potential $\rho \in (0, \frac{3\,\pi}{2\,T}]$.

$$u'' + \rho^2 u = f(u),\ t \in (0,T), \quad u(0) = u(T),\ u'(0) = u'(T). \tag{6}$$

In this case, it is very well known that the related Green's function $G_P(t,s) \geq 0$ for all $\rho \in (0, \frac{\pi}{T}]$ and it changes sign for $\rho \in (\frac{\pi}{T}, \frac{3\,\pi}{2\,T}]$ (see [2]). With this, it can be defined the constant

$$\delta = \begin{cases} \infty & \text{if} \quad \rho \in (0, \frac{\pi}{T}], \\ \inf_{t\in I} \frac{\int_0^T G_P^+(t,s)\,ds}{\int_0^T G_P^-(t,s)\,ds} & \text{if} \quad \rho \in (\frac{\pi}{T}, \frac{3\,\pi}{2\,T}] \end{cases} \tag{7}$$

 ISBN: 978-84-617-8694-7

and using the Krasnoselskii's fixed point Theorem, the authors prove the following existence result:

Theorem. *[10, Theorem 3] Suppose that the following assumptions are fulfilled:*

$(J1)$ $f : [0, \infty) \to [0, \infty)$ *is continuous.*

$(J2)$ $0 \leq m = \inf_{u \geq 0} \{f(u)\}$ *and* $M = \sup_{u \geq 0} \{f(u)\} \leq M \leq \infty$.

$(J3)$ $M/m \leq \delta$, *with* $M/m = \infty$ *when* $m = 0$.

Moreover, if $\delta = \infty$ *assume that*

$$\lim_{x \to \infty} \frac{f(x)}{x} < \rho^2 < \lim_{x \to 0^+} \frac{f(x)}{x}.$$

Then problem stated above has a positive solution on $[0, T]$.

Concerning this specific case, along this work we improve the range of the values ρ *for which the result is still valid. Furthermore, we apply our study to nonconstant potentials and nonautonomous nonlinear parts.*

Some of the positivity conditions imposed in the periodic BVP cannot be adapted for the Dirichlet BVP, so the approach that must be used needs to be considerably modified, by using, in this case, a different type of cones.

References

[1] Amann, H. *Fixed point equations and Nonlinear Problems in Ordered Banach Spaces*, SIAM Review, Vol. 18, 4 (1976), 620–709.

[2] A. Cabada, *Green's functions in the theory of ordinary differential equations. SpringerBriefs in Mathematics*. Springer, New York, 2014.

[3] A. Cabada, G. Infante, F. A. F. Tojo, *Nontrivial solutions of Hammerstein integral equations with reflections*, Bound. Value Prob. 2013, **2013:86**.

[4] J. Graef, L. Kong, H. Wang, *A periodic boundary value problem with vanishing Green's function*, Applied Mathematics Letters **21** (2008), 176–180.

[5] J. Graef, L. Kong, H. Wang, *Existence, multiplicity, and dependence on a parameter for a periodic boundary value problem*, J. Differential Equations **245** (2008), 1185–1197.

[6] G. Infante, P. Pietramala, F. A. F. Tojo, *Nontrivial solutions of local and nonlocal Neumann boundary value problems*, Proc. Roy. Soc. Edinburgh, **146A**, 337-369, 2016.

[7] R. Ma, *Nonlinear periodic boundary value problems with sign-changing Green's function*, Nonlinear Analysis **74** (2011), 1714–1720.

 ISBN: 978-84-617-8694-7

[8] J. Webb, *Boundary value problems with vanishing Green's function*, Communications in Applied Analysis 13 (2009), n 4, 587-596.

[9] A. Zettl, *Sturm-Liouville theory.* Mathematical Surveys and Monographs, 121. American Mathematical Society, Providence, RI, 2005.

[10] S. Zhong, Y. An, *Existence of positive solutions to periodic boundary value problems with sign-changing Green's function*, Boundary Value Problems 2011 (2011) DOI: 10.1186/1687-2770-2011-8.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference on Computational and Mathematical Methods in Science and Engineering, CMMSE 2017 4–8 July, 2017.

Variational Multiscale Proper Orthogonal Decomposition with Modular Regularization

Fatma G. Eroglu[1], Songul Kaya[1] and Leo G. Rebholz[2]

[1] *Department of Mathematics, Middle East Technical University*

[2] *Department of Mathematical Sciences, Clemson University*

emails: fguler@metu.edu.tr, smerdan@metu.edu.tr, rebholz@clemson.edu

Abstract

In this paper, we propose, analyze and test a post-processing implementation of a projection-based variational multiscale (VMS) method with proper orthogonal decomposition (POD) for the incompressible Navier-Stokes equations. The projection-based VMS stabilization is added as a separate post-processing step to the standard POD approximation, and since the stabilization step is completely decoupled, the method can easily be incorporated into existing codes, and stabilization parameters can be tuned independent from the time evolution step. We present a theoretical analysis of the method, and give results for several numerical tests on benchmark problems which both illustrate the theory and show the proposed method's effectiveness.

Key words: proper orthogonal decomposition, projection-based variational multiscale, reduced order models, post-processing

1 Introduction

We consider the incompressible Navier-Stokes equations (NSE) on a polyhedral domain $\Omega \subset \mathbb{R}^d, d \in \{2,3\}$ with boundary $\partial\Omega$:

$$\begin{aligned} \mathbf{u}_t - \nu\Delta\mathbf{u} + (\mathbf{u}\cdot\nabla)\mathbf{u} + \nabla p &= \mathbf{f} \quad \text{in } (0,T]\times\Omega, \\ \nabla\cdot\mathbf{u} &= 0 \quad \text{in } [0,T]\times\Omega, \\ \mathbf{u} &= \mathbf{0} \quad \text{in } [0,T]\times\partial\Omega, \\ \mathbf{u}(0,\mathbf{x}) &= \mathbf{u}_0 \quad \text{in } \Omega, \\ \int_\Omega p\, d\mathbf{x} &= 0, \quad \text{in } (0,T]. \end{aligned} \tag{1}$$

 ISBN: 978-84-617-8694-7

Here, $\mathbf{u}(t,\mathbf{x})$ is the fluid velocity and $p(t,\mathbf{x})$ the fluid pressure. The parameters in (1) are the kinematic viscosity $\nu > 0$, the prescribed body forces $\mathbf{f}(t,\mathbf{x})$ and the initial velocity field $\mathbf{u}_0(\mathbf{x})$. Simulating complex flows by a direct numerical simulation (DNS) can be very expensive. In this case reduced order models are needed to decrease computational cost. The proper orthogonal decomposition (POD) approach is most widely used reduced order model. POD only uses the most energetic base functions. However, POD causes numerical instability in turbulent flows. Using VMS in POD was pioneered in [1, 2], and their studies showed this could increased numerical accuracy for convection-dominated convection-diffusion equations [2] and for NSE [1].

2 Post-Processed VMS-POD Schemes

We proposes fully discrete VMS-POD methods. We analyze the backward Euler temporal discretization. We consider the extension to BDF2 time stepping. In our analysis, we assume that the eddy viscosity coefficient ν_T is known bounded, positive and element-wise constant. In that analysis that follows, we denote variables at time $t^n = n\Delta t, n = 0, 1, 2, \ldots, M, T := M\Delta t$ using superscripts, e.g. $\mathbf{f}^n := \mathbf{f}(t^n)$. The two step VMS-POD scheme equipped with backward Euler time stepping reads as follows:

Algorithm 2.1 *Let $\mathbf{f} \in L^2(0,T;\mathbf{H}^{-1}(\Omega))$ and $\mathbf{u}_r^0 = \mathbf{w}_r^0$ be given with L^2 projection of $\mathbf{u}_0$ in $\mathbf{X}^r$. Given $\mathbf{u}_r^n \in \mathbf{X}_r$ compute $\mathbf{u}_r^{n+1}$ by applying the following two steps:*
Step 1. *Calculate $\mathbf{w}_r^{n+1} \in \mathbf{X}^r$ satisfying $\forall \psi \in \mathbf{X}^r$,*

$$\left(\frac{\mathbf{w}_r^{n+1} - \mathbf{u}_r^n}{\Delta t}, \psi\right) + b(\mathbf{w}_r^{n+1}, \mathbf{w}_r^{n+1}, \psi) + \nu(\nabla \mathbf{w}_r^{n+1}, \nabla\psi) = (\mathbf{f}^{n+1}, \psi). \tag{2}$$

Step 2. *Post-process $\mathbf{w}_r^{n+1}$ by applying projection P_R to obtain $\mathbf{u}_r^{n+1} \in \mathbf{X}_r$, $\forall \psi \in \mathbf{X}^r$:*

$$\left(\frac{\mathbf{w}_r^{n+1} - \mathbf{u}_r^{n+1}}{\Delta t}, \psi\right) = (\nu_T(I - P_R)\nabla\frac{(\mathbf{w}_r^{n+1} + \mathbf{u}_r^{n+1})}{2}, (I - P_R)\nabla\psi), \tag{3}$$

We note that Step 1 is the standard Galerkin POD method, and Step 2 is completely decoupled VMS stabilization step. The projection in Step 2 is not a filter but constructed to recover VMS eddy viscosity term.

Lemma 2.1 *(Stability of Algorithm 2.1) The post-processed-VMS-POD approximation (2)-(3) is unconditionally stable in the following sense: for any $\Delta t > 0$,*

$$\|\mathbf{u}_r^M\|^2 + \sum_{n=0}^{M-1}\Big[2\nu_T\Delta t\|(I - P_R)\nabla\frac{(\mathbf{w}_r^{n+1} + \mathbf{u}_r^{n+1})}{2}\|^2 + \|\mathbf{w}_r^{n+1} - \mathbf{u}_r^n\|^2 + \nu\Delta t\|\nabla\mathbf{w}_r^{n+1}\|^2\Big] \le \|\mathbf{u}_r^0\|^2 + \nu^{-1}|||\mathbf{f}|||_{2,-1}^2.$$

 ISBN: 978-84-617-8694-7

We present the error analysis of the true solution of Navier Stokes equations and VMS-POD approximation (2)-(3). The optimal asymptotic error estimation requires the following regularity assumptions for the true solution:

$$\mathbf{u} \in L^\infty(0,T;H^{m+1}(\Omega)) \quad p \in L^\infty(0,T;H^m(\Omega)) \quad \mathbf{u}_{tt} \in L^2(0,T;H^1(\Omega))$$
$$\mathbf{f} \in L^2(0,T;H^{-1}(\Omega)) \tag{4}$$

Theorem 2.1 *Suppose (4) holds and $\mathbf{u}_r^n$ and $\mathbf{w}_r^n$ given by Algorithm 2.1. For sufficiently small Δt, i.e. $\Delta t \le [C\nu^{-3}\|\nabla \mathbf{u}\|_{\infty,0}^4]^{-1}$ we have the following asymptotic error estimation:*

$$\|\mathbf{u}^M - \mathbf{u}_r^M\|^2 + \sum_{n=0}^{M-1}\left[\frac{1}{4}\Delta t\nu_T\|(I-P_R)\nabla(\mathbf{u}^{n+1} - (\mathbf{u}_r^{n+1}+\mathbf{w}_r^{n+1})/2)\|^2\right.$$
$$\left.+\nu\Delta t\|\nabla(\mathbf{u}^{n+1}-\mathbf{w}_r^{n+1})\|^2\right] \le C\Big(h^{2m} + (\Delta t)^2 + (1+\|S_R\|_2+\|S_r\|_2)h^{2m+2}$$
$$+\sum_{j=R+1}^{d}\|\psi_j\|_1^2\lambda_j + \sum_{j=r+1}^{d}(1+\|\psi_j\|_1^2)\lambda_j\Big)$$

We consider now an extension of Algorithm 2.1 to BDF2 time stepping.

Algorithm 2.2 *Let $\mathbf{f} \in L^2(0,T;\mathbf{H}^{-1}(\Omega))$ and initial conditions $\mathbf{u}_r^0$ and $\mathbf{u}_r^{-1}$ be given in $\mathbf{X}^r$. Then for n=0,1,2,...*

Step 1. *Calculate $\mathbf{w}_r^{n+1} \in \mathbf{X}^r$ satisfying $\forall \psi \in \mathbf{X}^r$,*

$$\left(\frac{3\mathbf{w}_r^{n+1} - 4\mathbf{u}_r^n + \mathbf{u}_r^{n-1}}{2\Delta t}, \psi\right) + b(\mathbf{w}_r^{n+1}, \mathbf{w}_r^{n+1}, \psi) + \nu(\nabla \mathbf{w}_r^{n+1}, \nabla\psi) = (\mathbf{f}^{n+1}, \psi) \tag{5}$$

Step 2. *Post-process $\mathbf{w}_r^{n+1}$ to obtain $\mathbf{u}_r^{n+1} \in \mathbf{X}^r$ satisfying $\forall \psi \in \mathbf{X}^r$,*

$$\left(\frac{\mathbf{w}_r^{n+1} - \mathbf{u}_r^{n+1}}{\Delta t}, \psi\right) = (\nu_T(I-P_R)\nabla\frac{(\mathbf{w}_r^{n+1}+\mathbf{u}_r^{n+1})}{2}, (I-P_R)\nabla\psi). \tag{6}$$

We note the post-processing step is exactly the same as in the backward Euler case. Also as in the case of the backward Euler method above, without Step 2, Algorithm 2.2 reduces to the classical Galerkin POD formulation for the NSE, although now using BDF2 time stepping.

Lemma 2.2 *(Stability of Algorithm 2.2) The post-processed VMS-POD approximation (5)-(6) is stable for the eddy viscosity term $\nu_T < 4\nu$ in the following sense:*

$$\|\mathbf{u}_r^{M+1}\|^2 + \|2\mathbf{u}_r^{M+1} - \mathbf{u}_r^M\|^2 + 2\nu_T\Delta t\left\|(I-P_R)\nabla\frac{(\mathbf{w}_r^{M+1}+\mathbf{u}_r^{M+1})}{2}\right\|^2$$
$$+2\nu\Delta t\|\nabla\mathbf{w}_r^{M+1}\|^2 + \sum_{n=1}^{M}\|\mathbf{w}_r^{n+1} - 2\mathbf{u}_r^n + \mathbf{u}_r^{n-1}\|^2 + (4\nu-\nu_T)\frac{\Delta t}{2}\sum_{n=1}^{M-1}\|\nabla\mathbf{w}_r^{n+1}\|^2$$
$$\le \|\mathbf{u}_r^1\|^2 + \|2\mathbf{u}_r^1 + \mathbf{u}_r^0\|^2 + \frac{\nu_T\Delta t}{2}\|\nabla\mathbf{u}_r^1\|^2 + 2\nu^{-1}|||\mathbf{f}|||_{2,-1}^2.$$

 ISBN: 978-84-617-8694-7

3 Numerical Experiments

We test the VMS-POD approximate solution with three numerical experiments. In all cases we use Algorithm 2.2, i.e. the scheme with second order time stepping. Our first test considers the predicted convergence rates of the previous section, with respect to varying R and Δt. For the second test, we compare accuracy of the proposed VMS-POD scheme compared with the usual Galerkin POD method (i.e. unstabilized POD, computed by eliminating the post-processing step of the VMS-POD) in 2D channel flow past a cylinder. Finally we consider the VMS-POD for a 3D turbulent channel flow simulation.

4 Conclusion

We proposed, analyzed and tested a VMS-POD method for incompressible NSE simulation, where the stabilization is completely decoupled into the second step of a two step implementation at each time step. Decoupling of the stabilization has the advantage of easily being incorporated into existing POD-G codes, and also that stabilization parameters can be adjusted only in the stabilization step (and not as part of the evolution equation). We rigorously prove an error estimate for the model, in terms of the number of POD modes r, the stabilization parameters R (number of modes not to add stabilization to) and ν_T, as well as the time step size Δt and the mesh width h of the underlying FEM simulation that produced the POD modes. Results from several numerical experiments are provided that show how effective the method can be. In particular, we show for 2D channel flow past a step, POD-G has an energy growth that causes poor lift and drag prediction, especially for longer times. The proposed VMS-POD is able to fix this by stabilizing so that the energy matches the DNS energy, which in turn leads to excellent lift and drag prediction, even up to $t = 10$ (and from the plots, it appear the accurate predictions can continue for even longer time).

References

[1] T. Iliescu and Z. Wang, *Variational multiscale proper orthogonal decomposition: Navier-Stokes equations*, Numer. Meth. Partial. Diff. Eqs. **30(2)** (2014) 641–663.

[2] T. Iliescu and Z. Wang, *Variational multiscale proper orthogonal decomposition: Convection-dominated convection-diffusion-reaction equations*, Mathematics of Computation **82(283)** (2013) 1357–1378.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

Graded contractions of filiform Lie algebras

José M. Escobar[1], Juan Núñez[1] and Pedro Pérez-Fernández[2]

[1] *Departamento de Geometría y Topología, Facultad de Matemáticas. Universidad de Sevilla.*

[2] *Departamento de Física Aplicada III, Escuela Técnica Superior de Ingeniería. Universidad de Sevilla (Spain).*

emails: pinchamate@gmail.com, jnvaldes@us.es, pedropf@us.es

Abstract

The study of contractions of Lie algebras is profusely extended in the last decades. In this paper we study the graded contractions of some lower-dimensional filiform Lie algebras which have not been studied earlier. Particularly, we deal with graded contractions of model filiform Lie algebras of dimension less than or equal to 6 and with the ones of a non-model 6-dimensional filiform Lie algebra.

Key words: Graded contractions; filiform Lie algebras.
MSC 2000: 17B30; 17B40; 17B51.

1 Introduction

One of the most relevant and useful results in Physics is the so called *Correspondence principle*, which sets that a new theory should coincide with the old one in predictions for phenomena where these conditions are satisfied.

The mathematical formulation of this principle for relativistic mechanics was given by In*ö*n*ü* and Wigner [3] when introducing the In*ö*n*ü*-Wigner contractions (IW-contractions).

In this paper we concentrate on the algebraical approach to contractions, so called *graded contractions*, which were originally introduced in [4] as a generalization of IW-contractions.

There are two types of graded contractions: continuous graded contractions, which correspond to IW-contractions and discrete graded contractions, which possess no equivalent in continuous contractions. The general solution of the graded contractions, considering

only so called generic case, was achieved in [7, 8]. Since this solution depends solely on the grading group (the structure of the Lie algebra does not matter at all), it is obtained simultaneously for all Lie algebras which allow the given grading. However, this approach is in a certain sense too general. It motivates our study.

Indeed, graded contraction of several types of Lie algebras have been already dealt in previous papers. For instance, Novotny studied in deep graded contractions of the simple Lie algebra $sl(3,\mathbb{C})$ [5]. He showed 4 gradings for this algebra. Later, Novotny himself obtained the contractions for each grading. Bahturin, Goze y Remm [1] classified, up to isomorphism, gradings by abelian groups on nilpotent Lie algebras of nonzero rank and, in the case of rank 0, they described conditions to obtain non trivial Z_k-gradings.

So, by continuing with this study, which we began to deal with in [2], we show in the paper the graded contractions of the model filiform Lie algebras of dimension 3 and 4. These ones, together with those of dimensions 5 and 6, which have been also obtained by us, but are not included in the paper for reasons of length, have allowed us to study the general case of the contractions of n-dimensional model filiform Lie algebras, which is dealt with in Section 5. Moreover, with the objective of comparing the model case with the non-model one, the graded contractions of a non-model 6-dimensional filiform Lie algebra has been also obtained. It is convenient to say that our motivation for dealing with this type of algebras is due to the fact of that these algebras, which were introduced by Vergne in 1966, in her Ph. D. Thesis, later published in 1979 [6], constitute the most structured subset of nilpotent Lie algebras.

Let us now recall brief preliminaries on this subject.

2 Preliminaries

Let $\mathfrak{g}$ be a finite dimensional complex Lie algebra over $\mathbb{C}$. A decomposition $\Gamma : \mathfrak{g} = \oplus_{i\in I}\mathfrak{g}_i$ of the vector space $\mathfrak{g}$ into a direct sum of vector subspaces $\mathfrak{g}_i \neq 0$, $i \in I$, is called a *grading* of $\mathfrak{g}$ if for any pair of indices $i, j \in I$, there exists $k \in I$, such that $[\mathfrak{g}_i, \mathfrak{g}_j] \subset \mathfrak{g}_k$. Vector subspaces $\mathfrak{g}_i$ are called *grading subspaces*. The number of grading subspaces is equal to the cardinality $|I|$ of the index set I.

A grading $\Gamma : \mathfrak{g} = \oplus_{i\in I}\mathfrak{g}_i$ is called *group grading* (respectively, *semigroup grading*) if there exist an abelian group (resp. semigroup) G and an injective mapping $f : I \to G$ such that for any pair of indices $i, j \in I$, the equality $f(i \circ j) = f(i) + f(j)$ holds, where $+$ denotes the binary operation in G. The group (resp. semigroup) G is called *grading group (resp. semigroup)*.

The *universal group* is the Abelian finitely generated group U which contains the set of indices J of a grading group $J \subset U$.

Let $\Gamma : \mathfrak{g} = \oplus_{i\in I}\mathfrak{g}_i$ be a grading of the Lie algebra $\mathfrak{g}$, with $|I| = m \in \mathbb{N}$ grading subspaces. A complex Lie algebra $\mathfrak{g}^{\varepsilon}$ endowed with a Lie bracket $[,]_{\varepsilon}$ and satisfying the

 ISBN: 978-84-617-8694-7

two conditions: *i)* the underlying vector space of $\mathfrak{g}^\epsilon$ is the underlying vector space of $\mathfrak{g}$, i.e. $\mathfrak{g}^\epsilon = \oplus_{i\in I}\mathfrak{g}_i$, and *ii)* for all $i, j \in I$, there exists $\varepsilon_{ij} \in \mathbb{C}$ such that $[x, y]_\varepsilon = \varepsilon_{ij}[x, y]$, for all $x \in \mathfrak{g}_i$ and $y \in \mathfrak{g}_j$, is called Γ-*graded contraction* of the Lie algebra $\mathfrak{g}$.

We define the *contraction matrix* ε of the Lie algebra $\mathfrak{g}^\varepsilon$ as a matrix whose elements are ε_{ij}. This matrix determines the Lie algebra $\mathfrak{g}^\varepsilon$.

The elements ε_{ij} associated with Lie brackets verifying $[\mathfrak{g}_i, \mathfrak{g}_j] \neq 0$ will be called *relevant* elements. On the contrary, those elements verifying $[\mathfrak{g}_i, \mathfrak{g}_j] = 0$ will be called *non-relevant* elements and they will be considered null. The null relevant elements will be called *singular* elements. The set of the pairs (i, j) such that ε_{ij} is a relevant element of the contraction matrix ε will be denoted by $\mathcal{I}$.

Let P_n denote the symmetric group of the set $\{1, 2, ..., n\}$. We define an equivalence relation on I^n as follows: two n-tuples $(x_1, ..., x_n), (y_1, ..., y_n) \in I^n$ are equivalent if and only if there exists $\sigma \in \Pi_n$ such that $x_i = y_{\sigma(i)}$, for all $i = 1, ..., n$. The classes $(x_1\, x_2 \ldots x_n) = \{(x_{\sigma 1}, ..., x_{\sigma n}) | \sigma \in \Pi_n\}$ defined by this relation are called *unordered n-tuples* and the set of all unordered n-tuples with entries in I is denoted by I_u^n.

The *lower central series* of a Lie algebra $\mathfrak{g}$ is defined as $\mathfrak{g}^1 = \mathfrak{g},\ \mathfrak{g}^2 = [\mathfrak{g}^1, \mathfrak{g}],\ \ldots,\ \mathfrak{g}^k = [\mathfrak{g}^{k-1}, \mathfrak{g}],\ \ldots$.

If there exists $m \in \mathbb{N}$ such that $\mathfrak{g}^m \equiv 0$, then $\mathfrak{g}$ is called *nilpotent*. An n-dimensional nilpotent Lie algebra $\mathfrak{g}$ is said to be *filiform* if it is verified that $dim\, \mathfrak{g}^k = n - k$, for all $k \in \{2, \ldots, n\}$. A n-dimensional filiform Lie algebra is called *model* if the only nonzero brackets are $[e_1, e_k] = e_{k+1}$, for $2 \leq k \leq n - 1$, where $\{e_1, \ldots, e_n\}$ is an adapted basis.

3 Graded contractions of the model filiform Lie algebra of dimension 3

All the computations needed to obtain the graded contractions of model filiform Lie algebras will be only detailed in this dimension 3. For the next dimensions, these computations will be only indicated.

Let $\mathfrak{f}_3 : [e_1, e_2] = e_3$ be the model filiform Lie algebra of dimension 3. A grading of $\mathfrak{f}_3$ is given by (see [1])

$$\Gamma : \mathfrak{f}_3 = \mathfrak{f}^3_{(1,0)} \oplus \mathfrak{f}^3_{(0,1)} \oplus \mathfrak{f}^3_{(1,1)}$$

The universal group U of $\mathfrak{f}_3$ is $\mathbb{Z}_2 \otimes \mathbb{Z}_2$ and $\mathfrak{f}^3_{(1,0)} = \langle e_1 \rangle$, $\mathfrak{f}^3_{(0,1)} = \langle e_2 \rangle$ and $\mathfrak{f}^3_{(1,1)} = \langle e_3 \rangle$.

Let us now denote

$$I = \{(1, 0), (0, 1), (1, 1)\}$$

and let us consider a order relation on I

 ISBN: 978-84-617-8694-7

$$O : \begin{cases} 1 & \to & (1,0) \\ 2 & \to & (0,1) \\ 3 & \to & (1,1) \end{cases}$$

Let us consider

$$\Pi_3 = \begin{pmatrix} 1 & 0 \\ a & 1 \end{pmatrix}, \text{ with } a \in \mathbb{Z}_2,$$

and the set $G_{\Pi_3} = \{g \in Aut(\mathfrak{f}^3) \mid \exists \pi \in \Pi_3 \mid g(\mathfrak{f}_i^3) = \mathfrak{f}_{\pi(i)}^3, \text{ for all } i \in I\}$.

We define π as the mapping $I \mapsto I$, which maps an element $i \in I$ into the matrix product in I. Then, it is easy to see that G_{Π_3} is a subgroup of $Aut(\mathfrak{f}^3)$. Indeed, if $g_1, g_2 \in G_{\Pi_3}$, then there exist $\pi_1, \pi_2 \in \Pi_3$, such that $g_1(f_i^3) = f_{\pi_1(i)}^3$ and $g_2(f_i^3) = f_{\pi_2(i)}^3$, for all $i \in I$. Therefore, $(g_1 g_2^{-1}) f_i^3 = g_1(g_2^{-1}(f_{\pi_2(\pi_2^{-1}(i))}^3)) = g_1(f_{\pi_2^{-1}(i)}^3) = f_{\pi_1(\pi_2^{-1}(i))}^3$. Therefore, $g_1 g_2^{-1} \in G_{\Pi_3}$, and thus G_{Π_3} is a subgroup of de $Aut(\mathfrak{f}^3)$.

1. Orbits of I.

Let us recall that the concept of *orbit* is the following: If G is a group acting on a set I, the *orbit* of an element x in I is the set of elements in I to which x can be moved by the elements of G, that is, $G \cdot x = \{g \cdot x \mid g \in G\}$.

Now, let us see how we can obtain the orbits of I.

Let

$$\pi_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \qquad \pi_2 = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}$$

be the generators of Π_3. Then

$$\pi_1((1,0)) = (1,0)\pi_1 = (1,0) \in I, \quad \pi_2((1,0)) = (1,0)\pi_2 = (1,0) \in I.$$

Therefore, $(1,0)$ represents a orbit which contains itself. Similarly, as

$$\pi_1((0,1)) = (0,1)\pi_1 = (0,1) \in I, \pi_2((0,1)) = (0,1)\pi_2 = (1,1) \in I,$$

$(0,1)$ represents a orbit containing itself and the index $(1,1)$. So, we have

Represented by the points	Orbits
$(1,0)$	$(1,0)$
$(0,1)$	$(0,1)$ and $(1,1)$

Similarly, we obtain and show in the following tables the rest of orbits.

2. Orbits of the 3 points of I_u^2

These orbits are obtained by the following definition $\pi_i((p,q)(r,s)) = (\pi_i(p,q)\pi_i(r,s))$, for all $i \in \{1,2,3\}$. We obtain that

 ISBN: 978-84-617-8694-7

Orbits	Represented by the points
$((1,0)(0,1))$	$((1,0)(0,1))$ and $((1,0)(1,1))$
$((0,1)(1,1))$	$((0,1)(1,1))$

3. Orbit of the unique point of I_u^3

Similarly, these orbits are obtained by the following definition $\pi_i((m,n)(p,q)(r,s)) = (\pi_i(m,n)\pi_i(p,q)\pi_i(r,s))$, for all $i \in \{1,2,3\}$. We have

Orbits	Represented by the points
$((1,0)(0,1)(1,1))$	$((1,0)(0,1)(1,1))$

Let us observe that this orbit contains an unique triple, which is the one of indices of the vectors which satisfy Jacobi Identity.

4. Orbit of the 3 points of $\mathcal{I}$:

The non-relevant elements of the contraction matrix ε which might be different from zero are $\varepsilon_{(1,0)(0,1)}$ and $\varepsilon_{(1,0)(1,1)}$. We obtain

Orbits	Represented by the points
$((1,0)(0,1))$	$((1,0)(0,1))$ and $((1,0)(1,1))$

However, $\varepsilon_{(1,0)(1,1)} = 0$, because $0 = [\mathfrak{f}^3_{(1,0)}, \mathfrak{f}^3_{(1,1)}]_\varepsilon = \varepsilon_{(1,0)(1,1)}\,[\mathfrak{f}^3_{(1,0)}, \mathfrak{f}^3_{(1,1)}] = \varepsilon_{(1,0)(1,1)}$ $[\mathfrak{f}^3_{\pi_2(1,0)}, \mathfrak{f}^3_{\pi_2(1,1)}] = \varepsilon_{(1,0)(1,1)} g_2\,[\mathfrak{f}^3_{(1,0)}, \mathfrak{f}^3_{(0,1)}]$, but the bracket $[\mathfrak{f}^3_{(1,0)}, \mathfrak{f}^3_{(0,1)}] \neq 0$. So, the element $\varepsilon_{(1,0)(1,1)}$ of the contraction matrix is singular.

The explicit form of the contraction matrix ε with respect to the chosen order O is

$$\begin{pmatrix} 0 & \varepsilon_{(1,0)(0,1)} & 0 \\ \varepsilon_{(1,0)(0,1)} & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

Now, we are trying to find if the relevant elements of the contraction matrix verify some particular conditions. To do this, the elements of this matrix have to satisfy the following conditions

By imposing the Jacobi identity

$$[X,[Y,Z]_\varepsilon]_\varepsilon + [Y,[Z,X]_\varepsilon]_\varepsilon + [Z,[X,Y]_\varepsilon]_\varepsilon = 0 \qquad (1)$$

for all $X \in \mathfrak{f}^3_{(i,j)}$, $Y \in \mathfrak{f}^3_{kl}$, $Z \in \mathfrak{f}^3_{(m,n)}$ and for all $(i,j)(k,l)(m,n) \in I_u^3$, we obtain that

$$\varepsilon_{(i,j)(k+m,l+n)}\varepsilon_{(k,l)(m,n)}[X_{i,j},[X_{k,l},X_{m,n}]] + \varepsilon_{(k,l)(m+i,n+j)}\varepsilon_{(m,n)(i,j)}[X_{k,l},[X_{m,n},X_{i,j}]]$$

$$+\varepsilon_{(m,n)(i+k,j+l)}\varepsilon_{(i,j)(k,l)}[X_{m,n},[X_{i,j},X_{k,l}]] = 0, \qquad (2)$$

 ISBN: 978-84-617-8694-7

for all $((i,j)(k,l)(m,n)) \in I_u^3$.

From this expression and for all $\pi \in \Pi_3$, it is deduced that

$\varepsilon_{\pi(i,j)\pi(k+m,l+n)}\varepsilon_{\pi(k,l)\pi(m,n)}[X_{\pi(i,j)},[X_{\pi(k,l)},X_{\pi(m,n)}]] + \varepsilon_{\pi(k,l)\pi(m+i,n+j)}\varepsilon_{\pi(m,n)\pi(i,j)}$ $[X_{\pi(k,l)},[X_{\pi(m,n)},X_{\pi(i,j)}]] + \varepsilon_{\pi(m,n)\pi(i+k,j+l)}\varepsilon_{\pi(i,j)\pi(k,l)}[X_{\pi(m,n)},[X_{\pi(i,j)},X_{\pi(k,l)}]] = 0$, for all $((i,j)(k,l)(m,n)) \in I_u^3$.

Let now $g \in G_{\Pi_3}$ be such that $g(X_{(k,l)}) = X_{\pi(k,l)}$, for all $(k,l) \in I$. We have the three following restrictions for the elements of the contraction matrix

$\varepsilon_{\pi(i,j)\pi(k+m,l+n)}\varepsilon_{\pi(k,l)\pi(m,n)}\ g[X_{(i,j)},[X_{(k,l)},X_{(m,n)}]] + \varepsilon_{\pi(k,l)\pi(m+i,n+j)}\varepsilon_{\pi(m,n)\pi(i,j)}\ g[X_{(k,l)},$ $[X_{(m,n)},X_{(i,j)}]]+\varepsilon_{\pi(m,n)\pi(i+k,j+l)}\varepsilon_{(i,j)(k,l)}g[X_{(m,n)},[X_{(i,j)},X_{(k,l)}]] = 0$, for all $((i,j)(k,l)(m,n))$ $\in I_u^3$.

$g(\varepsilon_{\pi(i,j)\pi(k+m,l+n)}\varepsilon_{\pi(k,l)\pi(m,n)}\ [X_{(i,j)},[X_{(k,l)},X_{(m,n)}]] + \varepsilon_{\pi(k,l)\pi(m+i,n+j)}\varepsilon_{\pi(m,n)\pi(i,j)}\ [X_{(k,l)},$ $[X_{(m,n)},X_{(i,j)}]]\varepsilon_{\pi(m,n)\pi(i+k,j+l)}\varepsilon_{(i,j)(k,l)}[X_{(m,n)},[X_{(i,j)},X_{(k,l)}]]) = 0$, for all $((i,j)(k,l)(m,n))$ $\in I_u^3$.

$\varepsilon_{\pi(i,j)\pi(k+m,l+n)}\varepsilon_{\pi(k,l)\pi(m,n)}\ [X_{(i,j)},[X_{(k,l)},X_{(m,n)}]] + \varepsilon_{\pi(k,l)\pi(m+i,n+j)}\varepsilon_{\pi(m,n)\pi(i,j)}[X_{(k,l)},$ $[X_{(m,n)},X_{(i,j)}]]+\varepsilon_{\pi(m,n)\pi(i+k,j+l)}\varepsilon_{(i,j)(k,l)}[X_{(m,n)},[X_{(i,j)},X_{(k,l)}]] = 0$, for all $((i,j)(k,l)(m,n)) \in$ I_u^3.

These expressions allow us to obtain the restrictions which verify the rest of elements of the contraction matrix. It implies that the relevant elements $\varepsilon_{\pi(i,j)}$ are also elements of that matrix and satisfy the same conditions as $\varepsilon_{(i,j)}$.

Moreover, $[X_{i,j},[X_{k,l},X_{m,n}]]$, $[X_{m,n},[X_{i,j},X_{k,l}]]$ and $[X_{k,l},[X_{m,n},X_{i,j}]]$ are null for $\mathfrak{f}_3$. This implies that the element $\varepsilon_{(1,0)(0,1)}$ can take any complex value.

On the other hand, if $\varepsilon = (\varepsilon_{ij})$ is a contraction matrix, then we define $\tau = (\tau_{ij})$ such that $\tau_{ij} = \frac{1}{\varepsilon_{ij}}$, if $\varepsilon_{ij} \neq 0$ or $\tau_{ij} = 0$, otherwise. Besides, $\varepsilon \diamond \tau$ ($\diamond$ means the Hadamard product, that is the binary operation that takes two matrices of the same dimensions, and produces another matrix where each element pq is the product of elements pq of the original two matrices) is a contraction matrix in which all non-null elements are 1. We call *normalized contraction matrix* of ε to the matrix $\varepsilon \diamond \tau$, and we denote by $N(\mathfrak{f}_3)$ to the set of all normalized contraction matrices of $\mathfrak{f}_3$. This set has 2 elements, which are the 3×3 null matrix and the matrix $\begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$.

 ISBN: 978-84-617-8694-7

4 Graded contractions of the model filiform Lie algebra of dimension 4

Let $\mathfrak{f}_4 : [e_1, e_k] = e_{k+1}$, for $2 \leq k \leq 3$ be the model filiform Lie algebra of dimension 4. A grading of $\mathfrak{f}_4$ is given by

$$\Gamma : \mathfrak{f}^4 = \mathfrak{f}^4_{(1,0)} \oplus \mathfrak{f}^4_{(0,1)} \oplus \mathfrak{f}^4_{(1,1)} \oplus \mathfrak{f}^4_{(2,1)}$$

The universal group of $\mathfrak{f}_4$ is $\mathbb{Z}_3 \otimes \mathbb{Z}_2$ and $\mathfrak{f}^4_{(1,0)} = \langle e_1 \rangle$, $\mathfrak{f}^4_{(0,1)} = \langle e_2 \rangle$, $\mathfrak{f}^4_{(1,1)} = \langle e_3 \rangle$ and $\mathfrak{f}^4_{(2,1)} = \langle e_4 \rangle$.

Let us consider

$$\Pi_4 = \begin{pmatrix} 1 & 0 \\ a & 1 \end{pmatrix}, \text{ with } a \in \mathbb{Z}_3 \text{ and } \mathrm{I} = \{(1,0),(0,1),(1,1),(2,1)\}.$$

We now consider the following order O on $I : 1 \to (1,0), 2 \to (0,1), 3 \to (1,1)$ and $4 \to (2,1)$. By straightforward computations, we obtain that the elements of I constitute the following orbits

Represented by the points	Orbits
$(1,0)$	$(1,0)$
$(0,1)$	$(0,1),(1,1)$ and $(2,1)$

Similarly, we show in the following tables the following orbits

2. Orbits of the 6 points of I^2_u

Orbits	Represented by the points
$((1,0)(0,1))$	$((1,0)(0,1)),((1,0)(1,1))$ and $((1,0)(2,1))$
$((0,1)(1,1))$	$((0,1)(1,1)),((1,1)(2,1))$ and $((2,1)(0,1))$

3. Orbits of the 4 point of I^3_u

Orbits	Represented by the points
$((1,0)(0,1)(1,1))$	$((1,0)(0,1)(1,1)),((1,0)(1,1)(2,1))$ and $((1,0)(2,1)(0,1))$
$((0,1)(1,1)(2,1))$	$((0,1)(1,1)(2,1))$

Let us observe that these two orbits contain $\binom{4}{3} = 4$ triples, which correspond with the indices of the triples of vectors which must satisfy the Jacobi Identity.

4. Orbits of the 3 points of $\mathcal{I}$:

Orbits	Represented by the points
$((1,0)(0,1))$	$((1,0)(0,1)),((1,0)(1,1))$ and $((1,0)(2,1))$

 ISBN: 978-84-617-8694-7

The non-relevant elements of the contraction matrix ε that might be different from zero are $\varepsilon_{(1,0)(0,1)}$, $\varepsilon_{(1,0)(1,1)}$ and $\varepsilon_{(1,0)(2,1)}$. However, a similar reasoning as in the previous case shows that the element $\varepsilon_{(1,0)(2,1)}$ is singular. So, the explicit form of the contraction matrix ε with respect to the chosen order O is

$$\begin{pmatrix} 0 & \varepsilon_{(1,0)(0,1)} & \varepsilon_{(1,0)(1,1)} & 0 \\ \varepsilon_{(1,0)(0,1)} & 0 & 0 & 0 \\ \varepsilon_{(1,0)(1,1)} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

where the non-null elements of this matrix have to satisfy Equation (2).

Moreover, from Equation (1) is deduced that the elements $\varepsilon_{(1,0)(0,1)}$ and $\varepsilon_{(1,0)(1,1)}$ can take any complex value.

By reasoning as we did in the previous dimension, we deduce that the set $N(\mathfrak{f}_4)$ of all normalized contraction matrices of $\mathfrak{f}_4$ has 4 elements, which are the following matrices

$$\begin{pmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \quad \begin{pmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \quad \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \text{ and the } 3 \times 3 \text{ null matrix.}$$

5 Graded contractions of the n-dimensional filiform Lie algebra

Let $\mathfrak{f}_n : [e_1, e_k] = e_{k+1}$, for $2 \leq k \leq n-1$ be the model filiform Lie algebra of dimension n. A grading of $\mathfrak{f}_n$ is given by

$$\Gamma : \mathfrak{f}_n = \mathfrak{f}^n_{(1,0)} \oplus \mathfrak{f}^n_{(0,1)} \oplus \mathfrak{f}^n_{(1,1)} \oplus \mathfrak{f}^n_{(2,1)} \oplus \mathfrak{f}^n_{(3,1)} \oplus \mathfrak{f}^n_{(4,1)} \oplus \ldots \oplus \mathfrak{f}^n_{(n-2,1)}.$$

The universal group of $\mathfrak{f}_n$ is $\mathbb{Z}_{n-1} \otimes \mathbb{Z}_2$ and $\mathfrak{f}^n_{(1,0)} = \langle e_1 \rangle$, $\mathfrak{f}^n_{(0,1)} = \langle e_2 \rangle$, $\mathfrak{f}^n_{(1,1)} = \langle e_3 \rangle$, $\mathfrak{f}^n_{(2,1)} = \langle e_4 \rangle$, $\mathfrak{f}^n_{(3,1)} = \langle e_5 \rangle$, $\mathfrak{f}^n_{(4,1)} = \langle e_6 \rangle, \ldots, \mathfrak{f}^n_{(n-2,1)} = \langle e_n \rangle$.

Let us consider

$$\Pi_n = \begin{pmatrix} 1 & 0 \\ a & 1 \end{pmatrix}, \text{ with } a \in \mathbb{Z}_{n-1} \text{ and } \mathrm{I} = \{(1,0), (0,1), (1,1), (2,1), (3,1), \ldots, (n-2,1\}.$$

We now consider the following order O on $I : 1 \to (1,0)$, $2 \to (0,1)$, $3 \to (1,1)$, $4 \to (2,1)$, $\ldots$, and $n \to (n-2,1)$ and starting from this point and by using any symbolic computation package for computations we proceed in the same way as the indicated in the previous particular cases. Indeed, we obtain the orbits of the points of I, I^2_u and I^3_u and consider the elements of the $n \times n$ contraction matrix ε that might be different from zero.

 ISBN: 978-84-617-8694-7

These elements are $\varepsilon_{(1,0)(0,1)}$, $\varepsilon_{(1,0)(1,1)}$, $\varepsilon_{(1,0)(2,1)}$, $\varepsilon_{(1,0)(3,1)}, \ldots$ and $\varepsilon_{(1,0)(n-2\ 1)}$. The resting elements are null. Moreover, we find that $\varepsilon_{(1,0)(n-2\ 1)} = 0$. It allows us to obtain the explicit form of the contraction matrix ε with respect to the chosen order O.

Let us recall that for any n-dimensional filiform Lie algebra, the elements, the non-null ones have to verify the following conditions (Equation (2))

$$\varepsilon_{(i,j)(k+m,l+n)}\varepsilon_{(k,l)(m,n)}[X_{i,j},[X_{k,l},X_{m,n}]] + \varepsilon_{(k,l)(m+i,n+j)}\varepsilon_{(m,n)(i,j)}[X_{k,l},[X_{m,n},X_{i,j}]]$$
$$+\varepsilon_{(m,n)(i+k,j+l)}\varepsilon_{(i,j)(k,l)}[X_{m,n},[X_{i,j},X_{k,l}]] = 0, \forall((i,j)(k,l)(m,n)) \in I_u^3.$$

Moreover, $[X_{i,j},[X_{k,l},X_{m,n}]]$, $[X_{m,n},[X_{i,j},X_{k,l}]]$ and $[X_{k,l},[X_{m,n},X_{i,j}]]$ are null for $\mathfrak{f}_n$. This implies that the elements $\varepsilon_{(1,0)(0,1)}$, $\varepsilon_{(1,0)(0,1)}$, $\varepsilon_{(1,0)(1,1)}$, $\varepsilon_{(1,0)(2,1)}$, $\ldots$, $\varepsilon_{(1,0)(n-3\,1)}$ can take any complex value, which allow us to obtain the conclusions in each dimension.

In any case, $\varepsilon \diamond \tau$ is the contraction matrix in which all the non-null elements are 1 and we find that the set $N(\mathfrak{g}^n)$ of all normalized contraction matrices of $\mathfrak{f}^n$ has 2^{n-2} elements.

Now, we are going to study, as a particular example, the case of a non-model filiform Lie algebra.

6 Graded contractions of the 6-dimensional filiform Lie algebra Q_6

Let $\mathcal{Q}_6$, defined by the law $[e_1,e_k] = e_{k+1}$, for $2 \leq k \leq 5$, $[e_2,e_5] = -e_6$ and $[e_3,e_4] = e_6$ be the filiform Lie algebra of dimension 6. A grading of $\mathcal{Q}_6$ is given by

$$\Gamma : \mathcal{Q}_6 = Q^6_{(1,0)} \oplus Q^6_{(0,1)} \oplus Q^6_{(1,1)} \oplus Q^6_{(2,1)} \oplus Q^6_{(3,1)} \oplus Q^6_{(3,2)}$$

By proceeding in a similar way as in previous sections, we have that the universal group of $\mathcal{Q}_6$ is $\mathbb{Z}_4 \otimes \mathbb{Z}_3$. Besides, $\mathcal{Q}^6_{(1,0)} = \langle e_1 + e_2 \rangle$, $Q_{(0,1)} = \langle e_2 \rangle$, $\mathcal{Q}^6_{(1,1)} = \langle e_3 \rangle$, $\mathcal{Q}^6_{(2,1)} = \langle e_4 \rangle$, $\mathcal{Q}^6_{(3,1)} = \langle e_5 \rangle$ and $\mathcal{Q}^6_{(3,2)} = \langle e_6 \rangle$.

Let now consider

$$H_{\Pi_6} = \begin{pmatrix} 1 & 0 \\ 2\,a & 1 \end{pmatrix}, \text{ with } a \in \mathbb{Z}_4 \text{ and } \mathrm{I} = \{(1,0),(0,1),(1,1),(2,1),(3,1),(3,2)\}.$$

The considered order O on I is $\{1 \to (1,0); 2 \to (0,1); 3 \to (1,1); 4 \to (2,1); 5 \to (3,1); 6 \to (3,2)\}$. The tables now obtained for the elements of I and the orbits of the points of I_u^2 and I_u^3 are the following

Represented by the points	Orbits
$(1,0)$	$(1,0)$
$(0,1)$	$(0,1),(2,1)$
$(1,1)$	$(1,1),(3,1)$
$(3,2)$	$(3,2)$

 ISBN: 978-84-617-8694-7

Orbits of the 15 points of I_u^2

Orbit	Represented by the points
$((1,0)(0,1))$	$((1,0)(0,1)),((1,0)(2,1))$
$((1,0)(1,1))$	$((1,0)(1,1)),((1,0)(3,1))$
$((1,0)(3,2))$	$((1,0)(3,2))$
$((0,1)(1,1))$	$((0,1)(1,1)),((2,1)(3,1))$
$((0,1)(2,1))$	$((0,1)(2,1))$
$((0,1)(3,1))$	$((0,1)(3,1)),((2,1)(1,1))$
$((0,1)(3,2))$	$((0,1)(3,2)),((2,1)(3,1))$
$((1,1)(3,2))$	$((1,1)(3,2)),((3,1)(3,2))$
$((1,1)(3,1))$	$((1,1)(3,1))$

Orbits of the 20 points of I_u^3

Orbit	Represented by the points
$((1,0)(0,1)(1,1))$	$((1,0)(0,1)(1,1)),((1,0)(2,1)(3,1))$
$((1,0)(0,1)(2,1))$	$((1,0)(0,1)(2,1))$
$((1,0)(0,1)(3,1))$	$((1,0)(0,1)(3,1)),((1,0)(2,1)(1,1))$
$((1,0)(1,1)(3,1))$	$((1,0)(1,1)(3,1))$
$((1,0)(1,1)(3,2))$	$((1,0)(1,1)(3,2)),((1,0)(3,1)(3,2))$
$((1,0)(0,1)(3,2))$	$((1,0)(0,1)(3,2)),((1,0)(2,1)(3,2))$
$((0,1)(1,1)(2,1))$	$((0,1)(1,1)(2,1)),((2,1)(3,1)(0,1))$
$((0,1)(1,1)(3,1))$	$((0,1)(1,1)(3,1)),((2,1)(3,1)(1,1))$
$((0,1)(1,1)(3,2))$	$((0,1)(1,1)(3,2)),((2,1)(3,1)(3,2))$
$((0,1)(3,1)(3,2))$	$((0,1)(3,1)(3,2)),((2,1)(1,1)(3,2))$
$((1,1)(3,1)(3,2))$	$((1,1)(3,1)(3,2))$
$((0,1)(2,1)(3,2))$	$((0,1)(2,1)(3,2))$

Orbits of the 6 points of $\mathcal{I}$:

Orbit	Represented by the points
$((1,0)(0,1))$	$((1,0)(0,1)),((1,0)(1,1)),((1,0)(2,1)),((1,0)(3,1))$ and $((2,1)(1,1))$

The explicit form of the contraction matrix ε with respect to chosen order O is

$$\begin{pmatrix} 0 & \varepsilon_{(1,0)(0,1)} & \varepsilon_{(1,0)(1,1)} & \varepsilon_{(1,0)(2,1)} & 0 & 0 \\ \varepsilon_{(1,0)(0,1)} & 0 & 0 & 0 & 0 & 0 \\ \varepsilon_{(1,0)(1,1)} & 0 & 0 & \varepsilon_{(1,1)(2,1)} & 0 & 0 \\ \varepsilon_{(1,0)(2,1)} & 0 & \varepsilon_{(1,1)(2,1)} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

The conditions for the non-null elements of the contraction matrix are

$$\varepsilon_{(i,j)(k+m,l+n)}\varepsilon_{(k,l)(m,n)}[X_{ij},[X_{kl},X_{mn}]]+\varepsilon_{(k,l)(m+i,n+j)}\varepsilon_{(m,n)(i,j)}[X_{kl},[X_{mn},X_{ij}]]$$

$$+\varepsilon_{(m,n)(i+k,j+l)}\varepsilon_{(i,j)(k,l)}[X_{mn},[X_{ij},X_{kl}]]=0,\forall((i,j)(k,l)(m,n))\in I_u^3.$$

In this case, the following restriction $\varepsilon_{(1,0)(0,1)}\varepsilon_{(1,1)(2,1)}=0$ is obtained.

Moreover, the elements of any contraction matrix verify the following conditions

- If $\varepsilon_{(1,0)(0,1)}\neq 0$, then $\varepsilon_{(1,1)(2,1)}=0$ and the parameters $\varepsilon_{(1,0)(1,1)}$ and $\varepsilon_{(1,0)(2,1)}$ could be null. So, there are 2^2 different types of contraction matrices with $\varepsilon_{(1,0)(0,1)}\neq 0$.

 ISBN: 978-84-617-8694-7

- If $\varepsilon_{(1,0)(0,1)} = 0$, then the parameters $\varepsilon_{(1,0)(1,1)}$, $\varepsilon_{(1,0)(2,1)}$ y $\varepsilon_{(1,1)(2,1)}$ could be null. So, there are 2^3 different types of contraction matrices with $\varepsilon_{(1,0)(0,1)} = 0$.

If $\varepsilon = (\varepsilon_{ij})$ is a contraction matrix, then $\tau = (\tau_{ij})$, such that $\tau_{ij} = \frac{1}{\varepsilon_{ij}}$ if $\varepsilon_{ij} \neq 0$ or $\tau_{ij} = 0$, if $\varepsilon_{ij} = 0$ is also a contraction matrix. Moreover, $\varepsilon \diamond \tau$ is the contraction matrix in which all the non-null elements are 1. That matrix is the normalized contraction matrix of ε. If we denote by $N(\mathcal{Q}_6)$ to the set of all normalized contraction matrices of $\mathcal{Q}_6$, we find that this set has 12 elements.

7 Certain conclusions

In our research we have obtained the graded contractions of some lower-dimensional filiform Lie algebra, concretely, of the model filiform Lie algebras of dimensions $3, 4, 5$ and 6, although in this paper only the two first cases are shown, due to reasons of length. Then, as a consequence of the results obtained, we have dealt with the general case n-dimensional for this type of algebras. Moreover, we have repeated this study for the dimension 6 with a non-model filiform Lie algebra.

Two have been the reasons for the authors to deal with this subject. The first of them is to complete previous papers by different authors, like Inönü and Wigner [3] in 1953, Weimar-Woods [7] in 2006 or Bahturin, Goze and Remm [1] in 2013, for instance.

The second one is the possibility of setting new theoretical results on it. Indeed, as consequences of this study, we find some question which could make us think of giving an answer to the following facts, thereby determining some conjectures. For instance: a) will have the orbits of I_u^2, for the filiform Lie algebras $\mathfrak{f}_5$ and $\mathfrak{f}_6$, the same representatives? Have these orbits 4 points at most? b) have the contraction matrices of the filiform Lie algebras $\mathfrak{f}_n$ and Q_n an unique null relevant contraction parameter? and c) which is the form of the symmetry groups of filiform Lie algebras $\mathfrak{f}_n$ and Q_n?

We hope to give responses to these questions in future work.

Acknowledgements

This work has been partially supported by MTM2016-75024-P and FEDER.

References

[1] Y. Bahturin, M. Goze, E. Remm, Group gradings on filiform Lie algebras, arXiv: 1308.2396v1 (2013).

[2] J.M. Escobar, J. Núñez, P. Pérez-Fernández, *On contractions of Lie algebras*, Mathematics in Computer Science. In press (2017). DOI: 10.1007/s11786-016-0266-0.

[3] E. Inönü and E. Wigner, *On the contraction of groups and their representations*, Proc. Nat. Acad. Sci. U.S.A. **39** (1953), 510-524.

[4] M. de Montigny, J. Patera, *Discrete and continuous graded contractions of Lie algebras and superalgebras*, J. Phys. A: Math. Gen. **24** (1991), 525-547.

[5] Petr Novotný, Jirí Hrivnák, *On (α, β, γ)-derivations of Lie algebras and corresponding invariant functions*, Journal of Geometry and Physics **58**:2 (2008), 208-217

[6] M. Vergne, *Cohomologie des algèbres de Lie nilpotentes, Application à l'étude de la variété des algebres de Lie nilpotentes*. Bull. Soc. Math. France **98** (1970), 81-116.

[7] E. Weimar-Woods, *The general structure of G-graded contractions of Lie algebras I: The classification*, Canadian Journal of Mathematics **36**:6 (2006), 1291-1340

[8] E. Weimar-Woods, *The general structure of G-Graded contractions of Lie Algebras, II: The contracted Lie Algebra*, Rev. Math. Phys. **18**:06 (2006), 655-711.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

Auxiliary Point on the Semilocal Convergence of Newton's Method

J. A. Ezquerro[1] and M. A. Hernández-Verón[1]

[1] *Department of Mathematics and Computation, University of La Rioja*

emails: jezquer@unirioja.es, mahernan@unirioja.es

Abstract

We use an auxiliary point in the analysis of the semilocal convergence of Newton's method under center conditions on high order derivatives of the operator involved and use the majorant principle of Kantorovich to do it.

Key words: Newton's method, semilocal convergence, majorant principle
MSC 2000: 47H99, 65H10, 65J15

1 Introduction

The application of Newton's method to solve nonlinear equations has a long history and different types of conditions has been used over the past years to analyse the semilocal convergence of the method. Remember that semilocal convergence results for iterative methods, in general, and, for Newton's method, in particular, require conditions on the starting points and conditions on the operator involved. Different types of conditions on the operator can be required as we can see in the mathematical literature, but, if we pay attention to center conditions on the operator, we see that the starting points are used to center the conditions on the first or second derivative of the operator. In this work, we propose to require conditions on higher order derivatives of the operator and on an auxiliary point different from the starting point of Newton's method.

For the last, we consider a nonlinear operator $F : \Omega \subseteq X \to Y$ defined on a nonempty open convex domain Ω of a Banach space X with values in a Banach space Y in order to give sufficient generality to our study, so that problems from computational sciences, physics and other disciplines can be brought into the equation $F(x) = 0$ by using mathematical modelling, so that the unknowns of this equation can be functions (difference, differential, and integral equations), vectors (systems of linear or nonlinear algebraic equations), or real/complex numbers (single algebraic equations with single unknowns).

 ISBN: 978-84-617-8694-7

2 Preliminaries

The algorithm of Newton's method for solving equation $F(x) = 0$ is:

$$x_0 \in \Omega, \quad x_{n+1} = x_n - [F'(x_n)]^{-1} F(x_n), \quad n \geq 0,$$

and the first semilocal convergence result given for Newton's method in Banach spaces is due to the Russian mathematician L. V. Kantorovich, who proves, at the beginning of the 50s of the last century, the semilocal convergence of Newton's method using "the majorant principle" and under the following conditions [3]:

(K1) There exists $\Gamma_0 = [F'(x_0)]^{-1} \in \mathcal{L}(Y, X)$, for some $x_0 \in \Omega$, with $\|\Gamma_0\| \leq \beta$ and $\|\Gamma_0 F(x_0)\| \leq \eta$, where $\mathcal{L}(Y, X)$ is the set of bounded linear operators from Y to X,

(K2) $\|F''(x)\| \leq M$ for $x \in \Omega$,

(K3) $M\beta\eta \leq \frac{1}{2}$.

From that moment, a plethora of studies on the weakness and/or extension of conditions (K1)–(K3) have been made.

Focusing on center conditions on F, we underline the papers of Gutiérrez [1] and Gutiérrez and Hernández [2], where semilocal convergence results under conditions of type

$$\|F''(x) - F''(x_0)\| \leq L_0 \|x - x_0\|, \quad L_0 \geq 0, \quad x \in \Omega,$$

or

$$\|F'(x) - F'(x_0)\| \leq K_0 \|x - x_0\|, \quad K_0 \geq 0, \quad x \in \Omega,$$

are respectively given. In this work, we propose to study the semilocal convergence of Newton's method under conditions of the last two types, but with two significant variants: derivatives of higher order than two and center conditions on an auxiliary point $\widetilde{x}$ instead of on the starting point x_0; in particular, under a condition of type

$$\|F^{(k)}(x) - F^{(k)}(\widetilde{x})\| \leq \omega(\|x - \widetilde{x}\|), \quad x \in \Omega,$$

where $k \geq 3$ and $\omega : [0, +\infty) \to \mathbb{R}$ is a nondecreasing continuous function such that $\omega(0) = 0$. This variant of the previous known conditions leads us to modify the domain of starting points for Newton's method, so that we can guarantee the semilocal convergence of the method from starting at points where other results fail.

3 The majorant principle

To prove the semilocal convergence of Newton's method, we consider the conditions

(C1) There exists the operator $\Gamma_0 = [F'(x_0)]^{-1} \in \mathcal{L}(Y, X)$ with $\|\Gamma_0\| \leq \beta$ and $\|\Gamma_0 F(x_0)\| \leq \eta$; moreover, $\|F^{(i)}(x_0)\| \leq b_i$ with $i = 2, 3, \ldots, k-1$ and $k \geq 3$,

(C2) There exists $\widetilde{x} \in \Omega$ such that $\|x_0 - \widetilde{x}\| = \gamma$, where $x_0 \in \Omega$, and $\|F^{(k)}(\widetilde{x})\| \leq \delta$,

(C3) There exists a nondecreasing continuous function $\omega : [0, +\infty) \longrightarrow \mathbb{R}$ such that $\|F^{(k)}(x) - F^{(k)}(\widetilde{x})\| \leq \omega(\|x - \widetilde{x}\|)$ for $x \in \Omega$ and $\omega(0) = 0$,

and use "the majorant principle" of Kantorovich, that consists of looking a scalar sequence $\{t_n\}$,

$$t_0 = 0, \quad t_{n+1} = t_n - \frac{f(t_n)}{f'(t_n)}, \quad n \geq 0,$$

that *majorizes* Newton's sequence $\{x_n\}$ in the Banach space X. As a consequence of this fact, the convergence of $\{x_n\}$ follows immediately from the convergence of $\{t_n\}$, see [3].

So, we then look for a real function $f \in \mathcal{C}^j([\widetilde{t}, +\infty))$, with $\widetilde{t} \in \mathbb{R}_+$ and $j \geq k$, such that

$$\|F^{(k)}(x) - F^{(k)}(\widetilde{x})\| \leq f^{(k)}(t) - f^{(k)}(\widetilde{t}) \text{ with } \|x - \widetilde{x}\| \leq t - \widetilde{t}, x \in \Omega \text{ and } t \in [\widetilde{t}, +\infty).$$

So, from (C3) and the last, it follows

$$\|F^{(k)}(x) - F^{(k)}(\widetilde{x})\| \leq \omega(\|x - \widetilde{x}\|) \leq \omega(t - \widetilde{t}) = f^{(k)}(t) - f^{(k)}(\widetilde{t})$$

if $\|x - \widetilde{x}\| \leq t - \widetilde{t}$, since ω is a nondecreasing continuous function, and then

$$f^{(k)}(t) = f^{(k)}(\widetilde{t}) + \omega(t - \widetilde{t}).$$

In addition, if we consider $\widetilde{t} = -\gamma$, $f^{(k)}(\widetilde{t}) = \delta$ and take into account (C1)-(C2), we can solve the initial value problem

$$\begin{cases} y^{(k)}(t) = \delta + \omega(t + \gamma), \\ y(0) = \dfrac{\eta}{\beta}, \quad y'(t_0) = -\dfrac{1}{\beta}, \\ y''(0) = b_2, \quad y'''(0) = b_3, \quad \ldots, \quad y^{(k-1)}(0) = b_{k-1}, \end{cases}$$

to find $f(t)$, since we can choose, from (C1), $-\frac{1}{f'(0)} = \beta$, $-\frac{f(0)}{f'(0)} = \frac{\eta}{\beta}$ and $f^{(i)}(0) = b_i$, for $i = 2, 3, \ldots, k-1$. So, next result is given.

Theorem 1. *Suppose that the function $\omega(t + \gamma)$ is continuous in $[0, +\infty)$. Then, for any nonnegative real numbers γ, δ, $\beta \neq 0$, η, $b_2, b_3, \ldots, b_{k-1}$, the last initial value problem has a unique solution $f(t) \in \mathcal{C}^j([-\gamma, +\infty))$, with $j \geq k \geq 3$, which is given by*

$$f(t) = \int_0^t \int_0^{\theta_{k-1}} \cdots \int_0^{\theta_1} \omega(s + \gamma)\, ds\, d\theta_1 \cdots d\theta_{k-1} + \frac{\delta}{k!} t^k + \sum_{i=2}^{k-1} \frac{b_i}{i!} t^i - \frac{t}{\beta} + \frac{\eta}{\beta}.$$

 ISBN: 978-84-617-8694-7

4 Semilocal convergence

In the following result, we prove that the sequence $\{t_n\}$, defined from the function $f(t)$ given in Theorem 1, majorizes sequence $\{x_n\}$ in the Banach space X.

Theorem 2. *Let X and Y be two Banach spaces and $F : \Omega \subseteq X \longrightarrow Y$ a nonlinear q ($q \geq 2$) times continuously differentiable operator on a nonempty open convex domain Ω and $f(t)$ be function defined in Theorem 1. Suppose that conditions (C1)–(C3) are satisfied, there exists a root $\alpha > 0$ of $f'(t) = 0$ such that $f(\alpha) \leq 0$, and $B(x_0, t^*) \subset \Omega$, where t^* is the smallest positive root of $f(t) = 0$. Then, Newton's sequence $\{x_n\}$ satisfies:*

$$\|x_n - x_{n-1}\| \leq t_{n+1} - t_n, \quad \text{for all} \quad n \geq 0.$$

Once we have proved that $\{t_n\}$ majorizes $\{x_n\}$, we can prove the semilocal convergence of $\{x_n\}$ in the Banach space X.

Theorem 3. *Let X and Y be two Banach spaces and $F : \Omega \subseteq X \longrightarrow Y$ a nonlinear q ($q \geq 2$) times continuously differentiable operator on a nonempty open convex domain Ω and $f(t)$ be function defined in Theorem 1. Suppose that conditions (C1)–(C3) are satisfied, there exists a root $\alpha > 0$ of $f'(t) = 0$ such that $f(\alpha) \leq 0$, and $B(x_0, t^*) \subset \Omega$, where t^* is the smallest positive root of $f(t) = 0$. Then, Newton's sequence $\{x_n\}$ converges to a solution x^* of $F(x) = 0$ starting at x_0. Moreover, $x_n, x^* \in \overline{B(x_0, t^*)}$ and*

$$\|x^* - x_n\| \leq t^* - t_n, \quad \text{for all} \quad n \geq 0.$$

where $t_n = t_{n-1} - \frac{f(t_{n-1})}{f'(t_{n-1})}$, with $n \in \mathbb{N}$ and $t_0 = 0$.

Acknowledgements

This research was partially supported by Ministerio de Economía y Competitividad under grant MTM2014-52016-C2-1-P.

References

[1] J. M. Gutiérrez, *A new semilocal convergence theorem for Newton's method*, J. Comput. Appl. Math. **79** (1997) 131–145.

[2] J. M. Gutiérrez and M. A. Hernández, *Newton's method under weak Kantorovich conditions*, IMA J. Numer. Anal. **20** (2000) 521–532.

[3] L. V. Kantorovich and G. P. Akilov, *Functional analysis*, Pergamon Press, Oxford, 1982.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

Computing the sets of totally symmetric and totally conjugate orthogonal partial Latin squares by means of a SAT solver

Raúl M. Falcón[1], Óscar J. Falcón[2] and Juan Núñez[2]

[1] *Department of Applied Mathematics I, University of Seville (Spain)*

[2] *Department of Geometry and Topology, University of Seville (Spain)*

emails: rafalgan@us.es, oscfalgan@yahoo.es, jnvaldes@us.es

Abstract

Conjugacy and orthogonality of Latin squares have been widely studied in the literature not only for their theoretical interest in combinatorics, but also for their applications in distinct fields as experimental design, cryptography or code theory, amongst others. This paper deals with a series of binary constraints that characterize the sets of partial Latin squares of a given order for which their six conjugates either coincide or are all of them distinct and pairwise orthogonal. These constraints enable us to make use of a SAT solver to enumerate both sets. As an illustrative application, it is also exposed a method to construct totally symmetric partial Latin squares that gives rise, under certain conditions, to new families of Lie partial quasigroup rings.

Key words: Partial Latin square, conjugacy, orthogonality.
MSC 2000: 05B15, 20N05.

1 Introduction

A *quasigroup* [22] is a pair $(S, \cdot)$ formed by a nonempty set S endowed with a product $\cdot$ such that, if any two of the three symbols a, b and c in the equation $a \cdot b = c$ are given as elements of S, then the third one is uniquely determined. The size of S is the *order* of the quasigroup. The multiplication table of a quasigroup of order n constitutes a *Latin square* of the same order, that is, an $n \times n$ array in which each cell contains one symbol chosen from the set S, such that each symbol occurs exactly once in each row and in each column. The number of Latin squares is known [24, 26, 29, 30] for order up to 11.

Bruck [12] introduced the concept of *totally symmetric quasigroup* as a quasigroup $(S, \cdot)$ for which the equation $a \cdot b = c$ remains valid under every permutation of the three symbols $a, b, c \in S$. There exist six such permutations and each one of them gives rise to a new quasigroup, which is said to be *conjugate* to $(S, \cdot)$. Hence, a quasigroup is totally symmetric if its six conjugates coincide. If besides, the quasigroup is *idempotent*, that is, if $a \cdot a = a$, for all $a \in S$, then this notion is equivalent to that of a *Steiner triple system*. The distribution of totally symmetric quasigroups and Steiner triple systems into isomorphism classes is known [1, 25] for orders up to 10 and 19, respectively.

Two quasigroups of order n are said to be *orthogonal* if the juxtaposition of their corresponding multiplication tables gives rise to an $n \times n$ array containing n^2 distinct ordered pairs. Stein [34] posed the problem of constructing a quasigroup or Latin square that is orthogonal to one of its conjugates. It is known in this regard [6, 7, 11, 31] the existence of quasigroups that are orthogonal to the conjugate under consideration, distinct of themselves, for any order $n \notin \{2, 3, 6\}$. Much more recently, Bennett and Zhang [10] dealt with Latin squares for which each one of their conjugates is orthogonal to its transpose. They proved the existence of such Latin squares for all prime powers $n \notin \{2, 3, 5\}$. Further, Lindner et al. [28] focused on idempotent Latin squares for which their six conjugates are distinct and pairwise orthogonal. They proved in particular the existence of such Latin squares for every order being a prime power $n \geq 8$ and also for all sufficiently large orders n. Bennett [4] established $n > 5594$ as an upper bound for this last condition except possibly $n = 6810$, and enumerate a series of smaller orders for which these Latin squares also exist. Four years later, he improved [5] the previous upper bound to $n > 5074$. Much more recently, Belyavskaya and Popovich [3] introduced the equivalent notion of *totally conjugate orthogonal quasigroup* as a quasigroup for which its six conjugates are distinct and pairwise orthogonal. They proved the existence of such quasigroups for any order $n \geq 11$ that is relatively prime to 2, 3, 5, and 7. Their motivation to study this kind of quasigroups was mainly based on their application in error detecting codes [2].

The concept of quasigroup is straightforwardly generalized to that of *partial quasigroup* of order n, for which (a) the law $\cdot$ is a partial binary operation on a finite set S of n elements, and (b) if the equations $a \cdot x = b$ and $y \cdot a = b$, with $a, b \in S$, have solutions for x and y in S, then these solutions are unique. The multiplication table of a partial quasigroup of order n constitutes a *partial Latin square* of the same order, that is, an $n \times n$ array in which each cell is either empty or contains one element chosen from S, such that each symbol occurs at most once in each row and in each column. The number of partial Latin squares is known [17, 18, 19, 20] for order up to seven.

Since Evans [15] introduced the problem of embedding a partial quasigroup of order n into a quasigroup of order $2n$, a wide amount of authors have dealt with the embedding of distinct types of partial quasigroups; particularly, that of a partial totally symmetric quasigroup into a totally symmetric quasigroup [13, 27, 32, 33]. Further, the orthogonality

 ISBN: 978-84-617-8694-7

among conjugates of a partial Latin square was indirectly contemplated [8, 9, 23] by focusing on the existence of incomplete Latin squares that are orthogonal to one of their conjugates and have an empty subsquare that can be filled by means of a Latin square that is orthogonal in turn to its corresponding conjugate. A more general case was recently proposed by the first author [18], who makes use of computational algebraic geometry to enumerate the set of self-orthogonal partial Latin squares of order $n \leq 4$. This paper delves into this topic by dealing with the sets of partial Latin squares of a given order for which their six conjugates either coincide or are all of them distinct and pairwise orthogonal, respectively. In order to improve the computational efficiency, it is proposed to focus on techniques to solve Boolean satisfiability problems instead of those on algebraic geometry.

As an illustrative application of the exposed study, we also delve into a recent work developed by the authors [16] about the enumeration of partial quasigroup rings over finite fields derived from partial Latin squares. Bruck [12] introduced the concept of *quasigroup ring* related to a quasigroup $(S, \cdot)$ as an algebra of basis $\{e_a \mid a \in S\}$ over a base field $\mathbb{K}$ such that $e_a e_b = e_{a \cdot b}$, for all $a, b \in S$. This concept is straightforwardly generalized to that of *partial quasigroup ring* in case of being the pair $(S, \cdot)$ a partial quasigroup. In this paper, we describe a totally symmetric partial Latin square of order $3n$, derived from a given partial Latin square of order n, that enables us to introduce in turn a Lie partial quasigroup ring over a finite field of characteristic two.

The paper is organized as follows. In Section 2, we expose some preliminary concepts and results on partial Latin squares that are used throughout our study. In Section 3, we introduce a pair of series of binary constraints that characterize, respectively, the sets of totally symmetric and totally conjugate orthogonal partial Latin squares of given order and weight. Finally, Section 5 deals with an illustrative example that enables us to construct a family of Lie partial quasigroup rings from a totally symmetric partial Latin square satisfying certain conditions.

2 Preliminaries

This section deals with some basic concepts and notations on partial Latin squares that are used throughout the paper. We refer the reader to the monographs of Dénes and Keedwell [14] for more details about this topic.

Hereafter, the set of partial Latin squares of order n is denoted as PLS_n, whereas the set of symbols of any such a partial Latin square $P = (p_{ij}) \in \mathrm{PLS}_n$ is assumed to be the set $[n] = \{1, \ldots, n\}$. An *entry* of P is any triple $(i, j, p_{i,j}) \in [n] \times [n] \times [n]$. The partial Latin square P is uniquely determined by the set of all its entries, which is called its *entry set* and denoted as $E(P)$. The size of this set coincides, therefore, with the number of non-empty cells of P, which constitutes its *weight*. From here on, $\mathrm{PLS}_{n;m}$ denotes the set of partial Latin squares of order n having weight m. Thus, for instance, the partial

 ISBN: 978-84-617-8694-7

Latin square P in Figure 1 belongs to the set $\mathrm{PLS}_{3;4}$ and has as entry set the set $E(P) = \{(1,1,2),(1,2,1),(2,1,1),(3,3,3)\}$.

$$P \equiv \begin{array}{|c|c|c|}\hline 2 & 1 & \\ \hline 1 & & \\ \hline & & 3 \\ \hline \end{array}$$

Figure 1: Partial Latin square in $\mathrm{PLS}_{3;4}$.

Let S_3 denote the symmetric group of three elements. Let P be a partial Latin square in $\mathrm{PLS}_{n;m}$ and let π be a permutation in S_3. The π-*conjugate* of P is defined as the partial Latin square $P^\pi \in \mathrm{PLS}_{n;m}$ such that $E(P^\pi)\{(p_{\pi(1)}, p_{\pi(2)}, p_{\pi(3)}) \colon (p_1,p_2,p_3) \in E(P)\}$. There exist, therefore, six conjugates: $P^{\mathrm{Id}} = P$, $P^{(12)} = P^t$, $P^{(13)}$, $P^{(23)}$, $P^{(123)} = (P^{(23)})^t$ and $P^{(132)} = (P^{(13)})^t$, where the notation ^{t} denotes the transpose of the corresponding partial Latin square. In order to illustrate these conjugates, let us consider the partial Latin square $P \in \mathrm{PLS}_{3;4}$ in Figure 2. Particularly, $E(P) = \{(1,1,1),(1,2,2),(2,2,3),(3,3,1)\}$ and hence, once the corresponding permutations among the three components of each entry are done, we obtain the conjugates therein exposed. Observe that all of them are distinct. Figure 1 shows instead a partial Latin square in $\mathrm{PLS}_{3;4}$ for which all its conjugates coincide. In this case, the partial Latin square under consideration is said to be *totally symmetric*. Hereafter, the set of totally symmetric partial Latin squares of order n and its subset of partial Latin squares of weight m are respectively denoted as TSPLS_n and $\mathrm{TSPLS}_{n;m}$.

$$P \equiv \begin{array}{|c|c|c|}\hline 1 & 2 & \\ \hline & 3 & \\ \hline & & 1 \\ \hline \end{array} \qquad P^{(12)} \equiv \begin{array}{|c|c|c|}\hline 1 & & \\ \hline 2 & 3 & \\ \hline & & 1 \\ \hline \end{array} \qquad P^{(13)} \equiv \begin{array}{|c|c|c|}\hline 1 & & 3 \\ \hline & 1 & \\ \hline & 2 & \\ \hline \end{array}$$

$$P^{(23)} \equiv \begin{array}{|c|c|c|}\hline 1 & 2 & \\ \hline & & 2 \\ \hline 3 & & \\ \hline \end{array} \qquad P^{(123)} \equiv \begin{array}{|c|c|c|}\hline 1 & & 3 \\ \hline 2 & & \\ \hline & 2 & \\ \hline \end{array} \qquad P^{(132)} \equiv \begin{array}{|c|c|c|}\hline 1 & & \\ \hline & 1 & 2 \\ \hline 3 & & \\ \hline \end{array}$$

Figure 2: Partial Latin square in $\mathrm{PLS}_{3;4}$ and its conjugates.

Let $P = (p_{ij})$ and $Q = (q_{ij})$ be two partial Latin squares of order n. They are said to be *orthogonal* if all the ordered pairs on non-empty entries that are obtained when both arrays are superimposed are distinct. Equivalently, given $i, i', j, j' \in [n]$ such that $p_{ij} = p_{i'j'} \in [n]$, then q_{ij} and $q_{i'j'}$ are not the same symbol of $[n]$. Thus, for instance, the partial Latin squares P and $P^{(13)}$ in Figure 2 are orthogonal, but the partial Latin squares P and $P^{(12)}$ in the same figure are not. In this regard, given a permutation $\pi \in S_3 \setminus \{\mathrm{Id}\}$, a partial Latin square $P \in \mathrm{PLS}_n$ is said to be π-*orthogonal* if it is orthogonal to its π-conjugate.

 ISBN: 978-84-617-8694-7

Particularly, if $\pi = (12)$, then P is said to be *self-orthogonal*. Thus, for instance, the partial Latin square $P^{(23)}$ in Figure 2 is self-orthogonal. Finally, we say that a partial Latin square is *totally conjugate orthogonal* if its six conjugates are distinct and pairwise orthogonal. This is the case, for instance, of the partial Latin square shown in Figure 3. From here on, the set of totally conjugate orthogonal partial Latin squares of order n and its subset of partial Latin squares of weight m are respectively denoted as TCOPLS_n and $\mathrm{TCOPLS}_{n;m}$.

$$P \equiv \begin{array}{|c|c|c|}\hline & & 3 \\ \hline & & 2 \\ \hline 1 & 3 & \\ \hline \end{array} \quad P^{(12)} \equiv \begin{array}{|c|c|c|}\hline & & 1 \\ \hline & & 3 \\ \hline 3 & 2 & \\ \hline \end{array} \quad P^{(13)} \equiv \begin{array}{|c|c|c|}\hline 3 & & \\ \hline & & 2 \\ \hline & 3 & 1 \\ \hline \end{array}$$

$$P^{(23)} \equiv \begin{array}{|c|c|c|}\hline & & 3 \\ \hline & 3 & \\ \hline 1 & & 2 \\ \hline \end{array} \quad P^{(123)} \equiv \begin{array}{|c|c|c|}\hline & & 1 \\ \hline & 3 & \\ \hline 3 & & 2 \\ \hline \end{array} \quad P^{(132)} \equiv \begin{array}{|c|c|c|}\hline 3 & & \\ \hline & & 3 \\ \hline & 2 & 1 \\ \hline \end{array}$$

Figure 3: Totally conjugate orthogonal partial Latin square in $\mathrm{PLS}_{3;4}$.

The set PLS_n is identified [18] with the set of zeros of the following system of equations in the set of n^3 variables $\{X\} = \{x_{ijk} \mid i, j, k \in [n]\}$.

$$\begin{cases} x_{ijk}x_{i'jk} = 0, \text{ for all } i, i', j, k \leq n \text{ such that } i \neq i', \\ x_{ijk}x_{ij'k} = 0, \text{ for all } i, j, j', k \leq n \text{ such that } j \neq j', \\ x_{ijk}x_{ijk'} = 0, \text{ for all } i, j, k, k' \leq n \text{ such that } k \neq k', \\ x_{ijk} \in \{0, 1\}, \text{ for all } i, j, k \leq n. \end{cases} \tag{1}$$

Specifically, every partial Latin square $P = (p_{ij}) \in \mathrm{PLS}_n$ is uniquely identified with a zero $(x_{111}, \ldots, x_{nnn})$, where $x_{ijk} = 1$ if $p_{ij} = k$ and 0, otherwise. Hereafter, in order to avoid degeneracy, partial Latin squares are assumed to have at least one entry in each row, at least one entry in each column, and at least one copy of each symbol. To get this condition, the following inequations are added to (1)

$$\begin{cases} \sum_{j,k\in[n]} x_{ijk} \geq 1, \text{ for all } i \in [n], \\ \sum_{i,k\in[n]} x_{ijk} \geq 1, \text{ for all } j \in [n], \\ \sum_{i,j\in[n]} x_{ijk} \geq 1, \text{ for all } k \in [n]. \end{cases} \tag{2}$$

Based on (1) and (2), we establish in Section 3 some equations to deal, respectively, with the sets TSPLS_n and TCOPLS_n. To this end, let us introduce the following notation

$$x^{\pi}_{i_1 i_2 i_3} := x_{i_{\pi(1)} i_{\pi(2)} i_{\pi(3)}},$$

for all $\pi \in S_3$ and $x_{i_1 i_2 i_3} \in \{X\}$. Besides, we label the six permutations in S_3 as

$$S_3 := \{\pi_1 = \mathrm{Id}, \pi_2 = (12), \pi_3 = (13), \pi_4 = (23), \pi_5 = (123), \pi_6 = (132)\}.$$

 ISBN: 978-84-617-8694-7

3 Binary constraints related to the sets TSPLS_n and TCOPLS_n

This section deals with a series of binary constraints that characterize the sets of totally symmetric and totally conjugate orthogonal partial Latin squares of given order and weight.

Lemma 3.1. *Let n and m be two positive integers such that $n \leq m \leq n^2$.*

a) If $m > n$, then every pair of orthogonal conjugates of a partial Latin square in the set $\mathrm{TCOPLS}_{n;m}$ are distinct.

b) If $|\mathrm{TCOPLS}_{n;m}| = 0$, then $|\mathrm{TCOPLS}_{n;m'}| = 0$, for all $m' \in \{m+1, \ldots, n^2\}$.

Proof. Let us prove each statement separately.

a) Let $P \in \mathrm{PLS}_{n;m}$ and $\pi, \pi' \in S_3$ be such that $\pi \neq \pi'$ and $P^\pi = P^{\pi'}$. Since $m > n$, there exists one symbol $k \in [n]$ and a distinct pair of elements (i_1, j_1) and, (i_2, j_2) in $[n] \times [n]$ such that $\{(i_1, j_1, k), (i_2, j_2, k)\} \subseteq E(P^\pi) \cap E(P^{\pi'})$. As a consequence, $P^\pi = P^{\pi'}$ is not orthogonal to itself.

b) Otherwise, the partial Latin square that results after emptying any $m' - m$ filled cells of the partial Latin square in $\mathrm{TCOPLS}_{n;m'}$ would be in $\mathrm{TCOPLS}_{n;m}$, which is a contradiction. □

Proposition 3.2. *Let n and m be two positive integers such that $n < m \leq n^2$. Then,*

a) The set TSPLS_n is identified with the set of zeros of (1)–(2) and

$$x_{ijk}^{\pi_s} = x_{ijk}, \text{ for all } i, j, k \in [n] \text{ and } s \in \{1, 2, 3\}. \tag{3}$$

b) The set $\mathrm{TSPLS}_{n;m}$ is identified with the set of zeros of (1)–(3) and

$$\sum_{i,j,k \in [n]} x_{ijk} \leq m. \tag{4}$$

c) The set TCOPLS_n is identified with the set of zeros of (1)–(2) and

$$x_{ijp}^{\pi_s} x_{klp}^{\pi_s} x_{ijq}^{\pi_t} x_{klq}^{\pi_t} = 0, \text{for all } i, j, k, l, p, q \leq n; s, t \leq 3; \text{ such that } (i, j) \neq (k, l), s \leq t. \tag{5}$$

d) The set $\mathrm{TCOPLS}_{n;m}$ is identified with the set of zeros of (1), (2), (4) and (5).

Proof. The result follows straightforwardly from the definitions exposed in Section 2 once each partial Latin square $P = (p_{ij}) \in \mathrm{PLS}_{r,s,n}$ is identified with a zero $(x_{111}, \ldots, x_{rsn})$ such that $x_{ijk} = 1$ if $p_{ij} = k$ and 0, otherwise. Thus, for instance, if we focus on the proof of statement (c), then, given $1 \leq s < t \leq 3$, the system of equations determined by (5) involves the π_s^{-1}- and π_t^{-1}-conjugates of P to be orthogonal. Besides, from Lemma 3.1.a, both conjugates are distinct. □

 ISBN: 978-84-617-8694-7

Proposition 3.2 has been implemented in the SAT solver MINION [21] to obtain the numerical data exposed in Table 1. Further, Table 2 indicates the run time that is required in a system with an *Intel Core i7-2600, with a 3.4 GHz processor and 16 GB of RAM* to determine one specific example in the sets $\mathrm{TSPLS}_{n;m}$ and $\mathrm{TCOPLS}_{n;m}$.

m	$\lvert\mathrm{TSPLS}(n;m)\rvert$				$\lvert\mathrm{TCOPLS}(n;m)\rvert$	
	n				n	
	3	4	5	6	3	4
3	1				36	
4	6	1			216	576
5	6	12	1		12	45168
6	10	24	20	1	0	315048
7	12	64	80	30	0	391824
8	3	60	220	210	0	95028
9	3	100	380	680	0	2616
10		148	910	1980		0
11		72	1010	4380		0
12		90	1630	7660		0
13		72	2740	17820		0
14		36	2040	23370		0
15		16	2784	37476		0
16		16	3395	68850		0
17			2195	68190		
18			2080	96660		
19			2320	145560		
20			900	122040		
21			900	146040		
22			480	196200		
23			240	132480		
24			30	148710		
25			30	157320		
26				101430		
27				81540		
28				86310		
29				35820		
30				33390		
31				20340		
32				11340		
33				4560		
34				3960		
35				720		
36				480		
Total	41	711	24385	1755547	264	850260

Table 1: Distribution of the sets $\mathrm{TSPLS}_{n;m}$ and $\mathrm{TCOPLS}_{n;m}$.

4 Lie partial quasigroup rings derived from the conjugate-extension of a partial Latin square

The inclusion of new binary constraints into (1)–(5) enables us to determine families of partial Latin squares in the sets TSPLS_n and TCOPLS_n with possible applications in distinct fields. As an illustrative example, we conclude this paper by describing in this section a new family of Lie partial quasigroup rings related to a totally symmetric partial Latin square of order $3n$, which is derived in turn from a given partial Latin square of order n. Recall that a *Lie algebra* is an anti-commutative algebra A that holds the so-called *Jacobi identity*

$$J(a,b,c) := (ab)c + (bc)a + (ca)b = 0, \text{ for all } a,b,c \in A. \tag{6}$$

 ISBN: 978-84-617-8694-7

n	m	Run time (seconds) $\mathrm{TSPLS}_{n;m}$	Run time (seconds) $\mathrm{TCOPLS}_{n;m}$
5	5	0	22
	10	0	3
6	6	0	8561
	12	0	10
	15	0	74
10	10	69	Out of memory
	50	0	”
15	15	> 3 hours	”
	60	2	”
20	100	Out of memory	”

Table 2: Run times required to get exactly one totally symmetric or totally conjugate orthogonal partial Latin square of a given order and weight.

Let $P = (p_{ij}) \in \mathrm{PLS}_{n;m}$. We define the $n \times n$ arrays $P' = (p'_{ij})$ and $P'' = (p''_{ij})$ such that

$$p'_{ij} := \begin{cases} p_{ij} + n, \text{ if } p_{ij} \in [n], \\ 0, \text{ otherwise.} \end{cases} \quad \text{and} \quad p''_{ij} := \begin{cases} p_{ij} + 2n, \text{ if } p_{ij} \in [n], \\ 0, \text{ otherwise.} \end{cases} \tag{7}$$

Then, we define the partial Latin square $\overline{P} = (\overline{p}_{ij}) \in \mathrm{PLS}_{3n;6m}$ by means of nine $n \times n$ blocks as

$$\overline{P} :\equiv \begin{array}{|c|c|c|} \hline \mathbf{0} & P'' & P'^{(23)} \\ \hline P''^{(12)} & \mathbf{0} & P^{(132)} \\ \hline P'^{(123)} & P^{(13)} & \mathbf{0} \\ \hline \end{array} \tag{8}$$

where $\mathbf{0}$ denotes the $n \times n$ array with all its entries being zero. We call this new partial Latin square the *conjugate-extension* of P. Thus, for instance, Figure 4 shows the conjugate-extension of the partial Latin square exposed in Figure 2.

Lemma 4.1. *If $P \in \mathrm{PLS}_{n;m}$, then $\overline{P} \in \mathrm{TSPLS}_{3n;6m}$.*

Proof. The result follows from the entry set $E(\overline{P})$ once we keep in mind (7) and (8). □

Let $A_{\mathbb{K}}(P)$ denote the partial quasigroup ring over a finite field $\mathbb{K}$ of characteristic two that is related to $\overline{P}$. Particularly, we focus on the case of being $P \in \mathrm{TSPLS}_n$. If this is the case, then the definition (8) of the partial Latin square $\overline{P}$ results

$$\overline{P} \equiv \begin{array}{|c|c|c|} \hline \mathbf{0} & P'' & P' \\ \hline P'' & \mathbf{0} & P \\ \hline P' & P & \mathbf{0} \\ \hline \end{array} \tag{9}$$

 ISBN: 978-84-617-8694-7

			7	8		4	5	
				9				5
					7	6		
7						1		
8	9						1	2
		7				3		
4		6	1		3			
5				1				
	5			2				

Figure 4: Conjugate-extension of the partial Latin square $P \in \mathrm{PLS}_3$ of Figure 2.

Theorem 4.2. *Let $\mathbb{K}$ be a finite field of characteristic two and let $P \in \mathrm{TSPLS}_n$ be the multiplication table of a quasigroup $([n], \cdot)$ satisfying the* left invertive law

$$(a \cdot b) \cdot c = (c \cdot b) \cdot a, \text{ for all } a, b, c \in [n]. \tag{10}$$

Then, the partial quasigroup ring $A_{\mathbb{K}}(P)$ is a Lie algebra.

Proof. The symmetry of the partial Latin square $\overline{P} = (\overline{p}_{ij})$, with $p_{ii} = 0$, for all $i \leq 3n$, together with the fact of being $\mathbb{K}$ a finite field of characteristic two, involves $A_{\mathbb{K}}(P)$ to be anti-commutative. Now, in order to prove that the Jacobi identity (6) holds, suppose $\{e_1, \ldots, e_{3n}\}$ to be the basis of $A_{\mathbb{K}}(P)$, which we partition into the three sets $\{e_1, \ldots, e_n\}$, $\{e_{n+1}, \ldots, e_{2n}\}$ and $\{e_{2n+1}, \ldots, e_{3n}\}$. Let $S(e_i)$ denote which one of these three sets contains each basis vector e_i. From (9), we have that, if $S(e_i) = S(e_j)$, then $e_i e_j = 0$. Besides, if $S(e_i) \neq S(e_j)$ and $e_i e_j \neq 0$, then $S(e_i) \neq S(e_i e_j) \neq S(e_j)$. As a consequence, $J(e_i, e_j, e_k) = 0$, for all $i, j, k \leq 3n$ such that the three sets $S(e_i)$, $S(e_j)$ and $S(e_k)$ either coincide or are pairwise distinct. Then, from the symmetry of the Jacobi identity, it is enough to focus on the expression $J(e_i, e_j, e_k)$ in case of being $S(e_i) = S(e_j) \neq S(e_k)$. If this is the case, $e_i e_j = 0$ and hence, $J(e_i, e_j, e_k) = (e_j e_k)e_i + (e_k e_i)e_j = e_{(j \cdot k) \cdot i} + e_{(k \cdot i) \cdot j}$. The result follows from the symmetry of the partial Latin square $\overline{P}$ and the left invertive law. □

Every totally symmetric partial Latin square satisfying (10) constitutes the multiplication table of a partial totally symmetric group. In order to compute this kind of partial Latin squares, we include the following equations to (1)–(4)

$$x_{ijk} x_{kls} x_{ljt} (x_{tis} - 1) = 0, \text{ for all } i, j, k, l, s, t \in [n] \tag{11}$$

$$\left(\sum_{k \leq n} x_{ijk} - 1 \right) \left(\sum_{k \leq n} x_{ljk} \right) x_{ljt} \left(\sum_{k \leq n} x_{tik} \right) = 0, \text{ for all } i, j, l, t \in [n] \tag{12}$$

$$x_{ijk} \left(\sum_{s \leq n} x_{kls} - 1 \right) \left(\sum_{s \leq n} x_{ljs} \right) x_{ljt} \left(\sum_{s \leq n} x_{tis} \right) = 0, \text{ for all } i, j, k, l, t \in [n] \tag{13}$$

 ISBN: 978-84-617-8694-7

The implementation of these equations into our SAT solver determines, for instance, the pair of partial Latin squares exposed in Figure 5, which give rise in turn, according to Theorem 4.2, to a pair of Lie partial quasigroup rings as we have previously described.

3		1
	2	
1		3

2	1				
1	2				
		4	3		
		3	4		
				6	5
				5	6

Figure 5: Totally symmetric partial Latin squares satisfying the left invertive law.

5 Conclusion and further studies

We have described in this paper a series of binary constraints that enable us to determine the distribution of the sets TSPLS_n and TCOPLS_n of totally symmetric and totally conjugate partial Latin squares of order n, respectively, according to their weights. By means of the SAT solver Minion, we have computed the former, for all $2 \leq n \leq 6$, and the latter, for all $2 \leq n \leq 4$. A further study to improve the efficiency of the proposed method is required to deal with higher orders. Besides, we have introduced the conjugate-extension of a given partial Latin square, which gives rise to a totally symmetric partial Latin square. Particularly, the description of a family of Lie partial quasigroup rings derived from the conjugate-extension of a totally symmetric partial Latin square that holds the left invertive law has enabled us to delve into the open problem of constructing examples of this type of Lie algebras.

References

[1] R. A. Bailey, *Enumeration of totally symmetric Latin squares*, Utilitas Math. **15** (1979) 193–216. *Corrigendum*, Utilitas Math. **16** (1979) 302.

[2] G. B. Belyavskaya, *Check character systems and totally conjugate orthogonal T-quasigroups*, Quasigroups Related Systems **18** (2010) 7–16.

[3] G. B. Belyavskaya, T. V. Popovich, *Totally conjugate-orthogonal quasigroups and complete graphs*, J. Math. Sci. **185** (2012) 184–191.

[4] F. E. Bennett, *Latin squares with pairwise orthogonal conjugates*, Discrete Math. **36** (1981) 117–137.

 ISBN: 978-84-617-8694-7

[5] F. E. Bennett, *On conjugate orthogonal idempotent Latin squares*, Ars Combin. **19** (1985) 37–49.

[6] F. E. Bennett, *Conjugate orthogonal Latin squares and Mendelsohn designs*, Ars Combin. **19** (1985) 51–62.

[7] F. E. Bennett, L. S. Wu, L. Zhu, *Some new conjugate orthogonal Latin squares*, J. Combin. Theory Ser. A **46** (1987) 314–318.

[8] F. E. Bennett, L. Zhu, *On the existence of incomplete conjugate orthogonal idempotent Latin squares*, Ars Combin. **20** (1985) 193–210.

[9] F. E. Bennett, L. Zhu, *Further results on incomplete* $(3, 2, 1)$*-conjugate orthogonal idempotent Latin squares*, Discrete Math. **84** (1990) 1–14.

[10] F. E. Bennett, H. Zhang, *Latin squares with self-orthogonal conjugates*, Discrete Math. **284** (2004) 45–55.

[11] R. K. Brayton, D. Coppersmith, A. J. Hoffman, *Self-orthogonal Latin squares of all orders* $n \neq 2$*,* 3 *or* 6, Bull. Amer. Math. Soc. **80** (1974) 116–118.

[12] R. H. Bruck, *Some results in the theory of quasigroups*, Trans. Amer. Math. Soc. **55** (1944) 19–52.

[13] D. Bryant, M. Buchanan, *Embedding partial totally symmetric quasigroups*, J. Combin. Theory Ser. A **114** (2007) 1046–1088.

[14] J. Dénes, A. D. Keedwell, *Latin squares and their applications*, Academic Press, New York-London, 1974.

[15] T. Evans, *Embedding incomplete latin squares*, Amer. Math. Monthly **67** (1960) 958–961.

[16] O. J. Falcón, R. M. Falcón, J. Núñez, A. Pacheco, M. T. Villar, *Computation of isotopisms of algebras over finite fields by means of graph invariants*, J. Comput. Appl. Math., **318** (2017), 307–315.

[17] R. M. Falcón, *The set of autotopisms of partial Latin squares*, Discrete Math. **313** (2013) 1150–1161.

[18] R. M. Falcón, *Enumeration and classification of self-orthogonal partial latin rectangles by using the polynomial method*, European J. Combin. **48** (2015) 215–223.

[19] R. M. Falcón, R. J. Stones, *Classifying partial Latin rectangles*, Electron. Notes Discrete Math. **49** (2015) 765–771 .

 ISBN: 978-84-617-8694-7

[20] R. M. Falcón, R. J. Stones, *Enumerating partial Latin rectangles*, preprint.

[21] I. P. Gent, C. Jefferson, I. Miguel, *Minion: a fast scalable constraint solver*. In: G. Brewka, S. Coradeschi, A. Perini, P. Traverso. (eds.), Proceedings of the 17th European Conference on Artificial Intelligence ECAI 2006, IOS, Amsterdam (2006) 98–102.

[22] B. A. Hausmann, Øre, *Theory of Quasi-Groups*, Amer. J. Math. **59** (1937) 983–1004.

[23] K. Heinrich, L. Zhu, *Incomplete self-orthogonal Latin squares*, J. Austral. Math. Soc. Ser. A **42** (1987) 365–384.

[24] A. Hulpke, P. Kaski, P. R. J. Östergård, *The number of Latin squares of order 11*, Math. Comp. **80** (2011) 1197–1219.

[25] P. Kaski, P. R. J.Östergård, *The Steiner triple systems of order 19*, Math. Comp. **73** (2004) 2075–2092.

[26] G. Kolesova, C. W. H. Lam, L. Thiel, *On the number of* 8×8 *Latin squares*, J. Combin. Theory Ser. A **54** (1990) 143–148.

[27] C. C. Lindner, A. B. Cruse, *Small embeddings for partial semisymmetric and totally symmetric quasigroups*, J. London Math. Soc. **(2) 12** (1976) 479–484.

[28] C. C. Lindner, E. Mendelsohn, N. S. Mendelsohn, B. Wolk, *Orthogonal Latin square graphs*, J. Graph Theory **3** (1979) 325–338.

[29] B. D. McKay, A. Meynert, W. Myrvold, *Small Latin Squares*, Quasigroups and Loops, J. Combin. Des. **15** (2007) 98–119.

[30] B. D. McKay, I. M. Wanless, *On the number of Latin squares*, Ann. Comb. **9** (2005) 335–344.

[31] K. T. Phelps, *Conjugate orthogonal quasigroups*, J. Combin. Theory Ser. A **25** (1978) 117–127.

[32] M. E. Raines, *More on embedding partial totally symmetric quasigroups*, Australas. J. Combin. **14** (1996) 297–309.

[33] M. E. Raines, C. A. Rodger, *Embedding partial extended triple systems and totally symmetric quasigroups*, Discrete Math. **176** (1997) 211–222.

[34] S. K. Stein, *On the foundations of quasigroups*, Trans. Amer. Math. Soc. **85** (1957) 228–256.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

A Multi-physics Forest Fire Spread Model on Multi-core Systems

Angel Farguell[1], Ana Cortés[1], Tomàs Margalef[1], Josep R. Miró[2] and J. Mercader[2]

[1] *Computer Architecture and Operating Systems Department , Universitat Autónoma de Barcelona*

[2] *Servei Meteorològic de Catalunya,*

emails: angel.farguell@uab.cat, ana.cortes@uab.cat, tomas.margalef@uab.cat, jrmiro@meteo.cat, jmercaderc@meteo.cat

Abstract

Advances in High Performance Computing (HPC) have led to an improvement in modelling multi-physic systems because of the capacity to solve complex numerical systems in a reasonable time. WRF-SFIRE is a multi-physics system that couples the atmospheric model WRF and the forest fire spread model called SFIRE with the objective of considering the interactions atmosphere-fire. In systems like WRF-SFIRE, the trade off between result accuracy and time required to deliver that result is crucial. So, in this work, we analyze the influence of the WRF-SFIRE settings (grid resolutions) into the forecasts accuracy and into the execution times on multi-core platform.

Key words: forest fire simulation, multi-physic model, HPC.

1 Introduction

There are several factors that affect the evolution of a wildland fire. It is well known that one of the parameters that most affects forest fire propagation is the wind. Intuitively, the meteorological wind speed and wind velocity tend to drive the main direction and the rate of spread of forest fires. However, in large forest fires that take place in complex terrain where the wind fluxes can vary due to the topography and the wind convections produced by the heat generated by the fire, the effective wind speed and wind direction can

 ISBN: 978-84-617-8694-7

be unpredictable. For that reason, a multi-physics forest fire spread model that considers the feedback between the atmospheric model and the forest fire spread model could capture the micro-weather generated by a large forest fire and provide more accurate wildland fire propagations. However, this improvement in accuracy has a cost in terms of execution time. This limitation could be a serious drawback when trying to use those complex systems as operational tools for being used during a real event. However, the predictability potential of a multi-physic forest fire spread system is supposed to be better than considering the atmosphere and the forest fire as two isolate systems. There are many works that analyze this problem [6] [2] showing that any strategy that takes into account the atmosphere effect into the forest fire evolution provides better forecasted results. In fact, there are three different approaches to tackle this relationship: the unidirectional, the bidirectional and the integration strategy. The unidirectional approach uses static meteorological information, such as wind speed and direction, to drive the fire propagation, but does not consider the effect of the fire on the atmosphere [1] [4]. The second approach, the bidirectional scheme, is a coupled system where the fire propagation considers meteorological information and also on the other way round, so that the effect of heat fluxes from the fire on the atmosphere evolution is also captured [6] [3]. Finally, the third approach, the integration scheme, tries to integrate all the processes using only one simulator which combines everything [5] [7]. In this work, we focuses on the multi-physics system WRF-SFIRE with the aim of analyzing its predictability capacity compared to the time incurred in delivering the results. The main objective is to determine the viability of using WRF-SFIRE as an operational tool. The main constraint of these complex multi-physic systems is their execution time. Fortunately, some of them have been parallelized using different parallel programming paradigms such as OpenMP and MPI. Therefore, an exhaustive analysis about the trade off between results accuracy and time incurred in provide those results should be done.
Section 2 describes the WRF-SFIRE system. In section 3 a real forest fire is used to analyze the time requirements when running on a multi-core platform and the accuracy of the results depending on the WRF-SFIRE initialization settings. Finally, the main conclusions of this work are reported in section 4.

2 WRF-SFIRE

WRF-SFIRE is a forest fire simulator which couples the meteorological model WRF-ARW and the fire spread model resolving Rothermel's equation through the level set method called SFIRE. Particularly, WRF-SFIRE solves the multi-physical problem related to forest fire propagation in a cyclic fashion. That is, meteorological data is obtained by running WRF for a certain period of time and, the obtained WRF data is used to expand the fire front numerically. In order to solve these two models properly, it is necessary to do a discretization of the domain where the hazard is taking place. This domain discretization is done in two

 ISBN: 978-84-617-8694-7

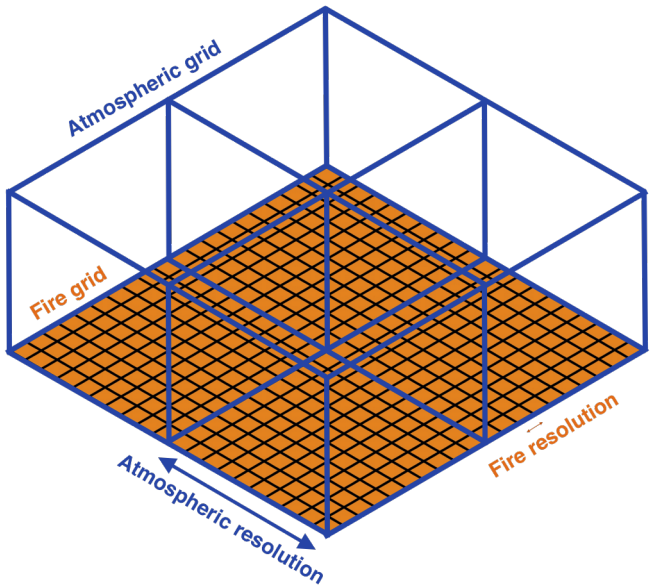


Figure 1: 3D atmospheric grid and 2D fire grid with a mesh ratio of 1:10

different meshes in order to distinguish atmospheric and fire processes. The atmospheric models works on a logically quadrilateral 3D grid on the Earth surface, whereas the fire model uses a refined 2D fire mesh posed in the Earth surface level. The recommended mesh ratio is $1:10$, so the fire grid resolution is 10 times thee atmospheric grid resolution. This grid scheme is depicted in figures 1

In order to capture the small scale meteorological processes, the atmospheric grid resolution should be as small as possible. However, reducing the grid resolution implies to increase the total system points (atmospheric points and, consequently, fire points) what has a direct effect in the WRF-SFIRE execution time. Therefore, the elections of the proper grid resolution is a critical feature to analyze. In the subsequently section, a real case has been used to study this relation, trying to highlight the weak points and the good features in order to determine a trade off among them.

 ISBN: 978-84-617-8694-7

3 Experimental study

The real case analyzed is a forest fire occurred in Catalonia (North-East of Spain) on 8th July 2005. The 2005 Cardona fire burned a total surface of 1439 ha. and it lasted 6 hours. The fire started at 14:30 and it keeps until 20:30 approximately. This particular case was labeled by the firefighter as a forest fire driven by the winds generated by the fire itself, thus, it is a perfect example to show the benefits of using a multi-physics system that takes into account the feedback between the atmosphere and fire processes. To run WRF-SIFRE, it is required to determine the domain to simulate, that is, to define a 3D cut of the Earth surface where the evolution of the forest fire is going to be forecasted. Once the domain has been determined, the equations of the atmospheric model are resolved in all domain points, so, a smart cutting would, eventually, save unnecessary execution time. The relationship between the domain size and the atmospheric and fire grid resolutions used is not the focus of this work, therefore, in this study, we used a fixed domain size of a 49 Km^2 (7 Km $\times$ 7 Km) forest area that surrounds the real fire perimeters and allows to properly simulate the propagation of the fire.

Regarding the atmospheric mesh initialization, the initial weather data used as initial conditions for the simulations is a weather data set provided by the SMC (Servei Meteorologic de Catalunya), which are at 3 Km horizontal resolution and interpolated at 19 vertical levels. The atmosphere grid resolution tested have been 236, 118, 100 and 59 meters and, consequently, the corresponding fire grid resolutions used have been 24, 12, 10 and 6 meters respectively. The coarser grid resolutions has been selected as initial grid resolutions because it corresponds to the minimum data resolution used in the system, in particular, the fuel map resolution, which has a resolution of 25 meters.

The main objectives of the experimental study reported in this section, consist of analyzing the relation between the prediction quality and the time incurred to obtain it. WRF-SFIRE has been parallelized using OpenMP and MPI but, in this work, we focused on the OpenMP parallelization. In particular, we have studied the relationship between the scalability improvements in terms of execution time due to the used of shared memory versus the quality improvements due to the different grid resolution used to solve the system.

3.1 Quality results

The main advantage of using a coupled multi-physic system to predict the evolutions of a forest fire is the ability of capturing the effect of meteorological events that happens a high resolution and, on the way back, the capacity to observe the influence of the heat fluxes generated by the fire into the atmosphere. Figure 2 shows the forest fire spread predicted by WRF-SFIRE when using different grid resolutions. The green shape corresponds to the final burnt area and the doted lines are the forecasted forest fire spread. As we can observe,

 ISBN: 978-84-617-8694-7

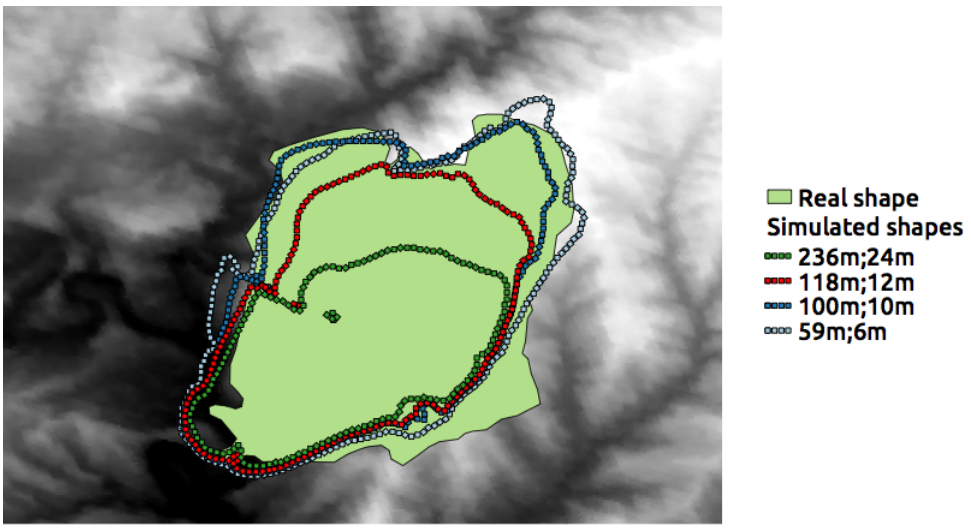


Figure 2: Cardona forest fire final burnt area compared to the forest fire prediction when using WRF-SFIRE with atmospheric grid resolutions equal to 236, 118, 100 and 59 meters

the atmospheric grid resolution plays a relevant role in the final results in terms of quality. Lower resolutions provide forest fire evolutions that are more similar to the real fire spread than the ones obtained using higher resolutions. The main reason of these results is the capacity of WRF to better detect local winds convections due, not only to the atmosphere effects but also because of the forest fire. However, a question that arises at this point is, can we increase the mesh resolution as much as we want in order to obtain better prediction results? The answer is not straight forward. Figure 2 depicts the predicted fire evolutions when using 236, 118, 100 and 59 meters as atmospheric grid resolution. If we focus on the cases where this value is 236, 118 and 59 meters, we can observe that each grid resolution is half the previous one. This variation clearly affects the quality of the results. However, as we will see in the next section, low resolution simulations imply higher execution times, for that reason, we also test an intermediate case with the atmospheric grid resolution equal to 100 meters. As we can see, this case provides a forest fire spread similar to the one obtained with a grid resolutions equal to 59 meters but, as it is following explained, the execution time is significantly lower. Therefore, it is important to be able to determine a relation between grids resolutions and quality if we want to use this multi-physics system in an operational way.

 ISBN: 978-84-617-8694-7

3.2 WRF-SFIRE scalability

As it has been introduced, WRF-SFIRE is a multi-physics system that couples the WRF atmospheric model and the SFIRE forest fire spread model. WRF requires a 3D grid to solve the atmospheres processes, meanwhile SFIRE works in a 2D mesh. The execution platform used for the experiments reported in this section is a multi-core system composed of 2 sockets integrating Intel Xeon processors with 8 cores and multithreading. Figure 3 shows the execution time spent when simulating the Cardona fire using a single thread approach for all tested grid resolutions. Moreover, the execution time spent for each individual model is also depicted. As we can see, lower resolutions imply higher executions times varying from, approximately, 1000 to 10000 minutes depending on the resolution used. Whatever grid resolution we use, the execution times obtained are prohibitively for operational purposes. It is also remarkable what happened when using a atmospheric grid resolution of 236 meters. As we can see, the total execution time for this case and for the experiment with atmospheric grid resolution equal to 118 meters are quite similar. This could be seen as an anomaly, however, there is a reason to explain this behavior. As it was previously mentioned, WRF uses a 3D grid to solve the atmospheric equations and, for all the experiments, the number of vertical level was initially set to 29. However, when executing the case with a grid resolution equal to 236 meters the system was not able to converge to a solution due to the reduce number of grid points used. For that reason, in this particular case, the number of vertical level was increased to 200 to include the minimum required number of points that allows the system to converge. Since the final number of points in the atmospheric grid has been similar to the case where the grid resolution is equal to 118 with 19 vertical levels, the total execution time is quite similar. However, if we analyze the execution times of WRF and SFIRE separately, we can observe that for the case of 236 meters grid resolution the time required to execute WRF is almost 90% of the total execution time, whereas in the case of 118 meters grid resolution the total execution time is balanced among both models.

Fortunately, WRF-SFIRE has been parallelized using OpenMP and MPI so, in order to be able to use the system during an ongoing event, we have studied the behavior of the multi-physics system on multi-core platforms. Since the simulations that provide better quality results are the ones with atmospheric grid resolutions equal to 100 and 59 meters, we have focused the study on these two cases. Figures 4 and 5 show, respectively, the ideal execution time, the real total execution time and the execution time of the two models (WRF and SFIRE) when using 2, 4, 8 and 16 threads locating each thread in a different core. As we can observe, the scalability of the system is quite good because the total execution time has almost the same tendency that the ideal case.

However, in terms of absolute execution time incurred in both cases, it is clear that the one that could reach the execution time requirements for operational purposes, is the experiment with atmospheric grid resolution equal to 100 and running using either 8 or 16

 ISBN: 978-84-617-8694-7

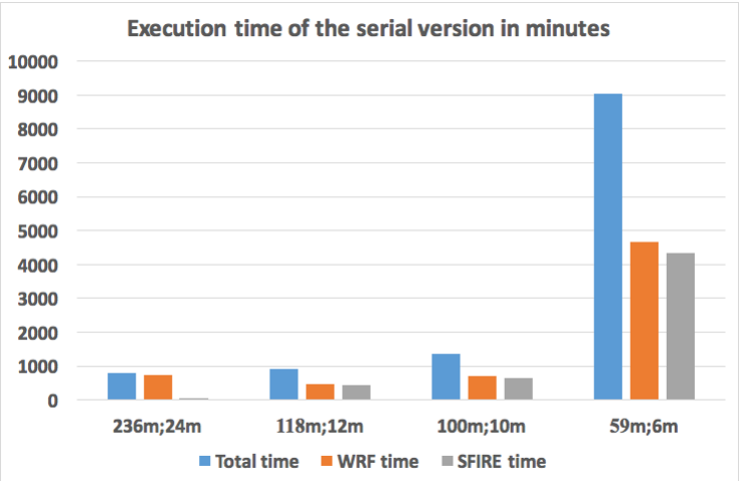


Figure 3: Total serial execution time of simulating the Cardona fire using WRF-SFIRE with atmospheric grid resolutions equal to 236, 118, 100 and 59 meters

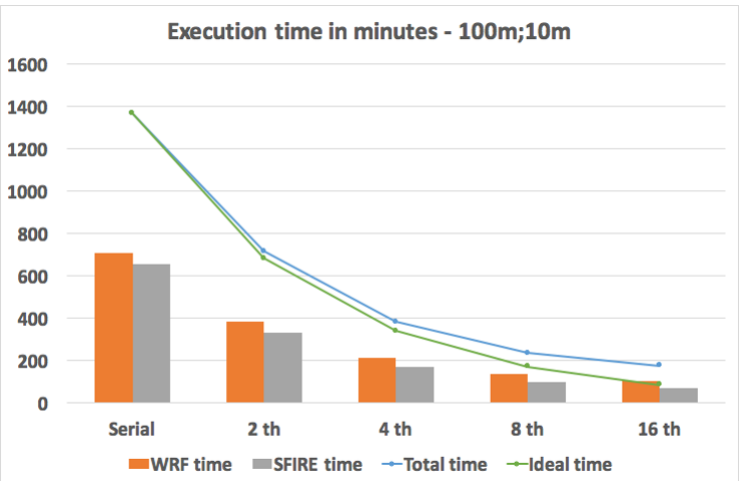


Figure 4: Execution time of simulation Cardona fire using WRF-SFIRE with atmospheric grid resolutions 100 meters and OpenMP with 2, 4, 8 and 16 threads

threads. Consequently, a deeper study to be able to asses in advance the appropriate grid settings and the hardware requirements to cope with operational constraints must be done.

 ISBN: 978-84-617-8694-7

Figure 5: Execution time of simulation Cardona fire using WRF-SFIRE with atmospheric grid resolutions 59 meters and OpenMP with 2, 4, 8 and 16 threads

4 Conclusions

Atmospheric wind is one of the parameters that most affect the forest fire behavior. However, this atmospheric phenomenon can be hardly affected by the forest fire micro-weather generated by the fire itself. WRF-SFIRE is a coupled atmospheric-fire system that takes into account the continuous relation between both models. However, the penalty of continuously evaluating the evolution of the atmosphere processes and the fire processes is the execution time. WRF-SFIRE has an OpenMP parallel implementation that allows to exploit multi-core systems. This parallel approach has been tested using as study case: the Cardona fire. This forest fire was a fire driven by the winds generated by the fire itself, so to run the simulations with low resolutions is the best approach to cope the local effect of the fire in the atmosphere. However, to reach operational times, it is necessary to exploit to the maximum the OpenMP parallelization of WRF-SFIRE taking into account convergence features of the system that can go from modifying the grid resolutions to improve the vertical levels of the meteorological grid in order to have enough point to allow the system to converge. Therefore, a further analysis of this relation is required to be able to develop a model that could predict in advance the required grids settings and number of threads required to accomplish the predetermined deadline times.

 ISBN: 978-84-617-8694-7

Acknowledgements

This work has been supported by Ministerio de Economía y Competitividad (Spain) under contract number TIN2014-53234-C2-1-R.

References

[1] P. L. Andrews. Behaveplus fire modeling system: past, present, and future. *7th Symposium on Fire and Forest Meteorology*, Paper J2.1, 2007.

[2] Carlos Brun, Tomàs Artés, Tomàs Margalef, and Ana Cortés. Coupling wind dynamics into a DDDAS forest fire propagation prediction system. *Procedia Computer Science*, 9:1110–1118, 2012.

[3] Jean-Baptiste Filippi, Frdric Bosseur, Xavier Pialat, Paul-Antoine Santoni, Susanna Strada, and Cline Mari. Simulation of coupled fire/atmosphere interaction with the mesonh-forefire models. *Journal of Combustion*, 2011, 2011.

[4] M. A. Finney. Farsite: Fire area simulator-model development and evaluation. *U.S. Department of Agriculture: Forest Service*, Research Paper RMRS-RP-4, 1998.

[5] R. Linn, J. Reisner, J. J. Colman, and J. Winterkamp. Studying wildfire behaviour using firetec. *Int. J. Wildland Fire*, 11: 233–246, 2002.

[6] J. Mandel, J. D. Beezley, and a. K. Kochanski. Coupled atmosphere-wildland fire modeling with WRF 3.3 and SFIRE 2011. *Geoscientific Model Development*, 4(3):591–610, 2011.

[7] W. Mell, M. A. Jenkins, J. Gould, and P. Cheney. A physics-based approach to modelling grassland fires. *Int. J. Wildland Fire*, 16: 1–22, 2007.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference on Computational and Mathematical Methods in Science and Engineering, CMMSE 2017 4–8 July, 2017.

High-Performance Computing for Optimizing High-Pressure Thermal Treatments in Food Processing

M.R. Ferrández[1], S. Puertas-Martín[1], J.L. Redondo[1], B. Ivorra[2], A.M. Ramos[2] and P.M. Ortigosa[1]

[1] *Department of Computer Sciences & Agrifood Campus of International Excellence (ceiA3), University of Almería*

[2] *Department of Applied Mathematics & Institute of Interdisciplinary Mathematics (IMI), Complutense University of Madrid*

emails: mrferrandez@ual.es, savinspm@ual.es, jlredondo@ual.es, ivorra@mat.ucm.es, angel@mat.ucm.es, ortigosa@ual.es

Abstract

In this work, we deal with the optimization of a High-Pressure Thermal (HPT) process for treating food samples. In particular, this industrial problem has several conflicting objectives as the vitamin retention and the enzymatic reduction or the temperature control. Recently, this optimization problem has been solved by using a multi-objective optimization algorithm called WASF-GA. Its main objective is to provide a good size approximation of the Pareto-front, i.e., a fixed number of well-distributed solutions which cover a region of interest determined by the preferences of a decision maker. WASF-GA works with a list of individuals, which evolves during the optimization procedure. In this scenario, the evaluation of a single individual is really high from a computational point of view, since it consists of solving a partial differential equation system that simulates the HPT treatment. Therefore, when the set approximating the Pareto-front must have many points (because a high precision is required), the number of function evaluations can explode and hence, the computational time needed by WASFGA may not be negligible at all. Furthermore, the computational resources needed may be so high that a PC may run out of memory. In those cases, parallelizing the algorithm and running it in a supercomputer may be the best way forward. In this work, we propose a parallelization of the optimization algorithm by using a master-slave strategy based on assigning a fraction of the population to each available processing element. The efficiency of this parallel strategy has been analyzed and the preliminary results are optimistic.

Key words: high-performance computing, multi-objective optimization, genetic algorithms, food industry, high-pressure

1 Introduction

High-Pressure (HP) technology is widely used in food treatment processes. In the last decades, its popularity has grown significantly due to the increasing demand for healthy and safe products, minimally processed, and at the same time, ready for immediate consumption. Among other food treatments, HP stands out for the two following advantages: it does not use additives that consumers prefer to avoid, and it is not based on extremely high or low temperatures, which can affect nutritional and organoleptic properties of the food.

In this work, we focus on a particular food treatment device (a cylindrical chamber) which is used to apply a combination of High-Pressure and Thermal processes to treat food samples. We model the behavior of this device by considering the heat transfer equations, to model the variations of pressure and temperature, coupled with a first–order kinetic equation, that describes the effect of such variations on the activity of certain enzymes and vitamins [5]. Consequently, given the pressure profile and the initial and refrigeration temperature, this model simulates the variation of temperatures and the evolution of enzymatic and vitamin activities in the food sample. Then, based on this mathematical model, we define a multi-objective optimization problem, associated with this High-Pressure Thermal treatment, which consists of determining the initial and refrigeration temperatures and the pressure profile provided to the equipment in order to minimize the final enzymatic activity and the maximum temperature reached during the whole process and to maximize the final vitamin activity.

In multi-objective optimization problems, as the one considered here, there is usually no single optimal solution, but a set of alternative results with different trade-offs. Such a set of solutions is called the Pareto set (or efficient set), and the corresponding set of objective vectors, the Pareto front. Then, solving a multi-objective optimization problem consists of finding the nondominated subset formed by the efficient decision vectors whose corresponding objective vectors represent the Pareto optimal front.

For the problem at hand, obtaining an exact description of the efficient set (or Pareto front) is practically impossible, since those sets are usually a continuum and include an infinite number of points. Furthermore, the computing cost may be excessive, and this is an important aspect, mainly for hard-to-solve optimization problems, such as the one considered in this work. In this context, the Preference-based Multi-Objective Evolutionary Algorithms (PMOEAs) allow us to obtain 'good approximations' of the region of interest (ROI). A good *Pareto front approximation* is defined as a finite set of non-dominated objective vectors which cover the whole ROI evenly.

In particular, to solve our optimization problem, we have used the preference-based algorithm called WASF-GA. Nevertheless, even when WASF-GA focuses on a region of interest and, thus, avoids unnecessary computations, its execution time is not negligible. On the contrary, WASF-GA needs nearly 55 hours to find the Pareto front approximation with 300 individuals and 36 iterations (i.e. 10800 evaluations), and more than 100 hours for

 ISBN: 978-84-617-8694-7

obtaining an approximated set composed by 500 individuals in 40 iterations of the algorithm (i.e. 20000 evaluations)). This is mainly due to the evaluation of the objective function is very costly.

Nevertheless, it is important to achieve a representative set of optimal individuals so that the food engineer receives as many solutions as possible for covering the industry requirements at each particular moment. Therefore, once the Pareto front approximation is computed, the decision maker (DM) has available a set of points that are individually good solutions for many different constrained mono-objective problems. Then, the DM can choose the most preferable one depending on the specific quality requirements that he/she wants to satisfy at each moment without having to execute a new optimization procedure. This translates into popupulations with larger number of individuals and hence, longer computing times. A parallelization of WASF-GA may allow us to deal with these inconvenients. In this work, WASF-GA is parallelized by considering a a master-slave paradigm.

The rest of the paper is structured as follows. First of all, in Section 2, we present the multi-objective optimization problem under study. Next, in Section 3, we describe the sequential algorithm WASF-GA used to solve it. Then, in Section 4, we explain the master-slave strategy proposed for parallelizing WASF-GA. Finally, Section 5 is devoted to summing up the preliminary results and the main conclusions.

2 The multi-objective problem

The industrial problem concerning us is obtaining the best configuration for a particular HPT equipment and a specific food sample so that, after the treatment, the food satisfies some quality requirements. More precisely, the HPT configuration that we are interested in determining is given by the initial and refrigeration temperatures and the pressure profile provided to the equipment. Additionally, the considered quality requirements are based on the inactivation of the enzyme called *Bacillus Subtilis* α*–Amylase* (BSAA) [2], the preservation of the vitamin C [7], and the control of the food temperature for maintaining it as lower as possible. Remember that high temperatures damage the food properties, and, in particular, the vitamins. However, the pressure has to be increased for fighting against the enzyme and, as a consequence, the temperature also experiments a rise.

Therefore, for managing those conflicting objectives and optimizing them simultaneously, our problem is formulated as a multi-objective problem as follows:

$$\begin{cases} \min & f_{\text{bsaa}}(T_0, T_r, P), \\ \max & f_{\text{vit}}(T_0, T_r, P), \\ \min & f_{\text{Tmax}}(T_0, T_r, P), \end{cases} \tag{1}$$

where the decision variables are the initial, T_0, and refrigeration, T_r, temperatures and a piece-wise linear function P (depending on time), defining the equipment pressure evolution.

 ISBN: 978-84-617-8694-7

Additionally, those decision variables are constrained due to the equipment restrictions detailed below (e.g., temperature and pressure admissible range).

Given a particular configuration (T_0, T_r, P), it determines a specific HPT process whose behavior is simulated solving numerically the heat transfer system explained at [5]. Moreover, two first-order kinetic equations for describing the activity evolution of both the considered enzyme and vitamin are also involved. As a consequence, the activity values of enzyme A_{bsaa} and vitamin A_{vit} at the end of the process (i.e. at time t_f) can be obtained from this numerical simulation. Then, they are averaged in the food sample domain Ω_{F} to evaluate the first and the second objective functions, respectively:

$$f_{\text{bsaa}}(T_0, T_r, P) = \frac{1}{|\Omega_{\text{F}}|} \int_{\Omega_{\text{F}}} A_{\text{bsaa}}(r, z, t_f) dr dz,$$

$$f_{\text{vit}}(T_0, T_r, P) = \frac{1}{|\Omega_{\text{F}}|} \int_{\Omega_{\text{F}}} A_{\text{vit}}(r, z, t_f) dr dz.$$

The third objective function f_{Tmax} is the maximum temperature reached in the food sample during the whole HPT treatment, which can be expressed by:

$$f_{\text{Tmax}}(T_0, T_r, P) = \max_{(r,z)\in\Omega_{\text{F}},\ t\in[t_0,t_f]} T(r, z, t).$$

In order to design the pressure evolution function P, we consider an initial equipment pressure $P_0 = 0.1$ (MPa) (i.e., atmospheric pressure) at time $t_0 = 0$ seconds. Then, we consider $n+1$ time intervals $[t_i, t_{i+1}]$, with $t_i = i \cdot \frac{900}{n}$, $i \in \{1, \cdots, n\}$, and, at each interval, we consider a constant pressure variation $\Delta P_i \in [\Delta P_{n,\text{dec}}, \Delta P_{n,\text{inc}}]$ (MPa), where $\Delta P_{n,\text{dec}}$ and $\Delta P_{n,\text{inc}}$ are the maximum variations allowed by the equipment for the decrease and for the increase in pressure, respectively, during $900/n$ seconds. However, those pressure variations cannot generate pressure out of the equipment admissible pressure range $[P_{\min}, P_{\max}]$ (MPa). Finally after 900 seconds, the pressure is decreased at a constant fixed rate of $\Delta P_{n,\text{dec}} \cdot (n/900)$ (MPa$\cdot$s^{-1}) up to reaching 0.1 MPa. Thus, $P(t)$ is built by considering the linear interpolation through the points $\{(i*900/n\ (\text{s}), P_i\ (\text{MPa})\ )\}_{i=0}^{n} \cup \{(900+(P_n-0.1)/|\Delta P_{n,\text{dec}}|\ (\text{s}), 0.1\ (\text{MPa})\ )\}$, with $P_i = P_{i-1} + \Delta P_i$ for $i \in \{1, \cdots, n\}$.

In this work, taking into account the restriction of the considered pressure vessel, we consider that the range of admissible temperature is [10, 50] (°C), the range of admissible pressure is [0.1, 900] (MPa) and the range of admissible variations for the pressure is $[\Delta P_{n,\text{dec}}, \Delta P_{n,\text{inc}}] = [-250, 250]$ (MPa).

Thus, the optimization problem considered is the following discrete version of Problem (1):

$$\begin{cases} \min & f_{\text{bsaa}}(T_0, T_r, \Delta P_1, \ldots, \Delta P_n), \\ \max & f_{\text{vit}}(T_0, T_r, \Delta P_1, \ldots, \Delta P_n), \\ \min & f_{\text{Tmax}}(T_0, T_r, \Delta P_1, \ldots, \Delta P_n). \\ \text{s.t.} & T_0, T_r \in [10, 50](^\circ\text{C}) \\ & \Delta P_1, \ldots, \Delta P_n \in [-250, 250](\text{MPa}) \end{cases} \tag{2}$$

 ISBN: 978-84-617-8694-7

3 The sequential method: the weighting achievement scalarizing function genetic algorithm

The weighting achievement scalarizing function genetic algorithm, shortly called WASF-GA, is a preference-based evolutionary algorithm aimed at multi-objective optimization. For taking into account the decision maker preferences, it considers a reference point (RP) $\mathbf{q} = (q_1, \ldots, q_m)$, consisting of a desirable value q_i for each one of the objective functions f_i, with $i = 1, \ldots, m$.

Additionally, the DM should provide a set of N_μ weight vectors $\{\mu^1, \ldots, \mu^{N_\mu}\}$, where each weight vector has a number of components equal to the number m of objective functions, i.e., $\mu^j = (\mu_1^j, \ldots, \mu_m^j)$ and $\mu_i^j > 0$ for $j = 1, \ldots, N_\mu$ and $i = 1, \ldots, m$.

Then, WASF-GA uses the following Wierzbicki's achievement scalarizing function (ASF) based on the L_∞ distance:

$$s(\mathbf{q}, \mathbf{f}(\mathbf{x}), \mu^j) = \max_{i=1,\ldots,m} \{\mu_i^j (f_i(\mathbf{x}) - q_i)\} + \eta \sum_{i=1}^{m} \mu_i^j (f_i(\mathbf{x}) - q_i), \tag{3}$$

where the parameter $\eta > 0$, called augmentation coefficient, must be a small positive value.

According to Equation (3), for evaluating the ASF, $\mathbf{q}$ is projected onto the optimal front in the direction determined by the inverses of the weights. Therefore, in order to obtain a well-distributed approximation of the region of interest, the set of weight vectors must be selected so that the inverse of its components μ_i^j must be evenly distributed in the space $(0, 1)^m$.

The general behaviour of the WASF-GA can be summarized as follows. First of all, an initial population of N individuals is randomly generated with a uniform distribution. Next, at each iteration, an offspring population with the same number of individuals is obtained from the previous population by applying the crossover and mutation operators. Then, the individuals belonging to both parent and offspring populations are classified into different fronts. To do so, all of those individuals are evaluated through the objective function and, for each one, the N_μ values of the ASFs for the N_μ weight vectors are calculated. Attending to those values, for each weight vector, the individual with the lowest value of its ASF is selected and copied into the first front. Once a particular individual is chosen for minimizing an ASF associated to a weight vector, this individual will not be considered for the remaining ASFs. Thus, the first front will be complete when it contains N_μ individuals, one minimizing each ASF. For filling the following fronts, the same procedure is carried out. Finally, the parent population for the next iteration is composed of the N first selected individuals belonging to the first fronts. In our study, we have considered $N = N_\mu$, so that, at each iteration, the parent population will be the first front of the previous iteration.

 ISBN: 978-84-617-8694-7

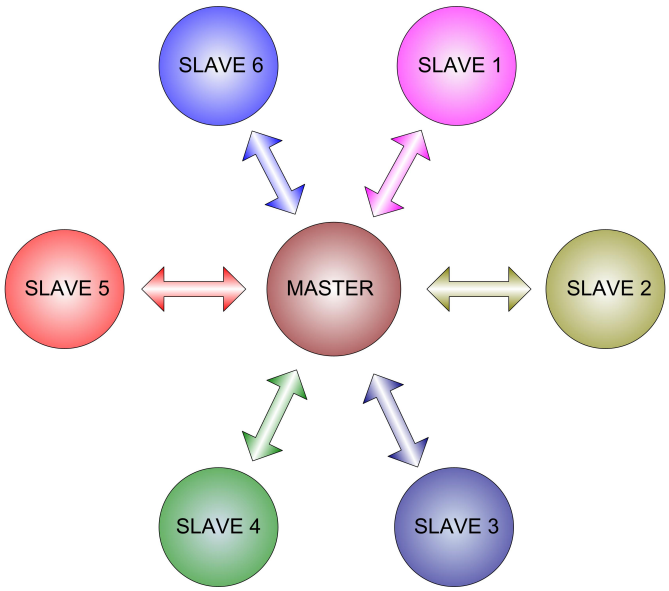


Figure 1: Master-slave model

4 Parallelization Approach: A master-slave strategy

Master-slave is a parallel communication model where one processing element called the *master* has unidirectional control over one or more processing elements known as the *slaves*. This technique is a "global parallel model" in the sense that all the decisions taken during the algorithm consider the whole population. Indeed, the global decisions are made by the master processor which also distributes the information among all the processing elements (including itself) [1]. On the other hand, the slave processors are in charge of run specific tasks as the evaluation of the assigned population.

Figure 1 depicts the communication procedure among processors. The central core is the *master*, which interchanges some information with the *slaves*. Such interchanges are carried out using communications, that are represented by arrows.

The execution time of a master-slave model has three basic components: the time used in computations, the time used to communicate information among processing elements, and the waiting time due to the synchronization points. The first one is largely determined by the size of the population. However, the population size is also a major factor in the effectiveness of population-based methods and if the population is reduced, then the probability that the algorithm will find good solutions would decrease [3, 4]. The second one depends directly on the number of slaves, on the particular hardware used to execute the algorithm and on the size of the messages. Finally, the third one depends on the degree of parallelism of the

 ISBN: 978-84-617-8694-7

algorithm, i.e. if there exist or not parts that have to be executed in a sequential way, and if the different parallel subtasks can evolve independently or, on the contrary, they need some information from the other ones to continue.

In the master-slave strategy that we have implemented, the master processor executes WASF-GA sequentially. The parallelism derives from the simultaneous resolution of the HPT system to evaluate the new candidate solutions. This is because the evaluation of an individual is independent from the rest of the population, and there is no need to communicate during this phase. The evaluation of individuals is parallelized by assigning a fraction of the population to each available processing element (master and slaves processors). Communications occur only as each slave receives its subset of individuals for evaluation and when the slaves return the objective function values. In our proposal, the master processor, after evaluating its assigned subpopulation, receives those objective function values from the slave processors. This is a synchronization point due to the necessity of having the objective function values of the whole population before applying the selection procedure to compose the different fronts. However, this selection task is not as time-consuming as the evaluation stage. Therefore, it is important to emphasize that, the master-slave method does not affect the behavior of the algorithm, i.e. the selection mechanism takes into account the entire population and it is possible to mate with any individual.

5 Preliminary results and conclusions

The preliminary results are promising, i.e. for the instances considered, the proposed master-slave strategy applied to WASF-GA allows to reduce significantly the computational time. This is an important aspect, mainly for hard-to-solve optimization problems, such as the one considered in this work. Thanks to the proposed parallelization, we can solve our industrial problem using a larger number of evaluations. As a consequence, the set of optimal points provided to the food engineer could have more individuals, since we can use larger populations, and it could be more accurate since we can carry out more iterations of the optimization algorithm.

Acknowledgements

This research has been funded by grants from the Spanish Ministry of Economy and Competitiveness (TIN2015-66680-C2-1-R and MTM2015-64865P); Junta de Andalucía (P11-TIC7176 and P12-TIC301), in part financed by the European Regional Development Fund (ERDF). Juana López Redondo is a fellow of the Spanish "Ramón y Cajal" contract program, co-financed by the European Social Fund.

References

[1] E. Cantú-Paz, *Designing efficient master-slave parallel genetic algorithms.*, Genetic Programming 1998: Proceedings of the Third Annual Conference (1998) 455.

[2] S. Denys, L.R. Ludikhuyze, A.M. Van Loey, M.E. Hendrickx, *Modeling Conductive Heat Transfer and Process Uniformity during Batch High-Pressure Processing of Foods*, Biotechnology Progress **16**(1) (2000), 92–101.

[3] D.E. Goldberg, K. Deb, J.H. Clark, *Genetic algorithms, noise, and the sizing of populations*, Complex Systems **6** (1992) 333–362.

[4] G. Harik, E. Cantu-Paz, D.E. Goldberg, B. Miller, *The gambler's ruin problem, genetic algorithms, and the sizing of populations*, Evolutionary Computation **7**(3) (1999) 231–253.

[5] J.A. Infante, B. Ivorra, A.M. Ramos, J.M. Rey, *On the Modelling and Simulation of High Pressure Processes and Inactivation of Enzymes in Food Engineering*, M3AS **19**(12) (2009), 2203–2229.

[6] A. Ruiz, R. Saborido, M. Luque, *A preference-based evolutionary algorithm for multiobjective optimization: the weighting achievement scalarizing function genetic algorithm*, JOGO **62**(1) (2015) 101–129.

[7] L. Verbeyst, R. Bogaerts, I. Van der Plancken, M. Hendrickx, A.M. Van Loey, *Modelling of vitamin C degradation during thermal and high-pressure treatments of red fruit*, Food and Bioprocess Technology **6**(4) (2013), 1015–1023.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

Inference in models with two-layer block compound symmetry covariance structure

Miguel Fonseca[1] **and Carlos A. Coelho**[1]

[1] *Center of Mathematics and Applications, Department of Mathematics, Faculty of Sciences and Technology, Nova University of Lisbon*

emails: fmig@fct.unl.pt, cmac@fct.unl.pt

Abstract

This paper deals with estimation and hypothesis testing in multivariate mixed linear models in which observation vectors have block compound symmetry covariance structure.

Key words: block compound symmetry, hypothesis testing, multivariate mixed linear models

1 Unstructured Mean Vector Within k Groups

1.1 Model

Consider an observations vector of size mp with distribution $\mathbf{Y}_{ij} \sim \mathrm{N}(\boldsymbol{\mu}_0 + \boldsymbol{\mu}_i, \boldsymbol{\Sigma})$, $i = 0, \ldots, k$, $j = 1, \ldots, n_i$, such that $\sum_{i=1}^{k} n_i \boldsymbol{\mu}_i = \mathbf{0}_{mp}$, where $\boldsymbol{\mu}_i = \begin{bmatrix} \mu_{i,1} & \ldots & \mu_{i,mp} \end{bmatrix}'$, $m = m_1 m_2$ and $n = \sum_{i=1}^{k} n_i$. The joint distribution of the model will be $\mathbf{Y} \sim \mathrm{N}\big(\boldsymbol{\mu}, \mathrm{diag}_{i=1}^{k}(\mathbf{I}_{n_i} \otimes \boldsymbol{\Sigma})\big)$, with $\boldsymbol{\mu} = \begin{bmatrix} \big(\mathbf{1}_{n_1} \otimes (\boldsymbol{\mu}_0 + \boldsymbol{\mu}_1)\big)' & \cdots & \big(\mathbf{1}_{n_k} \otimes (\boldsymbol{\mu}_0 + \boldsymbol{\mu}_k)\big)' \end{bmatrix}'$, vector of size nmp, with $n = \sum_{i=1}^{k} n_i$. The covariance matrix $\boldsymbol{\Sigma}$ has a 2-level BCS structure, with $\boldsymbol{\Sigma}_0, \boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$ matrices of size $p \times p$, with each "layer" of size m_1 and m_2, respectively. In algebraic form:

$$\begin{aligned} \mathbf{I}_{m_1 m_2} \otimes (\boldsymbol{\Sigma}_0 - \boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2) + \mathbf{I}_{m_1} \otimes (m_2 \mathbf{J}_{m_2}) \otimes (\boldsymbol{\Sigma}_1 - \boldsymbol{\Sigma}_2) + (m_1 m_2 \mathbf{J}_{m_1 m_2}) \otimes \boldsymbol{\Sigma}_2 \\ = \mathbf{Q}_0 \otimes \boldsymbol{\Delta}_0 + \mathbf{Q}_1 \otimes \boldsymbol{\Delta}_1 + \mathbf{Q}_2 \otimes \boldsymbol{\Delta}_2, \quad (1) \end{aligned}$$

with

$$\mathbf{Q}_0 = \mathbf{K}_{m_2} \otimes \mathbf{I}_{m_1}, \ \mathbf{Q}_1 = \mathbf{J}_{m_2} \otimes \mathbf{K}_{m_1}, \ \mathbf{Q}_2 = \mathbf{J}_{m_2} \otimes \mathbf{J}_{m_1}, \quad (2)$$

 ISBN: 978-84-617-8694-7

and

$$\boldsymbol{\Delta}_0 = \boldsymbol{\Sigma}_0 - \boldsymbol{\Sigma}_1,\ \boldsymbol{\Delta}_1 = \boldsymbol{\Sigma}_0 - (m_1-1)\boldsymbol{\Sigma}_1 - m_1\boldsymbol{\Sigma}_2,$$
$$\boldsymbol{\Delta}_2 = \boldsymbol{\Sigma}_0 - (m_1-1)\boldsymbol{\Sigma}_1 + m_1(m_2-1)\boldsymbol{\Sigma}_2, \quad (3)$$

where $\mathbf{J}_s$ is a matrix whose elements are all $\frac{1}{s}$ and $\mathbf{K}_s = \mathbf{I}_s - \mathbf{J}_s$. Consider the matrix

$$\mathbf{H} = \begin{bmatrix} \mathbf{H}_{m_2} \otimes \mathbf{I}_{m_1} & \mathbf{H}_{m_2} \otimes \mathbf{J}_{m_1} & \mathbf{J}_{m_2} \otimes \mathbf{J}_{m_1} \end{bmatrix} \quad (4)$$

where $\mathbf{H}_n : \mathbf{H}_n'\mathbf{H}_n = \mathbf{I}_{n-1} \wedge \mathbf{H}_n\mathbf{H}_n' = \mathbf{K}_n$. Taking matrix $\mathbf{D} = \operatorname{diag}_{i=1}^{k}(\mathbf{I}_{n_i} \otimes \mathbf{H} \otimes \mathbf{I}_p)$, which is an orthogonal transformation, we get that

$$\mathbf{D}'\mathbf{Y} \sim \mathrm{N}\left(\underset{i=1}{\overset{k}{\operatorname{diag}}} (\mathbf{1}_{n_i} \otimes \mathbf{I}_{mp}) \begin{bmatrix} (\mathbf{H}' \otimes \mathbf{I}_p)(\boldsymbol{\mu}_0 + \boldsymbol{\mu}_1) \\ \vdots \\ (\mathbf{H}' \otimes \mathbf{I}_p)(\boldsymbol{\mu}_0 + \boldsymbol{\mu}_k) \end{bmatrix}, \underset{i=1}{\overset{k}{\operatorname{diag}}} (\mathbf{I}_{n_i} \otimes \boldsymbol{\Delta}) \right). \quad (5)$$

Note that $\mathbf{H}\boldsymbol{\Sigma}\mathbf{H}' = \operatorname{diag}_{i=0}^{2}(\mathbf{I}_{g_i} \otimes \boldsymbol{\Delta}_i) = \boldsymbol{\Delta}$, where

$$g_i = \begin{cases} m_2(m_1-1), & i = 0 \\ m_1 - 1, & i = 1 \\ 1, & i = 2 \end{cases} \quad (6)$$

Considering the transformation $\mathbf{F} = \mathbf{H}_n \otimes \mathbf{I}_{mp}$, where $\mathbf{H}_s$ is a Helmert matrix of size s, the mean vector of $\mathbf{Z} = \mathbf{F}'\mathbf{D}'\mathbf{Y}$ has the form

$$\mathrm{E}[\mathbf{Z}] = \operatorname{diag} \begin{bmatrix} \boldsymbol{\tau}_0 \\ \left(\left((\mathbf{H}_n)' \operatorname{diag}_{i=1}^{k} \left(\frac{1}{\sqrt{n_i}} \mathbf{1}_{n_i} \mathbf{N} \right) \right) \otimes \mathbf{I}_{mp} \right) \boldsymbol{\tau}_F \end{bmatrix}, \quad (7)$$

with $\boldsymbol{\tau}_0 = (\mathbf{H}' \otimes \mathbf{I}_p)(\sqrt{n}\boldsymbol{\mu}_0)$ and $\boldsymbol{\tau}_F = \begin{bmatrix} \boldsymbol{\tau}_1' & \cdots & \boldsymbol{\tau}_{k-1}' \end{bmatrix}'$, and $\operatorname{Cov}(\mathbf{Z}) = \mathbf{I}_n \otimes \boldsymbol{\Delta}$.

1.2 Maximum Likelihood Estimators

The log-likelihood ℓ of $\mathbf{Z}$ can be expressed as

$$-2\ell = nmp\log(2\pi) + n\sum_{i=0}^{2} g_i \log\left(|\boldsymbol{\Delta}_i|\right) + \operatorname{tr}\left((\mathbf{z}_0 - \boldsymbol{\tau}_0)'\boldsymbol{\Delta}^{-1}(\mathbf{z}_0 - \boldsymbol{\tau}_0)\right)$$
$$+ \operatorname{tr}\Big(\left(\mathbf{z}_F - (\mathbf{X}_F \otimes \mathbf{I}_{mp})\boldsymbol{\tau}_F\right)'(\mathbf{I}_{n-1} \otimes \boldsymbol{\Delta}^{-1})(\mathbf{z}_F - (\mathbf{X}_F \otimes \mathbf{I}_{mp})\boldsymbol{\tau}_F) \Big), \quad (8)$$

with

$$\mathbf{z}_0 = \operatorname{diag}(\mathbf{I}_{mp}, \mathbf{0}_{(n-1)mp,(n-1)mp})\mathbf{z}, \quad (9)$$
$$\mathbf{z}_F = \operatorname{diag}(\mathbf{0}_{mp,mp}, \mathbf{I}_{(n-1)mp})\mathbf{z}, \quad (10)$$
$$\mathbf{X}_F = (\mathbf{H}_n^{\perp})' \underset{i=1}{\overset{k}{\operatorname{diag}}} \left(\frac{1}{\sqrt{n_i}} \mathbf{1}_{n_i} \mathbf{N} \right). \quad (11)$$

 ISBN: 978-84-617-8694-7

Taking

$$\mathbf{B} : \mathbf{B}'\mathbf{B} = \mathbf{I}_{k-1} \wedge \mathbf{B}\mathbf{B}' = \mathbf{X}_F(\mathbf{X}_F'\mathbf{X}_F)^{-1}\mathbf{X}_F', \tag{12}$$

$$\mathbf{B}^\perp : (\mathbf{B}^\perp)'\mathbf{B}^\perp = \mathbf{I}_{n-1} \wedge \mathbf{B}^\perp(\mathbf{B}^\perp)' = \mathbf{I}_{n-1} - \mathbf{X}_F(\mathbf{X}_F'\mathbf{X}_F)^{-1}\mathbf{X}_F', \tag{13}$$

$$h_s = \begin{cases} 0, & s < 0 \\ \sum_{t=0}^{s} g_t, & s \geq 0 \end{cases}, \ \boldsymbol{\delta}_{r,s}(k,l) = \boldsymbol{\delta}_r(k) \otimes \boldsymbol{\delta}_s(l), \tag{14}$$

$$\mathbf{Z}_{\boldsymbol{\Delta}_i} = \left[\left(\left((\boldsymbol{\delta}_{n-k,m}(1, h_{i-1}+1))'(\mathbf{B}^\perp \otimes \mathbf{I}_m)'\right) \otimes \mathbf{I}_p\right)\mathbf{z} \quad \cdots \quad \left(\left((\boldsymbol{\delta}_{n-k,m}(n-k, h_{i|l}))'(\mathbf{B}^\perp \otimes \mathbf{I}_m)'\right) \otimes \mathbf{I}_p\right)\mathbf{z}\right], \tag{15}$$

it is then possible to rewrite the log-likelihood as

$$\begin{aligned} -2\ell = {} & nmp\log(2\pi) + \operatorname{tr}\big((\mathbf{z}_0 - \boldsymbol{\tau}_0)'\boldsymbol{\Delta}^{-1}(\mathbf{z}_0 - \boldsymbol{\tau}_0)\big) \\ & + \operatorname{tr}\Big(\Big((\mathbf{B}' \otimes \mathbf{I}_{mp})\mathbf{z}_F - \big((\mathbf{B}'\mathbf{X}_F) \otimes \mathbf{I}_{mp}\big)\boldsymbol{\tau}_F\Big)'(\mathbf{I}_{k-1} \otimes \boldsymbol{\Delta}^{-1})\Big((\mathbf{B}' \otimes \mathbf{I}_{mp})\mathbf{z}_F - \big((\mathbf{B}'\mathbf{X}_F) \otimes \mathbf{I}_{mp}\big)\boldsymbol{\tau}_F\Big)\Big) \\ & + n\sum_{i=0}^{2} g_i \log\big(|\boldsymbol{\Delta}_i|\big) + \operatorname{tr}(\boldsymbol{\Delta}_i^{-1}\mathbf{Z}_{\boldsymbol{\Delta}_i}\mathbf{Z}_{\boldsymbol{\Delta}_i}'). \end{aligned} \tag{16}$$

Hence, the parameters that minimize -2ℓ are

$$\hat{\boldsymbol{\tau}}_0 = \mathbf{z}_0, \hat{\boldsymbol{\tau}} = \Big(\big((\mathbf{X}_F'\mathbf{X}_F)^{-1}\mathbf{X}_F\big) \otimes \mathbf{I}_m\Big)\mathbf{z}_F, \hat{\boldsymbol{\Delta}}_i^{(1)} = \frac{1}{ng_i}\mathbf{Z}_{\boldsymbol{\Delta}_i}\mathbf{Z}_{\boldsymbol{\Delta}_i}', \tag{17}$$

taking the value $\ell_1 = -\frac{nmp}{2}\big(\log(2\pi)+1\big) - \sum_{i=0}^{2}\frac{ng_i}{2}\log\big(|\hat{\boldsymbol{\Delta}}_i^{(1)}|\big)$.

As for distributions of these statistics, we can ascertain that

$$\mathbf{z}_0 \sim \mathrm{N}(\boldsymbol{\tau}_0, \boldsymbol{\Delta}), \tag{18}$$

$$\Big(\big((\mathbf{X}_F'\mathbf{X}_F)^{-1}\mathbf{X}_F\big) \otimes \mathbf{I}_m\Big)\mathbf{z}_F \sim \mathrm{N}(\boldsymbol{\tau}, \mathbf{I}_{k-1} \otimes \boldsymbol{\Delta}), \tag{19}$$

$$\mathbf{Z}_{\boldsymbol{\Delta}_i}\mathbf{Z}_{\boldsymbol{\Delta}_i}' \sim \mathrm{W}\big((n-k)g_i, \boldsymbol{\Delta}_l\big), \tag{20}$$

all of these statistics being independent.

2 Unstructured Mean Vector

Consider an observations vector of size mp with distribution $\mathbf{Y}_j \sim \mathrm{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma})$, $j = 1, \ldots, n$, where $\boldsymbol{\mu}_0 = \begin{bmatrix}\mu_{0,1} & \ldots & \mu_{0,mp}\end{bmatrix}'$. The joint distribution of the model will be $\mathbf{Y} \sim \mathrm{N}(\mathbf{1}_n \otimes \boldsymbol{\mu}_0, \mathbf{I}_n \otimes \boldsymbol{\Sigma})$.

 ISBN: 978-84-617-8694-7

Taking

$$\mathbf{Z}_{\mathbf{\Delta}_i,F} = \left[\left(\left((\delta_{n-k,m}(1,h_{i-1}+1))'(\mathbf{B}^{\perp}\otimes \mathbf{I}_m)'\right)\otimes \mathbf{I}_p\right)\mathbf{z} \quad \cdots \quad \left(\left((\delta_{n-k,m}(k-1,h_i))'(\mathbf{B}^{\perp}\otimes \mathbf{I}_m)'\right)\otimes \mathbf{I}_p\right)\mathbf{z}\right], \quad (21)$$

we get

$$-2\ell = nmp\log(2\pi) + \mathrm{tr}\big((\mathbf{z}_0-\boldsymbol{\tau}_0)'\mathbf{\Delta}^{-1}(\mathbf{z}_0-\boldsymbol{\tau}_0) + \sum_{i=0}^{2} ng_i \log\left(|\mathbf{\Delta}_i|\right) + \mathrm{tr}\big(\mathbf{\Delta}_i^{-1}(\mathbf{Z}_{\mathbf{\Delta}_i}\mathbf{Z}'_{\mathbf{\Delta}_i} + \mathbf{Z}_{\mathbf{\Delta}_i,F}\mathbf{Z}'_{\mathbf{\Delta}_i,F})\big). \quad (22)$$

The parameters that minimize the above function are

$$\hat{\boldsymbol{\tau}}_0 = \mathbf{z}_0,\ \hat{\mathbf{\Delta}}_i^{(0)} = \frac{1}{ng_i}(\mathbf{Z}_{\mathbf{\Delta}_i}\mathbf{Z}'_{\mathbf{\Delta}_i} + \mathbf{Z}_{\mathbf{\Delta}_i,F}\mathbf{Z}'_{\mathbf{\Delta}_i,F}), \quad (23)$$

with the log-likelihood taking the value $\ell_0 = -\frac{nmp}{2}\big((\log(2\pi)+1\big) - \sum_{i=0}^{2}\frac{ng_i}{2}\log\left(|\hat{\mathbf{\Delta}}_i^{(0)}|\right)$. We also have that

$$\mathbf{z}_0 \sim \mathrm{N}(\boldsymbol{\tau}_0, \mathbf{\Delta}),\ \mathbf{Z}_{\mathbf{\Delta}_i}\mathbf{Z}'_{\mathbf{\Delta}_i} \sim \mathrm{W}\big((n-k)g_i, \mathbf{\Delta}_i\big),\ \mathbf{Z}_{\mathbf{\Delta}_i,F}\mathbf{Z}'_{\mathbf{\Delta}_i,F} \sim \mathrm{W}\big((k-1)g_i, \mathbf{\Delta}_i\big), \quad (24)$$

all independent.

3 Likelihood Ratio Test

In order to test the hypothesis $H_0 : \boldsymbol{\mu}_q = \mathbf{0}_{mp}, q = 1,\ldots,k$ vs. $H_1 : \exists q, \boldsymbol{\mu}_q \neq \mathbf{0}_{mp}$, we consider the likelihood ratio test statistic $T = \left(\prod_{i=0}^{2}\frac{|\mathbf{W}_i|}{|\mathbf{W}_i+\mathbf{W}_{i,F}|}\right)^{-\frac{1}{2}}$, with $\mathbf{W}_i = \mathbf{Z}_{\mathbf{\Delta}_i}\mathbf{Z}'_{\mathbf{\Delta}_i}$ and $\mathbf{W}_{i,F} = \mathbf{Z}_{\mathbf{\Delta}_i,F}\mathbf{Z}'_{\mathbf{\Delta}_i,F}$.

Acknowledgements

Funded by UID/MAT/00297/2013.

References

[1] T. W. Anderson, *An Introduction to Multivariate Statistical Analysis*, Wiley, 2003.

[2] T. Kollo and D. von Rosen, *Advanced Multivariate Statistics with Matrices*, Springer, 2010.

[3] A. Roy, R. Zmyślony, M. Fonseca and R. Leiva, *Optimal estimation for doubly multivariate data in blocked compound symmetric covariance structure*, Journal of Multivariate Analysis, **144** (2016) 81–90.

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

On invariant manifolds of saddle points for 3D multistable models

Elisa Francomano[1] **and Marta Paliaga**[1]

[1] *Scuola Politecnica, DIID, Universita' degli Studi di Palermo*

emails: elisa.francomano@unipa.it, marta.paliaga@unipa.it

Abstract

In dynamical systems a particular solution is completely determined by the parameters considered and the initial conditions. Indeed, when the model shows a multistability, starting from different initial state, the trajectories can evolve towards different attractors. The invariant manifolds of the saddle points separate the vector field into the basins of attraction of different stable equilibria.
The aim of this work is the reconstruction of these separation surfaces in order to know in advance the geometry of the basins. In this paper three-dimensional models with three or more stable fixed points is investigated. To this purpose a procedure for the detection of the scattered data lying on the manifolds is proposed. Then a Moving Least Squares meshfree method is involved to approximate the surfaces. Numerical results are presented in order to assess the method.

Key words: Dynamical systems, Invariant manifolds, Separatrix, Meshfree method, Moving Least Squares.

1 Introduction

Mathematical modeling and experimental investigations in biomedical sciences are popular instruments to explain the biological or physics process developed from a particular observed phenomenon. The systems of ordinary differential equations are analyzed in order to predict the temporal evolutions of the involved variables.
For a designated set of parameters, the solutions of the models evolve towards a set of numerical values called "*attractor*" [10]. Althought persistent oscillations or chaotic behaviors can happen, in this work we focus only on fixed stable points. When the model presents

more than one stable equilibrium (multistability), the analysis of the basins of attraction is fundamental for a detailed knowledge of the different evolutions of the trajectories. In fact, nearby the boundaries of the basins, initial points could indeed lead to completely different system's outcomes [2], [3].
To this purpose we analyze the invariant manifolds of the saddle points, which represent the separatrices for the vector field.
Extending our preliminary results [6], we proceed on the reconstruction of the basins boundaries for 3D models with three or more equilibria concurrently stable. In order to find the points lying on the invariant manifolds we operate by employing the eigenspace generated by the saddle node's eigenvectors.
Therefore we reconnect the manifolds points by using the Moving Least Squares meshless method presented in [7],[8]. Finally the study of a tristable epidemiological model is proposed [5].

2 Invariant manifolds reconstruction

In order to approximate the basins of attraction, the general idea is finding the points of the invariant manifolds of the saddle points.
We focus our attention on three-dimensional dynamical models where a saddle node is a fixed point which admits three eigenvalues such that at least one of them has real part positive and at least one has real part negative.
To integrate the separatrix surface of a saddle E_s, we consider M equispaced points on an ellipse centered on E_s whose semi-axes are the eigenvectors corresponding to the eigenvalues with the same sign. Therefore the ellipse lies on the eigenspace V that is tangent to the invariant manifold W of the saddle . Furthermore we backward integrate the model starting from the M seeding points obtaining the separatrix points ([11], [12],[9]) that we reconnect by using the Moving Least Squares method.
We start projecting the scattered data on the plane xy: $X = \{\mathbf{x}_i, i = 1, ..., N\} \in \mathbb{R}^2$ and considering z_i as the corresponding height. To approximate the manifolds we have to construct the quasi-interpolant [1]:

$$P(\mathbf{y}) = \sum_{i=1}^{N} z_i \Phi_i(\mathbf{y}) \tag{1}$$

where $\Phi_i(\mathbf{y}) = \Phi(x_i, \mathbf{y})$ are the *generating functions*.
Let $Q = span\{p_1, ..., p_m\}$ with $m < N$ the approximation space with $p_m \in \prod_2^s$, the bivariate polynomial of degree at most d.

 ISBN: 978-84-617-8694-7

The generating functions are [4]:

$$\Phi_i(\mathbf{y}) = \omega(\mathbf{x}_i, \mathbf{y}) \sum_{j=1}^{m} \lambda_j p_j(\mathbf{x}_i),\ i = 1, ...N \tag{2}$$

where ω represent the weighted functions governing the influence of the scattered data X and λ_j are the Lagrangian multipliers that solve the system:

$$\sum_{i=1}^{N} p_k(\mathbf{x}_i) p_l(\mathbf{x}_i) \omega(\mathbf{x}_i, \mathbf{y}) \lambda(\mathbf{y}) = p(\mathbf{y})\ k, l = 1, ..., m. \tag{3}$$

This meshfree approximant is very attractive especially when one is only interested in a few evaluations. Indeed it is not necessary to set up and solve a large interpolation system [13]. Furthermore when $m < 3$ it is possible find the explicit formulas for the Lagragian multipliers avoiding the resolution of the system (3).

3 Tristable model analysis

In order to test the algorithm presented in the previous section, we consider the following model [5]:

$$\frac{dS}{dt} = S\left[(S-\theta)(1-S-I) - \beta I - aP\right], \tag{4}$$

$$\frac{dI}{dt} = \beta SI - aIP - \mu I, \tag{5}$$

$$\frac{dP}{dt} = P\left[bs + \alpha I - d\right] \tag{6}$$

that analyzes a predator-prey interaction with prey subjected to Allee effect and disease. Therefore these latter are divided into susceptible (S) and infected (I) individual and P represents the predators. We resume all the parameters on the following table:

Parameter	Biological Meaning
θ	Allee threshold
β	Infection Rate
a	Attack rate of predator
b	Total effect to predator by consuming susceptible prey
μ	Death rate of infected prey
α	Total effect to predator by consuming infected prey
d	Natural death rate of predator.

 ISBN: 978-84-617-8694-7

Letting $\beta = 1.5$, $\theta = 0.2$, $a = 2$, $b = 1.35$, $\mu = 1$ and $d = 1$ the system admits three stable equilibria: the origin $E_0 \equiv (0, 0, 0)$, the disappearance of the disease $E_1 \approx (0.7407, 0, 0.0701)$ and the predator extinction $E_2 \approx (0.6667, 0.0791, 0)$.
For the reconstruction of their basins of attraction we consider the invariant manifolds of the attractive saddle points $E_{s_1} \equiv (\theta, 0, 0)$ and $E_{s_2} \approx (0.7329, 0.0211, 0.0497)$.
We start with the first saddle that admits the stable eigenvectors $v_1 \approx (0.4099, 0, 0.9121)$ and $v_2 \approx (0.329, 0.9442, 0)$. We integrate the separatrix considering $M = 20$ equispaced on the ellipse generated by v_1 and v_2 rescaled by a factor 0.7. Then, throught a backward integration we obtain all the scattered data on the manifold (Figure 1A).
Applying the same procedure to the saddle E_{s_2} we generate the separatrix points on the second manifold by rescaling the stable complex conjugated eigenvectors $v_{1,2} \approx (0.9817, -0.0088 \pm 0.08731i, -0.0552 \pm 0.1599i)$ with a factor 0.6 (Figure 1B).
Finally we approximate the two surfaces (Figure 1C) applyng the MLS approximant using the Wendland C2 compactly supported function:

$$\omega(\mathbf{x}_i, \mathbf{y}) = (1 - \epsilon||\mathbf{y} - \mathbf{x}_i||_2)_+^4 (4\epsilon||\mathbf{y} - \mathbf{x}_i||_2 + 1) \tag{7}$$

where the *shape parameter* $\epsilon = 3$ and 60 evaluation points $\mathbf{y}$ are taking into account (Figure 1D).

4 Conclusions

In this work we have presented a method for the reconstruction of the invariant manifolds of the saddle points. These surfaces play an important role on the analysis of the vector field because they partition it into the basins of attraction of the different fixed stable equilibria. Therefore, studying a dynamical model, when the parameters involved are chosen, one can understand a priori the possible evolution of the initial state. In this paper we have investigated the basins of attraction of a tristable model, but also systems with more stable equilibria can be treated.

Acknowledgements

The research has been supported by the Istituto Nazionale di Alta Matematica - INDAM - GNCS Project 2017 and it has been accomplished within the RITA "Research Italian network on Approximation".

 ISBN: 978-84-617-8694-7

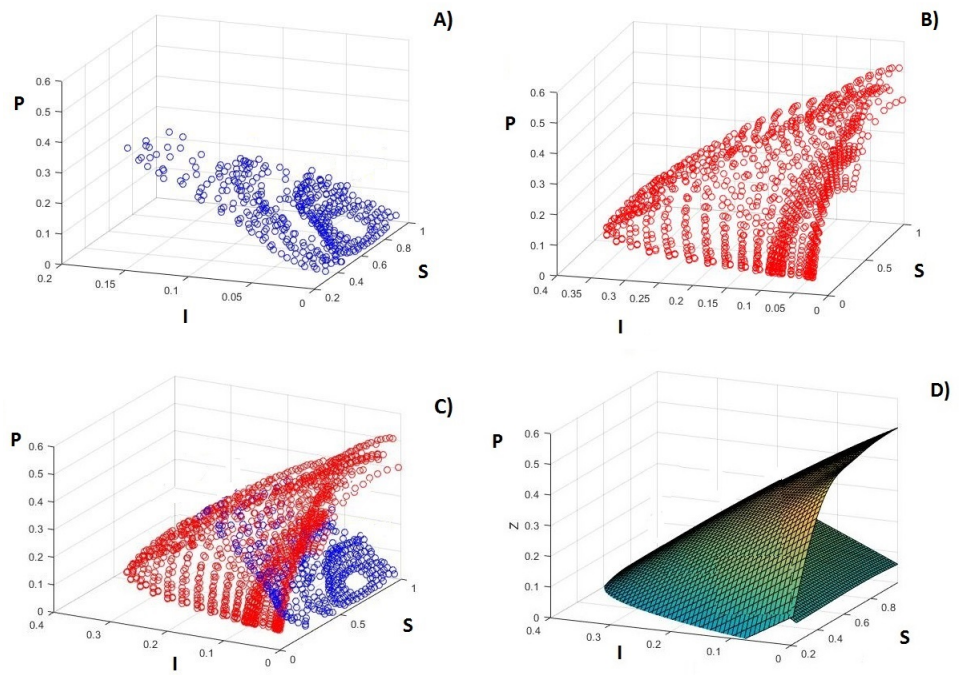


Figure 1: A) Scattered data on invariant manifold of saddle E_{s_1}; B) Scattered data on invariant manifold of saddle E_{s_2}; C) Intersection of the two separatricies; D) Reconstruction of the two surfaces.

References

[1] G. Backus and F. Gilbert, The resolving power of gross earth data, Geophys. J. R. Astr. Soc., Vol. 16, pp. 169–205 (1968).

[2] R. Cavoretto, A. De Rossi, E. Perracchione and E. Venturino, Graphical Representation of Separatrices of Attraction Basins in Two and Three-Dimensional Dynamical Systems, International Journal of Computational Methods, Vol. 14, Issue 1 (2016).

[3] R. Cavoretto, A. De Rossi, E. Perracchione and E. Venturino, Robust Approximation Algorithms for the Detection of Attraction Basins in Dynamical Systems, Journal of Scientific Computing, Vol. 68, pp. 395–415 (2016).

[4] G. Fassahuer, Meshfree Approximation Methods with MATLAB, World Scientific Publishing Co., Singapore, pp. 190–205 (2007).

[5] Y. Kang, S.K. Sasmal, A.R. Bhowmick and J. Chattopadhyay, Dynamics of a predator-prey system with prey subject to Allee effects and disease, Mathematical Biosciences and Engineering, Vol. 11, Issue 4, pp. 877–918 (2014).

[6] E. Francomano, F.M. Hilker, M. Paliaga and E. Venturino, An efficient method to reconstruct invariant manifolds of saddle points, Dolomites Research Notes on Approximation, in press (2017).

[7] E. Francomano, F.M. Hilker, M. Paliaga and E. Venturino, On Basins of Attraction for a Predator-Prey Model Via Meshless Approximation, AIP Conference Proceedings 1776, NUMTA (2016).

[8] E. Francomano, F.M. Hilker, M. Paliaga and E. Venturino, Separatrix reconstruction studying the Allee effect in a predator-prey model, Applied Mathematics and Computation, submitted (2016).

[9] G. Moore, Laguerre approximation of stable manifolds with application to connecting orbits, Mathematics and Computation, Vol. 73, pp. 211-242 (2003).

[10] J.D. Murray, Mathematical Biology I: An Introduction, third edition, Springer, New York (2002).

[11] R. Precup, M.A. Serban and D. Trif, Asymptotic stability for a model of cell dynamics after allogenic bone marrow transplation, Nonlinear Dynamics and Systems Theory, Vol. 13, Issue 1, pp. 79–92 (2013).

[12] H. Theisel, T. Weinkauf, H.C. Hege and H.P. Seidel, Saddle Connectors- an approach to visualizing the topological skeleton of complex 3D vector fields, Visualization, 2003, IEEE (2003).

[13] H. Wendland, Scattered Data Approximation, Cambridge University Press, Cambridge, pp. 35–43 (2010).

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

Nonparametric wavelet-based estimation from strongly spatially correlated data

María Pilar Frías[1] and María Dolores Ruiz-Medina[2]

[1] *Department of Statistics and O.R., University of Jaén, Spain*

[2] *Department of Statistics and O.R., University of Granada, Spain*

emails: mpfrias@ujaen.es, mruiz@ugr.es

Abstract

Wavelet-based estimation methodologies are considered for spatial prediction from strong correlated high-dimensional date. Different approaches are proposed for spatial estimation of annual mean ocean surface temperature maps. Specifically, spatial wavelet kernel penalized nonparametric regression and wavelet shrinkage are applied on daily ocean surface temperature curves from the Hawaii ocean stations, available at latitude-longitude interval $[22.7, 22.8] \times [-158.1, -157.94]$, during the period 2000 – 2007, from *The World-Wide Ocean Optics Database (WOOD).*

Key words: Wavelet-based estimation methodology, spatial models, ocean surface temperatures maps, functional data.

1 Introduction

Nowadays, prediction of future climate change based on statistical models play a key role in climate sciences. Alterations in sea-surface temperatures may lead to changes in atmospheric circulation and precipitation, causing more hurricanes and drought, particularly in the tropics. Statistical techniques for high- dimensional data can be applied to analyze long records of observations available at different stations of ocean islands. In this context, when only sample information from long records is incorporated in the estimation procedure, the most important issue is the application of estimation procedures adequate to the nature of the data, e.g., spatial correlation structure and spatial regularity through time, among others characteristics.

ISBN: 978-84-617-8694-7

The present paper provides an extended formulation of the wavelet kernel penalized univariate nonparametric regression approach studied in [1] to the bivariate spatial design point case, applied to ocean surface temperature estimation. Wavelet shrinkage is also considered here for spatial estimation of ocean surface temperature (see, for example, [2] and [3]). The results obtained are compared with those ones derived, under the spatial autoregressive Hilbertian framework (SARH(1) framework), considered in [6], [7], [8] and [9], and under the semiparametric approach presented in [10], based on Gaussian linear processes with heavy tail covariance functions.

2 Non-linear estimation methods

2.1 Wavelet kernel penalized nonparametric regression

Let us consider the following spatial nonparametric regression model:

$$Y_i = f(x_1^i, x_2^i) + \sigma\varepsilon(x_1^i, x_2^i), \quad i = 1, \ldots, N^2,$$

where f is a deterministic function to be estimated or approximated, and $\varepsilon(x_1^i, x_2^i)$, $(x_1^i, x_2^i) \in I \subset \mathbb{R}^2$, $i = 1, \ldots, N^2$, are independent and identically distributed standard Gaussian random variables. Here, parameter σ^2 is considered unknown, and estimated from the data. Furthermore, the following model is assumed to be satisfied by function f :

$$f(x_1, x_2) = \mu(x_1, x_2) + \sqrt{b}z(x_1, x_2),$$

where μ is a spatial deterministic function, and z is a spatial Gaussian process, both defined on $I = [a, b]\times[c, d] \subset \mathbb{R}^2$, with $[0, 1]\times[0, 1] \subseteq I = [a, b]\times[c, d]$. Function μ and spatial process z respectively admit the following orthogonal series representations: For all $(x_1, x_2) \in [0, 1]\times[0, 1]$,

$$\begin{aligned} \mu(x_1, x_2) &= \sum_{k_1=0}^{2^J-1}\sum_{k_2=0}^{2^J-1} \alpha_{J,k_1,k_2}\phi_{J,k_1,k_2}(x_1, x_2) \\ \sqrt{b}z(x_1, x_2) &= \sum_{s=h,v,d}\sum_{j\geq J}\sum_{k_1=0}^{2^j-1}\sum_{k_2=0}^{2^j-1} \beta^s_{j,k_1,k_2}\psi^s_{j,k_1,k_2}(x_1, x_2), \end{aligned}$$

where β^s_{j,k_1,k_2} are independent random variables satisfying $\beta^s_{j,k_1,k_2} \sim N(0, \lambda_j)$, $k_1 = 0, \ldots 2^j - 1,\ k_2 = 0, \ldots 2^j - 1,\ j \geq J,\ s = h, v, d$. Here, $\{\phi_{J,k_1,k_2},\ k_1 = 0, \ldots 2^J - 1,\ k_2 = 0, \ldots 2^J - 1\}$ denotes the compactly supported scaling function basis on the interval $[0, 1] \times [0, 1]$, generating the space V_J, and $\{\psi^s_{j,k_1,k_2},\ k_1 = 0, \ldots 2^j-1,\ k_2 = 0, \ldots 2^j-1,\ j \geq J,\ s = h, v, d\}$ are compactly supported wavelet bases, respectively generating the spaces $\{W_j,\ j \geq J\}$.

 ISBN: 978-84-617-8694-7

The predictor $\hat{f}$ can be expressed in the form,

$$\hat{f}(\mathbf{x}) = \Phi_{\mathbf{x}}\hat{\alpha} + b^{1/2}\hat{z}(\mathbf{x}), \quad \mathbf{x} \in [0,1] \times [0,1],$$

where

$$\hat{\alpha} = (\Phi' M^{-1}\Phi)^{-1}\Phi' M^{-1}\mathbf{Y}$$

is the least square weighted predictor for the model

$$Y(\mathbf{x}) = \Phi_{\mathbf{x}}\alpha + \varepsilon'(\mathbf{x}), \quad \mathbf{x} \in [0,1] \times [0,1],$$

with $\varepsilon'(\mathbf{x}) = b^{1/2}z(\mathbf{x}) + \sigma\varepsilon(\mathbf{x}), \quad \mathbf{x} \in [0,1] \times [0,1]$, and

$$b^{1/2}\hat{z}(\mathbf{x}) = \Sigma_{\mathbf{x}} M^{-1}(I - \Phi_{\mathbf{x}}(\Phi' M^{-1}\Phi)^{-1}\Phi' M^{-1})\mathbf{Y}, \quad \mathbf{x} \in [0,1] \times [0,1],$$

is the plug-in predictor of the centered Gaussian random effect, based on the least square weighted estimation of the fixed effect parameter vector α. Φ is the $n \times 2^{2J}$, with $\Phi_{i,\mathbf{k}} = \phi^2_{J,\mathbf{k}}(\mathbf{x}^i) = \phi^2_{J,\mathbf{k}}(x^i_1, x^i_2) = \phi_{J,k_1}(x^i_1)\phi_{J,k_2}(x^i_2)$, for $i = 1, \ldots, n$, and $\mathbf{k} = (k_1, k_2)$, $k_1 = 0, \ldots, 2^J - 1$, $k_2 = 0, \ldots, 2^J - 1$, and, for each $\mathbf{x} = (x_1, x_2) \in [0,1] \times [0,1]$, $\Phi_{\mathbf{x}}$ is then the 1×2^{2J} row matrix, with entries $(\Phi_{1,\mathbf{k}} = \phi^2_{J,\mathbf{k}}(\mathbf{x}))_{\mathbf{k}=(k_1,k_2),k_1=0,\ldots,2^J-1,k_2=0,\ldots,2^J-1}$. For each $\mathbf{x} = (x_1, x_2) \in [0,1]\times[0,1]$, matrix $\Sigma_{\mathbf{x}}$ is the $1\times n$ row matrix with entries $(K^1(\mathbf{x}, \mathbf{x^t}))_{t=1,\ldots,n}$, where,

$$\begin{aligned} K^1(\mathbf{x^r}, \mathbf{x^t}) &= \sum_{s=h,v,d} \sum_{j\geq J} \sum_{k_1=0}^{2^j-1} \sum_{k_2=0}^{2^j-1} \lambda_j \psi^s_{j,\mathbf{k}}(\mathbf{x_r})\psi^s_{j,\mathbf{k}}(\mathbf{x_t}) \\ &= \sum_{j\geq J} \sum_{k_1=0}^{2^j-1} \sum_{k_2=0}^{2^j-1} \lambda_j \phi_{j,k_1}(x^r_1)\psi_{j,k_2}(x^r_2)\phi_{j,k_1}(x^t_1)\psi_{j,k_2}(x^t_2) \\ &\quad +\lambda_j \psi_{j,k_1}(x^r_1)\phi_{j,k_2}(x^r_2)\psi_{j,k_1}(x^t_1)\phi_{j,k_2}(x^t_2) \\ &\quad +\lambda_j \psi_{j,k_1}(x^r_1)\psi_{j,k_2}(x^r_2)\psi_{j,k_1}(x^t_1)\psi_{j,k_2}(x^t_2). \end{aligned}$$

Finally, $M = \Sigma + (\sigma^2/b)I_n$, with Σ the $n \times n$ matrix, with entries $(\Sigma_{\mathbf{r},\mathbf{t}})_{r=1,\ldots,n,\ t=1,\ldots,n} = (K^1(\mathbf{x^r}, \mathbf{x^t}))_{r=1,\ldots,n,\ t=1,\ldots,n}$, with $\mathbf{x^r} = (x^r_1, x^r_2)$, $r = 1, \ldots, n$, and $\mathbf{x^t} = (x^t_1, x^t_2)$, $t = 1, \ldots, n$, (see, [4]).

2.2 Spatial shrinkage and thresholding

In this section, we briefly introduce an alternative wavelet-based nonparametric regression framework in space, based on the application of spatial non-linear *shrinkage* rules, that allows to eliminate the additive Gaussian white noise ε, in the following equation:

$$Y_i = f(x^i_1, x^i_2) + \sigma\varepsilon(x^i_1, x^i_2), \quad i = 1, \ldots, n. \tag{1}$$

 ISBN: 978-84-617-8694-7

Specifically, the shrinkage estimator $\hat{f}$ of function f minimizes the empirical mean-square error

$$E(\hat{f}) = \frac{1}{n}\sum_{i=1}^{n} E\left(\hat{f}(x_1^i, x_2^i) - f(x_1^i, x_2^i)\right)^2.$$

In the next section, spatial *wavelet thresholding* will be applied (see, for example, [2]). The following steps will then be implemented for the computation of the spatial wavelet thresholding estimators of function f in equation (1):

Step 1 The two-dimensional discrete wavelet transform is computed.

Step 2 A nonlinear *shrinkage* rule is applied. In particular, hard-thresholding or soft-thresholding could be applied (see, for example, [11]). We can also distinguish between *global thresholding* and *level dependent thresholding*.

Step 3 The corresponding shrinkage estimator is obtained from the inverse two-dimensional discrete wavelet transform, given by

$$\begin{aligned}\hat{f}(x_1^i, x_2^i) &= \sum_{k_1=0}^{2^{J_0}-1}\sum_{k_2=0}^{2^{J_0}-1} \hat{\alpha}_{J_0,\mathbf{k}}\phi_{J_0,\mathbf{k}}(x_1^i, x_2^i) \\ &+ \sum_{s=h,v,d}\sum_{j=J_0}^{J}\sum_{k_1=0}^{2^j-1}\sum_{k_2=0}^{2^j-1} \hat{\beta}^s_{j,\mathbf{k}}\psi^s_{j,\mathbf{k}}(x_1^i, x_2^i).\end{aligned}$$

3 Estimation of ocean surface temperature of Hawaii ocean

Here, ocean surface temperature is estimated in Hawaii Ocean by applying different wavelet-based methodologies. Data are collected from the oceanographic bio-optical database: *The Worldwide Ocean Optics Database (WOOD)*. Particularly, the area analyzed corresponds to the latitude-longitude interval $[22.7, 22.8] \times [-158.1, -157.94]$, during the period 2000 – 2007. In that area and period, 1564 weather stations are available, approximately between 190 and 292 stations are available per year, in the period studied. Spatial interpolation method based on exponential covariance function is applied to allocating data on a 16×16 spatial regular grid. The influence of the spatial interpolation technique in the wavelet-based methodologies studied here is investigated in [4]. Moreover, they show the presence of strong correlations in space and time in annual mean ocean surface temperature data in the area studied here. Specifically, the spatial dependence structure of the yearly averaged data reflects spatial and temporal correlations of daily data within each year studied. Then, kriging with exponential covariance model is suitable for interpolate yearly averaged ocean surface temperature values because it takes into account the spatial dependence structure. Figure 1 displays boxplots of the absolute error spatial L^∞-norms for the wavelet-based

 ISBN: 978-84-617-8694-7

non parametric regression techniques describes in sections 2.1 and 2.2. The yearly averaged ocean surface temperature maps estimated with the referred estimation methods can be found in Figure 2. Spatial wavelet thresholding methods at coarser resolution levels seems to provide the best estimation results against spatial wavelet kernel penalized regression, where local smoothing is performed. Table 1 shows mean and standard deviation of the absolute error spatial L^∞-norms of the methodologies considered in section 2 compared with two linear model based estimation methodologies implemented in the wavelet domain, previously proposed in the literature. Specifically, in the linear model context, we consider here curve estimation under the SARH(1) model framework (see [5]). The semiparametric estimation methodology, based on fractional-order pseudodiffential models, proposed in [10], is considered as well. The means of L^∞-norms of the absolute error curves, obtained from the application of the selected spatial wavelet thresholding technique, and the semiparametric estimation methodology proposed in [10], are lower than in the case of the wavelet kernel penalized nonparametric regression and SARH(1) prediction. Whether stations in tropical and subtropical ocean islands are very concentrated in space. Hence, the linear or non-linear character of the spatial functional statistical model fitted to the data is not as important as its capability for spatial decorrelation, i.e., for efficient processing of strong spatial correlated functional data, by removing redundant information (see, [4]).

Estimation Method	Mean (standard deviation) of L^∞-norms of the absolute error curves
SARH(1)	0.5911 (0.0994)
SEM	0.192 (0.1142)
SWKPR	0.4792 (0.2829)
SWTHR (db2, 1°)	0.1782 (0.0809)
SWTHR (db2, 2°)	0.4582 (0.2188)
SWTHR (db4, 1°)	0.1051 (0.0463)
SWTHR (db4, 3°)	0.6573 (0.3178)

Table 1: Mean and standard deviation of L^∞-norms of the absolute error curves, applying spatial wavelet level dependent thresholding (SWTHR), spatial wavelet kernel penalized regression (SWKPR), SARH(1) prediction, and semiparametric estimation methodology (SEM).

 ISBN: 978-84-617-8694-7

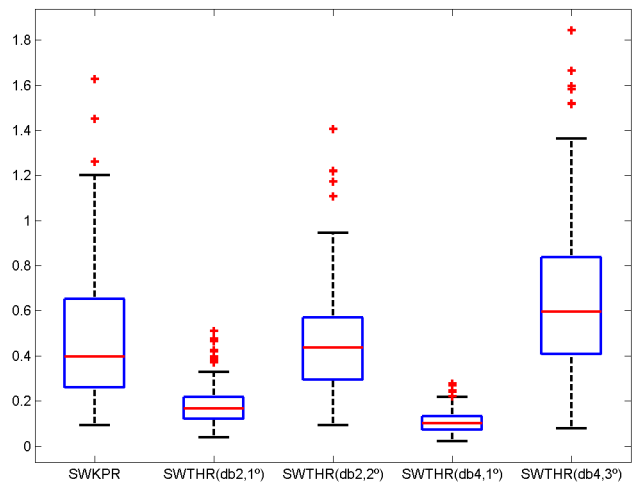


Figure 1: Boxplots of spatial L^{∞}-norms of absolute errors, from the application of Spatial Wavelet Kernel Penalized Regression (SWKPR), Spatial Wavelet Thresholding (SWTHR) in terms of Daubechies 2 at first resolution level (db2, 1°), in terms of Daubechies 2 at second resolution level (db4, 1°), in terms of Daubechies 4 at first resolution level (db4, 1°), and in terms of Daubechies 4 at third resolution level (db4, 3°).

 ISBN: 978-84-617-8694-7

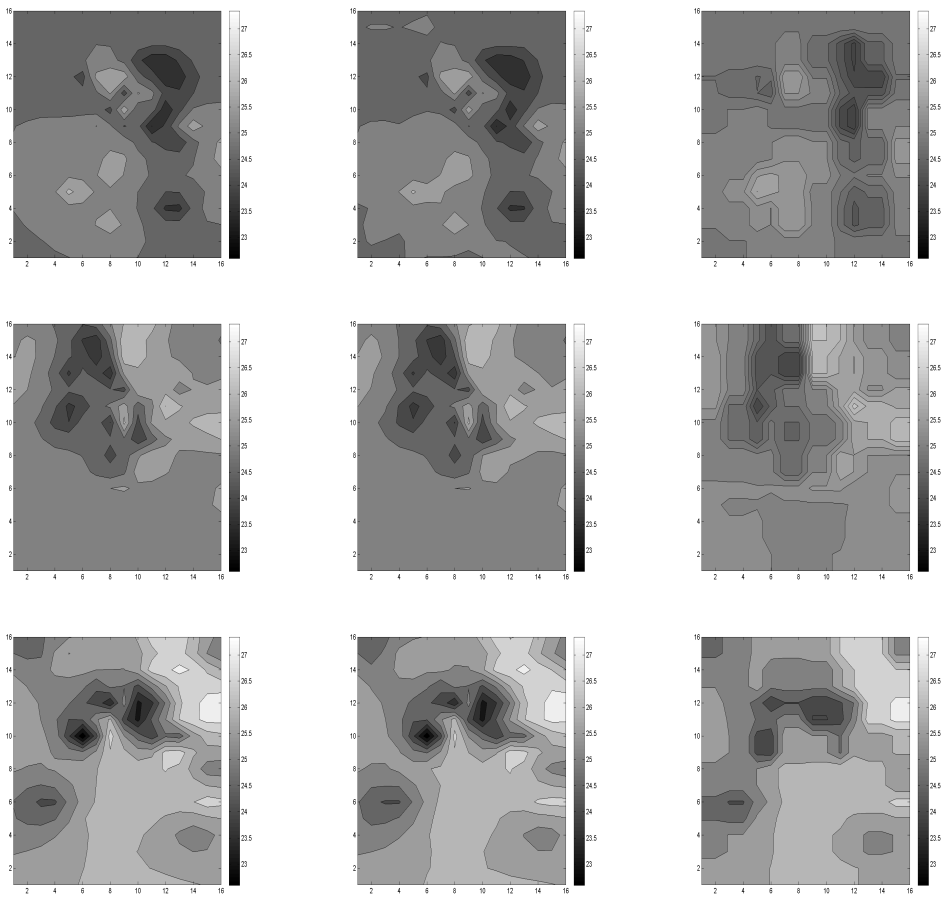


Figure 2: Interpolated (left), and estimated temperature maps applying Spatial Wavelet Kernel Penalized Regression (middle), Spatial Wavelet Thresholding in terms of Daubechies 4 at first resolution level (right), from the top to the bottom for years 2002, 2003 and 2004.

 ISBN: 978-84-617-8694-7

4 Conclusions

In order to find an optimal non-linear estimation of ocean surface temperature in a nonparametric framework, from the discrete wavelet transform of data, we have explained two different methodologies, spatial wavelet thresholding and spatial wavelet kernel penalized regression. The smooth local variation of ocean surface daily temperature in tropical and subtropical regions through the year allows annual spatial prediction of averaged daily temperature, from a coastal station network, covering a very small latitude-longitude interval, when a suitable statistical model is applied. Thus, in that case, the patterns observed in a small area can be reproduced in an extensive area of ocean, in certain subtropical and tropical regions. The results derived here illustrate the fact that, the most important property required on spatial estimation from slow varying strong correlated high-dimensional data is the capability for eliminating redundant information and avoiding over smoothing. While spatial wavelet kernel penalized nonparametric regression methods, like the one derived here, are required, for the analysis of high-dimensional data displaying singular (erratic) spatial patterns over time, to avoid, for example, false positives in risk assessment. Alternatively, a flexible wavelet-based non-linear methodology, allowing the adaptation of scale to the the dependence range of the data, for performing a suitable decorrelation, can be considered to solve the filtering problem. This is the case of spatial wavelet level-dependent thresholding methods considered here, leading to an optimal selection of the most relevant features of the data for estimation, removing redundant information.

Acknowledgements

This work has been supported in part by project MTM2015–71839–P, of the DGI, MINECO, Spain.

References

[1] C. Angelini, D. De Canditiis, and F. Leblanc, *Wavelet regression estimation in nonparametric mixed effect models*, J. Multivariate Anal. **85** (2003) 267–291.

[2] D. L. Donoho and I. M. Johnstone, *Ideal spatial adaptation by wavelet shrinkage*, Biometrika **81** (1994) 425–455.

[3] D. L. Donoho and I. M. Johnstone, *Adapting to unknown smoothness via wavelet shrinkage*, Journal of the American Statistical Association **90** (1995) 1200–1224.

[4] M. P. Frías and M. D. Ruiz-Medina, *Wavelet nonparametric estimation from strong spatial correlated high-dimensional data*, Spatial Statistics **18** (2016) 363–385.

 ISBN: 978-84-617-8694-7

[5] M. D. Ruiz-Medina, *Spatial autoregressive and moving average Hilbertian processes*, Journal of Multivariate Analysis **102** (2011) 292–305.

[6] M. D. Ruiz-Medina, *New challenges in spatial and spatiotemporal functional statistics for high-dimensional data*, Spatial Statistics **1** (2012a) 82–91.

[7] M. D. Ruiz-Medina, *Spatial functional prediction from spatial autoregressive Hilbertian processes*, Environmetrics **23** (2012b) 119–128.

[8] M. D. Ruiz-Medina and R. M. Espejo, *Spatial autoregressive functional plug-in prediction of ocean surface temperature*, Stoch. Environ. Res. Risk Assess. **26** (2012) 335–344.

[9] M. D. Ruiz-Medina and R. M. Espejo, *Integration of spatial functional interaction in the extrapolation of ocean surface temperature anomalies due to global warming*, Int. J. Appl. Earth. Obs. **22** (2013) 27–39.

[10] M. D. Ruiz-Medina and M. P. Frías, *Wavelet-based semiparametric estimation of ocean surface temperature*, Math. Geosci. **47** (2015) 149–171.

[11] B. Vidakovic, *Nonlinear wavelet shrinkage with Bayes rules and Bayes factors*, Journal of the American Statistical Association **93** (1998) 173–179.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

Existence and multiplicity results for systems of first order differential equations

Marlène Frigon[1]

[1] *Department of Mathematics and Statistic, University of Montréal*

emails: frigon@dms.umontreal.ca

Abstract

We present existence and multiplicity results for systems of first order differential equations. To this aim, we introduce the method of solution-regions. It generalizes the method of upper and lower solutions and the method of solution-tubes. Our results can also be seen as viabily results since we obtain solutions remaining in suitable regions. We give conditions insuring the existence of at least three viable solutions of a system of first order differential equations.

Key words: System of differential equations; multiplicity results; viability results
MSC 2000: AMS codes (optional)

References

[1] M. Frigon, *Existence and multiplicity results for systems of first order differential equations via the method of solution-regions*, (to appear).

[2] M. Frigon, M. Lotfipour, *Multiplicity results for systems of first order differential inclusions*. J. Nonlinear Convex Anal. **16** (2015), 1025–1040.

[3] J.R. Graef, L. Kong, *Existence of multiple periodic solutions for first order functional differential equations*. Math. Comput. Modelling **54** (2011), 2962–2968.

[4] J. Mawhin, *First order ordinary differential equations with several periodic solutions*. Z. Angew. Math. Phys. **38** (1987), 257–265.

[5] S. Plaskacz, *Periodic solutions of differential inclusions on compact subsets of* $\mathbb{R}^n$. J. Math. Anal. Appl. **148** (1990), 202–212.

 ISBN: 978-84-617-8694-7

[6] R.L. Pouso, *Nonordered discontinuous upper and lower solutions for first-order ordinary differential equations*. Nonlinear Anal. Ser. A: Theory Methods, **45** (2001), 391–406.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

Equivalence transformations and symmetry analysis for a generalized Fisher equation

M.L Gandarias[1], M. Rosa[1] and R. Tracinà[2]

[1] *Departamento de Matemáticas, University of Cádiz*

[2] *Department of Mathematics and Computer Sciences, University of Catania, Italy*

emails: marialuz.gandarias@uca.es, maria.rosa@uca.es, tracina@dmi.unict.it

Abstract

Equivalence transformations and Lie symmetries are found for a generalized Fisher equation.

1 Introduction

In recent years, interactions between the mathematical and biological sciences have been increasing rapidly. Consequently, it has stimulated developments in the theory of non-linear differential equations with a variety of challenging problems. Many physicists and mathematicians have paid much attention to the Fisher equations due to their importance in mathematical physics. The equation analyzed in this work is a generalized Fisher equation with variable coefficients

$$u_t = f(u) + \frac{1}{c(x)} \left(c(x)g(u)u_x\right)_x , \tag{1}$$

where $u(x,t)$ denote the density of tumor cells, g is the diffusion coefficient depending on the variable u, being x and t the independent variables, $f(u)$ an arbitrary function and $c(x)$ an arbitrary function depending on the space variable x. For special cases, this equation has been studied by other authors using symmetry analysis and equivalence transformations [1, 3, 8] has attracted considerable interest in studies of tumor growth and their applications [4, 7, 5, 2]. For a particular case of equation (1), a complete classification of the classical symmetries and exact solutions of was obtained in [6]. In this paper, a generalized Fisher equation is studied from the point of view of the theory of symmetry reductions in partial differential equations.

 ISBN: 978-84-617-8694-7

2 Lie symmetries

We perform the Lie group classification for equation (1) in the case $fg_uc' \neq 0$. That is we classify all the Lie symmetries depending on the form of the arbitrary elements (functions $f(u)$, $g(u)$, and $c(x)$). For the sake of simplicity, in the following analysis we prefer to introduce the new function $\alpha(x) = \frac{c'(x)}{c(x)}$ In this way equation (1) can be written as

$$u_t = f(u) + \alpha g u_x + g_u u_x^2 + g u_{xx}. \tag{2}$$

We look for infinitesimal generator of the equivalence transformations of equation (2) of the form

$$\mathbf{Y} = \Xi^1\partial_t + \Xi^2\partial_x + \phi\partial_u + \mu^1\partial_f + \mu^2\partial_\alpha + \mu^3\partial_g \tag{3}$$

where the infinitesimal components Ξ^1, Ξ^2, and ϕ, are depending on t, x, and u, while the infinitesimal components μ^i, $(i = 1, 2, 3)$ can also depend on f, g, and α. We find that the class (2) admits a continuous group of equivalence transformations generated by the following operators

$$\begin{aligned} &\mathbf{Y_1} = \partial_t, \quad \mathbf{Y_2} = \partial_x, \quad \mathbf{Y_3} = \partial_u, \\ &\mathbf{Y_4} = t\partial_t - f\partial_f - g\partial_g, \quad \mathbf{Y_5} = x\partial_x - \alpha\partial_\alpha + 2g\partial_g, \quad \mathbf{Y_6} = u\partial_u + f\partial_f. \end{aligned} \tag{4}$$

To apply the classical method to equation (2), one looks for infinitesimal generators of the form

$$V = \xi(x,t,u)\partial_x + \eta(x,t,u)\partial_t + \psi(x,t,u)\partial_u,$$

that leave invariant this equation. For f, g and α arbitrary the only symmetry generator admitted by (2) is $\mathbf{v}_1 = \partial_t$. Moreover whenever the function $\alpha(x)$ is constant, the equations (2) admit also the symmetry generator $\mathbf{v}_2 = \partial_x$.
Now, we look for the functional forms of f, g and α which yield extra symmetry generators and we distinguish the following four cases depending on the function g. We write the corresponding results in the following tables:

Table 1: $g = g_0u^{g_1}$

i	α	f	$\mathbf{v}_k$
1.1	$\forall$	$f_0u^{g_1+1}$	$\mathbf{v}_3 = t\partial_t - \frac{u}{g_1}\partial_u$
1.2	$\forall$	$f_0u^{g_1+1} + f_1u,\ f_1 \neq 0$	$\mathbf{v}_4 = \frac{e^{-f_1g_1t}}{f_1}\partial_t + e^{-f_1g_1t}u\partial_u$
1.3	(1)	$f_0u^{g_1+1},\ g_1 \neq -4/3$	$\mathbf{v}_5 = c_2\mathbf{v}_3 + c_4\mathbf{v}_{51} + \frac{c_2+c_1g_1}{g_1(4+3g_1)}\mathbf{v}_{52}$, (2) $\mathbf{v}_{51} = e^{-A}\left(\partial_x - \frac{2\alpha u}{3g_1+4}\partial_u\right)$, with $A = \frac{g_1}{3g_1+4}\int \alpha\, dx$ $\mathbf{v}_{52} = \left(2g_1e^{-A}\int e^A\, dx\right)\partial_x + 4u\left(1 - \frac{g_1\alpha e^{-A}}{3g_1+4}\int e^A\, dx\right)\partial_u$
1.4	(3)	$f_0u^{g_1+1} + f_1u,\ f_1 \neq 0,\ g_1 \neq -4/3$	$\mathbf{v}_4, \mathbf{v}_5 = c_4\mathbf{v}_{51} + \frac{c_1}{4+3g_1}\mathbf{v}_{52}$, (4)
1.5	$\frac{\alpha_1}{x}$	$f_0u^{f_1},\ f_1 \neq g_1 + 1$	$\mathbf{v}_6 = \frac{2(1-f_1)t}{1+g_1-f_1}\partial_t + x\partial_x + \frac{2u}{1+g_1-f_1}\partial_u$

 ISBN: 978-84-617-8694-7

(1) In this case α, f_0 and g_1 must satisfy the condition

$$e^{-2A}\left(c_4(3g_1+4)+2(g_1c_1+c_2)\int e^A\,dx\right)^2 H(x)=const, \tag{5}$$

where

$$H(x)=2g_0((2+g_1)\alpha^2+(3g_1+4)\alpha_x)-f_0(3g_1+4)^2. \tag{6}$$

(2) The constants c_1, c_2 and c_4 are linked to α, f_0 and g_1 by condition (5).
(3) In this case α, f_0 and g_1 must satisfy the condition

$$e^{-2A}\left(c_4(3g_1+4)+2g_1c_1\int e^A\,dx\right)^2 H(x)=const, \tag{7}$$

where $H(x)$ is given by (6).
(4) The constants c_1 and c_4 are linked to α, f_0 and g_1 by condition (7).

Table 2: $g=g_0u^{-4/3}$

i	α	f	$\mathbf{v}_k$
2.1	(5)	$f_0u^{-1/3}$	$\mathbf{v}_3, \mathbf{v}_{50}=\frac{1}{\alpha}\partial_x+\frac{3\alpha_x}{2\alpha^2}u\partial_u$
2.2	(5)	$f_0u^{-1/3}+f_1u,\ f_1\neq 0$	$\mathbf{v}_4, \mathbf{v}_{50}$

(5) In this case α and f_0 must satisfy the equation

$$3g_0(\alpha^3\alpha_{xx}-2\alpha^2\alpha_x^2+6\alpha_x^3-6\alpha\alpha_x\alpha_{xx}+\alpha^2\alpha_{xxx})-4f_0\alpha^2\alpha_x=0. \tag{8}$$

Table 3: $g=g_0e^{ug_1}$

i	α	f	$\mathbf{v}_k$
3.1	$\forall$	$f_0e^{g_1u}+f_1,\ f_1\neq 0$	$\mathbf{v}_1, \mathbf{v}_7=\frac{e^{-f_1g_1t}}{f_1}\partial_t+e^{-f_1g_1t}\partial_u$
3.2	$\forall$	$f_0e^{g_1u}$	$\mathbf{v}_1, \mathbf{v}_8=t\partial_t-\frac{1}{g_1}\partial_u$
3.3	(6)	$f_1+f_0e^{g_1u},\ f_1\neq 0$	$\mathbf{v}_1, \mathbf{v}_7, \mathbf{v}_9=c_5\mathbf{v}_{91}+c_1\mathbf{v}_{92},$ (7) $\mathbf{v}_{91}=e^{-B}\left(\partial_x-\frac{2\alpha}{3g_1}\partial_u\right)$, with $B=\frac{1}{3}\int\alpha\,dx$ $\mathbf{v}_{92}=\left(\frac{2}{3}e^{-B}\int e^B\,dx\right)\partial_x-\frac{4}{9g_1}\left(\alpha e^{-B}\int e^B\,dx-3\right)\partial_u$
3.4	(8)	$f_0e^{g_1u}$	$\mathbf{v}_1, \mathbf{v}_9=c_2\mathbf{v}_8+c_5\mathbf{v}_{91}+c_1\mathbf{v}_{92},$ (9)
3.5	$\frac{\alpha_1}{x}$	$f_0e^{f_1u}$	$\mathbf{v}_1, \mathbf{v}_{10}=\frac{2t}{f_1-g_1}\partial_t+\frac{x}{f_1}\partial_x-\frac{2}{f_1(f_1-g_1)}\partial_u$

(6) In this case α, f_0 and g_1 must satisfy the condition

$$e^{-2B}\left(c_5+\frac{2}{3}c_1\int e^B\,dx\right)^2\left(9g_1f_0-2g_0(3\alpha_x+\alpha^2)\right)=const. \tag{9}$$

 ISBN: 978-84-617-8694-7

(7) The constants c_1 and c_5 are linked to α, f_0 and g_1 by condition (9).
(8) In this case α, f_0 and g_1 must satisfy the condition

$$e^{-2B}\left(c_5+\frac{2}{3}(c_1+c_2)\int e^B\,dx\right)^2\left(9g_1f_0-2g_0(3\alpha_x+\alpha^2)\right)=const. \tag{10}$$

(9) The constants c_1, c_2 and c_5 are linked to α, f_0 and g_1 by condition (10).

Acknowledgements

The support of Junta de Andalucía group FQM-201 is gratefully acknowledged.

References

[1] M. J. Ablowitz and A. Zeppetella, *Explicit solutions of Fisher's equation for a special wave speed*, Bull. Math. Biol, **41** (1979) 835–840.

[2] J. Belmonte-Beitia, G. F. Calvo and V. M. Pérez-García, *Effective particle methods for the Fisher-Kolmogorov equations: Theory and applications to brain tumor dynamics*, Communications in Nonlinear Science and Numerical Simulation, **19** (2014) 3267-3283.

[3] A. H. Bokhari, M. T. Mustafà and F. D. Zaman, *An exact solution of a quasi-linear Fisher equation in cylindrical coordinates*, Nonlinear Analysis, **41** (2008) 4803–4805.

[4] J. D. Murray, *Mathematical Biology*, Third Edition Springer-Verlag, New York Berlin Heidelberg, (2002).

[5] V. M. Pérez-García, G.F. Calvo, J. Belmonte-Beitia, D. Diego and L. A. Pérez- Romasanta, *Bright solitons in malignant gliomas*, Phys. Rev. E, **84** (2011) 01921.

[6] M. Rosa, M. S. Bruzón and M. L. Gandarias *Symmetry analysis and exact solutions for a generalized Fisher equation in cylindrical coordinates*, Commun Nonlinear Sci Numer Simulat, **25** (2015) 74–83.

[7] K. R. Swanson, C. Bridgea, J. D. Murray, C. Ellsworth and Jr. Alvord, *Virtual and real brain tumors: using mathematical modeling to quantify glioma growth and invasion*, Journal of the Neurological Sciences, **216** (2003) 1-10.

[8] M. Torrisi and R. Tracinà, *An Application of Equivalence Transformations to Reaction Diffusion Equations*, Symmetry, **7** (2015) 1929-1944.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

Recursive filtering algorithm from observations with delays modeled by finite state Markov chains

M.J. García-Ligero[1], A. Hermoso-Carazo[1] and J. Linares-Pérez[1]

[1] *Departamento de Estadística e Investigación Opertiva, Universidad de Granada*

emails: mjgarcia@ugr.es, ahermoso@ugr.es, jlinares@ugr.es

Abstract

This paper addresses the least-squares linear filtering problem of signals from measurements which can be randomly delayed by one or two sampling times. The delays are modeled by homogeneous discrete-time Markov chains to capture the dependence between them. Assuming that the evolution equation generating the signal is not available and that only the first and second-order moments of the processes involved in the observation model are known, a recursive filtering algorithm is derived using an innovation approach. Recursive formulas for filtering error covariance matrices are obtained to measure the goodness of the proposed estimators.

Key words: Markovian delays, filtering
MSC 2000: 60G35, 62M20, 93E10, 93E11

1 Introduction

An unavoidable problem in communication networks is the existence of delays in the arrival measurements. Indeed, in practical situations such as wireless communication channels is common the existence of errors during the transmission due to heavy network traffic, which can lead delayed measurements. The delay may be deterministic or random, although in the most practical cases, such as mobil communications, exploration seismology, between others, the delay is random, being modeled by a stochastic process.

The problem of signal estimation in linear stochastic systems has been widely studied considering different delayed observation models. Most papers on signal estimation from randomly delayed measurements assume the delay is modeled by independent random variables. In this context, attention has been focused on investigating estimation problems from

 ISBN: 978-84-617-8694-7

measurement subject to a random delay which does not exceed one sample time; so, the available observation at any time is considered to be either the current or previous one. In this line, Hounkpevi and Yaz [1] derived a centralized linear minimum variance unbiased filter considering that the delays are modeled by independent Bernoulli variables with different delay probabilities, indicating that the model considered could be generalized to the case of multiple-sample delay. Thus recently, the research is addressed to signal estimation problem from measurements that can be randomly delayed by more one sampling time. In this situation, Linares-Pérez et al. [2] derived a recursive filtering algorithm from observations which can be randomly delayed by one or two sampling times.

On the other hand, in real communication systems, current time delays are usually correlated with the previous ones; a reasonable way to model the dependence on the delays is to consider them as homogeneous Markov chains. State estimation algorithms from measurements with Markovian delays have been derived considering that the delay steps were known on-line via time-stamped data (see, for example, [3]). Later, García-Ligero et al. [4] proposed recursive filtering and smoothing algorithms when the measurements can be delayed by one sampling time assuming known the statistic characteristics of the Markov chain without the need to know if a particular measurement is delayed or updated.

In this paper, a recursive algorithm is proposed for the least-squares linear filter of signals from measurements that can be randomly delayed by one or two sampling times. Assuming the state-space model of the signal is not available and that the delay is modeled by a homogeneous discrete-time Markov chain with three states, the filter is derived by using the information provided by the covariance functions of the process involved in the observations, as well as the probability distribution of Markov chain modeling the delay.

2 Observation model

Consider an n-dimensional signal, x_k, whose scalar measured output, z_k, at each sampling time is perturbed by an additive noise:

$$z_k = H_k x_k + v_k, \quad k \geq 1,$$

where H_k is a known matrix and v_k is the measurement noise.

A common problem in communication theory is the existence of failures during the transmission, which can lead to delays in the arrivals of the measurements. In this paper, it is assumed that the measurements may be randomly delayed by one or two sampling time during the transmission; that is, the observed measurement at time k is $z_{k-a},\ a = 1, 2$, if there is delay, or z_k, if there is not. The delay is modeled by a homogeneous Markov chain, $\{\theta_k,\ k \geq 1\}$, that takes values in the state space $S = \{0, 1, 2\}$. If $\theta_k = a,\ a = 1, 2$, means that the k-th measurement is delayed by a sampling periods; otherwise, if $\theta_k = 0$, there is

 ISBN: 978-84-617-8694-7

not delay in arrival. This situation can be modeled by the following observation model:

$$y_k = \sum_{a=0}^{(k-1)\wedge 2} \delta(\theta_k, a) z_{k-a}, \quad k \geq 1,$$

where $\delta(\cdot,\cdot)$ is the Kronecker delta function.

The aim of this paper is to study the least-squares linear estimation problem of the signal, x_k, from the randomly delayed observations $y_1, \cdots, y_k$. This problem is addressed using the information provided by the covariance functions of the signal and involved noises in the observation model. For this purpose, the following hypotheses about the signal and noise processes are assumed:

(i) The signal process, $\{x_k,\ k \geq 1\}$, has zero mean and its covariance function is given by

$$E[x_k x_s^T] = A_k B_s^T,\ s \leq k,$$

where A and B are known $n \times M$ matrix functions.

(ii) The measurement noise, $\{v_k,\ k \geq 1\}$, is a white process with zero mean and known variances $E[v_k^2] = R_k$.

(iii) $\{\theta_k,\ k \geq 1\}$ is a homogenous Markov chain that takes values in $S = \{0, 1, 2\}$, with known probability distribution $\pi_a^{(k)} = P[\theta_k = a],\ a \in S$, and transition probability matrix $\mathbf{P} = \begin{pmatrix} p_{00} & p_{01} & p_{02} \\ p_{10} & p_{11} & p_{12} \\ p_{20} & p_{21} & p_{22} \end{pmatrix}$, where $p_{ab} = P\left[\theta_k = b/\theta_{k-1} = a\right],\ a, b \in S$.

(iv) The signal process, $\{x_k,\ k \geq 1\}$, the noise, $\{v_k,\ k \geq 1\}$, and $\{\theta_k,\ k \geq 1\}$, are mutually independent.

Given the randomly delayed measurements up to time k, $y_1, \cdots, y_k$, our aim is to determine the least-squares linear estimator, $\widehat{x}_{k/k}$, of the signal, x_k. For this purpose an innovation approach is used.

The innovation approach is based on an orthogonalization procedure by means of which the observation process is transformed into an equivalent one, the innovation process, defined as the differences between each observation and its estimation from the previous ones; the linear estimation problem is then approached by replacing the observation process by the innovation one since both processes provide the same information. As the innovation process is white, the estimator calculated as linear combination of innovations provides a simpler form to obtain the algorithms than that obtained when it is expressed as linear combination of observations. In order to apply this approach, the first step is to calculate the explicit

 ISBN: 978-84-617-8694-7

expression for the innovation and its covariance matrix, and afterwards determining the estimator expression.

According to this approach, the estimator, $\widehat{x}_{k/k}$, is expressed as a linear combination of the innovation, $\mu_h = y_h - \widehat{y}_{h/h-1}$, where $\widehat{y}_{h/h-1}$ is the one-stage predictor of the observation y_h; specifically, the least-squares linear estimator, $\widehat{x}_{k/k}$, is given as follows:

$$\widehat{x}_{k/k} = \sum_{h=1}^{k} E[x_k \mu_h] \Pi_h^{-1} \mu_h, \quad k \geq 1$$

where $\Pi_h = E[\mu_h^2]$ denotes the innovation covariance matrix.

In order to simplify the expressions of the filtering algorithm, the following notations are used:

$$\mathbb{A}_k = \begin{cases} (H_1 A_1 \mid 0 \mid 0)\,(\mathbf{P} \otimes I_M)^T, & k = 1 \\ (H_2 A_2 \mid H_1 A_1 \mid 0)\,(\mathbf{P}^2 \otimes I_M)^T, & k = 2 \\ (H_k A_k \mid H_{k-1} A_{k-1} \mid H_{k-2} A_{k-2})\,(\mathbf{P}^k \otimes I_M)^T, & k \geq 3 \end{cases}$$

$$\mathbb{B}_k = \begin{cases} \left(\pi_0^{(1)} H_1 B_1 \mid 0 \mid 0\right)(\mathbf{P}^{-1} \otimes I_M)^T, & k = 1 \\ \left(\pi_0^{(2)} H_2 B_2 \mid \pi_1^{(2)} H_1 B_1 \mid 0\right)(\mathbf{P}^{-2} \otimes I_M)^T, & k = 2 \\ \left(\pi_0^{(k)} H_k B_k \mid \pi_1^{(k)} H_{k-1} B_{k-1} \mid \pi_2^{(k)} H_{k-2} B_{k-2}\right)(\mathbf{P}^{-k} \otimes I_M)^T, & k \geq 3 \end{cases}$$

$$\mathbb{F}_k = \begin{cases} p_{01}^{(1)} \pi_0^{(1)} R_1, & k = 1 \\ p_{02}^{(1)} \pi_0^{(k)} H_{k-1}(B_{k-1} A_k^T - A_{k-1} B_k^T) H_k^T + p_{01}^{(1)} \pi_0^{(k)} R_k + p_{12}^{(1)} \pi_1^{(k)} R_{k-1}, & k \geq 2 \end{cases}$$

where $\otimes$ denotes the Kronecker product and I_M is the $M \times M$ identity matrix.

3 Filtering algorithm

The filter, $\widehat{x}_{k/k}$, of the signal x_k is obtained by

$$\widehat{x}_{k/k} = A_k O_k^x, \ k \geq 1,$$

where the vectors O_k^x are recursively calculated as

$$O_k^x = O_{k-1}^x + J_k^x \Pi_k^{-1} \mu_k, \ k \geq 1; \quad O_0^x = 0,$$

with

$$J_k^x = \sum_{a=0}^{(k-1)\wedge 2} \pi_a^{(k)} B_{k-a}^T H_{k-a}^T - r_{k-1}^{xy} \mathbb{A}_k^T - \sum_{j=1}^{(k-1)\wedge 2} J_{k-j}^x \Pi_{k-j}^{-1} \mathbb{G}_{k-j}^{(j)}, \quad k \geq 2;$$
$$J_1 = \pi_0^{(1)} B_1^T H_1^T,$$

 ISBN: 978-84-617-8694-7

being

$$\mathbb{G}_k^{(1)} = \mathbb{F}_k + \mathbb{G}_{k-1}^{(2)}\Pi_{k-1}^{-1}(\mathbb{A}_k J_{k-1}^y + \mathbb{G}_{k-1}^{(1)}), \quad k \geq 2;\ \mathbb{G}_1^{(1)} = \mathbb{F}_1,$$
$$\mathbb{G}_k^{(2)} = p_{02}^{(2)}\pi_0^{(k)} R_k, \quad k \geq 1.$$

The innovation, μ_k, is given by

$$\mu_k = y_k - \mathbb{A}_k O_{k-1}^y + \sum_{j=1}^{(k-1)\wedge 2} \mathbb{G}_{k-j}^{(j)}\Pi_{k-j}^{-1}\mu_{k-j},\ k \geq 2; \quad \mu_1 = y_1,$$

where the vectors O_k^y are recursively calculated as

$$O_k^y = O_{k-1}^y + J_k^y \Pi_k^{-1}\mu_k,\ k \geq 1; \quad O_0^y = 0,$$

with

$$J_k^y = \mathbb{B}_k^T - r_{k-1}^y \mathbb{A}_k^T - \sum_{j=1}^{(k-1)\wedge 2} J_{k-j}^y \Pi_{k-j}^{-1}\mathbb{G}_{k-j}^{(j)}, \quad k \geq 2; \quad J_1^y = \mathbb{B}_1^T.$$

The matrices $r_k^{xy} = E[O_k^x O_k^{yT}]$ and $r_k^y = E[O_k^y O_k^{yT}]$ are obtained by

$$r_k^{xy} = r_{k-1}^{xy} + J_k^x \Pi_k^{-1} J_k^{yT}, \quad k \geq 1; \quad r_0^{xy} = 0,$$
$$r_k^y = r_{k-1}^y + J_k^y \Pi_k^{-1} J_k^{yT}, \quad k \geq 1; \quad r_0^y = 0.$$

The innovation covariance matrix, Π_k, is given by

$$\Pi_k = \sum_{a=0}^{(k-1)\wedge 2} \pi_a^{(k)}[H_{k-\theta}A_{k-a}B_{k-a}H_{k-a}^T + R_{k-a}] - \mathbb{A}_k\left[\mathbb{B}_k^T - J_k^y\right]$$
$$- \sum_{j=1}^{(k-1)\wedge 2} \mathbb{G}_{k-j}^{(j)}\Pi_{k-j}^{-1}\left[\mathbb{A}_k J_{k-j}^y + \mathbb{G}_{k-s}^{(j)}\right]^T, \quad k \geq 2;$$
$$\Pi_1 = \pi_0^{(1)}[H_1 A_1 B_1 H_1^T + R_1].$$

The filtering covariance matrix, $\Sigma_{k/k}$, verifies

$$\Sigma_{k/k} = A_k(B_k - A_k r_k^x)^T,\ k \geq 1.$$

where $r_k^x = E[O_k^x O_k^{xT}]$ is recursively calculated by

$$r_k^x = r_{k-1}^x + J_k^x \Pi_k^{-1} J_k^{xT}, \quad k \geq 1; \quad r_0^x = 0.$$

 ISBN: 978-84-617-8694-7

4 Numerical simulation results

In this section, the efficiency of the proposed filtering algorithm is illustrated by a numerical example. For the simulation, a zero-mean scalar signal $\{x_k,\ k \geq 0\}$ with covariance function

$$E[x_k x_s] = 1.025641 \times 0.95^{k-s}, \quad s \leq k,$$

is considered. Clearly this covariance function, according to Hypothesis (i), can be factorized taking

$$A_k = 1.025641 \times 0.95^k \text{ and } B_s = 0.95^{-s}.$$

The measured outputs are affected by a white noise, $\{v_k,\ k \geq 1\}$, with zero mean and variances $R_k = 0.9, \forall k$.

According to the proposed observation model, it is assumed that the available measurements of the signal can be delayed by one or two sample periods during the transmission; that is, the processed observations are modeled by

$$y_k = \sum_{a=0}^{(k-1)\wedge 2} \delta(\theta_k, a) z_{k-a}, \quad k \geq 1.$$

As in Hypothesis (iii), it is assumed that $\{\theta_k,\ k \geq 1\}$ is a homogeneous Markov chain with initial distribution $\pi_0^{(1)} = 1,\ \pi_1^{(1)},\ \pi_2^{(1)} = 0$, (the first observation is not delayed) and transition probability matrix $\mathbf{P} = \begin{pmatrix} 0.99 & 0.006 & 0.004 \\ 0.15 & 0.98 & 0.005 \\ 0.002 & 0.028 & 0.97 \end{pmatrix}$.
Moreover, the signal and noise processes are assumed to be mutually independent.

In order to realize the simulation process, the signal is assumed to be generated from the following first-order autoregressive model,

$$x_{k+1} = 0.95 x_k + w_k$$

where $\{w_k,\ k \geq 0\}$ is a zero-mean white Gaussian noise with $E[w_k^2] = 0.1,\ \forall k$.

Figure 1 shows a simulated signal together with the filtering estimates, $\widehat{x}_{k/k}$. In this figure, we notice the closeness between the evolution of filtering estimates and the signal, which reveals the good performance of the proposed estimators.

Moreover, we have also calculated the filtering error variances assuming that the delay is modeled by different Markov chains. Concretely, we assume the same initial distribution (first observation is not delay) and the following transition probability matrices:

$$\mathbf{P_1} = \begin{pmatrix} 0.95 & 0.03 & 0.02 \\ 0.05 & 0.89 & 0.06 \\ 0.03 & 0.07 & 0.9 \end{pmatrix}, \quad \mathbf{P_2} = \begin{pmatrix} 0.9 & 0.04 & 0.06 \\ 0.07 & 0.87 & 0.06 \\ 0.05 & 0.06 & 0.89 \end{pmatrix}.$$

 ISBN: 978-84-617-8694-7

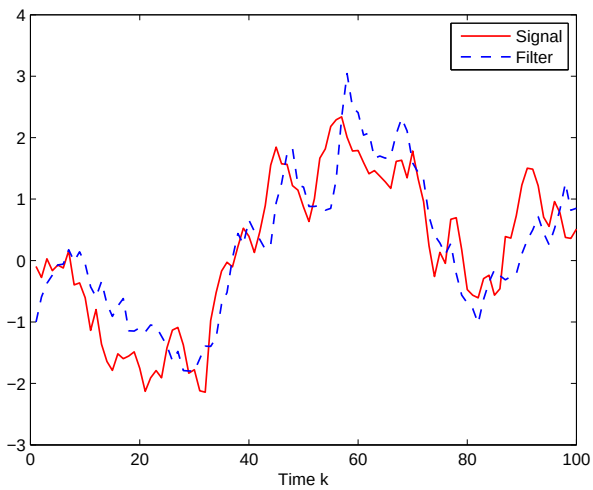


Figure 1: Simulated signal, filtering estimates

The properties of the Markov chains lead us to conclude that the no delay probabilities converge to constant values; in our case these values are $0.58, 0.44$, and 0.37, for the different considered transition probability matrices, $\mathbf{P}, \mathbf{P_1}$ and $\mathbf{P_2}$, respectively. Figure 2 shows the filtering error variances for these models; this figure reveals that as the limit probability of no delay increases, the filtering error variances become smaller and, consequently, the performance of the estimator improves.

 ISBN: 978-84-617-8694-7

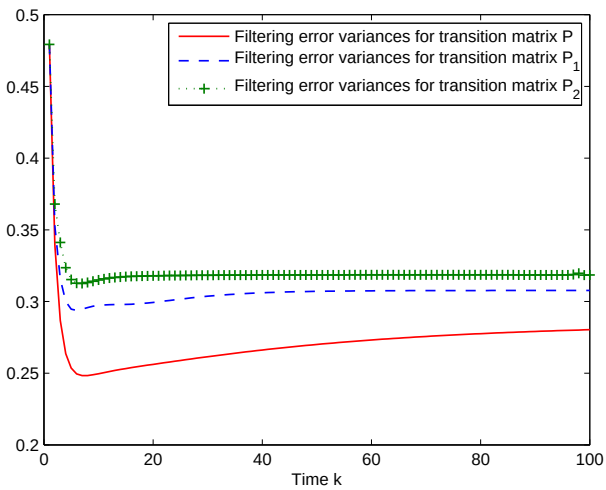


Figure 2: Filtering error variances for different transition probability matrix

Acknowledgements

This research is supported by Ministerio de Economía y Competitividad and Fondo Europeo de Desarrollo Regional FEDER (grant No. MTM2014-52291-P).

References

[1] F. O. Hounkpevi, E. E. Yaz, *Minimum variance generalized state estimators for multiple sensors with different delay rates*, Signal Process. **87** (2007) 602–613.

[2] J. Linares-Pérez, A. Hermoso-Carazo, R. Caballero-Águila and J. D. Jiménez-López, *Least-squares linear filtering using observations coming from multiple sensors with one or two-step random delay*, Signal. Process. **89** (2009) 2045–2052.

[3] J. S. Evans, V. Krishnamurthy, *Hidden Markov model state estimation with randomly delayed observations*, IEEE Trans. Signal Process. **47(8)** (1999), 2157–2166.

[4] M. J. García-Ligero, A. Hermoso-Carazo and J. Linares-Pérez, *Least-squares linear estimation of signals from observations with Markovian delays*, J. Comput. Appl. Math. **236** (2011) 234–242.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

Spectral Decomposition of Skew-symmetric Matrices and Partitioning of Oriented Graphs

Juan Luis García-Zapata[1] **and Juan Antonio Rico-Gallego**[2]

[1] *Department of Mathematics, University of Extremadura*

[2] *Department of Computer Systems Engineering and Telematics, University of Extremadura*

emails: jgzapata@unex.es, jarico@unex.es

Abstract

The symmetric matrices have particular spectral properties, that is, with respect to their eigenvalues (they are real) and eigenvectors (they are orthogonal and form a basis for diagonalization). In particular we are interested in the partitioning of symmetric graphs through a specific eigenvector of the Laplacian matrix. In this work we expose similar spectral properties regarding skew-symmetric matrices, some known, such that their eigenvalues are purely imaginary and not defective, and other seemingly new, that they verify a minimax variational formula. As application we use it in the problem of partitioning non-symmetric, oriented graphs in two parts of equal number of vertices, with the minimum number of edges between the parts.

Key words: Variational Eigenvalue, Spectral Graph Partition

1 Introduction

Among the spectral properties of the symmetric matrices are that they have real eigenvalues and eigenvectors forming an orthonormal basis, in addition to the extreme minimax relations. We propose to generalize these properties to skew-symmetric matrices. Some results are known, but the variational formula of Theorem 2, as far as we know, is not. This is the main development of the paper, and it is motivated by a problem of oriented graph partition that we will describe.

On terminology, we speak of matrices instead of operators (or of symmetric matrices instead of self-adjoint operators) for convenience. Also, although the matrices, of real

entries, are considered as operators in $\mathbb{R}^n$, it will sometimes be considered as operators in $\mathbb{C}^n$. For example, when speaking about the eigenvector corresponding to a complex eigenvalue.

In the following section we recall the spectral properties of symmetrical matrices of interest for us. In the section 3 we develop the results for skew-symmetric matrices, although the formulation of the statements requires some modifications, like the extremal formula containing pairs of purely imaginary eigenvalues. Another modification is the use of real Jordan blocks, instead of diagonal form, to have a real orthonormal basis. Finally in section 4 we suggest the application of these results to the problem of partition of oriented graphs. We delay until that section the exposition of the problem.

2 Spectral Properties of Symmetric Matrices

A generic real matrix A of size $n \times n$ is expressed as:

$$A = \begin{pmatrix} a_{11} & a_{12} & \cdots \\ a_{21} & \ddots & \\ \cdots & & a_{nn} \end{pmatrix} = (a_{ij})_{\substack{i=1,\ldots,n \\ j=1,\ldots,n}}$$

with $a_{ij} \in \mathbb{R}$. We will consider the Euclidean scalar product, defined for vectors $x = (x_1, \ldots, x_n)$ and $y = (y_1, \ldots, y_n)$ of $\mathbb{R}^n$ as $(x, y) = \sum_{i=1}^n x_i y_i$. We will also use the Hermitian product, if $x, y \in \mathbb{C}^n$, defined as $(x, y) = \sum_{i=1}^n x_i \bar{y_i}$. Both products are denoted equally, the context identify the meaning. The Euclidean product allows us to characterize symmetric and skew-symmetric matrices:

Lemma 1. *Let A be a real matrix $n \times n$. $(Ax, y) = (x, Ay)$ for all $x, y \in \mathbb{R}^n$ if and only if A is symmetric. $(Ax, y) = -(x, Ay)$ for all $x, y \in \mathbb{R}^n$ if and only if A is skew-symmetric.*

The proof of the facts of this section can be found in almost any standard book of Linear Algebra, for example [3].

We will denote as λ an eigenvalue of A, that is, a number $\lambda \in \mathbb{C}$ such that there is a non null vector $v \in \mathbb{C}^n$ (called eigenvector corresponding to λ) with $Av = \lambda v$. As it is know a real matrix can have complex eigenvalues, in conjugate pairs because they are the roots of the characteristic polynomial. The set of eigenvalues $\{\lambda_1, \lambda_2, \ldots, \lambda_n\}$ (not necessarily different) is called the spectrum of A. We resume some spectral properties of symmetrical matrices in the following:

Proposition 1. *If A is a $n \times n$ symmetric matrix:*

a) The eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$ are reals.

b) The eigenvectors corresponding to different eigenvalues are orthogonal.

 ISBN: 978-84-617-8694-7

c) There is a basis of eigenvectors (that is, the matrix A is diagonalizable).

In order to calculate the eigenvalues effectively, for example in numerical analysis [2] or in the theory of vibrations [4], variational methods like the following theorem are often used.

The Rayleigh quotient of a symmetric matrix A is

$$R_A(x) = \frac{(Ax, x)}{(x, x)}$$

defined for $x \neq 0$ in $\mathbb{R}^n$.

Theorem 1. *(Courant-Fisher). Let $\mathcal{S}_k$ be the set of subspaces of $\mathbb{R}^n$ of dimension lesser or equal than k, for $k = 1, 2, \ldots, n$. If the eigenvalues are sorted $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$,then*

$$\lambda_k = \min_{E \in \mathcal{S}_k} \max_{\substack{x \in E \\ ||x||=1}} R_A(x)$$

and, dually,

$$\lambda_{n-k+1} = \max_{E \in \mathcal{S}_k} \min_{\substack{x \in E \\ ||x||=1}} R_A(x)$$

Besides, the extreme value is reached in a corresponding eigenvector.

3 Spectral Properties of Skew-symmetric Matrices

We will follow the path outlined by the above results, but in the case of skew-symmetric matrices. Being this an extended abstract, the proofs are omitted. In this section, A will be a real skew-symmetric matrix. We consider it both an lineal operator in $\mathbb{R}^n$ or in $\mathbb{C}^n$. For example a non real eigenvalue $\lambda \notin \mathbb{R}$ has a corresponding eigenvector with non null imaginary part in some components.

Proposition 2. *The eigenvalues λ_k of A are purely imaginary, that is $\lambda_k = \tau_k i$, $\tau_i \in \mathbb{R}$.*

Note that 0 is purely imaginary. In particular, as the eigenvalues appear in conjugate pairs, if the order n is odd the real 0 must be an eigenvalue.

Proposition 3. *Two eigenvectors v_{k_1}, v_{k_2} corresponding to different eigenvalues $\lambda_{k_1}, \lambda_{k_2}$, respectively, are orthogonal.*

In the following proposition, the index starts from zero to ease the notation that will be used later to match pairs of vectors.

 ISBN: 978-84-617-8694-7

Proposition 4. *(Spectral Decomposition). For a skew-symmetric matrix A there exists an orthonormal basis of $\mathbb{C}^n$ that consists of eigenvectors $v_0, v_1, \ldots, v_{n-1}$ corresponding to the eigenvalues $\lambda_0, \lambda_1, \ldots, \lambda_{n-1}$. Moreover, for every $x \in \mathbb{C}^n$ we have:*

$$Ax = \sum_{k=0}^{n-1} \lambda_k (x, v_k) v_k$$

and

$$(Ax, x) = \sum_{k=0}^{n-1} \lambda_k (x, v_k)^2.$$

Propositions 3 and 4 are standards results for normal complex matrices, in particular for skew-symmetric real ones [4]. We have chosen to explicitly state it because, in conjunction with Proposition 2 and the following lemmas, we can arrive at the variational Theorem 2, which is the new result that we need for graph partitioning.

The above expression of Ax and of the quadratic form (Ax, x) in Proposition 4 is diagonal, but of complex coefficients in a basis of complex vectors. We will modify it to have a real expression, on a real orthonormal basis of $\mathbb{R}^n$. It will not be diagonal, but in the form of real Jordan blocks, that is, tridiagonal.

3.1 Orthonormal Real Basis

For a set of vectors $e_1, \ldots, e_k \in \mathbb{R}^n$, $E = \langle e_1, \ldots, e_k \rangle$ is the subspace generated over $\mathbb{R}$ by these vectors. Likewise, if the vectors belong to $\mathbb{C}^n$, $\langle e_1, \ldots, e_k \rangle_{\mathbb{C}}$ is the subspace generated over $\mathbb{C}$. We also denote with $E_\lambda = \mathrm{Ker}(A - \lambda I)$ the eigenspace of the eigenvalue λ.

Lemma 2. *If A is a skew-symmetric real matrix:*

a) If λ is a simple eigenvalue of A, $\bar{\lambda}$ is also simple. If v_λ is a unit vector with $E_\lambda = \langle v_\lambda \rangle_{\mathbb{C}}$, then there is a unit vector $v_{\bar{\lambda}}$ with $E_{\bar{\lambda}} = \langle v_{\bar{\lambda}} \rangle_{\mathbb{C}}$ and such that $u_0 = (v_\lambda + v_{\bar{\lambda}}) \dfrac{\sqrt{2}}{2}$ and $u_1 = (v_\lambda - v_{\bar{\lambda}}) i \dfrac{\sqrt{2}}{2}$ are unit real vectors with $E_\lambda \oplus E_{\bar{\lambda}} = \langle u_0, u_1 \rangle_{\mathbb{C}}$.

b) If λ is an eigenvalue with multiplicity n_λ, $\bar{\lambda}$ has the same multiplicity. For each unit $v_\lambda \in E_\lambda$ there is a unit $v_{\bar{\lambda}} \in E_{\bar{\lambda}}$ such that $u_0 = (v_\lambda + v_{\bar{\lambda}}) \dfrac{\sqrt{2}}{2}$ and $u_1 = (v_\lambda - v_{\bar{\lambda}}) i \dfrac{\sqrt{2}}{2}$ are unit real vectors. Besides if $v_\lambda, v'_\lambda \in E_\lambda$ are linearly independent, its corresponding $v_{\bar{\lambda}}, v'_{\bar{\lambda}}$ are also linearly independent.

We order the complex vector of the orthonormal basis $v_0, v_1, \ldots, v_{n-1}$ of proposition 4 in such a way that the corresponding eigenvalues $\lambda_0, \lambda_1, \ldots, \lambda_{n-1}$ verify $\lambda_{2p} = \bar{\lambda}_{2p+1}$ for

$p = 0, \ldots, \frac{r}{2} - 1$, being r the rank of A, necessarily even. If n is odd, the last eigenvalue is 0. We also consider that $\lambda_{2p} = \tau_p i$ with $\tau_0 \geq \tau_1 \geq \cdots \geq 0$. In this way $\lambda_{2p+1} = -\tau_p i$.

In addition, $\lambda_{2p+1} = \bar{\lambda}_{2p}$ and we can choose the pair of vectors v_{2p}, v_{2p+1} verifying the claim of Lemma 2. If we call u_{2p}, u_{2p+1} to the pair of real vectors that arise from v_{2p}, v_{2p+1} in this Lemma, the set $u_0, u_1, \ldots, u_{n-1}$ is a orthonormal basis of $\mathbb{R}^n$. If we change A to this basis, it results

$$\begin{pmatrix}
0 & \tau_0 & 0 & 0 & & & & & & \\
-\tau_0 & 0 & 0 & 0 & & & & & & \\
0 & 0 & 0 & \tau_1 & & & & & & \\
0 & 0 & -\tau_1 & 0 & & & & & & \\
 & & & & \ddots & & & & & \\
 & & & & & 0 & \tau_{\frac{r}{2}-1} & & & \\
 & & & & & -\tau_{\frac{r}{2}-1} & 0 & & & \\
 & & & & & & & 0 & & \\
 & & & & & & & & \ddots & \\
 & & & & & & & & & 0
\end{pmatrix}$$

The pair u_{2p}, u_{2p+1} verifies $Au_{2p} = \tau_p u_{2p+1}, Au_{2p+1} = -\tau_p u_{2p}$. It is termed a hyperbolic pair in the classification of quadratic forms [5]. Calling $a_k = (x, u_k)$ we have $x = \sum_{k=0}^{n-1} a_k u_k$, as in any basis, but besides

$$Ax = \sum_{p=0}^{r/2-1} \tau_p (a_{2p} u_{2p+1} - a_{2p+1} u_p)$$

Therefore $(x, x) = \sum_{p=0}^{r/2-1} (a_{2p}^2 + a_{2p+1}^2) + \sum_{k=r}^{n-1} a_k^2$ and

$$(Ax, Ax) = \sum_{p=0}^{r/2-1} \tau_p^2 (a_{2p}^2 + a_{2p+1}^2)$$

3.2 Cuadratic Extremal Theorem

With the above expressions we can reach the main result of the paper. We define the quadratic Rayleigh quotient, for $x \neq 0$, as:

$$S_A(x) = \frac{(Ax, Ax)}{(x, x)}$$

Being $u_0, u_1, \ldots, u_{n-1}$ the above orthonormal real basis, we define $F_p = \langle u_0, u_1, \ldots, u_{2p}, u_{2p+1} \rangle$. We can proof:

 ISBN: 978-84-617-8694-7

Proposition 5.

$$\tau_p^2 = \min_{\substack{x \in F_p \\ x \neq 0}} S_A(x)$$

Based on that proposition, we can prove that:

Theorem 2. *If $\mathcal{S}_{2p}$ be the set of subspaces of $\mathbb{R}^n$ of dimension lesser or equal than $2p$, then*

$$\tau_p^2 = \max_{E \in \mathcal{S}_{2p}} \min_{\substack{x \in E \\ ||x||=1}} S_A(x)$$

and, dually,

$$\tau_{n-2p+1}^2 = \min_{E \in \mathcal{S}_{2p}} \max_{\substack{x \in E \\ ||x||=1}} S_A(x)$$

Besides, the extreme value is reached in a corresponding eigenvector.

4 Partition of Oriented Graphs

Our main motivation for developing the above variational formula of Theorem 2 is its use in spectral partition of oriented graphs. An undirected graph, commonly called graph without further specification, is a set V of vertices and a set E of edges. Each edge is a set $\{p_1, p_2\}$ of two vertices of V. Its graphic representation associates each vertex to a point of the representation medium and each edge to a line that joins the corresponding vertices. An oriented graph is an undirected graph in which in each edge $\{p_1, p_2\}$ has been chosen an orientation, or arrow, either (p_1, p_2) or (p_2, p_1). It is different from the concept of directed graph, where between two vertices can go both arrows (or also one of them, or none at all).

In this work we study oriented graphs, well represented by the oriented incidence matrix A, skew-symmetric, with entries:

$$a_{ij} = \begin{cases} +1 \text{ if the orientation is } (p_i, p_j) \\ -1 \text{ if the orientation is } (p_j, p_i) \\ 0 \text{ in other case} \end{cases}$$

We briefly describe now the spectral partition of non-directed graphs [7], and then sketch how Theorem 2 allows to develop an analogous theory for oriented graphs. The problem of graph partitioning is to divide the set of vertices in two equal parts, so that the total number of edges between the two parts is minimized. One of the fields where this problem arises is in high performance computing (HPC). The distributed HPC applications are composed of many processes, each with its computational load. These processes must communicate with each other, and that slows the computation. The application is modeled by a graph where the vertices are the processes and the edges represent the communications. When assigning

 ISBN: 978-84-617-8694-7

the processes to the different processors of a parallel architecture, this assignment must be done in equal parts (so that the computational load is balanced), but also minimizing the number of edges between parts (so that the communication delays are minimized) [1].

An approximate solution to this combinatorial optimization problem, in the case of non-directed graphs, is obtained by the spectral theory of the Laplacian matrix L. This matrix is $L = D - M$, where M is the adjacency matrix, of entries:

$$m_{ij} = \begin{cases} 1 \text{ if } \{p_i, p_j\} \text{ is an edge} \\ 0 \text{ in other case} \end{cases}$$

and D is the diagonal degree matrix: d_{ii} is the number of vertices adjacent to p_i. The meaning of this construction, $D - M$, is to add a valuated loop in each vertex, so that the resulting graph verifies the Kirchhoff's current law: the sum of the incidences in each vertex (the sum of the rows of L) is zero. As a mathematical consequence, one of the eigenvalues of the Laplacian matrix L is $\lambda_1 = 0$ and the corresponding eigenvector is the vector of ones $\mathbf{1} = (1, 1, \ldots, 1)$. This causes that the eigenvectors corresponding to nonzero eigenvalues are orthogonal to $\mathbf{1}$. The Fiedler eigenvector v_f is the corresponding to the minimum nonzero eigenvalue λ_2. By the extremal Theorem 1, $\lambda_2 = \min\limits_{E \in \mathcal{S}_2} \max\limits_{x \in E}(Lx, x)$, and the minimum of (Lx, x) between those x orthogonal to $\mathbf{1}$ is precisely the Fiedler eigenvector v_f. If c is the characteristic vector of a vertex partition (that is, the i-th component of c is $+1$ if p_i belongs to the first part, -1 if p_i belongs to the second), it can be viewed that (Lc, c) is proportional to the number of edges between the two parts. That is, the negative and positive components of the Fiedler vector approximates the characteristic vector of a partition that minimizes the number of edges. A bound on the error of this approximation is given by [6].

To transfer this spectral partitioning scheme, from the case of non-directed graphs to the case of oriented graphs, is the practical application that we search when developing Theorem 2. To make a construction similar to that of Kirchhoff, obtaining a matrix with $\mathbf{1}$ as eigenvector, keeping the skew-symmetry, we consider the vector d of total degrees, that is, the i-th component is the number of arrows coming out of p_i minus those coming in. The total degree can be negative, if there are more arrows arriving at the vertex than coming out. We define the skew-symmetric Laplacian L_s as the matrix with the following blocks:

$$L_s = \begin{pmatrix} A & d \\ -d^t & 0 \end{pmatrix}$$

being d^t the vector of total degrees considered as a row matrix. The skew-symmetric Laplacian has eigenvalue 0 and corresponding eigenvector $\mathbf{1}$. By the spectral properties of the skew-symmetric matrices described in section 3, the non null eigenvalues of L_s with minimal imaginary part, λ_{2f} and λ_{2f+1} with $f = \frac{r}{2} - 1$ (r the rank of L_s), have a pair v_{2f}, v_{2f+1} of complex eigenvectors whose real equivalents u_{2f}, u_{2f+1} by Lemma 2 are analogous to

 ISBN: 978-84-617-8694-7

the Fiedler vector: they are orthogonal to **1** (hence approximating the characteristic of a partition) and minimize the expression $(L_s x, L_s x)$, that can be seen proportional to the number of arrows between the two parts.

In summary, the spectral properties of the eigenvalues closest to zero in the matrices of Laplacian type, either symmetric or skew-symmetric, allow us to approximate a solution to the combinatorial problem of graph partitioning.

Acknowledgements

This work was supported by EU under the COST Program Action IC1305: Network for Sustainable Ultrascale Computing (NESUS).

References

[1] C.E. Bichot and P. Siarry. *Graph Partitioning.* ISTE. Wiley, 2013.

[2] G.H. Golub and C.F. Van Loan. *Matrix Computations.* Johns Hopkins Studies in the Mathematical Sciences. Johns Hopkins University Press, 1996.

[3] R.A. Horn and C.R. Johnson. *Matrix Analysis.* Cambridge University Press, 1990.

[4] P. Lancaster and M. Tismenetsky. *The Theory of Matrices: With Applications.* Computer Science and Scientific Computing Series. Academic Press, 1985.

[5] S. Lang. *Algebra.* Springer, 3rd edition, 2002.

[6] Bojan Mohar. Isoperimetric numbers of graphs. *Journal of Combinatorial Theory, Series B*, 47(3):274–291, 1989.

[7] Daniel A Spielman. Spectral graph theory and its applications. In *Foundations of Computer Science, 2007. FOCS'07. 48th Annual IEEE Symposium on*, pages 29–38. IEEE, 2007.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

GPU Classification for Hyperspectral Images based on Convolutional Neural Networks

Alberto S. Garea[1], Dora B. Heras[1] and Francisco Argüello[2]

[1] *Centro singular de Investigación en Tecnoloxías da Información (CiTIUS), Universidade de Santiago de Compostela*

[2] *Departamento de Electrónica y Computación, Universidade de Santiago de Compostela*

emails: jorge.suarez.garea@usc.es, dora.blanco@usc.es, francisco.arguello@usc.es

Abstract

Recently, deep learning techniques based on Convolutional Neural Networks (CNN) have started to be used for the classification of hyperspectral images. These techniques present high computational cost when preprocessing stages are applied. In this paper, a GPU (Graphical Processor Unit) implementation of a spatial-spectral supervised classification scheme based on CNNs and applied to remote sensing datasets is presented. The scheme comprises convolution filters for processing the spectral information and a patch around each pixel to take the spatial information into account. To reduce the size of the filters, the dimensionality of the dataset is previously reduced using Principal Component Analysis (PCA). In order to achieve an efficient GPU projection, different techniques and optimizations have been applied such as the use of the deep learning framework Caffe. Speedups of up to 38.66× over the Pavia University dataset are obtained together with competitive classification accuracies.

Key words: Hyperspectral, Classification, Convolutional neural network, Deep learning, Caffe, GPU.

1 Introduction

Hyperspectral images contain a large amount of information that can be exploited during the processing. This information is not only spectral but there is also a lot of spatial information in the neighborhood of each pixel. Hyperspectral techniques that can exploit

 ISBN: 978-84-617-8694-7

both types of information are known by the name of spectral-spatial techniques [1]. When these techniques are introduced in the classification of hyperspectral images, experimental results show great improvements in the accuracy results.

Recently, deep-learning techniques have started to be introduced in the field of classification of hyperspectral datasets [2, 3, 4, 5, 6]. These classifiers consist of several layers with nonlinear processing units to extract and transform different features. Each layer uses the output of the previous layer as input and the network can be trained in a supervised or unsupervised manner. Applications include pattern recognition and statistical classification. The proposed methods extract spatial information using structures such as Multilayer Perceptrons (MLP) or Convolutional Neural Networks (CNN). Usually before the extraction of spatial information, a dimensionality reduction is performed using techniques such as Principal Component Analysis (PCA), Independent Component Analysis (ICA) or wavelets in order to obtain moderately small vectors.

A CNN contains convolutional layers that can be used to perform spatial convolutions on the hyperspectral image bands. Usually pooling layers are also included in order to apply some kind of decimation and reduce the number of coefficients. A CNN may have one or more convolutional layers, but the final classification is performed using one or more fully-connected layers. Activation functions to introduce non-linearity, usually of sigmoid type, can be included in convolutional layers. Such functions are similar to those used in MLPs. Usually the backpropagation algorithm is used to set the coefficients of both, neurons of fully-connected layers and convolution filters.

Some published deep-learning schemes applied to hyperspectral images use only the spectral information. Thus, Hu *et. al* [4] propose a scheme based on CNNs, which does not consider spatial information since each input is a single pixel-vector. Other schemes incorporate the spectral and spatial information separately to the classifier, often constructing a stack-vector for input to the neural network and using PCA [2, 3, 6, 5].

Remote sensing hyperspectral applications are computationally demanding and, therefore, good candidates to be projected in high performance computing infrastructures such as clusters or specialized hardware devices [7]. GPUs provide a cost-efficient solution to carry out onboard real-time processing of remote sensing hyperspectral data for performing hyperspectral unmixing, classification or change detection, among others [8]. In the case of deep learning techniques, different high-level frameworks optimized for GPU computing are available, such as Theano, Caffe [9], TensorFlow or Torch. The implementations of deep learning methods for hyperspectral images are in some cases presented in terms of execution times but without an analysis of the computational cost [10]. In other cases the use of an optimized framework such as Caffe [11] is mentioned but without including execution times or a detailed analysis of the implementations.

In this paper we propose a CUDA GPU spectral-spatial classification scheme for hyperspectral images based on CNNs and implemented mainly by using the Caffe analyzing

 ISBN: 978-84-617-8694-7

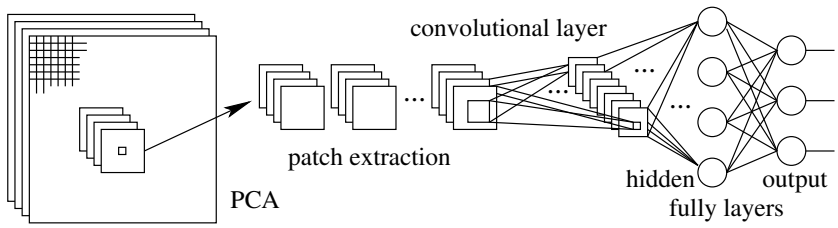


Figure 1: HYCNN scheme for the classification of hyperspectral images.

the details of the implementation. In order to reduce the size of the convolution filters, the image dimensionality is previously reduced using PCA, as it will be explained in the next section.

The paper is organized as follows: section 2 presents the proposed spectral-spatial classification scheme in CPU, section 3 presents the GPU code. The evaluation is performed in section 4, and, finally, section 5 presents the conclusions.

2 Spectral-Spatial CNN-Based Classification

In this section we present a scheme for the classification of hyperspectral images based on PCA, patch extraction, and CNNs, that we called HYCNN. Fig. 1 shows the operations performed and the network structure. These are described in more detail in the pseudocode of Algorithm 1, which also indicates the adjustable parameters in the scheme.

Algorithm 1 Steps of the HYCNN scheme

Input: Hyperspectral image
Output: Classification map
Parameters:
N_1: number of principal components
$H \times V$: patch size
N_2: number of convolution filters
$F_1 \times F_2$: spatial size of filters
$D_1 \times D_2$: decimation factor
N_3: number of neurons in hidden layer
η: learning parameter

1. Preprocessing
1.1 PCA on the image
2. Patch extraction
2.1 Patch around each pixel
3. Convolutional layer
3.1 Convolution filtering
3.2 Pooling (average)
3.3 Activation function (sigmoid)
4. Fully-connected layers
4.1 Hidden layer (with sigmoid)
4.2 Output layer (with sigmoid)

As a first step, HYCNN performs a reduction of image dimensionality using PCA. It extracts the most significant information from the hyperspectral image in the spectral dimension and reduces the number of components, which progress to the next step of the algorithm.

A patch is then extracted around each pixel to be classified. This step aims at getting the spatial information in the neighborhood of a pixel in addition to the spectral information. Accordingly, the patch has the same number of components as those retained from the PCA, and comprises a window around the pixel. The window size is an adjustable parameter of the classification. Each patch is considered a sample and used as the unit of information during the training and classification phases by the CNN.

The next step is the processing of each patch by the CNN. This consists of three parts: convolutional filters, pooling layer and activation function. A convolutional layer is a locally connected structure which is convolved with the image to produce several feature maps, one for each filter. Each filter consists of a rectangular grid of neurons. Unlike a fully-connected layer, the filter coefficients used in all the nodes are the same.

The convolutional layer of our scheme processes several components (spectral bands). The inputs to the filters are the patches, which we assume to have in this sequential algorithm a size of $H \times V \times N_1$, being H and V the size of the spatial dimensions, and N_1 the number of bands. In order to extract multiple features, the convolutional layer comprises N_2 filters, so we will have this same number of maps (planes) at the output. Regarding the size of the filters, if $F_1 \times F_2$ is the size of the spatial grid, each filter will have $F_1 \times F_2 \times N_1$ coefficients.

The pooling layer takes small rectangular blocks from the convolutional layer and subsamples them to produce a single output from each block. For the pooling layers each map is subsampled with mean pooling over blocks of size $D_1 \times D_2$. After the subsampling, a sigmoidal nonlinearity is applied to each feature map.

The last part of the scheme consists of fully-connected layers, which perform the high-level reasoning of the CNN. A fully connected layer takes all the outputs in the previous layer and connects then to every single neuron it has. This type of layer is arranged in one dimension, so they are not spatially located operations anymore. In this paper we use the typical two-layer MLP, with hidden and output layers. The number of neurons in the hidden layer is the adjustable parameter N_3, while the output layer has a number of neurons equal to the number of classes in the hyperspectral image. The activation function in both, convolutional and fully-connected layers, is of sigmoid type.

The learning of all the layers of the CNN in this scheme is conducted using a backpropagation algorithm. The error is computed at the output of the network using the training samples and comparing the results to the reference data. Then, the error is propagated backwards through the network. The backpropagation is used in conjunction with an optimization method, in this case a gradient descent. It calculates the gradient of a cost function with respect to all the weights of the network, and then updates the weights in an attempt to minimize the cost function. The learning parameter, usually denoted as η, indicates how much the weights are adjusted at each update.

 ISBN: 978-84-617-8694-7

3 Spectral-Spatial CNN-Based Classification in GPU

In this section we introduce some Compute Unified Device Architecture (CUDA) programming fundamentals as well as the CUDA GPU implementation of the scheme proposed in Sect. 2.

3.1 CUDA GPU programming fundamentals

CUDA is a parallel computing platform and programming model that enables NVIDIA GPUs to execute programs invoking parallel functions called kernels [12]. Each kernel launches a user-defined number of threads that are organized into blocks. The blocks are arranged in a grid that is mapped to a hierarchy of CUDA cores in the GPU. Threads can access data from multiple memory spaces. Each block has a shared memory that is visible exclusively to the threads within this block and whose lifetime is equal to the block lifetime. The shared memory lifetime makes it difficult to share data among thread blocks. This implies the use of global memory whose access is slower than shared memory access. The new Pascal architecture has introduced changes regarding the memory hierarchy [13].

Different performance optimization strategies have been applied in this work. The most important is to reduce the data transfers between the CPU and the GPU memories. Another key is to improve the efficiency in the use of the memory hierarchy by performing the maximum number of computations on the data already stored in shared memory. The search for the best kernel configurations is also fundamental. To get the highest possible occupancy is the only way to hide latencies and keep the hardware busy. To achieve this, the maximum block size for each kernel is selected with the requirement that the number of registers and the shared memory usage do not act as occupancy limiters. Finally, the existing CUDA optimized libraries must be used. CULA [14], MAGMA [15], and CUBLAS [16] are used for algebra operations. For the deep learning calculations the Caffe framework is used. It performs calls to CuDNN [17], CUBLAS and MAGMA. CuDNN is a GPU-accelerated library for deep neural networks.

3.2 CUDA implementation

In this section the GPU implementation of the HYCNN algorithm described in section 2 is detailed. The pseudocode in Algorithm 2 shows a detailed description of the classification scheme. The kernels executed in GPU are placed between <> symbols. The pseudocodes also include the GM and SM acronyms to indicate kernels executed only in global memory and kernels that only use shared memory, respectively. The whole forward-backward process for the training phase of the algorithm is detailed. The CNN is implemented using Caffe. Since the calls to Caffe functions produce a high number of calls to libraries, these are grouped in the pseudocode by steps of the scheme and only the most repeated kernels are

included pointing out the call sequence.

Algorithm 2 HYCNN classifier for hyperspectral images (GPU) → Training step

Input: Hyperspectral image

GPU EVD-PCA algorithm

1: **for** each epoch **do**
Forward
Convolution filtering
2: **for** each training sample **do**
3: im2col_gpu() → <im2col_gpu> ▷ GM
4: caffe_gpu_gemm() → cublasSgemm() → <gemmSN_NN>, <gemmK1> ▷ SM + GM
5: **end for**
Average Pooling
6: PoolingLayer::Forward_gpu() → <AvePoolForward> ▷ GM
Convolution Activation
7: CuDNNSigmoidLayer::Forward_gpu() → cudnnActivationForward() → <activation_fw_4d> ▷ GM
First Inner
8: InnerProductLayer::Forward_gpu() → caffe_gpu_gemm() → <sgemm_largeK>, <gemmk1> ▷ SM + GM
First Inner activation
9: CuDNNSigmoidLayer::Forward_gpu() → cudnnActivationForward() → <activation_fw_4d> ▷ GM
Second Inner
10: InnerProductLayer::Forward_gpu() → caffe_gpu_gemm() → <sgemm>, <gemmk1> ▷ SM + GM
Second Inner Activation
11: CuDNNSigmoidLayer::Forward_gpu() → cudnnActivationForward() → <activation_fw_4d> ▷ GM
SoftMax with Loss
12: CuDNNSoftmaxLayer::Forward_gpu() → cudnnSoftmaxForward() → <softmax_fw> ▷ SM + GM
13: SoftmaxLossForwardGPU() → <SoftmaxLossForwardGPU>, <cublasSasum> ▷ GM

Backward
14: SoftmaxLossBackwardGPU() → <SoftmaxLossBackwardGPU>, <cublasSscal> ▷ GM
Second Inner Activation
15: CuDNNSigmoidLayer::Backward_gpu() → cudnnActivationBackward() → <activation_bw_4d> ▷ GM
Second Inner
16: InnerProductLayer::Backward_gpu() → <sgemmNT2>, <gemmv2N>, <sgemm_128x64> ▷ SM + GM
First Inner Activation
17: CuDNNSigmoidLayer::Backward_gpu() → cudnnActivationBackward() → <activation_bw_4d> ▷ GM
First Inner
18: InnerProductLayer::Backward_gpu() → <sgemm_128x64>, <gemmv2N>, <sgemm_128x64> ▷ SM + GM
Convolution Activation
19: CuDNNSigmoidLayer::Backward_gpu() → cudnnActivationBackward() → <activation_bw_4d> ▷ GM
Pooling
20: PoolingLayer::Backward_gpu() → <AvePoolBackward> ▷ GM
Convolution filtering
21: **for** each training sample **do**
22: ConvolutionLayer::Backward_gpu():
23: backward_gpu_bias() → <gemv2T> ▷ SM + GM
24: weight_gpu_gemm() → <im2col_gpu> ▷ GM
25: backward_gpu_gemm() → <gemmSN_TN> ▷ SM + GM
26: **end for**
Weights update
27: caffe::SGDSolver() → <SGDUpdate> ▷ GM
28: **end for**

 ISBN: 978-84-617-8694-7

As a first step, the PCA algorithm using EVD (EVD-PCA) is applied to reduce the dimensionality of the dataset. For details of the GPU implementation see [18].

A patch is then extracted around each pixel and stored into a two different Lightning Memory-Mapped Databases (LMDBs) to be accessed from the Caffe framework. The first database stores the training patches whereas the second one stores the test patches. As shown in the pseudocode, both, the CNN steps and the fully-connected layers steps are applied to each patch N times (epochs).

The training phase is divided into two main steps: forward and backward. The forward step computes all the training patches through the full network to obtain a classification result and the backward step updates the network weights to adjust the obtained classification result.

The forward step starts applying the convolution filters to each training patch. Unlike in the CPU version where the $H \times V$ pixels of the patch are computed through the convolution filters sequentially, in the GPU version the patch is converted first into a matrix using the im2col_gpu function (line 3 in the pseudocode 2). Then, it is multiplied by a matrix containing the convolutional values using the cublasSgemm function (line 4).

Next, a pooling substep is performed using a Caffe kernel called AvePoolForward. This function computes the pooling over all the training patches at the same time. The last substep of the CNN is the activation. The CuDNNSigmoidLayer::Forward_gpu() calls the CuDNN funtion to perform the sigmoid activation (line 7).

Once the CNN has finished, two fully-connected layers perform the classification. First, an inner product function (line 8) that uses CuBLAS multiplies the CNN output matrix by a matrix of learned weights. Next, a sigmoid activation function (line 9) is applied over the previous result using a CuDNN function. The previous two operations are repeated over the last fully-connected layer (lines 10 and 11).

At this point, the output of the full network contains the classification of each training patch. Then, a softmax function (line 12) is applied to get a probability distribution over classes. This function takes a vector of arbitrary real-valued scores and converts it to a vector of values between zero and one that sum one. The last substep of the forward is to compute the loss of the network using the function SoftmaxLossForwardGPU (line 13).

Regarding the backward step, it includes all the substeps of the forward step but applied in reverse order (lines 14-26). This allows to update the values of all the neurons in the full network based on the results of the loss function computed in the forward step. At the end of the loop, the update of all the weights of the full network is performed (line 27). This task is carried out by a Caffe function.

 ISBN: 978-84-617-8694-7

4 Results

This section shows the experimental results obtained for the GPU HYCNN scheme comparing to the CPU scheme in terms of computation time and classification accuracy.

The proposed algorithms have been evaluated on a PC with a quad-core Intel i5-6600 at 3.3GHz and 32 GB of RAM. The codes have been compiled using the gcc 4.8.4 version with OpenMP (OMP) 3.0 support under Linux using four threads. The OPENBLAS library has been used to accelerate the algebra operations included in the algorithms. Regarding the GPU implementation, CUDA codes run on an Pascal NVIDIA GeForce GTX 1070 with 15 Streaming Multiprocesors (SMs) and 128 CUDA cores each. The CUDA codes have been compiled under Linux using the nvcc version 8.0.26 of the toolkit. As usual in remote sensing [19], measures of classification accuracy are given in terms of overall accuracy (OA), which is the percentage of correctly classified pixels comparing to the reference data information available. The computational performance results are expressed in terms of execution times and speedups. The results are the average of 10 independent executions.

The algorithms have been used over two remote sensing datasets: a 103-band ROSIS image of the University of Pavia (Pavia Univ.) and a 220-band AVIRIS image taken over Northwest Indiana (Indian Pines). The images and the corresponding reference data are shown in Fig. 2.

For each dataset the samples are randomly distributed between the training [18] and testing sets. During the testing stage all pixels of the image are classified, but the samples used in the training stage are excluded for the calculation of the accuracy results (see Table 1).

Datasets	Sensor	classes	Dimensions	samples	training samples
Pavia Univ.	ROSIS	9	610×340×103	42776	3921 (9.17%)
Indian Pines	AVIRIS	16	145×145×220	10249	695 (6.78%)

Table 1: Information for the test remote sensing datasets.

The configuration parameters were determined by performing experiments varying the number of principal components, the batch size and the filter size for the code executed in CPU. These parameters are also used for the GPU implementation when both codes are compared from the computational point of view. The base parameters considered for the comparison are $H = V = 28$ (patch size), $N_1 = 4$ (number of principal components), $N_2 = 16$ (number of filters), $N_3 = 100$ (neurons in the hidden layer), $F_1 = F_2 = 5$ (filter size), and $D_1 = D_2 = 2$ (decimation factor). The backpropagation algorithm was performed with learning parameter $\eta = 0.2$, batch size equal to the number of training samples, and a total of 200 epochs. The Caffe framework using a block size of 512 is used to execute the GPU version of the HYCNN scheme except for the initial PCA algorithm.

 ISBN: 978-84-617-8694-7

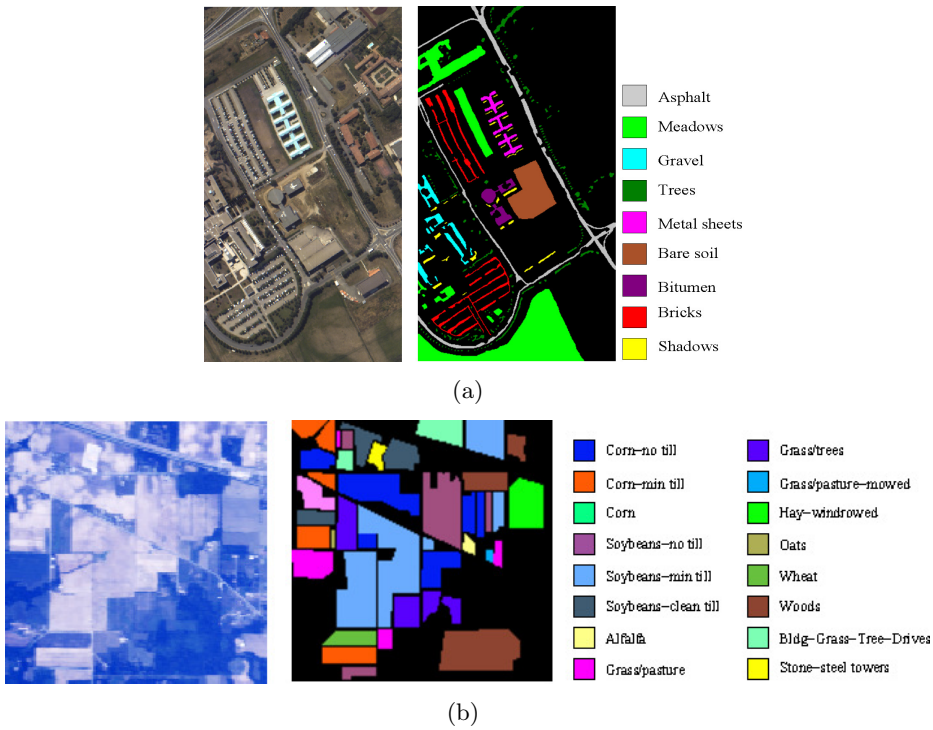


Figure 2: Hyperspectral datasets: (a) *Pavia Univ.*, (b) *Indian Pines*.

Table 2 shows the comparison between the CPU and the GPU implementations of the scheme when it is applied to the Pavia Univ. image. For a better understanding of the results, different parts of the code have been grouped into higher abstraction level functions. The times are split following the functions of the pseudocode in Algorithm 2 but aggregating the results for the backward step. The speedups are calculated as the number of times that the GPU code is faster than the CPU code. The biggest speedup is observed for the First Inner function that comprises the update of the hidden layer neurons in the fully-connected network. The speedup for the Second Inner is lower because the size of the matrix by matrix multiplication is smaller as it corresponds to layers with fewer neurons. The most time consuming function is the Convolution in the Forward step. Its speedup is only 47.41× because this function includes a group of kernels with low occupancy.

Table 3 shows the execution times and speedups for the whole classification scheme for the two test datasets (including the training and testing steps and also the PCA step) as well as the classification accuracies for both implementations. It is important to enhance that the same configuration parameters were used for both implementations in order to compare the computational time in the same conditions. Nevertheless, for the GPU accuracies the

Step	Lines	CPU	GPU	Speedup
Forward step				
Convolution	2-5	2.06537s	0.04356s	47.41×
Average Pooling	6	0.03557s	0.00556s	6.40×
Convolution Act.	7	0.26165s	0.01389s	18.83×
First Inner	8	1.92316s	0.00225s	854.74×
First Inner Act.	9	0.01281s	0.00063s	20.33×
Second Inner	10	0.00621s	0.00006s	103.50×
Second Inner Act.	11	0.00126s	0.00005s	25.20×
Loss	12-13	0.00023s	0.00015s	1.53×
Backward step				
Second Inner	14-16	0.00350s	0.00007s	50.00×
First Inner	17-18	0.95480s	0.00537s	177.80×
Convolution	19-26	1.61309s	0.04792s	33.66×
Total		6.87742s	0.11939s	57.60×

Table 2: CPU and GPU execution times and speedups for the training step of the HYCNN scheme for Pavia Univ. dataset. The column Lines shows the lines in Algorithm 2

	CPU		GPU		Speedup
Dataset	Time	Accuracy (%)	Time	Accuracy (%)	
Pavia Univ.	1404.26s	98.50	36.32s	97.15	38.66×
Indian Pines	252.33s	97.14	7.60s	84.84	33.20×

Table 3: Execution times,speedups and classification accuracies for the HYCNN scheme.

parameters were optimized for the GPU code separately. The patch size was reduced to 256. In addition, the number of epochs was set to 1300 and 3683 for the Pavia Univ. and the Indian Pines images respectively. The differences in classification accuracy among the CPU and the GPU schemes are produced by the weights update during the backpropagation. For the CPU case the update is carried out for each sample separately, on the contrary for the GPU case the updates are performed by blocks of samples.

5 Conclusions

In this paper we propose a spectral-spatial scheme in GPU based on convolutional neural networks for the classification of hyperspectral images and evaluate its results on several public datasets used in remote sensing for land-cover applications. The scheme consists of principal component analysis, patch extraction, convolution filters and fully-connected layers. The learning is performed using the standard backpropagation algorithm.

 ISBN: 978-84-617-8694-7

The CUDA GPU implementation is based on the use of the Caffe optimized framework for deep learning and other optimization strategies including calls to optimized libraries such as CULA, CUBLAS and MAGMA. Details on the Caffe implementation are given. The experiments obtain speedups of up to 38.66% for the Pavia Univ. dataset with accuracies of up to 97.15%.

Acknowledgments

This work was supported in part by the Consellería de Cultura, Educación e Ordenación Universitaria [grant numbers GRC2014/008 and ED431G/08] and Ministry of Education, Culture and Sport, Government of Spain [grant numbers TIN2013-41129-P and TIN2016-76373-P]. These grants are co-funded by the European Regional Development Fund (ERDF).

References

[1] M. Fauvel, Y. Tarabalka, J.A. Benediktsson, J. Chanussot, and J.C Tilton, *Advances in spectral-spatial classification of hyperspectral images*, Proceedings of the IEEE, vol. 101, no. 3, pp. 652-675, 2013.

[2] Y. Chen, Z. Lin, X. Zhao, G. Wang, and Y. Gu (2014), *Deep learning-based classification of hyperspectral data*, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 7(6), 2094-2107.

[3] J. Yue, W. Zhao, S. Mao, and H. Liu (2015), *Spectral-spatial classification of hyperspectral images using deep convolutional neural networks* Remote Sensing Letters, 6(6), 468-477.

[4] W. Hu, Y. Huang, L. Wei, F. Zhang, and H. Li (2015), *Deep Convolutional Neural Networks for Hyperspectral Image Classification*, Journal of Sensors, 2015, 258619.

[5] K. Makantasis, K. Karantzalos, A. Doulamis, and N. Doulamis (2015), *Deep supervised learning for hyperspectral data classification through convolutional neural networks*, Proc. IEEE Int. Geoscience and Remote Sensing Symposium (IGARSS), 4959-4962.

[6] W. Zhao, Z. Guo, J. Yue, X. Zhang, and L. Luo (2015), On combining multiscale deep learning features for the classification of hyperspectral remote sensing imagery, International Journal of Remote Sensing, 36(13), 3368-3379.

[7] E. Christophe, J. Michel, and J. Inglada, *Remote sensing processing: From multicore to GPU*, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 4, no. 3, pp. 643-652, 2011

 ISBN: 978-84-617-8694-7

[8] A. Plaza, Q. Du, Y. Chang, and R.L. King, *High performance computing for hyperspectral remote sensing*, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 4, no. 3, pp. 528-544, 2011

[9] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, *Caffe: Convolutional Architecture for Fast Feature Embedding*, arXiv preprint arXiv:1408.5093, 2014

[10] Y. Chen, H. Jiang, C. Li, X. Jia, and P. Ghamisi, *Deep feature extraction and classification of hyperspectral images based on convolutional neural networks*, IEEE Transactions on Geoscience and Remote Sensing, vol. 54, no. 10, pp. 6232-6251, 2016

[11] E. Aptoula, M.C. Ozdemir, and B. Yanikoglu, *Deep Learning With Attribute Profiles for Hyperspectral Image Classification*, IEEE Geoscience and Remote Sensing Letters, vol. 13, no. 12, pp. 1970-1974, 2016

[12] David B. Kirk and Wen-mei W. Hwu, *Programming Massively Parallel Processors A Hands-on Approach*, Morgan Kaufmann, 2016.

[13] Nvidia, *Whitepaper: NVIDIA Tesla P100 (2017)*, Available: https://www.nvidia.com/content/PDF/kepler/NVIDIA-Kepler-GK110-Architecture-Whitepaper.pdf, accessed: December 2, 2016

[14] Nvidia, *CULA Tools (2015)*, Available: http://www.culatools.com/, accessed: January 13, 2015

[15] MAGMA, *Matrix Algebra on GPU and Multicore Architectures (2015)*, Available: http://icl.cs.utk.edu/projectsfiles/magma/doxygen/, accessed: January 13, 2017

[16] Nvidia, *CUDA Toolkit Documentation: CUBLAS (2015)*, Available: http://docs.nvidia.com/cuda/cublas/index.html, accessed: January 11, 2017

[17] Nvidia, *CuDNN*, Available: https://developer.nvidia.com/cudnn, accessed: March 22, 2017

[18] A. S. Garea, D. B. Heras, and F. Argüello, *GPU classification of remote sensing images using kernel ELM and extended morphological profiles*, International Journal of Remote Sensing, vol. 37, no. 24, pp. 5918-5935, 2016

[19] M. Fauvel, Y. Tarabalka, J.A. Benediktsson, J. Chanussot, and J.C. Tilton, *Advances in spectral-spatial classification of hyperspectral images*, Proceedings of the IEEE, vol. 101, no. 3, pp. 652-675, 2013

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

A Minimax Approach for the Study of Constrained Variational Equations

A.I. Garralda-Guillem[1] **and M. Ruiz Galán**[1]

[1] *Department of Applied Mathematics, University of Granada, Spain*

emails: agarral@ugr.es, mruizg@ugr.es

Abstract

We analyze the existence of a solution for a quite general variational inequalities system, which includes some variational problems with constraints. The fundamental tool to carry out this study is a minimax theorem. In addition, a stable numerical method is introduced for approximating the solution of such a system.

Key words: Minimax problems, vartiational inequalities.
MSC 2000: 49K35, 49A29.

1 Main results

Although originally, minimax inequalities arose in the context of game theory, they have turned out to be powerful tools in other fields: let us mention, for instance, [2, 3, 6, 7, 8, 9, 10, 11]. We start from the classical minimax inequality of von Neumann–Fan for studying the existence of a solution for a certain system of variational inequalities.

In order to introduce the class of systems that we are going to analyze, let us mention that the study of variational equations with constraints emerges naturally, among others, from the context of the elliptic boundary value problems, when their essential boundary conditions are treated as constraints in their standard variational formulation. For instance, let Ω be an open and bounded subset of $\mathbb{R}^n$ with a Lipschitz boundary $\partial\Omega$, and let $f_0 \in L^2(\Omega)$ and $g_0 \in H^{1/2}(\partial\Omega)$. Let us also assume that $h_0 \in L^\infty(\Omega)$ and that there exists $\delta > 0$ such that

$$\delta \leq h_0 \quad \text{in } \Omega.$$

ISBN: 978-84-617-8694-7

Let us consider the variational problem: find $x_0 \in H^1(\Omega)$ such that

$$x_0 = g_0 \text{ on } \partial\Omega \text{ and for all } x \in H_0^1(\Omega), \quad \int_\Omega h_0 \nabla x_0 \cdot \nabla x = \int_\Omega f_0 x,$$

that is, the weak formulation of the elliptic boundary problem associated with the Poisson equation with non-homogeneous Dirichlet boundary condition

$$\left\{ \begin{array}{rl} -\text{div}\,(h_0 \nabla x) = f_0 & \text{in } \Omega \\ x = g_0 & \text{on } \partial\Omega \end{array} \right. .$$

Equivalently, we can impose weakly the boundary condition: let $X := H^1(\Omega)$, $Y := H^{-1/2}(\partial\Omega)$, $Z := H_0^1(\Omega)$, let $a : X \times X \longrightarrow \mathbb{R}$ and $b : X \times Y \longrightarrow \mathbb{R}$ be the continuous bilinear forms

$$a(x_1, x_2) := \int_\Omega h_0 \nabla x_1 \cdot \nabla x_2, \qquad (x_1, x_2 \in X)$$

and

$$b(x, y) :=< \text{tr}(x), y >, \qquad (x \in X,\ y \in Y)$$

$< \cdot, \cdot >$ being the canonical bilinear form in $H^{1/2}(\Gamma) \times H^{-1/2}(\Gamma)$ and $\text{tr} : H^1(\Omega) \longrightarrow H^{1/2}(\Gamma) \subset L^2(\Gamma)$ the trace operator in $H^1(\Omega)$; and let $f : X \longrightarrow \mathbb{R}$ and $g : Y \longrightarrow \mathbb{R}$ be the continuous and linear functionals defined by

$$f(x) := \int_\Omega f_0 x, \qquad (x \in X)$$

and

$$g(y) :=< g_0, y >, \qquad (y \in Y).$$

Then, the variational formulation coincide with this constrained variational equation:

$$\text{find } x_0 \in X \text{ such that } \left\{ \begin{array}{ll} z \in Z & \Rightarrow f(z) = a(x_0, z) \\ y \in Y & \Rightarrow g(y) = b(x_0, y) \end{array} \right. .$$

In a more general way, we deal with the following problem: let E be a real reflexive Banach space, $n \in \mathbb{N}$ and suppose that for each $j = 1, \ldots, n$, F_j is a real normed space, $y_j^* \in F_j^*$ ("*" stands for "topological dual space"), C_j is a convex subset of F_j with $0 \in C_j$, and $a_j : E \times F_j \longrightarrow \mathbb{R}$ is a bilineal form satisfying

$$y_j \in C_j \Rightarrow a_j(\cdot, y_j) \in E^*;$$

find $x_0 \in E$ such that

$$\left\{ \begin{array}{c} y_1 \in C_1 \Rightarrow y_1^*(y_1) \leq a_1(x_0, y_1) \\ \cdots \\ y_n \in C_n \Rightarrow y_n^*(y_n) \leq a_n(x_0, y_n) \end{array} \right. .$$

 ISBN: 978-84-617-8694-7

As mentioned above, our fundamental tool is the von Neumann–Fan minimax theorem, which allows us to characterize the existence of a solution for this variational inequalities system in terms of that of a positive constant. Moreover, that minimax result implies the stability of numerical schemes of the Galerkin type for approximating the solution. The corresponding finite dimensional subspaces are generated from adequate biorthogonal systems depending on the concrete problem.

The variational systems under consideration are so general that include mixed variational formulations of some elliptic problems, those in the so-called *Babuška–Brezzi theory* (see, for instance [1, 5] and some of its generalizations [4]).

Finally, we illustrate our results with some numerical examples.

Acknowledgement

Research partially supported by project MTM2016-80676-P (AEI/FEDER, UE) and by Junta de Andalucía Grant FQM359.

References

[1] D. Boffi et al., *Mixed finite elements, compatibility conditions and applications*, Lecture Notes in Mathematics **1939**, Springer–Verlag, Berlin, 2008.

[2] J.M. Borwein and O. Giladi, *Some remarks on convex analysis in topological groups*, J. Convex. Anal. **23** (2016), 313–332.

[3] X.T. Deng, Z.F. Li and S.Y. Wang, *A minimax portfolio selection strategy with equilibrium*, European J. Oper. Res. **166** (2005), 278–292.

[4] A.I. Garralda Guillem and M. Ruiz Galán, *Mixed variational formulations in locally convex spaces*, J. Math. Anal. Appl. **414** (2014), 825–849.

[5] G.N. Gatica, *A simple introduction to the mixed finite element method. Theory and applications*, SpringerBriefs in Mathematics, Springer, Cham, 2014.

[6] P.Q. Khanh and N.H. Quan, *General existence theorems, alternative theorems and applications to minimax problems*, Nonlinear Anal. **72** (2010), 2706–2715.

[7] N. Kenmochi, *Monotonicity and compactness methods for nonlinear variational inequalities, Handbook of differential equations: stationary partial differential equations, IV*, 203–298, Elsevier/North-Holland, Amsterdam, 2007.

[8] Y. Polyanskiy, *Saddle point in the minimax converse for channel coding*, IEEE Trans. Inform. Theory **59** (2013), 2576–2595.

[9] M. Ruiz Galán, *The Gordan theorem and its implications for minimax theory*, J. Nonlinear Convex Anal. **17** (2016), 2385–2405.

[10] M. Ruiz Galán, *An intrinsic notion of convexity for minimax*, J. Convex Anal. **21** (2014), 1105–1139.

[11] S. Simons, *Minimax and monotonicity*, Lecture Notes in Mathematics **1693**, Springer–Verlag, Berlin, 1998.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

Nonclassical symmetries, potential symmetries and conservation laws of the generalized Drinfeld-Sokolov equations

T.M. Garrido[1], R. de la Rosa[1], E. Recio[1] and M.S. Bruzón[1]

[1] *Department of Mathematics, Faculty of Sciences, University of Cádiz, Spain*

emails: tamara.garrido@uca.es, rafael.delarosa@uca.es, elena.recio@uca.es, m.bruzon@uca.es

Abstract

In the present work we study the generalized Drinfeld-Sokolov equations. Previously, we got its classical Lie symmetries and obtained exact solutions. Now we continue the analysis by applying the nonclassical method proposed by Bluman and Cole to deduce new symmetries of the equation. In addition, we look for nonlocal symmetries by applying the potential symmetry method introduced by Bluman *et al.* Finally, we obtain conservation laws depending on the parameters by using the multiplier method.

Key words: generalized Drinfeld-Sokolov, nonclassical symmetries, potential symmetries, conservation laws.

1 Introduction

Nowadays Drinfeld-Sokolov equations are getting a lot of attention because of their extensive applications. Initially they were presented by Drinfeld and Sokolov in [8] as some examples of generalized KdV and mKdV corresponding to classical Kats-Moody algebras, and these days we have organised them in two categories.

The first one is the Drinfeld–Sokolov–Wilson (DSW) system

$$\begin{cases} u_t + (v^n)_x & = 0 \\ v_t - av_{xxx} + 3bu_xv + 3kuv_x & = 0 \end{cases}$$

where a, b, k and n are constants. This system was introduced by Drinfeld and Sokolov as an example of a system of nonlinear equations possessing Lax pairs of a special form [10]

and it is used to describe nonlinear surface gravity waves propagating over an horizontal sea bed. It has been extensively studied in [3, 11, 16] and so on.

The second one is formed by the generalized Drinfeld-Sokolov equations (gDS) [14]

$$\begin{cases} u_t + \alpha_1 u u_x + \beta_1 u_{xxx} + \gamma \left(v^\delta\right)_x & = 0 \\ V_t + \alpha_2 u v_x + \beta_2 V_{xxx} & = 0 \end{cases} \tag{1}$$

where $\alpha_1, \alpha_2, \beta_1, \beta_2, \gamma$ and δ are constants. This system models one dimensional nonlinear wave processes in two-component media.

In a previous paper, Garrido and Bruzón [9] studied this couple of equations (1) obtaining its classical lie symmetries classification, reductions, and few travelling wave solutions applying the sine–cosine method introduced by Wazwaz [15]. So the aim of the present paper is to continue with the previous research of (1) extending it to the study of nonclassical symmetries, potential symmetries and conservation laws.

There are several reasons to be interested in the study of symmetries admitted by a partial differential equation and probably the most important is that they are useful for finding invariant solutions. Lie group theory provides a method to search for these special group invariant solutions [12]. However, sometimes not all of these invariant solutions can be found by using the Lie classical method and that is why we have looked for the nonclassical and potential symmetries of (1) too.

Bluman and Cole [4] proposed the so–called nonclassical method of group-invariant solutions in which since the number of determining equations is smaller, the set of solutions is larger than the one for the classical method. In the same way in [5] Bluman et al. have introduced the concept of potential symmetry, which is a nonlocal symmetry, for any differential equation which can be written as a conservation law.

To conclude and due to the physical importance of conservation laws in the study of properties that do not change in the course of time, especially for determining conserved quantities and constants of motion, we have applied the multiplier method proposed by Bluman and Anco [1, 2] which provides a general treatment to find all local conservation laws admitted by any given evolution equation. Some examples can be found in [6, 7, 13].

Acknowledgements

This work has been partially supported by University of Cádiz.

References

[1] S. C. Anco and G. W. Bluman, *Direct construction method for conservation laws of partial differential equations part 1: Examples of conservation law classifications*,

European Journal of Applied Mathematics **5** (2002) 545–566.

[2] S. C. Anco and G. W. Bluman, *Direct construction method for conservation laws of partial differential equations part 2: General treatment*, European Journal of Applied Mathematics **5** (2002) 567–585.

[3] R. Arora and A. Kumar, *Solution of the Coupled Drinfeld's-Sokolov-Wilson (DSW) System by Homotopy Analysis Method*, Advanced Science, Engineering and Medicine **5** (2013) 1–7.

[4] G. W. Bluman and J. Cole, *General similarity solution of the heat equation*, Journal of Mathematics and Mechanics **18** (1969) 1025–1042.

[5] G. W. Bluman, S. Kumei and G. J. Reid, *New classes of symmetries for partial differential equations*, Journal of Mathematics and Mechanics **29** (1988) 806–811.

[6] J. C. Camacho and M. S. Bruzón, *Exact solutions and conservation laws of a generalized Fornberg–Whitham equation*, Proceedings of the 16th International Conference on Computational and Mathematical Methods in Science and Engineering Volume 1, Rota, 2016.

[7] R. de la Rosa , M. L. Gandarias and M. S. Bruzón, *An study for the Microwave Heating of a Half–Space through Lie symmetries and conservation laws*, Proceedings of the 14th International Conference on Computational and Mathematical Methods in Science and Engineering Volume 2, Rota, 2014.

[8] V. G. Drinfeld and V. V. Sokolov, *Lie algebras and equations of Korteweg-de Vries type*, Journal of Soviet Mathematics **30(2)** (1985) 1975–2036.

[9] T. M. Garrido and M. S. Bruzón, *Lie Point Symmetries and Travelling Wave Solutions for the Generalized Drinfeld–Sokolov System*, Journal of Computational and Theoretical Transport **45(4)** (2016) 290–298.

[10] U. Goktas and F. Hereman, *Symbolic computation of conserved densities for systems of nonlinear evolution equations*, Journal of Symbolic Computation **24(5)** (1997) 591–622.

[11] R. Naz, *Conservation laws for a complexly coupled KdV system, coupled Burgers' system and Drinfeld-Sokolov-Wilson system via multiplier approach*, Communications in Nonlinear Science and Numerical Simulation **15(5)** (2010) 1177–1182.

[12] P. J. Olver, *Applications of Lie groups to differential equations*, 2nd ed. Graduate Texts in Mathematics, Springer–Verlag, Berlin, 1993.

 ISBN: 978-84-617-8694-7

[13] E. Recio and S. C. Anco, *Conservation laws and symmetries of radial generalized nonlinear p–Laplacian evolution equations*, Journal of Mathematical Analysis and Applications **452** (2017) 1229–1261.

[14] E. Sweet and R. A. Van Gorder, *Analytical solutions to a generalized Drinfel'd–Sokolov equation related to DSSH and KdV6*, Applied Mathematics and Computation **216(10)** (2010) 2783 - 2791.

[15] A. M. Wazwaz, *A sine-cosine method for handling nonlinearwave equations*, Mathematical and Computer Modelling,**40** (2004) 499–508.

[16] Z. Zhao, Y. Zhang and Z. Han, *Symmetry analysis and conservation laws of the Drinfeld-Sokolov-Wilson system*, The European Physical Journal Plus **129(7)** (2014) 1–7.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

Solving second order non-linear parabolic pde's using generalized finite difference method (GFDM)

L.Gavete[1], F.Ureña[2], J.J.Benito[2] and A.García[2]

[1] *Departamento de Matemática Aplicada a los Recursos Naturales, Universidad Politécnica de Madrid (UPM)*

[2] *Departamento de Construcción y Fabricación, Universidad Nacional de Educación a Distancia (UNED)*

emails: lu.gavete@upm.es, fuprieto@terra.com, jbenito@ind.uned.es, angelochurri@gmail.com

Abstract

The generalized finite difference method (GFDM) has been proved to be a good meshless method to solve several linear partial differential equations (pde's): wave propagation, advection-diffusion, plates, beams, etc.
The GFDM allows us to use irregular clouds of nodes that can be of interest for modelling non-linear parabolic pde's.
This paper illustres that the GFD explicit formulae developed to obtain the different derivatives of the pde's are based in the existence of a positive definite matrix that it is obtained using moving least squares approximation and Taylor series development. Criteria for convergence of fully explicit method using GFDM for different non linear parabolic pdes are given.
This paper shows the application of the GFDM to solving different non-linear problems including applications to heat transfer, acoustics and problems of mass transfer.

Key words: meshless methods, generalized finite difference method, non-linear parabolic partial differential equations.

1 Introduction

Modern numerical methods, in particular those for solving non-linear PDEs, have been developed in recent years using finite differences, finite elements, finite volume or spectral

methods. A review of numerical methods for non-linear partial differential equations is given by Polyanin [1] and Tadmor [2]. In this paper we use a meshless method called generalizad finite difference method (GFDM) for solving different parabolic non-linear pdes.
Benito, Gavete and Ureña [3,4,5] have developed the explicit formulae and h-adaptive method for the solution of the pdes in 2-D.
In this paper, this meshless method is used for solving non-linear parabolic partial differential equations in 2-D. Parabolic equations have been solved using an explicit method and the convergence has been studied taking into account the irregularity of the cloud of points. The numerical results show the high accuracy obtained.
The paper is organized as follows. In section 2 the explicit method and GFDM to solve non-linear parabolic partial differential are showed. In section 3, convergence is studied. Sections 4 exposes the results obtained for solving different non-linear problems. Finally, in section 5, some conclusions are obtained.

2 Explicit method and GFDM: application to non-linear parabolic partial differential equations

Consider the following non-linear problem in the domain $D = [0, T] \times \Omega$ with $\Omega \subset \mathbb{R}^2$

$$\frac{\partial U}{\partial t} = L_\Omega[U] \tag{1}$$

where $L_\Omega[U]$ is a non-linear operator, Γ is the boundary of the domain Ω, with boundary condition:

$$U_\Gamma = f(t) \tag{2}$$

and initial condition

$$U(x, y, 0) = g(x, y) \tag{3}$$

where f and g are two known functions.
To solve the problem described by Eqs.(1), (2) and (3), using the explicit method, time derivative is approximated by

$$\frac{\partial u(x_0, y_0, n\triangle t)}{\partial t} = \frac{u_0^{n+1} - u_0^n}{\triangle t} + \Theta(\triangle t) \tag{4}$$

 ISBN: 978-84-617-8694-7

and spatial derivatives are approximated using GFD [3,4,5], are denoted by

$$\begin{cases} \dfrac{\partial u(x_0,y_0,n\triangle t)}{\partial x} = -\lambda_0 u_0 + \displaystyle\sum_{i=1}^{s} \lambda_i u_i + \Theta(h_i^2,k_i^2),\ with\ \lambda_0 = \sum_{i=1}^{s} \lambda_i \\ \dfrac{\partial u(x_0,y_0,n\triangle t)}{\partial y} = -\mu_0 u_0 + \displaystyle\sum_{i=1}^{s} \mu_i u_i + \Theta(h_i^2,k_i^2),\ with\ \mu_0 = \sum_{i=1}^{s} \mu_i \\ \dfrac{\partial^2 u(x_0,y_0,n\triangle t)}{\partial x^2} + \dfrac{\partial^2 u(x_0,y_0,n\triangle t)}{\partial y^2} = -m_0 u_0 + \displaystyle\sum_{i=1}^{s} m_i u_i + \Theta(h_i^2,k_i^2) \\ with\ m_0 = \displaystyle\sum_{i=1}^{s} m_i \end{cases} \tag{5}$$

3 Convergence of the scheme for non-linear parabolic pde's

In this section convergence of non-linear parabolic pde's, using GFDM, is studied. We will do so by introducing the following definitions:

- A partial differential equation is **semilinear** if the coefficients of its highest derivatives are functions of the spaces variables only.
- A partial differential equation is **quasi-linear** if it is linear in its highest derivatives.

In this paper, the following cases are considered for $L_\Omega[U]$ of Eq.(1), for their importance in applied science and technology:

Semilinear

$$L_\Omega[U] = \frac{\partial^2 U}{\partial x^2} + \frac{\partial^2 U}{\partial y^2} + F(U) \tag{6}$$

Quasilinear

$$L_\Omega[U] = K(U)(\frac{\partial^2 U}{\partial x^2} + \frac{\partial^2 U}{\partial y^2}) \tag{7}$$

The method which has been used consists in finding an error bound directly from the discretized expressions and, therefore, the concept of stability is not used. However this concept appears when we consider that the error must be limited.

3.1 Semilinear parabolic PDEs

Let us consider Eq.(6) where $F(U)$ is a differentiable function. By considering the non-linear scheme together with the explicit expressions Eqs.(4) and Eq.(5)

$$\frac{u_0^{n+1} - u_0^n}{\Delta t} = -m_0 u_0^n + \sum_{i=1}^{N} m_i u_i^n + F(u_0^n) \tag{8}$$

 ISBN: 978-84-617-8694-7

and let the same expression for the exact solution be

$$\frac{U_0^{n+1}-U_0^n}{\Delta t}=-m_0U_0^n+\sum_{i=1}^{N}m_iU_i^n+F(U_0^n) \tag{9}$$

If we define the error as $e_i^n=u_i^n-U_i^n$ y $e_0^n=u_0^n-U_0^n$, using the mean value theorem for F , it is obtained

$$F(u_0^n)-F(U_0^n)=e_0^n\frac{\partial F}{\partial U}(\chi_j) \tag{10}$$

where $\chi_j=U_0^n+\xi e_0^n,\quad \xi\in[0,1]$ Subtracting Eqs.(8) and (9) and introducing Eq.(10) it is obtained

$$\frac{e_0^{n+1}-e_0^n}{\Delta t}=-m_0e_0^n+\sum_{i=1}^{N}m_ie_i^n+e_0^n\frac{\partial F}{\partial U}(\chi_j) \tag{11}$$

$$e_0^{n+1}=(1-m_0\Delta t+\Delta t\frac{\partial F}{\partial U}(\chi_j))e_0^n+\Delta t\sum_{i=1}^{N}m_ie_i^n \tag{12}$$

$$|e_0^{n+1}|\leq|(1-m_0\Delta t+\Delta t\frac{\partial F}{\partial U}(\chi_j))||e_0^n|+\Delta t|\sum_{i=1}^{N}m_ie_i^n| \tag{13}$$

Let $e^n=\max_{i=0,..,N}|e_i^n|$, taking into account eq.(4) and eq.(1) it is obtained

$$e^{n+1}\leq e^n(|1-m_0\Delta t+\Delta t\frac{\partial F}{\partial U}|+\sum_{i=1}^{N}\Delta t|m_i|)+\Theta_1(\Delta t(h_i^2+k_i^2+\Delta t)) \tag{14}$$

$$\alpha=|1-m_0\Delta t+\Delta t\frac{\partial F}{\partial U}|+\sum_{i=1}^{N}\Delta t|m_i| \tag{15}$$

by considering $n=1$ and including a constant C

$$e^1=\alpha e^0+C(\Delta t(h_i^2+k_i^2+\Delta t)) \tag{16}$$

as $e^0=0$ by the initial condition Eq.(3)

$$e^1=C(\Delta t(h_i^2+k_i^2+\Delta t)) \tag{17}$$

and similarly to eq.(17)

$$e^2=\alpha e^1+C(\Delta t(h_i^2+k_i^2+\Delta t))=C(\Delta t(h_i^2+k_i^2+\Delta t))(1+\alpha) \tag{18}$$

$$e^{n+1}=C(\Delta t(h_i^2+k_i^2+\Delta t))(1+\alpha+\alpha^2+\cdots+\alpha^n) \tag{19}$$

 ISBN: 978-84-617-8694-7

$1+\alpha+\alpha^2+\cdots+\alpha^n+\cdots$ is a geometric series, where the condition of convergence is $|\alpha|<1$. Hence, in order for the error not to diverge as time increment tends to 0, it must be:

$$|\alpha|<1 \Leftrightarrow |1-m_0\Delta t+\Delta t\frac{\partial F}{\partial U}|+\sum_{i=1}^{N}\Delta t|m_i|<1 \tag{20}$$

Thus,

$$|1-m_0\Delta t+\Delta t\frac{\partial F}{\partial U}|<1-\sum_{i=1}^{N}\Delta t|m_i| \tag{21}$$

$$-1+\sum_{i=1}^{N}\Delta t|m_i|<1-m_0\Delta t+\Delta t\frac{\partial F}{\partial U}<1-\sum_{i=1}^{N}\Delta t|m_i| \tag{22}$$

by comparing first and second terms of inequality eq.(29) and second and third terms the following two conditions can be obtained.

$$\left\{\begin{array}{l} 1-m_0\Delta t+\Delta t\frac{\partial F}{\partial U}<1-\sum_{i=1}^{N}\Delta t|m_i| \\ -1+\sum_{i=1}^{N}\Delta t|m_i|<1-m_0\Delta t+\Delta t\frac{\partial F}{\partial U}\end{array}\right. \Leftrightarrow \left\{\begin{array}{l} \Delta t(\frac{\partial F}{\partial U}+\sum_{i=1}^{N}|m_i|-m_0)<0 \\ \Delta t(\sum_{i=1}^{N}|m_i|+m_0-\frac{\partial F}{\partial U})<2\end{array}\right. \tag{23}$$

as $\Delta t>0$, the criteria of convergence are

$$\left\{\begin{array}{l} \frac{\partial F}{\partial U}+\sum_{i=1}^{N}|m_i|-m_0<0 \\ \Delta t<\frac{2}{\sum_{i=1}^{N}|m_i|+m_0-\frac{\partial F}{\partial U}}\end{array}\right. \tag{24}$$

From first inequality of eq.(24), taking account eq.(5) thus $\frac{\partial F}{\partial U}<0$ is obtained. The second inequality of eq.(24) gives us the limit Δt of convergence of each one of the stars of the domain. Then the minimum value obtained between all the stars is taken as limit of convergence.

3.2 Quasilinear parabolic PDEs

Let us consider eq.(7) representing an heat transmission case where U is the temperature and $K(U)$ is the conductivity.
Using fully explicit approximations

$$\frac{u_0^{n+1}-u_0^n}{\Delta t}=K(u_0^n)(-m_0u_0^n+\sum_{i=1}^{N}m_iu_i^n) \tag{25}$$

and similarly for the exact solution

$$\frac{U_0^{n+1}-U_0^n}{\Delta t}=K(U_0^n)(-m_0U_0^n+\sum_{i=1}^{N}m_iU_i^n) \tag{26}$$

Defining the error $e^n = u^n - U^n$, it is obtained

$$\frac{e_0^{n+1} - e_0^n}{\Delta t} = K(u_0^n)(-m_0 u_0^n + \sum_{i=1}^{N} m_i u_i^n) - K(U_0^n)(-m_0 U_0^n + \sum_{i=1}^{N} m_i U_i^n) \tag{27}$$

$$\begin{aligned} e_0^{n+1} = e_0^n + \Delta t[K(u_0^n)(-m_0 u_0^n + \sum_{i=1}^{N} m_i u_i^n) - K(u_0^n)(-m_0 U_0^n + \sum_{i=1}^{N} m_i U_i^n) \\ + K(u_0^n)(-m_0 U_0^n + \sum_{i=1}^{N} m_i U_i^n) - K(U_0^n)(-m_0 U_0^n + \sum_{i=1}^{N} m_i U_i^n)] \end{aligned} \tag{28}$$

and using the mean value theorem

$$e_0^{n+1} = e_0^n + \Delta t[K(u_0^n)(-m_0 e_0^n + \sum_{i=1}^{N} m_i e_i^n) + e_0^n \frac{\partial K}{\partial U}(\chi_j)(-m_0 U_0^n + \sum_{i=1}^{N} m_i U_i^n)] \tag{29}$$

and rearranging

$$e_0^{n+1} = e_0^n[1 + \Delta t \frac{\partial K}{\partial U}(-m_0 U_0^n + \sum_{i=1}^{N} m_i U_i^n) - m_0 \Delta t K(u_0^n)] + \Delta t K(u_0^n) \sum_{i=1}^{N} e_i^n m_i \tag{30}$$

If, again, we make $e^n = \max_{i=0,..,N} |e_i^n|$ in every point of the cloud of points, we shall obtain the relation:

$$\begin{aligned} e^{n+1} \leq e^n(|1 + \Delta t \frac{\partial K}{\partial U}(-m_0 U_0^n + \sum_{i=1}^{N} m_i U_i^n) - m_0 \Delta t K(u_0^n)| + \Delta t |K(u_0^n)| \sum_{i=1}^{N} |m_i|) \\ + \Theta_3(\Delta t(h_i^2 + k_i^2 + \Delta t)) \end{aligned} \tag{31}$$

Denoting:

$$R = \Delta t \frac{\partial K}{\partial U}(-m_0 U_0^n + \sum_{i=1}^{N} m_i U_i^n) \tag{32}$$

$$\gamma = |1 + R\Delta t - m_0 \Delta t K(u_0^n)| + \Delta t |K(u_0^n)| \sum_{i=1}^{N} |m_i| \tag{33}$$

$$e^{n+1} \leq e^n \gamma + \Theta_3(\Delta t(h_i^2 + k_i^2 + \Delta t)) \tag{34}$$

Thus,

$$e^1 \leq e^0 \gamma + C_2(\Delta t(h_i^2 + k_i^2 + \Delta t)) = C_2(\Delta t(h_i^2 + k_i^2 + \Delta t)) \tag{35}$$

 ISBN: 978-84-617-8694-7

where $e^0 = 0$ by initial condition eq.(3)

$$e^n \leq C_2(\Delta t(h_i^2 + k_i^2 + \Delta t))(1 + \gamma + \cdots) \tag{36}$$

Taking account that $1+\gamma+\gamma^2+\cdots$ is a geometric series, then the condition of convergence is $|\gamma| < 1$.

$$|1 + R\Delta t - m_0\Delta t K(u_0^n)| + \Delta t|K(u_0^n)| \sum_{i=1}^{N} |m_i| < 1 \tag{37}$$

$$-1 + \Delta t|K(u_0^n)| \sum_{i=1}^{N} |m_i| < 1 + R\Delta t - m_0\Delta t K(u_0^n) < 1 - \Delta t|K(u_0^n)| \sum_{i=1}^{N} |m_i| \tag{38}$$

by relating second and third terms and first and second ones

$$\begin{cases} 1 + R\Delta t - m_0\Delta t K(u_0^n) < 1 - \Delta t|K(u_0^n)| \sum_{i=1}^{N} |m_i| \\ -1 + \Delta t|K(u_0^n)| \sum_{i=1}^{N} |m_i| < 1 + R\Delta t - m_0\Delta t K(u_0^n) \end{cases} \tag{39}$$

then the following two inequalities are obtained

$$\begin{cases} \Delta t(R + |K(u_0^n)| \sum_{i=1}^{N} |m_i| - m_0 K(u_0^n)) < 0 \\ \Delta t(|K(u_0^n)| \sum_{i=1}^{N} |m_i| + K(u_0^n)m_0 - R) < 2 \end{cases} \tag{40}$$

and operating in eq.(40)

$$\begin{cases} -R + m_0 K(u_0^n) > |K(u_0^n)| \sum_{i=1}^{N} |m_i| \\ \Delta t < \dfrac{2}{|K(u_0^n)| \sum_{i=1}^{N} |m_i| + m_0 K(u_0^n) - R} \end{cases} \tag{41}$$

by taking first inequality of the eq.(41) and substituting it in the second one

$$\begin{cases} -R + m_0 K(u_0^n) > |K(u_0^n)| \sum_{i=1}^{N} |m_i| \\ \Delta t < \dfrac{1}{m_0 K(u_0^n) - R} \end{cases} \tag{42}$$

As it is shown in the second inequality of eq.(42) the values of the coefficients (m_o, m_i) are important to determine the Δt for each one of the stars. The limit of convergence Δt for each cloud of nodes is the minimum value obtained considering all the stars of the domain.

4 Numerical results

In this section it will be shown the numerical results obtained by solving five non-linear parabolic pde with Dirichlet boundary conditions the initial conditions (which are obtained

 ISBN: 978-84-617-8694-7

by substitution in the exact solution), which correspond to non-linear problems in physics and engineering: heat equation, diffusion of liquids in porous medium, advection-diffusion, etc.

Firstly we consider 3 examples and apply them in 6 different discretizations (see figure 1), obtaining the solution for $t = 0.25$ sec. To do it we divide the total time into a different number of steps, in particular, 50, 100, 250, 500 and 1000 steps, which is equivalent to use as time increments 0.005, 0.0025, 0.001, 0.0005 and 0.00025 sec. respectively.

In all the cases considered the global error has been calculated according to the formula

$$\%\text{Global Error } = \frac{\sqrt{\sum_{i=1}^{NI}\Big(sol(i) - exac(i)\Big)^2}}{|exac_{max}|} \times 100 \tag{43}$$

where $sol(i)$ is the GFD solution in node i, $exac(i)$ is the exact value of the solution at node i, $exac_{max}$ is the maximum value of the exact values in the cloud of nodes considered and NI is the number of nodes of the domain Ω.

Note that in the numerical results the time step used, Δt, have been selected according to

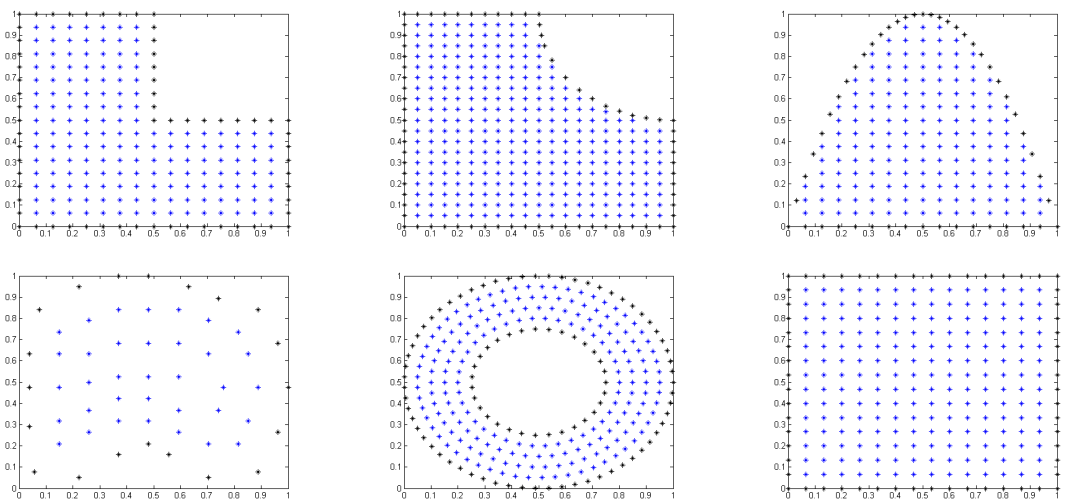


Figure 1: Clouds of points 1, 2, 3, 4 , 5 and 6

the convergence limits defined in eq.(24) and eq.(42). However these limits depend of the stars of nodes, and hence they are different for the six clouds of points considering in fig.1 Then as consequence only the converging cases have been considered in tables [1-3].

 ISBN: 978-84-617-8694-7

Example 1

Let us consider the equation

$$\frac{\partial U}{\partial t} = \frac{\partial^2 U}{\partial x^2} + \frac{\partial^2 U}{\partial y^2} - Uln(U) \tag{44}$$

with exact solution:

$$U(x, y, t) = e^{e^{-t}} e^{\frac{x^2}{5} + \frac{xy}{5} + \frac{y^2}{20} + \frac{1}{2}} \tag{45}$$

This example corresponds to a semilinear case whose convergence has been studied in previous section.

Time increments $\triangle t$ calculated by using the rate of convergence for each mesh are,

mesh \ $\triangle t$	0.005	0.0025	0.001	0.0005	0.00025
1	-	-	$2.7582 \cdot 10^{-5}$	$8.2719 \cdot 10^{-6}$	$2.4258 \cdot 10^{-5}$
2	-	-	$2.7802 \cdot 10^{-3}$	$2.0688 \cdot 10^{-5}$	$6.8425 \cdot 10^{-6}$
3	-	-	$5.5248 \cdot 10^{-5}$	$2.7699 \cdot 10^{-5}$	$2.5353 \cdot 10^{-5}$
4	$3.8457 \cdot 10^{-4}$	$1.6820 \cdot 10^{-4}$	$1.2141 \cdot 10^{-4}$	$4.4682 \cdot 10^{-1}$	$1.5638 \cdot 10^{-4}$
5	-	-	$1.6033 \cdot 10^{-5}$	$1.0019 \cdot 10^{-5}$	$9.2320 \cdot 10^{-6}$
6	-	-	$4.1119 \cdot 10^{-5}$	$2.3447 \cdot 10^{-5}$	$5.3133 \cdot 10^{-5}$

Table 1: Table of mean squared errors for Example 1

respectively: 0.0022, 0.00125, 0.001, 0.0078, 0.00125 and 0.0022. Thus, the numerical results obtained Table 1 and Fig. 2 are according with the limit previously developed.

Example 2

Let us consider the equation

$$\frac{\partial U}{\partial t} = U(\frac{\partial^2 U}{\partial x^2} + \frac{\partial^2 U}{\partial y^2}) \tag{46}$$

with the exact solution:

$$U(x, y, t) = \frac{1}{1-t}(\frac{x^2 + y^2 + 1}{4} + xy + x + y) \tag{47}$$

This case corresponds to a heat transmission problem where the conductivity K depends on the temperature U.
This example corresponds to a quasilinear whose convergence has been studied in previous section.

$\triangle t$ has been calculated by using the rate of convergence for each cloud of points

 ISBN: 978-84-617-8694-7

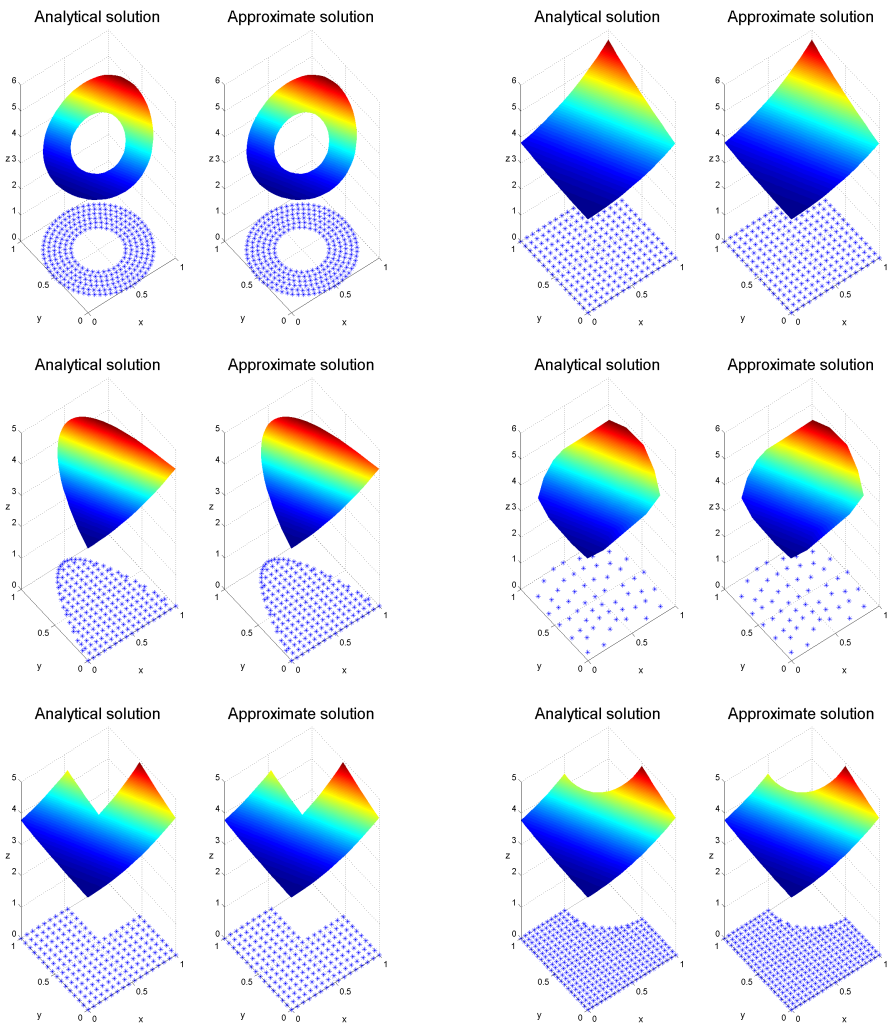


Figure 2: Exact and approximated solutions of the example 1 on meshes 1, 2, 3, 4 5 and 6

being, respectively: 0.00088, 0.00035, 0.00067, 0.0038, 0.00076 and 0.00088. Thus, the numerical results obtained Table 2 and Fig. 3 are according with the convergence theory, previously developed.

Example 3

 ISBN: 978-84-617-8694-7

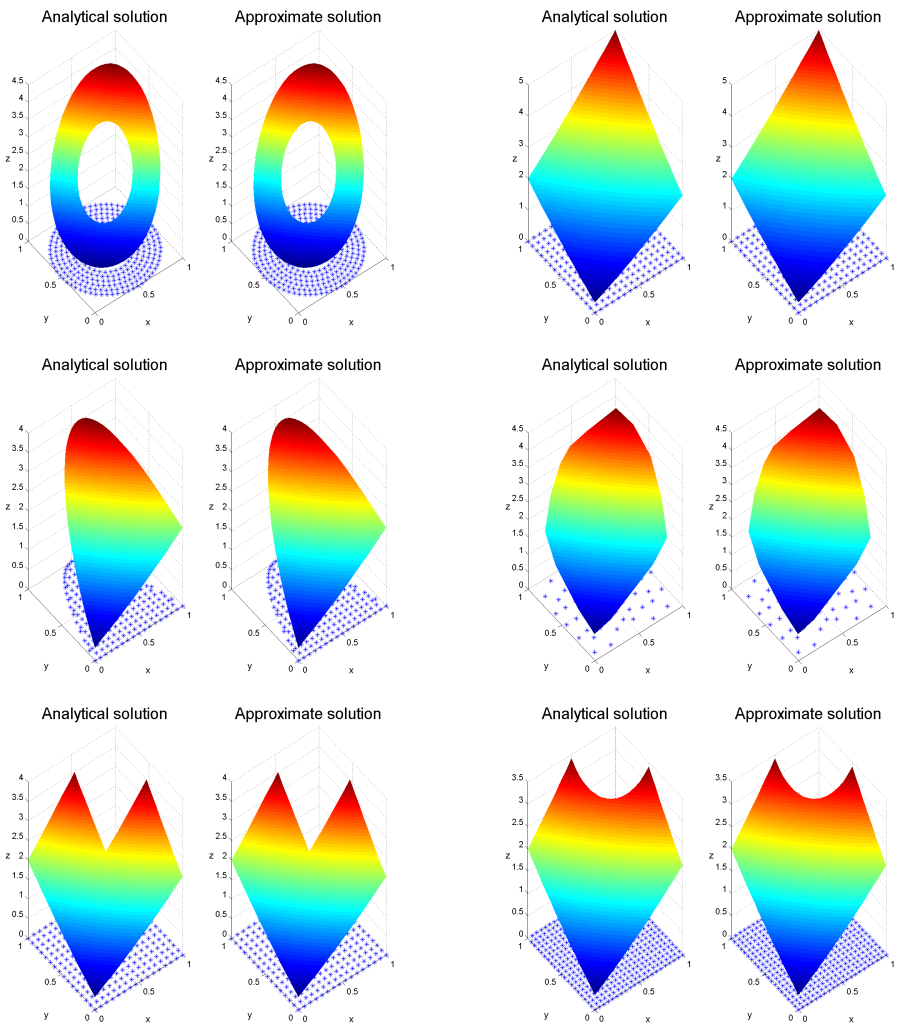


Figure 3: Exact and approximated solutions of the example 2 on clouds of points 1, 2, 3, 4, 5 and 6

Let us consider the equation

$$\frac{\partial U}{\partial t} = U(\frac{\partial^2 U}{\partial^2 x} + \frac{\partial^2 U}{\partial^2 y}) + U \tag{48}$$

with the exact solution:

$$U(x, y, t) = e^t(1 + e^x \sin(y)) \tag{49}$$

 ISBN: 978-84-617-8694-7

mesh \ $\triangle t$	0.005	0.0025	0.001	0.0005	0.00025
1	-	-	-	$2.1974 \cdot 10^{-5}$	$1.1149 \cdot 10^{-5}$
2	-	-	-	-	$1.3456 \cdot 10^{-5}$
3	-	-	-	$9.8553 \cdot 10^{-2}$	$1.2612 \cdot 10^{-5}$
4	-	$1.5380 \cdot 10^{-4}$	$6.1535 \cdot 10^{-5}$	$3.0915 \cdot 10^{-5}$	$1.5726 \cdot 10^{-5}$
5	-	-	-	$9.8553 \cdot 10^{-2}$	$1.2612 \cdot 10^{-5}$
6	-	-	-	$3.8040 \cdot 10^{-5}$	$1.9095 \cdot 10^{-5}$

Table 2: Table of mean squared errors for Example 2

This example corresponds to a quasilinear whose convergence has been studied in previous section.

mesh \ Δt	0.005	0.0025	0.001	0.0005	0.00025
1	-	-	$2.2895 \cdot 10^{-1}$	$1.3241 \cdot 10^{-5}$	$1.0168 \cdot 10^{-5}$
2	-	-	-	$1.497 \cdot 10^{-1}$	$1.3643 \cdot 10^{-5}$
3	-	-	-	$4.9862 \cdot 10^{-5}$	$4.7147 \cdot 10^{-5}$
4	-	$2.1002 \cdot 10^{-4}$	$1.9940 \cdot 10^{-4}$	$1.9655 \cdot 10^{-4}$	$1.9527 \cdot 10^{-4}$
5	-	-	-	$3.8149 \cdot 10^{-1}$	$2.1619 \cdot 10^{-5}$
6	-	-	$5.2962 \cdot 10^{-1}$	$2.838 \cdot 10^{-5}$	$2.2815 \cdot 10^{-5}$

Table 3: Table of mean squared errors for Example 3

$\triangle t$ computed by using the rate of convergence for each mesh are, respectively: 0.0017, 0.00065, 0.00062, 0.0031, 0.00072 and 0.0017. Thus, the numerical results obtained Table 3 and Fig 4 are according with the convergence, previously developed.

5 Conclusions

In this paper explicit and implicit methods using GFDM for solving different non-linear parabolic PDEs, have been considered.
Convergence has been studied, for semilinear and quasilinear equations, and limits of convergence have been developed and implemented in different examples.
The examples provided illustrates the viability of the application of GFDM for solving parabolic non-linear PDEs in 2D. The efficiency of the developed methods is clearly shown.
The accuracy of the GFDM has been tested in different non-linear PDEs, including different cases related with acoustics, heat transfer, mass transfer, heat extinction, combustion.
Numerical results for several non-linear problems validate the use of GFDM to solve this type of practical problems.

 ISBN: 978-84-617-8694-7

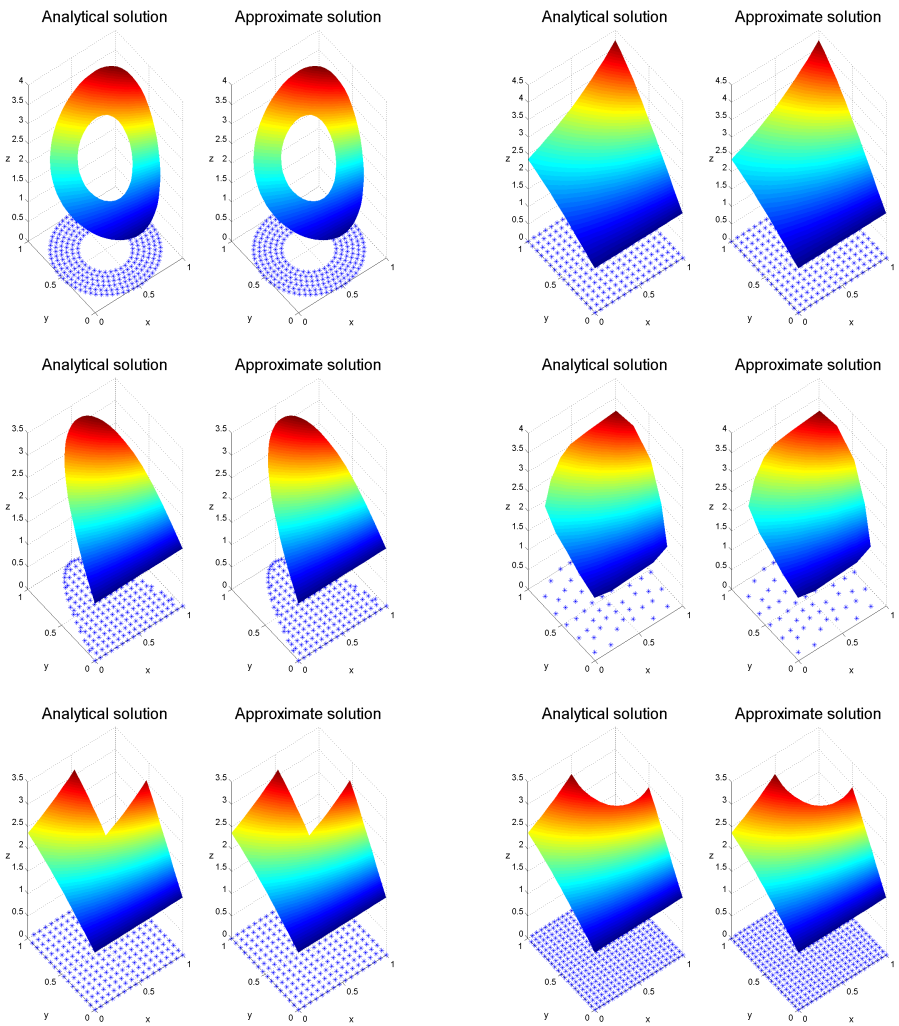


Figure 4: Exact and approximated solutions of the example 3 on clouds of points 1, 2, 3, 4, 5 and 6

Acknowledgements

The authors acknowledge the support of the Escuela Técnica Superior de Ingenieros Industriales (UNED) of Spain, project 2017-IFC02, and of the Technical University of Madrid (Research groups 2017).

References

[1] A. D. Polyanin, V. F. Zaitsev, *Handbook of nonlinear partial differential equations*, Chapman & Hall/CRC, ISBN 1-58488-355-3.

[2] A. Tadmor, *A review of numerical methods for non-linear partial differential equations*, Bulletin of the American Mathematical Society 2012; 42(4):507-554.

[3] F. Ureña, J. J. Benito, L. Gavete, *Application of the generalized finite difference method to solve the advection-diffusion equation*, Journal of Computational and Applied Mathematics **235** (2011) 1849–1855

[4] J. J. Benito, F. Ureña, L. Gavete, *Solving parabolic and hyperbolic equations by the generalized finite difference method*, Journal of Computational and Applied Mathematics **209** (2007) 208–233.

[5] L. Gavete, F. Ureña, J. J. Benito, A. Garcia, M. Ureña, E. Salete, *Solving second order non-linear elliptic partial differential equations using generalized finite difference method*, Journal of Computational and Applied Mathematics, **318** (2017) 378–387.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

Rational Interpolation, Newton Correction and Zero-Finding Methods

Luca Gemignani[1]

[1] *Dipartimento di Informatica, Università di Pisa, Pisa, Italy*

emails: luca.gemignani@unipi.it

Abstract

In this talk we discuss the numerical properties of structured linearizations formed by Bezoutian-like matrices as well as of their rank structured counterparts generated from the values of the Newton correction applied for the solution of a nonlinear equation. Numerical results are presented and some open questions are listed.

1 Introduction

We consider the problem of estimating zeros of an analytic function $f:\Omega \subset \mathbb{C} \to \mathbb{C}$ from the values of a related rational function at sample points. Approaches based on interpolation techniques are known for this problem [11]. Recently, the interest on these methods has been renewed in [1], where approximations of the zeros of $f(z)$ inside the unit circle are computed by solving certain structured generalized eigenvalue problems directly constructed from the values attained by $f(z)$ at the roots of unity. Efficient eigensolvers based on rank-structured matrix algorithms are presented in [7, 5, 4]. Differently, the associated linearized eigenvalue problems might be solved using the Ehrlich-Aberth method [3, 8] for simultaneous polynomial root-finding. Numerical experiments carried out in [7, 6] show that the weakness of the resulting procedure lies in the evaluation of $f(z)$ at the interpolation points which can be prone to round-off errors. Approximation schemes based on computing the ratio $f(z)/f'(z)$ –generally referred as the Newton correction of $f(z)$– are numerically more reliable. It is an immediate observation that the function value and the derivative might overflow/underflow while the ratio may still be a reasonable machine number. In this talk we elaborate upon the construction of structured linearizations of the zerofinding problem using the values of the Newton correction.

ISBN: 978-84-617-8694-7

More precisely, for a given fixed $n \in \mathbb{N}$ let $\mathcal{Z}^{(1)} = \{z_j\}_{j=1}^n$ and $\mathcal{Z}^{(2)} = \{w_j\}_{j=1}^n$ be two disjoint sets of pairwise distinct points. Let us consider rational approximations of the meromorphic function $N(z) = f'(z)/f(z)$, which gives the reciprocal of the Newton correction applied to $f(z)$. The rational interpolant $r(z) = p(z)/q(z)$ of of type $(n-1, n)$ on $\mathcal{Z} = \mathcal{Z}^{(1)} \cup \mathcal{Z}^{(2)}$ is defined by the conditions

$$\frac{p(t_j)}{q(t_j)} = N(t_j) = \frac{f'(t_j)}{f(t_j)}, \quad t_j \in \mathcal{Z}; \quad p(z) \in \mathcal{P}_{n-1},\ q(z) \in \mathcal{P}_n,\ q(0) = 1,$$

where $\mathcal{P}_\ell$ denotes the set of univariate polynomials of degree at most ℓ.

Let us introduce the two matrices $B_1 = (b_{i,j}^{(1)}), B_2 = (b_{i,j}^{(2)}) \in \mathbb{C}^{n\times n}$ given by

$$b_{i,j}^{(1)} = \frac{N(z_i) - N(w_j)}{z_i - w_j}, \quad z_i \in \mathcal{Z}^{(1)}, w_j \in \mathcal{Z}^{(2)}, \quad 1 \le i, j \le n,$$

and

$$b_{i,j}^{(2)} = \frac{z_i N(z_i) - w_j N(w_j)}{z_i - w_j}, \quad z_i \in \mathcal{Z}^{(1)}, w_j \in \mathcal{Z}^{(2)}, \quad 1 \le i, j \le n,$$

Pairs of matrices of the form (B_1, B_2) are considered in [9, 10]. Properties of these matrices can be enlightened by using some classical connections with other well known classes of displacement structured matrices. From

$$b_{i,j}^{(1)} = \frac{N(z_i) - N(w_j)}{z_i - w_j} = \frac{\frac{p(z_i)}{q(z_i)} - \frac{p(w_j)}{q(w_j)}}{z_i - w_j} = q(z_i)\frac{p(z_i)q(w_j) - p(w_j)q(z_i)}{z_i - w_j}q(w_j).$$

There follows that

$$B_1 = \operatorname{diag}\left[q(z_i)\right] \cdot \mathcal{B}_1 \cdot \operatorname{diag}\left[q(w_j)\right],$$

where

$$\mathcal{B}_1 = \left(\frac{p(z_i)q(w_j) - p(w_j)q(z_i)}{z_i - w_j}\right)_{1\le i,j\le n}.$$

Recall that the symmetric $n \times n$ matrix $\mathcal{B}_{p,q} = (f_{i,j})$ such that

$$\frac{p(x)q(y) - p(y)q(x)}{x - y} = \left[1, x, \ldots, x^{n-1}\right] \mathcal{B}_{p,q} \left[1, y, \ldots, y^{n-1}\right]^T = \sum_{i,j=0}^{n-1} f_{i+1,j+1} x^i y^j,$$

is called the *Bezoutian* associated with the polynomial pair $(p(z), q(z))$. Hence, it is found that

$$\mathcal{B}_1 = \mathcal{V}^T(z_1, \ldots, z_n) \cdot \mathcal{B}_{p,q} \cdot \mathcal{V}(w_1, \ldots, w_n),$$

where

$$\mathcal{V}(x_1, \ldots, x_n) = \begin{bmatrix} 1 & \ldots & \ldots & 1 \\ x_1 & \ldots & \ldots & x_n \\ \vdots & \vdots & \vdots & \vdots \\ x_1^{n-1} & \ldots & \ldots & x_n^{n-1} \end{bmatrix},$$

 ISBN: 978-84-617-8694-7

is the classical Vandermonde matrix generated from the nodes $x_1, \ldots, x_n$. The matrix $\mathcal{B}_{p,q}$ can further be decomposed by using the properties of Vandermonde matrices [12]. For the sake of simplicity, let us assume that the zeros $\xi_1^{(n)}, \ldots, \xi_n^{(n)}$ are pairwise distinct. Then we have

$$\mathcal{B}_{p,q} = \mathcal{V}^{-T}(\xi_1^{(n)}, \ldots, \xi_n^{(n)}) \cdot \mathrm{diag}\left[\varepsilon_{1,i}^{(n)}\right] \cdot \mathcal{V}^{-1}(\xi_1^{(n)}, \ldots, \xi_n^{(n)}),$$

where

$$\varepsilon_{1,i}^{(n)} = p(\xi_i^{(n)})q'(\xi_i^{(n)}), \quad 1 \le i \le n.$$

A similar analysis relates the matrix B_2 with $\mathcal{B}_{zp,q}$ such that

$$\mathcal{B}_{zp,q} = \mathcal{V}^{-T}(\xi_1^{(n)}, \ldots, \xi_n^{(n)}) \cdot \mathrm{diag}\left[\varepsilon_{2,i}^{(n)}\right] \cdot \mathcal{V}^{-1}(\xi_1^{(n)}, \ldots, \xi_n^{(n)}),$$

where

$$\varepsilon_{2,i}^{(n)} = \xi_i^{(n)} p(\xi_i^{(n)})q'(\xi_i^{(n)}), \quad 1 \le i \le n.$$

In this way, we arrive at the following two-step procedure for approximating the zeros of $q(z)$ and therefore, hopefully, the zeros of $f(z)$:

1. for two given set of points $\mathcal{Z}^{(1)} = \{z_j\}_{j=1}^n$ and $\mathcal{Z}^{(2)} = \{w_j\}_{j=1}^n$, evaluate $N(z_i)$ and $N(w_i)$, $1 \le i \le n$, and then form the matrices B_1 and B_2;

2. compute the generalized eigenvalues of the matrix pair (B_2, B_1).

A rank structured linearization can also be defined starting from the matrix pair (B_1, B_2). If we set $D_2 = \mathrm{diag}([w_1, \ldots, w_n])$ then we have

$$B_2 - B_1 D_2 = \boldsymbol{e}\,[N(z_1), \ldots, N(z_n)]\,, \quad \boldsymbol{e}^T = [1, \ldots, 1]\,.$$

The resulting representation of $B_1^{-1}B_2$ as a diagonal plus a rank-one matrix corresponds with the linearization provided in [2] based on Lagrange interpolation.

In this talk we analyze the role of structured linearizations formed by Bezoutian-like matrices as well as of their rank structured counterparts generated from the values of the Newton correction applied for the solution of a nonlinear equation. Numerical results are presented and some open questions are listed.

References

[1] A. P. Austin, P. Kravanja, and L. N. Trefethen, *Numerical algorithms based on analytic function values at roots of unity*, SIAM J. Numer. Anal. **52** (2014), no. 4, 1795–1821. MR 3240851

[2] D. A. Bini, L. Gemignani, and V. Y. Pan, *Improved initialization of the accelerated and robust QR-like polynomial root-finding*, Electron. Trans. Numer. Anal. **17** (2004), 195–205. MR 2113008

[3] D. A. Bini and V. Noferini, *Solving polynomial eigenvalue problems by means of the Ehrlich-Aberth method*, Linear Algebra Appl. **439** (2013), no. 4, 1130–1149. MR 3061758

[4] P. Boito, Y. Eidelman, and L. Gemignani, *Implicit QR for rank-structured matrix pencils*, BIT **54** (2014), no. 1, 85–111. MR 3177956

[5] _____, *Implicit QR for companion-like pencils*, Math. Comp. **85** (2016), no. 300, 1753–1774. MR 3471106

[6] _____, *A real QZ algorithm for structured companion pencils*, Tech. report, arXiv:1608.05395v3, 2017.

[7] L. Gemignani, *Zerofinding of analytic functions by structured matrix methods*, A panorama of mathematics: pure and applied, Contemp. Math., vol. 658, Amer. Math. Soc., Providence, RI, 2016, pp. 47–66. MR 3475271

[8] L. Gemignani and V. Noferini, *The Ehrlich-Aberth method for palindromic matrix polynomials represented in the Dickson basis*, Linear Algebra Appl. **438** (2013), no. 4, 1645–1666. MR 3005247

[9] A. C. Ionită, *Lagrange rational interpolation and its applications to approximation of large-scale dynamical systems*, Ph.D. thesis, Rice University, Houston, Texas, 2013.

[10] S. Ito and Y. Nakatsukasa, *Stable polefinding and rational least-squares via eigenvalues.*

[11] S. Ito and M. Sugihara, *A note on the convergence of pole estimation by rational interpolation*, Tech. Report METR 2015-02, Department of Mathematical Informatics, Graduate School of Information Science and Technology, University of Tokio, 2015.

[12] G. Sansigre and M. Alvarez, *On Bezoutian reduction with the Vandermonde matrix*, Linear Algebra Appl. **121** (1989), 401–408, Linear algebra and applications (Valencia, 1987). MR 1011748 (90i:15012)

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

Decision making modelling process to optimize the power unit maintenance in mining excavators

S. Gerassis[1], J.F. García[2], Á. Saavedra[3], J.E. Martín[1] and J. Taboada[1]

[1] *Department of Natural Resources and Environmental Engineering, University of Vigo, Spain*

[2] *CIPP Internationacional SL, Gijón Llanera, Spain*

[3] *Department of Department of Statistics and Operational Research, University of Vigo, Spain*

emails: sakis@uvigo.es, jgarcia@cippinternacional.com, saavedra@uvigo.es, jmartinsuarez@vigo.es, jtaboada@uvigo.es

Abstract

The power unit is the key element of heavy machinery. Its actual technological complexity makes it difficult either for constructors or engineers to predict properly its failure. The main contribution of this work is, firstly, to provide a suitable model to obtain the probability distribution that better reflects the fault occurrence on the power unit for mining excavators from a work management approach. Secondly, an optimum maintenance strategy is modelled through an influence diagram in terms of repair costs and production losses, representing the direct and indirect costs engineers have to face when a machine breaks down. Results show Weibull distribution as the best probabilistic model for the estimation of prior fault probabilities of the power unit elements. Indirect costs are demonstrated to be about 4.5 times bigger than direct costs, reflecting the necessity for a maintenance strategy capable to reduce faults in the early stages avoiding costs to become expansive over time. This approach tries to minimize production losses at the same time engineers gain knowledge in the management of risks involved with the severity and time of appearance of certain types of faults.

Key words: model selection, reliability and life testing, decision making, maintenance
MSC 2000: 62N02, 90B25, 90B50

ISBN: 978-84-617-8694-7

1 Introduction

Excavators (Hydraulic Excavators) are heavy construction equipment decisive for earth-moving operations both in mining and civil works. The engine consists of three parts. The engine block, the engine head and the lower engine. Each part is constituted respectively for a large number of components which are assembled in order to obtain the intended part. When talking about failures, from the set of faults that can be recorded during the machine operating life, according to their prevalence, those that can be associated with the power unit can be classified as: direct engine faults, injection system faults and starter engine faults.

During the past 20 years there has been a heightened improvement in the manufacture of heavy machinery engines. They are more powerful and fuel efficient with a minimized impact of emissions. The reliability has also been significantly enhanced with the inclusion of sophisticated electronic settings able to detect and predict nearby faults, but with a high price. Commonality and simplicity of design has turned now into complex structures with multiple sensors and an increased number of components. On many occasions, when faults occur time to repair takes a considerable time due to disassembly process which has to be carried out by qualified technical staff.

However, from a management perspective when a fault occurs the problem for engineers is not so much which component has failed, rather than how long the machine is going to be stopped or how much is the reparation cost. This situation, creates the necessity to define maintenance strategies oriented toward a wider scale represented by the set of power unit elements failure instead of power unit components failure. This new approach provides a great simplification of the domain problem, although is widely known that the decision-model results and conclusions are reliable as the probability model is well specified.

The motivation behind this work comes from trying to figure out whether different distributions change maintenance decisions in a considerably way. In this context, this paper evaluates exponential and Weibull distributions due to its extensive use in data analysis and reliability engineering [1]. The exponential distribution excels by its simplicity in calculation, but might not be appropriate to model the overall lifetime power unit elements, because the failure rates are not constant, and a constant failure rate approximation could not be representative enough. Alternatively, an important aspect of the Weibull distribution is how the values of the shape parameter k and the scale parameter λ affect the probability density function (PDF) and how they properly represent the power unit reliability [2]. For this latter issue, a suitable estimation of Weibull parameters is needed in order to obtain a reliable analysis of the occurrence of faults in the power unit.

The rest of the paper is structured in the following way. Section 2 presents the mathematical methods used to determine Weibull parameters and obtain the cumulative distribution function (CDF), both for Weibull and exponential distribution. Section 3 explains the influence diagram design. Section 4 shows the decision model results and discusses

the findings obtained about how these distributions may influence maintenance decisions. Section 5 concludes the paper and provides future work.

2 Model selection

The Weibull distribution is a suitable probability distribution for modelling survival analysis and has been widely used in reliability engineering and failure analysis due to its versatility. The density function of a Weibull distribution is given by the following expression:

$$f\left(x|\theta\right)=\frac{k}{\lambda}\left(\frac{x}{k}\right)^{k-1}e^{-(x/\lambda)^{k}},\ \ \text{if } x\geq 0,$$

where $\theta=(k,\lambda)^{t}$, $k>0$ is the shape parameter, $\lambda>0$ the scale parameter and t denotes transposed. On the other hand, the exponential distribution is a particular case with $k=1$ indicating a constant failure rate $\eta=\frac{1}{\lambda}>0$ over time and therefore that random external events are causing the components failure.

2.1 Estimation methods

In order to select the probabilistic model that better determines the fault of the power unit in mining excavators, several methods have been implemented for the estimation of the parameter θ. Given a sample $\{x_1,x_2,,x_n\}$ of size n drawn from a random variable X, the following estimation methods can be defined:

Method 1. Maximum likelihood estimation. Under i.i.d. assumptions, θ is estimated by maximizing the likelihood function defined as:

$$L\left(\theta\right)=\prod_{i=1}^{n}f\left(x_i|\theta\right).$$

Method 2. Moment matching estimation. This technique is based on matching the sample moments with the corresponding distribution moments:

$$E\left(X^{r}\right)=\frac{1}{n}\sum_{i=1}^{n}x_i^{r},\ r=1,2.$$

Method 3. Quantile matching estimation. Based on matching the sample quantiles, Q_{n,p_r}, with the corresponding distribution moments:

$$F^{-1}\left(p_r|\theta\right)=Q_{n,p_r},\ 0<p_r<1.$$

Method 4. Maximum goodness-of-fit estimation with the Cramer-von Mises goodness-of-fit distance. Assuming that an ordered sample, $x_1\leq x_2\leq\ldots\leq x_n$, θ is estimated by minimizing:

 ISBN: 978-84-617-8694-7

$$\frac{1}{12n}+\sum_{i=1}^{n}\left[F\left(x_i|\theta\right)-\frac{2i-1}{n}\right]^2. \tag{1}$$

Method 5. Maximum goodness-of-fit estimation with the Kolmogorov-Smirnov goodness-of-fit distance. Assuming that an ordered sample , $x_1 \leq x_2 \leq \ldots \leq x_n$, θ is estimated by minimizing:

$$max\left\{max_{i=1,\ldots,n}\left[\frac{1}{n}-F\left(x_i|\theta\right)\right], max_{i=1,\ldots,n}\left[F\left(x_i|\theta\right)-\frac{1-1}{n}\right]\right\}. \tag{2}$$

2.2 Goodness-of-fit statistics

Different goodness-of-fit statistics were calculated to measure the distance between the adjusted parametric distribution and the empirical distribution. Firstly, three goodness-of-fit statistics which are classically considered when fitting continuous distributions: Cramer-von Mises, based on (1), Kolmogorov-Smirnov, whose distance was given in (2), and Anderson-Darling with the following goodness-of-fit distance, assuming an ordered sample:

$$-n-\frac{1}{n}\sum_{i=1}^{n}\left(2i-1\right)\log\left\{F\left(x_i|\theta\right)\left[1-F\left(x_{n+1-i}|\theta\right)\right]\right\}.$$

Secondly, Akaike information criterion or the Bayesian information criterion were calculated. These loglikehood criteria are often appropriate to avoid overfitting when small samples are available:

$$\begin{aligned} AIC &= 2r-2\ln\left[L\left(\hat{\theta}\right)\right], \\ BIC &= \ln\left(n\right)r-2\ln\left[L\left(\hat{\theta}\right)\right], \end{aligned}$$

with $r=2$, number of estimated parameters.

Finally, root mean square error:

$$RMSE=\left\{\frac{1}{n}\sum_{i=1}^{n}\left[f\left(x_i|\hat{\theta}\right)-\frac{1}{n}\right]^2\right\}^{1/2}$$

and Chi-squared distance:

$$\chi^2=\frac{\sum_{i=1}^{n}\left[f\left(x_i|\hat{\theta}\right)-\frac{1}{n}\right]^2}{n-r}$$

were also calculated.

 ISBN: 978-84-617-8694-7

2.3 Estimated parameters and selected probability distributions

All calculations were obtained using the open source programming language R [3]. The results obtained with the different goodness-of-fit statistics allowed to identify Method 5 for the engine, Method 1 for the starter engine and Method 3 for the injection system as the best estimation methods for the obtainment of Weibull parameters. The values of the parameters required for the representation of the cumulative distribution function for each power unit element with the winning methods are shown in Table 1.

Table 1: Parameters values for Weibull and exponential distributions representation

Distribution	Parameters	Engine	Starter Engine	Injection System
Weibull	Shape (k)	0.91290	1.44447	1.94212
	Scale (λ)	1222.22	3244.58	3658.132
Exponential	Failure Rate (η)	0.00027586	0.00015517	0.00012069

Note that the cumulative distribution function for the Weibull distribution is:

$$F(x) = 1 - e^{-(x/\lambda)^k},\ x \geq 0,$$

whose representation requires the estimation of the scale λ and shape k parameters. Whereas, the exponential distribution only needs the failure rate η, being its cumulative distribution function:

$$F(x) = 1 - e^{-\eta x},\ \eta = \frac{1}{\lambda}, x \geq 0.$$

Table 1 showns the constant failure rate for each power unit element estimated by means of the average of faults in the machine operating life. It is interesting to analyse the payoff between its simplicity of calculation and the magnitude of the difference in the results, Figure 1, especially when they are transferred to an influence diagram affecting strategic decisions.

In view of Figure 1, it can be seen how the Weibull distribution gives higher probabilities to the failure of the units during the first years of its operating time.

3 Influence diagram for maintenance strategy evaluation

The maintenance strategy is modelled through an influence diagram (ID). An ID is a directed acyclic graph representing a generalization of a Bayesian network, in which probabilistic inference can be applied to solve decision making problems [4]. In this case, the ID was created using the decision modelling software BayesFusion, LLC [5]. The problem design depicted in Figure 2 involves 4 variable types for notation:

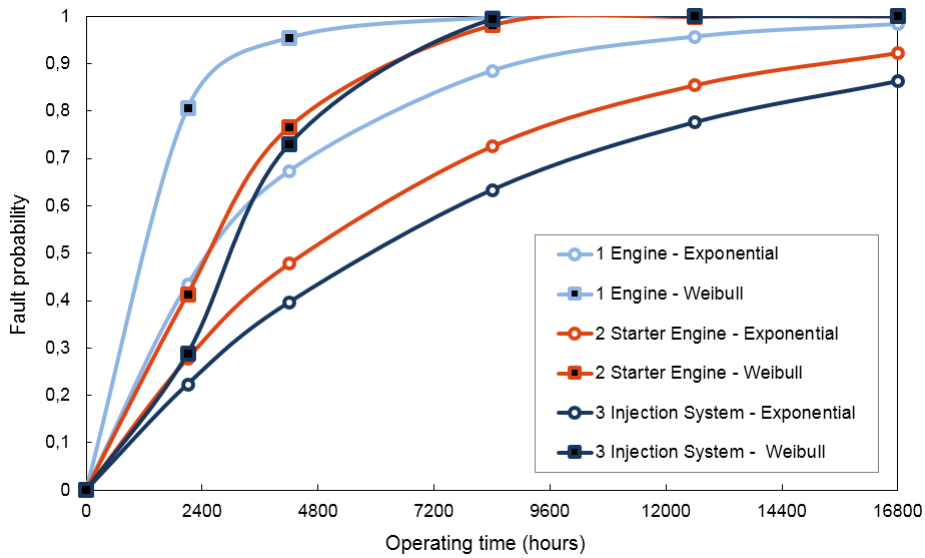


Figure 1: Cumulative distribution functions comparison for the power unit elements

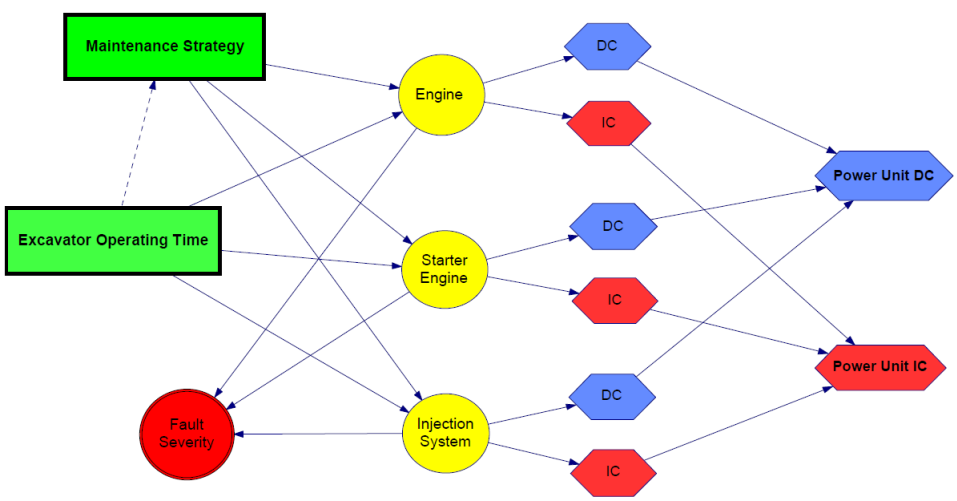


Figure 2: Influence diagram (ID) for power unit maintenance evaluation

 ISBN: 978-84-617-8694-7

Table 2: Deterministic node definition. Fault occurrences (yes or no) determining power unit fault severity (very high, high, medium or low)

Engine	Yes	Yes	Yes	Yes	No	No	No
Starter Engine	Yes	Yes	No	No	Yes	Yes	No
Injection System	Yes	No	Yes	No	Yes	No	Yes
Very high	✓		✓				
High		✓		✓	✓		
Moderate							✓
Low						✓	

- 2 decision nodes (green rectangles). The *excavator operating time* is evaluated in hours. Every operation year the excavator works 4800 hours, considering a service life at full performance up to 16800 hours. The *maintenance strategy* node offers the possibility to assess the maintenance strategy according to the fault probabilities for each element of the power unit obtained with exponential and Weibull distributions (Figure 1).

- 3 chance nodes (yellow circles). The *engine* itself, the *starter engine* and the *injection system*. They are quantified by the probabilities (Figure1) which integrate the uncertainty associated to the failure of the power unit.

- 1 deterministic node (red double circle). It represents the *fault severity* of the power unit. Once all their parents are known, there is no uncertainty about the outcome. The quantification is similar to chance nodes. The only difference now is that when a fault event takes place, the outcome is known with certainty. The definition is done with a probability table (Table 2) that contains the fault severity depending on the combination of fault elements in the power unit, according to the criteria of mining engineers consulted.

- 8 value nodes (blue and red hexagons). Blue hexagons represent the direct cost (DC) and red hexagons the indirect cost (IC) for each power unit element fault. DCs refer to the economic cost of fault repair. On the other hand, ICs imply a broader concept. They compute the economic cost associated with the loss of production due to the failure of the machine. The loss of production depends on the repair time, which is in turn dependent on the severity of the fault. For this reason, the utility costs for every element fault are computed independently. Moreover, for a generalized analysis it is usually easier for the decision maker to combine them in a single multi-attribute utility function (MAU) [6]. Thus, the influence diagram has two final MAUs, the

Power Unit DC and the Power Unit IC, summarizing the direct and indirect costs expected for the power unit elements failure over time.

In accordance with the expected utility theory, the goal of an influence diagram is to choose a decision alternative that has the highest expected gain or utility [7]. Utility is, however, by assumption subjective. In this particular case, the influence diagram enables engineers and decision makers assess the expected costs of suffering a failure over time. This approach means that direct and indirect costs need now to be minimized knowing that utility has not a meaningful zero point because maintenance has always an associated cost and it is very rare the case, not to say impossible, that an excavator has no faults during its operating time.

Various decision makers facing the same problem and even sharing the same set of beliefs may choose differently because of their preference structure and different utility functions. This can be especially noticed in a field such as engineering. Taking this into account, the utility elicitation of fault costs was made through an extensive review of repair invoices from power units of several mining excavators in the last 5 years.

4 Decision model results

The results obtained offer two main contributions. Firstly, how the estimations of prior probabilities can affect decision making for maintenance policy in this new approach based on the power unit segmentation into three main failure elements. Secondly, how the risk for fault severity changes depending on the model selected.

4.1 Distribution influence on decision making

The expected direct and indirect costs associated with the failure of the power unit modelled either with exponential or Weibull distribution show a significant growth in the first two years (see iceberg chart in Figures 3 and 4). After the second year, when the machine has been operating more than 9600 hours, the expected costs present a certain stabilization.

Comparing the results obtained with each model, a big difference lies within the first 3 years. The influence diagram, when is modelled with exponential distribution, gives rise to lower direct and indirect costs for that period. These differences reduce over time, from 48% less for the first year to 20% less the second and just 10% less the third. This highlights that even if both models present a good similarity from the third year onwards, Weibull distribution, more mathematically appropriate, can better respond to the expected costs during the initial stages. Therefore, even though for some real life scenarios a constant failure rate can represent a good approximation, a Weibull distribution has proven to be a worthwhile distribution for modelling power unit faults, although the estimations of its parameters may be more time-consuming in terms of calculation.

 ISBN: 978-84-617-8694-7

This last point is also supported by the fact that Weibull distribution enables a better understanding of power unit elements. The shape parameter k represents the failure rate behaviour. A value of $k < 1$ indicates that the failure rate decreases over time. This is the case of the engine block k=0.91290 (Table 1). When $k > 1$ the failure rate increases with the passing of time. The starter engine k=1.44447 and the injection system k= 1.94212 (Table 1) present this condition reflecting the existence of an aging process. It is noticeable that the engine itself is the one with the most similar behaviour to an exponential distribution (k=1), while the injection system gets closer to a Rayleigh distribution (k=2) with the starter engine in the middle of these two.

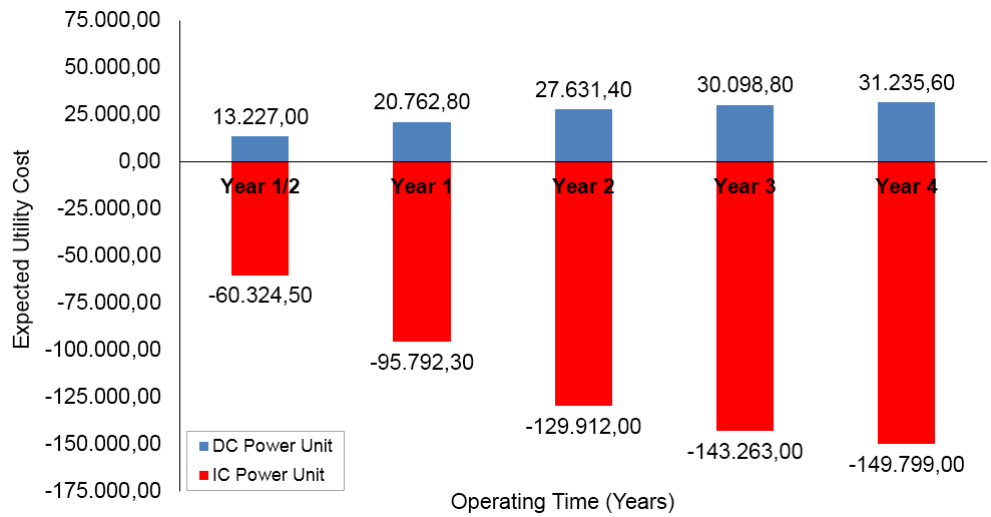


Figure 3: Direct and indirect expected fault costs for the exponential probabilistic model

When designing an optimum maintenance policy this knowledge is crucial. The engine is known now as more sensitive to suffer initial faults, while the starter engine and the injection system present the inverse condition. A right balance could be found with an extensive maintenance that pays special attention to the machine in its initial stage moving toward a less exhaustive maintenance when the machine have reached its half-life. This could ensure a good adaptation of the engine to the work environment whilst promoting a healthy aging for the starter engine and the injection system.

From a management perspective, a fault minimization at the beginning of the excavator operating life not only involves a reduction in direct costs associated with repair, but also would contribute to quickly reach the required hours to complete the amortization of the machine. Indirect costs show up in a ratio of 4.5 to 1 compared with direct cost. They are larger from what many engineers can imagine, often hidden behind the shadow of the faults having a huge role in the whole system performance.

 ISBN: 978-84-617-8694-7

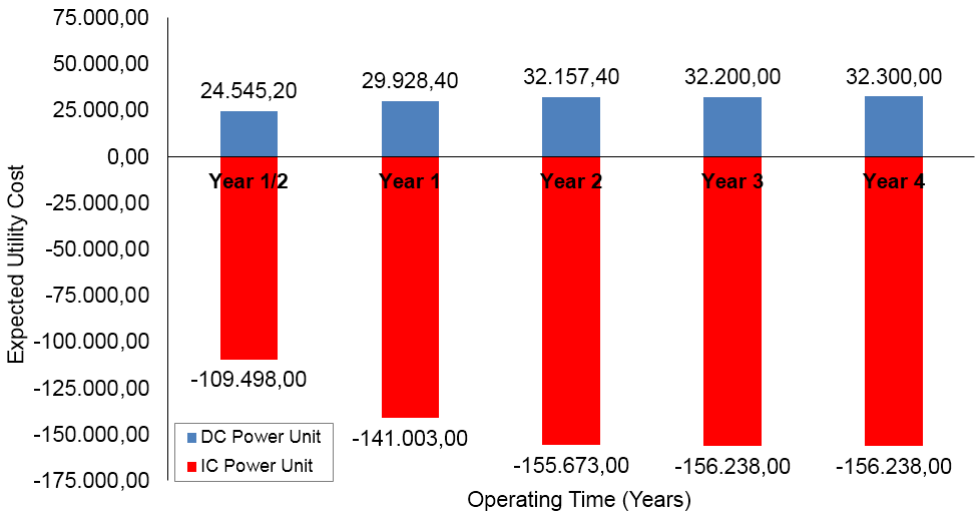


Figure 4: Direct and indirect expected fault cost for the Weibull probabilistic model

4.2 Fault severity

Aleatory or stochastic uncertainty due to the faults randomness is always at some point inherent to the variability of the system regardless of how good the maintenance strategy is. However, epistemic or subjective uncertainty arising from the lack of knowledge about the system and its behaviour can be certainly reduced by acquiring knowledge through probability and decision models like the one shown here. One aspect that holds special importance is the risk of suffering a fault with a high degree of severity, because of its huge cost and long time to repair.

The deterministic node incorporated in the influence diagram (Figure 2) with engineers' criteria for the fault degree of severity regarding the particular damaged elements in the power unit (Table 2) makes it possible to calculate the risk profile for the machine operating time (Figure 5). The risk profile is important for identifying the acceptable level of risk an individual or corporation is able to accept. It is expressed in terms of risk probability including the results of using in the influence diagram the prior probabilities for both the exponential and Weibull model.

As it can be seen in Figure 5, from year 2 the likelihood of having a fault with a very high severity exceeds the 50%. Year 2 can, therefore, set the point for which the risk level determines the maintenance strategy. During the two first years could be implemented an extensive maintenance based on a predictive approach that tries to minimize fault risk levels for the later years. From the third year a preventive maintenance, less exhaustive, can be applied. The aging effect is already present and expensive maintenance approaches could

 ISBN: 978-84-617-8694-7

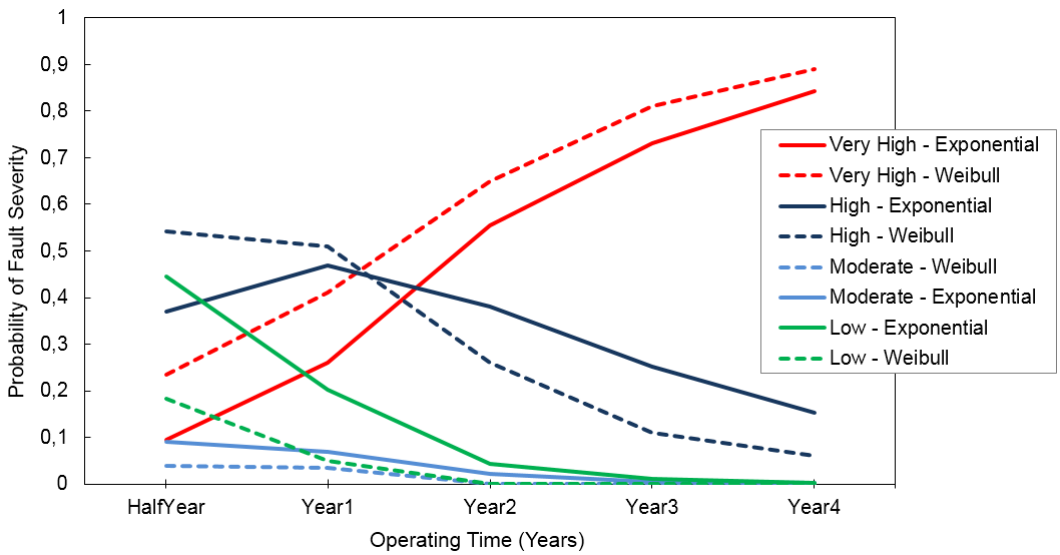


Figure 5: Probability risk profile for the power unit fault severity

not really stop faults appearance.

5 Conclusions and future work

In this paper, it was tried to develop a new process for the estimation of the power unit failure in mining excavators using an innovative management perspective. The power unit was divided in three main fault elements and from data of faults collected the last years in different mining excavators throughout its operation life, exponential and Weibull probabilistic models were used in order to obtain the prior fault probabilities for each element. The analysis of the prior probabilities into an influence diagram showed that the Weibull model offers a more accurate representation of the expected direct and indirect costs for the power unit.

A risk profile for the faults severity was calculated proposing an optimized maintenance solution for this machinery. Maintenance strategies should be designed under the assumption of a certain probability model that does not influence decisions. Since the selection of the probability model is carried out at an early stage of the design, one might expect a low impact on the final selection of the best strategy. This paper shows how a misspecification of the probability model can lead to erroneous conclusions since early stages causing expansive economic losses. Future work is required to analyse and optimize maintenance strategies in other crucial parts of excavators and other mining and civil machinery, especially those

exposed to a high level of wear.

References

[1] N. Sedransk, J. Sedransk: *Distinguishing among distributions using data from complex sample designs*, J. Amer. Stat. Assoc. **74** (1979) 754–760.

[2] B. de Jonge, W. Klingenberg, R. Teunter, T. Tinga: *Optimum maintenance strategy under uncertainty in the lifetime distribution*, Reliab. Eng. Sysyt. Saf. **133** (2015) 59–67.

[3] R Core Team: *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria, (2016).

[4] R. A. Howard: *The foundations of decision analysis revisited. In Advances in decision analysis: from foundations to applications*, ed. W. Edwards, R.F. Miles and D. Von Winterfeldt. Cambridge, UK: Cambridge University Press, (2007).

[5] BayesFusion: *GeNIe Modeler*, BayesFusion, LLc, Data Analytics, Mathematical Modelling, Decision Support. Pittsburgh, PA, (2016). http://www.bayesfusion.com/

[6] Z. Pawlak: *Rough set approach to multi-attribute decision analysis*, Eur. J. Oper. Res. **72** (1994) 443–459.

[7] J. von Neumann, O. Morgenstern: *Theory of games and economic behaviour*, Princenton, NJ. Princenton University, (1953).

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

Optimal Control with linear versus quadratic cost functions in disease prevention: From analytically treatable toy models to numerical analysis

Peyman Ghaffari[1], Karunia Putra Wijaya[2], Maíra Aguiar[1], Luís Mateus[1], Thomas Götz[2] and Nico Stollenwerk[1]

[1] *Biomathematics and Statistics Group, Centro de Matemática, Aplicações Fundamentais e Investigação Operacional CMAF-CIO, Department of Mathematics, Universidade de Lisboa, Campo Grande, Lisboa, Portugal*

[2] *Mathematical Institute, University of Koblenz, 56070 Koblenz,Germany*

emails: pgsaid@fc.ul.pt, karuniaputra@uni-koblenz.de, maira@ptmat.fc.ul.pt, lgmateu@fc.ul.pt, goetz@uni-koblenz.de, nico.biomath@gmail.com

Abstract

We investigate some simple models to describe prevention measures for mosquito borne diseases like dengue fever. We introduce the concepts of optimal control in very simple models which can in most aspects be treated analytically, namely the linear infection model, which has been used previously to describe for example vaccine trial data and could be as well applied to other control measures. We then relax the conditions of the linear infection model, and have to perform now most steps numerically, until we reach realistic models for mosquito borne diseases, namely the SISUV model as a simplest model with susceptible and infected humans, S and I, and susceptible and infected mosquitoes, U and V.

The SISUV model has already sufficient nonlinearities to not only apply quadratic cost functions for optimization but also linear cost functions, which have nontrivial optimal control equations due to the nonlinearities in the disease model. The classically used quadratic cost functions to be optimized are compared to linear cost functions, which are more realistical from the health economics point of view than the quadratic cost functions, but tend to the occasional occurence of bang-bang control signals. However, we can give examples of non-linear models, like the SISUV model, in which qudratic and linear cost functions give qualitatively very similar control signals, avoiding any

 ISBN: 978-84-617-8694-7

complications of the optimum reaching any unreasonable boundaries, which could lead to bang-bang controls. The small numerical differences are due to different scalings in quadratic and linear cost functions, and could eventually be removed. But these numerical differences are of little practical relevance.

Key words: optimal control theory, dengue fever, vaccination, mosquito repellents, linear infection model, SISUV model

1 Introduction

Zika infection, dengue fever, chikungunya and yellow fever are examples of vector-borne diseases transmitted by day-time active mosquitoes. In 128 countries, in particular in tropic and sub-tropic regions of Asia and Latin America these diseases are a major health risk and a negative economic factor.

In recent years, however, vector-borne diseases and especially dengue fever are occurring in Europe. Some reasons for this are the worldwide flow of trade and travelling tourism. Increasing urbanization, as well as regional warming due to global climate change, have amplified the spread of mosquitoes like *Aedes albopictus* in Europe. In 2010, infections with the dengue virus were registered in Croatia, France and Italy. In 2012 and 2013, more than 2000 autochtonous cases having dengue fever were found on the isle of Madeira/Portugal transmitted by *Aedes aegypti*. Chikungunya infections occurred in Italy (2007) and Spain/France (2015). Eggs, larvae, pupae and adult mosquitoes of *Aedes albopictus* were repeatedly detected in the south of Germany in autumn 2014 and in 2015. Researchers assume from these findings that Asian tiger mosquitoes can survive the winter and settle in Germany.

Over the past few decades, the incidences of dengue have grown dramatically. Recent studies indicate the existence of approximately 390 million dengue infections per year and that 3,9 billion people, in 128 countries including Thailand, Brazil, India and Pakistan, are at risk of being infected with the dengue virus. The WHO has set the goal to constrain and control the spreading of dengue fever by 2020, however there are major obstacles in achieving this goal. Some vaccines are in advanced trial stages, but not effective against all serotypes, with the Phase 3 results of the Sanofi Pasteur vaccine as front runner just concluded, and have negative effects in some age classes. WHO guidelines for vaccine trials are very detailed and specific in their requirements of scientific investigation before licensing, with phases 1, 2, 2b and 3, and finally phase 4 after licensing.

As already mentioned, for dengue fever first vaccine trials are running, but the results are not satisfactory. In general regarding mosquito vector-borne diseases vaccines are quite imperfect like DengVaxia for dengue fever, recently licenced by Sanofi-Pasteur, or vaccines do not yet exist as is the case for the Zika virus. In relation to yellow fever the vaccine is even in some cases lethal. Classical mosquito control measures, like bed-nets and municipal spraying in the streets, have proven to be of little effectiveness in combating disease cases. In

 ISBN: 978-84-617-8694-7

mosquito control, some activities in demonstration of efficacy using bed-nets via the WHO are performed. However bed nets are not very efficient against the disease. One reason is that vectors of dengue, the species *Aedes aegypti* and *Aedes albopictus* are active in the morning and evening, but not very active at night. Another important aspect in eliminating mosquitoes by classical pesticides and insecticides, beside the danger to human health, is that the elimination of mosquitoes, would also deprive many fish, birds, and reptiles of a food source and even destroy critical pollinator for plants.

In future research we will investigate SIR-type models with repellency and vaccination and analyse with optimal control theory. Here we first study a toy model which can in many aspects be treated analytically, and can already capture some simplest aspects of repellents respectively vaccination. Then numerical methods are studied to relax the need for analyticity in the models. In order to calculate numerically the influence of repellency and vaccination in the model we will use the gradient method. For some aspects see also [1, 2]. Ones the effectiveness of different control measures in known, like e.g. in the case of the dengue vaccine [3] using the experimental data obtained during the phase 3 trial, e.g. [4] or like new generations of mosquito repellent applications (including nano-particles in textiles, wall paints etc.) the next step, of course, is the suggestion of optimal strategies to combat vector-borne diseases like dengue fever, and it would be a matter of optimal control as a mathematical field.

2 The mathematical model for dengue fever

2.1 The general SIRUV-model for coupling mosquito to human epidemiology including repellency and vaccination

For pure human disease epidemiology, we assume the usual SIR model given by the ODEs

$$\begin{aligned} \frac{d}{dt}S &= \mu(N-S) - \frac{\beta}{N}SI \\ \frac{d}{dt}I &= \frac{\beta}{N}SI - (\gamma+\mu)I \\ \frac{d}{dt}R &= \gamma I - \mu R \end{aligned} \tag{1}$$

with state variables S for susceptible humans, I for infected and R for recovered humans. The population of humans $N = S + I + R$ is assumed constant. The infection rate is given by β, recovery rate γ and birth and death rate for humans by μ.

The stationary states are easily otained. For the human infection we obtain the trivial disease-free equilibrium stationary state $I_1^* = 0$ and the non-trivial case $I_2^* = (\mu/(\gamma+\mu)(1-(\gamma+\mu)/\beta)N$. Respectly we have the two stationary states for the susceptibles $S_1^* = N$ and $S_2^* = ((\gamma+\mu)/\beta)N$ and for the recovered in both cases $R^* = N - S^* - I^*$.

 ISBN: 978-84-617-8694-7

Now we add to the ODEs (1) the susceptible mosquito population U and the infected mosquitos as disease vectors V. In the easiest case the total population size M adds up to $M = U + V$. We assume, since mosquitoes do not have an immune system, they cannot recover from the infection. So the resulting ordinary differential equations give the following system

$$\begin{aligned} \frac{d}{dt}S &= \mu(N-S) - \frac{\beta}{M}SV - \upsilon S \\ \frac{d}{dt}I &= \frac{\beta}{M}SV - (\gamma+\mu)I \\ \frac{d}{dt}R &= \gamma I - \mu R + \upsilon S \\ \frac{d}{dt}U &= \psi - \nu U - \frac{\vartheta}{N}UI \\ \frac{d}{dt}V &= \frac{\vartheta}{N}UI - \nu V \end{aligned} \tag{2}$$

with mosquito birth rate ψ, infection rate from human to mosquito ϑ and mosquito mortality ν, mosquito population size $M = U(t) + V(t)$ assumed constant, hence for now $\psi := \nu \cdot M$. We also included in the above equations (2) the vaccination factor υ. Here the repellent acts as reducing contact rates between humans and mosquitoes, hence β and ϑ, where in time scale separation or center manifold analysis the simplest version already gives $\beta \cdot \vartheta$ as contact parameter in the effective SIR model [5].

2.2 The simple case of linear infection model and optimal control

Using the described model in the last section we use the simple case of the linear infection model to understand the application of the optimal control method in these above mentioned models. The linear infection model with reaction scheme

$$S + I^* \quad \stackrel{\beta}{\longrightarrow} \quad I + I^* \tag{3}$$

with susceptibles S meeting outside infected I^* assumed in equilibrium with infection rate β brings the different aspects of modelling and analysing disease control measures together, since

$$\begin{array}{lcl} S_c + I^* & \stackrel{\beta}{\longrightarrow} & I_c + I^* \\ S_v + I^* & \stackrel{(1-k_v)\beta}{\longrightarrow} & I_v + I^* \end{array}$$

is used in its stochastic version to describe vaccine trials [3]. Here k_v is the vaccine efficacy which can be estimated e.g. in a Bayesian framework from empirically given numbers of

 ISBN: 978-84-617-8694-7

infected in the vaccine group I_v, ones β is estimated from the number of infected in the control group I_c. And S_c and S_v are given by the group sizes of control group and vaccine group. Vaccines, however, are not always completely effective or even in some age groups like the dengue vaccine dangerous with occasionally negative vaccine efficacy $k_v < 0$, giving quite interesting effects in more complex multi-strain models including such imperfect vaccines in itself [7, 8].

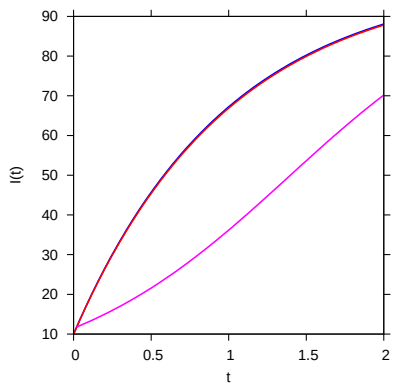


Figure 1: *Number of infected with time for the linear infection model, blue and red on top of each other for the uncontroled system, blue with the direct integration of the ODE, red with the discretized solution. The controlled system gives the number of infected with time as shown by the pink curve.*

Hence, other control measures like mosquito repellents, modelled as

$$\begin{array}{ccc} S_c + I^* & \xrightarrow{\beta} & I_c + I^* \\ S_M + I^* & \xrightarrow{(1-k_M)\beta} & I_M + I^* \end{array}$$

should be studied, with an as yet to investigate mosquito repellent efficacy k_M.

The reaction scheme Eq. (3) gives the rate equation

$$\frac{d}{dt}I = \frac{\beta}{N}SI^* = \beta^* S \tag{4}$$

with $\beta^* = (\beta/N)I^*$ and then omiting the star, but keeping conservation of population size N constant, hence

$$\frac{d}{dt}I = \beta S = \beta(N - I) \tag{5}$$

defining the linear infection model [6]. For the optimal control problem the linear infection model is given by

$$\frac{d}{dt}I = \beta_0(N - I) \tag{6}$$

 ISBN: 978-84-617-8694-7

with natural infection rate β_0, as for example measured from a control group study. For any control measure, vaccination or mosquito control, we obtain then

$$\frac{d}{dt}I = (1-k)\beta_0(N-I) = (\beta_0 - k\beta_0)(N-I) \tag{7}$$

with a maximally possible control $u_0 := k\beta_0$ for a vaccination of mosquito repellency group with all members vaccinated (hence 100% coverage). Any actual control strategy $u(t)$ then should be larger zero and smaller than the maximum, hence $0 < u(t) < u_0$.

We introduce the control signal $u(t)$ in $\beta(t) = (\beta_0 - u(t))$ in the above optimal control problem (7) with the simple solution of the ODE $\dot{I} = (\beta_0 - u(t))(N-I)$ given by

$$\begin{aligned} I(t) &= N - (N-I_0)e^{-\int_{t_0}^{t}\beta(\tilde{t})d\tilde{t}} \\ &= N - (N-I_0)e^{-\int_{t_0}^{t}\{\beta_0 - u(\tilde{t})\}d\tilde{t}} \qquad (8) \\ &=: I[u(\tilde{t})] \qquad (9) \end{aligned}$$

and then the optimal control problem minimizes the cost function

$$\mathcal{J} = \int_{t_0}^{T}\{\frac{1}{2}kI^2 + \frac{1}{2}\ell u^2\}\,dt \quad . \tag{10}$$

By using the variation of the cost function to obtain the optimal control signal $u(t)$

$$\frac{\delta\mathcal{J}[u(t)]}{\delta u(t)} = 0 \tag{11}$$

and for numerical analysis discretisation we can then solve the above problem. For the solution we discretize the total time T into n steps with size Δt

$$T = n\Delta t \tag{12}$$

and hence the continuous $I(t)$ becomes time discretisized using Eq. (8)

$$I_i = N - (N-I_0)e^{-\sum_{k=0}^{i-1}\{\beta_0 - u_k\}\,\Delta t} \tag{13}$$

and for convenience $B_i := e^{-\sum_{k=0}^{i-1}\{\beta_0 - u_k\}\,\Delta t}$. Note that we run the sum from $k=0$ to $k = i-1$. The cost function becomes plugging the above into Eq. (10), and using $m := \ell/k$ with unimportant factor k in the cost function

$$\begin{aligned} \mathcal{J}(T) &= \sum_{i=1}^{n}\{\frac{1}{2}I_i^2 + \frac{1}{2}mu_i^2\}\,\Delta t \\ &= \mathcal{J}\underbrace{(u_0, u_1, \cdots, u_{n-1})}_{=:\underline{u}} = \mathcal{J}(\Big\{u_k\Big\}_{k=0}^{n-1}) \qquad (14) \end{aligned}$$

 ISBN: 978-84-617-8694-7

hence

$$\mathcal{J}(\underline{u}) = \sum_{i=1}^{n} \{\frac{1}{2} I_i^2(\underline{u}) + \frac{1}{2} m u_i^2\} \, \Delta t \tag{15}$$

so the optimization problem reduces to optimizing $\mathcal{J}$ in respect to $\underline{u}$, i.e.

$$\frac{\partial \mathcal{J}(\underline{u})}{\partial u_j} = 0 \tag{16}$$

for all u_j.

After some calculations one finds the following expressions for the partial derivatives for the cost function $\mathcal{J}$

$$\frac{\partial \mathcal{J}(\underline{u})}{\partial u_j} = m u_j \cdot \Delta t + \sum_{i=j+1}^{n+1} I_i(\underline{u}) \cdot (-(N - I_0) B_i) \, (\Delta t)^2 \tag{17}$$

with $I_i(\underline{u}) = N - (N - I_0)B_i$ and the B_i being functions of the control signal $\underline{u}$. These partial derivatives either could be analyzed to be zero, but no solution is easily visible, or they can be used for a gradient method numerically, the way we went finally. Hence for the update of the $(\nu + 1)$ step of the control signal u_j from the ν step with initially zero control $u_j^{(0)} := 0$ is given by

$$u_j^{(\nu+1)} = u_j^{(\nu)} + h_j \tag{18}$$

with $h_j = c \cdot (-\partial \mathcal{J} / \partial u_j)$, i.e. maximizing $G := -\mathcal{J}$, and a step size c initially quite large, and then halfening every time it overshoots the minimization process. Numerical results are given in Figs. 1 and 2.

2.3 Generalization to not directly solvable disease dynamics

In cases when the disease dynamics $\dot{I} = f(I, u)$ cannot be solved analytically, we can use Lagrange multipliers $\lambda(t)$ for the optimization of the cost function, taking the disease dynamics as constraint into account, hence

$$\mathcal{L}[I(t), u(t), \lambda(t)] = \int_{t_0}^{T} \{\frac{1}{2} I^2 + \frac{1}{2} m u^2 + \lambda(\dot{I} - f(I, u))\} \, dt \tag{19}$$

to be minimized.

Now the discretization, done in the same way as before, reveals some interesting insights, namely the variation in respect to the Lagrange multipiers $\partial \mathcal{L} / \partial \lambda_j = 0$ gives back the constraint of the disease dynamcis

$$\frac{1}{\Delta t}(I_{j+1} - I_j) = f(I_j, u_j) \tag{20}$$

 ISBN: 978-84-617-8694-7

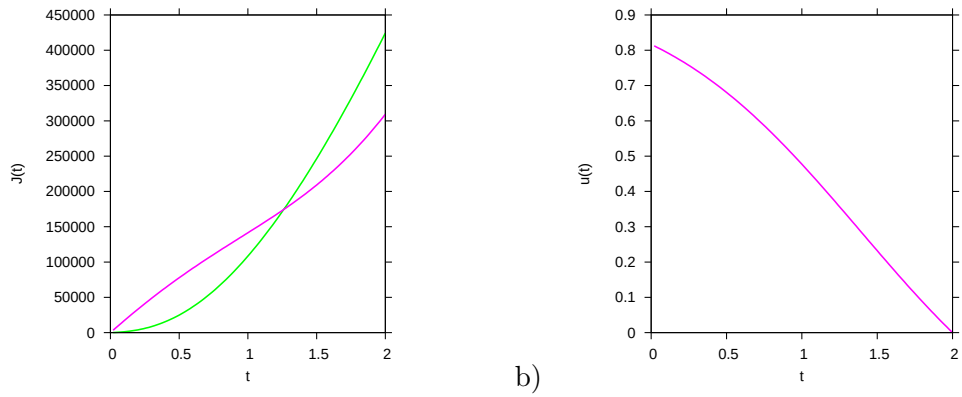


Figure 2: *In a) we show the cost function of the uncontrolled system in green, and the result after optimization in pink, reducing the cost significantly at the end of the target time, while initially the costs were increased due to control costs and not yet achieved reduction in disease cases. In b) we show the optimal control signal itself. The main effect has a control at the beginning of the controling time, and then relaxing, since the initial increase of infected has been successfully dampened.*

and then variation in respect to the infecteds $\partial\mathcal{L}/\partial I_j = 0$ gives a backward in time dynamics for the Lagrange multipliers

$$\frac{1}{\Delta t}(\lambda_{j-1} - \lambda_j) = \lambda_j \frac{\partial f}{\partial I_j} - I_j \tag{21}$$

with the upper boundary condition $\lambda_n = -I_{n+1} \cdot \Delta t \to 0$ for small time steps $\Delta t \to 0$. Finally the variation in respect to the control signal $\partial\mathcal{L}/\partial u_j = 0$ gives an algebraic equation system

$$mu_j - \lambda_j \frac{\partial f}{\partial u_j} = 0 \tag{22}$$

which in the case of a linear control in f, like we have here in the linear infection model, gives the control signal u_j as a function of λ_j and I_j only, due to $\partial f/\partial u_j$ is then only a function of I_j, and not of u_j any more. Hence then the forward dynamics for the infected and the backward dynamics for the Lagrange multipliers are sufficient to solve the optimal control problem, and

$$u_j = \frac{1}{m}\lambda_j \frac{\partial f}{\partial u_j} \tag{23}$$

determins the new updated control signal.

We implemented this method and in this case of the linear infection method we obtain exactly the same result as with the known analytic solution $I(\underline{u})$. Numerically, the results are indistinguishable from Figs. 1 and 2.

2.4 Optimal control for the SIS system

Now we use the resulting method for general disease models $\dot{I} = f(I, u)$ and generalize to the next simplest model of the SIS system, susceptibles becoming infected with rate β and recovering with rate α, hence

$$\dot{I} = f(I, u) = (\beta_0 - u(t)) \cdot \frac{I}{N}(N - I) - \alpha I \tag{24}$$

which is nonlinear in the disease variable I. In principal the direct method with

$$u_j = \frac{1}{m}\lambda_j \frac{\partial f}{\partial u_j} \tag{25}$$

could still be used, but this gives not converging cost functions, but oscillations between different values of the cost function, hence no clear minimum can be found.

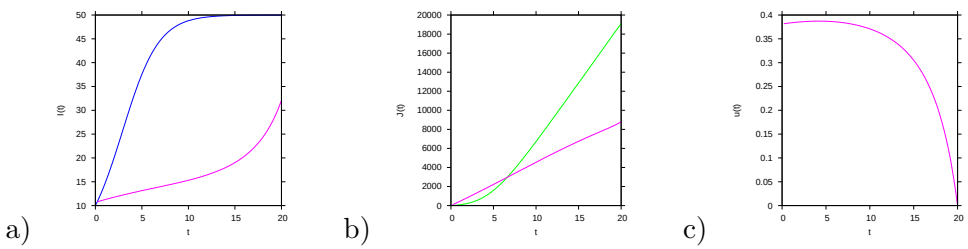


Figure 3: *Optimal control of the SIS system for cost function weight* $m = 5000$.

However, the constraint of the disease function has to be fulfilled always, so we keep the disease dynamics with given control signal in an Euler-forward scheme. Then, we also use the optimal solution of the backward dynamics for the Lagrange multipliers, and only replace the optimization of the control signal, which is most likely causing oscillations between minimal and maximal solutions of the cost function, by an explicit gradient method, as described before, i.e. with large initial step size and on rejection halfening this step size.

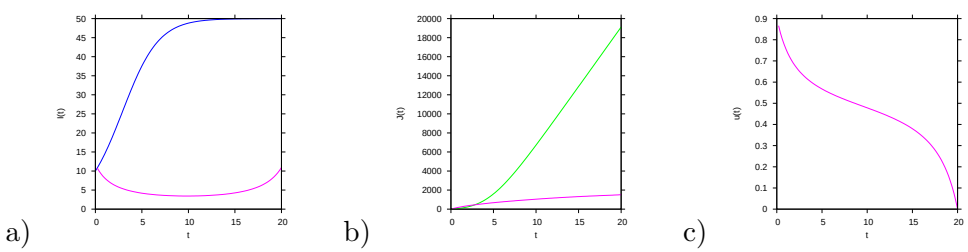


Figure 4: *Optimal control of the SIS system for reduced cost function weight* $m = 500$.

 ISBN: 978-84-617-8694-7

In this way we found a valid optimal solution also for the nonlinear SIS system, as given in Fig. 3 for the cost function weight of $m = 5000$, as before used for the linear infection model, and Fig. 4 for reduced costs of the control with weight $m = 500$.

2.5 Optimal control for the SISUV system: non-linear dynamics and non-linear control allow quadratic and also linear cost-functions with similar results

We now give some first preliminar results on the optimal control of the SISUV model [5], with percentual control in the contact parameters β and ϑ. Due to the non-linearities not only in the disease dynamics but also in the control signal, we can obtain similar optimal control results for the above given quadratic cost function and for a more economically reasonable linear cost function.

The SISUV model, describing in a simple way the time scale separation of mosquito dynamics versus the human disease dynamics and serving as a simplified version of the SIRUV model described above [5], is in its result given by the differential equation

$$\dot{I} = (\beta \cdot \vartheta) \cdot \frac{I}{N}(N - I) \cdot \frac{1}{\nu + \vartheta \frac{I}{N}} - \alpha I \tag{26}$$

with all notations also used and described in more detail in [5]. Now we will control the two contact rates, β of human susceptibles being infected by infected mosquitoes, and ϑ of susceprible mosquitoes being infected by infected humans, both in percentage of the rate in order to have comparable control signals in both parameters, hence

$$\beta(t) = \beta_0 \cdot (1 - u(t)) \tag{27}$$

and

$$\vartheta(t) = \vartheta_0 \cdot (1 - u(t)) \tag{28}$$

such that we have finally the controllable disease dynamics given by

$$\dot{I} = f(I, u) = (\beta_0 \cdot \vartheta_0) \cdot \frac{I}{N}(N - I) \cdot \frac{(1 - u)^2}{\nu + (1 - u) \cdot \vartheta_0 \frac{I}{N}} - \alpha I \tag{29}$$

hence a nonlinear dynamics $f(I, u)$ as well in the number of infected I as in the control signal u.

For the evaluation of the optimal control as described above in the section for the SIS model, we need the partial derivatives of $f(I, u)$, which can be calculated analytically as

$$\frac{\partial f}{\partial u} = (\beta_0 \cdot \vartheta_0) \cdot \frac{I}{N}(N - I) \cdot \frac{(u - 1)\left((1 - u)\vartheta_0 \frac{I}{N} + 2\nu\right)}{\left(\nu + (1 - u) \cdot \vartheta_0 \frac{I}{N}\right)^2} \tag{30}$$

 ISBN: 978-84-617-8694-7

and

$$\frac{\partial f}{\partial I} = (\beta_0 \cdot \vartheta_0) \cdot \frac{(1-u)^2}{N} \cdot \frac{(N-2I)\nu - (1-u)\vartheta_0 \frac{I^2}{N}}{\left(\nu + (1-u)\cdot \vartheta_0 \frac{I}{N}\right)^2} - \alpha \tag{31}$$

to be inserted in the conditions $\partial\mathcal{L}/\partial\lambda_j = 0$ for the disease forward dynamics, $\partial\mathcal{L}/\partial I_j = 0$ for the Langrange multiplier backward dynamics and for the gradient method of updating the control signal u the gradient $\partial\mathcal{L}/\partial u_j$, ones the cost function $\mathcal{J}$ and hence the Lagrangian $\mathcal{L}$ are specified.

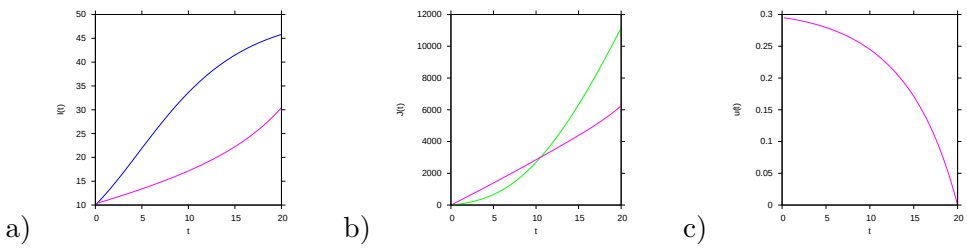


Figure 5: *Optimal control of the SISUV system for quadratic cost function with weight* $m_2 = 5000$.

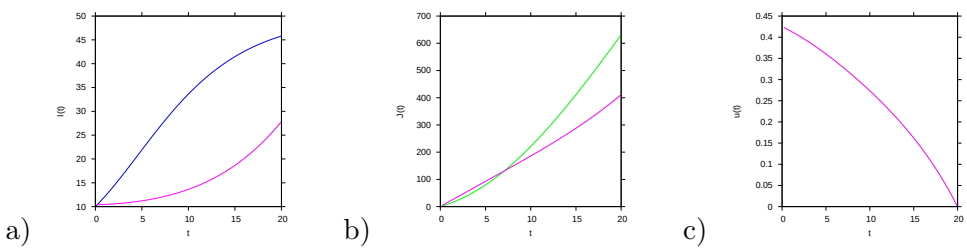


Figure 6: *Optimal control of the SISUV system for linar cost function with weight* $m_1 = 20$, *giving similar results to those using the quadratic cost function in the previous figure.*

Here we can now compare the results for the already above mentioned quadratic cost function

$$\mathcal{J}_2(\underline{u}) = \sum_{i=1}^{n} \left(\frac{1}{2} I_i^2(\underline{u}) + \frac{1}{2} m_2 u_i^2 \right) \Delta t \tag{32}$$

and an in economical terms better justifiable linear cost function, which however might be less convenient mathematically,

$$\mathcal{J}_1(\underline{u}) = \sum_{i=1}^{n} \left(I_i(\underline{u}) + m_1 u_i \right) \Delta t \tag{33}$$

both already given in their discretized version.

For the quadratic cost function and hence the Lagrangian function respectively we obtain the above mentioned results: The variation in respect to the Lagrange multipiers $\partial\mathcal{L}/\partial\lambda_j = 0$ gives back the constraint of the disease dynamcis

$$\frac{1}{\Delta t}(I_{j+1} - I_j) = f(I_j, u_j) \tag{34}$$

and then variation in respect to the infecteds $\partial\mathcal{L}/\partial I_j = 0$ gives the backward dynamics for the Lagrange multipliers

$$\frac{1}{\Delta t}(\lambda_{j-1} - \lambda_j) = \lambda_j \frac{\partial f}{\partial I_j} - I_j \tag{35}$$

while the variation in respect to the control signal $\partial\mathcal{L}/\partial u_j$ gives the gradient to be used in updating the control signal $\underline{u}$

$$\partial\mathcal{L}/\partial u_j = \left(m u_j - \lambda_j \frac{\partial f}{\partial u_j}\right) \cdot \Delta t \tag{36}$$

with the results after optimization given in Fig. 5 with a weight of the control signal in the quadratic cost function as $m_2 = 5000$.

For the linear cost function we obtain from the variation the following results: The variation in respect to the Lagrange multipiers $\partial\mathcal{L}/\partial\lambda_j = 0$ gives again the constraint of the disease dynamcis

$$\frac{1}{\Delta t}(I_{j+1} - I_j) = f(I_j, u_j) \tag{37}$$

and then variation in respect to the infecteds $\partial\mathcal{L}/\partial I_j = 0$ gives the backward dynamics for the Lagrange multipliers now in the slightly different form

$$\frac{1}{\Delta t}(\lambda_{j-1} - \lambda_j) = \lambda_j \frac{\partial f}{\partial I_j} - 1 \tag{38}$$

while the variation in respect to the control signal $\partial\mathcal{L}/\partial u_j$ gives the gradient to be used in updating the control signal $\underline{u}$ now also slightly different as

$$\partial\mathcal{L}/\partial u_j = \left(m - \lambda_j \frac{\partial f}{\partial u_j}\right) \cdot \Delta t \tag{39}$$

with the results after optimization given in Fig. 6, when using a weight of the control signal in the linear cost function as $m_1 = 20$ as a first guess to obtain similar control results as for the quadratic cost function given in Fig. 5 with $m_2 = 5000$. The results for both cost functions, the quadratic and the linear, are qualitatively in good agreement. No effort was given to find an m_1 for the linear cost function which could closer match the results obtained with m_2 in the case of the quadratic cost function.

 ISBN: 978-84-617-8694-7

Acknowledgements

This work has been supported by the European Union under FP7 in the DENFREE project, and the kind support by FCT, Portugal, in various ways. We also thank the DAAD and FCT to support the interaction between University of Koblenz and University of Lisbon in an DAAD-FCT exchange grant. Many thanks go to Gustav Feichtinger, Vienna, for pointing us to ways of achieving optimal control with linear cost functions.

References

[1] D.Aldila, E.Soewono, N.Nuraini, *On the Analysis of Effectivness in Mass Application of Mosquito Repellent for Dengue Disease Prevention*, *AIP Conf. Proc.* **1450** (2012), 103–109

[2] Karunia Putra Wijaya, Thomas Goetz, Edy Soewono, *An Optimal Control Model of Mosquito Reduction Managment in Dengue Endemic Region*, *International Journal of Biomathematics* Vol. 7, No. 5 (2014) 1450056, 22 pages, *DOI: 10.1142/SI793524514500569*

[3] Luís Mateus, Maíra Aguiar and Nico Stollenwerk (2015) Bayesian estimation of vaccine efficacy. *Proceedings of the 15th International Conference on Mathematical Methods in Science and Engineering - CMMSE 2015, Cadiz, Spain*, pp. 794–802, ISBN: 978-84-617-2230-3, edited by Jesus Vigo et al.

[4] M. R. Capeding et al., *Clinical efficacy and safety of a novel tetravalent dengue vaccine in healthy childen in Asia: a phase 3, randomised, observer-masked, placebo-controlled trial*, Lancet **384** (2014) 1358–65.

[5] Rocha, F., Aguiar, M., Souza, M., & Stollenwerk, N. (2013) Time-scale separation and center manifold analysis describing vector-borne disease dynamics, *Int. Journal. Computer Math.* **90**, 2105–2125.

[6] Stollenwerk, N., & Jansen, V. (2011) *Population Biology and Criticality: From critical birth–death processes to self-organized criticality in mutation pathogen systems* (Imperial College Press, World Scientific, London).

[7] Aguiar, M., Stollenwerk, N., & Halstead, S. (2016) The impact of the newly licensed dengue vaccine in endemic countries, *accepted for publication in "PLOS Neclected Tropical Diseases", published online December 21, 2016.*

[8] Aguiar, M., Stollenwerk, N., & Halstead, S. (2016) The risks behind Dengvaxia recommendation, *The Lancet Infectious Diseases*, **16**, 882–883.

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

On viable solutions of differential inclusions with fractional derivative without singular kernel

Ewa Girejko[1]

[1] *Faculty of Computer Science, Department of Mathematics, Bialystok University of Technology*

emails: e.girejko@pb.edu.pl

Abstract

In the paper we provide sufficient conditions assuring existence of viable solutions of differential inclusions with fractional derivative without singular kernel, namely Caputo–Fabrizio derivative of order $\alpha \in (0, 1)$. A modified condition of tangency, according to specifity of system with this new fractional derivative is given.

Key words: viability, differential inclusion, Caputo–Fabrizio derivative

1 Introduction and preliminaries

Viability theory as well as fractional calculus have a wide range of applications. And so, viablity theory designs and develops mathematical and algorithmic methods that can be found, for example, in biological evolution, economics, environmental sciences, financial markets or control theory and robotics (see [2] and the references within), while fractional calculus finds numerous applications in viscoelasticity, capacitor theory, electrical circuits (see for example [8, 9, 10, 11]). It was a motivation for the author to examine a combination of these two so applicative fields of mathematics. The goal of this research is to deliver conditions under which solutions of fractional differential inclusion are constrained to be in a given set (viable solutions). To the best of our knowledge, there are only few papers devoted to this subject [5, 6, 7, 12]. In the paper we give tangency conditions for the existence of viable solutions of differential inclusions with fractional derivative without singular kernel - Caputo–Fabrizio derivative.

 ISBN: 978-84-617-8694-7

We start with the definition of Caputo fractional derivative (for details see [9]). For given $b > 0$ and $\alpha \in (0,1)$ the Caputo fractional derivative is given by

$$ {}^{C}D^{\alpha} f(t) = \frac{1}{\Gamma(1-\alpha)} \int_{t_0}^{t} (t-s)^{-\alpha} f'(s) ds \,, \quad t > t_0. $$

Now changing the kernel $(t-s)^{-\alpha}$ by the function $\exp^{\frac{-\alpha(t-s)}{1-\alpha}}$ and $\frac{1}{\Gamma(1-\alpha)}$ by $\frac{1}{\sqrt{2\pi(1-\alpha^2)}}$ we obtain the new Caputo–Fabrizio fractional derivative of order $\alpha \in (0,1)$ introduced racently by Caputo and Fabrizio in [4] in the following form:

$$ {}^{CF}D^{\alpha} f(t) = \frac{(2-\alpha)M(\alpha)}{2(1-\alpha)} \int_{t_0}^{t} \exp\left(-\frac{\alpha}{1-\alpha}(t-s)\right) f'(s) ds \,, \quad t \geq t_0 \,, \tag{1} $$

where $M(\alpha)$ is a normalization constant depending on α.

Let F be a multifunction from $[t_0, T] \times \mathbb{R}^n$ into $\mathbb{R}^n$. We consider the differential inclusion with Caputo–Fabrizio derivative as follows:

$$ {}^{CF}D^{\alpha} y(t) \in F(t, y(t)) \,, \quad y(t_0) = y_0 \,, \quad t \in [t_0, T] \,. \tag{2} $$

Definition 1. An absolutely continuous function $y(\cdot)$ is said to be a solution to differential inclusion (2) if there exists a measurable selection $f_y(t) \in F(t, y(t))$ such that for all $t \in [t_0, T]$ one has

$$ y(t) = c + a_\alpha f_y(t) + b_\alpha \int_{t_0}^{t} f_y(s) ds \,, \tag{3} $$

where $c = -a_\alpha f(t_0, y_0) + y_0$, $a_\alpha = \frac{2(1-\alpha)}{(2-\alpha)M(\alpha)}$, $b_\alpha = \frac{2\alpha}{(2-\alpha)M(\alpha)}$ and $M(\alpha) = \frac{2}{2-\alpha}$, $0 \leq \alpha \leq 1$ and $y' \in L^{\infty}([t_0, T], \mathbb{R}^n)$.

2 Main results

The main result of the paper is theorem that delivers viability conditions for fractional differential inclusions with Caputo–Fabrizio derivative. Before we start with the assumptions on a set-valued map F and a function f, we define a multifunction $G(t,x) := \overline{co} F([t_0, T], x + \mathbb{B})$ (then G has convex closed values) and consider a differential inclusion with an initial condition as follows

$$ {}^{CF}D^{\alpha} y(t) \in G(t, y(t)) \,, \quad y(t_0) = y_0 \,, \quad t \in [t_0, T] \,. \tag{4} $$

Now we propose an assumption:

(A1) Assume that $F : [t_0, T] \times \mathbb{R}^n \rightsquigarrow \mathbb{R}^n$ is upper semi–conitnuous (short: u.s.c., for definition and details see [1, 2, 3]) set–valued map with nonempty convex and compact values. Moreover, let F be Lipschitz with the constant $k(t)$.

 ISBN: 978-84-617-8694-7

Next proposition is crucial for the main results.

Proposition 2. *Let us assume that* $(A1)$ *is fullfiled. Then there exists a constant* N_α *such that* $\|G(t,y(t))\| \leq N_\alpha$ *a.e. in* $[t_0,T]$ *for every solution* $y(\cdot)$ *to* (4) *and the solution set of* (4) *is equicontinuous.*

Now we introduce the notion of viability.

Definition 3. Let $K \subset \mathbb{R}^n$ be a closed set and let $y_0 \in K$. We say that K is *viable* with respect to (2) if there exists a solution $y(\cdot)$ to (2) such that $y(t) \in K$ for $t \in [t_0,T]$. This kind of solutions we call *viable*.

One of the main difficulties is to establish the tangency condition taking into account the specificity of the fractional system with Caputo–Fabrizio derivative. In order to deliver approximate solutions remaining close to the set K we propose the following tangency condition.

Definition 4. Suppose $t_0 \leq \bar{t} < T$. Let $E \in \mathbb{R}^n$ be bounded, $f(\cdot) \in L^\infty([t_0,T],\mathbb{R}^n)$ and let us define $y(t) = c + a_\alpha f(t) + b_\alpha \int_{t_0}^t f(s)ds$. The pair (f,E) is said to be tangent to $[t_0,T]\times K$ at $(\bar{t},\bar{y}) \in [t_0,T]\times K$ if $y(\bar{t}) = \bar{y}$ and

$$\liminf_{h\to 0^+} \frac{1}{h} d(\bar{y} + a_\alpha r(\bar{t},f(\cdot))(h) + h\cdot b_\alpha E, K) = 0\,, \tag{5}$$

where $r(\bar{t},f(\cdot))(h) = f(\bar{t}+h, y(\bar{t}+h)) - f(\bar{t},\bar{y})$.

Remark 5. Observe that the above tangency condition one can understand as follows. For $t_0 \leq \bar{t} < \bar{t}+h \leq T$ a solution to (2) is of the form $y(\bar{t}+h) = c + a_\alpha f_y^F(\bar{t}+h) + b_\alpha \int_{t_0}^{\bar{t}+h} f_y^F(s)ds$
$= c + a_\alpha f_y^F(\bar{t}) + b_\alpha \int_{t_0}^{\bar{t}} f^F(s,y(s))ds + a_\alpha r(\bar{t},f_y^F(\cdot))(h) + b_\alpha \int_{\bar{t}}^{\bar{t}+h} f^F(s,y(s))ds\,,$
where $r(\bar{t},f_y^F(\cdot))(h) = f_y^F(\bar{t}+h) - f_y^F(\bar{t})$ and $f_y^F(\cdot)$ is a measurable function such that $f_y^F(t) \in F(t,y(t))$, $t \in [t_0,T]$. Then having in mind that $F(\cdot,\cdot)$ is u.s.c. we get $\liminf_{h\to 0^+} \frac{1}{h} dist(\bar{y} + a_\alpha r(\bar{t},f_y^F(\cdot))(h) + h\cdot b_\alpha F(\bar{t},\bar{y}), K) = 0$ for $y(\bar{t}) = \bar{y}$. It clearly implies that for every viable solution $y(\cdot)$ to (2) the pair $\big({}^{CF}D^\alpha y, F(t,y(t))\big)$ is tangent to $[t_0,T]\times K$ at $(t,y(t))$ for almost all $t \in [t_0,T]$.

In order to prove existence of viable solution we need a suitable construction of approximate solutions.

Proposition 6. *Let assumption* $(A1)$ *be fullfiled and let* $y(\cdot)$ *be an absolutely continuous function on* $[t_0,\bar{t}]$ *with* $y(t_0) = y_0$ *and* $t_0 \leq \bar{t} < T$. *If the pair* $\big({}^{CF}D^\alpha y(\cdot), F(\bar{t},\bar{y})\big)$ *is tangent to* $[t_0,T]\times K$ *at* $(\bar{t},\bar{y}) \in [t_0,T]\times K$, *then for every* $\varepsilon > 0$ *there exist* $\delta > 0$ *and* ε*-solution* $z(\cdot)$ *to* (2) *on* $[\bar{t},\bar{t}+\delta]$, *which is an extension of* $y(\cdot)$, *such that* $z(\bar{t}+\delta) \in K$.

Finally we are in the position to present the main result of the paper.

Theorem 7. *Let us assume that* $(A1)$ *hold. If the tangency condition is fullfiled by system* (2) *at every* $(\bar{t},\bar{y}) \in [t_0,T]\times K$, *then there exists at least one viable solution to* (2).

 ISBN: 978-84-617-8694-7

Acknowledgements

This work has been supported by Bialystok University of Technology Grant No. S/WI/1/2016.

References

[1] J. P. Aubin and H. Frankowska, *Set-valued analysis*, Boston-Basel-Berlin, Birkäuser, 1990.

[2] J. P. Aubin, A. M. Bayen and P. Saint-Pierre, *Viability Theory, New Directions*, Second edition, Springer-Verlag Berlin Heidelberg, 2011.

[3] Z. Bartosiewicz and E. Girejko, *On generalized differentials, viability and invariance of differential inclusions*, J. Convex Ann. **15**, (4) (2008) 819–830.

[4] M. Caputo and F. Fabrizio, *A new definition of fractional derivative without singular kernel*, Progr. Fract. Differ. Appl. **1:2** (2015) 1–13.

[5] O. Carja, T. Dinchev, M. Rafaqat and R. Ahmed, *Viability of fractional differential inclusions*, Applied Math. Lett. **38** (2014) 48–51.

[6] E. Girejko, D. Mozyrska and M. Wyrwas, *A sufficient condition of viability for fractional differential equations with the Caputo derivative*, J. Math. Anal. and Appl. **381** (2011) 146–154.

[7] D. Mozyrska, E. Girejko and M. Wyrwas,*A necessary condition of viability for fractional differential equations with initialization*, Comp. Math. with Appl. **62** (9) (2011) 3642–3647.

[8] T. Kaczorek, *Selected problems of fractional systems theory*, volume 411, Springer, 2011.

[9] A. A. Kilbas, H. M. Srivastava, and J. J. Trujillo, *Theory and applications of fractional differential equations*, Amsterdam: North–Holland Mathematics Studies, 204. Elsevier Science B. V., 2006.

[10] A. B. Malinowska and D. F. M. Torres, *Introduction to the Fractional Calculus of Variations*, Imperial College Press London, 2012.

[11] I. Podlubny, *Fractional Differential Equations*, Academic Press, San Diego-Boston-New York-London-Tokyo-Toronto, 1999.

[12] J. Vasundgaradevi and V. Lakshmikantham, *Nonsmooth analysis and fractional differential equations*, Nonlinear Anal. **70** (2009) 4151–4157.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

Probabilistic evolution theory for explicit autonomous ordinary differential equations: recursion of squarified telescope matrices and optimal space extension

Coşar Gözükırmızı[1] and Metin Demiralp[2]

[1] *Computer Engineering Department, Beykent University*

[2] *Informatics Institute, Istanbul Technical University*

emails: cosargozukirmizi@beykent.edu.tr, metin.demiralp@gmail.com

Abstract

Probabilistic evolution theory facilitates the solution of initial value problem of explicit autonomous ordinary differential equations with second degree multinomial right hand side functions. Its formulation has components we call telescope matrices. The matrices grow in size very rapidly and has many zeroes and repeating structures. In order to avoid the computational complexity coming from telescope matrices, squarified telescope matrices are utilized. Their calculation is through a recursion. This recursion has been used in several works by the authors and their colleagues but its proof was not given. This work gives the proof of the recursion and all the surrounding details. A second purpose of this work is to provide a method for most facilitative (optimal) space extension: using probabilistic evolution theory when degree of multinomiality of the right hand side functions is more than two. For this purpose, an approach using method of exhaustion (brute-force) is proposed.

Key words: probabilistic evolution theory, squarification, Kronecker multiplication, telescope matrices, space extension

1 Probabilistic Evolution Theory

The problem under consideration is the initial value problem of explicit autonomous ODE set where the right hand side functions are second degree multinomials [1, 2]. It is in the compact form

$$\begin{aligned} \dot{\mathbf{x}}(t) &= \mathbf{F}_0 + \mathbf{F}_1\mathbf{x}(t) + \mathbf{F}_2\mathbf{x}(t)^{\otimes 2}, \\ &\quad \mathbf{x}(0) = \boldsymbol{\alpha} \end{aligned} \tag{1}$$

 ISBN: 978-84-617-8694-7

where $\otimes 2$ on the exponent represents taking the second Kronecker power: Kronecker product of a vector by itself. ODEs with a higher degree of multinomiality may be put in the form in (1) using space extension. Section 3 has information about space extension. In (1), $\mathbf{F}_0$ is a vector, $\mathbf{F}_1$ is a square matrix and $\mathbf{F}_2$ is a rectangular matrix. Considering that $\mathbf{x}(t)$ in (1) is a vector with n elements, then $\mathbf{F}_0$ is of type $n \times 1$, $\mathbf{F}_1$ is of type $n \times n$ and $\mathbf{F}_2$ is of type $n \times n^2$. $\mathbf{F}_0$, $\mathbf{F}_1$ and $\mathbf{F}_2$ have constants (independent of t) as elements. These constants represent the parameters of the dynamics of the system whose state at time t is represented by the vector $\mathbf{x}(t)$.

Constancy adding space extension (CASE) is utilized to eliminate $\mathbf{F}_0$ at the expense of incrementing each dimension of matrices and vectors by 1. It relies on appending a constant element to $\mathbf{x}(t)$. In addition, the new (augmented) $\mathbf{F}_1$ may be put in a form so that it is proportional to the identity matrix. Therefore, using CASE, (1) may be written in the form

$$\begin{aligned} &\overline{\mathbf{x}}(t) \equiv \begin{bmatrix} \mathbf{x}(t) \\ \alpha_{n+1} \end{bmatrix}, \\ &\dot{\overline{\mathbf{x}}}(t) = \beta\, \overline{\mathbf{x}}(t) + \mathbf{F}\, \overline{\mathbf{x}}(t)^{\otimes 2}, \\ &\overline{\mathbf{x}}(0) \equiv \mathbf{a} \end{aligned} \tag{2}$$

where α_{n+1} is any real nonzero constant. The value of α_{n+1} will not show itself in the final solution: norm minimization makes the result independent of α_{n+1}. β may also be calculated using norm minimization: unlike α_{n+1}, the value of β has an effect on the solution [2]. If a factor $\mathrm{e}^{\beta t}$ is pulled out from $\overline{\mathbf{x}}(t)$, then

$$\begin{aligned} &\overline{\mathbf{x}}(t) \equiv \mathrm{e}^{\beta t}\, \overline{\overline{\mathbf{x}}}(t) \\ &\dot{\overline{\overline{\mathbf{x}}}}(t) = \mathrm{e}^{\beta t}\, \mathbf{F}\, \overline{\overline{\mathbf{x}}}(t)^{\otimes 2}, \\ &\overline{\overline{\mathbf{x}}}(0) = \mathbf{a} \end{aligned} \tag{3}$$

appears. The initial value problem in (3) has only second degree terms. On the other hand, it is not autonomous. If the independent variable is taken as a certain function of t instead of t itself, autonomous behavior may be obtained. This may be shown as

$$\begin{aligned} &u \equiv \frac{\mathrm{e}^{\beta t} - 1}{\beta}, \qquad \mathbf{y}(u) \equiv \overline{\overline{\mathbf{x}}}(t), \\ &\frac{d\mathbf{y}}{du} = \mathbf{F}\, \mathbf{y}(u)^{\otimes 2}, \\ &\mathbf{y}(0) = \mathbf{a} \end{aligned} \tag{4}$$

where taking β as 0 is a special case in which u is equal to t. The initial value problem in (4) is autonomous. In order to form the solution, the derivative of Kronecker power of a

 ISBN: 978-84-617-8694-7

vector is necessary. This may be performed using Leibniz rule as follows

$$\begin{aligned}\frac{d}{du}\left(\mathbf{y}(u)^{\otimes m}\right) &= \sum_{j=0}^{m-1}\mathbf{y}(u)^{\otimes j}\otimes\frac{d\mathbf{y}(u)}{du}\otimes\mathbf{y}(u)^{\otimes m-j-1},\\ &= \left(\sum_{j=0}^{m-1}\mathbf{I}_n^{\otimes j}\otimes\mathbf{F}\otimes\mathbf{I}_n^{\otimes m-j-1}\right)\mathbf{y}(u)^{\otimes m+1},\\ &\qquad m=0,1,2,...\end{aligned}\tag{5}$$

where $\mathbf{I}_n$ is $n\times n$ identity matrix. In (5) and from here on, n shows the number of elements in the system vector after constancy adding space extension (CASE). (5) relies on distributive property of product over Kronecker product. There is also the distributive property of Kronecker product over product. The matrix coefficients in (5) are named "monocular matrices". They are defined as follows.

$$\begin{aligned}&\mathbf{M}_m\equiv\left(\sum_{j=0}^{m-1}\mathbf{I}_n^{\otimes j}\otimes\mathbf{F}\otimes\mathbf{I}_n^{\otimes m-j-1}\right),\qquad m=0,1,2,...\\ &\frac{d}{du}\left(\mathbf{y}(u)^{\otimes m}\right)=\mathbf{M}_m\mathbf{y}(u)^{\otimes m+1},\\ &\mathbf{y}(0)^{\otimes m}=\mathbf{a}^{\otimes m+1}\qquad m=0,1,2,...\end{aligned}\tag{6}$$

Integrating both sides of the ODE in (6),

$$\mathbf{y}(u)^{\otimes m}=\mathbf{a}^{\otimes m+1}+\mathbf{M}_m\int_0^u d\upsilon\mathbf{y}(\upsilon)^{\otimes m+1},\quad m=0,1,2,...\tag{7}$$

may be obtained. Proceeding in the same manner,

$$\begin{aligned}\mathbf{y}(u) &= \mathbf{a}+\mathbf{M}_1\int_0^u d\upsilon\mathbf{y}(\upsilon)^{\otimes 2}\\ &= \mathbf{T}_0\mathbf{a}+u\mathbf{T}_1\mathbf{a}^{\otimes 2}+\mathbf{T}_2\int_0^u d\upsilon_1\int_0^{\upsilon_1}d\upsilon_2\mathbf{y}\left(\upsilon_2\right)^{\otimes 3}\\ &= \sum_{j=0}^{2}\frac{u^j}{j!}\mathbf{T}_j\mathbf{a}^{\otimes j+1}+\mathbf{T}_3\int_0^u d\upsilon_1\int_0^{\upsilon_1}d\upsilon_2\int_0^{\upsilon_2}d\upsilon_3\mathbf{y}\left(\upsilon_3\right)^{\otimes 4}\end{aligned}\tag{8}$$

appears where $\mathbf{T}$ matrices are telescope matrices: they are ordered products of monocular matrices defined as follows.

$$\mathbf{M}_0\equiv\mathbf{I}_n,\qquad\mathbf{T}_m\equiv\mathbf{M}_0...\mathbf{M}_m,\qquad m=0,1,2,...\tag{9}$$

 ISBN: 978-84-617-8694-7

The generalization of (8) gives

$$\mathbf{y}(u)=\sum_{j=0}^{m}\frac{u^{j}}{j!}\mathbf{T}_{j}\mathbf{a}^{\otimes j+1}+\mathbf{T}_{m+1}\int_{0}^{u}dv_{1}...\int_{0}^{v_{m}}dv_{m+1}\mathbf{y}\left(v_{m+1}\right)^{\otimes m+2},\quad m=0,1,2,... \tag{10}$$

and taking m to infinity by taking into consideration that the integral in (10) approaches zero when m is increased,

$$\mathbf{y}(u)=\sum_{j=0}^{\infty}\frac{u^{j}}{j!}\mathbf{T}_{j}\mathbf{a}^{\otimes j+1} \tag{11}$$

appears. (11) may be used in order to calculate the solution of the ODE given in (1). The only drawback is the size of telescope matrices: $\mathbf{T}_j$ is an $n\times n^{j+1}$ matrix.

2 Recursion of Squarified Telescope Matrices

In previous works, recursion of squarified telescope matrices is given [3] and is applied to many different problems. On the other hand, proof of recursion of squarified telescope matrices is given in this paper. Squarification is utilized so that (11) is simplified. Assume that there is a matrix $\mathbf{S}_m(\mathbf{a})$ such that

$$\mathbf{T}_{m}\mathbf{a}^{\otimes m+1}\equiv\mathbf{S}_{m}\left(\mathbf{a}\right)\mathbf{a},\qquad m=0,1,2,... \tag{12}$$

holds. $\mathbf{S}_m(\mathbf{a})$ is named "squarified telescope matrix" (SquTelMat). The calculation of these matrices may be performed by the use of a recursion. This section shows the proof of the recursion between SquTelMats. Using (12) in (11),

$$\mathbf{y}(u)=\left\{\sum_{j=0}^{\infty}\frac{u^{j}}{j!}\mathbf{S}_{j}\left(\mathbf{a}\right)\right\}\mathbf{a} \tag{13}$$

is obtained. The derivative of $\mathbf{y}(u)$ may be found using (13). Differentiating (13), j comes down as a factor and forms $(j-1)!$ on the denominator. The exponent becomes $(j-1)$. The zeroth term of the sum is 0 since $(j-1)!$ goes to infinity when j is 0. Therefore, the infinite sum starts from 1. In order to make it start from 0, all instances of j may be replaced by $(j+1)$. Then,

$$\frac{d\mathbf{y}(u)}{du}=\left\{\sum_{j=0}^{\infty}\frac{u^{j}}{j!}\mathbf{S}_{j+1}\left(\mathbf{a}\right)\right\}\mathbf{a} \tag{14}$$

 ISBN: 978-84-617-8694-7

may be obtained. Therefore, differentiation corresponds to increasing the subindex of SquTelMats. The right hand side of the ODE is in the form

$$\mathbf{F}\mathbf{y}(u)^{\otimes 2} = \lfloor\, \mathbf{F}, \mathbf{y}(u) \,\rceil\, \mathbf{y}(u) \tag{15}$$

where the first factor on the right hand side of (15) is a squarification. A squarification takes two operands: a matrix and a vector. The result of a squarification is a vector which is a linear combination of the square blocks of the matrix where the elements of the vector are the linear combination coefficients. Squarification is a linear operation with respect to its vector operand. The squarification in (15) is

$$\lfloor\, \mathbf{F}, \mathbf{y}(u) \,\rceil = \sum_{j=0}^{\infty} \frac{u^j}{j!} \lfloor\, \mathbf{F}, \mathbf{S}_j(\mathbf{a})\, \mathbf{a} \,\rceil \tag{16}$$

by using (13) and linearity of squarification with respect to its vector operand. Multiplying both sides of (16) by $\mathbf{y}(u)$ from right and using Cauchy product of two infinite series,

$$\begin{aligned}
\lfloor\, \mathbf{F}, \mathbf{y}(u) \,\rceil\, \mathbf{y}(u) &= \left\{ \sum_{k=0}^{\infty} \frac{u^k}{k!} \lfloor\, \mathbf{F}, \mathbf{S}_k(\mathbf{a})\, \mathbf{a} \,\rceil \sum_{j=0}^{\infty} \frac{u^j}{j!} \mathbf{S}_j(\mathbf{a}) \right\} \mathbf{a} \\
&= \left\{ \sum_{k=0}^{\infty} \frac{u^k}{k!} \lfloor\, \mathbf{F}, \mathbf{S}_k(\mathbf{a})\, \mathbf{a} \,\rceil \sum_{j=k}^{\infty} \frac{u^{j-k}}{(j-k)!} \mathbf{S}_{j-k}(\mathbf{a}) \right\} \mathbf{a} \\
&= \left\{ \sum_{j=0}^{\infty} \frac{u^j}{j!} \sum_{k=0}^{j} \binom{j}{k} \lfloor\, \mathbf{F}, \mathbf{S}_k(\mathbf{a})\, \mathbf{a} \,\rceil\, \mathbf{S}_{j-k}(\mathbf{a}) \right\} \mathbf{a}
\end{aligned} \tag{17}$$

is obtained. Substituting (17) in (15) and using (14) and (4),

$$\sum_{j=0}^{\infty} \frac{u^j}{j!} \mathbf{S}_{j+1}(\mathbf{a})\, \mathbf{a} = \sum_{j=0}^{\infty} \frac{u^j}{j!} \sum_{k=0}^{j} \binom{j}{k} \lfloor\, \mathbf{F}, \mathbf{S}_k(\mathbf{a})\, \mathbf{a} \,\rceil\, \mathbf{S}_{j-k}(\mathbf{a})\, \mathbf{a} \tag{18}$$

appears. (18) has the following consequence: for a certain power of u, the coefficient vector on the left side of (18) should be equal to the coefficient vector on the right side of (18). Therefore,

$$\begin{aligned}
\mathbf{S}_{j+1}(\mathbf{a})\, \mathbf{a} &= \sum_{k=0}^{j} \binom{j}{k} \lfloor\, \mathbf{F}, \mathbf{S}_k(\mathbf{a})\, \mathbf{a} \,\rceil\, \mathbf{S}_{j-k}(\mathbf{a})\, \mathbf{a} \\
j = 0, 1, 2, \ldots \qquad & \mathbf{S}_0 = \mathbf{T}_0 = \mathbf{I}_n
\end{aligned} \tag{19}$$

should be the case in order for (18) to hold for any value of u. (19) is a recursion between vectors. The initial value of the recursion is $\mathbf{S}_0\mathbf{a}$. (19) has been used for the solution of

 ISBN: 978-84-617-8694-7

certain initial value problems in previous proceedings by the authors of this paper and their colleagues [4, 5]. The fact that it is possible to obtain (19) from the original ODE set and the definition of squarification is important: the direct proof for the recursion between the images of the (augmented) initial vectors under SquTelMats is complete. The recursion in (19) suffices for finding the solution of an initial value problem by probabilistic evolution theory. In fact, the use of the vector recursion in (19) is more efficient than the matrix recursion between the squarified telescope matrices. On the other hand, it is also necessary to show that the vector recursion in (19) may be used in order to prove the matrix recursion between the squarified telescope matrices. This is done by proof by contradiction. Assuming that the solution to the matrix recursion is not unique, thus adding an $\mathbf{R}_j(\mathbf{a})$ matrix to each $\mathbf{S}_j(\mathbf{a})$, it is possible to find that $\mathbf{R}_j(\mathbf{a})$ matrices should all be zero matrices.

$$\begin{aligned}
&\mathbf{S}_{j+1}(\mathbf{a}) = \sum_{k=0}^{j} \binom{j}{k} \lfloor\, \mathbf{F}, \mathbf{S}_k(\mathbf{a})\,\mathbf{a} \,\rceil\, \mathbf{S}_{j-k}(\mathbf{a}) \\
&\mathbf{S}_j(\mathbf{a}) \equiv \overline{\mathbf{S}}_j(\mathbf{a}) + \mathbf{R}_j(\mathbf{a}), \qquad \mathbf{R}_j(\mathbf{a})\,\mathbf{a} = 0, \\
&\overline{\mathbf{S}}_{j+1}(\mathbf{a}) = \sum_{k=0}^{j} \binom{j}{k} \lfloor\, \mathbf{F}, \overline{\mathbf{S}}_k(\mathbf{a})\,\mathbf{a} \,\rceil\, \overline{\mathbf{S}}_{j-k}(\mathbf{a}) \\
&\mathbf{R}_{j+1}(\mathbf{a}) = \sum_{k=0}^{j} \binom{j}{k} \lfloor\, \mathbf{F}, \overline{\mathbf{S}}_k(\mathbf{a})\,\mathbf{a} \,\rceil\, \mathbf{R}_{j-k}(\mathbf{a}) \\
&\qquad j = 0, 1, 2, ... \qquad \overline{\mathbf{S}}_0 = \mathbf{I}_n, \qquad \mathbf{R}_0(\mathbf{a}) = \mathbf{0}
\end{aligned} \tag{20}$$

The recursion at the end of (20) is a linear recursion. Its solution may be performed by the following steps. Creating infinite sums at both sides of the recursion in (20),

$$\sum_{j=0}^{\infty} \frac{u^j}{j!} \mathbf{R}_{j+1}(\mathbf{a}) = \sum_{j=0}^{\infty} \frac{u^j}{j!} \sum_{k=0}^{j} \binom{j}{k} \lfloor\, \mathbf{F}, \overline{\mathbf{S}}_k(\mathbf{a})\,\mathbf{a} \,\rceil\, \mathbf{R}_{j-k}(\mathbf{a}) \tag{21}$$

and then considering that the left side is the derivative of a certain entity,

$$\begin{aligned}
\frac{d}{du}\left(\sum_{j=0}^{\infty} \frac{u^j}{j!} \mathbf{R}_j(\mathbf{a})\right) &= \sum_{k=0}^{\infty} \frac{u^k}{k!} \lfloor\, \mathbf{F}, \overline{\mathbf{S}}_k(\mathbf{a})\,\mathbf{a} \,\rceil \sum_{j=0}^{\infty} \frac{u^j}{j!} \mathbf{R}_j(\mathbf{a}) \\
\sum_{j=0}^{\infty} \frac{u^j}{j!} \mathbf{R}_j(\mathbf{a}) &= \exp\left(\sum_{k=0}^{\infty} \frac{u^k}{k!} \lfloor\, \mathbf{F}, \overline{\mathbf{S}}_k(\mathbf{a})\,\mathbf{a} \,\rceil\right) \mathbf{R}_0(\mathbf{a}) \\
&\mathbf{R}_j(\mathbf{a}) = \mathbf{0}, \qquad j = 0, 1, 2, ...
\end{aligned} \tag{22}$$

 ISBN: 978-84-617-8694-7

appears. The last equality of (22) shows that all $\mathbf{R}_j(\mathbf{a})$ must be zero since $\mathbf{R}_0(\mathbf{a})$ is $\mathbf{0}$. Vector recursion yields the matrix recursion.

$$\mathbf{S}_{j+1}(\mathbf{a})\,\mathbf{a} = \sum_{k=0}^{j} \binom{j}{k} \lfloor\, \mathbf{F}, \mathbf{S}_k(\mathbf{a})\,\mathbf{a} \,\rceil\, \mathbf{S}_{j-k}(\mathbf{a})\,\mathbf{a}$$

$$\mathbf{S}_{j+1}(\mathbf{a}) = \sum_{k=0}^{j} \binom{j}{k} \lfloor\, \mathbf{F}, \mathbf{S}_k(\mathbf{a})\,\mathbf{a} \,\rceil\, \mathbf{S}_{j-k}(\mathbf{a})$$

$$j = 0, 1, 2, ... \qquad \mathbf{S}_0 = \mathbf{T}_0 = \mathbf{I}_n \tag{23}$$

Therefore, the matrix recursion is

$$\mathbf{S}_{j+1}(\mathbf{a}) = \sum_{k=0}^{j} \binom{j}{k} \lfloor\, \mathbf{F}, \mathbf{S}_k(\mathbf{a})\,\mathbf{a} \,\rceil\, \mathbf{S}_{j-k}(\mathbf{a})$$

$$j = 0, 1, 2, ... \qquad \mathbf{S}_0 = \mathbf{I}_n \tag{24}$$

and it may be used to calculate all squarified telescope matrices.

Another way to see that the recursion is valid is to start with the recursion and come back to the ODE. Multiplying both sides of (24) by $\mathbf{a}$ from right

$$\mathbf{S}_{j+1}(\mathbf{a})\,\mathbf{a} = \sum_{k=0}^{j} \binom{j}{k} \lfloor\, \mathbf{F}, \mathbf{S}_k(\mathbf{a})\,\mathbf{a} \,\rceil\, \mathbf{S}_{j-k}(\mathbf{a})\,\mathbf{a}$$

$$j = 0, 1, 2, ... \qquad \mathbf{S}_0\mathbf{a} = \mathbf{a} \tag{25}$$

and forming an infinite sum in both sides of (25),

$$\sum_{j=0}^{\infty} \frac{u^j}{j!} \mathbf{S}_{j+1}(\mathbf{a})\mathbf{a} = \sum_{j=0}^{\infty} \frac{u^j}{j!} \sum_{k=0}^{j} \binom{j}{k} \lfloor\, \mathbf{F}, \mathbf{S}_k(\mathbf{a})\,\mathbf{a} \,\rceil\, \mathbf{S}_{j-k}(\mathbf{a})\,\mathbf{a} \tag{26}$$

may be obtained. Simplifying the factorials and rewriting the summation in the triangular region

$$\sum_{j=0}^{\infty} \frac{u^j}{j!} \mathbf{S}_{j+1}(\mathbf{a})\mathbf{a} = \sum_{k=0}^{\infty} \sum_{j=k}^{\infty} \frac{u^j}{k!(j-k)!} \lfloor\, \mathbf{F}, \mathbf{S}_k(\mathbf{a})\,\mathbf{a} \,\rceil\, \mathbf{S}_{j-k}(\mathbf{a})\,\mathbf{a} \tag{27}$$

appears. Using Cauchy product rule in (27),

$$\sum_{j=0}^{\infty} \frac{u^j}{j!} \mathbf{S}_{j+1}(\mathbf{a})\mathbf{a} = \sum_{k=0}^{\infty} \frac{u^k}{k!} \lfloor\, \mathbf{F}, \mathbf{S}_k(\mathbf{a})\,\mathbf{a} \,\rceil\, \sum_{j=0}^{\infty} \frac{u^j}{j!} \mathbf{S}_j(\mathbf{a})\mathbf{a} \tag{28}$$

 ISBN: 978-84-617-8694-7

is obtained. Using the linearity of squarification

$$\sum_{j=0}^{\infty}\frac{u^j}{j!}\mathbf{S}_{j+1}(\mathbf{a})\mathbf{a}=\left\lfloor\ \mathbf{F},\sum_{k=0}^{\infty}\frac{u^k}{k!}\mathbf{S}_k\left(\mathbf{a}\right)\mathbf{a}\ \right\rceil\sum_{j=0}^{\infty}\frac{u^j}{j!}\mathbf{S}_j(\mathbf{a})\mathbf{a} \tag{29}$$

appears. The next step is to see that the left hand side of (29) is the derivative of a term having $\mathbf{S}_j(\mathbf{a})$. Therefore,

$$\frac{d}{du}\left(\sum_{j=0}^{\infty}\frac{u^j}{j!}\mathbf{S}_j(\mathbf{a})\mathbf{a}\right)=\left\lfloor\ \mathbf{F},\sum_{k=0}^{\infty}\frac{u^k}{k!}\mathbf{S}_k\left(\mathbf{a}\right)\mathbf{a}\ \right\rceil\sum_{j=0}^{\infty}\frac{u^j}{j!}\mathbf{S}_j(\mathbf{a})\mathbf{a} \tag{30}$$

which is a matrix ODE. Making the following definition

$$\mathbf{g}(u)\equiv\sum_{j=0}^{\infty}\frac{u^j}{j!}\mathbf{S}_j(\mathbf{a})\mathbf{a} \tag{31}$$

(30) may be written in the compact form

$$\frac{d}{du}(\mathbf{g}(u))=\lfloor\ \mathbf{F},\mathbf{g}(u)\ \rceil\,\mathbf{g}(u) \tag{32}$$

and using the definition of squarification

$$\frac{d\mathbf{g}(u)}{du}=\mathbf{F}\mathbf{g}(u)^{\otimes 2},\quad \mathbf{g}(0)=\mathbf{a} \tag{33}$$

may be obtained. This was the original ODE. Consequently, it is also possible to show that the recursion is valid by starting with the recursion and finding out that it is a recursion of the matrices in the solution of the ODE in (33). It is possible to replace u by t using (4), to obtain

$$\overline{\mathbf{g}}(t)\equiv \mathrm{e}^{\beta t}\sum_{j=0}^{\infty}\frac{(\mathrm{e}^{\beta t}-1)^j}{j!}\mathbf{T}_j\mathbf{a}^{\otimes j+1} \tag{34}$$

which is the solution via telescope matrices.

3 Optimal Space Extension

In the vector form of the ODE, the matrix coefficient corresponding to the first degree terms is forced to be proportional to the identity matrix. If the proportionality constant is chosen as 0, the exponential term does not appear in the probabilistic evolution theory solution. In that case, the distinction between space extension and constancy adding space extension (CASE) disappears; space extension by itself suffices to form purely second degree multinomial right hand side functions. This special case is worth studying, because it gives a lot of information about what the most facilitative space extension may be. This section focuses on finding the most facilitative space extension where the proportionality constant corresponding to first degree terms (β) is 0.

 ISBN: 978-84-617-8694-7

3.1 Basis Set Utilization

A multivariate basis set is used for showing the multinomial right hand side functions. It is given as follows

$$\begin{aligned} u^{(\ell_1,\ldots,\ell_n)}(x_1,\ldots,x_n) &= x_1^{\ell_1} x_2^{\ell_2} x_3^{\ell_3} \ldots x_n^{\ell_n} \\ \ell_1 &= 0,\ldots,m_1 \\ \vdots &\qquad \vdots \\ \ell_n &= 0,\ldots,m_n \end{aligned} \tag{35}$$

where each function has n upper indices. The functions are linearly independent. Then, the ODE set with n equations and n unknowns may be shown in the form

$$\begin{aligned} \dot{x}_1(t) &= \sum_{\ell_1,\ldots,\ell_n} a_1^{(\ell_1,\ldots,\ell_n)} u^{(\ell_1,\ldots,\ell_n)}(x_1(t),\ldots,x_n(t)) \\ \dot{x}_2(t) &= \sum_{\ell_1,\ldots,\ell_n} a_2^{(\ell_1,\ldots,\ell_n)} u^{(\ell_1,\ldots,\ell_n)}(x_1(t),\ldots,x_n(t)) \\ \vdots &\qquad\qquad \vdots \\ \dot{x}_n(t) &= \sum_{\ell_1,\ldots,\ell_n} a_n^{(\ell_1,\ldots,\ell_n)} u^{(\ell_1,\ldots,\ell_n)}(x_1(t),\ldots,x_n(t)) \end{aligned} \tag{36}$$

where the right hand sides have n summations. The coefficients also have n upper indices. The lower indices of the coefficients show which equation a certain coefficient belongs to. The idea here is to show the ODE set in terms of the basis set and then use space extension to form purely second degree structures. The sought functions, the derivatives of which appear on the left hand side of (36), may also be shown in terms of the basis set with basis functions having one upper index being 1 and all others 0. The temporal derivatives of the basis functions are

$$\dot{u}^{(\ell_1,\ldots,\ell_n)}(x_1,\ldots,x_n) = \frac{\partial u^{(\ell_1,\ldots,\ell_n)}(x_1,\ldots,x_n)}{\partial x_1}\dot{x}_1(t) + \cdots + \frac{\partial u^{(\ell_1,\ldots,\ell_n)}(x_1,\ldots,x_n)}{\partial x_n}\dot{x}_n(t) \tag{37}$$

where the derivatives with respect to sought functions are

$$\begin{aligned} \frac{\partial u^{(\ell_1,\ldots,\ell_n)}(x_1,\ldots,x_n)}{\partial x_1} &= \ell_1 x_1^{\ell_1-1}\left(x_2^{\ell_2}\ldots x_n^{\ell_n}\right) \\ \frac{\partial u^{(\ell_1,\ldots,\ell_n)}(x_1,\ldots,x_n)}{\partial x_2} &= x_1^{\ell_1}\ell_2 x_2^{\ell_2-1}\left(x_3^{\ell_3}\ldots x_n^{\ell_n}\right) \\ \vdots &\qquad\qquad \vdots \\ \frac{\partial u^{(\ell_1,\ldots,\ell_n)}(x_1,\ldots,x_n)}{\partial x_n} &= \left(x_1^{\ell_1}\ldots x_{n-1}^{\ell_{n-1}}\right)\ell_n x_n^{\ell_n-1}. \end{aligned} \tag{38}$$

 ISBN: 978-84-617-8694-7

Using (38) in (37), temporal derivatives of the basis functions appear as

$$\begin{aligned}\dot{u}^{(\ell_1,\ldots,\ell_n)}(x_1,\ldots,x_n) &= \ell_1 u^{(\ell_1-1,\ell_2,\ell_3,\ldots,\ell_n)}(x_1,\ldots,x_n)\sum_{j_1,\ldots,j_n} a_1^{(j_1,\ldots,j_n)} u^{(j_1,\ldots,j_n)}(x_1,\ldots,x_n)\\ &+ \ell_2 u^{(\ell_1,\ell_2-1,\ell_3,\ldots,\ell_n)}(x_1,\ldots,x_n)\sum_{j_1,\ldots,j_n} a_2^{(j_1,\ldots,j_n)} u^{(j_1,\ldots,j_n)}(x_1,\ldots,x_n)\\ &+ \cdots\\ &+ \ell_n u^{(\ell_1,\ell_2,\ell_3,\ldots,\ell_n-1)}(x_1,\ldots,x_n)\sum_{j_1,\ldots,j_n} a_n^{(j_1,\ldots,j_n)} u^{(j_1,\ldots,j_n)}(x_1,\ldots,x_n)\ .\end{aligned} \tag{39}$$

This representation will be utilized when introducing new equations in space extension.

3.2 Basis Set with a Singularity

The philosophy stays the same when working with singularities. Consider the case where the basis set is

$$\begin{gathered}u^{(\ell_1,\ldots,\ell_n)}(x_1,\ldots,x_n) = x_1^{\ell_1}x_2^{\ell_2}x_3^{\ell_3}\ldots x_n^{\ell_n}\\ \ell_1 = 0,\ldots,m_1\\ \vdots \qquad\qquad \vdots\\ \ell_{i-1} = 0,\ldots,m_{i-1}\\ \ell_i = -s,\ldots,m_i\\ \ell_{i+1} = 0,\ldots,m_{i+1}\\ \vdots \qquad\qquad \vdots\\ \ell_n = 0,\ldots,m_n\end{gathered} \tag{40}$$

Then, (37), (38) and (39) are still valid.

3.3 Applications

A brute-force method based on the formulation above is formed and it is used for the solution of classical quartic anharmonic oscillator and gravitational two body problem. A computer program using Maxima language is formed. The program finds the optimal space extension of classical quartic anharmonic oscillator and it is adaptable to other systems. The application details will be given in the conference presentation.

 ISBN: 978-84-617-8694-7

4 Concluding Remarks

Recursion of squarification is one of the primary reasons why probabilistic evolution theory is powerful. This recursion was conjectured in a previous work by the authors and it was applied to different problems with success. The step by step direct proof of the recursion is given in this paper.

Also, a brute-force approach for making ODE sets with multinomial right hand side functions (with a degree greater than two) into an ODE set with purely second degree multinomial right hand side functions is proposed. It relies on finding the number of equations in all the possible space extensions and choosing the one with the smallest number of equations: therefore minimal number of equations is guaranteed.

References

[1] Coşar Gözükırmızı and Metin Demiralp. Probabilistic evolution approach for the solution of explicit autonomous ordinary differential equations. part 1: Arbitrariness and equipartition theorem in kronecker power series. *Journal of Mathematical Chemistry*, 52(3):866–880, 2014.

[2] Coşar Gözükırmızı and Metin Demiralp. Probabilistic evolution approach for the solution of explicit autonomous ordinary differential equations. part 2: Kernel separability, space extension, and, series solution via telescopic matrices. *Journal of Mathematical Chemistry*, 52(3):881–898, 2014.

[3] Coşar Gözükırmızı, Melike Ebru Kırkın, and Metin Demiralp. Probabilistic evolution theory for the solution of explicit autonomous ordinary differential equations: squarified telescope matrices. *Journal of Mathematical Chemistry*, 55(1):175–194, 2017.

[4] Coşar Gözükırmızı and Melike Ebru Kırkın. Classical symmetric fourth degree potential systems in probabilistic evolution theoretical perspective: Most facilitative conicalization and squarification of telescope matrices. *AIP Conference Proceedings*, 1798(1):020061, 2017.

[5] Coşar Gözükırmızı and Elif Tataroğlu. Squarification of telescope matrices in the probabilistic evolution theoretical approach to the two particle classical mechanics as an illustrative implementation. *AIP Conference Proceedings*, 1798(1):020062, 2017.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference on Computational and Mathematical Methods in Science and Engineering, CMMSE 2017 4–8 July, 2017.

Digital Image Sequence Processing via Tridiagonal Folmat Enhanced Multivariance Products Representation (TFEMPR)

Zeynep Gündoğar [1] **and Metin Demiralp**[1]

[1] *Informatics Institute, Computational Science and Engineering Department, İstanbul Technical University*

emails: gundogarz@itu.edu.tr, metin.demiralp@gmail.com

Abstract

In this paper, we focus on the recently developed decomposition method we have called "Tridiagonal Folmat Enhanced Multivariance Products Representation (TFEM PR)". This method uses recently defined mathematical objects called folded matrix (folmat) and folded vector (folvec), and, is designed to reduce the computation cost by avoiding the decomposition on each way (direction, dimension). This method has now been applied on digital image sequence data like digital videos in our recent works including the present one. The method mentioned here is applicable to many areas such as digital image, digital image sequence, signal processing and also compressing problems due to the convenience of its nature.

Key words: Tridiagonal Folmat Enhanced Multivariance Products Representation (TFEMPR), Folded Vectors, Folded Matrices, Folded Arrays, Multiway Arrays, Decomposition, Factorization, Digital Image Processing

MSC 2000: 94A08, 41Axx, 49M27

1 Introduction

Multiway array decomposition is a commonly focused issue in image processing, neuroscience (like on FMRI data), data science, mathematics etc. Tucker Decomposition [1, 2],Lathauwer's Multilinear Singular Value Decomposition [3–6], PARAFAC, SVD are amongst the well-known decomposition methods [7–10]. One of commonly used decomposition method is High Dimensional Model Representation (HDMR) which has been proposed by Sobol [11] and developed by Rabitz and Demiralp [12–15] beside some others whose works can be found in scientific literature. Enhanced Multivariance Products Representation (EMPR) is another decomposition method arisen from reinforcing HDMR with supports. This method is developed by Demiralp and his group [16–20] during HDMR implementations on the image reconstruction problem.

Basically EMPR of a matrix which is two way array is explicitly given as,

$$\mathbf{A} = a_0 \mathbf{u}\mathbf{v}^T + \mathbf{a_1}\mathbf{v}^T + \mathbf{u}\mathbf{a_2}^T + \mathbf{A_{1,2}}. \tag{1}$$

The representation includes $2^2 = 4$ components; respectively constant component (a_0), the univariate components respectively in direction of row and column ($\mathbf{a_1}$ and $\mathbf{a_2}$) and

 ISBN: 978-84-617-8694-7

bivariate component which is also called as remainder term($\mathbf{A_{1,2}}$). Here, $\mathbf{u}$ and $\mathbf{v}$ are known as the support vectors which enhance the multivariance of each term to 2th order multivariance. The representation turns out to be High Dimensional Model Representation (HDMR) of the matrix under consideration by taking each support equal to a normalized vector whose all elements are same. We do not intend to give details about the determination of EMPR components. Details can be found in papers [16–20].

Remainder term has gained more importance during the image reconstruction problems in consequence of being located major information of image in. To extract this information, EMPR is applied remainder term recursively. When whole information is extracted from remainder, a factorization is obtained. This factorization and decomposition method is developed by Demiralp and his group and called as Tridiagonal Matrix Enhanced Multivariance Products Representation (TMEMPR) [21–23].

The paper is organised as follows. In section 2 we give a short explanation about folded matrices proposed by Demiralp [24,25]. Tridiagonal Folmat Enhanced Multivariance Product Representation (TFEMPR), which is the main method of this work, will be given in third section [26–28]. Fourth section includes the implementations about the method for real data and Peak Signal to Noise Ratio (PSNR) concept which is commonly used and known in literature as a measurer of reconstruction quality. The concluding remarks will be summarized in the final section.

2 Folded Matrices (Folmats)

Vectors and matrices in ordinary linear algebra are respectively considered as one-way and two-way arrays in literature. Matrices have two spaces (row and column space) and also have mapping characters. Folded matrices (Folmats) are defined for making analogy between matrices and multiway arrays and adapt the features of a matrix such as inner product, norm, outer product etc. to multiway arrays. The folarr, folmat and folvec concepts have been proposed by Demiralp [24–28] to facilitate the multiway array analyses.

2.1 Certain Basic Issues in Folmat Definitions

A folmat is in fact a matrix whose rows and columns are folded. In its present format, it can be defined as follows

$$\mathbf{A}_{G_L;G_R} \equiv \left[A_{(i);(j)} \right]_{\forall(i)\in G_L\downarrow}^{\forall(j)\in G_R\rightarrow}, \qquad L \equiv I_1 \times \cdots \times I_m, \quad R \equiv J_1 \times \cdots \times J_n, \tag{2}$$

where (i) and (j) stand for the following tuples

$$(i) \equiv (i_1, ..., i_m), \quad m = 1, 2, 3, ... \qquad (j) \equiv (j_1, ..., j_n), \quad n = 1, 2, 3, ... \tag{3}$$

The general elements of these tuples, i_k and j_ℓ are assumed to take integer values between 1 and I_k inclusive, and, 1 and J_ℓ inclusive respectively. As long as all I_ks are independent of the elements of the m-tuple (i), this implies that the domain of (i) is an orthogonal multidimensional grid whose kth edge contains I_k number of nodes located at the integer values. Same thing also remains valid for the other tuple (j) whose domain is also an orthogonal grid and its ℓth edge contains J_ℓ number of nodes located at the integer values as long as J_ℓs are all independent of the elements of (j).

If we denote the abovementioned grids whose general nodes are (i) and (j) respectively by G_L and G_R then their definitions can be given as follows

$$G_L \equiv \mathbb{Z}_{I_1}^+ \times \cdots \times \mathbb{Z}_{I_m}^+, \qquad G_R \equiv \mathbb{Z}_{J_1}^+ \times \cdots \times \mathbb{Z}_{J_n}^+ \tag{4}$$

 ISBN: 978-84-617-8694-7

where, for a positive integer I, $\mathbb{Z}_I^+$ stands for the finite set composed of first I number of positive integers. These grids can also be introduced by using following open set representations.

$$G_L \equiv \left\{ (i_1, ..., i_m) | \, i_p \in \mathbb{Z}_{I_p}^+, p \in \mathbb{Z}_m^+ \right\}, \qquad G_R \equiv \left\{ (j_1, ..., j_n) | \, j_p \in \mathbb{Z}_{J_p}^+, p \in \mathbb{Z}_n^+ \right\}. \tag{5}$$

As can be noticed immediately, the grids G_L and G_R are sets of m-tuples and n-tuples respectively such that the number of their elements are $I_1 \cdots I_m$ and $J_1 \cdots J_n$ respectively. In these definitions, m and n denotes the dimensions of the Cartesian spaces where G_L and G_R lay respectively. These numbers somehow correspond to the folding orders of rows and columns of an ordinary linear algebraic matrix. Hence we call m and n "row folding level" and "column folding level" respectively. Since they are fundamental properties of a folmat, they should be explicitly or implicitly declared somewhere in the definition of a folmat. We prefer to use types of grids for these declarations. The type definition for a grid is analogue to the type definitions of ordinary linear algebraic matrices. In this connection, the types of the grids G_L and G_R are denoted by $I_1 \times \cdots \times I_m$ and $J_1 \times \cdots \times J_n$ respectively. The folding levels are hidden in these representations such that they are equal to the number of "times" symbol plus 1. Hence, the types define the grids uniquely without leaving any doubt.

There are two ways to declare the type of a grid: (i) to use it as the subscript of the grid symbol G, like $G_{I_1 \times \cdots \times I_m}$ or $G_{J_1 \times \cdots \times J_n}$; (ii) to declare it as the value of a variable like in G_L, $L \equiv I_1 \times \cdots \times I_m$ or in G_R, $R \equiv J_1 \times \cdots \times J_n$. The second option has been used in (2) for typographical reasons. The subscripts L and R in grid symbols have been used to recall the statements "left grid type" and "right grid type" respectively.

We can also now define a "Global Grid" as the ordered pair of these newly defined grids through $G \equiv (G_L, G_R)$

In ordinary linear algebra a matrix can be given by its general term in the form like $a_{i,j}$ where the comma is generally exchanged with the space character by many authors since comma symbols in indices of continuum mechanical objects can be used to denote some kind of derivatives (covariant, and/or contravariant). Continuum mechanics is a quite specific external area for our work, hence we can exclude the danger of confusion in the comma utilization. Beyond this, the comma utilization accentuates the fact that the indices are truely corresponding to separate entities. On the other hand comma separation fits very well to tuple representation. i and j in $a_{i,j}$ correspond to the row and column locations respectively in an ascending ordering starting from 1. Hence we may also call i and j "row locator" and "column locator" respectively. Analogously, the m-tuple (i) and the n-tuple (j) can be considered the row and column locators of the folmat defined in (2) since their values taken from G_L and G_R specify somehow rowwise and columnwise locations. However, the use of grids in the domains of the indices somehow brings a geometrical structuring, "folding" to the issue. Hence we respectively call (i) and (j) "Folded Row Locator" and "Folded Column Locator".

At the right hand side of (2) the left and right brackets have been used to imply that the encompassed entity is a two way array whose rows and columns are folded to more than one ways. The symbol encompassed by the brackets stands for the general element of the multiway array at the focus. This is somehow a matrix, however it is composed of folded rows and columns each of which is denoted by a tuple like m-tuple (i) and n-tuple (j). That is, (i) and (j) stand for pointing to the folded rows and columns of this generalized or geometrically structured matrix.

Since we have used comma as an element separator in tuples we have used semicolon to distinguish these so-called folded row and column locators. The down and right arrows

 ISBN: 978-84-617-8694-7

in the right hand side of (2) imply the downward positionings (folded rows) and rightward positionings (folded columns) respectively. The inclusions of the tuples in the relevant grids have also been explicitly shown in the notation of (2) as sub and super indices. Beyond these the "forall" symbol accentuates that the left and right brackets encompass an array for all elements the relevant grids.

We use these notations for simplicity. We use folmats to get rid of very high computational complexity of each-way-separation-style decomposition of multiway arrays. Semicolon enables binary decomposition by separating indices into two groups. As can be noticed immediately, the important components in the folmat definition are the row and column folding levels and the grids for locating a specific element of folmat. These are for positionings of course and the other but very important entities are the elements of the folmat.

2.2 Folvec Definition

If we specifically choose $n = 0$ in the above definitions then we obtain no grid for columns. In other word, the folmat under consideration becomes an entity which has just a single folded column. We call these entities "folvec"s within an analogy to the matrix-vector discrimination in ordinary linear algebra. Folvec is in fact a special form of folmat. The general term of a folded vector (folvec) can be represented by $a_{(i);}$ where (i) belongs to the grid G_t, $t \equiv I_1 \times \cdots \times I_m$ where t stands for a variable name as any variable name can be used as long as its value declares the type of the relevant grid and unless there appears any confusion matter. The shorthand notation for this is $\mathbf{a}_{G_t;}$, $t \equiv I_1 \times \cdots \times I_m$. The further specification $m = 1$ corresponds to the vector of ordinary linear algebra.

The explicit folvec definition can be given as follows

$$\mathbf{a}_{G_t;} = \left[a_{(i);} \right]_{\forall (i) \in G_t \downarrow}, \qquad t \equiv I_1 \times \cdots \times I_m \tag{6}$$

where the folding level of this folvec is m as we can deduce from the type of the relevant grid.

3 Tridiagonal Folmat Enhanced Multivariance Products Representation (TFEMPR)

Tridiagonal Folmat Enhanced Multivariance Products Representation (TFEMPR) which is one of the recently developed decomposition methods for multiway arrays has the all features of binary decompositions due to the usage of folmat concept [26–28] (also see Sect. 2). TFEMPR method can be considered as higher order analogues of matrix decomposition.

First we will apply EMPR to a general folmat. We can write EMPR of a folmat as,

$$\mathbf{A}_{G_L;G_R} \;=\; a_0 \mathbf{U}_{G_L;} \mathbf{V}^T_{G_R;} + \mathbf{a}^{(1)}_{G_L;} \mathbf{V}^T_{G_R;} + \mathbf{U}_{G_L;} {\mathbf{a}^{(2)}_{G_R;}}^T + \mathbf{A}^{(1,2)}_{G_L;G_R} \tag{7}$$

where $\mathbf{A}_{G_L;G_R}$ stands for a folmat and G_L is called left grid which includes row space indices, G_R is called right grid including column space indices. The subscript L and R stand for two type variables and take the values $I_1 \times \cdots \times I_m$ and $J_1 \times \cdots \times J_n$ respectively, and, Is and Js are the upper limits of the corresponding indices. Beyond these we denote the sizes of grids. G_L and G_R, by μ and ν respectively. These parameters are defined through the equalities, $\mu \equiv I_1 \cdots I_m$, $\nu \equiv J_1 \cdots J_n$. The representation in (7) consists of four terms respectively, the constant term, the term towards left grid, the term towards

 ISBN: 978-84-617-8694-7

right grid and the remainder term ($\mathbf{A}^{(1,2)}_{G_L;G_R}$). $\mathbf{U}_{G_L;}$ and $\mathbf{V}_{G_R;}$ are preselected support folded vectors (folvecs) respectively on left and right grids.

In (7) there are four components to be determined, a_0, $\mathbf{a}^{(1)}_{G_L;}$, $\mathbf{a}^{(2)}_{G_R;}$, $\mathbf{A}^{(1,2)}_{G_L;G_R}$. It is necessary to provide two preconditions for determining the components. First condition is the unit norm standardization and can be given as follows for the left and right grid support folvecs

$$\mathbf{U}^T_{G_L;}\mathbf{U}_{G_L;} = \sum_{(i)\in G_L} U^2_{(i)} = 1, \qquad \mathbf{V}^T_{G_R;}\mathbf{V}_{G_R;} = \sum_{(j)\in G_R} V^2_{(j)} = 1 \tag{8}$$

Support folded vectors should satisfy these unit norm standardization. The second constraints on the components are the vanishing conditions:

$$\mathbf{U}^T_{G_L;}\mathbf{a}^{(1)}_{G_L;} = 0, \qquad {\mathbf{a}^{(2)}_{G_R;}}^T\mathbf{V}_{G_R;} = 0 \tag{9}$$

These conditions point out the orthogonality of support folvec and the relevant component through the same grid. And also $\mathbf{U}_{G_L;}$ should be in the left null space of $\mathbf{A}^{(1,2)}_{G_L;G_R}$ and $\mathbf{V}_{G_R;}$ is in the right null space of $\mathbf{A}^{(1,2)}_{G_L;G_R}$, that is,

$$\mathbf{U}^T_{G_L;}\mathbf{A}^{(1,2)}_{G_L;G_R} = \mathbf{0}_{;G_R}, \qquad \mathbf{A}^{(1,2)}_{G_L;G_R}\mathbf{V}_{G_R;} = \mathbf{0}_{G_L;} \tag{10}$$

Under these conditions we obtain components as follows.

$$a_0 = \mathbf{U}^T_{G_L;}\mathbf{A}_{G_L;G_R}\mathbf{V}_{G_R;} \tag{11}$$

$$\mathbf{a}^{(1)}_{G_L;} = \left(\mathbf{I}_{G_L;G_L} - \mathbf{U}_{G_L;}\mathbf{U}^T_{G_L;}\right)\mathbf{A}_{G_L;G_R}\mathbf{V}_{G_R;} \tag{12}$$

$$\mathbf{a}^{(2)}_{G_R;} = \left(\mathbf{I}_{G_R;G_R} - \mathbf{V}_{G_R;}\mathbf{V}^T_{G_L;}\right)\mathbf{A}^T_{G_L;G_R}\mathbf{U}_{G_L;} \tag{13}$$

$$\mathbf{A}^{(1,2)}_{G_L;G_R} = \left(\mathbf{I}_{G_L;G_L} - \mathbf{U}_{G_L;}\mathbf{U}^T_{G_L;}\right)\mathbf{A}_{G_L;G_R}\left(\mathbf{I}_{G_R;G_R} - \mathbf{V}_{G_R;}\mathbf{V}^T_{G_L;}\right) \tag{14}$$

Details about determination components of the representation (7) can be seen in [26–28]. From now on operations to be taken is for building recursive structure to give TFEMPR.

We use the following definitions to get the first step of the recursion we desire to construct

$$\mathbf{A}^{(0)}_{G_L;G_R} \equiv \mathbf{A}_{G_L;G_R}, \qquad \mathbf{A}^{(1)}_{G_L;G_R} \equiv \mathbf{A}^{(1,2)}_{G_L;G_R}, \qquad \mathbf{U}^{(1)}_{G_L;} = \mathbf{U}_{G_L;}, \qquad \mathbf{V}^{(1)}_{G_R;} = \mathbf{V}_{G_R;}. \tag{15}$$

For each recursion step we can get rid of the folvec components by defining new two support folvecs for rowwise and columnwise directioning such that each of which are orthogonal to its initially given counterpart

$$\mathbf{U}^{(2)}_{G_L;} = \frac{1}{\left\|\mathbf{a}^{(1)}_{G_L;}\right\|}\mathbf{a}^{(1)}_{G_L;}, \qquad \mathbf{V}^{(2)}_{G_R;} = \frac{1}{\left\|\mathbf{a}^{(2)}_{G_R;}\right\|}\mathbf{a}^{(2)}_{G_R;} \tag{16}$$

We also use the following definitions for simplicty

$$\alpha_1 = a_0, \qquad \beta_1 = \left\|\mathbf{a}^{(1)}_{G_L;}\right\|, \qquad \gamma_1 = \left\|\mathbf{a}^{(2)}_{G_R;}\right\| \tag{17}$$

Now, (7) is rewritten by using these definitions

$$\mathbf{A}^{(0)}_{G_L;G_R} = \alpha_1\mathbf{U}_{G_L;}{}^{(1)}{\mathbf{V}^{(1)}_{G_R;}}^T + \beta_1\mathbf{U}_{G_L;}{}^{(2)}{\mathbf{V}^{(1)}_{G_R;}}^T + \gamma_1\mathbf{U}_{G_L;}{}^{(1)}{\mathbf{V}^{(2)}_{G_R;}}^T + \mathbf{A}^{(1)}_{G_L;G_R} \tag{18}$$

 ISBN: 978-84-617-8694-7

which can be considered as the first step of a recursion.

The goal of the recursion we want to construct is to get rid of remainder term. To this end, in each step, we apply EMPR to the remainder term under the support folarrs which have been obtained one step before. This leads us to the following general recursion

$$\mathbf{A}^{(\ell)}_{G_L;G_R} = \alpha_{\ell+1}\mathbf{U}^{(\ell+1)}_{G_L;}\mathbf{V}^{(\ell+1)^T}_{G_R;} + \beta_{\ell+1}\mathbf{U}^{(\ell+2)}_{G_L;}\mathbf{V}^{(\ell+1)^T}_{G_R;} + \gamma_{\ell+1}\mathbf{U}^{(\ell+1)}_{G_L;}\mathbf{V}^{(\ell+2)^T}_{G_R;} + \mathbf{A}^{(\ell+1)}_{G_L;G_R} \tag{19}$$

The rank of the remainder term decreases by 1 in each step as ℓ increases. So this recursion stops when ℓ takes (at most) the value of $\min\{\mu,\nu\} = \min\{IJ, K\}$ where, in that specific case, remainder identically vanishes. Hence, this is a finite recursion. So we can write the following exact relation

$$\begin{aligned}\mathbf{A}_{G_L;G_R} &= \sum_{\ell=1}^{n_\alpha}\alpha_\ell \mathbf{U}_{G_L;}{}^{(\ell)}\mathbf{V}^{(\ell)}_{G_R;}{}^{T} + \sum_{\ell=1}^{n_\beta}\beta_\ell \mathbf{U}_{G_L;}{}^{(\ell+1)}\mathbf{V}^{(\ell)}_{G_R;}{}^{T} \\ &\quad + \sum_{\ell=1}^{n_\gamma}\gamma_\ell \mathbf{U}_{G_L;}{}^{(\ell)}\mathbf{V}^{(\ell+1)^T}_{G_R;} = \overline{\mathbf{U}}_{G_L;G_L}\boldsymbol{\Sigma}_{\mathbf{G_L};\mathbf{G_R}}\overline{\mathbf{V}}_{G_R;G_R}{}^{T}.\end{aligned} \tag{20}$$

where

$$\begin{aligned} &n_\alpha \equiv \min\{\mu,\nu\} \equiv \min\{IJ, K\}, \\ &n_\beta \equiv \begin{cases} n_\alpha & \mu < \nu \\ n_\alpha - 1 & \mu \geq \nu \end{cases} \qquad n_\gamma \equiv \begin{cases} n_\alpha & \mu > \nu \\ n_\alpha - 1 & \mu \leq \nu \end{cases} \end{aligned} \tag{21}$$

Here the entities of $\overline{\mathbf{U}}_{G_L;G_L}$ and $\overline{\mathbf{V}}_{G_R;G_R}$ are orthonormal folmats. Because the entities of these folmats are normalized support folvecs and orthogonal to each other we can write

$$\overline{\mathbf{U}}_{G_L;G_L} \equiv \left[\begin{array}{ccc}\mathbf{U}^{(1)}_{G_L;} & \ldots & \mathbf{U}^{(\mu)}_{G_L;}\end{array}\right], \qquad \overline{\mathbf{V}}_{G_R;G_R} \equiv \left[\begin{array}{ccc}\mathbf{V}^{(1)}_{G_R;} & \ldots & \mathbf{V}^{(\nu)}_{G_R;}\end{array}\right] \tag{22}$$

where the superscripts(μ) and (ν) respectively represents the sizes of the left and right grids and therefore equal to IJ and K respectively. We have used capital letters for denoting the left and right supports since the left one is folded vector hence has somehow a hidden matrix nature even though the right one is in fact a truly ordinary vector.

$\boldsymbol{\Sigma}_{\mu;\nu}$ consists of αs, βs and γs calculated at each step of recursion. These parameters represent the contribution of outer products. Under assumption of $\mu < \nu$ we can write

$$\boldsymbol{\Sigma}_{\mu;\nu} = \left[\begin{array}{ccccccccc} \alpha_1 & \gamma_1 & 0 & \cdots & \cdots & 0 & 0 & \cdots & 0 \\ \beta_1 & \alpha_2 & \gamma_2 & \cdots & \cdots & 0 & 0 & \cdots & 0 \\ 0 & \beta_2 & \alpha_3 & \ddots & \vdots & \vdots & \vdots & \cdots & \vdots \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots & \vdots & \cdots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \ddots & \gamma_{\mu-1} & 0 & \cdots & 0 \\ 0 & 0 & \cdots & \cdots & \beta_{\mu-1} & \alpha_\mu & \gamma_\mu & \cdots & 0 \end{array}\right] \tag{23}$$

By using α, β and γ parameters we can define quality measurers at each recursion step to measure how good the approximation is. Quality measurer definition is given as follows for the ℓth step of the recursion

$$\sigma_\ell = \sum_{i=1}^{\ell}\frac{\alpha_i^2 + \beta_i^2 + \gamma_i^2}{\|\mathbf{A}\|^2} \tag{24}$$

 ISBN: 978-84-617-8694-7

4 Numerical Implementations

In this section TFEMPR is applied on a digital image sequence and results given by quality measurer of the method which comes from method's nature and by PSNR which is commonly used in literature as a measurer on these issues. Illustrative numerical examples given previous works, you can see in [27].

4.1 Peak Signal to Noise Ratio (PSNR)

PSNR is a widely used parameter to measure how efficient the approximation to a signal sequence is. It is defined as follows

$$PSNR \equiv 10 \log_{10} \left(\frac{\max_{(G_L,G_R)} \left(\chi^2_{i,j;k} \right)}{MSE} \right), \quad MSE \equiv \frac{1}{\mu\nu} \left\| \mathbf{A}_{G_L;G_R} - \mathbf{A}^{TFEMPR}_{G_L;G_R} \right\|^2 \quad (25)$$

where μ and ν stand for the number of elements in G_L and G_R respectively while $\mathbf{A}_{G_L;G_R}$ and $\mathbf{A}^{TFEMPR}_{G_L;G_R}$ denote the target folmat and its TFEMPR approximant under consideration.

If we assume that $\mu < \nu$ then we can use (20) and its truncated form for N recursive step truncation and write the following equality

$$\mathbf{A}_{G_L;G_R} - \mathbf{A}^{TFEMPR}_{G_L;G_R} = \sum_{\ell=N+1}^{n_\alpha} \alpha_\ell \mathbf{U}^{(\ell)}_{G_L;} \mathbf{V}^{(\ell)\ T}_{G_R;} + \sum_{\ell=N+1}^{n_\beta} \beta_\ell \mathbf{U}^{(\ell+1)}_{G_L;} \mathbf{V}^{(\ell)\ T}_{G_R;} + \sum_{\ell=N+1}^{n_\gamma} \gamma_\ell \mathbf{U}^{(\ell)}_{G_L;} \mathbf{V}^{(\ell+1)T}_{G_R;} \quad (26)$$

which implies

$$\left\| \mathbf{A}_{G_L;G_R} - \mathbf{A}^{TFEMPR}_{G_L;G_R} \right\| = \sum_{\ell=N+1}^{n_\alpha} \alpha^2_\ell + \sum_{\ell=N+1}^{n_\beta} \beta^2_\ell + \sum_{\ell=N+1}^{n_\gamma} \gamma^2_\ell \quad (27)$$

and therefore

$$MSE = \frac{1}{mn} \left(\sum_{\ell=N+1}^{n_\alpha} \alpha^2_\ell + \sum_{\ell=N+1}^{n_\beta} \beta^2_\ell + \sum_{\ell=N+1}^{n_\gamma} \gamma^2_\ell \right). \quad (28)$$

where n_α, n_β, n_γ are defined as before. As we can see the number of the additive terms decreases as N grows such that MSE vanishes at the limit where none of the α, β, γ appears in the sums.

4.2 Real Data Implementations: Digital Video

In this part we will present application of TFEMPR on the real data obtained from digital videos. Grayscale video data will be taken into consideration as implementation in this part and the results will be provided numerically in table and visually in figures.

Grayscale video is basically taken into consideration as a three-way array. Two of these ways represent respectively the horizontal and vertical positions of each pixel in frames and the third way represents the location of the frame in the frame sequence. We can imagine video as a rectangular prism which is obtained by aligning the images consecutively. Despite the highly many components of digital video data, we consider only these three ways; height, width and the location of frame in these implementations. We will work on the grayscale video data first for simplicity.

 ISBN: 978-84-617-8694-7

The grayscale video data taken for TFEMPR implementation here is a $240 \times 320 \times 36$ type multiway array. We use the video file under the name "xylophone.mp4" which is a default image file in Matlab. Movement of a mallet during playing xylophone is tapen in this video. The results are given for the 18th frame, which produces the middle instantaneous display of video, to compare approximation and exact digital video in this paper. The resulting frames are illustrated such that the left frame represents approximation while the right frame characterizes the exact frame.

Figure 1 depicts the 18th approximation and exact frame of digital video for the 5 recursion steps under the use of Normalised Directional Average Support (NDAS). For 5 steps of TFEMPR, background is almost clear, but the mallet and hand are fuzzy due to their moving object characters.

Figure 2 presents the result for 10 recursion steps under the same conditions. Background is sharper, mallet and hand are still unsharp, but much sharper than Figure 1. The moving objects obtained for 20 recursion steps are visibly clear in Figure 3, and, as the last comparison in Figure 4, moving objects become clear and also background is sharp as in the exact frame. These results indicate that TFEMPR works pretty good on background reconstruction and video processing problems.

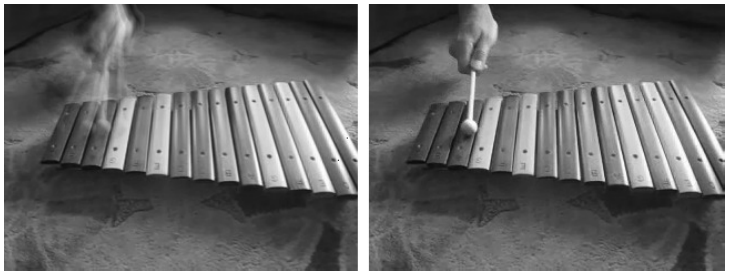


Figure 1: The result of 18th frame for 5 Recursion Step with NDAS

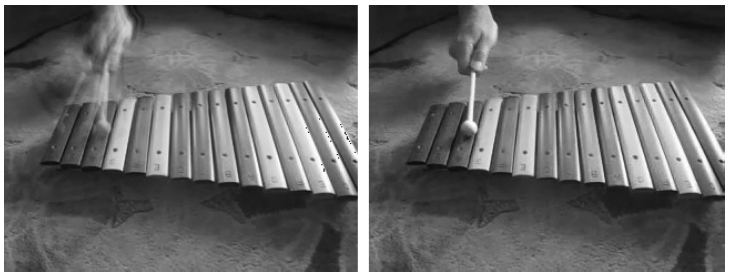


Figure 2: The result of 18th frame for 10 Recursion Step with NDAS

 ISBN: 978-84-617-8694-7

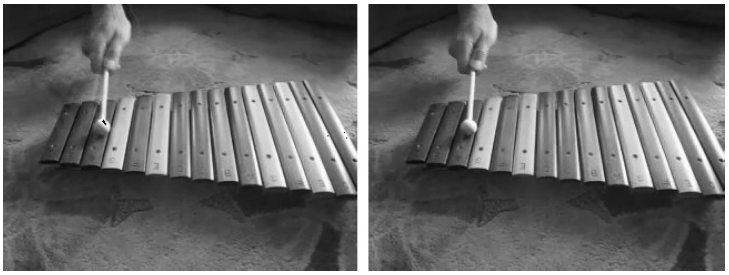


Figure 3: The result of 18th frame for 20 Recursion Step with NDAS

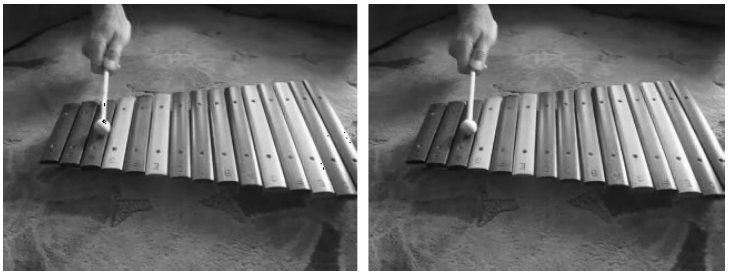


Figure 4: The result of 18th frame for 30 Recursion Step with NDAS

Quality measurers and PSNR results of TFEMPR method is tabulated in Table 1. It is clearly seen in Table 1 that approximation quality and PSNR get better as long as the number of recursion step increases.

Table 1: TFEMPR Quality Measurers and PSNR values for Grayscale Video Implementation

RecSN	5	10	20	30
σ	0,9949	0,9972	0,9990	0,9998
PSNR	30,2559	32,9384	37,5787	43,8265

TFEMPR is also applicable to multilinear arrays having more-than-three ways. Especially for digital video data we can increase dimension via some parameters. One of these is the color parameters which is well-known as RGB. RGB is abbreviation for the names of its main color components, Red, Green and Blue, each of which takes values from nonnegative integers in $\mathbb{Z}_{255}$ (the subset of integers composed of first natural numbers between 0 and 255 inclusive). All possible colors obtained from the mixtures of these colors and a color is represented by three values, for example red is $[255, 0, 0]$

We can increase the number of ways or dimensions by adding the audio, brightness, opaqueness components and so on. Audio data can be represented in 8 bit or 16 bit format and the data range is the interval $[-1, 1]$. Audio data is a sequence of digital signals in fact, as long as digital audio devices and relevant processors can be used. According to today's technology analog sound signals can also be processed as if they are digital data. However to this end a widely-used-in-computers method Pulse Code Modulation (PCM)

 ISBN: 978-84-617-8694-7

is utilized. Due to finite bit formats these parameters take also discrete and limited values too. So these may be the source of error accumulations for the TFEMPR at small level truncations. This can be cured perhaps by using modular arithmetic even though we have not attempted to do so yet.

5 Concluding Remarks

In this work we have focused on the decomposition of multiway arrays to outer products lower multivariance multiway arrays. To this end we have used the folmat and folvec concepts we have recently developed. The scheme has been designed in such a way that the additive components of decompositions are binary (two-factor) outer products. We have given illustrative and also confirmative implementations. We are planning to publish a more comprehensive report on these types digital image sequence decompositions pretty soon.

References

[1] L.R. Tucker, *The extension of factor analysis to three-dimensional matrices*, Contributions to Mathematical Psychology, (1964) 110–127.

[2] L. R. Tucker, *Some mathematical notes on three-mode factor analysis*, Psychometrika, **31**, (1966) 279–311.

[3] L. De Lathauwer, B. De Moor and J. Vandewalle, *A multilinear singular value decomposition*, SIAM J. Matrix Anal. Appl., **21**(4), (2000) 1253–1278.

[4] L. De Lathauwer, B. De Moor and J. Vandewalle, *Independent component analysis and (simultaneous) third-order tensor diagonalization*, IEEE Transactions on Signal Processing, **49**, (2001) 2262–2271.

[5] L. De Lathauwer, *Decompositions of a higher-order tensor in block terms Part I: Lemmas for partitioned matrices*,SIAM Journal on Matrix Analysis and Applications (SIMAX), **30**(3), (2008) 1022–1032.

[6] L. De Lathauwer, *Decompositions of a higher-order tensor in block terms Part II: Definitions and uniqueness*, SIAM Journal of Matrix Analysis and Applications, **30**(3), (2008)1033–1043.

[7] S. E. Leurgans, R. T. Ross and R. B. Abel, *A decomposition for three-way arrays*, SIAM J. Matrix Anal. Appl., **14**, (1993) 1064–1083.

[8] T. G. Kolda and B. W. Bader, *Tensor decompositions and applications*, SIAM Review, **51** (3), (2008) 455–500.

[9] T. G. Kolda, *Orthogonal tensor decompositions*, SIAM Journal on Matrix Analysis and Applications, **23**, (2001) 243–255.

[10] A. Cichocki, R. Zdunek, A. H. Phan and S. Amari, *Nonnegative matrix and tensor factorizations-Applications to exploratory multiway data Analysis and blind source seperation*, Wiley and Sons Publication, (2009).

[11] I. M. Sobol, *Sensitivity estimates for nonlinear mathematical models*, Mathematical Modelling and Computational Experiments,**1**, (1993) 407–414.

 ISBN: 978-84-617-8694-7

[12] H. Rabitz and O. F. Alis, *General foundations of high dimensional model representations*, J. Math. Chem., **25**, (1999) 197–233.

[13] O. F. Alis and H. Rabitz, *Efficient implementation of high dimensional model representations*, J. Math. Chem.,**29**, (2001) 127–142.

[14] G. Li, C. Rosenthal and H. Rabitz, *High dimensional model representations*, J. Phys. Chem. A, **105**, (2001) 7765–7777.

[15] M. Demiralp, *High dimensional model representation and its application varieties*, Mathematical Research, **9**, (2003) 146–159.

[16] M. Demiralp, *New generation HDMR based multiway array decomposers: enhanced multivariance products representation(EMPR)*, Proceedings for 1st IEEEAM Conference on Applied Computer Science(ACS), **16**, (2010).

[17] M. Demiralp, *High dimensional model representation (HDMR) and enhanced multivariate product representation(EMPR) as small scale multivariate decomposition methods*, Proceedings of the 4th WSEAS International Conference on Finite Differences-Finite Elements-Finite Volumes-Boundary Elements(ECC'11 and F-and-B'11), (2011) 12–13.

[18] B. Tunga and M. Demiralp, *The influence of the support functions on the quality of enhanced multivariate product representation*, Journal of Mathematical Chemistry, **48**, (2010) 827–840.

[19] E. K. Özay and M. Demiralp, *A new multiway array decomposition via enhanced multivariance product representation*, Numerical Analysis and Applied Mathematics ICNAAM, **1479**(1), (2012) 2015–2018.

[20] Z. Gündoğar and M. Demiralp, *Enhanced multivariance products respresentation under an integral operator weight with a product type kernel of univariate factors*, 1st International Conference on Optimization Techniques in Engineering (OTENG'13), Antalya, Turkey, (2013) 23–28.

[21] E. Demiralp and M. Demiralp, *Tridiagonal matrix enhanced multivariance products representation(TMEMPR) for matrix decomposition*, Proceedings of 14th International Conference Computational and Mathematical Methods in Science and Engineering, **2**, (2014) 446–455.

[22] E. Demiralp, *Weighted tridiagonal matrix enhanced multivariance products representation of finite interval data (TMEMPR)*, Proceedings of 14th International Conference Computational and Mathematical Methods in Science and Engineering (CMMSE'14), **2**, (2014) 441–445.

[23] E. K. Özay and M. Demiralp, *Tridiagonal matrix enhanced multivariance products representation (TMEMPR) studies: Decomposing the planarly unfolded three-way arrays*, Proceedings of 14th International Conference Computational and Mathematical Methods in Science and Engineering, **3**, (2014) 785–793.

[24] M. Demiralp and E. Demiralp, *An orthonormal decomposition method for multidimensional matrices*, AIP Proceedings for the International Conference of Numerical Analysis and Applied Mathematics (ICNAAM 2009), 18-22 September 2009, Rethymno, Crete, Greece, **1168**, (2009) 428–431.

 ISBN: 978-84-617-8694-7

[25] M. Demiralp and E. Demiralp, *A contemporary linear representation theory for ordinary differential equations: multilinear algebra in folded arrays (folarrs) perspective and its use in multidimensional case*, Journal of Mathematical Chemistry, **51**(1), (2013) 38–57.

[26] Z. Gündoğar and M. Demiralp, *Formulation of tridiagonal folmat enhanced multivariance products representation (TFEMPR)*, AIP Conference Proceedings, **1702**, (2015).

[27] Z. Gündoğar and M. Demiralp, *Certain illustrative numerical implementations of tridiagonal folmat enhanced multivariance products representation (TFEMPR) for 3-Way Array*, International Journal of Signal Processing, **1**, (2016) 108-113.

[28] Z. Gündoğar and M. Demiralp, *Tridiagonal folmat enhanced multivariance products representation (TFEMPR) under subspace supported rational transformations (SsSRT)*, AIP Conference Proceedings, **1798**, 020064 (2017).

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

Function Approximation via Contour Integration and Tridiagonal Kernel Enhanced Multivariance Products Representation (TKEMPR)

Ercan Gürvit[1] and N.A. Baykara[2]

[1] *Faculty of Sciences and Letters, Department of Mathematics, Marmara University*

[2] *"Interinstitutional Group for Science and Methods of Computing" located at Informatics Institute, İstanbul Technical University*

emails: ercangurvit@gmail.com, nabaykara@gmail.com

Abstract

This work takes into consideration a single variable function to be approximated. Then making a series of transformations it is represented as a contour integral where the kernel fuction is expressed in terms of two variables. This new representation allows us to apply Tridiagonal Kernel Enhanced Multivariance Products Representation (TKEMPR) method onto it and provides us with a good approximation of the kernel function which then is integrated to obtain an approximation for the initial function in question. Because of the overall conceptual approach to the problem through this article, applications which are done are intended to be given during the oral representation.

Key words: Univariate Approximation, multivariate approximation, EMPR, TKEMPR, Contour integrals
MSC 2000: 00A69, 15A12, 65D15

1 Introduction

The intended work uses the TKEMPR method which decomposes a linear integral operator on univariate functions by using high dimensional modelling with the basic idea to use Enhanced Multivariance Products Representation (EMPR) technique created and conjectured by Demiralp. The representation used here is not based on the general EMPR and is a specific EMPR construction for bivariate function decomposition. We call this decomposition Tridiagonal Kernel Enhanced Multivariance Products Representation (TKEMPR). It

 ISBN: 978-84-617-8694-7

uses EMPR [1-18] bivariate function decomposition consecutively such that in each step the remainder term is expanded to again a bivariate EMPR but with different support functions. To make a function approximation using TKEMPR, first we represent our function in terms of contour integration, then following a procedure which allows us to obtain the kernel to be used in TKEMPR [19-24], we obtain a two variable form and finally use it to approximate our function.

2 Basics of EMPR

Enhanced Multivariance Products Representation (EMPR) is a decomposition method recently developed by M. Demiralp. For a given multivariate function $f(x_1, ..., x_N)$ it can be written as

$$\begin{aligned} f(x_1, ..., x_N) \;=\; & f_0 \prod_{i=1}^{N} s_i(x_i) + \sum_{j=1}^{N} f_j(x_j) \prod_{i=1, i\neq j}^{N} s_i(x_i) \\ & + \sum_{\substack{j_1, j_2 = 1 \\ j_1 < j_2}}^{N} f_{j_1, j_2}(x_{j_1}, x_{j_2}) \prod_{i=1, i\neq j_1, j_2}^{N} s_i(x_i) \\ & + \sum_{\substack{j_1, j_2, j_3 = 1 \\ j_1 < j_2 < j_3}}^{N} f_{j_1, j_2, j_3}(x_{j_1}, x_{j_2}, x_{j_3}) \prod_{i=1, i\neq j_1, j_2, j_3}^{N} s_i(x_i) \\ & + \cdots \end{aligned} \tag{1}$$

where f_js stand for the EMPR components ordered in ascending multivariance. The univariate functions denoted by s are called support functions and finally

$$x_i \in [\, a_i, b_i \,], \qquad i = 1, ..., N \tag{2}$$

where the interval may or may not be a finite one. There are 2^N unknown components given in a single equation. Hence some constraints apply to uniquely determine EMPR components. These constraints are constructed via vanishing integrals over the EMPR components except f_0. To define these integrals univariate weight functions the product of which defines a single multivariate weight function, are used. These univariate weight function integrals over the relevant intervals given above are set equal to 1 to facilitate further analysis; these constraints are not necessary but provide the averaging property to the weight function.

$$\int_{a_i}^{b_i} dx_i W_i(x_i) = 1, \qquad i = 1, ..., N \tag{3}$$

 ISBN: 978-84-617-8694-7

$$W(x_1, ..., x_N) \equiv \prod_{i=1}^{N} W_i(x_i). \tag{4}$$

nd the vanishing conditions can be written as follows

$$\int_{a_i}^{b_i} dx_i W_i(x_i) f_{j_1,...,j_k}(x_{j_1}, ..., x_{j_k}) = 0,$$
$$i \in \{j_1, ..., j_k\}, \qquad k = 1, ..., N \tag{5}$$

By using them after integrating both sides of the EMPR definitions over certain chosen x independent variables over their intervals and under the product of the corresponding univariate weight factors produces the unique values of the EMPR components.

EMPR is an extended form of the High Dimensional Model Representation (HDMR) first proposed by Sobol with all unit intervals $[0, 1]$ and unit constant weight functions whose values are all one throughout the relevant intervals. This form can be called Sobol's plain HDMR. A lot of new extensions have been developed in Demiralp's group studies. Amongst them, the most important one is the extension to EMPR which becomes HDMR when all univariate support functions are taken equal to just unit constant function over the relevant interval.

3 EMPR for Bivariate Functions

When matrices have been taken as target entities a new decomposition method called "Tridiagonal Matrix Enhanced Multivariance Products Representation ", has been developed. The elements of a matrix being located via two indices which take positive integer values, independently from the other index values, this allows us to construct, so-called "Discrete Bivariate EMPR, over the matrices. Following this step we have to search for a method to apply this construction to corresponding bivariate functions. To this end we can write the following equality for a given bivariate function $f(x, y)$

$$f(x, y) = f_0 u(x) v(y) + f_1(x) v(y) + f_2(y) u(x) + f_{1,2}(x, y) \tag{6}$$

where u and v are the support functions. In this construction we will use the unit interval $[0, 1]$ for both independent variables, x and y without any loss of generality Beyond that the weight factors are defined as follows

$$W_1(x) \equiv 1, \quad W_2(y) \equiv 1, \qquad x, y \in [0, 1]. \tag{7}$$

The support functions are assumed to have unit norms over the above intervals and under the above weights

$$\int_0^1 dx u(x)^2 = 1, \qquad \int_0^1 dy v(y)^2 = 1 \tag{8}$$

 ISBN: 978-84-617-8694-7

For this case the vanishing conditions should be taken as constraints imposed on $f_1(x)$, $f_2(y)$ and, $f_{1,2}(x,y)$. They can be written as follows

$$\int_0^1 dx f_1(x)u(x) = 0 \tag{9}$$

$$\int_0^1 dy f_2(y)v(y) = 0 \tag{10}$$

$$\int_0^1 dx f_{1,2}(x,y)u(x) = 0 \tag{11}$$

$$\int_0^1 dy f_{1,2}(x,y)v(y) = 0 \tag{12}$$

These equations allow us to evaluate the EMPR terms uniquely as

$$f_0 = \int_0^1 dx \int_0^1 dy f(x,y)\, u(x)\, v(y) \tag{13}$$

$$f_1(x) = \int_0^1 dy f(x,y)\, v(y) - f_0 u(x) \tag{14}$$

$$f_2(y) = \int_0^1 dx f(x,y)\, u(x) - f_0 v(y) \tag{15}$$

$$\begin{aligned} f_{1,2}(x,y) &= f(x,y) - f_0 u(x)\, v(y) \\ &\quad -f_1(x)\, v(y) - f_2(y)\, u(x). \end{aligned} \tag{16}$$

These four equations can be used to obtain more concise expressions for EMPR terms and EMPR's itself. To this end we can define the following integral operators through their actions on a given arbitrary function g which may be taken depending on x and y as we need

$$\begin{aligned} \widehat{P}g(x) &\equiv u(x) \int_0^1 d\xi u(\xi)\, g(\xi) \\ \widehat{Q}g(y) &\equiv v(y) \int_0^1 d\eta v(\eta)\, g(\eta) \end{aligned} \tag{17}$$

 ISBN: 978-84-617-8694-7

This enables us to write the following equalities

$$f_0 u(x)v(y) = \widehat{P}\widehat{Q}f(x,y) \tag{18a}$$

$$f_1(x)v(y) = \left(\widehat{I}_x - \widehat{P}\right)\widehat{Q}f(x,y) \tag{18b}$$

$$f_2(y)u(x) = \widehat{P}\left(\widehat{I}_y - \widehat{Q}\right)f(x,y) \tag{18c}$$

$$f_{1,2}(x,y) = \left(\widehat{I}_x - \widehat{P}\right)\left(\widehat{I}_y - \widehat{Q}\right)f(x,y) \tag{18d}$$

and finally

$$\begin{aligned} f(x,y) &= \widehat{P}\widehat{Q}f(x,y) + \left(\widehat{I}_x - \widehat{P}\right)\widehat{Q}f(x,y) + \widehat{P}\left(\widehat{I}_y - \widehat{Q}\right)f(x,y) \\ &\quad + \left(\widehat{I}_x - \widehat{P}\right)\left(\widehat{I}_y - \widehat{Q}\right)f(x,y) \end{aligned} \tag{19}$$

where $\widehat{I}_x$ and $\widehat{I}_y$ stand for the unit operators in the spaces of univariate functions depending on x and y respectively.

From (19) the EMPR is an orthogonal decomposition of the target function into two components laying in the space spanned by $u(x)$ and in its complementary space and another two components laying in the space spanned by $v(y)$ and in its complementary space. Therefore "EMPR is a projective decomposition projecting onto the spaces spanned by support functions and their complementary spaces.

Finally to reach (TKEMPR) construct which consists of using EMPR bivariate function decomposition consecutively such that in each step the remainder term is expanded to again a bivariate EMPR but with different support functions.

4 Contour Integration

A function $f(x)$ can be represented in terms of a contour integration over an interval $[0,1]$ which may in fact easily be converted to any finite interval $[a,b]$ without loss of generality

$$f(x) = \frac{1}{2\pi i}\oint_C d\xi \frac{f(\xi)}{\xi - x}, \quad x \in [0,1] \tag{20}$$

$$C \to \xi = x + e^{i\theta}$$
$$r > 0, \quad \theta \in [0, 2\pi)$$

 ISBN: 978-84-617-8694-7

The corresponding integration interval $[0, 2\pi]$ can be divided into two halves.

$$\begin{aligned} f(x) &= \frac{1}{2\pi}\int_0^{2\pi} d\theta f\left(x + re^{i\theta}\right) \\ &= \frac{1}{2\pi}\int_0^{\pi} d\theta f\left(x + re^{i\theta}\right) + \frac{1}{2\pi}\int_{\pi}^{2\pi} d\theta f\left(x + re^{i\theta}\right) \end{aligned} \tag{21}$$

Via a simple change of variable, the second half can be written in terms of the first one and the overall expression is expressed in a more compact manner.

$$f(x) = \int_{\pi}^{2\pi} d\theta f\left(x + re^{i\theta}\right) \overset{2\pi-\theta}{\rightarrow} \int_0^{\pi} d\theta f\left(x + re^{-i\theta}\right) \tag{22}$$

$$f(x) = \frac{1}{\pi}\int_0^{\pi} d\theta \frac{1}{2}\left[f\left(x + re^{i\theta}\right) + f\left(x + re^{-i\theta}\right)\right] \tag{23}$$

The function with two different arguments can be written as below, by representing their real and imaginary components separately.

$$\begin{aligned} f\left(x + re^{i\theta}\right) &= f_g + if_s \\ f\left(x + re^{-i\theta}\right) &= f_g - if_s \end{aligned} \tag{24}$$

Thus f_g can be written as

$$f_g = \frac{1}{2}\left[f\left(x + re^{i\theta}\right) + f\left(x + re^{-i\theta}\right)\right] \tag{25}$$

By making a change of variable $\theta \rightarrow \pi\theta$ we obtain

$$f(x) = \int_0^{1} d\theta \frac{1}{2}\left[f\left(x + re^{i\pi\theta}\right) + f\left(x + re^{-i\pi\theta}\right)\right] \tag{26}$$

Now, let us make a definition which corresponds to Heaviside step function

$$1_f \equiv \begin{cases} 1 & 0 \leq \theta \leq 1 \\ 0 & other \end{cases} \tag{27}$$

Correspondingly we obtain the expression below.

$$\mathcal{I}1_f = \int_0^{1} d\theta K(\theta, x) 1_f \tag{28}$$

The final form of the kernel to be used in TKEMPR can be written as follows:

$$K(\theta, x) = \frac{1}{2}\left[f\left(x + re^{i\theta}\right) + f\left(x + re^{-i\theta}\right)\right] \tag{29}$$

 ISBN: 978-84-617-8694-7

5 Method

Now that we have constructed a bivariate form for a function $f(x)$ needed to be used it in (TKEMPR), we have to adapt it to the proper form mentioned in bivariate (EMPR)

$$K(\theta, x) = K_0 u(\theta) v(x) + K_1(\theta) v(x) + K_2(x) u(\theta) + K_{1,2}(\theta, x) \tag{30}$$

where u and v are support functions which will be chosen as 1 for the initial step. θ and x are taken on the unit interval as mentioned before. The same will be valid for the weight functions also. Now when we proceed for the (EMPR) we will obtain

$$K_{1,2}(\theta, x) = K(\theta, x) - K_0 u(\theta) v(x) - K_1(\theta) v(x) - K_2(x) u(\theta). \tag{31}$$

At this stage, this is again a bivariate function which needs our attention in terms of (EMPR). This process can be repeated as many times as we need and each time we add a new step a better approximation will be obtained.

A final step after making sufficient number of iterations to approximate the bivariate function $K_{1,2}(\theta, x)$ is to integrate this latter over the interval $[0, 1]$ to obtain back the intended approximation of our original function $f(x)$

$$f(x) = \int_0^1 d\theta K(\theta, x) \tag{32}$$

6 Conclusion

The decomposition obtained by applying (TKEMPR) to a univariate function after having obtained a bivariate counterpart of it via contour integration provides us with a good approximation to the bivariate kernel. Finally if we integrate this kernel over the unit interval we obtain an approximation to the original function. An intended future work consists of applying the same method to the remainder term of a Taylor expansion of a function, expressed in integral form and obtain a better approximation enforced by the Taylor expansion itself, after necessary transformations

Acknowledgements

Both authors are grateful to Prof. Metin Demiralp who had an invaluable scientific support in this present work

References

[1] Ö. F Alış and H. Rabitz: *General Foundations of High Dimensional Model Representation*, Journal of Mathematical Chemistry, 25, pp.197-233, **(1999)**.

[2] M. Demiralp : *High Dimensional Model Representation and Its Application Varieties*, The Fourth International Conference on Tools for Mathematical Modelling, St. Petersburg, Russia, June 23-28, **(2003)**.

[3] M. Demiralp and E. Demiralp : *An Orthonormal Decomposition Method for Multidimensional Matrices* in AIP Proceedings for the International Conference of Numerical Analysis and Applied Mathematics (ICNAAM 2009), vol. 1168, Rethymno, Crete, Greece, 18-22 September 2009, pp. 424427, doi:http://dx.doi.org/10.1063/1.3241487, **(2009)**.

[4] M. Demiralp and E. Demiralp : *Dimensionality Reduction and Approximation via Space Extension and Multilinear Array Decomposition* in AIP Proceedings for the International Conference of Computational Methods in Science and Engineering (ICCMSE 2009), Mini Symposium on Recent Developments in Numerical Schemes for Hilbert Space Related Issues in Science and Engineering, Rhodes, Greece, 29 September-4 October 2009, pp. 837 840, doi:http://dx.doi.org/10.1063/1.4771824, **(2009)**.

[5] M. Demiralp and E. Demiralp : *A New Straightforward Decomposition Method without Iteration to Approximate Matrices via Dominant Basis Matrices* in The Interna- tional Conference on Scientific Computing - WorldComp09 (CSC09), Las Vegas, Nevada, USA, 13-16 July 2009, pp. 7983, **(2009)**.

[6] B. Tunga and M. Demiralp : *An Iterative Scheme for Enhanced Multivariance Product Representation Method* in Proceedings for the 1st IEEEAM Conference on Applied Computer Science (ACS), Malta, pp. 247255, **(2010)**.

[7] M. Demiralp : *New Generation HDMR Based Multiway Array Decomposers: Enhanced Multivariance Products Representation (EMPR)* in Proceedings for the 1st IEEEAM Conference on Applied Computer Science (ACS), ser. ISBN: 978-960-474-225-7, Malta, pp. 1616, keynote Speech, **(2010)**.

[8] E. Demiralp and M. Demiralp : *Reductive Multilinear Array Decomposition Based Support Functions in Enhanced Multivariance Products Representation (EMPR)* in Proceedings for the 1st IEEEAM Conference on Applied Computer Science (ACS), Malta, pp. 448454, **(2010)**.

[9] C. Gözükırmızı and M. Demiralp : *Numerical Studies on the Use of Enhanced Multivariance Product Representation as a Multiway Array Decomposer* in AIP Proceedings for the International Conference of Numerical Analysis and Applied Mathematics (ICNAAM 2010), Symposium 112, Recent Developments in Hilbert Space Tools

and Methodology for Scientific Computing, vol. 1281, Rhodes, Greece, pp. 19221925, doi:http://dx.doi.org/10.1063/1.3498300, **(2010)**.

[10] L. Divanyan and M. Demiralp : *Weighted Reductive Multilinear Array Decomposition* in AIP Proceedings for the 9th International Conference on Numerical Analysis and Applied Mathematics (ICNAAM2011), vol. 1389, Halkidiki, Greece, pp. 11561159, doi:http://dx.doi.org/10.1063/1.3637820, **(2011)**.

[11] S. Tuna, N. A. Baykara, and M. Demiralp : *Weighted Singular Value Decomposition for Folded Matrices* in Proceedings of the 2nd Inter- national Conference on Applied Informatics and Computing Theory (AICT11), IEEEAM, ser. ISBN: 978-1-61804-034-3, N. Mastorakis, M. Demiralp, and N. A. Baykara, Eds., Prague, Czech Republic, pp. 7075, **(2011)**.

[12] M. Demiralp : *Decomposing Functions, Arrays, Function Arrays* in Lecture Talk based on the symposium 48s preface, AIP Proceedings for the 9th International Conference on Numerical Analysis and Applied Math- ematics (ICNAAM2011), vol. 1389, Halkidiki, Greece, pp. 11381138, doi:http://dx.doi.org/10.1063/1.3637815, **(2011)**.

[13] M. Ayvaz and M. Demiralp : *Towards a New Multiway Array Decomposition Algorithm: Elementwise Multiway Array High Dimensional Model Representation (EMAHDMR)* in Proceedings of the 2nd International Conference on Applied Informatics and Computing Theory (AICT11), IEEEAM, ser. ISBN: 978-1-61804-034-3, N. Mastorakis, M. Demiralp, and N. A. Baykara, Eds., Prague, Czech Republic, pp. 7681, **(2011)**.

[14] E. K. Özay : *A New Multi-way Array Decomposition via Enhanced Multivariance Product Representation*, AIP Proceedings for the 10th International Conference of Numerical Analysis and Applied Mathematics (ICNAAM), Kos, Greece, pp. 2015-2018,Volume 1479, **(2012)**

[15] E. K. Özay and M. Demiralp : *A New Multi-way Array Decomposition*, 2012 SIAM Conference on Applied Linear Algebra, Valencia, Spain, pp. 78–78, **(2012)**

[16] E. K. Özay and M. Demiralp : *Tridiagonal Matrix Enhanced Multivariance Products Representation (TMEMPR) Studies: Decomposing the Planarly Unfolded Three-way Arrays*, Proceedings of the 14th International Conference on Computational and Mathematical Methods in Science and Engineering (CMMSE), Cadiz, Spain, **(2014)**.

[17] B. Tunga and M. Demiralp : *The Influence of the Support Functions on the Quality of Enhanced Multivariance Product Representation* Journal of Mathematical Chemistry, Volume 48, Issue 3, pp 827-840, 2010.

[18] A. Okan, N.A. Baykara and M. Demiralp : *Weight Optimization in Enhanced Multivariance Product Representation (EMPR) Method* Int. Conf. on Numer. Anal. and Appl. Math., AIP Conference Proceedings, Volume 1281, pp. 1935-1938,Rhodes, Greece, 2010.

[19] A. Okan and M. Demiralp, *Tridiagonal Kernel Enhanced Multivariance Products Representation (TKEMPR) for Univariate Integral Operator Kernels*,The 2014 International Conference Mathematics and Computers in Sciences and Industry (MCSI 2014), 2014 International Conference,13-15 Sept., Varna, Bulgaria, pp. 195-200, doi: 10.1109/MCSI.2014.26, print ISBN: 978-1-4799-4744-7

[20] A. Okan and M. Demiralp, *Tridiagonal Kernel Enhanced Multivariance Products Representation (TKEMPR) for Outer Product Sums: Arrowheading EMPR for Kernel (AEMPRK)*, The 12th International Conference of Numerical Analysis and Applied Mathematics (ICNAAM 2014), 22-28 September 2014, Rhodes, Greece.

[21] A. Okan and M. Demiralp, *Arrowheading Enhanced Multivariance Products Representation for a Kernel (AEMPRK) in a Taylor Series Expansion*, 11th International Conference of Computational Methods in Sciences and Engineering, ICCMSE 2015, 20-23 March 2015, Athens, Greece, doi: 10.1063/1.4912452

[22] A. Okan and M. Demiralp, *Numerical Implementations for Tridiagonal Kernel Enhanced Multivariance Products Representation (TKEMPR) Method: Bivariate Case*, in The Proceedings of International Journal of Signal Processing, ISSN: 2367-8984, pages 102 - 107, Volume 1, 2016

[23] M. Demiralp and E. Demiralp : *A New Straightforward Decomposition Method without Iteration to Approximate Matrices via Dominant Basis Matrices* in The International Conference on Scientific Computing - WorldComp09 (CSC09), Las Vegas, Nevada, USA, 13-16 July 2009, pp. 7983, 2009.

[24] S. Tuna, and M. Demiralp : *Zero Interval Limit Perturbation Expansion for the Spectral Entities of Hilbert-Schmidt Operators Combined with Most Dominant Spectral Component Extraction: Convergence and Confirmative Implementations*, Journal of Mathematical Chemistry: 1-23., doi:10.1007/s10910-017-0740-1, 2017.

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

Time Valuation in Cancer Optimal Therapies: A Study of Chronic Myeloid Leukemia

P.J. Gutiérrez Diez[1], M.A. López-Marcos[2] and J. Martínez-Rodríguez[3]

[1] *Department of Economic Theory and IMUVA, University of Valladolid, Spain*

[2] *Department of Applied Mathematics and IMUVA, University of Valladolid, Spain*

[3] *Department of Applied Economics and IMUVA, University of Valladolid, Spain*

emails: pedrojos@fae.uva.es, malm@mac.uva.es, julia@eco.uva.es

Abstract

This paper analyzes how optimal therapies are affected when the time evolution of cancer is envisaged as an additional element determining malignancy. We introduce a time valuation factor that measures the increase of malignancy associated to the quick development of the disease and the persistent negative effects of initial drug doses. Taking as reference a mathematical model of Chronic Myeloid Leukemia (CML), we solve and simulate the model with and without the new time valuation factor. We conclude that the consideration of this factor allows more efficient optimal therapies to be designed.

Key words: Optimal Control Problem; Objective Function; Time Valuation Factor; Chronic Myeloid Leukemia; Imatinib Therapy.
MSC 2000: AMS 49N90;90C90.

1 Introduction

The design of optimal therapies to fight cancer is an important research field in today's Biomathematics (see [5] and [7]). Basically, an optimal therapy problem is a control problem consisting of: first, a set of difference/differential equations describing the biological dynamics of the disease under the specified treatment; and, second, an objective function measuring the malignancy of the treated cancer. By solving this optimal control problem it is possible to find the optimal therapy, i.e, the drug doses that minimize the malignancy of the treated disease. Until now, the mathematical formulation adopted to measure the

 ISBN: 978-84-617-8694-7

malignancy of cancer exclusively considers as malignancy elements the tumor size (given by the number of cancer cells) and the administered drug concentration measured in a given instant. From this assumption, a given number of cancer cells and a given drug dosage always cause the same malignancy, independently of the moment of time in which these magnitudes have been observed/administered. However, from the biomedical perspective, it is a well established fact that the faster the cancer grows the worse the cancer is. To take into account this malignancy factor, ignored by the literature, would entail assigning higher malignancy to cancer cells at the beginning of the cancer and lower malignancy to those appeared afterwards, that is, a time valuation factor weighting cancer size. On the other hand, it is also widely accepted by biomedical researchers and practitioners that drugs are not totally eliminated by patients and remain in their bodies, and that, additionally, the apparition of drug resistance is a direct consequence of early and persistent treatments. Subsequently, a given drug dose involves more negative effects if it is administered at the beginning of treatment than at the end, and then, to evaluate malignancy associated to drug doses, it would also be necessary to consider a time valuation factor assigning higher malignancy to the initial drug doses. We will study these questions for a standard optimal therapy problem. To do so, after formulating a discrete version of the model of Chronic Myeloid Leukemia (CML) proposed in [1], we introduce a time valuation factor that allows the malignancy elements associated to time and commented on above to be considered. We analyze its consequences with the purpose of clarifying whether or not this time dependent malignancy factor modifies the optimal therapies and the population of normal and cancer cells, and quantifying these modifications.

2 A Mathematical Model of Treated CML

In [1] a continuous time model was introduced to analyze the global dynamics of CML. This model incorporates the existence of nonlinear effects of imatinib treatment over a fixed period of time. Here, in order to gain consistency in the analysis, comparisons with real data and simulations, we propose a discrete time version of that model incorporating a daily schedule in the dynamics, since in the empirical literature on CML and in clinical practice the parameters and variables involved in the model are measured in per day values. We start by considering two different populations: that of hematopoietic stem cells (HSC), and that of differentiated cells (DC). In addition, each of these populations is divided into normal cells and cancer cells. Then, at time instant t, it is possible to distinguish between four different populations: Normal HSC, denoted by $x_0(t)$; cancer HSC, denoted by $y_0(t)$; normal DC, represented by $x_1(t)$; and cancer DC, denoted by $y_1(t)$. The evolution over time of CML is described by a system of difference equations which incorporates the most relevant biomedical facts. First, the populations of all the considered types of cells naturally decrease at fairly constant rates. Upon this fact, let d_0, g_0, d and g be, respectively, the per

 ISBN: 978-84-617-8694-7

day decrease rates of normal HSC, cancer HSC, normal DC, and cancer DC. In addition, since DC are produced not only by proliferation of DC but also by HSC, it is necessary to distinguish between these two mechanisms of increase in the number of DC for both normal and cancer cells. In particular, let d_2 and g_2 be the per day rates at which normal and cancer DC proliferate and originate, respectively, normal and cancer DC; and let r and q denote the rates at which normal and cancer HSC produce normal and cancer DC, in this order. Finally, through the self-renewal process, normal and cancer HSC produce similar cells by division. Then, we let normal and cancer HSC divide at rates n and m per day, respectively. In this self-renewing activity, there underlies a homeostatic process that controls the proliferation of HSC. In this respect, the division of normal HSC x_0 is directed by homeostasis, depending on the total level of HSC $(x_0 + y_0)$, and given by

$$\Phi(x_0 + y_0) = 1 - \frac{x_0 + y_0}{K},$$

where K represents the carrying capacity of bone marrow. In the same way, homeostasis for cancer cells y_0 is governed by the function

$$\Psi(x_0 + \alpha y_0) = 1 - \frac{x_0 + \alpha y_0}{K},$$

where $\alpha \in (0, 1]$ measures the fall in the homeostatic efficiency due to the disease (see [5], chapter 9, for an analysis of this fall).

Drug treatment is described by a positive time-dependent function $u(t)$. This function $u(t)$ captures the drug dose, and given that there is a dosage limitation due to the drug's toxicity, $u(t)$ is supposed to be bounded in $[0, u_{max}]$ for all t in $[0, T]$, where T denotes the treatment duration, measured in days. The effects of imatinib treatment are introduced through nonlinear functions, which affect the lifetime of cancer cells, and imply their maximum effect only for an intolerable dosage. In [1], different scenarios were studied depending on the distinct effects of imatinib on the dynamics. That work concluded that the disease completely remits only when imatinib causes an additional mortality of cancer HSC, represented by the function $h(u)$, which is a nonlinear increasing function satisfying $h(0) = 0$ (this means that cancer HSC decline at rate g_0 without treatment), and h attains the maximum value 1 for an intolerable dosage. Since this is the case with highest plausibility from the biomedical point of view, we limit our study to this situation. The behavior under treatment of CML is therefore described by the following system of difference equations:

$$\left.\begin{array}{l} x_0(t+1) = x_0(t) + n\Phi(x_0(t) + y_0(t))x_0(t) - d_0 x_0(t) \\ x_1(t+1) = x_1(t) + r x_0(t) - (d - d_2)x_1(t) \\ y_0(t+1) = y_0(t) + m\Psi(x_0(t) + \alpha y_0(t))y_0 - g_0 y_0(t) - \beta h(u(t))y_0(t) \\ y_1(t+1) = y_1(t) + q y_0(t) - (g - g_2)y_1(t) \\ t = 0, 1, \ldots, T \end{array}\right\},$$

where $\beta \in [0, 1]$ is a parameters measuring the proportion for additional decline in cancer HSC.

 ISBN: 978-84-617-8694-7

3 The Optimal Therapy Problem

As in any optimal therapy problem, the objective is to find the therapy $u^*(t)$ minimizing the malignancy of the disease under treatment. Concerning the objective function measuring malignancy, [1] considers quadratic terms representing the nonlinear costs of the treatment and the malignancy of the cancer HSC and DC levels; however, it does not entail any valuation of time. Here we formulate the objective function

$$N(u) = \sum_{t=0}^{T} \rho^t [u^2(t) + y_0^2(t) + y_1^2(t)] + \rho^{T+1}[y_0^2(T+1) + y_1^2(T+1)],$$

where $\rho \in (0, 1]$ is a parameter measuring the increase of malignancy of early cancer development and drug administration. More specifically, we consider an objective function which incorporates a final addend measuring the malignancy at the end of the treatment, and that implies lower malignancy as time passes. In this respect, and given that a higher malignancy is associated with both early cancer growth and early drug administration, we introduce a time valuation malignancy factor, decreasing over time, that affects the objective function in the optimal therapy problem as a whole. This is consistent with the observed evidence for treated CML, obviously the only existing situation of the disease for which data exist. Indeed, most empirical survival analyses show decreasing rates of mortality as treated cancer persists, i.e., higher mortality rates at the beginning of the disease than in subsequent dates. This decreasing rate of mortality over time would also suggest introducing a time valuation factor weighting cancer malignancy, and implying decreasing malignancy as cancer persists. Summing up, we propose to solve and simulate the optimal therapy problem

$$\min \sum_{t=0}^{T} \rho^t [u^2(t) + y_0^2(t) + y_1^2(t)] + \rho^{T+1}[y_0^2(T+1) + y_1^2(T+1)]$$

$$s.t. \begin{cases} x_0(t+1) = x_0(t) + n(1 - \frac{x_0(t)+y_0(t)}{K})x_0(t) - d_0 x_0(t) \\ x_1(t+1) = x_1(t) + r x_0(t) - (d - d_2)x_1(t) \\ y_0(t+1) = y_0(t) + m(1 - \frac{x_0(t)+\alpha y_0(t)}{K})y_0 - g_0 y_0(t) - \beta h(u(t))y_0(t) \\ y_1(t+1) = y_1(t) + q y_0(t) - (g - g_2)y_1(t) \\ x_0(t) \geq 0, x_1(t) \geq 0, y_0(t) \geq 0, y_1(t) \geq 0, u_{max} \geq u(t) \geq 0 \\ t = 0, 1, \ldots, T \\ x_0(0), x_1(0), y_0(0), y_1(0) \quad \text{initially given} \end{cases} \tag{1}$$

 ISBN: 978-84-617-8694-7

4 Solving and Calibrating the Model

We investigate the role played by the malignancy parameter ρ in the control problem. As usually happens with optimal control problems in Biomedicine, the optimal therapy problem (1) does not have an explicit algebraic solution. We therefore numerically solve the problem taking account of the necessary conditions that the variables must satisfy. These necessary conditions involve the system of difference equations for the state variables describing the dynamics of CML and with the associated initial conditions, the system of difference equations for the Lagrange multipliers with the corresponding final conditions, and the nonlinear equation for the control variable (see [4]).

The numerical solution is obtained running an iterative procedure: we start with an initial guess for $u(t)$ at $0 \leq t \leq T$; with this initial iterant, we compute the corresponding values of the state and Lagrange multiplier variables; finally, from these approximations, we compute a new control variable. We iterate this procedure until convergence to the optimal control $u^*(t)$.

We implement this numerical method for a practical case. To this end, it is previously necessary to calibrate the parameters, i.e., to assign a value to the parameters in the system of difference equations providing the solution. This calibration is carried out on the basis of the available recent biomedical data ([2], [3], [6], [8], [9]). After the calibration, the optimal therapy is simulated for two referential situations, namely with and without the proposed time valuation factor.

5 Results and Conclusions

Regarding the consequences of the consideration of a time valuation factor capturing the higher malignancy associated to early growth of cancer and drug administration, our results are unequivocal: *this consideration allows more efficient therapies to be designed.* More specifically, when this time valuation factor is considered, the optimal drug doses are considerably lower, and do not entail significant increases in the number of cancer cells or in the disease duration.

Acknowledgements

Financial support from Spanish Office of Economy and Competitiveness and European FEDER Funds, research project MTM2014-56022-C2-2-P, is gratefully acknowledged. The authors also thank the valuable comments of J. Russo, Fox Chase Cancer Center, US.

 ISBN: 978-84-617-8694-7

References

[1] B.E. Aïnseba and C. Benosman, *Optimal control for resistance and suboptimal response in CML*, Math. Biosci. **227** (2010) 81–93.

[2] V.S. Carneiro Maia, *Mecanismos Moleculares Implicados en la Regulación de la Apoptosis y la Adhesión Celular por la Ruta c3g/p38 Mapk. Implicaciones en la Patogénesis de la Leucemia Mieloide Crónica,* PhD dissertation, Cancer Research Center, University of Salamanca, 2012.

[3] S.N. Catlin, L. Busque, RE. Gale, P. Guttorp and J.L. Abkowitz, *The Replication Rate of Human Hematopoietic Stem Cells in Vivo*, Blood **117** (2011) 4460–4466.

[4] T.L. Friesz, *Dynamic Optimization and Differential Games*, Springer, New York, 2010.

[5] P. J. Gutiérrez Diez, I. H. Russo and J. Russo, *The Evolution of the Use of Mathematics in Cancer Research*, Springer, New York, 2012.

[6] H. Kantarjian, S. O'Brien, E. Jabbour, G. Garcia-Manero, A. Quintas-Cardama, J. Shan, M.B. Rios, F. Ravandi, S. Faderl, T. Kadia, G. Borthakur, X. Huang, R. Champlin, M. Talpaz and J. Cortes, *Improved survival in chronic myeloid leukemia since the introduction of imatinib therapy: A single-institution historical experience*, Blood **119** (2012) 1981–1987.

[7] U. Ledzewicz, H. Schättler, A. Friedman and E. Kashdan (eds.), *Mathematical methods and models in Biomedicine*, Springer, New York, 2013.

[8] T.G. Parslow, D.P. Stites, A.I. Terr and J.B. Imboden (eds.), *Medical Immunology*, Lange, New York, 2001.

[9] P.R. Wheater, H.G. Burkitt, V.G. Daniels and P.J. Deakin, *Functional Histology: A Text and Colour Atlas*, Churchill Livingstone, Edinburgh, 1987.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

An acceleration of the continuous Newton's method

J. M. Gutiérrez[1] and M. Á. Hernández-Verón[1]

[1] *Department of Mathematics and Computer Sciences, University of La Rioja, Logroño, Spain*

emails: jmguti@unirioja.es, mahernan@unirioja.es

Abstract

In this work we study some numerical properties of the continuous Newton's method, the continuous version of the classical Newton's method for solving nonlinear equations $p(z) = 0$. In fact, continuous Newton's method is an initial value problem whose solutions flow to a root of the equation. We show the influence of the multiplicity of the roots of the considered equation in the Jacobian matrix related to the problem. In addition we study some modifications of the continuous Newton's method that allow us to increase the velocity of the convergence of the solutions towards the roots of $p(z) = 0$.

Key words: Continuous Newton's method; nonlinear equations; fixed points.
MSC 2000: 65J15

1 Introduction

In this work we are concerned with some numerical properties of the initial value problem

$$z(0) = z_0, \quad z'(t) = -\frac{p(z(t))}{p'(z(t))}, \tag{1}$$

where $z : \mathbb{R} \to \mathbb{C}$ and $p(z)$ is a given complex function. This problem is called continuous Newton's method and it was initially posed in the late 1980s (see [4], [6], [7]). Later, the method was popularized by Neuberger [5].

Note that the classical Newton's method for solving a nonlinear equation $p(z) = 0$, can be seen as an Euler step to the differential equation (1). Even more, a general Euler approximation of the differential equation (1) with step size $h > 0$ gives rise to the damped Newton's method

$$z_{n+1} = z_n - h\frac{p(z_n)}{p'(z_n)}. \tag{2}$$

 ISBN: 978-84-617-8694-7

This relationship between the continuous and discrete Newton's method has been analyzed by Jacobsen et al. ([3]). Actually, they have studied the basins of attraction of the roots for methods (2) and they have realized that their fractal structure shrinks away when $h \to 0$. This fact has been also pointed out by Epureanu and Greenside ([1]). In addition, in [3] other ODE solvers, like Runge-Kutta methods or Adams-Bashforth methods have been considered. The influence of the step size in the fractal dimension of the basin boundaries is the same than in the Euler's case. However, higher-order approximations for solving ODEs did not necessarily produce root-finding methods with a good numerical efficiency. In [2] the relationship between the step size and the order of convergence is particularly considered. For a constant step size, these iterative root-finding methods have a low order of convergence (linear) but a high computational cost. For a non-constant step size, a plethora of root-finding methods can be constructed. In particular, the famous Chebyshev-Halley family of methods can be constructed in this way.

2 Modified continuous Newton's method

Neuberger ([5]) used the following improved version of continuous Newton's method for obtaining roots of a nonconstant complex polynomial p. For a given $z_0 \in \mathbb{C}$, the target is to get a continuous function $z : \mathbb{R} \to \mathbb{C}$ such that

$$z(0) = z_0, \quad p(z)'(t) = -p(z(t)),\ t \in \mathbb{R}. \tag{3}$$

Neuberger characterized the convergence of continuous Newton's method to the roots of p.

In this paper we consider the modified continuous Newton's method

$$\omega(0) = \omega_0, \quad p(\omega)'(t) = -kp(\omega(t)), \quad k > 0. \tag{4}$$

Our aim is to extend the results given by Neuberger and to analyze the influence of the parameter k in the velocity of convergence of the flow to the solutions. In fact, we can state that if ω is a solution of (4), then $p(\omega(t)) = p(\omega_0)e^{-kt}$.

For each $k > 0$, let us introduce the set $\mathcal{Q}_k$ of functions $\omega : \mathbb{R} \to \mathbb{C}$ which solve the differential equation (4). If $\omega \in \mathcal{Q}_k$, then $u = \lim_{t\to\infty} \omega(t)$ exists and $p(u) = 0$. In addition, we can deduce that the modified continuous Newton's method has the same properties than the classical Neuberger's problem (3).

3 Fixed points and equilibrium solutions

The fixed points (x^*, y^*) of a system of differential equations

$$\begin{cases} x'(t) &= f(x(t), y(t)), \\ y'(t) &= g(x(t), y(t)), \end{cases} \tag{5}$$

 ISBN: 978-84-617-8694-7

represent equilibrium solutions of the problem, since if $x(t) = x^*$ and $y(t) = y^*$ initially, then $x(t) = x^*$ and $y(t) = y^*$ for all time. We are interested in the study of the equilibria of the modified continuous Newton's method (4). Actually, if we make $z = x + iy$ we obtain a system of differential equations (5) where

$$f(x,y) = \mathbb{R}e\left(-\frac{p(x+iy)}{p'(x+iy)}\right), \quad g(x,y) = \mathbb{I}m\left(-\frac{p(x+iy)}{p'(x+iy)}\right). \tag{6}$$

Our first result characterizes the equilibria of (5) and establishes their local dynamical behavior in terms of the corresponding Jacobian matrices.

Theorem 1 *Let $z^* = x^* + iy^*$ be a zero with multiplicity m of a differentiable enough function $p(z)$. Then (x^*, y^*) is an asymptotically stable fixed point of the system (5) with f and g given by (6). In addition, if we denote $F(x,y)$ to the vector field $F(x,y) = (f(x,y), g(x,y))$, we have that the Jacobian matrix of $F(x,y)$ at (x^*, y^*) is*

$$\operatorname{Jac} F(x^*, y^*) = \begin{pmatrix} -k/m & 0 \\ 0 & -k/m \end{pmatrix}. \tag{7}$$

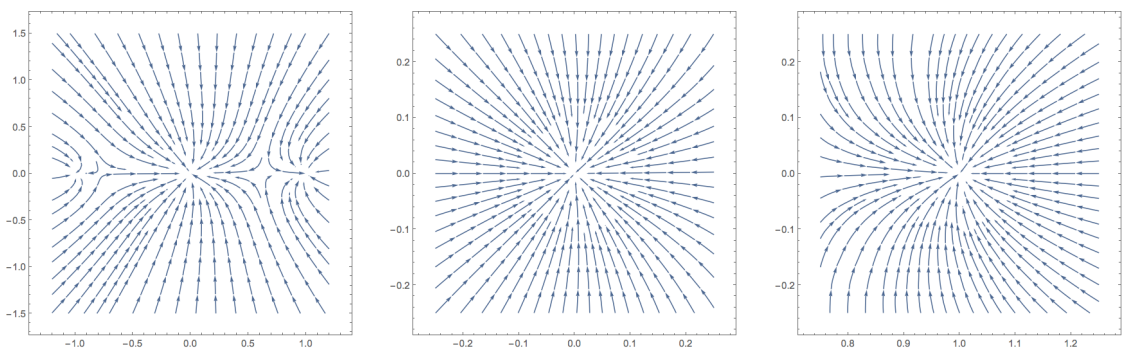


Figure 1: On the left, phase portrait of the continuous Newton's method applied to the polynomial $p(z) = z^3(z-1)^2(z+1)$. The other two figures are details around the points $(0,0)$ and $(1,0)$ respectively.

Remark 1 *Note that* $\operatorname{Jac} F(x^*, y^*)$ *is a diagonal matrix with a double eigenvalue of $\lambda = -k/m$. So from a local point of view, (x^*, y^*) behaves as a shrink star node whose typical phase portrait is shown in Figure 1. In fact, when the trajectories $(x(t), y(t))$ are close enough to the equilibrium point (x^*, y^*), these trajectories behaves as the functions*

$$\begin{cases} x(t) &= x^* + C_1 e^{-kt/m} \\ y(t) &= y^* + C_2 e^{-kt/m}. \end{cases}$$

 ISBN: 978-84-617-8694-7

So the convergence of the trajectories to the equilibrium point is slower when the multiplicity m increases. However the convergence of the trajectories to the equilibrium point is faster when the parameter k increases.

Acknowledgements

This work has been partially supported by the Spanish Ministry of Economy and Competitiveness under grant MTM2014-52016-C2-1-P.

References

[1] B. I. Epureanu and H. S. Greenside, *Fractal basins of attraction associated with a damped Newton's method*, SIAM Rev. **40** (1998) 102–109.

[2] J. M. Gutiérrez, *Numerical Properties of Different Root-Finding Algorithms Obtained for Approximating Continuous Newton's Method*, Algorithms **8** (2015) 1210–1218.

[3] J. Jacobsen, O. Lewis and B. Tennis, *Approximations of continuous Newton's method: An extension of Cayley's problem*, Electron. J. Diff. Equ. **15** (2007) 163–173.

[4] H. Jongen, P. Jonker and F. Twilt, *The continuous, desingularized Newton method for meromorphic functions*, Acta Appl. Math. **13** (1988) 81–121.

[5] J. W. Neuberger, *Continuous Newton's method for polynomials*, Math. Intell. **21** (1999) 18–23.

[6] H. Peitgen, M. Prufer and K. Schmitt, *Global aspects of the continuous and discrete Newton's method: A case study*, Acta Appl. Math. **13** (1988) 123–202.

[7] D. Saupe, *Discrete versus continuous Newton's method: A case study*, Acta Appl. Math. **13** (1988) 59–80.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

Dynamics of the FK3V cardiac cell model

Radek Halfar[1]

[1] *Department of Applied Mathematics, VŠB - Technical University of Ostrava,*
17. listopadu 15/2172, 708 33 Ostrava, Czech Republic.

emails: radek.halfar@vsb.cz

Abstract

The aim of this paper is to study evolution of transmembrane potencial on the cardiac cell under different rates of stimulation. For modeling this potential, the Fenton-Karma model was applied. It is a phenomenological model with three degree of freedom that corresponds to nondimensional transmembrane potencial and gating variables for regulation of inward and outward ion currents. Moreover, as a novelty, the model was forced by stimulus with shape of half sinus period. For solving the equations the explicit Runge-Kutta method was used. As the main aim of the paper it is shown that that the Fenton-Karma model is showing regular as well as irregular motion; periodic and also chaotic patterns are detected using bifurcation diagrams and 0-1 test for chaos.

Key words: cardiac cell model, bifurcation, 0-1 test for chaos
MSC 2000: 34H20, 34H10, 37N30

1 Introduction

The cardiac electrophysiology is the result of complex processes occurring on the heart cell membranes, which aim to ensure the proper progression of cardiac action potential inherent in heart muscle. Understanding of this process is very important for prevention of heart failure, incompatible with life, such as ventricular fibrillation.

The Fenton-Karma model (FK3V) is commonly used to describe a simplified model of the heart and electrophysiology. Studies that uses this model can be generally divided into several groups. Studies dealing with the determination of model parameters to better replicate the outputs obtained by using physiological models or experimental data, studies

 ISBN: 978-84-617-8694-7

that uses FK3V model to study the electrophysiology of the heart and studies that examine the characteristics of the model FK3V itself.

The first group can include for example the study [7], which describes a procedure for finding such parameters FK3V model to its restitution characteristics correspond restitution characteristics of Courtemanche-Ramirez Nattel model of atrial tissue. In [6], the authors then edited the parameters of detailed and simplified (FK3V) model to fit for five clinical data of patients undergoing ablation therapy.

Termination of fibrillation is studied in [9]. Their objective was to numerical validation the experimental techniques for terminating fibrillation presented in [8]. In his work, they found that synchronized defibrillation can create a low-energy alternative to traditional defibrillation. In [1] the authors studied electrophysiological and dynamic mechanism of spiral waves break up. The authors found several alternans unstable modes with different growth rates, frequencies and spatial structures. In a study [2], the authors examined the behavior of fibers in a computing FK3V monodomain anisotropic model of re-entrant ventricular fibrillation.

The main aim of this paper is the research of dynamical properties of the FK3V model in dependence of the stimulation frequency; moreover regular as well as chaotic moments were observed.

2 FK3V model

FK3V model is ionic model of cardiac action potencial which is based on the Luo-Rudy I model. For reproducing action potencial it uses three variables u, v and w. Variable u represent nondimensional transmembrane potential (so that $u = 0$ and $u = 1$ are the rest and peak voltages, respectively) Transmembrane potential changing depending on ionic currents according to equation:

$$\dot{u} = J_{stim} - J_{fi} - J_{so} - J_{si}$$

where J_{fi} (accountable for depolarization of the membrane), J_{so} (accountable for repolarization of the membrane) and J_{si} (balances J_{so} during the plateau phase) are cross-membrane currents named fast inward (fi), slow outward (so) and slow inward (si) that roughly corresponds to sodium, potassium and calcium ion currents respectively. But because they do not represent quantitatively measured currents, but only their activation, inactivation, and reactivation it is preferred to call these currents as fast and slow inward, and slow outward, rather than Na, Ca, and K as a reminder of these simplification. J_{stim} indicates externally applied current. In this study is external current composition of pulses created by first half period of sinus function followed by zero function. J_{stim} is therefore defined by equation:

$$J_{stim} = \begin{cases} 0.48\sin(t - n(c+1)) & t \in < n(c+1), n(c+1)+1 > \quad n \in \mathbb{N} \cup \{0\} \\ 0 & t \notin < n(c+1), n(c+1)+1 > \quad n \in \mathbb{N} \cup \{0\}. \end{cases}$$

 ISBN: 978-84-617-8694-7

Where c denotes length of time interval between pulses. Dot over a variable denotes derivative with respect to time.
Cross membrane currents are given by

$$J_{fi}(u;v) = \Theta(u-u_c)(1-u)(u-u_c)\frac{-v}{\tau_d},$$

$$J_{so}(u) = \Theta(u_c-u)\frac{u}{\tau_0} + \Theta(u-u_c)\frac{1}{\tau_r},$$

$$J_{si}(u;w) = \frac{(1+\tanh\left[k(u-u_c^{si})\right])}{2}\frac{-w}{\tau_{si}},$$

where $\Theta(x)$ is the Heaviside function which replaced the gating functions $h_\infty(V)$, $m_\infty(V)$, $j_\infty(V)$, and $f_\infty(V)$, in the Beeler-Reuter or Luo-Rudy-I models. Function

$$\frac{(1+\tanh\left[k(u-u_c^{si})\right])}{2}$$

is than smooth function that replaced $d_\infty(V)$. to provide good fit of APD restitution curves.
Another two variables v and w used in the model are gating variables which regulates inactivation of J_{fi} and J_{so} takes the following form:

$$\dot{v} = \Theta(u_c-u)(1-v)\frac{1}{\tau_v^-(u)} - \Theta(u-u_c)v\frac{1}{\tau_v^+},$$

$$\dot{w} = \Theta(u_c-u)(1-w)\frac{1}{\tau_w^-} - \Theta(u-u_c)w\frac{1}{\tau_w^+},$$

$\tau_v(u)$ is function for defining time constants for two voltage ranges ($u_v < u < u_c$ and $u < u_v$) and is introduced for proper reproducing CV restitution curve. It controls reactivation of J_{fi} and is given by equation

$$\tau_v^-(u) = \Theta(u-u_v)\tau_{v_1}^- + \Theta(u_v-u)\tau_{v_2}^-$$

The model contains several constants, which are used for fitting the output curves into requested shape, time constants τ_r, τ_{si}, τ_0 ... and threshold potentials u_c, u_c^{si} and u_v.
The original paper [3] describes four different sets of parameters to fit for different dataset.

- BR: obtained by stimuli of Beeler-Reuter model with standard parameter values.
- MBR: obtained by stimuli of modified Beeler-Reuter model with speeded up calcium kinetic.
- MLBR-I: stimuli of Luo-Ruby-I model with speeded up calcium kinetic.

 ISBN: 978-84-617-8694-7

- GP: experimental data extracted by Girouard et al. from measuring membrane potentials on the epicardial surface of left Ventricle of a guinea pig.

Table 1: Original published parameters of FK3V model [3] for BR parameters set, $k = 10$

Parameter	**Description**	**unit**	**value**
τ_d	setting influence J_{fi} for $u > u_c$	ms	0.25
τ_r	setting influence J_{so} for $u > u_c$	ms	33
τ_{si}	setting influence $J_s i$ on $\dot{u}$	ms	30
τ_0	setting decrease u to 0 for $u < u_c$	ms	12.5
τ_v^+	setting decrease v to 0 for $u > u_c$	ms	3.33
$\tau_{v_1}^-$	setting value for $\tau_v^-(u)$ for $u > u_v$	ms	1250
$\tau_{v_2}^-$	setting value for $\tau_v^-(u)$ for $u < u_v$	ms	19.6
τ_w^+	setting decrease w to 0 for $u > u_c$	ms	870
τ_w^-	setting increase w to 1 for $u < u_c$	ms	41
u_c	depolarization threshold	-	0.13
u_v	threshold for activation $\tau_{v_1}^-$ or $\tau_{v_2}^-$	-	0.04
u_c^{si}	threshold for opening J_{si}	-	0.85

3 Main results

In this work, the BR parameter set was used. In individual simulations, the heart cell was irritated with the half-sinus shaped current pulses with amplitude 0.48 (twice as needed to cause irritation) and duration 1 ms. The individual stimulation pulses were separated by the delay c (see equation defining J_{stim} current). Computations was done for stimulation delays from 10 to 300 ms with step of 5 ms. Each simulation was done for time from 0 to 5×10^5 ms. From the last 20 % of results was consequently created a phase diagram for each simulated frequency and bifurcation diagram from the entire simulated frequency spectrum.

Chaotic behavior of the model was observed on stimulation delays from 10 to 85 ms (see Figures 1 and 2). This behavior can also be seen in bifurcation diagrams in Figures 4 and 5. Next, regular responses to stimulation impulses are observable on delays from 85 to 300 ms and this model output is shown in Figure 3 and also in bifurcation diagrams in Figures 4 and 5.

 ISBN: 978-84-617-8694-7

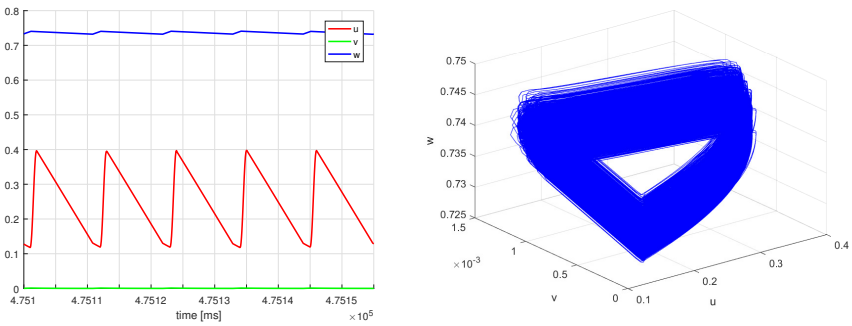


Figure 1: Time responses of u, v, w (left) and phase diagram (right) of the FK3V model for $c = 10$ ms.

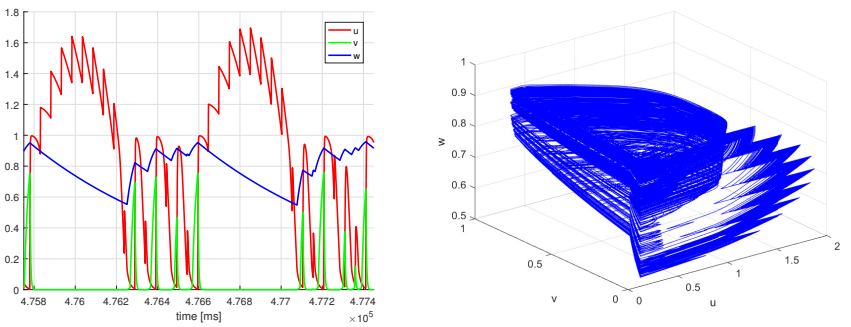


Figure 2: Time responses of u, v, w (left) and phase diagram (right) of the FK3V model for $c = 50$ ms.

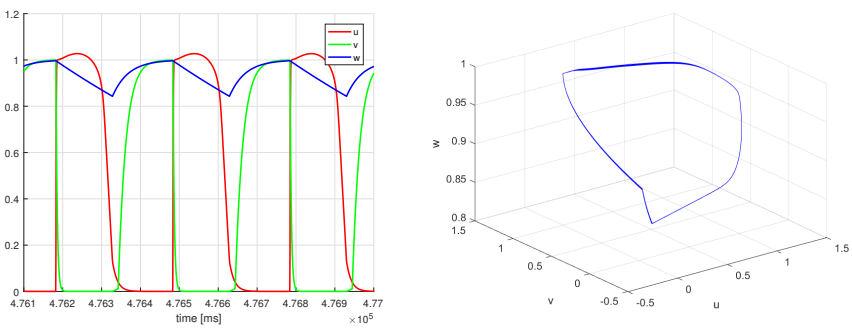


Figure 3: Time responses of u, v, w (left) and phase diagram (right) of the FK3V model for $c = 300$ ms

 ISBN: 978-84-617-8694-7

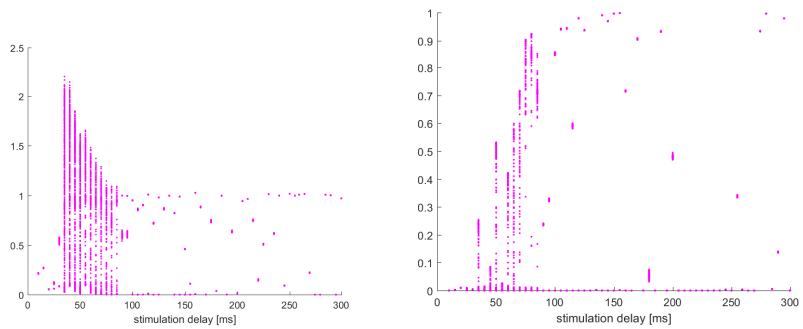


Figure 4: FK model bifurcation of variable u (lef) and v (right).

Figures 4, and 5 shows bifurcation diagrams of FK3V model. Values for diagrams was collected witch period of stimulation frequency (recorded was every fifth stimulus). In Figures can be seen that for stimulation delays from 35 to 85 ms is is response of the model chaotic.

Finally, 0-1 test for chaos was performed. This test, introduced in [4] (see also [5]), is used to distinguish regular and chaotic dynamic. It works with the time series and does not need any phase space reconstruction. The resulting value of this test can only be 0 (regular behavior) or 1 (chaos). The results of this test can bee seen in Figure 6. This test coincides just for variable w.

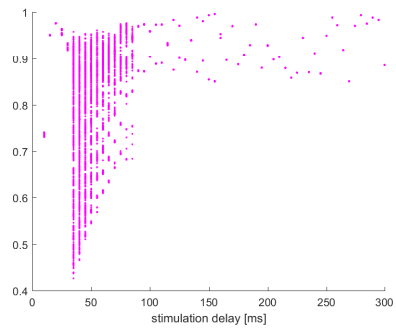


Figure 5: Bifurcation diagram of variable w.

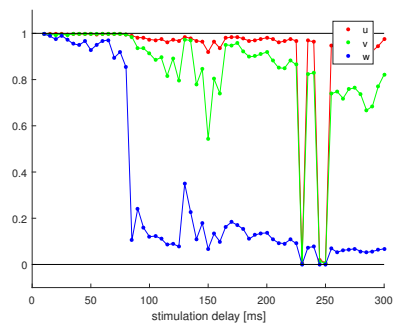


Figure 6: Results of the 0-1 test for chaos of variables u, v and u.

 ISBN: 978-84-617-8694-7

4 Conclusions

In this paper, responses of Fenton-Karma model of cardiac cell to various rates of stimulation frequencies were analyzed. Newly, the model was forced by stimulus with shape of half sinus period. The equations of potentials were solved numerically using Runge-Kutta method of the fourth order as *ode45* solver in Matlab.

It was observed that model is showing regular and also chaotic pattern for different range of stimulation delay. Chaotic behavior of variables u, v and w were confirmed by 0-1 test for chaos for suitable choices of stimulation delays.

Acknowledgements

The author would like to thank Prof. M. Lampart for the valuable suggestions that improved the paper.

This work was supported by The Ministry of Education, Youth and Sports from the Large Infrastructures for Research, Experimental Development and Innovations project IT4Innovations National Supercomputing Center LM2015070"; and by grant SGS No. SP2017/122, VŠB - Technical University of Ostrava, Czech Republic.

References

[1] G. D. Allexandre, N. Otani, *Preventing alternans-induced spiral wave breakup in cardiac tissue: An ion-channel-based approach*, Phys. Rev. E **70** (2004) 061903.

[2] G. R.H. Clayton, A.V. Holden, *Filament behavior in a computational model of ventricular fibrillation in the canine heart*, IEEE Transactions on Biomedical Engineering **51** (2004) 28-34.

[3] G. E. Flavio Fenton, Alain Karma, *Vortex dynamics in three-dimensional continuous myocardium with fiber rotation: Filament instability and fibrillation*, Chaos: An Interdisciplinary Journal of Nonlinear Science **8** (1998) 20-47.

[4] G. A Gottwald, I. Melbourne, *A new test for chaos in deterministic systems*, Proc. R. Soc. London A **460** (2004) 603–611.

[5] G. A Gottwald, I. Melbourne, *On the implementation of the 0-1 test for chaos*, SIAM J. Appl. Dyn. **8** (2009) 129–145.

[6] G. D. Lombardo, F. Fenton, S. Narayan, W. Rappel, *Comparison of detailed and simplified models of human atrial myocytes to recapitulate patient specific properties*, PLOS Computational Biology **12** (2016) 1-15.

 ISBN: 978-84-617-8694-7

[7] G. R. Oliver, W. Krassowska, *Reproducing cardiac restitution properties using the FentonKarma membrane model*, Annals of Biomedical Engineering **33** (2005) 907-911.

[8] G. H. Pak, Y. Liu, H. Hayashi, Y. Okuyama, P. Chen, S. Lin, *Synchronization of ventricular fibrillation with real-time feedback pacing: implication to low-energy defibrillation*, American Journal of Physiology - Heart and Circulatory Physiology **285** (2003) H2704-H2711.

[9] G. S. Puwal, B. Roth, *Numerical simulations of synchroniyed pacing*, Journal of Biological Systems **14** (2006) 101-112 .

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

Cloud implementation of logistic regression for hyperspectral image classification

Juan Mario Haut[1], Mercedes Eugenia Paoletti[1], Abel Paz-Gallardo[2], Javier Plaza[1] and Antonio Plaza[1]

[1] *Department of Technology of Computers and Communications, University of Extremadura, Escuela Politecnica, Avda. de la Universidad s/n, Cáceres*

[2] *Extremadura Research Centre for Advanced Technologies, CETA-CIEMAT, Calle Sola, 1, 10200, Trujillo, Cáceres*

emails: juanmariohaut@unex.es, mpaolett@alumnos.unex.es, abelfrancisco.paz@ciemat.es, jplaza@unex.es, aplaza@unex.es

Abstract

Classification of remotely sensed hyperspectral images is a challenging task due the enormous amount of information comprised in these images, that contain hundreds of continuous spectral bands. This creates a need to develop new techniques for hyperspectral classification using high performance computing architectures. Despite the availability of multiple algorithms adapted to parallel environments (such as multicore computers or accelerators like field programmable gate arrays or graphics processing units, the application of cloud computing techniques has not been as widespread, although there are many potential advantages in exploiting cloud computing architectures for distributed hyperspectral image analysis. In this paper, we present a cloud implementation (developed using Apache Spark) of a successful technique for hyperspectral image classification: the multinomial logistic regression probabilistic classifier. Our experimental results suggest that cloud computing architectures allow for the efficient classification of large hyperspectral image data sets.

Key words: Hyperspectral imaging, multinomial logistic regression, cloud computing, Apache Spark.

 ISBN: 978-84-617-8694-7

1 Introduction

Remotely sensed hyperspectral imaging is a popular technique for Earth observation (EO) [1], which allows for the simultaneous collection of images (at different wavelength channels) for the same area on the surface of the Earth. A characteristic of hyperspectral imagers is that they can collect data in thousands of narrow, contiguous spectral bands [2], providing so-called hyperspectral image data cubes [3].

An important property of hyperspectral instruments is their ability to acquire a complete reflectance spectrum for each pixel in the image (called contiguous spectral curves or spectral signatures). These signatures allow us to accurately distinguish different physical materials. For instance, the NASA's Jet Propulsion Laboratory's Airbone Visible/Infrared Imaging Spectrometer (AVIRIS) [4] measures the solar reflected spectrum from 0.4μm to 2.5μm at intervals of 0.01μm. The EO-1 Hyperion imaging spectrometer also collects bands in the range of 0.4μm to 2.5μm (more than 200 bands in both cases) [5, 6]. Several new satellite mission that will be soon operative and ready to collect data in a very similar spectral range. For instance, the German Environmental Mapping and Analysis Program (EnMAP [7]) is expected to collect data in the range 0.42μm to 2.45μm, as well as the Italian PRISMA program [8]. Other spectrometers acquire hyperspectral images in other regions of the spectrum, for instance the Reflective Optics System Imaging Spectrometer (ROSIS) takes images with a spectral range from 0.43μm to 0.96μm [9].[2].

Hyperspectral imaging has proved to be useful over a wide range of applications, such as agriculture, forestry, geology, ecological monitoring and disaster monitoring [10, 6]. However, due to the great dimensionality of hyperspectral data cubes, analysis techniques exhibit significant requirements in terms of storage and data processing [11, 12]. Therefore, the development of techniques that are computationally efficient becomes critical [6, 13, 14, 15].

Many efforts have been made within the field of hyperspectral image classification, both supervised and unsupervised [16]. Supervised techniques have been generally more popular due to their higher classification accuracy, but they require sufficient training information in order to perform properly. One of the supervised classifiers that can perform more accurately in the presence of limited training samples is the multinomial logistic regression (MLR) [17]. However, this classifier is computationally expensive, and available implementations have not considered the possibility of using cloud computing architectures [18]. These platforms can be greatly beneficial for hyperspectral image classification due to their advanced capabilities for internet-scale, service-oriented and high-performance computing. Specifically, the use of cloud computing for the classification of large hyperspectral data repositories can be considered a natural solution and an evolution of previously developed techniques for other kinds of computing platforms [19]. Still, there are few efforts in the recent literature oriented to the exploitation of cloud computing infrastructure for hyperspectral imaging techniques.

 ISBN: 978-84-617-8694-7

This work explores the possibility of using a distributed framework for classification of massive hyperspectral images based on cloud computing architectures. In particular we have focused on the discriminative MLR classifier [17] to demonstrate the applicability of utilizing cloud computing technologies to efficiently perform distributed classification of hyperspectral data.

The remainder of the paper is organized as follows. Section 2 first presents the theoretical principles of the MLR method (section2.1). Then, it describes our distributed framework design for this classifier (section 2.2). Finally, it describes our cloud implementation in detail (section 2.3). Section 3 validates the proposed cloud MLR algorithm by comparing it with other implementations. Finally, section 4 concludes with some remarks and hints at plausible future research lines.

2 Methodology

2.1 Multinomial Logistic Regression

To understand the operations of the MLR, we first we need to describe how logistic regression (LR) works. Given a collection of n linear-separable numeric samples $X = \{x_1, ..., x_n\}$ where each $x_i \in \mathbb{R}^d$, $x_i = [x_{i,1}, ..., x_{i,d}]$, the goal of classification methods is to categorize each x_i into a class or category y_i of those available in $Y = \{y_1, ..., y_k\}$, with $k < n$. But, in contrast to other classification methods, LR does not try to predict the value of a x_i given a set of inputs. Instead, the output is a probability that the input x_i belongs to a certain class y_i. It would be 0 when x_i does not belong to y_i and 1 if x_i belongs to y_i. Suppose that $x_i = [x_{i,1}, ..., x_{i,d}]$ and $Y = \{-, +\}$, LR assumes that the input d-space can be separated into two regions by a linear boundary: $\beta_0 + \beta_1 x_{i,1} + ... + \beta_d x_{i,d}$. This function outputs a value in $(-\infty, \infty)$ given an input data point, x_i[1]. To map the label probabilities with boundary values, LR applies log-odds functions[2] and calculates the predicted probabilities as $P(y_i = +|x_i, \beta) = \frac{\exp(\beta_0 + \sum_{j=1}^{d} \beta_j x_{i,j})}{1+\exp(\beta_0 + \sum_{j=1}^{d} \beta_j x_{i,j})}$. The goal of LR is to estimate the coefficients $\beta = \{\beta_0, \beta_1, ..., \beta_d\}$ through *maximum likelihood estimation* (MLE), that optimizes β in order to maximize the *log likelihood* (LL, i.e. the log odds).

MLR extends the binary problem of LR to any number of classes, $k > 2$. Specifically,

[1] If x_i lies in the region defined by the + class, the function's value is positive in the range $(0, +\infty)$ and its probability $P(y_i = +|x_i, \beta)$ is in $(0.5, 1]$. If x_i lies in the region defined by the − class, the function's value is negative in the range $(-\infty, 0)$ and its probability $P(y_i = -|x_i, \beta) = 1 - P(y_i = +|x_i, \beta)$ is in $[0, 0.5)$. Finally if we do not know whats x_i is, the function's value is 0, and probabilities of being + or − are exactly 0.5

[2] Given a probability function $P(x) \in [0, 1]$, the odds ratio is defined as $OR(x) = \frac{P(x)}{1-P(x)} \in [0, +\infty)$. Applying the logarithm to $OR(x)$ we obtain $\log(OR(x)) \in (-\infty, \infty)$. If the log-odds is $\log(\frac{P(x)}{1-P(x)}) = a + bx$, we can transform it into $\frac{P(x)}{1-P(x)} = \exp(a + bx) \rightarrow P(x) = \frac{\exp(a+bx)}{1+\exp(a+bx)}$, i.e the logistic function

 ISBN: 978-84-617-8694-7

MLR selects one category as the baseline, e.g. the k-th class, and calculates the regression coefficients for the $l = 1, ..., k-1$ non-baseline categories $(\beta^{(1)}, ..., \beta^{(k-1)}$ with $\beta^{(l)} = \{\beta_0^{(l)}, ..., \beta_d^{(l)}\})$ against the baseline class. The predicted probabilities are extended to $P(y_i = l|x_i, B) = \frac{\exp\left(\beta_0^{(l)} + \sum_{j=1}^{d} \beta_j^{(l)} x_{i,j}\right)}{1 + \sum_{l'=1}^{k-1} \left(\exp\left(\beta_0^{(l')} + \sum_{j=1}^{d} \beta_j^{(l')} x_{i,j}\right)\right)}$, where B is the $(d+1) \times (k-1)$ matrix of all the regression coefficients. The goal of MLR is then to estimate B given the samples dataset X and the categories Y, by minimizing the optimization function:

$$f(B; X, Y) = -\sum_{i=1}^{n} \log P(y_i|x_i B) + \frac{\lambda}{2} \sum_{j=1}^{d} \sum_{l=1}^{k-1} |\beta_j^{(l)}|^2, \tag{1}$$

where λ is a regularization term added in order to mitigate the overfitting problem.

2.2 Distributed framework design

To create our distributed environment, two frameworks have been used: 1) *OpenStack*[3] and 2) *Apache Spark*[4]. Each one of them will be in charge of the correct execution of the architecture in two aspects:

- On the one hand, *OpenStack* provides Infrastructure as a Service (IaaS), abstracting and manages the physical machines that will give the support to the virtual machines. *OpenStack* works like a cloud operating system that controls large pools of compute, storage, and networking resources throughout a datacenter, all managed through a dashboard that gives administrators control while empowering their users to provision resources through a web interface.

- On the other hand, *Apache Spark* is a distributed in-memory processing framework that works over the virtual machines (managed by *OpenStack*) and allows to implement MapReduce[5] distributed programming model [18]. Also, *Apache Spark* implements a fault-tolerant abstraction for in-memory cluster computing, and provides fast and general data processing on large distributed platforms. It supports simple one-pass computations and can also be extended to the case of multi-pass, iterative algorithms.

[3] https://wiki.openstack.org/wiki/Main_Page

[4] http://spark.apache.org/

[5] The MapReduce model takes full advantage of the high-performance capabilities provided by cloud computing architectures. The operation is easy: a task is processed by two distributed operations, map and reduce. The datasets are organized as key/value pairs, and the map function processes a key/value pair to generate a set of intermediate pairs, dividing a task into several independent subtasks to be run in parallel. The reduce function is in charge of processing all intermediate values associated with the same intermediate key, then collecting all the subtask results to gather the result for the whole task.

 ISBN: 978-84-617-8694-7

The full architecture of our newly developed system for hyperspectral image classification is shown in Fig. 1. In Fig. 2 we display the services offered by *OpenStack* and the *Apache Spark* framework used.

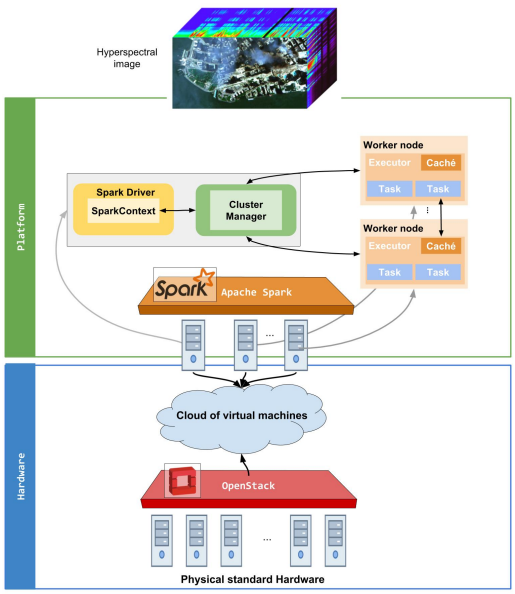


Figure 1: Integrated OpenStack and Apache Spark framework for Logistic Regression.

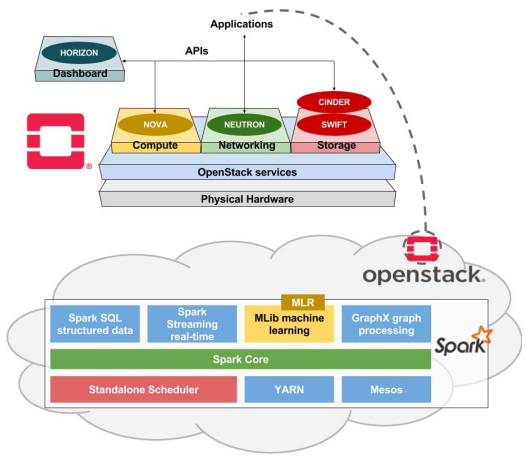


Figure 2: Description of the proposed Apache Spark and Open Stack architecture.

2.3 Cloud Implementation

Proposed cloud MLR divides the execution between one master (Spark driver) and several slaves (Spark executors). The driver prepares the environment (reserves resources), launches the executors and initializes β. Each executor loads their corresponding patch of the hyperspectral image and for each data calculates the loss function and the gradient (Map) that are summed up (Resume) and sent back to the driver. The driver calculates the loss and the gradient of regularizer and plugs the gradient and loss of the model and the regularizer into optimizer to get the new β. If the loss is less than the stopping criterion, the algorithm ends.

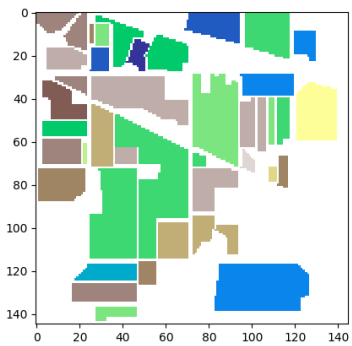
To execute our cloud implementation of MLR we need several parameters: the number of classes (k), the input training data (a percentage of hyperspectral image's pixels with which MLR will train), the number of maximum iterations and the tolerance of the L-BFGS optimizer. On the other hand, the input data is regularized by L2.

 ISBN: 978-84-617-8694-7

3 Experimental results

3.1 Experimental Configuration

In order to evaluate the performance of the adopted MLR implementation, we use a hardware environment composed by a Intel(R) Xeon(R) CPUs E5430 @ 2.66GHz (8 cores), 16 GB RAM, Shared storage, NetApp FAS3140. Virtual nodes have two virtual CPUs, 4GB of RAM and 40 GB hard disk each. In addition, we have developed a parallel version of the algorithm for comparative purposes. This version has been implemented on a paltform with Intel(R) Core(TM) i7-4790 CPUs @ 3.60GHz (8 cores), 16 GB RAM, SanDisk SDSSDA240G. In our experiments, we used Ubuntu 14.04 x64 LTS as operating system. For the parallel version of MLR, a virtual machine of the cluster with 2 cores, 4GB of RAM and 40GB of hard disk, and the same software configuration has been used.



Color	Land cover type	Samples
	Background	10776
	Alfalfa	46
	Corn-notill	1428
	Corn-min	830
	Corn	237
	Grass/Pasture	483
	Grass/Trees	730
	Grass/pasture-mowed	28
	Hay-windrowed	478
	Oats	20
	Soybeans-notill	972
	Soybeans-min	2455
	Soybean-clean	593
	Wheat	205
	Woods	1265
	Bldg-Grass-Tree-Drives	386
	Stone-steel towers	93
	Total samples	21025

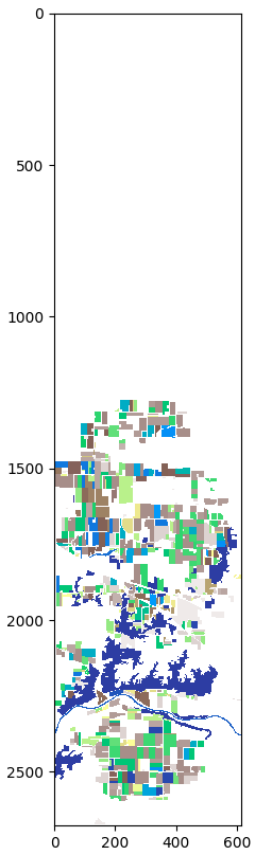
Figure 3: Original ground-truth of Small Indian Pines scene, with class labels and original number of samples per class.

3.2 Hyperspectral data sets

In our experiments, we use two different hyperspectral images. The first one was collected by AVIRIS [4] in 1992 over a set of agricultural fields with regular geometry and irregular patches of forest in Northwestern Indiana (Indian Pines image). This scene has 145×145 pixels with 224 spectral bands in the range 0.4-2.5μm, with 0.01μm of spectral resolution, 0.020μm moderate spatial resolution and 16 bits of radiometric resolution. 4 zero bands plus 20 bands with lower signal-to-noise ratio (SNR) have been removed, retaining 200 spectral

 ISBN: 978-84-617-8694-7

channels. The data has 16 ground-truth classes (Fig. 3). Also, we use a larger version of



Color	Land cover type	Samples	Color	Land cover type	Samples
	Background	1310047		BareSoil	57
	Buildings	17195		Concrete/Asphalt	69
	Corn	17783		Corn?	158
	Corn-EW	514		Corn-NS	2356
	Corn-CleanTill	12404		Corn-CleanTill-EW	26486
	Corn-CleanTill-NS	39678		Corn-CleanTill-NS-Irrigated	800
	Corn-CleanTilled-NS?	1728		Corn-MinTill	1049
	Corn-MinTill-EW	5629		Corn-MinTill-NS	8862
	Corn-NoTill	4381		Corn-NoTill-EW	1206
	Corn-NoTill-NS	5685		Fescue	114
	Grass	1147		Grass/Trees	2331
	Grass/Pasture-mowed	19		Grass/Pasture	73
	Grass-runway	37		Hay	1128
	Hay?	2185		Hay-Alfalfa	2258
	Lake	224		NotCropped	1940
	Oats	1742		Oats?	335
	Orchard	39		Pasture	10386
	pond	102		Soybeans	9391
	Soybeans?	894		Soybeans-NS	1110
	Soybeans-CleanTill	5074		Soybeans-CleanTill?	2726
	Soybeans-CleanTill-EW	11802		Soybeans-CleanTill-NS	10387
	Soybeans-CleanTill-Drilled	2242		Soybeans-CleanTill-Weedy	543
	Soybeans-Drilled	15118		Soybeans-MinTill	2667
	Soybeans-MinTill-EW	1832		Soybeans-MinTill-Drilled	8098
	Soybeans-MinTill-NS	4953		Soybeans-NoTill	2157
	Soybeans-NoTill-EW	2533		Soybeans-NoTill-NS	929
	Soybeans-NoTill-Drilled	8731		Swampy Area	583
	River	3110		Trees?	580
	Wheat	4979		Woods	63562
	Woods?	144			
				Total samples 1644292	

Figure 4: Original ground-truth of Big Indian Pines scene, with class labels and original number of samples per class.

the Indian Pines scene, with a size of 2678×614 pixels. It was collected over the same area, but spanning a much larger extent. It contains 220 spectral bands an the total number of classes is 58 (Fig. 4).

3.3 Performance Evaluation

To evaluate the performance of our cloud implementation of MLR, we make a comparison between the cloud version and a multi-core parallel implementation of the same MLR algorithm. Our experiments have been launched for each hyperspectral image, using different training percentages (15%, 25% and 50% of the training samples available in each

class). For the cloud version we have considered 2, 4 and 8 distributed nodes. The optimal number of iterations and λ value are obtained by cross-validation. Each configuration has been repeated five times, and the results reported are the average across the executions for statistical consistency.

Training percentage	**Parallel**	**Distributed**		
		2 nodes	**4 nodes**	**8 nodes**
	Time Execution			
5%	1.82 (2.143)	16.34 (1.876)	16.90 (2.102)	19.32 (2.201)
15%	4.48 (2.051)	18.65 (2.039)	20.01 (1.872)	23.50 (2.231)
25%	6.84 (1.942)	21.04 (2.052)	22.61 (2.312)	25.32 (2.214)
50%	13.42 (2.161)	30.34 (1.911)	32.39 (1.891)	32.84 (1.857)
	Accuracy Results			
5%	68.25 (1.0)	67.11 (0.9)	68.19 (1.0)	67.02 (0.8)
15%	77.15 (0.8)	75.74 (0.8)	75.00 (1.0)	73.77 (0.9)
25%	79.58 (0.9)	77.27 (0.9)	76.69 (1.1)	76.58 (0.9)
50%	82.05 (1.4)	78.40 (1.1)	79.36 (1.2)	79.11 (1.0)

Table 1: Average processing time, classification accuracy (and standard deviation) for different implementations of multinomial logistic regression using the Small Indian Pines Image.

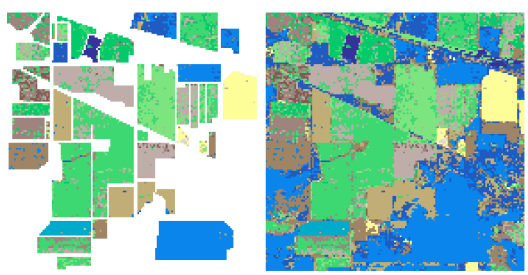


Figure 5: Classification results for the Small Indian Pines image: classification map without background (left) and classification map with background (right), obtained using 15% training.

Table 1 shows the results obtained by different implementations of the MLR using the Small Indian Pines dataset. The classification accuracies are worse as we add nodes to the cluster, due to the lack of data within the nodes, and the processing times tend to be worse too. This is because the nodes have not enough data to optimize. Fig. 5 shows the obtained classification result. As mentioned before, the processing times increase slightly as we add nodes to the distributed environment (with 8 nodes the weight of the communication prevents to improve the speed up). However, the increase of training samples affects non-uniformly the obtained classification results (this depends on the calculation of

 ISBN: 978-84-617-8694-7

λ).

Training percentage	Parallel	Distributed		
		2 nodes	4 nodes	8 nodes
	Time Execution			
5%	163.83 (2.620)	169.12 (2.527)	106.22 (1.403)	79.33 (2.300)
15%	400.20 (9.601)	364.00 (4.341)	218.46 (5.231)	127.36 (3.053)
25%	598.31 (3.310)	556.48 (3.254)	327.106 (1.900)	181.65 (1.024)
50%	1208.98 (11.708)	1087 (9.865)	584.65 (5.651)	374.42 (3.643)
	Accuracy Results			
5%	44.52 (1.1)	42.24 (0.9)	42.47 (0.9)	42.48 (1.0)
15%	45.39 (1.2)	43.06 (1.1)	43.02 (1.3)	43.88 (1.2)
25%	45.52 (1.1)	43.54 (1.2)	43.63 (0.9)	43.95 (1.1)
50%	45.65 (1.3)	44.29 (1.3)	44.34 (1.2)	44.96 (1.3)

Table 2: Average processing time, classification accuracy (and standard deviation) for different implementations of multinomial logistic regression using the Big Indian Pines Image.

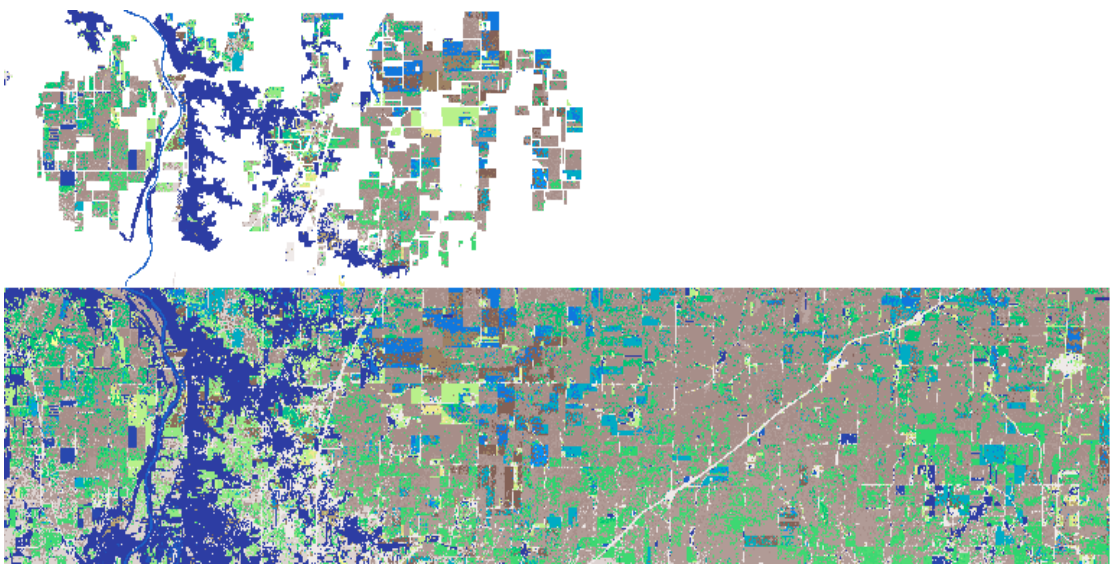


Figure 6: Classification results for the Big Indian Pines image: classification map without background (top) and classification map with background (bottom), obtained using 15% training.

On the other hand, Table 2 shows the results obtained by MLR using the Big Indian Pines dataset. As we can see, as we increase the number of nodes, time decreases. The highest speed up is achieved with 8 nodes (a 3.29), while the accuracy results are quite

acceptable given the complexity of this scene (although is less than in the parallel version). These results reveal that our cloud implementation benefits from the availability of large data volumes and complex analysis scenarios, such as the one given by the Big Indian Pines scene. The complexity of the classification of this scene can be appreciated in Fig. 6.

4 Conclusions and Future Lines

In this paper, we have discussed the possibility of exploiting cloud computing architectures for hyperspectral image classification. As a case study, we have presented a cloud computing implementation of the multinomial logistic regression classifier (a technique that has been used successfully for hyperspectral data interpretation) on the Apache Spark and Openstack platforms. Our experimental results show the effectiveness of the proposed distributed implementation with large hyperspectral datasets (i.e., the proposed technique provides satisfactory results with very large images and complex analysis scenarios given by a large numbers of samples and classes). As future work, we will implement other techniques for hyperspectral image classification using cloud computing platforms, as it is our feeling that there are many open and unexplored possibilities for the exploitation of these kind of platforms in remotely sensed hyperspectral imaging.

Acknowledgements

This work has been supported by Ministerio de Educación (Resolución de 26 de diciembre de 2014 y de 19 de noviembre de 2015, de la Secretaría de Estado de Educación, Formación Profesional y Universidades, por la que se convocan ayudas para la formación de profesorado universitario, de los subprogramas de Formación y de Movilidad incluidos en el Programa Estatal de Promoción del Talento y su Empleabilidad, en el marco del Plan Estatal de Investigación Científica y Técnica y de Innovación 2013-2016). This work has also been supported by Junta de Extremadura (decreto 297/2014, ayudas para la realización de actividades de investigación y desarrollo tecnológico, de divulgación y de transferencia de conocimiento por los Grupos de Investigación de Extremadura, Ref. GR15005).

This work was partially supported by the computing facilities of Extremadura Research Centre for Advanced Technologies (CETA-CIEMAT), funded by the European Regional Development Fund (ERDF). CETA-CIEMAT belongs to CIEMAT and the Government of Spain.

References

[1] D. Chutia, D. K. Bhattacharyya, K. K. Sarma, R. Kalita, and S. Sudhakar. Hyperspectral Remote Sensing Classifications: A Perspective Survey. *Transactions*

in GIS, 20(4):463–490, 2016.

[2] Alexander F H Goetz, Gregg Vane, Jerry E Solomon, and Barrett N Rock. Imaging Spectrometry for Earth Remote Sensing. *Science*, 228(4704):1147–1153, 1985.

[3] Chein-I Chang. *Hyperspectral Imaging: Techniques for Spectral Detection and Classification*. Springer US, 2003.

[4] Robert O. Green, Michael L. Eastwood, Charles M. Sarture, Thomas G. Chrien, Mikael Aronsson, Bruce J. Chippendale, Jessica A. Faust, Betina E. Pavri, Christopher J. Chovit, Manuel Solis, Martin R. Olah, and Orlesa Williams. Imaging spectroscopy and the Airborne Visible/Infrared Imaging Spectrometer (AVIRIS). *Remote Sensing of Environment*, 65(3):227–248, 1998.

[5] Amin Beiranvand Pour and Mazlan Hashim. ASTER, ALI and Hyperion sensors data for lithological mapping and ore minerals exploration. *SpringerPlus*, 3(1):130, 2014.

[6] A. Plaza, J. Plaza, A. Paz, and S. Sanchez. Parallel Hyperspectral Image and Signal Processing. *IEEE Signal Processing Magazine*, 28(3):119–126, 2011.

[7] H Kaufmann, L Guanter, K Segl, S Hofer, K.-P Foerster, T Stuffler, A Mueller, R Richter, H Bach, and P Hostert. Environmental Mapping and Analysis Program (EnMAP) – Recent Advances and Status. *IEEE International Geoscience & Remote Sensing Symposium, IGARSS*, 4:109–112, 2008.

[8] C. Galeazzi, A. Sacchetti, A. Cisbani, and G. Babini. The PRISMA Program. In *IGARSS 2008 - 2008 IEEE International Geoscience and Remote Sensing Symposium*, pages IV – 105–IV – 108, 2008.

[9] B. Kunkel, F. Blechinger, R. Lutz, R. Doerffer, and H. van der Piepen. ROSIS (Reflective Optics System Imaging Spectrometer) - A candidate instrument for polar platform missions. In J. Seeley and S. Bowyer, editors, *Optoelectronic technologies for remote sensing from space*, pages 134–141, 1988.

[10] Mustafa Teke, Hüsne Seda Deveci, Onur Haliloğlu, Sevgi Zübeyde Gürbüz, and Ufuk Sakarya. A Short Survey of Hyperspectral Remote Sensing Applications in Agriculture. In *Recent Advances in Space Technologies (RAST)*, 2013.

[11] Antonio Plaza, Javier Plaza, and David Valencia. Impact of platform heterogeneity on the design of parallel algorithms for morphological processing of high-dimensional image data. *Journal of Supercomputing*, 40(1):81–107, 2007.

 ISBN: 978-84-617-8694-7

[12] Antonio Plaza, Jon Atli Benediktsson, Joseph W Boardman, Jason Brazile, Lorenzo Bruzzone, Gustavo Camps-Valls, Jocelyn Chanussot, Mathieu Fauvel, Paolo Gamba, Anthony Gualtieri, Mattia Marconcini, James C Tilton, and Giovanna Trianni. Recent advances in techniques for hyperspectral image processing. *Remote Sensing of Environment*, 113(1):S110–S122, 2009.

[13] Javier Setoain, Manuel Prieto, Christian Tenllado, and Francisco Tirado. GPU for Parallel On-Board Hyperspectral Image Processing. *International Journal of High Performance Computing Applications*, 2008.

[14] Carlos González, Sergio Sánchez, Abel Paz, Javier Resano, Daniel Mozos, and Antonio Plaza. Use of FPGA or GPU-based architectures for remotely sensed hyperspectral image processing. *Integration, the VLSI Journal*, 46(2):89 – 103, 2013.

[15] Antonio J Plaza, Chein-I Chang, Liping Di, and Yuqi Bai. *High Performance Computing in Remote Sensing Book Review Book Review*. Chapman & Hall/CRC Press, Computer & Information Science Series, Boca Raton, Florida, 2008.

[16] P. Ghamisi, J. Plaza, Y. Chen, J. Li, and A. Plaza. Advanced Supervised Spectral Classifiers for Hyperspectral Images: A Review. *IEEE Geoscience and Remote Sensing Magazine*, 5(1):8–32, 2017.

[17] Balaji Krishnapuram, Lawrence Carin, Mário A T Figueiredo, and Alexander J. Hartemink. Sparse multinomial logistic regression: Fast algorithms and generalization bounds. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(6):957–968, 2005.

[18] Zebin Wu, Yonglong Li, Antonio Plaza, Jun Li, Fu Xiao, and Zhihui Wei. Parallel and Distributed Dimensionality Reduction of Hyperspectral Data on Cloud Computing Architectures. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 9(6):2270–2278, 2016.

[19] J. A. Martínez, E. M. Garzón, A. Plaza, and I. García. Automatic tuning of iterative computation on heterogeneous multiprocessors with ADITHE. *Journal of Supercomputing*, 58(2):151–159, 2011.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

General one-sided Clifford Fourier transform, convolution products in the spatial and frequency domains, and auto-correlation theorems

Eckhard Hitzer[1]

[1] *College of Liberal Arts, International Christian University, Mitaka, Tokyo, Japan*

emails: hitzer@icu.ac.jp

Abstract

In this paper we use the general steerable one-sided Clifford Fourier transform (CFT), and relate the classical convolution of Clifford algebra-valued signals over $\mathbb{R}^{p,q}$ with the (equally steerable) Mustard convolution. A Mustard convolution can be expressed in the spectral domain as the point wise product of the CFTs of the factor functions. In full generality do we express the classical convolution of Clifford algebra signals in terms of a linear combination of Mustard convolutions, and vice versa the Mustard convolution of Clifford algebra signals in terms of a linear combination of classical convolutions. Finally, we derive auto-correlation theorems for the cross-correlation and for the auto-correlation of Clifford signals.

Key words: Convolution, Clifford Fourier transform, Clifford algebra signals, spatial domain, frequency domain, auto-correlation theorem
MSC 2000: 44A35, 11E88, 15A66, 43A32, 30G35

1 Introduction

The steerable one-sided Clifford Fourier transformation (CFT) was introduced in [10]. It generalizes related transforms, like the classical complex Fourier transform, the one-sided single kernel quaternion Fourier transform [5], and the Clifford Fourier transforms with pseudoscalar kernels [4, 6] to higher dimensions. These CFTs essentially replace the imaginary unit $i \in \mathbb{C}$ by a general multivector square root of -1, which usually populate continuous Clifford algebra submanifolds [9, 11]. The classical complex Fourier transform needs only one fully commuting kernel factor, due to the commutativity of complex numbers. To have

a non-commutative kernel factor under the transform integral on one side of the signal function is meaningful due to the inherent non-commutativity in Clifford algebras. An extensive discussion of the historical development and the application relevance of the CFTs can be found in [1] and [14].

This paper is organized as follows. Section 2 gives some background on the steerable one-sided CFT. Next, Section 3 defines the classical convolution of two Clifford signal functions, their steerable Mustard convolution, their cross-correlation, and the auto-correlation of a Clifford signal function. The rest of the section is devoted to representing the classical convolution in terms of a sum of Mustard convolutions (Theorem 3.3) and dually to expressing the Mustard convolution in terms of a sum of classical convolutions (Theorem 3.4). Furthermore, direct single convolution product identities between classical and Mustard convolutions are established (Theorem 3.6), together with the theoretical equivalence (for general Clifford signal convolution product factor functions) of expressing the classical convolution in terms of the Mustard convolution and the reverse (Equation (3.17)). Finally, Section 4 derives auto-correlation theorems for the cross-correlation and for the auto-correlation of Clifford signals.

2 General steerable one-sided Clifford Fourier transforms

We will make use of the following notation.

Notation 2.1 (Argument reflection)**.** *For a function $h : \mathbb{R}^{p,q} \to Cl(p',q')$ we set*[1]

$$h^1(\boldsymbol{x}) := h(-\boldsymbol{x}). \tag{2.1}$$

Note that we obviously have

$$(h^1)^1(\boldsymbol{x}) = h^1(-\boldsymbol{x}) = h(\boldsymbol{x}). \tag{2.2}$$

We will make use of the following lemma [11].

Lemma 2.2. *Every multivector $A \in Cl(p,q)$ has, with respect to a square root $f \in Cl(p,q)$ of -1, i.e., $f^{-1} = -f$, the unique decomposition*

$$\begin{aligned} A_{+f} = \frac{1}{2}(A + f^{-1}Af), \qquad A_{-f} = \frac{1}{2}(A - f^{-1}Af) \\ A = A_{+f} + A_{-f}, \qquad A_{+f}\, f = f A_{+f}, \qquad A_{-f}\, f = -f A_{-f}, \end{aligned} \tag{2.3}$$

$A_{+f} \in \text{centralizer}(f, Cl_{p,q})$.

[1] We are aware that this notation could be confused with an ordinary taking to the power of 1, but as will be seen in the current context no danger of confusion is likely to arise.

 ISBN: 978-84-617-8694-7

The *general steerable one-sided Clifford Fourier transform* (CFT) [10], can be understood as a generalization of previously known one-sided CFTs [6], to a general Clifford algebra setting. Most known CFTs (prior to [10]) used in their kernels specific square roots of -1, like bivectors, pseudoscalars, unit pure quaternions, or sets of coorthogonal blades (commuting or anticommuting blades) [2]. All those restrictions on the square roots of -1 used in a CFT do not apply in our definition below. Note further, that the definition we are about to introduce is even more general than Definition 3.1 given in [10], because we generalize to multivector signal functions in $L^1(\mathbb{R}^{p,q};Cl(p',q'))$ and not only in $L^1(\mathbb{R}^{p,q};Cl(p,q))$.

Definition 2.3 (Steerable CFT with respect to one square root of -1)**.** Let $i \in Cl(p',q')$, $i^2=-1$, be any square root of -1. The general Clifford Fourier transform (CFT) of $f \in L^1(\mathbb{R}^{p,q};Cl(p',q'))$, with respect to i is

$$\mathcal{F}^i\{f\}(\boldsymbol{\omega}) = \int_{\mathbb{R}^{p,q}} f(\boldsymbol{x})\, e^{-iu(\boldsymbol{x},\boldsymbol{\omega})} d^n\boldsymbol{x}, \tag{2.4}$$

where $d^n\boldsymbol{x} = dx_1 \dots dx_n$, $\boldsymbol{x},\boldsymbol{\omega} \in \mathbb{R}^{p,q}$, and $u : \mathbb{R}^{p,q} \times \mathbb{R}^{p,q} \to \mathbb{R}$.

Since square roots of -1 in $Cl(p',q')$ populate *continuous submanifolds* in $Cl(p',q')$, the CFT of Definition 2.3 is generically *steerable* within these manifolds, see (2.6). In Definition 2.3, the square roots $i \in Cl(p',q')$ of -1 may be from any component of any conjugacy class. The choice of the Clifford's geometric product between multivector signal function f and the multivector kernel $e^{-iu(\boldsymbol{x},\boldsymbol{\omega})}$, in the integrand of (2.4) is very important. Because only this choice allowed, e.g. in [4], to define and apply a holistic vector field convolution, without loss of information.

Note that two-sided CFTs can be decomposed to pairs of one-sided CFTs [12].

Remark 2.4. *In order to avoid clutter we often drop the upper index i as in $\mathcal{F}\{h\} = \mathcal{F}^i\{h\}$, but in principle the one-sided CFT always depends on the particular choice i of the multivector square root of -1. Since square roots of -1 in $Cl(p',q')$ populate continuous submanifolds in $Cl(p',q')$, the CFT of Definition 2.3 is generically steerable within these submanifolds. In Definition 2.3, the square root $i \in Cl(p',q')$ of -1, may be from any conjugacy class and component, respectively.*

Within the same conjugacy class of square roots of -1 the CFTs of Definition 2.3 are related by the following equation, and therefore steerable. Let $i, i' \in Cl(p',q')$ be any two square roots of -1 in the same conjugacy class, i.e. $i' = a^{-1}ia$, $a \in Cl(p',q')$, a being invertible. As a consequence of this relationship we also have

$$e^{-i'u} = a^{-1}e^{-iu}a, \ \ \forall u \in \mathbb{R}. \tag{2.5}$$

This in turn leads to the following *steerability relationship* of all CFTs with square roots of -1 from the same conjugacy class:

$$\mathcal{F}^{i'}\{h\}(\boldsymbol{\omega}) = \mathcal{F}^i\{ha^{-1}\}(\boldsymbol{\omega})a, \tag{2.6}$$

 ISBN: 978-84-617-8694-7

where ha^{-1} means to multiply the signal function h by the constant multivector $a^{-1} \in Cl(p',q')$.

For establishing an *inversion* formula and other properties of the CFT in Definition 2.3, certain *assumptions* about the phase function $u(\boldsymbol{x},\boldsymbol{\omega})$ need to be made. In principle these assumptions could be made based on the desired properties of the resulting CFT. One possibility is, e.g., to assume

$$u(\boldsymbol{x},\boldsymbol{\omega}) = \boldsymbol{x} * \widetilde{\boldsymbol{\omega}} = \sum_{l=1}^{n} x^l \omega^l = \sum_{l=1}^{n} x_l \omega_l, \tag{2.7}$$

which will be assumed in the rest of this paper.

We then get the following *inversion* theorem[2].

Theorem 2.5 (Inversion of one-sided CFT)**.** *For $\mathcal{F}^i\{h\} \in L^1\left(\mathbb{R}^{p,q}; Cl(p',q')\right)$, provided that in any finite interval h and the partial coordinate derivatives of h are piecewise continuous, and have at most a finite number of extrema and discontinuities, and that h is continuous at $\boldsymbol{x} \in \mathbb{R}^{p,q}$, we have*

$$h(\boldsymbol{x}) = \mathcal{F}^i_{-1}\{\mathcal{F}^i\{h\}\}(\boldsymbol{x}) = \frac{1}{(2\pi)^n} \int_{\mathbb{R}^{p,q}} \mathcal{F}^i\{h\}(\boldsymbol{\omega})\, e^{iu(\boldsymbol{x},\boldsymbol{\omega})} d^n\boldsymbol{\omega}, \tag{2.8}$$

where $d^n\boldsymbol{\omega} = d\omega_1 \dots d\omega_n$, $\boldsymbol{x},\boldsymbol{\omega} \in \mathbb{R}^{p,q}$.

The proof of theorem 2.5 is strictly analogous to the proof of equation (4.8) on page 231 of [10], and therefore left as an exercise to the reader.

We further note the following useful relationship using the argument reflection of Notation 2.1

$$\mathcal{F}^{-i}\{h\} = \mathcal{F}^i\{h^1\} = \mathcal{F}\{h^1\}. \tag{2.9}$$

The main properties of the CFT of Definition 2.3 have been studied for the special case of multivector signal functions $f \in L^1(\mathbb{R}^{p,q}; Cl(p,q))$ in detail in [10], and can easily be generalized to the more general case of $f \in L^1(\mathbb{R}^{p,q}; Cl(p',q'))$.

3 Convolution and steerable Mustard convolution, cross-correlation and auto-correlation

We define the *convolution* of two Clifford (algebra) signals $a, b \in L^1(\mathbb{R}^{p,q};\, Cl(p',q'))$ as

$$(a \star b)(\boldsymbol{x}) = \int_{\mathbb{R}^{p,q}} a(\boldsymbol{y}) b(\boldsymbol{x}-\boldsymbol{y}) d^n\boldsymbol{y}, \tag{3.1}$$

[2]Note, that we show the inversion symbol -1 as lower index in $\mathcal{F}^i_{-1}$, in order to avoid a possible confusion by using two upper indice. The inversion could also be written with the help of the CFT itself as $\mathcal{F}^i_{-1} = \frac{1}{(2\pi)^n}\mathcal{F}^{-i}$.

 ISBN: 978-84-617-8694-7

provided that the integral exists.

Note that the real continuous Clifford geometric algebra wavelet transform can be written as a convolution of the multivector signal function with the daughter wavelet (a rotated, dilated and translated mother wavelet), essentially evaluated at the center of the daughter wavelet, see [7].

The *Mustard* convolution [15, 3] of two Clifford signals $a, b \in L^1(\mathbb{R}^{p,q};\, Cl(p',q'))$ is defined as

$$(a \star_M b)(\boldsymbol{x}) = (\mathcal{F}^i)^{-1}(\mathcal{F}^i\{a\}\mathcal{F}^i\{b\})(\boldsymbol{x}), \tag{3.2}$$

provided that the integral exists.

Remark 3.1. *The Mustard convolution has the conceptual and computational advantage to simply yield, independent of the particular Clifford algebra $Cl(p',q')$ involved and of the particular multivector square root of -1 in the CFT kernel, as spectrum in the CFT Fourier domain the point wise product of the CFTs of the two signals, just as for the classical complex Fourier transform. On the other hand, by its very definition, the Mustard convolution itself depends on the choice of i, i.e. of the multivector square root of -1, used in the Definition 2.3 of the CFT. The Mustard convolution* (3.2) *is therefore a steerable operator, dependent on the choice of i.*

We further define the *cross-correlation* of two Clifford signals $a, b \in L^1(\mathbb{R}^{p,q};\, Cl(p',q'))$ as

$$(a \star_c b)(\boldsymbol{x}) = \int_{\mathbb{R}^{p,q}} a(\boldsymbol{y})\widetilde{b(\boldsymbol{y}-\boldsymbol{x})}d^n\boldsymbol{y} = \int_{\mathbb{R}^{p,q}} a(\boldsymbol{y}+\boldsymbol{x})\widetilde{b(\boldsymbol{y})}d^n\boldsymbol{y}, \tag{3.3}$$

provided that the integral exists.

We finally define the *auto-correlation* of a Clifford signal $a \in L^1(\mathbb{R}^{p,q};\, Cl(p',q'))$ as

$$(a \star_a a)(\boldsymbol{x}) = \int_{\mathbb{R}^{p,q}} a(\boldsymbol{y})\widetilde{a(\boldsymbol{y}-\boldsymbol{x})}d^n\boldsymbol{y} = \int_{\mathbb{R}^{p,q}} a(\boldsymbol{y}+\boldsymbol{x})\widetilde{a(\boldsymbol{y})}d^n\boldsymbol{y}, \tag{3.4}$$

provided that the integral exists.

In the following two Subsections we will express the convolution (3.1) in terms of the Mustard convolution (3.2), and vice versa, and study the mutual relations of these expressions. Then we will give auto-correlation theorems for the cross- and auto-correlations of Clifford signals.

3.1 Expressing the convolution in terms of the Mustard convolution

In this Subsection we assume the use of the one-sided CFT with a general multivector square roots of -1, $i \in Cl(p',q')$. The definition of the classical convolution (3.1) is independent of the application of a CFT. The Mustard convolution of (3.2) depends on the definition of the CFT and in particular on the choice of the multivector square root i of -1.

 ISBN: 978-84-617-8694-7

In our approach we generalize equation (4.17) on page 233 of [10], which expresses the convolution of two Clifford signal functions in the Clifford Fourier domain with the help of the CFT of Definition 2.3. We generalize this equation to the case of multivector signal functions $a, b \in L^1(\mathbb{R}^{p,q}; Cl(p',q')$, and to the CFT of Definition 2.3. Nevertheless the proof works perfectly analogous to the one given in [10], we therefore leave this as an exercise to the reader.

Theorem 3.2 (CFT of convolution)**.** *We assume that the function u is linear with respect to its first argument. The CFT of the convolution* (3.1) *of two multivector signals $a, b \in L^1(\mathbb{R}^{p,q}; Cl(p',q')$ can then be expressed as*

$$\begin{aligned}\mathcal{F}^i\{a \star b\} &= \mathcal{F}^{-i}\{a\}\mathcal{F}^i\{b_{-i}\} + \mathcal{F}^i\{a\}\mathcal{F}^i\{b_{+i}\} \\ &= \mathcal{F}^i\{a^1\}\mathcal{F}^i\{b_{-i}\} + \mathcal{F}^i\{a\}\mathcal{F}^i\{b_{+i}\}. \end{aligned} \tag{3.5}$$

We can now easily express the convolution of two multivector signals $\mathcal{F}^i\{a \star b\}(\boldsymbol{\omega})$ in terms of only two Mustard convolutions (3.2), by applying the inverse CFT.

Theorem 3.3 (Convolution in terms of Mustard convolution)**.** *Assuming a general multivector square root i of -1, the convolution* (3.1) *of two Clifford functions $a, b \in L^1(\mathbb{R}^{p,q}; Cl(p',q'))$ can be expressed in terms of two Mustard convolutions* (3.2) *as*

$$a \star b = a^1 \star_M b_{-i} + a \star_M b_{+i}. \tag{3.6}$$

An alternative direct proof of Theorem 3.3 is the following:

$$\begin{aligned}(a \star b)(\boldsymbol{x}) &= \int_{\mathbb{R}^{p,q}} a(\boldsymbol{y}) b(\boldsymbol{x}-\boldsymbol{y}) d^n\boldsymbol{y} \\ &= \int_{\mathbb{R}^{p,q}} a(\boldsymbol{y}) \tfrac{1}{(2\pi)^n} \int_{\mathbb{R}^{p,q}} \mathcal{F}\{b\}(\boldsymbol{\omega}) e^{iu(\boldsymbol{x}-\boldsymbol{y},\boldsymbol{\omega})} d^n\boldsymbol{\omega} d^n\boldsymbol{y} \\ &= \tfrac{1}{(2\pi)^n} \int_{\mathbb{R}^{p,q}} a(\boldsymbol{y}) \int_{\mathbb{R}^{p,q}} [\mathcal{F}\{b_+\}(\boldsymbol{\omega}) + \mathcal{F}\{b_-\}(\boldsymbol{\omega})] e^{-iu(\boldsymbol{y},\boldsymbol{\omega})} e^{iu(\boldsymbol{x},\boldsymbol{\omega})} d^n\boldsymbol{\omega} d^n\boldsymbol{y} \\ &= \tfrac{1}{(2\pi)^n} \int_{\mathbb{R}^{p,q}} \int_{\mathbb{R}^{p,q}} [a(\boldsymbol{y}) e^{-iu(\boldsymbol{y},\boldsymbol{\omega})} d^n\boldsymbol{y} \mathcal{F}\{b_+\}(\boldsymbol{\omega}) \\ &\qquad + a(\boldsymbol{y}) e^{iu(\boldsymbol{y},\boldsymbol{\omega})} d^n\boldsymbol{y} \mathcal{F}\{b_-\}(\boldsymbol{\omega})] e^{iu(\boldsymbol{x},\boldsymbol{\omega})} d^n\boldsymbol{\omega} \\ &= \tfrac{1}{(2\pi)^n} \int_{\mathbb{R}^{p,q}} [\mathcal{F}\{a\}(\boldsymbol{\omega})\mathcal{F}\{b_+\}(\boldsymbol{\omega}) + \mathcal{F}\{a\}(-\boldsymbol{\omega})\mathcal{F}\{b_-\}(\boldsymbol{\omega})] e^{iu(\boldsymbol{x},\boldsymbol{\omega})} d^n\boldsymbol{\omega} \\ &= \tfrac{1}{(2\pi)^n} \int_{\mathbb{R}^{p,q}} [\mathcal{F}\{a\}(\boldsymbol{\omega})\mathcal{F}\{b_+\}(\boldsymbol{\omega}) + \mathcal{F}\{a^1\}(\boldsymbol{\omega})\mathcal{F}\{b_-\}(\boldsymbol{\omega})] e^{iu(\boldsymbol{x},\boldsymbol{\omega})} d^n\boldsymbol{\omega} \\ &= a \star_M b_{+i} + a^1 \star_M b_{-i}, \end{aligned} \tag{3.7}$$

where we have applied the inverse CFT by substituting for the second equality $b = \mathcal{F}^i\mathcal{F}^i_{-1}\{b\}$, and that for $\alpha \in \mathbb{R}$, $\mathcal{F}^i\{b_\pm\}e^{i\alpha} = e^{\pm i\alpha}\mathcal{F}^i\{b_\pm\}$ for the fourth equality.

 ISBN: 978-84-617-8694-7

3.2 Expressing the Mustard convolution in terms of the convolution

Now we will first simply write out the Mustard convolution (3.2) and simplify it until only standard convolutions (3.1) remain.

We begin by writing the Mustard convolution (3.2) of two multivector functions $a, b \in L^2(\mathbb{R}^{p,q}; Cl(p',q'))$

$$
\begin{aligned}
& a \star_M b(\boldsymbol{x}) \\
&= \tfrac{1}{(2\pi)^n} \int_{\mathbb{R}^{p,q}} \mathcal{F}\{a\}(\boldsymbol{\omega}) \mathcal{F}\{b\}(\boldsymbol{\omega}) e^{iu(\boldsymbol{x},\boldsymbol{\omega})} d^n\boldsymbol{\omega} \\
&= \tfrac{1}{(2\pi)^n} \int_{\mathbb{R}^{p,q}} \int_{\mathbb{R}^{p,q}} a(\boldsymbol{y}) e^{-iu(\boldsymbol{y},\boldsymbol{\omega})} d^n\boldsymbol{y} \int_{\mathbb{R}^{p,q}} b(\boldsymbol{z}) e^{-iu(\boldsymbol{z},\boldsymbol{\omega})} d^n\boldsymbol{z} e^{iu(\boldsymbol{x},\boldsymbol{\omega})} d^n\boldsymbol{\omega} \\
&= \tfrac{1}{(2\pi)^n} \iiint a(\boldsymbol{y}) e^{-iu(\boldsymbol{y},\boldsymbol{\omega})} [b_{+i}(\boldsymbol{z}) + b_{-i}(\boldsymbol{z})] e^{iu(\boldsymbol{x}-\boldsymbol{z},\boldsymbol{\omega})} d^n\boldsymbol{y} d^n\boldsymbol{z} d^n\boldsymbol{\omega} \\
&= \tfrac{1}{(2\pi)^n} \iiint a(\boldsymbol{y}) b_{+i}(\boldsymbol{z}) e^{-iu(\boldsymbol{y},\boldsymbol{\omega})} e^{iu(\boldsymbol{x}-\boldsymbol{z},\boldsymbol{\omega})} d^n\boldsymbol{y} d^n\boldsymbol{z} d^n\boldsymbol{\omega} \\
&\quad + \tfrac{1}{(2\pi)^n} \iiint a(\boldsymbol{y}) b_{-i}(\boldsymbol{z}) e^{iu(\boldsymbol{y},\boldsymbol{\omega})} e^{iu(\boldsymbol{x}-\boldsymbol{z},\boldsymbol{\omega})} d^n\boldsymbol{y} d^n\boldsymbol{z} d^n\boldsymbol{\omega} \\
&= \tfrac{1}{(2\pi)^n} \iiint a(\boldsymbol{y}) b_{+i}(\boldsymbol{z}) e^{iu(\boldsymbol{x}-\boldsymbol{y}-\boldsymbol{z},\boldsymbol{\omega})} d^n\boldsymbol{y} d^n\boldsymbol{z} d^n\boldsymbol{\omega} \\
&\quad + \tfrac{1}{(2\pi)^n} \iiint a(\boldsymbol{y}) b_{-i}(\boldsymbol{z}) e^{iu(\boldsymbol{x}+\boldsymbol{y}-\boldsymbol{z},\boldsymbol{\omega})} d^n\boldsymbol{y} d^n\boldsymbol{z} d^n\boldsymbol{\omega} \\
&= \iint a(\boldsymbol{y}) b_{+i}(\boldsymbol{z}) \delta(\boldsymbol{x}-\boldsymbol{y}-\boldsymbol{z}) d^n\boldsymbol{y} d^n\boldsymbol{z} \\
&\quad + \iint a(\boldsymbol{y}) b_{-i}(\boldsymbol{z}) \delta(\boldsymbol{x}+\boldsymbol{y}-\boldsymbol{z}) d^n\boldsymbol{y} d^n\boldsymbol{z} \\
&= \int_{\mathbb{R}^{p,q}} a(\boldsymbol{y}) b_{+i}(\boldsymbol{x}-\boldsymbol{y}) d^n\boldsymbol{y} + \int_{\mathbb{R}^{p,q}} a(\boldsymbol{y}) b_{-i}(\boldsymbol{x}+\boldsymbol{y}) d^n\boldsymbol{y} \\
&= \int_{\mathbb{R}^{p,q}} a(\boldsymbol{y}) b_{+i}(\boldsymbol{x}-\boldsymbol{y}) d^n\boldsymbol{y} + \int_{\mathbb{R}^{p,q}} a(\boldsymbol{y}) b_{-i}(-(-\boldsymbol{x}-\boldsymbol{y})) d^n\boldsymbol{y} \\
&= a \star b_{+i}(\boldsymbol{x}) + a \star b^1_{-i}(-\boldsymbol{x}) \\
&= a \star b_{+i}(\boldsymbol{x}) + a^1 \star b_{-i}(\boldsymbol{x}). \qquad (3.8)
\end{aligned}
$$

We have abbreviated $\int_{\mathbb{R}^{p,q}} \int_{\mathbb{R}^{p,q}}$ to $\iint$, and $\int_{\mathbb{R}^{p,q}} \int_{\mathbb{R}^{p,q}} \int_{\mathbb{R}^{p,q}}$ to $\iiint$. For the third equality we applied the split of Lemma 2.2 to $b(\boldsymbol{x})$ and used the linearity of u with respect to its first argument. For the fourth equality we used the linearity of Clifford's geometric product, the linearity of the triple integral, and we used the commutation and anti-commutation properties of $b_{\pm i}(\boldsymbol{x})$ with the multivector square root $i \in Cl(p',q')$, which produces the sign change $e^{-iu(\boldsymbol{y},\boldsymbol{\omega})} \to e^{+iu(\boldsymbol{y},\boldsymbol{\omega})}$ in the case of anti-commutation. For the fifth equality we again applied the linearity of u with respect to its first argument. The integrations

 ISBN: 978-84-617-8694-7

$\frac{1}{(2\pi)^n}\int e^{iu(\boldsymbol{x}\pm\boldsymbol{y}-\boldsymbol{z},\boldsymbol{\omega})}d^n\boldsymbol{\omega}$ produce the n-dimensional Dirac delta functions $\delta(\boldsymbol{x}\pm\boldsymbol{y}-\boldsymbol{z})$, giving the sixth equality.
We illustrate the last identity of (3.8), $a\star b^1_{-i}(-\boldsymbol{x}) = a^1\star b_{-i}(\boldsymbol{x})$, in the one-dimensional case $\mathbb{R}^{p,q}=\mathbb{R}$, the generalization to $\mathbb{R}^{p,q}$ is then straightforward

$$\begin{aligned} a\star b^1(-x) &= \int_{\mathbb{R}} a(y)b(-(-x-y))dy = \int_{-\infty}^{+\infty} a(y)b(x+y)dy \\ &= \int_{+\infty}^{-\infty} a(-g)b(x-g)(-1)dg = \int_{-\infty}^{+\infty} a(-g)b(x-g)dg \\ &= \int_{\mathbb{R}} a^1(g)b(x-g)dg = a^1\star b(x). \end{aligned} \tag{3.9}$$

where we have substituted $g=-y$, $dg=-dy$, including substitution of the integration boundaries for the third equality. The interchange of the integration boundaries eliminates the overall minus sign in the fourth equality of (3.9).

Note that in (3.8), $a\star b^1_{-i}(-\boldsymbol{x})$, means to first apply the convolution to the pair of functions a and b^1_{-i}, and only then to evaluate the result of the convolution integral with the argument $(-\boldsymbol{x})$. So in general $a\star b^1_{-i}(-\boldsymbol{x})\neq a\star b_{-i}(+\boldsymbol{x})$.

We finally obtain the desired decomposition of the Mustard convolution (3.2) in terms of the classical convolution.

Theorem 3.4 (Mustard convolution in terms of standard convolution)**.**
The Mustard convolution (3.2) *of two multivector signal functions* $a,b\in L^1(\mathbb{R}^{p,q};Cl(p',q'))$ *can be expressed in terms of two standard convolutions* (3.1) *as*

$$a\star_M b(\boldsymbol{x}) = a^1\star b_{-i}(\boldsymbol{x}) + a\star b_{+i}(\boldsymbol{x}). \tag{3.10}$$

Remark 3.5 (Theorem duality)**.** *Comparing Theorems 3.3 and 3.4 we notice an interesting duality: interchanging convolution and Mustard convolution in either theorem yields the other, independent over which vector space* $\mathbb{R}^{p,q}$ *the multivector signals are defined, independent from the signal value Clifford algebra* $Cl(p',q')$*, and independent from the particular choice of multivector square root of* -1*,* $i\in Cl(p',q')$*. The last form of independence also means, that the observed duality is stable with respect to steering the CFT and the Mustard convolution by changing* $i\in Cl(p',q')$*. Note further, that a corresponding duality will be valid for the left-sided version of the CFT in Definition 2.3, by placing the kernel factor on the left side and going analogously through all arguments up to Theorem 3.4.*

Yet, it is an interesting non-trivial question, whether a similar duality may hold for other forms of the CFT, e.g. with more than one kernel factor, see e.g. [12, 13].

 ISBN: 978-84-617-8694-7

3.3 Single convolution product identities for classical and Mustard convolutions

Let us now apply Theorem 3.3 to the three functions $a, b_{\pm i}$, observing that

$$(b_{+i})_{-i} = (b_{-i})_{+i} = 0, \quad (b_{+i})_{+i} = b_{+i}, \quad (b_{-i})_{-i} = b_{-i}. \tag{3.11}$$

Then we obtain

$$a \star b_{+i} = a^1 \star_M (b_{+i})_{-i} + a \star_M (b_{+i})_{+i} = 0 + a \star_M b_{+i} = a \star_M b_{+i}, \tag{3.12}$$

and similarly,

$$a \star b_{-i} = a^1 \star_M b_{-i} \quad \Longleftrightarrow \quad a^1 \star b_{-i} = a \star_M b_{-i}, \tag{3.13}$$

since double reflection of the argument returns the function itself (2.2). Note, that the very same identities are easily obtained by analogously applying Theorem 3.4 to $a, b_{\pm i}$. We therefore summarize them in the following theorem.

Theorem 3.6 (Partial identities between convolutions and Mustard convolutions)**.** *For pairs of functions (a, b_{-i}) and (a, b_{+i}) with $a, b_{\pm i} \in L^1(\mathbb{R}^{p,q}; Cl(p', q'))$, where the second factor either commutes or anti-commutes with the multivector square root of -1, $i \in Cl(p', q')$ of the Definition 2.3, the following convolution product identities between convolution* (3.1) *and Mustard convolution* (3.2) *hold*

$$\begin{gathered} a^1 \star b_{-i} = a \star_M b_{-i} \quad \Longleftrightarrow \quad a \star b_{-i} = a^1 \star_M b_{-i}, \\ a \star b_{+i} = a \star_M b_{+i}. \end{gathered} \tag{3.14}$$

Theorem 3.6 can therefore either be derived from Theorem 3.3 or from Theorem 3.4. Moreover, Theorem 3.6 can also be established independently by *direct* computation. Then adding two convolution terms would give the Mustard convolution

$$a^1 \star b_{-i} + a \star b_{+i} \overset{\text{Th. } 3.6}{=} a \star_M b_{-i} + a \star_M b_{+i} = a \star_M b. \tag{3.15}$$

And conversely adding two Mustard convolution terms would give the convolution

$$a^1 \star_M b_{-i} + a \star_M b_{+i} \overset{\text{Th. } 3.6}{=} a \star b_{-i} + a \star b_{+i} = a \star b. \tag{3.16}$$

This establishes the following important *threefold theorem equivalence*

$$\text{Theorem 3.3} \quad \Longleftrightarrow \quad \text{Theorem 3.6} \quad \Longleftrightarrow \quad \text{Theorem 3.4}. \tag{3.17}$$

Remark 3.7. *Note that the need to always decompose the right convolution product factor function $b = b_{-i} + b_{+i}$ is manifestly due to the kernel in Definition 2.3 being placed on the right side. Using a corresponding left side kernel CFT, would lead to analogous results with decomposing the left convolution product factor $a = a_{-i} + a_{+i}$.*

 ISBN: 978-84-617-8694-7

Furthermore, we can ask under what conditions we get a *full direct single convolution product identity* of the two convolution products $a \star b = a \star_M b$? This identity holds under any of the following conditions:

1. For all functions $a, b \in L^1(\mathbb{R}^{p,q}; Cl(p',q'))$, with $b_{-i} \equiv 0$. This condition depends on the choice of i.
2. For central multivector square roots $i \in Cl(p',q')$ of -1 and *all* functions $a, b \in L^1(\mathbb{R}^{p,q}; Cl(p',q'))$. An important practical example is $i = e_1e_2e_3 \in Cl(3,0)$ [4].
3. For all functions $a, b \in L^1(\mathbb{R}^{p,q}; Cl(p',q'))$, with reflection symmetry $a^1 = a$. This condition does not depend on the choice of i, and poses no restriction on b.

4 Auto-correlation theorem for Clifford signals

Using the classical complex Fourier transform, the auto-correlation of a signal can be expressed as the inverse Fourier transform of the spectral density. It provides information on how closely related a signal is to the same signal at some other time (or location). For unrelated noise, the auto-correlation function will be nearly zero, but if there is temporal (or spatial) structure in the signal, the auto-correlation will provide interesting information about it, see [16], Chapter 15.7.

In the following we assume $b \in L^1(\mathbb{R}^{p,q}; Cl(p',q'))$, that in any finite interval b and the partial coordinate derivatives of b are piecewise continuous, and have at most a finite number of extrema and discontinuities, that b is continuous at $\boldsymbol{x} \in \mathbb{R}^{p,q}$, that $\mathcal{F}\{b\} \in L^1(\mathbb{R}^{p,q}; Cl(p',q'))$, and that $\widetilde{i} = -i$. For establishing the auto-correlation theorem for Clifford signals we begin with the cross correlation (3.3) and apply the inverse CFT of Theorem 2.5.

$$\begin{aligned}
(a \star_c b)(\boldsymbol{x}) &= \int_{\mathbb{R}^{p,q}} a(\boldsymbol{y}) \widetilde{b(\boldsymbol{y}-\boldsymbol{x})} d^n\boldsymbol{y} \\
&= \tfrac{1}{(2\pi)^n} \int_{\mathbb{R}^{p,q}} a(\boldsymbol{y}) \int_{\mathbb{R}^{p,q}} [\mathcal{F}^i\{b\}(\boldsymbol{\omega}) e^{iu(\boldsymbol{y}-\boldsymbol{x},\boldsymbol{\omega})} d^n\boldsymbol{\omega}]^{\sim} d^n\boldsymbol{y} \\
&= \tfrac{1}{(2\pi)^n} \int_{\mathbb{R}^{p,q}} \int_{\mathbb{R}^{p,q}} a(\boldsymbol{y})\, e^{-iu(\boldsymbol{y}-\boldsymbol{x},\boldsymbol{\omega})} \widetilde{\mathcal{F}^i\{b\}}(\boldsymbol{\omega}) d^n\boldsymbol{\omega} d^n\boldsymbol{y} \\
&= \tfrac{1}{(2\pi)^n} \int_{\mathbb{R}^{p,q}} \int_{\mathbb{R}^{p,q}} a(\boldsymbol{y})\, e^{-iu(\boldsymbol{y},\boldsymbol{\omega})} d^n\boldsymbol{y}\, e^{iu(\boldsymbol{x},\boldsymbol{\omega})} \widetilde{\mathcal{F}^i\{b\}}(\boldsymbol{\omega})\, d^n\boldsymbol{\omega} \\
&= \tfrac{1}{(2\pi)^n} \int_{\mathbb{R}^{p,q}} \mathcal{F}^i\{a\}(\boldsymbol{\omega}) e^{iu(\boldsymbol{x},\boldsymbol{\omega})} \widetilde{\mathcal{F}^i\{b\}}(\boldsymbol{\omega})\, d^n\boldsymbol{\omega}
\end{aligned} \tag{4.1}$$

Setting $b = a$, we get an important identity for the auto-correlation

$$(a \star_a a)(\boldsymbol{x}) = \tfrac{1}{(2\pi)^n} \int_{\mathbb{R}^{p,q}} \mathcal{F}^i\{a\}(\boldsymbol{\omega}) e^{iu(\boldsymbol{x},\boldsymbol{\omega})} \widetilde{\mathcal{F}^i\{a\}}(\boldsymbol{\omega}) d^n\boldsymbol{\omega}. \tag{4.2}$$

 ISBN: 978-84-617-8694-7

Hence we obtain one theorem and one corollary.

Theorem 4.1 (Auto-correlation theorem for cross-correlation of Clifford signals)**.** *We assume two Clifford signals $a, b \in L^1(\mathbb{R}^{p,q}, Cl(p',q'))$, that in any finite interval b and the partial coordinate derivatives of b are piecewise continuous, and have at most a finite number of extrema and discontinuities, that b is continuous at $\boldsymbol{x} \in \mathbb{R}^{p,q}$, and that $\mathcal{F}\{b\} \in L^1(\mathbb{R}^{p,q}, Cl(p',q'))$. The cross-correlation of a, b can then be expressed as*

$$(a \star_c b)(\boldsymbol{x}) = \tfrac{1}{(2\pi)^n} \int_{\mathbb{R}^{p,q}} \mathcal{F}^i\{a\}(\boldsymbol{\omega}) e^{iu(\boldsymbol{x},\boldsymbol{\omega})} \widetilde{\mathcal{F}^i\{b\}}(\boldsymbol{\omega}) d^n\boldsymbol{\omega}. \tag{4.3}$$

Corollary 4.2 (Auto-correlation theorem for auto-correlation of Clifford signals)**.** *We assume a Clifford signal $a \in L^1(\mathbb{R}^{p,q}, Cl(p',q'))$, that in any finite interval a and the partial coordinate derivatives of a are piecewise continuous, and have at most a finite number of extrema and discontinuities, that a is continuous at $\boldsymbol{x} \in \mathbb{R}^{p,q}$, and that $\mathcal{F}\{a\} \in L^1(\mathbb{R}^{p,q}, Cl(p',q'))$. The auto-correlation of a can then be expressed as*

$$(a \star_a a)(\boldsymbol{x}) = \tfrac{1}{(2\pi)^n} \int_{\mathbb{R}^{p,q}} \mathcal{F}^i\{a\}(\boldsymbol{\omega}) e^{iu(\boldsymbol{x},\boldsymbol{\omega})} \widetilde{\mathcal{F}^i\{a\}}(\boldsymbol{\omega}) d^n\boldsymbol{\omega}. \tag{4.4}$$

Acknowledgements

Soli Deo Gloria. The author requests to apply the results of this paper only under the terms of the *Creative Peace License* [8].

References

[1] Brackx F, Hitzer E, Sangwine SJ. History of Quaternion and Clifford-Fourier Transforms. in: Hitzer E, Sangwine SJ (eds.). Quaternion and Clifford Fourier Transforms and Wavelets. Trends in Mathematics (TIM) **27**, Birkhäuser, Basel, 2013, pp. xi–xxvii.

[2] Bujack R, Scheuermann G, Hitzer E. A General Geometric Fourier Transform, in: Hitzer E, Sangwine SJ (eds.). Quaternion and Clifford Fourier Transforms and Wavelets. Trends in Mathematics (TIM) **27**, Birkhäuser, Basel, 2013, pp. 155–176.

[3] Bujack R, De Bie H, De Schepper N, Scheuermann G. *Convolution products for hypercomplex Fourier transforms*, J. Math. Imaging Vision 2014;48:606–624,

[4] Ebling J, Scheuermann G. Clifford Fourier transform on vector fields. IEEE Transactions on Visualization and Computer Graphics 2005;11(4):469–479, DOI: 10.1109/TVCG.2005.54

[5] Ell TA, Le Bihan N, Sangwine SJ. Quaternion Fourier Transforms for Signal and Image Processing. Digital Signal and Image Processing, Wiley-ISTE, Hoboken, 2014.

[6] Hitzer E, Mawardi B. Clifford Fourier Transform on Multivector Fields and Uncertainty Principles for Dimensions $n = 2 \pmod 4$ and $n = 3 \pmod 4$. AACA 2008;18(3–4):715–736.

[7] Hitzer E. Clifford (Geometric) Algebra Wavelet Transform. in V. Skala and D. Hildenbrand (eds.), Proc. of GraVisMa 2009, 02–04 Sep. 2009, Plzen, Czech Republic, pp. 94-101 (2009). Online: http://gravisma.zcu.cz/GraVisMa-2009/Papers_2009/!_2009_GraVisMa_proceedings-FINAL.pdf,

[8] Hitzer E. *Creative Peace License*. http://gaupdate.wordpress.com/2011/12/14/the-creative-peace-license-14-dec-2011/

[9] Hitzer E, Abłamowicz R. Geometric Roots of -1 in Clifford Algebras $Cl(p,q)$ with $p+q \leq 4$. AACA 2011;21(1):121–144, DOI: 10.1007/s00006-010-0240-x. Available as preprint: http://arxiv.org/abs/0905.3019 .

[10] Hitzer E. The Clifford Fourier transform in real Clifford algebras. in E. Hitzer, K. Tachibana (eds.), "Session on Geometric Algebra and Applications, IKM 2012", Special Issue of Clifford Analysis, Clifford Algebras and their Applications 2013;2(3):227–240.

[11] Hitzer E, Helmstetter J, Abłamowicz R. Square Roots of -1 in Real Clifford Algebras. in: Hitzer E, Sangwine SJ (Eds.), Quaternion and Clifford Fourier Transforms and Wavelets. Trends in Mathematics (TIM) **27**, Birkhäuser, 2013, pp. 123–153, DOI: 10.1007/978-3-0348-0603-9_7 .

[12] Hitzer E. Two-Sided Clifford Fourier Transform with Two Square Roots of -1 in $Cl(p,q)$. AACA 201424:313–332. DOI: 10.1007/s00006-014-0441-9 .

[13] Hitzer E. *General Steerable Two-Sided Clifford Fourier Transform, Convolution and Mustard Convolution*, AACA Online First 9 June 2016; 1–20. DOI: 10.1007/s00006-016-0687-5.

[14] Hitzer E. New Developments in Clifford Fourier Transforms. in N. E. Mastorakis et al (eds.), Advances in Applied and Pure Mathematics, Proceedings of the 2014 International Conference on Pure Mathematics, Applied Mathematics, Computational Methods (PMAMCM 2014), Santorini Island, Greece, July 17–21, 2014, Mathematics and Computers in Science and Engineering Series 2014;29:19–25.

[15] Mustard D. Fractional convolution. J. Aust. Math. Soc. Ser. B 1998;40:257–265.

[16] Nearing J. Mathematical Tools for Physics. Mineola, New York (US): Dover; 2010.

 ISBN: 978-84-617-8694-7

Proceedings of the 16th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

ELEMENTARY DISCRET HOLOMORPHIC FUNCTIONS

Angela Hommel[1]

[1] *Faculty of Economics, University of Applied Sciences Zwickau*

emails: angela.hommel@fh-zwickau.de

Abstract

Based on finite differences the discrete Laplacian can be factorized. By using one of these factors it is possible to study discrete holomorphic functions similar to the continuous case. In this paper discrete polynomials as well as an exponential function and a sine- and cosine function are defined such that all of them have the property to be discrete holomorphic. All investigations start in the one-dimensional case and extend the information to the complex case. An extension to the hypercomplex case seems to be possible but is not practiced here.

Key words: Finite Differences, Discrete Holomorphic Functions, Polynomials.

1 Introduction

Using finite differences and the factorization of the Laplacian, discrete holomorphic functions can be studied which are discrete harmonic, too. The paper starts with discrete polynomials which fulfill the so-called Appell property. Especially in the one-dimensional case such polynomials were already studied by Faustino und Kähler in [2]. The authors extend their polynomials to multi-index polynomials in order to obtain a Fischer decomposition with respect to the discrete Dirac operator and they use Euler- and Gamma operators in unbounded domains. The aim of this article is the extension of the polynomials to the complex case and the study of their properties. Furthermore the exponential function and the cosine and sine function are investigated. Starting again with the one-dimensional case it is possible to use some information to develop discrete holomorphic functions in the complex case.

 ISBN: 978-84-608-6082-2

2 Discrete polynomials and the Appell property

In the following section the Appell property is important so that the complex derivation of a basic function leads again to a multiple of a basic function (see [1]). In detail a system of polynomials $\{P^n(z)\}$ is called an Appell system if $\frac{d}{dz}P^n(z) = nP^{n-1}(z)$ holds with $n = 1, 2, \ldots$. Let $\{mh = (m_1h, m_2h)\ with\ m_1, m_2 \in \mathbf{Z}\}$ be an uniform lattice with step width h.

Theorem 2.1 *The polynomials* $P^n(m_1h, m_2h) = \begin{pmatrix} P_0^n(m_1h, m_2h) \\ P_1^n(m_1h, m_2h) \end{pmatrix}$ *with*

$$\begin{aligned} P_0^n &= \sum_{s=0(2)}^{n} \begin{pmatrix} n \\ s \end{pmatrix} (-1)^{s/2} \prod_{k=s/2}^{n-s/2-1} (m_1-k)h \prod_{l=1-s/2}^{s/2} (m_2+l)h \quad \textit{and} \\ P_1^n &= \sum_{s=1(2)}^{n} \begin{pmatrix} n \\ s \end{pmatrix} (-1)^{(s-1)/2} \prod_{k=(s-1)/2}^{n-s/2-3/2} (m_1-k)h \prod_{l=(1-s)/2}^{(s-1)/2} (m_2+l)h \end{aligned}$$

have in case $n \geq 1$ *the properties*

$$\frac{1}{2} \begin{pmatrix} D_h^{-1} & -D_h^2 \\ D_h^{-2} & D_h^1 \end{pmatrix} \begin{pmatrix} P_0^n(m_1h, m_2h) \\ P_1^n((m_1-1)h, m_2h) \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \quad \textit{and}$$

$$\frac{1}{2} \begin{pmatrix} D_h^1 & D_h^2 \\ -D_h^{-2} & D_h^{-1} \end{pmatrix} \begin{pmatrix} P_0^n(m_1h, m_2h) \\ P_1^n(m_1h, m_2h) \end{pmatrix} = n \begin{pmatrix} P_0^{n-1}(m_1h, m_2h) \\ P_1^{n-1}((m_1-1)h, m_2h) \end{pmatrix}.$$

In this notation D_h^1 *and* D_h^2 *denote forward differences and* D_h^{-1} *and* D_h^{-2} *denote backward differences with respect to the first or second component, respectively. The product of the two matrices leads to the factorization of the discrete Laplacian in the form*

$$D^{1h}\, D^{2h} = \begin{pmatrix} D_h^{-1} & -D_h^2 \\ D_h^{-2} & D_h^1 \end{pmatrix} \begin{pmatrix} D_h^1 & D_h^2 \\ -D_h^{-2} & D_h^{-1} \end{pmatrix} = \begin{pmatrix} \Delta_h & 0 \\ 0 & \Delta_h \end{pmatrix}.$$

Each complex function which fulfill the first property is called discret holomorphic. The second property is the Appell property in the discrete case. For more information about the difference operators which are used here we refer to [3]. A possibility for using the discrete polynomials is discussed in [4].

Before this theorem is proved, the first polynomials are specified in order to illustrate the structure of the them. It is

$$\begin{aligned} P^1(m_1h, m_2h) &= \begin{pmatrix} m_1h \\ m_2h \end{pmatrix} \\ P^2(m_1h, m_2h) &= \begin{pmatrix} m_1h\,(m_1-1)h - m_2h\,(m_2+1)h \\ 2m_1h\,m_2h \end{pmatrix} \\ P^3(m_1h, m_2h) &= \begin{pmatrix} m_1h\,(m_1-1)h\,(m_1-2)h - 3(m_1-1)h\,m_2h\,(m_2+1)h \\ 3m_1h\,(m_1-1)h\,m_2h - (m_2-1)h\,m_2h\,(m_2+1)h \end{pmatrix} \end{aligned}$$

 ISBN: 978-84-608-6082-2

Proof of Theorem 2.1: Having in mind the Appell property, the first component of the left-hand side is considered:

$$\begin{aligned}
&\frac{1}{2}(D_h^1 P_0^n(m_1h, m_2h) + D_h^2 P_1^n(m_1h, m_2h)) \\
= \;& \frac{1}{2h}\left(\sum_{s=0(2)}^{n} \binom{n}{s} (-1)^{s/2} \prod_{k=s/2}^{n-s/2-1} (m_1-(k-1))h \prod_{l=1-s/2}^{s/2} (m_2+l)h \right. \\
& - \sum_{s=0(2)}^{n} \binom{n}{s} (-1)^{s/2} \prod_{k=s/2}^{n-s/2-1} (m_1-k))h \prod_{l=1-s/2}^{s/2} (m_2+l)h \\
& + \sum_{s=1(2)}^{n} \binom{n}{s} (-1)^{(s-1)/2} \prod_{k=(s-1)/2}^{n-s/2-3/2} (m_1-k)h \prod_{l=(1-s)/2}^{(s-1)/2} (m_2+l+1)h \\
& \left. - \sum_{s=1(2)}^{n} \binom{n}{s} (-1)^{(s-1)/2} \prod_{k=(s-1)/2}^{n-s/2-3/2} (m_1-k)h \prod_{l=(1-s)/2}^{(s-1)/2} (m_2+l)h \right)
\end{aligned}$$

In case of $s = n$ for the first two summation signs it holds

$$\binom{n}{n} (-1)^{n/2} \prod_{l=1-n/2}^{n/2} (m_2+l)h - \binom{n}{n} (-1)^{n/2} \prod_{l=1-n/2}^{n/2} (m_2+l)h = 0.$$

Therefore it is possible to substitute the upper summation index by $n-1$. Furthermore it is possible to sum up the first two terms by using the property

$$\binom{n}{s} (n-s) = n \binom{n-1}{s}. \tag{1}$$

It follows

$$\begin{aligned}
& \sum_{s=0(2)}^{n-1} \binom{n}{s} (-1)^{s/2} \left(\prod_{k=s/2-1}^{n-s/2-2} (m_1-k)h - \prod_{k=s/2}^{n-s/2-1} (m_1-k)h \right) \prod_{l=1-s/2}^{s/2} (m_2+l)h \\
= \;& \sum_{s=0(2)}^{n-1} \binom{n}{s} (-1)^{s/2} \prod_{k=s/2}^{n-s/2-2} (m_1-k)h \prod_{l=1-s/2}^{s/2} (m_2+l)h \cdot (n-s)h \\
= \;& nh \sum_{s=0(2)}^{n-1} \binom{n-1}{s} (-1)^{s/2} \prod_{k=s/2}^{n-1-s/2-1} (m_1-k)h \prod_{l=1-s/2}^{s/2} (m_2+l)h\,.
\end{aligned}$$

In order to sum up the last two summands the property

$$\binom{n}{s} s = n \binom{n-1}{s-1} \tag{2}$$

 ISBN: 978-84-608-6082-2

is used. This leads to

$$\begin{aligned}
&\sum_{s=1(2)}^{n}\binom{n}{s}(-1)^{(s-1)/2}\prod_{k=(s-1)/2}^{n-s/2-3/2}(m_1-k)h\left(\prod_{l=(1-s)/2+1}^{(s-1)/2+1}(m_2+l)h-\prod_{l=(1-s)/2}^{(s-1)/2}(m_2+l)h\right)\\
=&\sum_{s=1(2)}^{n}\binom{n}{s}(-1)^{(s-1)/2}\prod_{k=(s-1)/2}^{n-s/2-3/2}(m_1-k)h\prod_{l=(3-s)/2}^{(s-1)/2}(m_2+l)h\cdot sh\\
=&\ nh\sum_{s=1(2)}^{n}\binom{n-1}{s-1}(-1)^{(s-1)/2}\prod_{k=(s-1)/2}^{n-s/2-3/2}(m_1-k)h\prod_{l=(3-s)/2}^{(s-1)/2}(m_2+l)h\\
=&\ nh\sum_{s=0(2)}^{n-1}\binom{n-1}{s}(-1)^{s/2}\prod_{k=s/2}^{n-1-s/2-1}(m_1-k)h\prod_{l=1-s/2}^{s/2}(m_2+l)h
\end{aligned}$$

Because both results are equal, it follows

$$\frac{1}{2}(D_h^1P_0^n(m_1h,m_2h)+D_h^2P_1^n(m_1h,m_2h))=n\,P_0^{n-1}(m_1h,m_2h)\,.$$

Now the second component is considered:

$$\begin{aligned}
&\frac{1}{2}(-D_h^{-2}P_0^n(m_1h,m_2h)+D_h^{-1}P_1^n(m_1h,m_2h))\\
=&\ \frac{1}{2h}\left(-\sum_{s=0(2)}^{n}\binom{n}{s}(-1)^{s/2}\prod_{k=s/2}^{n-s/2-1}(m_1-k)h\prod_{l=1-s/2}^{s/2}(m_2+l)h\right.\\
&+\sum_{s=0(2)}^{n}\binom{n}{s}(-1)^{s/2}\prod_{k=s/2}^{n-s/2-1}(m_1-k))h\prod_{l=1-s/2}^{s/2}(m_2+l-1)h\\
&+\sum_{s=1(2)}^{n}\binom{n}{s}(-1)^{(s-1)/2}\prod_{k=(s-1)/2}^{n-s/2-3/2}(m_1-k)h\prod_{l=(1-s)/2}^{(s-1)/2}(m_2+l)h\\
&\left.-\sum_{s=1(2)}^{n}\binom{n}{s}(-1)^{(s-1)/2}\prod_{k=(s-1)/2}^{n-s/2-3/2}(m_1-(k+1))h\prod_{l=(1-s)/2}^{(s-1)/2}(m_2+l)h\right)
\end{aligned}$$

In case of $s=0$ for the first two sums it follows

$$-\binom{n}{0}(-1)^{0/2}\prod_{k=0}^{n-1}(m_1-k)h+\binom{n}{0}(-1)^{0/2}\prod_{k=0}^{n-1}(m_1-k)h=0.$$

Therefore in both sums the lower summation index can be substituted by 2. Furthermore both expressions can be summarized because of property (2) in the following form:

$$\sum_{s=2(2)}^{n}\binom{n}{s}(-1)^{s/2}\prod_{k=s/2}^{n-s/2-1}(m_1-k)h\left(-\prod_{l=1-s/2}^{s/2}(m_2+l)h+\prod_{l=-s/2}^{s/2-1}(m_2+l)h\right)$$

 ISBN: 978-84-608-6082-2

$$
\begin{aligned}
&= \sum_{s=2(2)}^{n} \binom{n}{s} (-1)^{s/2} \prod_{k=s/2}^{n-s/2-1} (m_1-k)h \prod_{l=1-s/2}^{s/2-1} (m_2+l)h \cdot (-s)h \\
&= -nh \sum_{s=2(2)}^{n} \binom{n-1}{s-1} (-1)^{s/2} \prod_{k=s/2}^{n-s/2-1} (m_1-k)h \prod_{l=1-s/2}^{s/2-1} (m_2+l)h \\
&= -nh \sum_{s=1(2)}^{n-1} \binom{n-1}{s} (-1)^{(s+1)/2} \prod_{k=(s+1)/2}^{n-s/2-3/2} (m_1-k)h \prod_{l=(1-s)/2}^{(s-1)/2} (m_2+l)h \\
&= -nh \sum_{s=1(2)}^{n-1} \binom{n-1}{s} (-1)^{(s+1)/2} \prod_{k=(s-1)/2}^{n-1-s/2-3/2} (m_1-1-k)h \prod_{l=(1-s)/2}^{(s-1)/2} (m_2+l)h\,.
\end{aligned}
$$

In order to combine the last two summands in the second component the relation (1) is used. In case $s=n$ it holds additionally

$$
\binom{n}{n} (-1)^{(n-1)/2} \prod_{l=(1-n)/2}^{(n-1)/2} (m_2+l)h - \binom{n}{n} (-1)^{(n-1)/2} \prod_{l=(1-n)/2}^{(n-1)/2} (m_2+l)h = 0,
$$

such that the upper summation index can be replaced by $n-1$. These steps leads to

$$
\begin{aligned}
&\sum_{s=1(2)}^{n-1} \binom{n}{s} (-1)^{(s-1)/2} \left(\prod_{k=(s-1)/2}^{n-s/2-3/2} (m_1-k)h - \prod_{k=(s-1)/2+1}^{n-s/2-1/2} (m_1-k)h \right) \prod_{l=(1-s)/2}^{(s-1)/2} (m_2+l)h \\
&= \sum_{s=1(2)}^{n-1} \binom{n}{s} (-1)^{(s-1)/2} \prod_{k=(s+1)/2}^{n-s/2-3/2} (m_1-k)h \prod_{l=(1-s)/2}^{(s-1)/2} (m_2+l)h \cdot (n-s)h \\
&= nh \sum_{s=1(2)}^{n-1} \binom{n-1}{s} (-1)^{(s-1)/2} \prod_{k=(s+1)/2}^{n-s/2-3/2} (m_1-k)h \prod_{l=(1-s)/2}^{(s-1)/2} (m_2+l)h \\
&= nh \sum_{s=1(2)}^{n-1} \binom{n-1}{s} (-1)^{(s-1)/2} \prod_{k=(s-1)/2}^{n-1-s/2-3/2} (m_1-1-k)h \prod_{l=(1-s)/2}^{(s-1)/2} (m_2+l)h\,.
\end{aligned}
$$

Based on the property $(-1)^{(s+1)/2} = (-1)^{(s-1)/2+1} = -(-1)^{(s-1)/2}$ both results can be added and it follows

$$
\frac{1}{2}(-D_h^{-2} P_0^n(m_1h, m_2h) + D_h^{-1} P_1^n(m_1h, m_2h)) = n\, P_1^{n-1}((m_1-1)h, m_2h).
$$

In this way the Appell property is proved. Repeating all steps adapted to the changed finite differences it is easy to prove that the polynomials considered are discret holomorphic in

 ISBN: 978-84-608-6082-2

the sense of the property

$$\frac{1}{2}\begin{pmatrix} D_h^{-1} & -D_h^2 \\ D_h^{-2} & D_h^1 \end{pmatrix}\begin{pmatrix} P_0^n(m_1h,m_2h) \\ P_1^n((m_1-1)h,m_2h) \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \quad \blacksquare$$

From the last property it follows immediately

$$\begin{aligned} & \frac{1}{4}\begin{pmatrix} D_h^{-1} & -D_h^2 \\ D_h^{-2} & D_h^1 \end{pmatrix}\begin{pmatrix} D_h^1 & D_h^2 \\ -D_h^{-2} & D_h^{-1} \end{pmatrix}\begin{pmatrix} P_0^n(m_1h,m_2h) \\ P_1^n(m_1h,m_2h) \end{pmatrix} \\ = & \frac{n}{2}\begin{pmatrix} D_h^{-1} & -D_h^2 \\ D_h^{-2} & D_h^1 \end{pmatrix}\begin{pmatrix} P_0^{n-1}(m_1h,m_2h) \\ P_1^{n-1}((m_1-1)h,m_2h) \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}. \end{aligned}$$

Because of the factorization of the discrete Laplacian it is obviously shown that the discrete polynomials are discrete harmonic, too.

The next step is to investigate whether the polynomials are linear independent. In the whole plane it is possible to show that

$$\begin{pmatrix} P_0(m_1h,m_2h) \\ P_1(m_1h,m_2h) \end{pmatrix} := \sum_{i=0}^{n} a_i \begin{pmatrix} P_0^i(m_1h,m_2h) \\ P_1^i(m_1h,m_2h) \end{pmatrix} \equiv \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

holds for all (m_1h,m_2h) only if all a_i with $i=0,\ldots,n$ are identically to zero. Based on the structure of the polynomials it follows

$$\begin{pmatrix} P_0(0,0) \\ P_1(0,0) \end{pmatrix} = \begin{pmatrix} a_0 \\ 0 \end{pmatrix}.$$

Consequently, only in case of $a_0=0$ the real part is equal to zero. Using the Appell property of the polynomials and a small correction concerning the difference operator in order to do the transition from the mesh point $((m_1-1)h,m_2h)$ to the mesh point (m_1h,m_2h), it is easy to see that the identity

$$\begin{aligned} & \frac{1}{2}\begin{pmatrix} D_h^1 & D_h^2 \\ -D_h^{-2}-hD_h^1D_h^{-2} & D_h^{-1}+hD_h^1D_h^{-1} \end{pmatrix}\begin{pmatrix} P_0(m_1h,m_2h) \\ P_1(m_1h,m_2h) \end{pmatrix} \\ = & \sum_{i=0}^{n} ia_i \begin{pmatrix} P_0^{i-1}(m_1h,m_2h) \\ P_1^{i-1}(m_1h,m_2h) \end{pmatrix} \equiv \begin{pmatrix} 0 \\ 0 \end{pmatrix} \end{aligned}$$

in the mesh point $(m_1h,m_2h)=(0,0)$ is only fulfilled if $a_1=0$ holds. By repeated application of the modified difference operator it can be shown that for all a_i, $i=0,\ldots,n$ the identity $a_i=0$ must be fulfilled.

From the continuous case it is well-known that polynomials can not only be developed according to powers of z, but also according to powers of $z-z_0$, where z_0 is an arbitrary

 ISBN: 978-84-608-6082-2

fixed complex number. In the following this strategy is transferred to the discrete case. Let $N = N_1 \cdot N_2$. For a fixed degree of polynomials $n = N - 1$ the identity

$$\sum_{n_1=1}^{N_1} \sum_{n_2=1}^{N_2} a_{n_1,n_2} \begin{pmatrix} P_0^n((m_1-n_1)h,(m_2-n_2)h) \\ P_1^n((m_1-n_1)h,(m_2-n_2)h) \end{pmatrix} \equiv \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

is investigated. It is possible to add n further equations by applying step by step the modified difference operator

$$\frac{1}{2} \begin{pmatrix} D_h^1 & D_h^2 \\ -D_h^{-2} - hD_h^1 D_h^{-2} & D_h^{-1} + hD_h^1 D_h^{-1} \end{pmatrix}.$$

Based on the Appell property the degree of the polynomial decreases from equation to equation. The whole equation system can be written in matrix representation with complex elements. Because the right-hand side is in each component equal to zero, all coefficients a_{n_1,n_2} have to be zero if the matrix of the polynomials has a determinant different from zero. Therefore the structure of the matrix is to be investigated in more detail. By transposing the matrix it becomes obviously that it is a Vandermonde matrix. By investigating the column vectors of the transposed matrix it is possible to show that these column vectors are linear independent. Therefore a linear combination of the column vectors is formed and the resulting problem is considered line by line. In each line, polynomials of ascending degree are located in one and the same lattice point. For these polynomials, the linear independence has been shown already, so that the linear independence of the column vectors follows directly from this property.

3 Diskrete Exponential,- Sinus- und Cosinusfunktionen

In order to discretize the exponential function the investigations start with the one-dimensional case. The aim is to find a function, which difference derivation is the function themselve.

Lemma 3.1 *In the one-dimensional case it holds*

$$D_h^1(1+h)^{m_1} = (1+h)^{m_1} \qquad D_h^{-1}(1-h)^{-m_1} = (1-h)^{-m_1}$$
$$D_h^1(1-h)^{m_1} = -(1-h)^{m_1} \qquad D_h^{-1}(1+h)^{-m_1} = -(1+h)^{-m_1}.$$

Proof: The steps of the proof are demonstrated for the first equation. All other relations are proved in the same way. Using the definition of forward differences it holds

$$D_h^1(1+h)^{m_1} = h^{-1}[(1+h)^{m_1+1} - (1+h)^{m_1}] = h^{-1}(1+h)^{m_1}(1+h-1) = (1+h)^{m_1} \quad \blacksquare$$

 ISBN: 978-84-608-6082-2

The function $(1+h)^{m_1}$ considered in Lemma 3.1 is now studied in more detail. Using the binomial theorem it follows

$$\begin{aligned}(1+h)^{m_1} &= \sum_{k=0}^{m_1} \left(\begin{array}{c} m_1 \\ k \end{array} \right) h^k \\ = \quad 1 + m_1 h + \frac{m_1 h \cdot (m_1-1)h}{2!} &+ \frac{m_1 h \cdot (m_1-1)h \cdot (m_1-2)h}{3!} + \ldots + m_1 h^{m_1-1} + h^{m_1}.\end{aligned}$$

On the other hand the exponential function in the continuous case can be developed into the infinit series $e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \ldots$. Therefore it is obviously to consider polynomials of the form

$$x_+^n = \prod_{k=0}^{n-1} (m_1 - k)\, h.$$

These polynomials have the property $D_h^1\, x_+^n = n\, x_+^{n-1}$, since

$$\begin{aligned} D_h^1\, x_+^n &= \frac{1}{h} \left(\prod_{k=0}^{n-1} (m_1 + 1 - k)\, h - \prod_{k=0}^{n-1} (m_1 - k)\, h \right) \\ &= \frac{1}{h} \left(\prod_{k=0}^{n-1} (m_1 - (k-1))\, h - \prod_{k=0}^{n-1} (m_1 - k)\, h \right) \\ &= \frac{1}{h} \left(\prod_{k=-1}^{n-2} (m_1 - k)\, h - \prod_{k=0}^{n-1} (m_1 - k)\, h \right) \\ &= \frac{1}{h} \left(\prod_{k=0}^{n-2} (m_1 - k)\, h \right) \Big((m_1+1)h - (m_1 - (n-1))h \Big) \\ &= \frac{1}{h} \cdot x_+^{n-1} \cdot n \cdot h\ . \end{aligned}$$

In analogy to this strategy it is possible by using polynomials of the form

$$x_-^n = \prod_{k=0}^{n-1} (m_1 + k)\, h$$

to prove the property

$$\begin{aligned} D_h^{-1}\, x_-^n &= \frac{1}{h} \left(\prod_{k=0}^{n-1} (m_1 + k)\, h - \prod_{k=0}^{n-1} (m_1 - 1 + k)\, h \right) \\ &= \frac{1}{h} \left(\prod_{k=0}^{n-1} (m_1 + k)\, h - \prod_{k=-1}^{n-2} (m_1 + k)\, h \right) \end{aligned}$$

 ISBN: 978-84-608-6082-2

$$\begin{aligned} &= \frac{1}{h}\left(\prod_{k=0}^{n-2}(m_1+k)\,h\right)\Big((m_1+(n-1))h-(m_1-1)h\Big) \\ &= n\cdot x_-^{n-1}\,. \end{aligned}$$

With regard to the note in the introduction the here mentioned polynomials were already studied by Faustino und Kähler in [2]. They were the basis for the construction of the discrete holomorphic polynomials in Theorem 2.1.

The question is how to find a discret holomorphic exponential function in the complex case by using the information from the one-dimensional case. In detail the one-dimensional case should be generalized so that the following two properties remain intact: On the one hand side the function should be discret holomorphic. Using the notation from section 2 we are looking for a function $\mathrm{EXP}_h(m_1h,m_2h)=\begin{pmatrix}\mathrm{EXP}_h^0(m_1h,m_2h)\\ \mathrm{EXP}_h^1(m_1h,m_2h)\end{pmatrix}$ with the property

$$D^{1h}\,\mathrm{EXP}_h=0.$$

On the other hand the complex derivation of the function should return the function itself. Based on the information that the operator D^{2h} converges to $\left(\frac{\partial}{\partial x}-\mathbf{i}\frac{\partial}{\partial y}\right)$ and looking at the definition $\partial_z=\frac{1}{2}\left(\frac{\partial}{\partial x}-\mathbf{i}\frac{\partial}{\partial y}\right)$ the desired discrete function should have the property

$$\frac{1}{2}\,D^{2h}\,\mathrm{EXP}_h=\mathrm{EXP}_h.$$

The function described in the following theorem essentially satisfies these requirements, even if some neighboring grid points are included.

Theorem 3.1 *The function* $EXP_h(m_1h,m_2h)$ *with*

$$\begin{aligned} EXP_h^0(m_1h,m_2h) &= \frac{1}{2}(1+h)^{m_1-1}[(1+\mathbf{i}h)^{m_2+1}+(1-\mathbf{i}h)^{m_2+1}] \\ EXP_h^1(m_1h,m_2h) &= \frac{1}{2\mathbf{i}}(1+h)^{m_1}[(1+\mathbf{i}h)^{m_2}-(1-\mathbf{i}h)^{m_2}] \end{aligned}$$

has the properties

$$\begin{aligned} D_h^{-1}EXP_h^0((m_1+1)h,(m_2-1)h)-D_h^2EXP_h^1((m_1-1)h,m_2h) &= 0 \\ D_h^{-2}EXP_h^0(m_1h,m_2h)+D_h^1EXP_h^1((m_1-1)h,m_2h) &= 0 \end{aligned}$$

and

$$\begin{aligned} &\frac{1}{2}D_h^1EXP_h^0(m_1h,m_2h)+\frac{1}{2}D_h^2EXP_h^1(m_1h,m_2h) \\ =\;&\frac{1}{2}[EXP_h^0(m_1h,m_2h)+EXP_h^0((m_1+1)h,(m_2-1)h)] \end{aligned}$$

 ISBN: 978-84-608-6082-2

as well as

$$-\frac{1}{2}D_h^{-2}EXP_h^0((m_1+1)h,m_2h)+\frac{1}{2}D_h^{-1}EXP_h^1((m_1+1)h,m_2h)=EXP_h^1(m_1h,m_2h).$$

Proof: With regard to the first property it holds

$$\begin{aligned}
& D_h^{-1}\mathrm{EXP}_h^0((m_1+1)h,(m_2-1)h)-D_h^2\mathrm{EXP}_h^1((m_1-1)h,m_2h)\\
= \;& \frac{1}{2h}[(1+h)^{m_1}-(1+h)^{m_1-1}][(1+\mathbf{i}h)^{m_2}+(1-\mathbf{i}h)^{m_2}]\\
& -\frac{1}{2\mathbf{i}h}(1+h)^{m_1-1}[(1+\mathbf{i}h)^{m_2+1}-(1-\mathbf{i}h)^{m_2+1}-(1+\mathbf{i}h)^{m_2}+(1-\mathbf{i}h)^{m_2}]\\
= \;& \frac{1}{2h}(1+h)^{m_1-1}(1+h-1)[(1+\mathbf{i}h)^{m_2}+(1-\mathbf{i}h)^{m_2}]\\
& -\frac{1}{2\mathbf{i}h}(1+h)^{m_1-1}[(1+\mathbf{i}h)^{m_2}(1+\mathbf{i}h-1)+(1-\mathbf{i}h)^{m_2}(-1+\mathbf{i}h+1)]\\
= \;& 0.
\end{aligned}$$

The second property results from

$$\begin{aligned}
& D_h^{-2}\mathrm{EXP}_h^0(m_1h,m_2h)+D_h^1\mathrm{EXP}_h^1((m_1-1)h,m_2h)\\
= \;& \frac{1}{2h}(1+h)^{m_1-1}[(1+\mathbf{i}h)^{m_2+1}+(1-\mathbf{i}h)^{m_2+1}-(1+\mathbf{i}h)^{m_2}-(1-\mathbf{i}h)^{m_2}]\\
& +\frac{1}{2\mathbf{i}h}[(1+h)^{m_1}-(1+h)^{m_1-1}][(1+\mathbf{i}h)^{m_2}-(1-\mathbf{i}h)^{m_2}]\\
= \;& \frac{1}{2h}(1+h)^{m_1-1}[(1+\mathbf{i}h)^{m_2}((1+\mathbf{i}h-1)-(1-\mathbf{i}h)^{m_2}(-1+\mathbf{i}h+1)]\\
& +\frac{1}{2ih}(1+h)^{m_1-1}(1+h-1)[(1+\mathbf{i}h)^{m_2}-(1-\mathbf{i}h)^{m_2}]\\
= \;& 0.
\end{aligned}$$

Now the third term is investigated. It follows

$$\begin{aligned}
& \frac{1}{2}D_h^1\mathrm{EXP}_h^0(m_1h,m_2h)+\frac{1}{2}D_h^2\mathrm{EXP}_h^1(m_1h,m_2h)\\
= \;& \frac{1}{4h}[(1+h)^{m_1}-(1+h)^{m_1-1}][(1+\mathbf{i}h)^{m_2+1}+(1-\mathbf{i}h)^{m_2+1}]\\
& +\frac{1}{4\mathbf{i}h}(1+h)^{m_1}[(1+\mathbf{i}h)^{m_2+1}-(1-\mathbf{i}h)^{m_2+1}-(1+\mathbf{i}h)^{m_2}+(1-\mathbf{i}h)^{m_2}]\\
= \;& \frac{1}{4h}(1+h)^{m_1-1}(1+h-1)[[(1+\mathbf{i}h)^{m_2+1}+(1-\mathbf{i}h)^{m_2+1}]\\
& +\frac{1}{4\mathbf{i}h}(1+h)^{m_1}[(1+\mathbf{i}h)^{m_2}((1+\mathbf{i}h-1)+(1-\mathbf{i}h)^{m_2}(-1+\mathbf{i}h+1)].
\end{aligned}$$

 ISBN: 978-84-608-6082-2

Finally the last property is considered.

$$\begin{aligned}
& -\frac{1}{2}D_h^{-2}\mathrm{EXP}_h^0((m_1+1)h,m_2h)+\frac{1}{2}D_h^{-1}\mathrm{EXP}_h^1((m_1+1)h,m_2h)\\
= \;& \frac{1}{4h}(1+h)^{m_1}[-(1+\mathbf{i}h)^{m_2+1}-(1-\mathbf{i}h)^{m_2+1}+(1+\mathbf{i}h)^{m_2}+(1-\mathbf{i}h)^{m_2}]\\
& +\frac{1}{4\mathbf{i}h}[(1+h)^{m_1+1}-(1+h)^{m_1}][(1+\mathbf{i}h)^{m_2}-(1-\mathbf{i}h)^{m_2}]\\
= \;& \frac{1}{4h}(1+h)^{m_1}[(1+\mathbf{i}h)^{m_2}(-1-\mathbf{i}h+1)-(1-\mathbf{i}h)^{m_2}(1-\mathbf{i}h-1)]\\
& +\frac{1}{4\mathbf{i}h}(1+h)^{m_1}(1+h-1)[(1+\mathbf{i}h)^{m_2}-(1-\mathbf{i}h)^{m_2}] \quad \blacksquare
\end{aligned}$$

In the second part of this section we look for a discretization of the sine and cosine function. Again we start with the one-dimensional case. Using the relations $\cos(x)=\frac{1}{2}(e^{\mathbf{i}x}+e^{-\mathbf{i}x})$ and $\sin(x)=\frac{1}{2\mathbf{i}}(e^{\mathbf{i}x}-e^{-\mathbf{i}x})$ the discrete sine and cosine functions investigated in the following lemma were obtained.

Lemma 3.2 *The discrete functions* $\cos_h^+(m_1h)=\frac{1}{2}[(1+\mathbf{i}h)^{m_1}+(1-\mathbf{i}h)^{m_1}]$ *and* $\sin_h^+(m_1h)=\frac{1}{2\mathbf{i}}[(1+\mathbf{i}h)^{m_1}-(1-\mathbf{i}h)^{m_1}]$ *have the properties*

$$\begin{aligned}
& D_h^1\,D_h^1\,\cos_h^+(m_1h)=-D_h^1\sin_h^+(m_1h)=-\cos_h^+(m_1h) \quad \textit{and}\\
& D_h^1\,D_h^1\,\sin_h^+(m_1h)=D_h^1\cos_h^+(m_1h)=-\sin_h^+(m_1h).
\end{aligned}$$

On the other hand for functions $\cos_h^-(m_1h)=\frac{1}{2}[(1-\mathbf{i}h)^{-m_1}+(1+\mathbf{i}h)^{-m_1}]$ *and* $\sin_h^-(m_1h)=\frac{1}{2\mathbf{i}}[(1-\mathbf{i}h)^{-m_1}-(1+\mathbf{i}h)^{-m_1}]$ *it follows*

$$\begin{aligned}
& D_h^{-1}\,D_h^{-1}\,\cos_h^-(m_1h)=-D_h^{-1}\sin_h^-(m_1h)=-\cos_h^-(m_1h) \quad \textit{and}\\
& D_h^{-1}\,D_h^{-1}\,\sin_h^-(m_1h)=D_h^{-1}\cos_h^-(m_1h)=-\sin_h^-(m_1h).
\end{aligned}$$

Additionally it is possible to prove

$$\begin{aligned}
(\cos_h^+(m_1h))^2+(\sin_h^+(m_1h))^2 &= (1+h^2)^{m_1}\\
(\cos_h^-(m_1h))^2+(\sin_h^-(m_1h))^2 &= (1+h^2)^{-m_1}\,,
\end{aligned}$$

where the right-hand sides converges to one for $h\to 0$ *and the product of both left-hand sides gives exactly one.*

The proof of Lemma 3.2 as well as the proof of Lemma 3.1 is based on straight forward calculations using forward and backward differences. Now it is possible to explain the origin of the complex exponential function in Theorem 3.1. It is based on the equation $e^z=e^{x+\mathbf{i}y}=e^x(\cos(x)+\mathbf{i}\sin(x))$.

 ISBN: 978-84-608-6082-2

Fortunately, using the relation

$$\cos z = \frac{1}{2}(e^{\mathbf{i}z} + e^{-\mathbf{i}z})) = \frac{1}{2}(e^{\mathbf{i}x-y} + e^{-\mathbf{i}x+y}) = \frac{1}{2}(e^{y} + e^{-y})\cos x + \frac{\mathbf{i}}{2}(e^{-y} - e^{y})\sin x$$

in the complex case, a discrete cosine function $\mathrm{COS}_h(m_1h, m_2h) = \begin{pmatrix} \mathrm{COS}_h^0(m_1h, m_2h) \\ \mathrm{COS}_h^1(m_1h, m_2h) \end{pmatrix}$ can be defined. Even in the special combination of the grid points, the function fits excellently with the exponential function in Theorem 3.1. By applying the operator $\frac{1}{2}D^{2h}$ the correspon-ding discrete sine function $\mathrm{SIN}_h(m_1h, m_2h)$ is explained.

Theorem 3.2 *The function $COS_h(m_1h, m_2h)$ with*

$$\begin{aligned} COS_h^0(m_1h, m_2h) &= \frac{1}{4}[(1-h)^{m_2+1} + (1+h)^{m_2+1}][(1+\mathbf{i}h)^{m_1-1} + (1-\mathbf{i}h)^{m_1-1}] \\ COS_h^1(m_1h, m_2h) &= \frac{1}{4\mathbf{i}}[(1-h)^{m_2} - (1+h)^{m_2}][(1+\mathbf{i}h)^{m_1} - (1-\mathbf{i}h)^{m_1}] \end{aligned}$$

has the properties

$$\begin{aligned} D_h^{-1}COS_h^0((m_1+1)h, (m_2-1)h) - D_h^2 COS_h^1((m_1-1)h, m_2h) &= 0 \\ D_h^{-2}COS_h^0(m_1h, m_2h) + D_h^1 COS_h^1((m_1-1)h, m_2h) &= 0 \end{aligned}$$

and

$$\frac{1}{2}\,D^{2h}COS_h(m_1h, m_2h) = -\begin{pmatrix} \frac{1}{2}[SIN_h^0(m_1h, m_2h) + SIN_h^0((m_1+1)h, (m_2-1)h] \\ SIN_h^1((m_1-1)h, m_2h) \end{pmatrix}$$

with

$$\begin{aligned} SIN_h^0(m_1h, m_2h) &= \frac{1}{4\mathbf{i}}[(1+h)^{m_2+1} + (1-h)^{m_2+1}][(1+\mathbf{i}h)^{m_1-1} - (1-\mathbf{i}h)^{m_1-1}] \\ SIN_h^1(m_1h, m_2h) &= \frac{1}{4}[(1+h)^{m_2} - (1-h)^{m_2}][(1+\mathbf{i}h)^{m_1} + (1-\mathbf{i}h)^{m_1}]\,. \end{aligned}$$

Conversely, the equations

$$\begin{aligned} D_h^{-1}SIN_h^0((m_1+1)h, (m_2-1)h) - D_h^2 SIN_h^1((m_1-1)h, m_2h) &= 0 \\ D_h^{-2}SIN_h^0(m_1h, m_2h) + D_h^1 SIN_h^1((m_1-1)h, m_2h) &= 0 \end{aligned}$$

and

$$\frac{1}{2}\,D^{2h}SIN_h(m_1h, m_2h) = \begin{pmatrix} \frac{1}{2}[COS_h^0(m_1h, m_2h) + COS_h^0((m_1+1)h, (m_2-1)h)] \\ COS_h^1((m_1-1)h, m_2h) \end{pmatrix}$$

are fulfilled. When using the operator D^{2h}, there is only one shift in the two summands in the first component concerning the lattice points, which prevents a summation.

 ISBN: 978-84-608-6082-2

The proof of Theorem 3.2 is entirely analogous to the proof steps of Theorem 3.1.

In addition it is possible to get results like

$$\begin{aligned} & \frac{1}{4} D^{2h} D^{2h} \mathrm{SIN}_h(m_1 h, m_2 h) \\ = & \begin{pmatrix} -\mathrm{SIN}_h^0(m_1 h, (m_2-1)h) + O(h) \\ -\mathrm{SIN}_h^1((m_1-2)h, (m_2-1)h) + O(h) \end{pmatrix}. \end{aligned}$$

By the help of the discrete holomorphic functions a tool is created which is very helpful for the solution of difference equations. For instance the discrete exponential function can be used to extend the theory of discrete Cauchy-Riemann operators in the sense of Vekua. The next step can be the generalization of the results in the complex case to the hypercomplex case.

References

[1] P. E. Appell, *Sur une class de polynomes*, Ann. Sci École Norm. Sup **9**,119–144, 1880.

[2] N. Faustino U. Kähler, *Fischer Decomposition for Difference Dirac Operators* Advances in Applied Clifford Algebras, vol. **17**, Nr.1, 37-58, 2007.

[3] A. Hommel *A discrete theorem of Goursat* Advances in Applied Clifford Algebras, vol. **24**, issue 4,1039-1045, 2014.

[4] K. Gürlebeck A. Hommel *The relationship between linear elasicity theory and complex function theory studied on the basis of finite differences* Digital proceedings of the 20th International Conference on the Application of Computer Science and Mathematics in Architecture and Civil Engineering (IKM), 2015.

 ISBN: 978-84-608-6082-2

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

High order iterative methods with memory for nonlinear equations

Cory L. Howk[1], Jose L. Hueso[2], Eulalia Martínez[2] and Carles Teruel.[2]

[1] *Dept. of Mathematics and Computer Science, W. Carolina University, Cullowhee, North Carolina*

[2] *Instituto Universitario de Matemática Multidisciplinar, Universitat Politècnica de València, Spain*

emails: clhowk@email.wcu.edu, jlhueso@mat.upv.es, eumarti@mat.upv.es, cartefer@teleco.upv.es

Abstract

In this paper we obtain some theoretical results about iterative methods with memory for nonlinear equations. The main idea consists of using for the predictor step of each iteration a quantity that has already been calculated in the previously, usually the quantity governing the slope from the previous corrector step. In this way we do not introduce any extra computation, and more importantly we avoid new functional evaluations, allowing us to obtain high iterative methods in a simple way. A specific class of methods of this type is introduced, and we prove that the convergence order is $2^n + 2^{n-2}$ with $n + 1$ functional evaluations. Finally, an exhaustive efficiency study is performed to show the competitiveness of these methods.

Key words: Iterative methods with memory, convergence rate, efficiency, Kung-Traub conjecture.

MSC 2000: 47H99, 65H10, 65B05, 47H17, 49M15.

1 Introduction

Nonlinear equations appear in a natural way in many applications of science and engineering. The solutions are typically approximated via iterative methods. Newton's method and other "simple" techniques are common, and historically have been sufficient. However, nowadays high-order methods are very important, as many scientific applications (astronomy, climate

simulations, etc.) need high precision in their computations. These methods allow getting the required precession without a significant increase of the number of iterations.

The convergence order of an iterative method is directly related with the efficiency in the sense of the conjecture of Kung-Traub, [3], which states that a method without memory that uses $n+1$ function evaluations per iterate can have a convergence rate of at most 2^n.

This paper focuses on the consequences of memory reuse rather than just using the historical points in new ways. It is well known that iterative methods with memory can surpass the Kung-Traub conjecture, but often the computational cost of obtaining high-order methods is very expensive. Therefore we concentrate in this work on trying to use memory to obtain a high-order of convergence, but without introducing extra computation to the iterative expression.

Our idea in this work is to simplify this cost by constructing high-order methods in a way that the iterative expression remains as simple as possible. We use the idea of using for the predictor step of each iteration a quantity that has already been calculated in the previous iteration. While high-order methods are important, the operational cost of getting them must also be taken into account. For this reason an exhaustive efficiency study is performed to show the effectiveness of these methods.

2 Predictor-Corrector iterative methods

While Newton's Method is one of the most popular root-finding algorithms due both to its simplicity and its quadratic rate of convergence, many more powerful methods exist with faster convergence rates. The methods of derivation for these powerful methods often result in implicit equations. For example, if one integrates $f'(x)$ between the current iterate and the (unknown) root, $\int_{x_n}^{\alpha} f'(x)dx = -f(x_n)$ and apply the Midpoint Rule to the integral, we get $f'(\frac{1}{2}[x_n+\alpha])(\alpha - x_n) \approx -f(x_n)$, which can be rearranged into the implicit equation $\alpha \approx x_n - \dfrac{f(x_n)}{f'(\frac{1}{2}[x_n+\alpha])}$. These equations are typically applied algorithmically as a predictor-corrector set ([1]-[2]); for example,

$$\begin{aligned} y_n &= x_n - f(x_n)\Psi(x_n) \\ x_{n+1} &= x_n - \frac{f(x_n)}{f'(\frac{1}{2}[x_n+y_n])} \end{aligned} \tag{1}$$

for some function Ψ. It is common to use Newton's Method for the predictor step,

$$\begin{aligned} y_n &= x_n - \frac{f(x_n)}{f'(x_n)} \\ x_{n+1} &= x_n - f(x_n)\Phi(x_n, y_n). \end{aligned} \tag{2}$$

 ISBN: 978-84-617-8694-7

In this sense, one could view this type of algorithm as Newton's Method with a corrective factor. We would like to use for the predictor step a quantity that has already been calculated, namely $\Phi(x_{n-1}, y_{n-1})$, that is the quantity governing the slope from the previous corrector step.

$$\begin{aligned} y_n &= x_n - f(x_n)\Phi(x_{n-1}, y_{n-1}) \\ x_{n+1} &= x_n - f(x_n)\Phi(x_n, y_n). \end{aligned} \tag{3}$$

This type of method will be referred to as the *Standard 2-step Predictor-Corrector* algorithm, denoted by SA.

We again consider the effect of incorporating the previous slope into the predictor step

$$\begin{aligned} y_n &= x_n - f(x_n)\Phi(x_{n-1}, y_{n-1}) \\ x_{n+1} &= y_n - f(y_n)\Phi(x_n, y_n), \end{aligned} \tag{4}$$

and refer to this as the *Improved 2-step Predictor-Corrector* algorithm, denoted by IA. This can be further expanded into an *Improved multi-step Predictor-Corrector* algorithm with a similar scheme.

3 Convergence Analysis

We begin by establishing convergence results for the two-step predictor-corrector iterative schemes previously defined to locate a zero $x = \alpha$ of $f(x)$. Define the error terms $\varepsilon_n = x_n - \alpha$ and $\varepsilon_n^* = y_n - \alpha$ for all n. The sequence $\{x_n\}_{n=0}^{\infty}$ is said to converge to $x = \alpha$ with *rate of convergence* ρ if $\lim\limits_{n\to\infty} \dfrac{|\varepsilon_{n+1}|}{|\varepsilon_n|^\rho} = \lambda$ for positive constants ρ, λ. The relationship between ε_{n+1} and ε_n can therefore be written $\varepsilon_{n+1} = A_n\varepsilon_n^\rho + h.o.t.$ where $A_n \to \lambda$ as $n \to \infty$ and $h.o.t.$ denotes *higher-order terms* which satisfy $\lim\limits_{n\to\infty} \dfrac{h.o.t.}{\varepsilon_n^\rho} = 0$. This relationship will be denoted by $\varepsilon_{n+1} \sim \varepsilon_n^\rho$.

Theorem 1. *Suppose that the error condition for the corrector step of the Standard 2-step Predictor-Corrector algorithm is given by $\varepsilon_{n+1} \sim \varepsilon_n^p(\varepsilon_n^*)^q$, with p and q real positive numbers, and that f is sufficiently differentiable. Then the convergence rate for the algorithm is given by $\rho = \dfrac{p + q + \sqrt{(p+q)^2 + 4q}}{2}$.*

 ISBN: 978-84-617-8694-7

3.1 Applying theorem 1 to the Midpoint method with memory

We consider the Midpoint Method with memory, denoted by MP, derived at the beginning of Section 2. It is given by:

$$y_n = x_n - \frac{f(x_n)}{f'(\frac{1}{2}[x_{n-1} + y_{n-1}])}$$
$$x_{n+1} = x_n - \frac{f(x_n)}{f'(\frac{1}{2}[x_n + y_n])}. \tag{5}$$

This is a Standard 2-step Predictor-Corrector algorithm. The error relationship from the corrector step is derived in [5] and shown to be $\varepsilon_{n+1} \sim \varepsilon_n \varepsilon_n^*$, so that $p = q = 1$. Therefore the rate of convergence is $\rho = \dfrac{1 + 1 + \sqrt{(1+1)^2 + 4}}{2} = 1 + \sqrt{2} \approx 2.414$.

Theorem 2. *Suppose that the error condition for the corrector step of the Improved 2-step Predictor-Corrector algorithm is given by $\varepsilon_{n+1} \sim \varepsilon_n^p (\varepsilon_n^*)^q$, with p and q real positive numbers, and f is sufficiently differentiable. Then the convergence rate for the algorithm, denoted by IA, is given by $\rho = \dfrac{(p+q+1) + \sqrt{(p+q+1)^2 - 4p}}{2}$.*

3.2 Efficiency Indices

The *efficiency index* of an iterative algorithm as defined in [2] measures the balance between the rate of convergence (ρ) and the number of functional evaluations (m) required per iterate, and is defined by $EI = \rho^{1/m}$. Obviously methods are preferable when they have a higher efficiency index.

The algorithms stated in the previous section, SA and IA, have slower convergence order than their Newton analogs, NP. These algorithms reuse conveniently the predictor step, resulting in one fewer functional evaluation and surpassing the optimal convergence rate stated by the Kung-Traub conjecture.

The following theorem addresses the efficiency of the respective algorithms.

Theorem 3. *Suppose one iterative scheme has convergence rate ρ_1, requiring m functional evaluations per iterate, while a second iterative scheme has convergence rate ρ_2 with $1 < \rho_2 < \rho_1$, requiring $m - 1$ functional evaluations per iterate. The efficiency index of the second scheme is greater than the first scheme for $m < \dfrac{\ln(\rho_1)}{\ln(\rho_1) - \ln(\rho_2)}$.*

According to this Theorem, the Standard 2-step Predictor-Corrector is more efficient than its Newton analog when:

$$m < \frac{\ln(p+2q)}{\ln\left(\frac{2(p+2q)}{p+q+\sqrt{(p+q)^2+4q}}\right)},$$

 ISBN: 978-84-617-8694-7

while the Improved 2-step Predictor-Corrector is more efficient than its Newton analog when:

$$m < \frac{\ln(p+2q)}{\ln\left(\frac{2(p+2q)}{p+q+1\sqrt{(p+q+1)^2-4p}}\right)}.$$

Next we consider the convergence rate of the Improved multi-step Predictor-Corrector (IMS).

Theorem 4. *Suppose that the error conditions for the predictor steps* $(1),...,(k)$ *of the Improved k-step Predictor-Corrector algorithm are given by* $\varepsilon_n^{(\ell)} \sim \varepsilon_n^{r_\ell}(\varepsilon_n^{(0)})^{s_\ell}$*, that of the corrector step is given by* $\varepsilon_{n+1} \sim \varepsilon_n^p(\varepsilon_n^{(0)}))^q$*, and that* f *is sufficiently differentiable. Then the convergence rate for the algorithm is given by:*

$$\rho = \frac{(p+q+s_k) \pm \sqrt{(p+q+s_k)^2 - 4(ps_k - qr_k)}}{2}.$$

4 Proposed iterative methods

In this section we propose a new multistep algorithm that utilizes the idea that we have introduced in the previous sections, that is, to use the prior corrector slope for the current first predictor slope.

We consider the ZLH (Zheng-Li-Huang) family presented in [7]. The family uses $n+1$ points $(y_k^{(-1)}, y_k^{(0)}, \cdots, y_k^{(n-1)})$, utilizes $n+1$ function evaluations per iterate, and is shown to have the optimal convergence rate of 2^n. However, the algorithm that we propose interchanges the roles of $y_k^{(-1)}$and $y_k^{(0)}$ and reuses the slope from the previous corrector step to obtain the first step in a new iteration. This modification introduces memory into the expression, and this algorithm is shown in the following proof to have a convergence rate of $2^n + 2^{n-2}$ by using the same number of functional evaluations per iterate than that of the ZLH family. The function iteration for the proposed algorithm, (OP), has $n-1$ intermediate steps and is given by:

$$\begin{aligned}
y_k^{(0)} &= x_k - \frac{f(x_k)}{P'_{k-1,n-1}(y_k^{(n-1)})}, \qquad y_k^{(-1)} = x_k \\
y_k^{(\ell+1)} &= y_k^{(\ell)} - \frac{f(y_k^{(\ell)})}{P'_{k,\ell}(y_k^{(\ell)})}, \quad \ell = 0,1,...,n-2 \\
x_{k+1} &= y_k^{(n-1)} - \frac{f(y_k^{(n-1)})}{P'_{k,n-1}(y_k^{(n-1)})}.
\end{aligned} \tag{6}$$

 ISBN: 978-84-617-8694-7

Obviously we need for the first iteration a starting guess x_0 and a constant value for $P'_{0,n-1}(y_0^{(n-1)})$. The following theorem uses the result of Theorem 3 to obtain the convergence rate.

Theorem 5. *Suppose that α is a simple root of $f(x)$. Then the numerical algorithm outlined in equations (6), denoted by OP, has order of convergence $2^n + 2^{n-2}$, provided the method is convergent.*

4.1 An exhaustive efficiency study

It is preceptive in a new proposal to compare our methods with already existing methods exhibiting similar characteristics in order to show their efficiency. In this sense we cite a very interesting paper due to Džunić and Petković, see [6], and compare our proposed methods with theirs. In [6] the authors perform a complete study about iterative methods with memory by using Newton's interpolating polynomials, considering biparametric multipoint methods with the following form:

$$\begin{aligned} y_k^{(0)} &= x_k \\ y_k^{(1)} &= x_k + \gamma f(x_k) \\ y_k^{(2)} &= x_k - \frac{f(x_k)}{f[x_k, y_k^{(1)}] + pf(y_k^{(1)})} \\ y_k^{(\ell+1)} &= y_k^{(\ell)} - \frac{f(y_k^{(\ell)})}{P'_{k,\ell}(y_k^{(\ell)})}, \quad \ell = 1, ..., n-2 \\ x_{k+1} &= y_k^{(n-1)} - \frac{f(y_k^{(n-1)})}{P'_{k,n-1}(y_k^{(n-1)})} \end{aligned} \tag{7}$$

From this multipoint method are obtained two different families with memory. In the first one, denoted by $DP1$, it is shown that if $\gamma = \dfrac{-1}{f'(\alpha)}$, then the rate of convergence increases to $2^n + 2^{n-1}$.

The second parameter p can also be approximated by using historical data in order to increase the convergence order of the iterative method to $2^n + 2^{n-1} + 2^{n-2}$. The method obtained, denoted by $DP2$, is written as:

 ISBN: 978-84-617-8694-7

$$
\begin{aligned}
y_k^{(0)} &= x_k, \quad \gamma_k = \frac{-1}{N'_m(y_k^{(0)})} \\
y_k^{(1)} &= x_k + \gamma_k f(x_k), \quad p_k = \frac{-N''_{m+1}(y_k^{(0)})}{2N'_{m+1}(y_k^{(0)})} \\
y_k^{(2)} &= x_k - \frac{f(x_k)}{f[x_k, y_k^{(1)}] + p_k f(y_k^{(1)})} \\
y_k^{(\ell+1)} &= y_k^{(\ell)} - \frac{f(y_k^{(\ell)})}{P'_{k,\ell}(y_k^{(\ell)})}, \quad \ell = 1, ..., n-2 \\
x_{k+1} &= y_k^{(n-1)} - \frac{f(y_k^{(n-1)})}{P'_{k,n-1}(y_k^{(n-1)})}
\end{aligned} \tag{8}
$$

where $N'_{m+1}(y_k^{(0)})$ and $N''_{m+1}(y_k^{(0)})$ denote the evaluation in $y_k^{(0)}$ of the derivatives for the Newton interpolatory polynomial of degree $m+1$ based on two points of the new iteration and $m = n+1$ points of previous iteration. Obviously the iterative methods with memory, OP, $DP1$ and $DP2$, derived from the optimal iterative method without memory, ZLH, perform the same number of functional evaluations, $n+1$, so the corresponding efficiency indices, EI, increases when the order increases.

However, for high-order methods one can also take into account the computational efficiency, CE ([4]), which is defined in terms of the number of operations performed per iteration (N), specifically products and quotients, for a method with order of convergence ρ. It is obtained by $CE = \rho^{1/N}$.

In order to obtain these values we recall the number of products and quotients that the iterative methods OP, $DP1$ and $DP2$ perform per iteration. We have taken into account the information for obtaining polynomials of degree $n-1$ in order to obtain polynomials of degree n, which takes $n(n+1)/2$, plus the number of operations for computing its derivative if it is needed. Our proposed method has better computational efficiency for different values of n than methods $DP1$ and $DP2$.

For completing the efficiency study we have to analyze the total cost of computing a root, because normally high-order methods perform less iterations than a method with lower convergence order. We obtain the computational efficiency by taking into account the number of iterations performed (k) as well as the work performed per iterate (N). We define the total computational efficiency as $TCE = \rho^{\frac{1}{kN}}$.

In these terms we consider the following question: how many fewer iterations must be performed for the iterative methods $DP1$ and $DP2$ to be more efficient than OP?. We prove the following result that establish the relation between the total computational effi-

 ISBN: 978-84-617-8694-7

ciency for the methods of the previous sections.

Theorem 6. *Suppose one iterative scheme has convergence rate ρ_1, requiring k iterations for reaching the solution with tolerance tol and N_1 operations per iteration, while a second iterative scheme has convergence rate ρ_2 with $1 < \rho_1 < \rho_2$, requiring $k - i$ iterations and N_2 operations per iteration The TCE index of the second scheme is greater than the first scheme for:*

$$i < k\left(1 - \frac{N_1}{N2}\frac{\ln(\rho_1)}{\ln(\rho_2)}\right). \tag{9}$$

For the iterative methods under consideration we have that $TCE(OP) > TCE(DP1)$ if the difference between the number of iterations i is bounded for:

$$i < k\left(1 - \frac{n}{n+2}\frac{n-1+\log_2 3}{n-2+\log_2 5}\right)$$

For the highest order iterative method we have that $TCE(OP) > TCE(DP2)$ if:

$$i < k\left(1 - \frac{n(n+1)}{(n+2)(n+3)}\frac{n-2+\log_2 7}{n-2+\log_2 5}\right).$$

5 Conclusions

Numerical algorithms for approximating zeros of nonlinear functions have a rich history. The primary study of these methods has historically been focused on methods without memory, which possess an upper bound on their rate of convergence as stated by the conjecture of Kung-Traub. This bound does not apply to methods with memory, resulting in methods incorporating memory becoming a current topic of research. These methods typically utilize historical data in new ways to boost convergence rates. In this paper we aim to minimize computations while still allowing one to exceed the theoretical maximum that exists for methods without memory. In a predictor-corrector scheme, this is achieved by using the slope from the previous corrector step to generate the current predictor step. Convergence rates for various algorithms of this general form were calculated, followed by an analysis of their efficiency. Two specific cases were examined: the Midpoint method with memory ([5]) and the OP method. The OP method was compared to two other modifications of the ZLH algorithm.

Acknowledgements

2000 Mathematics Subject Classification: 47H99, 65H10, 65B05, 47H17, 49M15.
This work was supported in part by the project MTM2014-52016-C2-2-P of the Spanish

Ministry of Science and Innovation, and by the project of Generalitat Valenciana Prometeo/2016/089.

References

[1] R. L. Burden and J. D. Faires, *Numerical Analysis*, Cengage Learning, Boston, 2011.

[2] J. F. Traub, *Iterative Methods for the Solution of Equations*, Prentice-Hall, New Jersey, 1964.

[3] H. T. Kung, J. F. Traub, *Optimal Order of One-Point and Multipoint Iteration*, J. Appl. Math. Mech. **21** 4 (1974) 643–651.

[4] A. M. Ostrowski, *Solutions of Equations in Euclidean and Banach Spaces*, Academic Press, New York, 1973.

[5] T. J. McDougall, S. J. Wotherspoon, *A simple modification of Newton's method to achieve convergence of order* $1+\sqrt{2}$, Appl. Math. Lett. **29** (2014) 20–25.

[6] J. Džunić, M. S. Petković, *On generalized biparametric multipoint root finding methods with memory*, J. Comput. Appl. Math. **255** (2014) 362–375.

[7] Q. Zheng, J. Li, F. Huang, *An optimal Steffensen-type family for solving nonlinear equations*, Appl. Math. Comput. **217** (2011) 9592–9597.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

Stability of running localized waves in fluid-filled elastic membrane tubes: weakly nonlinear approach

Andrej T. Il'ichev[1]

[1] *Department of Mechanics, Steklov Mathematical Institute*

emails: ilichev@mi.ras.ru

Abstract

We examine the problem of stability of solitary waves, propagating in a fluid-filled membrane tube. We consider only waves with speeds close to those given by the linear dispersion relation (it is known that there may exist four families of solitary waves having such speeds), i. e. waves of a small (but finite) amplitude bifurcating from the quiescent state of the system. In other words we adopt weakly nonlinear description of solitary waves. It is shown that if a solitary wave speed is bounded away from zero, then the solitary wave itself is orbitally stable, either the fluid initially stationary or not.

Key words: memrane tubes, spectral stability, solitary waves
MSC 2000: AMS codes (74J30,76B25)

1 Introduction

Governing equations for quasi-one dimensional motion of the perfect fluid in an axisymmetric membrane tube were obtained in [1] by means of straightforward derivation. Study of spectral stability of a branch of steady solitary-wave solutions (so-called aneurysm solutions) in the absence of the fluid inside the tube (pressure-controlled case) is given in [2]. A bifurcation parameter was the inflation pressure, and the authors found that all family of solitary waves is always spectrally unstable (i. e. a perturbation of a wave form exponentially grows with time). In [3] stability of the whole branch aneurysm solutions is studied when the fluid inside the tube is present, but a mean flow (a constant speed of the fluid at infinity) is zero. It was found there that the aneurysm is still unstable, but the presence of fluid has a strong stabilizing effect. The authors of [4] undertook a stability analysis of

 ISBN: 978-84-617-8694-7

aneurysm solution in the presence of the mean flow and found that if a speed of the fluid at infinity is bounded away from zero, then the aneurysm is spectrally stable.

In the present paper we examine the problem of stability of solitary waves, propagating with a non-zero speed in a fluid-filled axisymmetric membrane tube. In this case a bifurcation parameter is not the inflation pressure any more (it may take arbitrary values), but a solitary wave's speed. We consider only waves with speeds close to those given by the linear dispersion relation, i. e. the waves of a small (but finite) amplitude bifurcating from the quiescent state of the system. In other words we adopt weakly nonlinear description of solitary waves.

2 Formulation of the problem

We model the tube as incompressible, isotropic, hyperelastic, cylindrical membrane. The tube has a constant undeformed radius R and a constant undeformed thickness H. The tube is assumed to be infinitely long, and end conditions are imposed at infinity. We use cylindrical coordinates, and undeformed configuration is given by coordinates R, Θ, Z.

We assume that the axisymmetry remains throughout the entire deformation; the deformed configuration is expressed using cylindrical polar coordinates r, θ, z, where $r = r(Z,t)$, $\theta = \theta(Z,t)$, $z = z(Z,t)$, and t denotes time.

The principal directions of the deformation correspond to the lines of latitude, the meridian and the normal to the deformed surface, and the principal stretches are given by

$$\lambda_1 = \frac{r}{R}, \quad \lambda_2 = (r'^2 + z'^2)^{\frac{1}{2}}, \quad \lambda_3 = \frac{h}{H}, \tag{1}$$

where the indices 1, 2, 3 are used for the circumferential, axial and radial directions respectively, a prime represents differentiation with respect to Z, and h denotes the deformed thickness.

The principal Cauchy stresses σ_1, σ_2, σ_3 in the deformed configuration for an incompressible material are given by

$$\sigma_i = \lambda_i W_i - p, \qquad i = 1,2,3 \quad \text{(no summation)},$$

where $W = W(\lambda_1, \lambda_2, \lambda_3)$ is the strain-energy function, $W_i = \partial W / \partial \lambda_i$, and p is a Lagrange multiplier, associated with the constraint of incompressibility. Utilizing the incompressibility constraint $\lambda_1\lambda_2\lambda_3 = 1$ and the membrane assumption of no stress through the thickness direction $\sigma_3 = 0$, we find [5]

$$\sigma_i = \lambda_i \hat{W}_i, \qquad i = 1,2$$

where $\hat{W}(\lambda_1, \lambda_2) = W(\lambda_1, \lambda_2, \lambda_1^{-1}\lambda_2^{-1})$ and $\hat{W}_1 = \partial \hat{W} / \partial \lambda_1$, etc.

 ISBN: 978-84-617-8694-7

As an example we give here frequently used three strain-energy functions, the Varga, Ogden and Gent materials, given respectively by,

$$W = 2\mu(\lambda_1 + \lambda_2 + \lambda_3 - 3),$$

$$W = \mu \sum_{r=1}^{3} \mu_r(\lambda_1^{\alpha_r} + \lambda_2^{\alpha_r} + \lambda_3^{\alpha_r} - 3)/\alpha_r,$$

$$W = -\frac{1}{2}\mu J_m \ln\left(1 - \frac{\lambda_1^2 + \lambda_2^2 + \lambda_3^2 - 3}{J_m}\right),$$

where μ is the shear modulus for infinitesimal deformations, $J_m > 0$ is a material constant representing the maximum stretch of the material and $\alpha_1 = 1.3$, $\alpha_2 = 5.0$, $\alpha_3 = -2.0$, $\mu_1 = 1.491$, $\mu_2 = 0.003$, $\mu_3 = -0.023$. The Ogden and Gent materials were proposed in [6] and [7] respectively, and are popularly used to model rubber.

The equations of motion for the tube can be derived from the exact field equations of general nonlinear shell theory, (see e. g. [8]), but in [1] a very readable self-contained derivation is given based on linear momentum balance applied to an infinitesimal material form of tube's wall. In terms of the Lagrangian coordinate Z and the time t the equations of motion for the fluid inside the tube were obtained in [1], (for derivation, see [3, 4]). In the dimensionless form these equations read

$$\left[\sigma_2 \frac{z'}{\lambda_2^2}\right]' - Prr' = \ddot{z}, \quad \left[\sigma_2 \frac{r'}{\lambda_2^2}\right]' - \frac{\sigma_1}{\lambda_1} + Prz' = \ddot{r}, \tag{2}$$

$$\dot{r}z' - r'\dot{z} + v_f r' + \frac{1}{2} r v_f' = 0, \quad b_f[\dot{v}_f z' - v_f'\dot{z} + v_f v_f'] + P' = 0, \tag{3}$$

where

$$z = z_\infty Z + u(Z,t), \quad r = r_\infty + w(Z,t), \quad u, w \to 0, \text{ as } Z \to \infty,$$

and P is the dimensionless internal pressure in the fluid, and the dot denotes the differentiation with respect to time; b_f is defined by

$$b_f = \frac{\rho_f R}{\rho H}.$$

where ρ_f is the fluid density and ρ is density of solid walls.

The governing equations (2) and (3) admit the uniform solution

$$r = r_\infty, \quad z' = z_\infty = \lambda_{2\infty}, v_f = v_{f\infty}, P = P_\infty \equiv \frac{W_1(c, r_\infty, \lambda_{2\infty})}{r_\infty \lambda_{2\infty}}. \tag{4}$$

 ISBN: 978-84-617-8694-7

We look for a general localized traveling wave solution for which the dependence on Z and t is through $Z - \hat{c}t$, where $\hat{c}$ denotes the wave speed of the wave. Localization means that as $Z - \hat{c}t \to \pm\infty$, the fluid-filled tube is in a uniform state given by (4).

We shall also need to refer to the dispersion relation for small-amplitude traveling waves superimposed on the uniform state (4). Assuming that the small-amplitude perturbations are proportional to

$$\exp\left(\frac{\lambda_{2\infty}}{r_\infty}k(Z-\hat{c}t)\right),$$

then the scaled wavenumber k and wave speed $c = \hat{c}\lambda_{2\infty}$ satisfy the dispersion relation [3]

$$\begin{aligned}&\left(k^2m+2\right)c^4 - 4v_{f\infty}c^3 - \left(m\alpha_0k^2 + m\gamma_1k^2 - 2v_{f\infty}^2 - m\beta_0 + m\beta_1 + 2\gamma_1\right)c^2\\ &+4v_{f\infty}\gamma_1 c - 2v_{f\infty}^2\gamma_1 + m\gamma_1(k^2\alpha_0 + \beta_1 - \beta_0) - m(\alpha_1-\beta_0)^2 = 0,\end{aligned} \tag{5}$$

where $m = 1/(b_f r_\infty^2 \lambda_{2\infty})$, and the expressions for the quantities α_0, α_1, γ_1, β_0, β_1 are given in [9] (with remarks in [4]). When $v_{f\infty} = 0$, the above equation becomes a bi-quadratic on c at the limit $k \to 0$, and the roots can be written as

$$4c^2 = 2\gamma_1 + m(\beta_1 - \beta_0) \pm \sqrt{[2\gamma_1 - m(\beta_1-\beta_0)]^2 + 8m(\alpha_1-\beta_0)^2}, \tag{6}$$

so that the four roots are all real. These characteristic speeds correspond to four branches of long waves: two of them propagate to the right, the other two propagate symmetrically to the left.

It can be easily shown [1] that the fluid equations (3) in this case can be integrated to yield

$$P = P_\infty + v_{f0}\left(1 - \frac{r_\infty^4}{r^4}\right), \quad v_f = \frac{v_{f\infty}r_\infty^2}{r^2}, \tag{7}$$

where the constant v_{f0} is defined by

$$v_{f0} = \frac{1}{2}b_f(v_{f\infty} - c)^2, \quad c = \hat{c}\lambda_{2\infty}.$$

It is also known [1, 4] that the equations in (2) together with (7) have two integrals. They are given by

$$W - \lambda_2 W_2 + \frac{1}{2}c^2\lambda_2^2 = C_1, \quad \frac{W_2 z'}{\lambda_2} - \frac{1}{2}P^* r^2 - c^2 z' = C_2, \tag{8}$$

where a prime again denotes differentiation with respect to $Z - \hat{c}t$,

$$P^* = P_\infty + v_{f0}\left(1 + \frac{\lambda_{1\infty}^4}{r^4}\right),$$

and the constants C_1 and C_2 can be determined by evaluating the corresponding left hands at the uniform state (4).

 ISBN: 978-84-617-8694-7

3 Weakly nonlinear theory of running solitary waves

We first note that the two integrals (8) are of the forms $f(\lambda_1, \lambda_2) = 0$ and $z' = g(\lambda_1, \lambda_2)$, respectively. These two equations always admit the trivial solution (4). To characterize non-trivial solutions, we write $\lambda_1 = r = r_\infty + w(Z - \hat{c}t)$ and proceed to derive a governing equation for $w(Z-\hat{c}t)$. To this end, we note that in principle we may solve the first integral to express λ_2 in terms of w. Although in general this expression cannot be obtained explicitly, a Taylor expansion of this expression valid for small w can be obtained in a straightforward manner [4]. The second integral can then be used to find the Taylor expansion of z' in terms of w as well. On substituting these expressions into (1), we then obtain

$$\left(w'\right)^2 = \omega(c, r_\infty, \lambda_{2\infty}, v_{f\infty})w^2 + \gamma(c, r_\infty, \lambda_{2\infty}, v_{f\infty})w^3 + O(w^4), \tag{9}$$

where the expression for $\omega(c, r_\infty, \lambda_{2\infty}, v_{f\infty})$ is given in [4], eqn (35), and the expression for $\gamma(c, r_\infty, \lambda_{2\infty}, v_{f\infty})$ is given by [9], eqn (3.7).

On differentiating (9), we obtain

$$w'' = \omega(c, r_\infty, \lambda_{2\infty}, v_{f\infty})\, w + \frac{3}{2}\gamma(c, r_\infty, \lambda_{2\infty}, v_{f\infty})w^2 + O(w^3). \tag{10}$$

It can then be seen that a bifurcation takes place when the coefficient $\omega(c, r_\infty, \lambda_{2\infty}, v_{f\infty})$ vanishes. Considering c as a bifurcation parameter, we further assume $v_{f\infty} = 0$, (because, as shown in [4], the mean flow stabilize the motion) and put r_∞ and $\lambda_{2\infty}^2$ arbitrary. The critical value of c, c_{cr} say, is then determined by the bifurcation condition $\omega(c_{cr}, r_\infty, \lambda_{2\infty}) = 0$, which can be reduced to (6).

On expanding the coefficients in (10) around $c = c_{cr}$, we obtain

$$V'' - V + V^2 + e(\epsilon, V) = 0, \tag{11}$$

where

$$w = \frac{2\epsilon\, \omega'_{cr}}{3\gamma_{cr}} V(\xi), \quad \xi = \sqrt{|\epsilon \omega'_{cr}|}(Z - \hat{c}t),$$

$$\omega'_{cr} = \left.\frac{d\omega}{dc}\right|_{c=c_{cr}}, \quad \gamma_{cr} = \gamma(c_{cr}, r_\infty, \lambda_{2\infty}), \quad e(\epsilon, V) = O(\epsilon),$$

$$\begin{aligned} &\varepsilon = |c_{cr} - c|, \quad \text{if } \omega'_{cr} > 0,\ \gamma_{cr} < 0, \\ &\varepsilon = -\mathrm{sgn}(\omega_{cr})|c_{cr} - c|, \quad \text{if } \gamma_{cr} > 0. \end{aligned}$$

Therefore, when $\gamma_{cr} < 0$ and $\omega'_{cr} > 0$ or $\gamma_{cr} > 0$ and $\omega'_{cr} < 0$ the solitary wave family bifurcate from the quiescent state in result of a supercritical bifurcation ($c > c_{cr}$), in the case $\gamma_{cr} > 0$ and $\omega'_{cr} > 0$ this bifurcation is a subcritical one ($c < c_{cr}$).

 ISBN: 978-84-617-8694-7

When the $O(\epsilon)$ term is neglected, equation (11) has an exact solitary wave-type solution given by

$$V = V_0 \equiv \frac{3}{2}\,\mathrm{sech}^2\frac{\xi}{2}. \tag{12}$$

Equation (11) is a special case of reversible differential equations for which general persistence results have previously been established in [10]. In other words, the solution of the full nonlinear equation (11) which is given by (12) for $\epsilon \to 0$ is close to the expression given by (12) for small enough $|\epsilon|$ and tend to zero as $\xi \to \pm\infty$, namely

$$|V - V_0| \le C\epsilon \exp(-|\xi|), \tag{13}$$

where $C > 0$ is some constant. Very good correspondence of the asymptotic (12) and exact result (which relates to a solution of (11), see [9], Fig. 1), without any doubt serves a demonstration of the persistence property (13).

4 Orbital stability of the weakly nonlinear running solitary waves

We now investigate the stability of the weakly nonlinear solitary wave (12). We assume that perturbations of the quiescent state (4) depend on the slow time variable $\tau = \epsilon^{3/2}t$, where $\epsilon = |c_{cr} - c|$, as well as on the variable ξ. We look for a perturbation solution of the form

$$\begin{aligned} &r = r_\infty + \epsilon[w_1(\xi,\tau) + \epsilon w_2(\xi,\tau) + ...],\\ &z = z_\infty Z + \sqrt{\epsilon}[u_1(\xi,\tau) + \epsilon u_2(\xi,\tau) + ...],\\ &v_f = \epsilon v_{f1}(\xi,\tau) + v_{f2} + ...,\\ &P = P_\infty + \epsilon[p_1(\xi,\tau) + \epsilon p_2(\xi,\tau) + ...]. \end{aligned} \tag{14}$$

For P_∞ given by Eq. (4) we have the Taylor expansion

$$P_\infty = \frac{W_1(c_{cr}, r_\infty, \lambda_{2\infty})}{r_\infty \lambda_{2\infty}} + \epsilon P_1 + ... \tag{15}$$

On substitution (14), (15) into Eqs. (2) and (3), and then equating the coefficients of ϵ we obtain

$$\begin{aligned} &\mathbf{L}\begin{bmatrix} w_1 \\ \sqrt{|\omega'|}u_{1\xi} \end{bmatrix} = \mathbf{0},\\ &\mathbf{L} = \begin{bmatrix} -W_1/z_\infty + W_{12} & W_{22} - z_\infty^2 {c_{cr}}^2 \\ z_\infty(W_1 - r_\infty W_{11}) - \frac{2b_f r_\infty^2}{z_\infty}c_{cr}^2 & r_\infty(W_1 - z_\infty W_{12}) \end{bmatrix}, \end{aligned} \tag{16}$$

 ISBN: 978-84-617-8694-7

where W_1, W_{11}, W_{12}, W_{22} are all evaluated at $c = c_{cr}$ and $(r, z) = (r_\infty, z_\infty)$. It is easy to see that $\omega(c_{cr}, r_\infty, z_\infty) = 0$ implies $\det \mathbf{L} = 0$, and thus the equation (16) has a nontrivial solution for w_1 and $u_{1\xi}$. Proceeding to the next order, we find

$$\mathbf{L} \begin{bmatrix} w_{2\xi} \\ \sqrt{|\omega'|} u_{2\xi\xi} \end{bmatrix} = \mathbf{b}, \tag{17}$$

$$\begin{aligned} v_{1f} &= \frac{2c_{cr}}{r_\infty} w_1, \quad p_{1\xi} = \frac{2b_f c_{cr}^2}{r_\infty} w_{1\xi}, \quad v_{2f} = \frac{2c_{cr}}{r_\infty} w_2 - \frac{6c_{cr}}{r_\infty^2} w_1 w_{1\xi} - \frac{2z_\infty}{r_\infty} \dot{w}_1, \\ p_{2\xi} &= \frac{2b_f c_{cr}^2}{r_\infty} w_{2\xi} - \frac{10 b_f c_{cr}^2}{r_\infty^2} w_1 w_{1\xi} - \frac{4 b_f c_{cr}}{r_\infty} \dot{w}_1 - \frac{4 b_f c_{cr}^2}{r_\infty z_\infty} u_{1\xi} w_{1\xi}, \end{aligned} \tag{18}$$

where after substitution of (18) the vector $\mathbf{b}$ only contains w_1 and its derivatives. On forming the dot product of the vector $\mathbf{b}$ in Eq. (17) with the left zero eigenvector of $\mathbf{L}$, we then obtain the evolution equation in the form

$$\frac{\partial w_1}{\partial \tau} - c_0 \frac{\partial w_1}{\partial \xi} + c_1 w_1 \frac{\partial w_1}{\partial \xi} + c_2 \frac{\partial^3 w_1}{\partial \xi^3} = 0, \tag{19}$$

where c_0, c_1, and c_2 are constants. Since (19) reduces to (11) when time independence is assumed, it must take the form

$$\frac{\partial \tilde{V}}{\partial \tau} - c_0 \frac{\partial \tilde{V}}{\partial \xi} + c_0 2\tilde{V} \frac{\partial \tilde{V}}{\partial \xi} + c_0 \frac{\partial^3 \tilde{V}}{\partial \xi^3} = 0, \tag{20}$$

where

$$\tilde{V} = -\mathrm{sgn}(\omega'_{cr} \gamma_{cr}) \frac{3\gamma_{cr}}{2\omega'_{cr}} w_1,$$

and the remaining constant can be found by looking for a traveling wave solution of the linearized form of (20) and then comparing the resulting dispersion relation with (5). However, for our purpose the explicit expression of c_0 is not needed.

Under the variable transformation $T = |c_0|\tau - \mathrm{sgn} c_0\, \xi$, $x = \mathrm{sgn} c_0\, \xi$, $u = 2\tilde{V}$ the equation (20) reduces to the standard form

$$\frac{\partial u}{\partial T} + u \frac{\partial u}{\partial x} + \frac{\partial^3 u}{\partial x^3} = 0. \tag{21}$$

The solution (12) of (20) corresponds to the following traveling solitary wave solution of (21)

$$u = 3\mathrm{sech}^2 \left(\frac{x - T}{2} \right). \tag{22}$$

 ISBN: 978-84-617-8694-7

A nonlinear small perturbation of a solitary wave of the KdV equation (21) (and, in particular, (22)) can yield a solitary wave with slightly different speed. Therefore it is reasonable to study the orbital stability of solitary waves, namely when a solution $u(x,T)$, which is initially close to a solitary wave $u_c(x-cT)$ in the Sobolev space of square integrable functions with their derivatives $H^1(\mathbb{R})$, will remain close to the set of translates of the solitary wave (or the orbit of the solitary wave under the group of time translations) for any time $t > 0$). This is true for solitary waves (solitons) of Eq. (21) as was established in [11] (see also [12]). In other words, it was shown that for sufficiently small ε, there exists δ, such that, if

$$||u(\cdot,0) - u_c(\cdot)||_{H^1(\mathbb{R})} < \delta,$$

then one has

$$\inf_s ||u(\cdot,T) - u_c(+s)| \cdot |_{H^1(\mathbb{R})} < \varepsilon.$$

The orbital stability property of solitary wave solutions to (21) was established in [11] (see also [12]). This definition of stability states that that a shape of the wave is stable, so it is often also called stability in form.

Acknowledgements

This work is supported by the Russian Science Foundation under grant 14-50-00005.

References

[1] M. Epstein and C. Johnston, *On the exact speed and amplitude of solitary waves in fluid-filled elastic tubes*, Proc. Roy. Soc. Lond. A. **457** (2001) 1195–1213.

[2] S.P. Pearse and Y.B. Fu, *Characterization and stability of localized bulging/necking in inflated membrane tubes*, IMA J. Appl. Math. **75** (2010) 581–602.

[3] A. T. Il'ichev and Y.-B. Fu, *Stability of aneurysm solutions in a fluid-filled elastic membrane tube*, Acta Mechanica Sinica **28** (2012) 1209-1218.

[4] Y. B. Fu and A. T. Il'ichev, *Localized standing waves in a hyperelastic membrane tube and their stabilization by a mean flow*, Math. Mech. Solids **20** (2015) 1198-2014.

[5] D. M. Haughton and R. W. Ogden, *Bifurcation of inflated circular cylinders of elastic material under axial loading. I. Membrane theory for thin-walled tubes*, J. Mech. Phys. Solids **27** (1979) 179–212.

 ISBN: 978-84-617-8694-7

[6] R. W. Ogden, *Large deformation isotropic elasticity-on the correlation of theory and experiment for incompressible rubber-like solids*, Proc. Roy. Soc. Lond. A **326** (1972) 565–584.

[7] A. N. Gent, *A new constitutive relation for rubber*, Rubber Chem. Nechnol. **69** (1996) 59–61.

[8] B. Budiansky, *Notes on nonlinear shell theory*, J. Appl. Mech. **35** (1968) 393–401.

[9] Y. B. Fu and A. T. Il'ichev, *Solitary waves in fluid-filled elastic tubes: existence, persistence, and the role of axial displacement*, IMA J. Appl. Math. **75** (2010) 257-268.

[10] G. Iooss and K. Kirchgässner, *Water waves for a small surface tension: an approach via normal form*, Proc. R. Soc. Edin. A **122** (1992) 267–299.

[11] T. B. Benjamin, *The stability of solitary waves*, Proc. Roy. Soc. Lond. A **328** (1972) 153–183.

[12] M. Grillakis, J. Shatah and W. Strauss, *Stability theory of solitary waves in the presence of symmetry, I*, J. Funct. Anal. **74** (1987) 160–197.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

A Quadrature-Difference Method for systems of second order Fredholm Integro-Differential Equations

João Janela[1], João Guerra[1] and Gilson Silva[1]

[1] *Departamento de Matemática and CEMAPRE, ISEG, Rua do Quelhas 6, 1200-781 Lisboa, Portugal, Universidade de Lisboa*

emails: jjanela@iseg.ulisboa.pt, jguerra@iseg.utl.pt, gsilva@iseg.ulisboa.pt

Abstract

In this paper we present a Quadrature-Difference Method (QDM) to solve a system of two second order Fredholm integro-differential equations with constant coefficients. We fully discretize the coupled equations using finite difference operators on the differential part of the equations and numerical quadrature formulas on the integral part. This procedure leads to two systems of linear algebraic equations, which can be solved simultaneously as a block matrix system. Moreover, we illustrate the method on a benchmark problem for which we construct the exact solution, and test its performance and accuracy, as well as the numerical order of convergence.

Key words: Coupled Systems, FIDE, Finite Difference, Direct Methods, Quadrature Formulas.

1 Introduction

Many relevant physical, biological and financial phenomena can be modelled using Integro-Differential Equations (IDEs) [17]. For instance, the use of IDEs is quite common in scattering theory, heat transfer in the presence of memory effects [9], floating structures and viscoelastic materials[18], economics [17] and finance [14]. More applications of IDEs can be found in [4]. Generally speaking, an IDE is an equation involving an unknown function u, its derivatives and integral terms associated with convolution kernels.

We focus on Fredholm integro-differential Equations (FIDEs), for which the integral terms have fixed integration limits. This class of problems has gained importance in the literature with a variety of applications [13]. For a comprehensive study on FIDEs and other types of IDEs we refer to [11].

 ISBN: 978-84-617-8694-7

Closed form analytical solution are rarely available for FIDEs and, in most cases of interest, we must resort to numerical methods. A significant amount of literature is available regarding the numerical solution of FIDEs, namely using Galerkin and Wavelet-Galerkin methods [2, 12], Tau method [15], shifted Legendre polynomials [10], Hybrid Functions [1] and Quadrature-Difference Methods [7].

In this paper we consider a Quadrature-Difference Method to solve a system of second order coupled FIDEs with constant coefficients, arising from a model in financial option pricing with memory effects [8]. The paper is organized as follows. In Section 2 we discretize the continuous problem using finite differences and numerical quadrature, obtaining a system of linear algebraic equations. In Sections 3 and 4, we analyse the numerical performance of the method, with respect to consistency and local truncation error. In Section 5, we present a toy problem for which we construct the exact solution and use it to show the accuracy of the method as well as the order of convergence. Conclusions and future perspectives are discussed in Section 6.

Preliminaries

We will consider the following system of Fredholm integro-differential equations:

$$\begin{cases} a_0u''(x) + a_1u'(x) + a_2u(x) + \lambda_0 \displaystyle\int_a^b K^0(x,z)v(z)dz = f(x), \\ b_0v''(x) + b_1v'(x) + b_2v(x) + \lambda_1 \displaystyle\int_a^b K^1(x,z)u(z)dz = g(x), \\ u(a) = u_a, u(b) = u_b, v(a) = v_a, v(b) = v_b. \end{cases} \tag{1}$$

where a_i, b_i, λ_0, λ_1 are real constants, $K^0, K^1 : [a,b]^2 \to \mathbb{R}$ and $f, g : [a,b] \to \mathbb{R}$ are given functions with suitable regularity. We search for a classical solution of the problem, $(u,v) \in (C^2(]a,b[))^2 \cap (C^0([a,b]))^2$, which we want to approximate numerically.

The existence and uniqueness of solution can be shown converting (1) to a system of first order FIDEs and using the results derived in [3]. This involves transforming our boundary value problem into an initial value problem with the same solution, which is achieved by setting up a shooting method. This requires additional regularity with respect to [3] and we refer to [19].

2 Discretization of the system of FIDEs

We use a computational grid over the interval $[a,b]$, over which our problem will be approximated. Let $x_1, \cdots, x_M$ be equally spaced points in the interval $[a,b]$, such that:

$$a = x_1, x_2, \cdots, x_M = b, \quad h = \frac{b-a}{M-1}, \quad x_{i+1} - x_i = h, \quad i = 1, ..., M. \tag{2}$$

 ISBN: 978-84-617-8694-7

At each grid point x_i the solutions $u(x_i)$ and $v(x_i)$ are approximated by $u_n(x_i)$ and $v_n(x_i)$, where the index n stand for numerically computed. In order to ease the writing, we set $u_n(x_i) = u_i$ and $v_n(x_i) = v_i$. The solution can be written as a single vector $W \in \mathbb{R}^{2M}$:

$$W = (U|V) = (u_1, u_2, \cdots, u_M, v_1, v_2, \cdots, v_M). \tag{3}$$

Approximation of the differential part

The differential parts of our system are discretized replacing the first and second order derivatives by centered finite differences on a stencil coinciding with the computational grid.

$$W'(x_i) \approx \frac{W_{i+1} - W_{i-1}}{2h}, \quad W''(x_i) \approx \frac{W_{i-1} - 2W_i + W_{i+1}}{h^2}, \quad i = 2, ..., 2M-1. \tag{4}$$

Both these approximations are second order accurate.

Approximation of the integral part

In order to obtain a full discretization of (1), the integral terms are approximated by using a quadrature rule which uses the grid points as integration nodes.

$$\int_a^b K^s(x_i, t)W(t)dt \approx \sum_{j=1}^{M} q_j K_{ij}^s W_{sM+j}, \quad s \in \{0, 1\}, \quad i = 2, ..., M-1, \tag{5}$$

where q_j are the weights of the chosen quadrature rule. In this work we make use of a composite trapezoidal rule and a composite Simpson's $\frac{1}{3}$ rule [5]. Substituting (4) and (5) in (1) we get a full discretization of the coupled system:

$$\begin{aligned} a_0\frac{W_{i+1} - 2W_i + W_{i-1}}{h^2} + a_1\frac{W_{i+1} - W_{i-1}}{2h} + a_2W_i + \lambda_0\sum_{j=1}^{M} q_j K_{ij}^1 W_{M+j} = f_i \\ , i = 2, ..., M-1, \\ b_0\frac{W_{i+1} - 2W_i + W_{i-1}}{h^2} + b_1\frac{W_{i+1} - W_{i-1}}{2h} + b_2W_i + \lambda_1\sum_{j=1}^{M} q_j K_{i-M,j}^2 W_j = g_i \\ i = M+1, ..., 2M-1. \end{aligned} \tag{6}$$

 ISBN: 978-84-617-8694-7

2.1 The System of Linear Algebraic Equations

Grouping the coefficients in (6), we obtain:

$$\begin{aligned}
&\left(\frac{a_0}{h^2}-\frac{a_1}{2h}\right)W_{i-1}+\left(-\frac{2a_0}{h^2}+a_2\right)W_i+\left(\frac{a_0}{h^2}+\frac{a_1}{2h}\right)W_{i+1}+\lambda_0\sum_{j=1}^{M}q_jK_{ij}W_{M+j}=f_i,\\
&\left(\frac{b_0}{h^2}-\frac{b_1}{2h}\right)W_{i-1}+\left(-\frac{2b_0}{h^2}+b_2\right)W_i+\left(\frac{b_0}{h^2}+\frac{b_1}{2h}\right)W_{i+1}+\lambda_1\sum_{j=1}^{M}q_jK_{i-M,j}W_j=g_i.
\end{aligned} \tag{7}$$

Let us define $Q=\mathrm{Diag}(q_1,\cdots,q_M)$, $[K^s_{ij}]=[K^s(x_i,x_j)]$, $D_s=K^sQ$, $F=(f(x_1),...,f(x_M))$, $G=(g(x_1),...,g(x_M))$, and A,B are tridiagonal matrices with line coefficients given by $\left(\frac{a_0}{h^2}-\frac{a_1}{2h},-\frac{2a_0}{h^2}+a_2,\frac{a_0}{h^2}+\frac{a_1}{2h}\right)$, and $\left(\frac{b_0}{h^2}-\frac{b_1}{2h},-\frac{2b_0}{h^2}+b_2,\frac{b_0}{h^2}+\frac{b_1}{2h}\right)$, respectively. The discretized system reduces to the following $2M\times 2M$ linear algebraic system of equations, written in matrix form:

$$\left(\begin{array}{c:c}A & D_0\\ \hdashline D_1 & B\end{array}\right)\underbrace{\left(\begin{array}{c}U\\ \hdashline V\end{array}\right)}_{W}=\left(\begin{array}{c}F\\ \hdashline G\end{array}\right) \tag{8}$$

Boundary conditions are enforced by replacing rows corresponding to boundary points, *i.e.* rows $1, M, M+$ and $2M$, by the corresponding lines in the identity matrix I_{2M}. For these lines, the RHS of the system is set to be the corresponding boundary data.

Different methods can be used for solving this system: direct methods such as Gaussian Elimination, LU-decomposition, ILU decomposition, etc; or iterative methods such as Gauss-Sidel, SOR or Krylov subspace methods. In this first approach, given the moderate dimension of the system, we used Gaussian elimination. Considering the sparsity pattern of system (8), illustrated in Fig. 1, we used sparse representations of the linear system, leading to a significant gain in memory use.

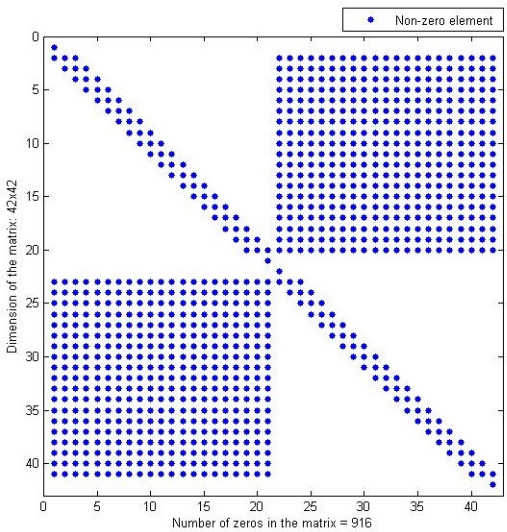
3 Consistency and Local Truncation Error

This section is dedicated to the analysis of consistency of the QDM. Consistency is a property of a discretization that ensures that the discrete equations converge to the continuous equations as the discretization parameter goes to zero ($h\to 0$). When a numerical scheme is not consistent with the continuous equations, it is probably modelling a different process and, even if convergent, is not suitable to approximate the desired equation [6]. We start by rewriting (1) splitting the differential and integral parts:

$$\left\{\begin{array}{l}\mathcal{D}^0u+\mathcal{J}^0v=f,\\ \mathcal{D}^1v+\mathcal{J}^1u=g,\end{array}\right. \tag{9}$$

 ISBN: 978-84-617-8694-7

Figure 1: The $(2N \times 2N)$ matrix block sparsity pattern of the discretized system (8). (N=21)



where

$$\mathcal{D}^0 w = a_0 w'' + a_1 w' + a_2 w, \quad \mathcal{D}^1 w = b_0 w'' + b_1 w' + b_2 w$$
$$\mathcal{J}^0 w = \lambda_0 \int_a^b K^0(x,z) w(z) dz, \quad \mathcal{J}^1 w = \lambda_1 \int_a^b K^1(x,z) w(z) dz.$$

The continuous operators $\mathcal{D}^{0,1}$ and $\mathcal{J}^{0,1}$ are approximated, at each grid point, by the discrete operators $\mathcal{D}_\Delta^{0,1}$ and $\mathcal{J}_\Delta^{0,1}$, *i.e* $\mathcal{D}^{0,1} w(x_i) \approx \mathcal{D}_\Delta^{0,1} w_i$ and $\mathcal{J}^{0,1} w(x_i) \approx \mathcal{J}_\Delta^{0,1} w_i$. The following theorem establishes the consistency of our QDM.

Theorem 1. *Let u and $v \in C^\infty([a,b])$ satisfy the boundary conditions of (1). Then, for sufficiently small $h > 0$, and for each $x_i, i = 2, \cdots, M-1$*

$$\begin{cases} \mathcal{D}^0 u(x_i) + \mathcal{J}^0 v(x_i) - \mathcal{D}_\Delta^0 u_i - \mathcal{J}_\Delta^0 v_i = \mathcal{O}(h^2) \\ \mathcal{D}^1 v(x_i) + \mathcal{J}^1 u(x_i) - \mathcal{D}_\Delta^1 v_i - \mathcal{J}_\Delta^1 u_i = \mathcal{O}(h^2) \end{cases} . \tag{10}$$

Proof. Starting with the integral operator $\mathcal{J}^0$ for the first equation on (10) and using, for instance, the composite trapezoidal rule, we have

$$J^0 v(x_i) = \int_a^b K^0(x_i, z) v(z) dz = \frac{h}{2} \left[K_{i,1}^0 v_1 + 2 \sum_{j=2}^{M-1} K_{i,j}^0 v_j + K_{i,M}^0 v_M \right] + \mathcal{O}(h^2) \tag{11}$$

Hence, the error of approximating the continuous $\mathcal{J}^0 v$ by the discretized $\mathcal{J}_\Delta^0 v$ is

$$\mathcal{J}^0 v(x_i) - \mathcal{J}_\Delta^0 v_i = \mathcal{O}(h^2). \tag{12}$$

 ISBN: 978-84-617-8694-7

The same argument can be used for the second equation in (10), leading to

$$\mathcal{J}^1 u(x_i) - \mathcal{J}^1_\Delta u_i = \mathcal{O}(h^2). \tag{13}$$

Regarding the differential operators $\mathcal{D}^{0,1}$, simple third order Taylor's expansions can be used to derive error bounds to the difference operators, yielding

$$\begin{aligned} \mathcal{D}^0 u(x_i) - \mathcal{D}^0_\Delta u_i &= \mathcal{O}(h^2) \\ \mathcal{D}^1 v(x_i) - \mathcal{D}^1_\Delta v_i &= \mathcal{O}(h^2). \end{aligned} \tag{14}$$

Finally, combining (12), (13) and (14) we obtain (10), showing the consistency of the quadrature-difference method.

□

4 Numerical Experiments

In this section we consider a benchmark problem in order to test the efficiency and accuracy of our QDM. We consider the system (1), where the given functions and constants are as follows:

$$a = 0, b = 2, a_0 = a_1 = a_2 = b_0 = b_1 = b_2 = 1, \lambda_0 = \lambda_1 = 1, \quad K^0(x,z) = K^1(x,z) = xz.$$

$$f(x) = -\frac{x}{\pi} - 2\pi \sin(2\pi x) - 4\pi^2 \cos(2\pi x) + \cos(2\pi x).$$

$$g(x) = -4\pi^2 \sin(2\pi x) + \sin(2\pi x) + 2\pi \cos(2\pi x),$$

$$u(0) = u(2) = 1, \quad v(0) = v(2) = 0.$$

The exact solution to this problem is $(u(x), v(x)) = (\cos(2\pi x), \sin(2\pi x))$. Tables 1 and 2 show a comparison between the numerical solution and the exact solution. Table 1 shows the absolute difference between QDM and the analytical solution. The maximum value is highlighted. Table 2 presents the errors measured in two different norms (l_2 and l_∞). In Figure 4, we plot in logarithmic scale the error measured in l^*_∞ and Root Mean Squared Error ($RMSE^*$), as function of the number of discretization points M, where

$$\begin{aligned} & l^*_\infty = \max\left\{\|u - u'\|_\infty, \|v - v'\|_\infty\right\} \\ & RMSE^* = \max\left\{\sqrt{\tfrac{1}{M}\sum_{i=1}^{M}\left(u(x_i) - u_i\right)^2}, \sqrt{\tfrac{1}{M}\sum_{i=1}^{M}\left(v(x_i) - v_i\right)^2}\right\} \\ & \text{end} \\ & u = \left(u(x_1), \ldots, u(x_M)\right), v = \left(v(x_1), \ldots, v(x_M)\right) \\ & u' = \left(u_1, \ldots, u_M\right), v' = \left(v_1, \ldots, v_M\right). \end{aligned} \tag{15}$$

 ISBN: 978-84-617-8694-7

Table 1: Values of the numerical and analytical solution. Absolute Error Comparison. $M = 200$ (Number of discretization points).

x_i	u_i	$u(x_i)$	$\|u(x_i) - u_i\|$	v_i	$v(x_i)$	$\|v(x_i) - v_i\|$
0	1.0000	1	0.0000	0.0000	0	0.0000
0.2	0.3131	0.3090	0.0041	0.9557	0.9511	0.0046
0.4	-0.8020	-0.8090	0.0070	0.5956	0.5878	0.0078
0.6	-0.7999	-0.8090	0.0092	-0.5782	-0.5878	0.0096
0.8	0.3197	0.3090	0.0107	-0.9405	-0.9511	0.0105
1	1.0111	1.0000	**0.0111**	0.0110	-0.0000	**0.0110**
1.2	0.3192	0.3090	0.0102	0.9616	0.9511	0.0106
1.4	-0.8009	-0.8090	0.0082	0.5967	0.5878	0.0089
1.6	-0.8033	-0.8090	0.0058	-0.5816	-0.5878	0.0062
1.8	0.3122	0.3090	0.0032	-0.9480	-0.9511	0.0031
2	1.0000	1	0.0000	0.0000	0	0.0000

Table 2: l_2 and l_∞ errors analysis

	$M = 100$	
	l_2- norm	l_∞- norm
$\|\|u(x_i) - u_i\|\|$	0.1096635	0.01115816
$\|\|v(x_i) - v_i\|\|$	0.1134598	0.01097373

Numerical estimation of the order of convergence

Here we intend to analyze the numerical convergence rate of our algorithm. Let W be the exact solution to our problem and W_h the approximate solution in a uniform grid with spacing h. The method that has order α if there is a constant K, independent of h, such that

$$|W - W_h| \leq Kh^\alpha, \tag{16}$$

for sufficiently small values of h [16]. Now, considering the sequence $h_j = \dfrac{h}{2^j}$, $j = 0, 1, \cdots, N$, the order of convergence can be estimated as

$$\alpha_j = \log_2 \frac{\|W - W_{h_j}\|_\infty}{\|W - W_{h_{j+1}}\|_\infty}, \tag{17}$$

where $\|W - W_{h_j}\|_\infty = \max\left(\|u(x_i) - u_i\|_\infty, \|v(x_i) - v_i\|_\infty\right)$. Table 3 shows that the last column of the sequence α_j is close to 2, which is consistent with second order convergence in the l_∞ norm.

 ISBN: 978-84-617-8694-7

Figure 2: Comparison of the analytical solution $u(x)$ and $v(x)$ with the numerical solution obtained with 201 discretization points. The dots represent the exact solution while the lines stands for the numerical solution

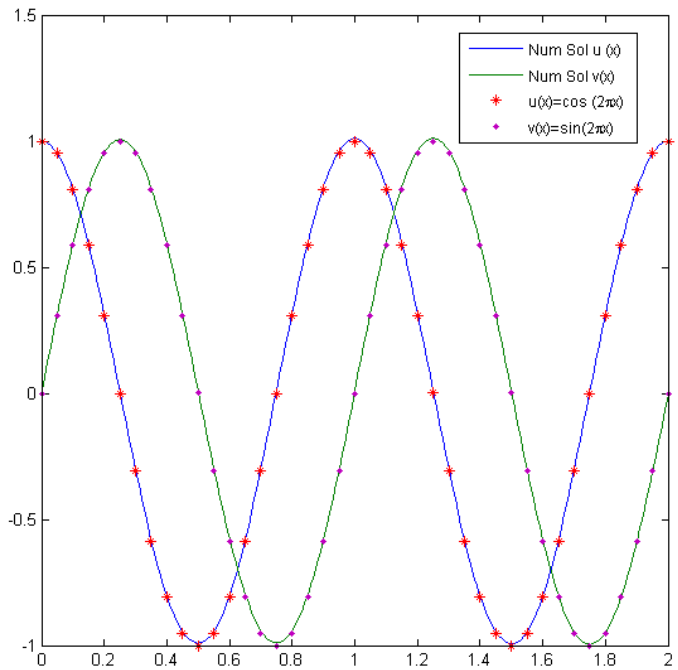


Table 3: The l_∞-norm of the errors obtained by the QDM. M is the number of points in $[a, b]$ and α is the estimated rate of convergence defined by (10).

M	h	$E_h = \|W - W_M\|_\infty$	α
16	0.125	1.82567	-
32	0.0625	4.40487×10^{-1}	2.0512
64	0.03125	1.09263×10^{-1}	2.0112
128	0.015625	2.72542×10^{-2}	2.0032
256	0.0078125	6.80971×10^{-3}	2.0008
512	0.00390625	1.70222×10^{-3}	2.0002
1024	0.00195313	4.25541×10^{-4}	2.00004
2048	0.000976563	1.06384×10^{-4}	2.00001

 ISBN: 978-84-617-8694-7

Figure 3: Absolute Error function for different values of M

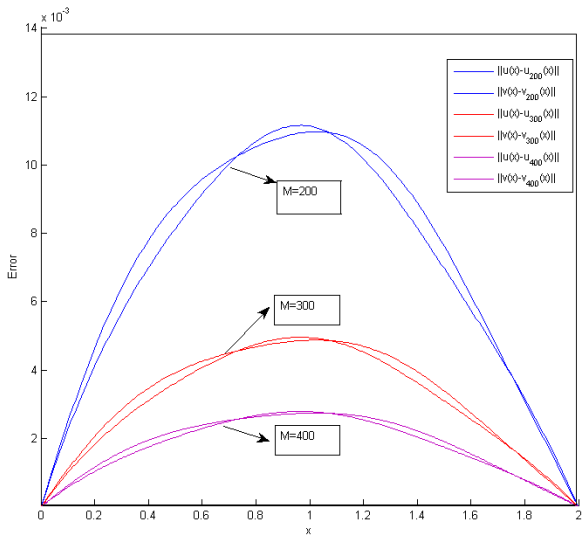
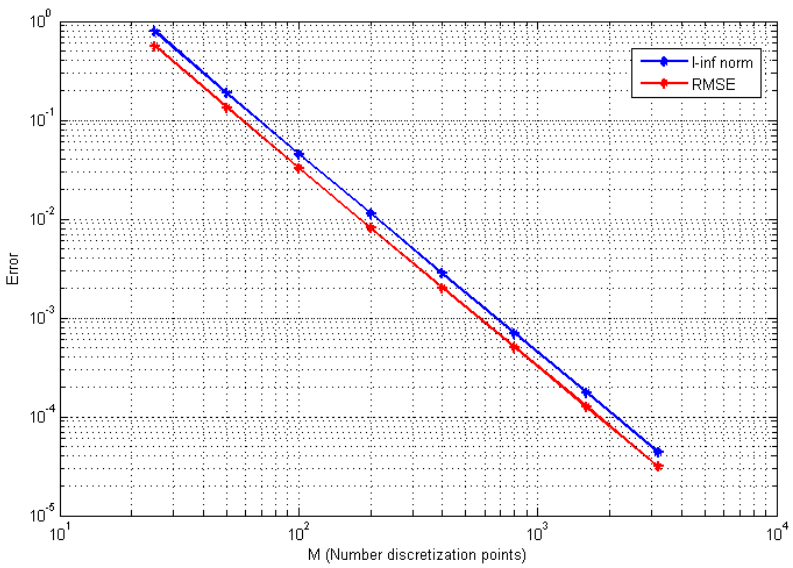


Figure 4: $l_\infty-$norm and RMSE as functions of M. The graph is plotted in logarithm scale in both axes



 ISBN: 978-84-617-8694-7

5 Conclusions and future work

This paper presents a simple method to solve numerically a class of systems of second order Fredholm integro-differential equations. A finite difference scheme and a numerical quadrature rule are used to fully discretize the system, and the resulting linear algebraic system is solved as a block matrix system. The accuracy and convergence order of the method was demonstrated in a benchmark problem. We are currently working on a direct proof of existence and uniqueness of the boundary value problem, which will require less regularity. On another direction, we are extending the procedure to a time evolution system of Fredholm integro-differential equations with possible applications to financial derivative markets [14].

Acknowledgements

The authors were partially supported by the Project CEMAPRE - UID/MULTI/00491/2013 financed by FCT/MCTES through national funds. The third author is extremely thankful to the Calouste Gulbenkian Foundation for the scholarship award to him through the program **Bolsas PALOP**.

References

[1] A. L. Ahmadieh and G. M. Eid. Using hybrid functions to solve a coupled system of fredholm integro-differential equations of the second kind. *Advances in Dynamical Systems and Applications*, 9(1):1–15, 2014.

[2] A. Avudainayagam and C. Vani. Wavelet-galerkin method for integro-differential equations. *Appl. Numer. Math.*, 32(3):247–254, 2000.

[3] M. I. Berenguer, D. Gámez, and A. J. L. Linares. Solution of systems of integro-differential equations using numerical treatment of fixed point. *Journal of Computational and Applied Mathematics*, 315:343–353, 2017.

[4] C. Constanda, D. Doty, and W. Hamill. *Boundary Integral Equation Methods and Numerical Solutions: Thin Plates on an Elastic Foundation*. Developments in Mathematics 35. Springer International Publishing, 1 edition, 2016.

[5] P. J. Davis and P. Rabinowitz. *Methods of numerical integration*. Courier Corporation, 2007.

[6] G. Evans, J. M. Blackledge, and P. Yardley. *Numerical Methods for Partial Differential Equations*. Springer Undergraduate Mathematics Series. Springer, 1999.

 ISBN: 978-84-617-8694-7

[7] A. I. Fedotov. Quadrature-difference methods for solving linear and nonlinear singular integro-differential equations. *Nonlinear Analysis: Theory, Methods & Applications*, 71(12):e303–e308, 2009.

[8] J. Guerra, J. Janela, and G. Silva. Option pricing under a jump-telegraph diffusion model with jumps of random size. In *Proceedings of International Conference on Stochastics and Computational Finance*, volume 1, page 39, Lisbon, Portugal, July 6-10 2015.

[9] H. Jorquera. Simple algorithm for solving linear integrodifferential equations with variable limits. *Computer Physics Communications*, 86(1–2):91 – 96, 1995.

[10] H. Khalil and R. A. Khan. Numerical scheme for solution of coupled system of initial value fractional order fredholm integro differential equations with smooth solutions. *Journal of Mathematical Extension*, 9:39–58, 2015.

[11] P. K. Kythe and P. Puri. *Computational Methods for Linear Integral Equations*. Birkhäuser Basel, 1 edition, 2002.

[12] K. Maleknejad and M. T. Kajani. Solving linear integro-differential equation system by galerkin methods with hybrid functions. *Applied Mathematics and Computation*, 159(3):603–612, 2004.

[13] P. K. Pandey. Numerical solution of linear fredholm integro-differential equations by non-standard finite difference method. *Applications & Applied Mathematics*, 10(2), 2015.

[14] H. Pham and P. Tankov. A coupled system of integrodifferential equations arising in liquidity risk model. *Applied Mathematics and Optimization*, 59(2):147, 2008.

[15] J. Pour-Mahmoud, M. Y. Rahimi-Ardabili, and S. Shahmorad. Numerical solution of the system of fredholm integro-differential equations by the tau method. *Applied Mathematics and Computation*, 168(1):465–478, 2005.

[16] A. Quarteroni, R. Sacco, and F. Saleri. *Numerical mathematics*, volume 37. Springer Science & Business Media, 2010.

[17] S. Sekar and C. Jaisankar. Numerical investigation of fredholm integro- differential equations by sthws method. *International Journal of Scientific Engineering Research*, 2014.

[18] S. W. Sirlin and R. W. Longman. Control of systems governed by integro-differential equations with application to floating structures and viscoelastic material dynamics. In *1982 American Control Conference*, pages 232–234, 1982.

 ISBN: 978-84-617-8694-7

[19] J. Stoer and R. Bulirsch. *Introduction to numerical analysis*, volume 12. Springer Science & Business Media, 2013.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

Null distribution approximations for a class of statistics for testing independence

M. Dolores Jiménez-Gamero[1] and M. Virtudes Alba-Fernández[2]

[1] *Department of Statistics and O.R., University of Sevilla, Spain*

[2] *Department of Statistics and O.R., University of Jaén, Spain*

emails: dolores@us.es, mvalba@ujaen.es

Abstract

A class of tests for testing independence whose test statistic is an L_2-norm of the difference between the joint empirical characteristic function and the product of the marginal empirical characteristic functions associated with a sample is considered. Since the null distribution of these test statistics is unknown, some approximations are investigated. Specifically, the permutation, bootstrap and weighted bootstrap estimators are examined. All of them provide consistent estimators. A simulation study analyzes the performance of these approximations for small and moderate sample sizes.

Key words: testing for independence, permutation, bootstrap, weighted bootstrap.

1 Introduction

Independence is a key concept in Statistics, which plays a fundamental role in many statistical procedures. Because of this reason, a number of tests have been proposed for testing the null hypothesis that two or more random vectors are independent. This paper deals with the issue of approximating the null distribution of a certain class of test statistics for such testing problem. To keep the notation as simple as possible, in our development we will only consider the case of a bivariate random vector (X, Y). Nevertheless, the methods proposed here can be applied in an obvious way to testing for the independence of any collection of subvectors from vectors with arbitrary dimensions. With this notation, the hypothesis of interest is

$$H_0 : X \text{ and } Y \text{ are independent.}$$

 ISBN: 978-84-617-8694-7

The problem of testing for independence by using the familiar equation linking the joint characteristic function (CF) and the product of component CFs has been exploited in several papers (see, for example, Csörgő [3], Kankainen and Ushakov [5], Bilodeau and Lafaye de Micheaux [1], Székely et al. [7], Meintanis and Iliopoulos [6], Hlávka et al. [4], and the references therein). Here we consider the approach in [4, 6, 7], that proposed to reject the null hypothesis for large values of

$$T_n(w) = \|\sqrt{n}\{C_n(t,s) - C_n(t,0)C_n(0,s)\}\|_w^2,$$

where C_n is the empirical CF (ECF) associated with a sample, $(X_1, Y_1), \ldots, (X_n, Y_n)$, which are independent, identically distributed (IID) from (X, Y),

$$C_n(t,s) = \frac{1}{n}\sum_{j=1}^{n} \exp(\mathrm{i}tX_j + \mathrm{i}sY_j),$$

and $\|\cdot\|_w$ stands for the norm in the separable Hilbert space $L_2(w) = \{f : \mathbb{R}^2 \to \mathbb{C} : \|f\|_w^2 = \int |f(t)|^2 w(t)dt < \infty\}$, for some nonnegative function w satisfying

$$\int w(t,s)dtds < \infty \tag{1}$$

where an unspecified integral denotes integration over $\mathbb{R}^2$ and for any complex number $x = a + \mathrm{i}b$, $|x| = \sqrt{a^2 + b^2}$.

In order to determine what are large values of $T_n(w)$, one must calculate or approximate its null distribution. Székely et al. [7] proposed to approximate the null distribution of $T_n(w)$ by a permutation estimator. Here we consider two further estimators: a bootstrap estimator and a weighted bootstrap (WB) estimator, in the sense of Burke [2]. This piece of research shows that each of these three estimators provide a consistent approximation to the null distribution of $T_n(w)$. Their finite sample behaviour is numerically compared in a simulation study.

Before ending this section we introduce some notation: all limits in this paper are taken when $n \to \infty$; $\xrightarrow{\mathcal{L}}$ denotes convergence in distribution; P_0 denotes probability under the null hypothesis; P_* denotes the conditional probability law given the data $(X_1, Y_1), \ldots, (X_n, Y_n)$.

2 Approximations

2.1 The permutation estimator

Let $\pi(1), \ldots, \pi(n)$ and $\sigma(1), \ldots, \sigma(n)$ be two independent random permutations of $1, \ldots, n$. Let $(X_1^*, Y_1^*) = (X_{\pi(1)}, Y_{\sigma(1)}), \ldots, (X_n^*, Y_n^*) = (X_{\pi(n)}, Y_{\sigma(n)})$. Let $T_n^*(w)$ be defined as $T_n(w)$ with (X_j, Y_j) replaced by (X_j^*, Y_j^*), $1 \le j \le n$. The permutation null distribution estimator of $T_n(w)$ is defined as the conditional distribution, given the data, of $T_n^*(w)$.

 ISBN: 978-84-617-8694-7

The next theorem shows that the permutation distribution of $T_n(w)$ consistently estimates its null distribution when any weight function satisfying (1) is used.

Theorem 1 *Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be a random sample from a random vector (X, Y) and suppose that (1) holds, then*

$$\sup_{x \in \mathbb{R}} |P_*\{T_n^*(w) \leq x\} - P_0\{T_n(w) \leq x\}| \to 0, \ \ a.s.$$

2.2 The bootstrap estimator

Let $X_1^\dagger, \ldots, X_n^\dagger$ be a random sample from $F_{n,X}$ and let $Y_1^\dagger, \ldots, Y_n^\dagger$ be a random sample from $F_{n,Y}$, selected independently from $X_1^\dagger, \ldots, X_n^\dagger$, where $F_{n,X}$ and $F_{n,Y}$ stand for the empirical distribution function associated to $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$, respectively. Let $T_n^\dagger(w)$ be defined as $T_n(w)$ with (X_j, Y_j) replaced by $(X_j^\dagger, Y_j^\dagger)$, $1 \leq j \leq n$.

The next theorem shows that the bootstrap distribution of $T_n(w)$ consistently estimates its null distribution when any weight function satisfying (1) is used.

Theorem 2 *Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be a random sample from a random vector (X, Y) and suppose that (1) holds, then*

$$\sup_{x \in \mathbb{R}} \left| P_*\{T_n^\dagger(w) \leq x\} - P_0\{T_n(w) \leq x\} \right| \to 0, \ \ a.s.$$

2.3 The weighted bootstrap estimator

Let $S_n(t,s) = \sqrt{n}\{C_n(t,s) - C(t,s)\}$, where C stands for the CF of (X, Y). Since $S_n \xrightarrow{\mathcal{L}} Z$ in $L_2(w)$, where Z is a zero-mean Gaussian process on $L_2(w)$, it follows that under H_0

$$\begin{aligned} \xi_n(t,s) &= \sqrt{n}\{C_n(t,s) - C_n(t,0)C_n(0,s)\} \\ &= \sqrt{n}\{C_n(t,s) \pm C(t,0)C(0,s)\} - \sqrt{n}\{C_n(t,0) \pm C(t,0)\}\{C_n(0,s) \pm C(0,s)\} \\ &= \xi_{0n}(t,s) + r_n(t,s), \end{aligned}$$

where $\|r_n\|_w^2 = o_P(1)$ and

$$\xi_{0n}(t,s) = \frac{1}{\sqrt{n}} \sum_{j=1}^{n} U_j(t,s),$$

with

$$U_j(t,s) = \exp(\mathrm{i}tX_j + \mathrm{i}sY_j) - C(t,0)\exp(\mathrm{i}sY_j) - C(0,s)\exp(\mathrm{i}tX_j) + C(t,0)C(0,s),$$

 ISBN: 978-84-617-8694-7

$1 \leq j \leq n$. Let $\varsigma_1, \ldots, \varsigma_n$ be IID random variables with mean 0 and variance 1, which are independent of the data $(X_1, Y_1), \ldots, (X_n, Y_n)$. Let

$$\xi^{\star}_{1n}(t,s) = \frac{1}{\sqrt{n}} \sum_{j=1}^{n} U_{n,j}(t,s)\varsigma_j,$$

where

$$U_{n,j}(t,s) = \exp(\mathrm{i}tX_j + \mathrm{i}sY_j) - C_n(t,0)\exp(\mathrm{i}sY_j) - C_n(0,s)\exp(\mathrm{i}tX_j) + C_n(t,0)C_n(0,s),$$

$1 \leq j \leq n$. The WB distribution estimator of $T_n(w)$ is defined as the conditional distribution, given the data, of $T^{\star}_n(w) = \|\xi^{\star}_{1n}\|^2_w$. The weighted bootstrap estimator provides a consistent approximation to the null distribution of $T_n(w)$.

The next theorem shows that the weighted bootstrap estimator provides a consistent approximation to the null distribution of $T_n(w)$.

Theorem 3 *Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be a random sample from a random vector (X, Y), suppose that (1) holds and that H_0 is true, then*

$$\sup_{x \in \mathbb{R}} |P_*\{T^{\star}_n(w) \leq x\} - P_0\{T_n(w) \leq x\}| \to 0, \ \ a.s.$$

2.4 Practical calculation

The calculation of the exact permutation (bootstrap, WB) distribution of $T_n(w)$ is, from a practical point of view, unaffordable, so the permutation (bootstrap, WB) p-value is usually approximated by simulation as follows:

1. Compute the test statistic $T_{n,obs}(w)$.
2. For some large integer B, repeat for every $b \in \{1, \ldots, B\}$:
   (a) Generate $(X^{*b}_1, Y^{*b}_1), \ldots, (X^{*b}_n, Y^{*b}_n)$ $((X^{\dagger b}_1, Y^{\dagger b}_1), \ldots, (X^{\dagger b}_n, Y^{\dagger b}_n)$, $\varsigma^b_1, \ldots, \varsigma^b_n)$.
   (b) Calculate $T^{*b}_n(w)$ $(T^{\dagger b}_n(w)$, $T^{\star b}_n(w))$,
3. Approximate the p-value of the observed value of the test statistic by

$$\hat{p}^* = \frac{1}{B} \sum_{b=1}^{B} I\{T^{*b}_n(w) > T_{n,obs}(w)\}$$

$$(\hat{p}^{\dagger} = \tfrac{1}{B}\textstyle\sum_{b=1}^{B} I\{T^{\dagger b}_n(w) > T_{n,obs}(w)\},\ \hat{p}^{\star} = \tfrac{1}{B}\textstyle\sum_{b=1}^{B} I\{T^{\star b}_n(w) > T_{n,obs}(w)\}).$$

 ISBN: 978-84-617-8694-7

3 Numerical comparisons

This section summarizes the results of an intensive simulation experiment whose objective is to investigate and compare the goodness of these approximations for small and moderate sample sizes. All computations in this paper were performed by using programs written in the R language.

The first step is to study the goodness of the bootstrap, the permutation and the WB approximations to the null distribution of the test statistic $T_n(w)$. With this aim, we generated 5000 samples with size n, for $n = 20, 30, 40, 50, 100$, from a bivariate normal distribution with independent components; the approximations of the p-value were calculated by simulation, as indicated in Subsection 2.4 with $B = 1000$. As weight function $w(t)$, several choices are possible, here we considered two of the most used in other papers related to the use of the ECF for inferential purposes: the probability density function (PDF) of a univariate normal distribution with zero mean and standard deviation $\sigma = a$ and the PDF of a univariate Laplace distribution with zero mean and variance $\sigma^2 = 2a^2$. The obtained results are shown in Table 1, that displays the actual levels for nominal values 0.05 (left column) and 0.10 (right column) for each method. The approximations of the p-value are labeled as P, B and WB for the permutation, bootstrap and WB approximations, respectively. The function w considered is indicated in the tables as N, for the normal PDF, and L, for the Laplace PDF. The same experiment were repeated for data from other populations, obtaining quite similar results. Looking at this table we see that the bootstrap and the permutation give significance levels which are quite close to the nominal values in all tried cases; the goodness of the WB approximation depends on the weight function, giving quite satisfactory results for the normal weight and rather conservative tests for the Laplace weight.

To compare the powers of the resulting tests, we repeated the above experiment for data from a bivariate normal distribution with correlated marginals with correlation coefficient $\rho = 0.2$. Table 2 displays the obtained results. Looking at this table we conclude that the bootstrap and permutation approximation behave very closely; that these two approximations with normal weight and $a = 0.5$ yield the most powerful results; as the sample size increases, the differences between the three studied approximations becomes negligible.

Finally, we also compared the three studied approximations in terms of the required time to get a p-value. Table 3 displays the average time consumed in the seconds needed to obtain each approximation to the p-value (Intel(R) Core(TM) i7-4710MQ, 2.5 Ghz). In this respect, the bootstrap approximation is a little more time consuming than the permutation one. Anyway, both approximations require much more time than the WB.

Summarizing, in the light of the numerical results, our recommendation is to use the permutation approximation for $n < 100$ and the WB for $n \geq 100$; as for the weight function w, it is advisable to use the normal PDF with standard deviation 0.5.

 ISBN: 978-84-617-8694-7

Table 1: Estimated type I error probabilities.

		$a=0.5$				$a=1$				$a=1.5$			
n		N		L		N		L		N		L	
20	P	.052	.105	.051	.105	.047	.095	.045	.100	.051	.104	.052	.106
	B	.050	.107	.044	.104	.050	.115	.048	.111	.053	.105	.053	.110
	WB	.043	.102	.040	.099	.042	.095	.029	.080	.036	.088	.020	.067
30	P	.059	.103	.056	.104	.040	.097	.048	.101	.054	.105	.054	.103
	B	.053	.108	.050	.106	.050	.108	.046	.108	.050	.105	.051	.102
	WB	.042	.084	.045	.097	.041	.088	.034	.080	.037	.086	.025	.069
40	P	.049	.096	.049	.104	.055	.109	.051	.109	.054	.111	.054	.109
	B	.053	.105	.050	.105	.048	.101	.046	.099	.049	.106	.046	.104
	WB	.047	.089	.042	.095	.044	.097	.037	.084	.036	.085	.028	.074
50	P	.048	.101	.049	.103	.056	.103	.051	.105	.052	.104	.052	.103
	B	.052	.106	.051	.107	.047	.098	.046	.099	.045	.101	.047	.098
	WB	.051	.102	.044	.097	.046	.097	.038	.086	.037	.085	.031	.076
100	P	.053	.103	.055	.106	.055	.107	.052	.106	.051	.101	.052	.101
	B	.050	.097	.047	.101	.053	.104	.053	.108	.049	.102	.051	.102
	WB	.050	.100	.042	.095	.049	.093	.048	.098	.048	.092	.036	.081

Table 2: Estimated power.

		$a=0.5$				$a=1$				$a=1.5$			
n		N		L		N		L		N		L	
20	P	.134	.221	.118	.192	.097	.178	.086	.165	.086	.157	.082	.156
	B	.134	.230	.115	.209	.105	.192	.095	.180	.075	.151	.067	.150
	WB	.099	.196	.083	.161	.068	.137	.047	.108	.040	.104	.031	.082
30	P	.163	.261	.142	.230	.123	.210	.108	.192	.089	.164	.087	.159
	B	.178	.283	.155	.247	.122	.209	.109	.191	.089	.168	.087	.163
	WB	.145	.249	.112	.208	.062	.127	.074	.140	.063	.128	.048	.108
40	P	.230	.336	.201	.295	.149	.242	.132	.215	.105	.183	.105	.179
	B	.225	.332	.187	.292	.160	.255	.141	.231	.116	.199	.111	.198
	WB	.190	.297	.156	.259	.078	.153	.094	.183	.077	.159	.057	.127
50	P	.275	.396	.231	.349	.184	.281	.158	.253	.123	.205	.118	.202
	B	.262	.376	.224	.333	.195	.291	.168	.264	.129	.213	.122	.205
	WB	.241	.354	.195	.301	.106	.189	.118	.203	.079	.153	.082	.153
100	P	.489	.604	.419	.537	.324	.445	.282	.397	.218	.322	.212	.314
	B	.476	.607	.409	.534	.319	.434	.269	.387	.214	.316	.205	.309
	WB	.455	.588	.387	.507	.301	.418	.251	.371	.187	.292	.171	.278

Table 3: Average CPU time.

	n=20			n=30			n=40			n=50			n=100		
	P	B	WB	P	B	WB	P	B	WB	P	B	WB	P	B	WB
N	.03	.04	.01	.05	.05	.01	.08	.11	.02	.12	.15	.04	.70	.77	.06
L	.03	.05	.01	.05	.07	.01	.07	.10	.01	.12	.14	.03	.70	.75	.06

 ISBN: 978-84-617-8694-7

Acknowledgements

The research in this paper has been partially funded by grants: MTM2014-55966-P of the Spanish Ministry of Economy and Competitiveness (M.D. Jiménez-Gamero) and CTM2015–68276–R of the Spanish Ministry of Economy and Competitiveness (M.V. Alba-Fernández).

References

[1] M. Bilodeau, P. Lafaye de Micheaux, *A multivariate empirical characteristic function test of independence with normal marginals*, J. Multivariate. Anal. **95** (2005) 345–369.

[2] M.D. Burke, *Multivariate tests-of-fit and uniform confidence bands using a weighted bootstrap*, Statist. Probab. Lett. **46** (2000) 13–20.

[3] S. Csörgő, *Testing for independence by the empirical characteristic function*, J. Multivariate. Anal. **16** (1985) 290–299.

[4] Z. Hlávka, M. Hušková, S.G. Meintanis, *Tests for independence in non-parametric heteroscedastic regression models*, J. Multivariate. Anal. **102** (2011) 816–827.

[5] A. Kankainen, N.G. Ushakov, *A consistent modification of a test for independence based on the empirical characteristic function*, J. Math. Sci. **89** (1998) 1486–1493.

[6] S.G. Meintanis, G. Iliopoulos, *Fourier methods for testing multivariate independence*, Comput. Stat. Data. Anal. **52** (2008) 1884–1895.

[7] G.J. Székely, M. Rizzo, N.K. Bakirov, *Measuring and Testing dependence by correlation of distances*, Ann. Stat. **35** (2007) 2769–2794.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

A Compact Splitting Scheme for Highly Oscillatory Subwavelength Metamaterials Computations

Tiffany Jones[1] **and Qin Sheng**[1]

[1] *Department of Mathematics and Center for Astrophysics, Space Physics and Engineering Research, Baylor University*

emails: Tiffany_Jones1@baylor.edu, Qin_Sheng@baylor.edu

Abstract

Recent advances in subwavelength metal optics, *e.g.* nanophotonics, metamaterials, and plasmonics, provide several new examples where nanostructured metals perform the separate tasks of absorption and charge separation necessary for solar power conversion. Nanostructured metals are extremely efficient broadband absorbers of radiation, with tailorable optical properties throughout the visible and infrared spectrum.

This preliminary report concerns a compact splitting difference method for solving Helmholtz partial differential equations in subwavelength metal in radially symmetric fields. We consider a highly accurate transverse approximation for nanostructured performance simulations. Proper auxiliary expansions are carried out, and a decomposition strategy is employed. It is proven that the highly reliable and efficient compact splitting algorithm is asymptotically stable. Simulation illustrations are given.

Key words: compact method, splitting approach, wave equations, oscillatory solutions
MSC 2000: AMS Subject Classification: 65M06, 65M12

1 Introduction

Maxwell's field equations have been playing a crucial role in modeling electromagnetic fields and subwavelength optical waves. The equations provide a ultimate theoretical backbone for numerous cutting-edge technologies such as high efficiency energy transforms. Maxwell's equations describe how electric and magnetic fields propagate and interact. Although these equations are difficult to use for immediate initial-boundary value problem computations,

 ISBN: 978-84-617-8694-7

if being decoupled properly, they yield straightforward computational models such as the following time-dependent Helmholtz equation,

$$u_{tt} = c^2 \left(u_{xx} + u_{yy} + u_{zz}\right), \quad (x, y) \in \mathcal{D}_2,\ z > 0,\ t > t_0, \tag{1.1}$$

where $u = u(x, y, z, t)$ is the intensity function of the electric field, z is the beam propagation direction, c is the speed of light, and $\mathcal{D}_2$ is the two-dimensional transverse domain. When a monochromatic light is assumed, we may define $u(x, y, z, t) = v(x, y, z)e^{2\pi i\kappa_0 t}$, where $i = \sqrt{-1}$, κ_0 is the optical wave frequency in the free space, and v is the complex wavefunction [1, 6]. Thus, from (1.1) we have the following:

$$v_{xx} + v_{yy} + v_{zz} = -\kappa^2 v, \quad (x, y) \in \mathcal{D}_2,\ z > 0,$$

where $\kappa = 2\pi\kappa_0/c > 10^2$ is the wave number of the light beam. We continue letting w be the complex envelope of v. Therefore, from the above we observe that

$$2i\kappa w_z = w_{xx} + w_{yy} + w_{zz}, \quad (x, y) \in \mathcal{D}_2,\ z > 0. \tag{1.2}$$

Explorations and investigations of fast and effective numerical methods for computing oscillatory solutions of (1.1), (1.2) with relatively small wave numbers can be found in numerous recent publications. However, the investigation of highly efficient and accurate algorithms beyond finite-difference time-domain methods for solving highly oscillatory optical beam propagation problems is still in its early stages [5, 8]. Diffraction and Fourier integral formalism is still dominating solution procedures for paraxial optical systems.

In this presentation, we are primarily interested in decomposition strategies which offer high accuracy in transverse directions while providing both computational efficiency and effectiveness. We turn our main attention to cases where radially symmetric electric fields in transverse directions are expected. The singularity that emerges in the resulting Helmholtz equation in polar coordinates is also successively removed via a decomposition strategy in the transverse direction [7, 11, 13]. Special consideration is given to the investigation of asymptotic stability of the compact splitting scheme derived.

2 Approximations in Transverse Direction

For the simplicity in discussions, we consider only monochromatic lights, and cases in which electric fields in transverse directions may be treated as radially symmetric [6, 10]. Now, if $\mathcal{D}_2$ is radially symmetric, then (1.2) can be conveniently reformulated to yield

$$w_z = \beta \left(w_{rr} + \frac{1}{r} w_r + w_{zz}\right), \quad 0 < r \leq R,\ z \geq 0, \tag{2.1}$$

 ISBN: 978-84-617-8694-7

where $z_0 > 0$ and $\beta = -i/(2\kappa)$ via a standard polar transformation. Consider a transparent boundary condition [9, 14]:

$$w_r|_{r=R} = 0, \quad z \geq 0, \tag{2.2}$$

where w_r is the outgoing normal derivative of w along the boundary of $\mathcal{D}_2$. We likewise adopt a typical Gaussian beam type initial function [1, 3],

$$w(r, 0; z_0) = \frac{e^{-r^2/[2(1+iz_0)]}}{1 + iz_0}, \quad 0 \leq r \leq R. \tag{2.3}$$

Theorem 2.1. *Let $0 < \tilde{r} \ll R$ and w be sufficiently smooth in the transverse direction. Then for $0 < r < \tilde{r}$ and $z > z_0$,*

$$\frac{1}{r}w_r(r, z) - w_{rr}(r, z) = g(w) - \frac{r^4}{30}w_{r^6}(\xi_1, z) - \frac{r^4}{24}w_{r^7}(\xi_0, z)(\xi_2 - \xi_1),$$

where

$$g(w) = -\frac{r}{2}w_{rrr}(0, z) - \frac{r^2}{3}w_{r^4}(0, z) - \frac{r^3}{8}w_{r^5}(0, z)$$

and $0 < \xi_k < \tilde{r},\ k = 0, 1, 2$.

To eliminate the singularity in (2.1) as $r \to 0$, we replace the equation by the following coupled decomposed equations:

$$w_z = \beta\left(2w_{rr} + w_{zz} + g(w)\right), \quad 0 \leq r < \tilde{r}, \tag{2.4}$$

$$w_z = \beta\left(w_{rr} + \frac{1}{r}w_r + w_{zz}\right), \quad \tilde{r} \leq r \leq R. \tag{2.5}$$

Note that the error induced is of $\mathcal{O}(r^4)$ where $0 < r < \tilde{r} \ll R$. Moreover, since the singularity has been removed, for calculations involving (2.4) we may consider a slightly extended transverse domain, $[0, \tilde{r})$.

The structures of the underlying differential equations often play a critical role in developing proper compact or essentially compact algorithms [4, 12, 13, 15]. Continuing differentiations of (2.4) yield

$$w_{rrrr} = \frac{1}{2\beta}w_{zrr} - \frac{1}{2}w_{zzrr} + \frac{1}{3}w_{r^4}(0, z) + \frac{3r}{8}w_{r^5}(0, z), \quad 0 \leq r < \tilde{r},\ z > z_0. \tag{2.6}$$

By the same token, from (2.5) we acquire

$$\begin{aligned} w_{rrrr} + \frac{2}{r}w_{rrr} &= \frac{1}{\beta}w_{zrr} + \frac{1}{\beta r}w_{zr} - w_{zzrr} - \frac{1}{r}w_{zzr} \\ &\quad + \frac{1}{r^2}w_{rr} - \frac{1}{r^3}w_r, \quad \tilde{r} \leq r \leq R,\ z > z_0. \end{aligned} \tag{2.7}$$

 ISBN: 978-84-617-8694-7

Now, let $h = R/(n-1) < \tilde{r}$ be sufficiently small, and denote $\mathcal{D}_h = \{r_k : \ r_k = (k-1)h, \ k = 1, 2, \ldots, n\}$. Since $w(-h, z) = w(h, z), \ z \geq z_0$, we derive from (2.4), (2.5) that

$$\begin{aligned}
(w_z)_k &= 2\beta \frac{w_{k+1} - 2w_k + w_{k-1}}{h^2} - \frac{\beta h^2}{3!}(w_{r^4})_k - \frac{4\beta h^4}{6!}(w_{r^6})_k \\
&\quad + \beta(w_{zz})_k + \beta g(w_0) + \mathcal{O}\left(h^6\right), \quad k = 1, 2, \ldots, \tilde{k}, \qquad (2.8) \\
(w_z)_k &= \beta \frac{w_{k+1} - 2w_k + w_{k-1}}{h^2} + \frac{\beta}{r_k}\left(\frac{w_{k+1} - w_{k-1}}{2h}\right) + \beta(w_{zz})_k \\
&\quad - \beta\left[\frac{h^2}{3!}\left(\frac{1}{2}(w_{r^4})_k + \frac{1}{r_k}(w_{r^3})_k\right) + \frac{h^4}{5!}\left(\frac{1}{3}(w_{r^6})_k\right.\right. \\
&\quad \left.\left. + \frac{1}{r_k}(w_{r^5})_k\right)\right] + \mathcal{O}\left(h^6\right), \quad k = \tilde{k}+1, \tilde{k}+2, \ldots, n; \qquad (2.9)
\end{aligned}$$

where $\tilde{k} \geq 1$ is a small integer for which $0 < \tilde{k}h < \tilde{r}$ and $(\tilde{k}+1)h \geq \tilde{r}$ [11]. Note that the values of w_0 and w_{n+1} can be determined via (2.3) and (2.2), respectively.

A substitution of (2.6), (2.7) into (2.8), (2.9) leads to

$$\begin{aligned}
(w_z)_k &= 2\beta\left(\frac{w_{k+1} - 2w_k + w_{k-1}}{h^2}\right) + \beta(w_{zz})_k - \frac{(w_z)_{k+1} - 2(w_z)_k + (w_z)_{k-1}}{12} \\
&\quad + \beta\frac{(w_{zz})_{k+1} - 2(w_{zz})_k + (w_{zz})_{k-1}}{12} + \beta\tilde{g}(w_0) + \mathcal{O}\left(h^4\right), \\
&\quad k = 1, 2, \ldots, \tilde{k}, \ z > z_0; \qquad (2.10) \\
(w_z)_k &= \beta\left(\frac{w_{k+1} - 2w_k + w_{k-1}}{h^2} + \frac{w_{k+1} - w_{k-1}}{2r_k h}\right) + \beta(w_{zz})_k \\
&\quad - \frac{\beta h}{12}\left[\frac{(w_z)_{k+1} - 2(w_z)_k + (w_z)_{k-1}}{\beta h} + \frac{(w_z)_{k+1} - (w_z)_{k-1}}{2\beta r_k}\right. \\
&\quad - \frac{(w_{zz})_{k+1} - 2(w_{zz})_k + (w_{zz})_{k-1}}{h} - \frac{(w_{zz})_{k+1} - (w_{zz})_{k-1}}{2r_k} \\
&\quad \left. + \frac{w_{k+1} - 2w_k + w_{k-1}}{r_k^2\, h} - \frac{w_{k+1} - w_{k-1}}{2r_k^3}\right] + \mathcal{O}\left(h^4\right), \\
&\quad k = \tilde{k}+1, \tilde{k}+2, \ldots, n, \ z > z_0, \qquad (2.11)
\end{aligned}$$

where

$$\tilde{g}(w_0) = g(w) - \frac{h^2}{18}w_{r^4}(0, z) - \frac{h^2 r_k}{16}w_{r^5}(0, z).$$

 ISBN: 978-84-617-8694-7

3 Approximations in Propagation Direction

Adopt a standard central difference approximation for the z-derivatives, from (2.10), (2.11) we have

$$\begin{aligned}
\frac{w_k^{j+1}-w_k^{j-1}}{2\tau} &= 2\beta\left(1+\frac{h^2}{12\tau^2}\right)\left[\frac{w_{k+1}^j-2w_k^j+w_{k-1}^j}{h^2}\right]+\beta\left[\frac{w_k^{j+1}-2w_k^j+w_k^{j-1}}{\tau^2}\right]\\
&\quad-\frac{1}{12}\left[\frac{w_{k+1}^{j+1}-2w_k^{j+1}+w_{k-1}^{j+1}-w_{k+1}^{j-1}+2w_k^{j-1}-w_{k-1}^{j-1}}{2\tau}\right]\\
&\quad+\frac{\beta}{12}\left[\frac{w_{k+1}^{j+1}-2w_k^{j+1}+w_{k-1}^{j+1}+w_{k+1}^{j-1}-2w_k^{j-1}+w_{k-1}^{j-1}}{\tau^2}\right]\\
&\quad+\beta\tilde{g}\left(w_0^j\right)+\mathcal{O}\left(h^4+\tau^2\right),\quad k\in\theta_0; \qquad (3.1)\\
\frac{w_k^{j+1}-w_k^{j-1}}{2\tau} &= \beta\left(1-\frac{h^2}{12r_k^2}\right)\left[\frac{w_{k+1}^j-2w_k^j+w_{k-1}^j}{h^2}\right]+\beta\left[\frac{w_k^{j+1}-2w_k^j+w_k^{j-1}}{\tau^2}\right]\\
&\quad+\frac{\beta}{r_k}\left(1+\frac{h^2}{12r_k^2}\right)\left[\frac{w_{k+1}^j-w_{k-1}^j}{2h}\right]-\frac{h}{24r_k}\left[\frac{w_{k+1}^{j+1}-w_{k-1}^{j+1}-w_{k+1}^{j-1}+w_{k-1}^{j-1}}{2\tau}\right]\\
&\quad-\frac{1}{12}\left[\frac{w_{k+1}^{j+1}-2w_k^{j+1}+w_{k-1}^{j+1}-w_{k+1}^{j-1}+2w_k^{j-1}-w_{k-1}^{j-1}}{2\tau}\right]\\
&\quad+\frac{\beta}{12}\left[\frac{w_{k+1}^{j+1}-2w_k^{j+1}+w_{k-1}^{j+1}-2w_{k+1}^j+4w_k^j-2w_{k-1}^j+w_{k+1}^{j-1}-2w_k^{j-1}+w_{k-1}^{j-1}}{\tau^2}\right]\\
&\quad+\frac{\beta h}{24r_k}\left[\frac{w_{k+1}^{j+1}-w_{k-1}^{j+1}-2w_{k+1}^j+2w_{k-1}^j+w_{k+1}^{j-1}-w_{k-1}^{j-1}}{\tau^2}\right]\\
&\quad+\mathcal{O}\left(h^4+\tau^2\right),\quad k\in\theta_1, \qquad (3.2)
\end{aligned}$$

where $\theta_0=\left\{1,\ldots,\tilde{k}\right\}$, $\theta_1=\left\{\tilde{k}+1,\tilde{k}+2,\ldots,n\right\}$ and $p=p^j$ and τ is the variable z-step.

Denote

$$\sigma=\frac{\tau}{h^2},\ \lambda=24\beta\sigma,\ \mu_k^{\pm}=2\pm\frac{h}{r_k},\ \nu_k^{\pm}=1\pm\frac{h^2}{12r_k^2},\ \eta^{\pm}=1\pm\frac{2\beta}{\tau}.$$

In an effort to organize the scheme in a convenient manner, drop all truncation errors from (3.1), (3.2) and rearrange terms. We consequently arrive at the following *compact splitting algorithm* for the numerical solution of the subwavelength energy transfering problem (2.2)-(2.5),

$$Aw^{j+1}=Bw^j+Cw^{j-1}+p^{\,j},\quad j=1,2,3,\ldots, \qquad (3.3)$$

 ISBN: 978-84-617-8694-7

where $p, w \in \mathbb{C}^n$ and $A, B, C \in \mathbb{C}^{n\times n}$ are tridiagonal,

$$A = \begin{bmatrix} b_1 & (a_1+c_1) & & \\ a_2 & b_2 & c_2 & \\ \ddots & \ddots & \ddots & \\ & a_{n-1} & b_{n-1} & c_{n-1} \\ & & (a_n+c_n) & b_n \end{bmatrix}, \tag{3.4}$$

$$B = \begin{bmatrix} e_1 & (d_1+f_1) & & \\ d_2 & e_2 & f_2 & \\ \ddots & \ddots & \ddots & \\ & d_{n-1} & e_{n-1} & f_{n-1} \\ & & (d_n+f_n) & e_n \end{bmatrix}, \tag{3.5}$$

$$C = \begin{bmatrix} m_1 & (l_1+n_1) & & \\ l_2 & m_2 & n_2 & \\ \ddots & \ddots & \ddots & \\ & l_{n-1} & m_{n-1} & n_{n-1} \\ & & (l_n+n_n) & m_n \end{bmatrix}, \tag{3.6}$$

where

$$a_k = \begin{cases} \eta^-, & k\in\theta_0, \\ \mu_k^-\eta^-/2, & k\in\theta_1; \end{cases} \qquad b_k = \begin{cases} 10\eta^-, & k\in\theta_0, \\ 10\eta^-, & k\in\theta_1; \end{cases}$$

$$c_k = \begin{cases} \eta^-, & k\in\theta_0, \\ \mu_k^+\eta^-/2, & k\in\theta_1; \end{cases} \qquad d_k = \begin{cases} 4\beta\left(12\sigma - 1/\tau\right), & k\in\theta_0, \\ \lambda\nu_k^- - \dfrac{\lambda h}{2r_k}\nu_k^+ - \dfrac{2\beta}{\tau}\mu_k^-, & k\in\theta_1; \end{cases}$$

$$e_k = \begin{cases} -8\beta\left(12\sigma + 5/\tau\right), & k\in\theta_0, \\ -2\left(\lambda\nu_k^- + 20\beta/\tau\right), & k\in\theta_1; \end{cases} \qquad f_k = \begin{cases} 4\beta\left(12\sigma - 1/\tau\right), & k\in\theta_0, \\ \lambda\nu_k^- + \dfrac{\lambda h}{2r_k}\nu_k^+ - \dfrac{2\beta}{\tau}\mu_k^+, & k\in\theta_1; \end{cases}$$

$$g_k^j = \begin{cases} \tilde{g}(E_0^j), & k\in\theta_0, \\ 0, & k\in\theta_1; \end{cases} \qquad l_k = \begin{cases} \eta^+, & k\in\theta_0, \\ \mu_k^-\eta^+/2, & k\in\theta_1; \end{cases}$$

$$m_k = \begin{cases} 10\eta^+, & k\in\theta_0, \\ 10\eta^+, & k\in\theta_1; \end{cases} \qquad n_k = \begin{cases} \eta^+, & k\in\theta_0, \\ \mu_k^+\eta^+/2, & k\in\theta_1. \end{cases}$$

4 Asymptotic Stability

Consider an evolutional finite difference method with an amplification matrix M for solving an oscillatory problem associated with a high wave number κ. We say that the numerical

 ISBN: 978-84-617-8694-7

method is asymptotically stable if there exists a constant $d \in \mathbb{R}^+$ such that

$$\|M\| < 1 + \mathcal{O}\left(\kappa^{-d}\right), \quad \kappa \gg 1,$$

where $\|\cdot\|$ is an appropriate matrix norm and the constant $d > 0$ is the asymptotical stability index of the method [12, 14].

Theorem 4.1. *The compact splitting scheme* (3.3) *is unconditionally asymptotically stable with the index* $d = 1$.

Proof. Without loss of generality, we begin with our perturbation equation

$$\epsilon^{\,j+1} = A^{-1}B\epsilon^{\,j} + A^{-1}C\epsilon^{\,j-1}.$$

It follows that

$$\left\|\epsilon^{\,j+1}\right\| \;\leq\; \left(\left\|A^{-1}B\right\| + \left\|A^{-1}C\right\|\right)\left\|\epsilon^{\,j}\right\|.$$

Thus for the amplification matrix M, we have $\|M\| \leq \left\|A^{-1}B\right\| + \left\|A^{-1}C\right\|$.

Define

$$\tilde{A} = \begin{bmatrix}
10 & 1 & & & & & & \\
1 & 10 & 1 & & & & & \\
 & \ddots & \ddots & \ddots & & & & \\
 & & 1 & 10 & 1 & & & \leftarrow \left(\tilde{k}\right)^{th} \text{ row} \\
 & & & 1-\dfrac{h}{2r_{\tilde{k}+1}} & 10 & 1+\dfrac{h}{2r_{\tilde{k}+1}} & & \leftarrow \left(\tilde{k}+1\right)^{st} \text{ row} \\
 & & & & \ddots & \ddots & \ddots & \\
 & & & & 1-\dfrac{h}{2r_{n-1}} & 10 & 1+\dfrac{h}{2r_{n-1}} & \\
 & & & & & 1-\dfrac{h}{2r_n} & 10 &
\end{bmatrix}, \quad (4.1)$$

It is not difficult to see that

$$\left\|A^{-1}C\right\| = 1, \quad \left\|A^{-1}B\right\| \leq \left|\frac{1}{1+\frac{i}{\kappa\tau}}\right| \left|-\frac{i}{2\kappa}\right| \left\|\tilde{A}^{-1}\right\| \left\|\tilde{B}\right\|.$$

Since $\left|\left(1+\dfrac{i}{\kappa\tau}\right)^{-1}\right| < 1$, we conclude that

$$\left\|A^{-1}B\right\| < \frac{1}{2\kappa}\left\|\tilde{A}^{-1}\right\| \left\|\tilde{B}\right\| = \mathcal{O}\left(1/\kappa\right).$$

 ISBN: 978-84-617-8694-7

Thus

$$\|M\| \leq \left\|A^{-1}B\right\| + \left\|A^{-1}C\right\| < 1 + \mathcal{O}\left(1/\kappa\right), \quad \kappa \gg 1.$$

This completes our proof. ∎

The above proof follows for any desired matrix norm.

5 Example and Conclusion

Consider the problem (2.1)-(2.3) with $\kappa = 10^4$ and $z_0 = 100$. For a CFL number 0.01, we simulate the numerical solution w at the wave focusing location in Figures 5.1 and 5.2. The vector generated 3D figures clearly demonstrate the superior numerical convergence as suggested by the asymptotic stability of the compact splitting algorithm. The computational experiments are carried out on a high performance HP C3000BL HPC cluster system running CentOS 5. The cluster is comprised of 128 computer nodes, each with 16GB of RAM and dual quad-core Intel 2.6GHz processors giving a total of 1024 cores. An Infiniband ConnectX DDR network is used for message passing and networked storage. Shared storage capacity in the cluster is 123TB.

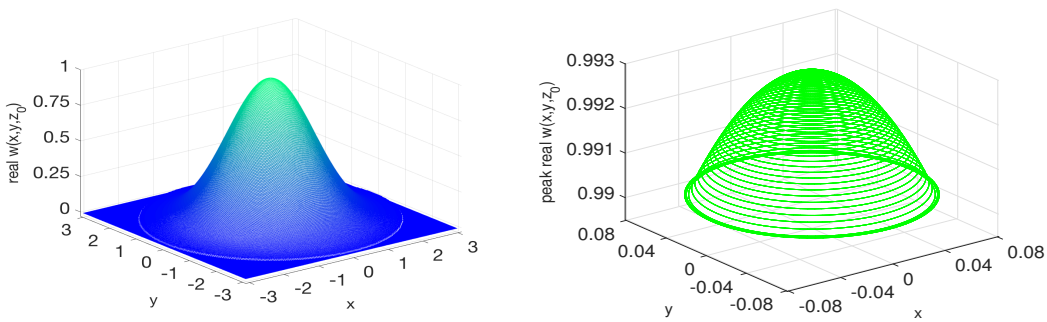


Fig. 5.1. Three-dimensional views of the real part of the focusing numerical solution $w(x, y, z_0) = w(r, z_0)$ (LEFT) and its enlarged inner solution (RIGHT).

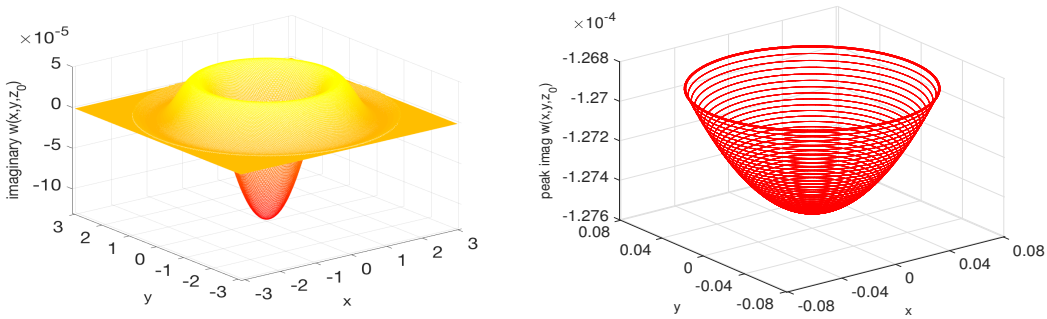


Fig. 5.2. Three-dimensional views of the imaginary part of the focusing numerical solution $w(x, y, z_0) = w(r, z_0)$ (LEFT) and its enlarged inner solution (RIGHT).

 ISBN: 978-84-617-8694-7

In summary, a splitting method for solving Helmholtz partial differential equations in subwavelength metal is developed on decomposed radially symmetric transverse fields successfully. The compact structures of the numerical method ensures higher accuracy and simplicity in nanostructured applications. Proper auxiliary expansions are utilized and discussed on decomposed domains. It is proven that the highly reliable and efficient compact splitting algorithm is asymptotically stable. The corresponding algorithms have been intensively used for numerous computational projects involving subwavelength metal optics and applications.

References

[1] Y. B. Band, *Light and Matter: Electromagnetism, Optics, Spectroscopy and Lasers,* John Wiley & Sons, West Sussex, 2006.

[2] M. A. Beauregard and Q. Sheng, *A fully adaptive method to approximate reaction-diffusion equations of the quenching type over circular domains,* Numer. Meth. Partial Diff. Equs., **30**, (2014) 472–489.

[3] P. P Crooker, W. B. Colson and J. Blau, *Representation of a Gaussian beam by rays,* Amer. J. Phys., **74**, (2006) 722–727.

[4] W. E and J.-G. Liu, *Essentially compact schemes for unsteady viscous incompressible flows,* rem J. Comput. Phys., **126**, (1996) 122–138.

[5] B. Engquist, A. Fokas, E. Hairer and A. Iserles, *Highly Oscillatory Problems*, London Math Society, London, 2009.

[6] S. Guha, *Validity of the paraxial approximation in the focal region of a small-f -number lens*, Optical Lett., **26**, (2001) 1598–1600.

[7] T. Jones, L. P. Gonzalez, S. Guha and Q. Sheng, *A continuing exploration of a decomposed compact method for highly oscillatory wave problems,* J. Comp. Appl. Math., **299**, (2016) 207–220.

[8] A. C. Newell and J. V. Moloney, *Nonlinear Optics,* Addison-Wesley Pub. Comp., New York, 1992.

[9] D. Ruprecht, A. Schädle, F. Schmidt and L. Zschiedrich, *Transparent boundary conditions for time-dependent problems*, SIAM J. Sci. Comput., **30**, (2008), 2358–2385.

[10] M. N. O. Sadiku, *Numerical Techniques in Electromagnetics,* CRC Press, London and New York, 2000.

 ISBN: 978-84-617-8694-7

[11] Q. Sheng, *Adaptive decomposition finite difference methods for solving singular problems*, Frontiers Math. China, **4**, (2009) 599–626.

[12] Q. Sheng, *ADI, LOD and modern decomposition methods for certain multiphysics applications*, J. Algorithms Comput. Tech., **9**, (2015) 105–120.

[13] Q. Sheng, *The ADI Method, Encyclopedia of Applied and Computational Mathematics,* Editor-in-Chief: Bjorn Engquist, Springer Verlag GmbH, Heidelberg, 2015.

[14] Q. Sheng, S. Guha and L. P. Gonzalez, *A short note on the asymptotic stability of certain oscillation-free eikonal splitting schemes*, Appl. Math. Lett., **25**, (2012) 1539–1543.

[15] T. W. H. Sheu, L. W. Hsieh and C. F. Chen, *Development of a three-point sixth-order Helmholtz scheme,* J. Comput. Acoustics, **16**, (2008) 343–359.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

Smooth Cubic Pythagorean Hodograph Splines

Kryštof Kadlec[1] **and Zbyněk Šír**[1]

[1] *Faculty of Mathematics and Physics, Charles University, Prague*

emails: kadlec.krystof@gmail.com, zbynek.sir@mff.cuni.cz

Abstract

In this paper we describe how to control the transformations, reparameterizations and continuity of planar cubic Pythagorean Hodograph splines. More precisely we show how these features can be ensured by conditions on the curve preimage.

Key words: PH curve, spline, signed curvature, rotation

1 Introduction and preliminaries

Pythagorean hodograph (PH) curves (see [1, 3, 2, 4, 5] and the references cited therein), form a remarkable subclass of polynomial parametric curves. They have a piecewise polynomial arc length function and, in the planar case, rational offset curves. These curves provide an elegant solution of various difficult problems occurring in applications, in particular in the context of CNC (computer-numerical-control) machining. Our paper is devoted to the modification of the planar PH curves via their preimage and on the connection of two PH cubics in a smooth way.

A Bézier curve is called *Pythagorean Hodograph (PH)* if the length of its tangent vector, taken in the appropriate metric, depends in a polynomial way on the parameter. In particular in the planar case $\mathbf{p}(t) = [x(t), y(t)]$ is called *planar PH curve* if there exists a polynomial $\sigma(t)$ such that

$$x'(t)^2 + y'(t)^2 = \sigma^2(t). \tag{1}$$

The degree of $\sigma(t)$ equals $n-1$, where n is the degree of the PH curve. The curve $\mathbf{h}(t) = [x'(t), y'(t)]$ is called the *hodograph* of $\mathbf{p}(t)$.

The planar polynomial curve $\mathbf{p}(t)$ can be identified with complex valued polynomial $\mathbf{p}(t) = x(t) + iy(t)$. The hodograph $\mathbf{h}(t) = x'(t) + iy'(t)$ then satisfy the equation (1) if and

 ISBN: 978-84-617-8694-7

only if it is of the form $\mathbf{h}(t) = \lambda(t)\mathbf{w}(t)^2$, where $\mathbf{w}(t) = u(t) + iv(t)$ is a complex valued polynomial called *preimage*, [1, 6].

In order to study the $\mathcal{G}^2$ continuity we will need the notion of the signed curvature.

Definition 1.1 Let $\mathbf{c} : I \to \mathbb{R}^2$ be a regular parameterized curve. We define its *s*igned curvature at the point $t \in I$ by the formula

$$\kappa_z(t) = \frac{\det(c'(t),\, c''(t))}{\|\|c'(t)\|\|^3}. \tag{2}$$

The signed curvature is not changed by the orientation-preserving reparameterizations.

2 Controling PH curves via their preimage

In this section we will show how a PH curve can be transformed and reparameterized via its preimage. We will also provide the formula for the signed curvature based on the preimage.

Lemma 2.1 Let $\mathbf{p} : I \to \mathbb{R}^2$ be a PH curve nad $\mathbf{q} : I \to \mathbb{R}^2$ its preimage. Then $\mathbf{p}$ is regular if and only if $\forall t \in I : \mathbf{q}(t) \neq \mathbf{0}$.

Proof: Let the complex preimage is of the form $\mathbf{q}(t) = a(t) + ib(t)$. Suppose

$$\forall t \in I : \mathbf{q}(t) \neq \mathbf{0} \Leftrightarrow \forall t \in I : a(t) \neq 0 \vee b(t) \neq 0.$$

Thus for

$$\mathbf{h}(t) = q^2(t) = a^2(t) - b^2(t) + i2a(t)b(t)$$

we get $\forall t \in I : \mathbf{h}(t) \neq (0,0)$, because if for some $t_0 \in I : a(t_0) = b(t_0) \neq 0$, then $a(t_0)b(t_0) \neq 0$. If on the other hand for some $t_1 \in I : a(t_1)b(t_1) = 0$, then $a^2(t_1) - b^2(t_1) \neq 0$. As $\mathbf{p}'(t) = \mathbf{h}(t) \neq \mathbf{0} t \in I$, we obtain the regularity.
For the other implication consider the curve $\mathbf{p}$ with the hodograph

$$\mathbf{h}(t) = a^2(t) - b^2(t) + i2a(t)b(t),$$

which due to the regularity satisfies $\mathbf{h}(t) \neq \mathbf{0}, t \in I$. Then

$$\forall t \in I : a^2(t) - b^2(t) \neq 0 \vee 2a(t)b(t) \neq 0 \Leftrightarrow \forall t \in I : a(t) \neq 0 \vee b(t) \neq 0.$$

Thus we get

$$\forall t \in I : \mathbf{q}(t) = a(t) + ib(t) \neq \mathbf{0}.$$

□

Lemma 2.2 If the peimage is rotated clockwise by an angle α then the resulting PH curve is rotated by the angle 2α.

 ISBN: 978-84-617-8694-7

Proof: Rotation by the angle α can be realized by the multiplication with the complex unit $\cos\alpha + i\sin\alpha$. Suppose that the starting preimage is $\mathbf{q} = a + ib$ (we omit the parameter t on which the functions a, b depend) and obtain the rotated preimage

$$\begin{aligned}\mathbf{q}_r &= (a+ib)(\cos\alpha + i\sin\alpha) = \\ &= (a\cos\alpha - b\sin\alpha) + i(a\sin\alpha + b\cos\alpha) = c + id.\end{aligned}$$

Taking its square we obtain the rotated hodograph

$$\begin{aligned}&\mathbf{h}_r = (c^2 - d^2) + i(2cd) = \\ &= a^2\cos^2\alpha - 2ab\cos\alpha\sin\alpha + b^2\sin^2\alpha - a^2\sin^2\alpha - 2ab\sin\alpha\cos\alpha - \\ &- b^2\cos^2\alpha + i2(a^2\sin\alpha\cos\alpha + ab\cos^2\alpha - ab\sin^2\alpha - b^2\sin\alpha\cos\alpha) = \\ &= a^2\cos 2\alpha - 2ab\sin 2\alpha - b^2\cos 2\alpha + i(a^2\sin 2\alpha - b^2\sin 2\alpha + 2ab\cos(2\alpha) \\ &= ((a^2 - b^2) + i(2ab))(\cos 2\alpha + i\sin 2\alpha) = \mathbf{h}(\cos 2\alpha + i\sin 2\alpha).\end{aligned}$$

We thus obtain the starting hodograph rotated by the angle 2α. The same holds for the PH curve, as the integration commutes with the rotation. □

Translation can be realized by the integration constant. We have seen, that the rotation is obtained via rotating the preimage. The following lemma shows that the linear reparameterization of the preimage provides a scaled PH curve.

Lemma 2.3 Let the PH curve $\mathbf{p}$ has the r $\mathbf{q}$ and let $k \neq 0, l \in \mathbb{R}$. Then the linearly reparameterized preimage $\mathbf{q}(kt + l)$ provides the PH curve $\frac{1}{k}\mathbf{p}(kt + l)$.

Proof: We obtain this result by a direct use of the substitution in the integral. □

The following result is straightforward, too.

Lemma 2.4 Let $\mathbf{q}$ is the preimage of the PH curve $\mathbf{p}$. The $k\mathbf{q}, k \in \mathbb{R}$, is the preimage of the curve $k^2\mathbf{p}$.

Combining the two previous observation we get the following method for obtaining the pure reparameterization of the curve.

Proposition 2.5 Let $\mathbf{q}(t)$ be the preimage of the PH curve $\mathbf{p}(t)$ and k, $l \in \mathbb{R}, k > 0$. Then $\sqrt{k} \cdot \mathbf{q}(kt + l)$ is the preimage of the PH curve $\mathbf{p}(kt + l)$.

The signed curvature of the PH curve can be obtained from its preimage in the following way.

Proposition 2.6 Let a PH kivka
$\mathbf{p} : I \to \mathbb{R}^2$ has the preimage $\mathbf{q}$. Then the signed curvature of $\mathbf{p}$ can be expressed as

$$\kappa_z(t) = 2\frac{\mathrm{Im}(\bar{\mathbf{q}}(t)\mathbf{q}'(t))}{|\mathbf{q}(t)|^4}. \tag{3}$$

 ISBN: 978-84-617-8694-7

Proof: This formula can be found at [1] without the proof. Let us denote $\mathbf{h} = \mathbf{p}' = (x', y')$, $\mathbf{h} = \mathbf{q}^2$, thus $\mathbf{h}' = 2\mathbf{q}\mathbf{q}' = (x'', y'')$. using the formula (2) we get

$$\kappa_z = \frac{\det(\mathbf{p}',\, \mathbf{p}'')}{\|\mathbf{p}\|^3} = \frac{x'y'' - x''y'}{(x'^2 + y'^2)^{3/2}}.$$

Moreover

$$2\frac{\mathrm{Im}(\bar{\mathbf{q}}\mathbf{q}')}{|\mathbf{q}|^4} = \frac{\mathrm{Im}(2\mathbf{q}\mathbf{q}'\bar{\mathbf{q}}^2)}{|\mathbf{q}|^6} = \frac{\mathrm{Im}(\mathbf{h}'\bar{\mathbf{h}})}{|\mathbf{h}|^3} =$$
$$= \frac{\mathrm{Im}((x'' + iy'')(x' - iy'))}{(x'^2 + y'^2)^{3/2}} = \frac{x'y'' - x''y'}{(x'^2 + y'^2)^{3/2}} = \kappa_z.$$

□

3 Smooth joints of PH cubics

In the reminder of the paper we will restrict our investigations to the cubic PH curves. In particular we will consider the joint between two cubic PH curves and the sufficient end necessary conditions for the joint to be smooth. We will suppose that the curves are connected in the $\mathcal{C}^0$ way, which can be achieved by setting suitably the integration constants. We will use the following notation: Consider two PH cubics $\mathbf{p}_1$ a $\mathbf{p}_2$ with the linear preimages $\mathbf{q}_1$ and $\mathbf{q}_2$ of the form

$$\begin{gathered}\mathbf{q}_1(t) = \mathbf{w}_0(1-t) + \mathbf{w}_1 t, \mathbf{q}_2(t) = z_0(1-t) + z_1 t, \\ \mathbf{w}_0,\, \mathbf{w}_1,\, \mathbf{z}_0,\, \mathbf{z}_1 \in \mathbb{C}, \mathbf{w}_0 = w_{01} + iw_{02},\, \mathbf{w}_1 = w_{11} + iw_{12}.\end{gathered} \tag{4}$$

The following proposition gives the conditions for the $\mathcal{G}^1$ and $\mathcal{C}^1$ continuity.

Proposition 3.1 Let $\mathbf{p}_1$, $\mathbf{p}_2$ are PH cubics with the preimages of the form (4). The curves $\mathbf{p}_1$, $\mathbf{p}_2$ are connected in the $\mathcal{G}^1$ way if and only if $\mathbf{z}_0 = c\mathbf{w}_1, c \in \mathbb{R}, c \neq 0$.
In particular the curves are connected in the $\mathcal{C}^1$ way if and only if $c = \pm 1$.

Proof: Let $\mathbf{h}_1 = \mathbf{p}_1'$, $\mathbf{h}_2 = \mathbf{p}_2'$.
First suppose $\mathbf{z}_0 = c\mathbf{w}_1, c \in \mathbb{R}, c \neq 0$. Then

$$\mathbf{h}_1(t) = \mathbf{w}_0^2(1-t)^2 + \mathbf{w}_0\mathbf{w}_1 2t(1-t) + \mathbf{w}_1^2(1-t)^2,$$
$$\mathbf{h}_2(t) = c^2\mathbf{w}_1^2(1-t)^2 + c\mathbf{w}_1\mathbf{z}_1 2t(1-t) + \mathbf{z}_1^2(1-t)^2.$$

We thus have $\mathbf{p}_1'(1) = \mathbf{h}_1(1) = \mathbf{w}_1^2$ a $\mathbf{p}_2'(0) = \mathbf{h}_2(0) = c^2\mathbf{w}_1^2$. The tangent vectors thus have the same direction and orientation and we have the $\mathcal{G}^1$ continuity.

 ISBN: 978-84-617-8694-7

Moreover the reparameterization of $\mathbf{p}_2$ by the function $\phi(t) = t/c^2$ provides the $\mathcal{C}^1$ continuity. Denote

$$\mathbf{P}_0 = \frac{\mathbf{w}_0^2 + \mathbf{w}_0\mathbf{w}_1 + \mathbf{w}_1^2}{3}, \mathbf{P}_1 = \mathbf{P}_0 + \frac{c^2\mathbf{w}_1^2}{3},$$
$$\mathbf{P}_2 = \mathbf{P}_1 + \frac{c\mathbf{w}_1\mathbf{z}_1}{3}, \mathbf{P}_3 = \mathbf{P}_2 + \frac{\mathbf{z}_1^2}{3}.$$

It holds $\mathbf{p}_2(t) = \mathbf{P}_0(1-t)^3 + \mathbf{P}_1 3t(1-t)^2 + \mathbf{P}_2 3t^2(1-t) + \mathbf{P}_3 t^3$, tedy

$$(\mathbf{p}_2 \circ \phi)(t) = \mathbf{P}_0\left(1 - \frac{t}{c^2}\right)^3 + \mathbf{P}_1\frac{3t}{c^2}\left(1 - \frac{t}{c^2}\right)^2 + \mathbf{P}_2\frac{3t^2}{c^4}\left(1 - \frac{t}{c^2}\right) + \mathbf{P}_3\frac{t^3}{c^6}.$$

Differentiating we get

$$(\mathbf{p}_2 \circ \phi)'(t) = \mathbf{P}_0\left(-\frac{3}{c^2} + \frac{6t}{c^4} - \frac{3t^2}{c^6}\right) + 3\mathbf{P}_1\left(\frac{1}{c^2} - \frac{4t}{c^4} + \frac{3t^2}{c^6}\right) +$$
$$+ 3\mathbf{P}_2\left(\frac{2t}{c^4} - \frac{3t^2}{c^6}\right) + \mathbf{P}_3\frac{3t^2}{c^6},$$

and thus $(\mathbf{p}_2 \circ \phi)'(0) = -\frac{3\mathbf{P}_0}{c^2} + \frac{3\mathbf{P}_1}{c^2} = \mathbf{w}_1^2 = \mathbf{p}_1'(1)$.
For the inverse implication if the connection is $\mathcal{G}^1$ we have $\mathbf{p}_1'(1) = k\mathbf{p}_2'(0), k > 0$, and thus $\mathbf{h}_1(1) = \mathbf{w}_1^2 = k\mathbf{h}_2(0) = k\mathbf{z}_0^2$. Setting $c = \pm\sqrt{k} \neq 0$ we get the statement.
For $c = \pm 1$ there holds $\phi(t) = t/c^2 = t$, the reparameterization is thus the identity and we have the $\mathcal{C}^1$connection. □

The following proposition gives the conditions for the $\mathcal{G}^2$ and $\mathcal{C}^2$ continuity.

Proposition 3.2 Let $\mathbf{p}_1$, $\mathbf{p}_2$ are PH cubics with the preimages (4) and let $\mathbf{z}_0 = c\mathbf{w}_1, c \neq 0$. The curves $\mathbf{p}_1, \mathbf{p}_2$ are connected in the $\mathcal{G}^2$ way if and only if the point $\mathbf{z}_1$ is on the straight line

$$w_{12}x - w_{11}y + c^3(w_{01}w_{12} - w_{11}w_{02}) = 0.$$

The curves are connected in the $\mathcal{C}^2$ way if and only if $c = \pm 1$ and
$c(\mathbf{z}_1 - \mathbf{z}_0) = c(\mathbf{z}_1 - c\mathbf{w}_1) = \mathbf{w}_1 - \mathbf{w}_0$, tedy $\mathbf{z}_1 = c(2\mathbf{w}_1 - \mathbf{w}_0)$.

Proof: The $\mathcal{G}^1$ connection is ensured by the previous proposition. We must obtain the same signed curvature on both segments. Let us denote κ_1 (resp. κ_2) the signed curvature of the curve $\mathbf{p}_1$ (resp. $\mathbf{p}_2$). For $\mathbf{q}_1 = \mathbf{w}_0(1-t) + \mathbf{w}_1 t$ we get using 2.6

$$\kappa_1(t) = 2\frac{\mathrm{Im}(\bar{\mathbf{q}_1}(t)\mathbf{q}_1'(t))}{|\mathbf{q}_1(t)|^4} =$$
$$= \frac{2[(w_{01} + t(w_{11} - w_{01}))(w_{12} - w_{02}) - (w_{11} - w_{01})(w_{02} + t(w_{12} - w_{02}))]}{[(w_{01} + t(w_{11} - w_{01}))^2 + (w_{02} + t(w_{12} - w_{02}))^2]^2}$$

 ISBN: 978-84-617-8694-7

and for $\mathbf{q}_2 = c\mathbf{w}_1(1-t) + \mathbf{z}_1 t$, $\mathbf{z}_1 = [z_{11}, z_{12}]$ we get

$$\kappa_2(t) = 2\frac{\mathrm{Im}(\bar{\mathbf{q}_2}(t)\mathbf{q}_2'(t))}{|\mathbf{q}_2(t)|^4} =$$
$$= \frac{2[(cw_{11} + t(z_{11} - cw_{11}))(z_{12} - cw_{12}) - (z_{11} - cw_{11})(cw_{12} + t(z_{12} - cw_{12}))]}{[(cw_{11} + t(z_{11} - cw_{11}))^2 + (cw_{12} + t(z_{12} - cw_{12}))^2]^2}.$$

We require $\kappa_1(1) = \kappa_2(0)$, which leads to

$$\frac{2(w_{01}w_{12} - w_{11}w_{02})}{(w_{11}^2 + w_{12}^2)^2} = \frac{2c(w_{11}z_{12} - w_{12}z_{11})}{((cw_{11})^2 + (cw_{12})^2)^2},$$

which gives the condition for $\mathbf{z}_1$, namely

$$c^3(w_{01}w_{12} - w_{11}w_{02}) = w_{11}z_{12} - w_{12}z_{11}.$$

For the $\mathcal{C}^2$ continuity we must have $\mathbf{p}_1''(1) = \mathbf{p}_2''(0)$. There holds

$$\mathbf{p}_1''(1) = \mathbf{p}_2''(0) \Leftrightarrow \mathbf{h}_1'(1) = \mathbf{h}_2'(0) \Leftrightarrow 2\mathbf{w}_1(\mathbf{w}_1 - \mathbf{w}_0) = 2c\mathbf{w}_1(\mathbf{z}_1 - c\mathbf{w}_1) \Leftrightarrow$$
$$\Leftrightarrow (\mathbf{w}_1 - \mathbf{w}_0) = c(\mathbf{z}_1 - c\mathbf{w}_1), c = \pm 1.$$

□

4 Examples

Let us demonstrate the continuity results on several examples.

Example 4.1 *Let*

$$\mathbf{w}_0 = [1,\, 1],\ \mathbf{w}_1 = [0,\, -1],\ \mathbf{z}_0 = c\mathbf{w}_1 = [0,\, -3/2],\ \mathbf{z}_1 = [-1,\, 0], c = 3/2.$$

By Proposition 3.1 the resulting curves should be connected in the $\mathcal{G}^1$ way. Indeed

$$\mathbf{p}_2'(0) = (-9/4,\, 0) = c^2\mathbf{p}_1'(1) = \frac{9}{4}(-1, 0).$$

On Figure 1 there is the linear preimage $\mathbf{q}_1$, $\mathbf{q}_2$, and on Figure 2 the resulting PH curves. $\mathcal{C}^1$ connection is a special case of the $\mathcal{G}^1$ connection and is obtained by setting $c = \pm 1$. For $c = 1$ we get $\tilde{\mathbf{z}}_0 = [0, -1]$ and indeed $\tilde{\mathbf{p}}_2'(0) = \mathbf{p}_1'(1) = (-1, 0)$, see Figures 3 and 4. The preimage $\tilde{\mathbf{q}}_2$ has the control points $\tilde{\mathbf{z}}_0$, $\mathbf{z}_1$ and $\tilde{\mathbf{p}}_2$ is the resulting PH curve.

 ISBN: 978-84-617-8694-7

Example 4.2 *Let*

$$\mathbf{w}_0 = [1,\, 0],\, \mathbf{w}_1 = [2,\, 2],\, \mathbf{z}_0 = c\mathbf{w}_1 = [4,\, 4],\, \mathbf{z}_1 = [0,\, 8], c = 2.$$

For $\mathbf{z}_1$ *on the line* $x - y + 8 = 0$ *we get by 3.2 the* $\mathcal{G}^2$ *connectivity. On the figure 5 we see the preimage of the resulting PH along the the condition line (red). At the Figure 6 we see the resulting PH curves having at the joint points identical signed curvatures* $\kappa_{z1}(1) = \kappa_{z2}(0) = 1/16$.
In order to have the $\mathcal{C}^1$ *continuity let us set* $c = -1$, *and thus* $\tilde{\mathbf{z}}_0 = [-2,\, -2]$. *To obtain the* $\mathcal{G}^2$ *continuity the control point* $\mathbf{z}_1$ *must be on the line* $x - y - 1 = 0$. *If we set* $\mathbf{z}_1 = -2\mathbf{w}_1 + \mathbf{w}_0 = [-3,\, -4]$ *we obtain even the* $\mathcal{C}^2$ *continuity. We can see the results on Figures 7 and 8.*

5 Conclusion

We have solved the problem of the connection of order one and two for planar PH cubics. In the future we plan to investigate higher degree PH curves and express the connection results in the formalism of B-splines.

Acknowledgements

This work was supported by the project number 17-01171S of the Czech Science Foundation.

References

[1] *Farouki, R.:* Pythagorean-hodograph curves: algebra and geometry inseparable, Geometry and Computing Vol. 1, Springer, Berlin, ISBN 978-3-540-73397-3 (2008)

[2] *Šír, Z. Feichtinger, R. and Jüttler, B.:* Approximating curves and their offsets using biarcs and Pythagorean hodograph quintics. In CAD Computer Aided Design, 38 (6): 608-618, 2006.

[3] *Farouki, R., Moon H.P. and Choi, H.:* Construction and shape analysis of PH quintic Hermite interpolants, Computer Aided Geometric Design 18, 93-115 (2001)

[4] *Farouki, R.T., Neff A.C.:* Hermite Interpolation by Pythagorean Hodograph Quintics, Mathematics of Computation, Volume 64, Numver 212, October 1995, pages 1589-1609

[5] *Šír, Z., Kosinka, J.:* Low Degree Euclidean and Minkowski Pythagorean Hodograph Curves, M. Daehlen et al. (Eds.): MMCS 2008, LNCS 5862, pp. 394418, 2010.

 ISBN: 978-84-617-8694-7

[6] *Kubota, K.K.:* Pythagorean Triples in Unique Factorization Domains. The American Mathematical Monthly, Vol. 79, No. 5 (May, 1972), pp. 503-505.

 ISBN: 978-84-617-8694-7

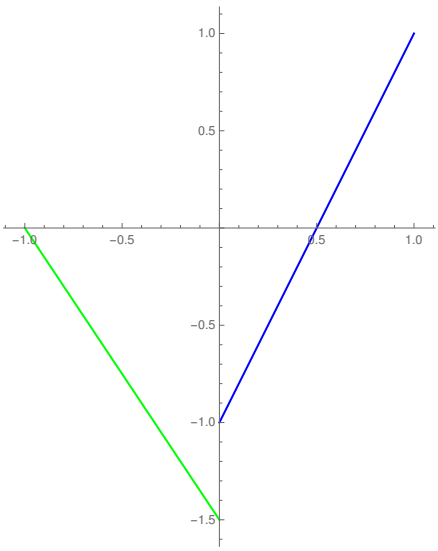


Figure 1: Linear preimage $\mathbf{q}_1$ (blue) a $\mathbf{q}_2$ (green)

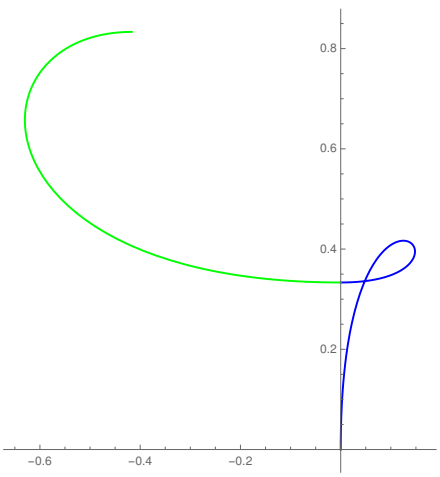


Figure 2: PH curves $\mathbf{p}_1$ a $\mathbf{p}_2$ connected in the $\mathcal{G}^1$ way.

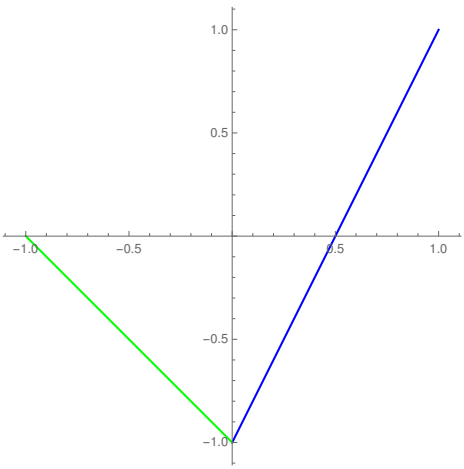


Figure 3: Linear preimage $\mathbf{q}_1$ and $\tilde{\mathbf{q}}_2$

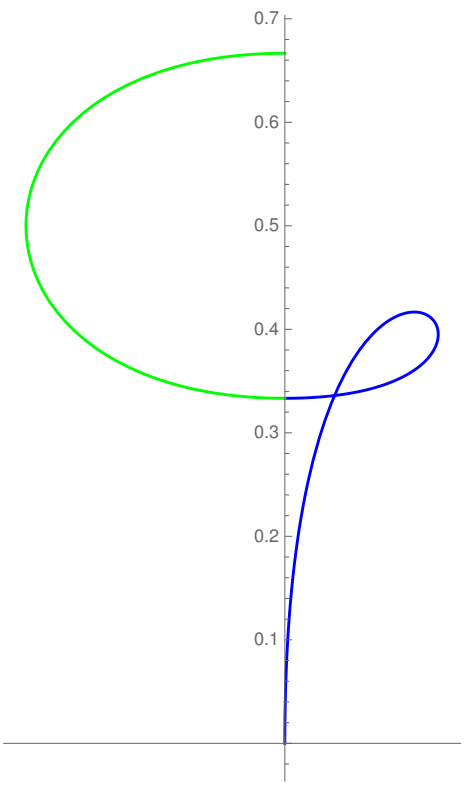


Figure 4: PH curves $\mathbf{p}_1$ a $\tilde{\mathbf{p}}_2$ connected $\mathcal{C}^1$

 ISBN: 978-84-617-8694-7

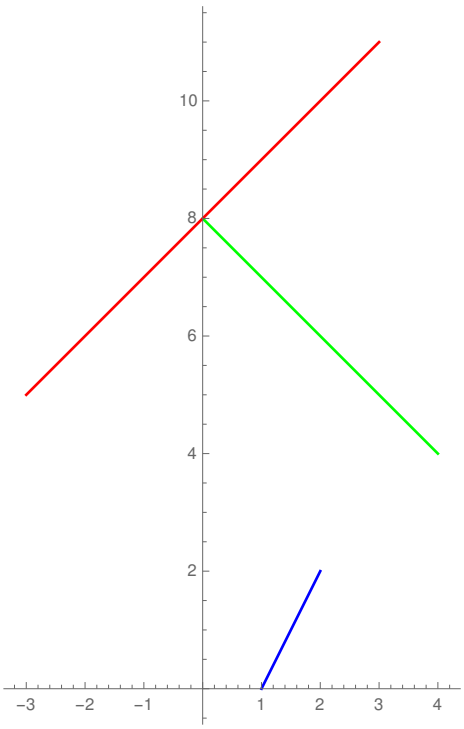


Figure 5: Linear preimage $\mathbf{q}_1$ a $\mathbf{q}_2$

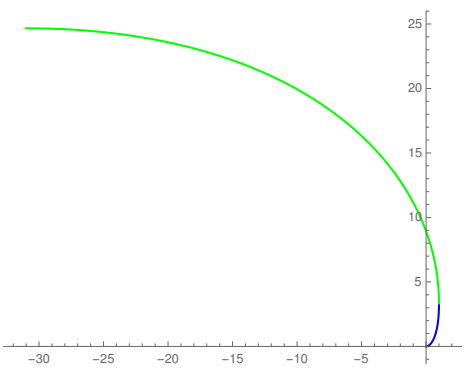


Figure 6: PH curves $\mathbf{p}_1$ and $\mathbf{p}_2$ connected in the $\mathcal{G}^2$ way

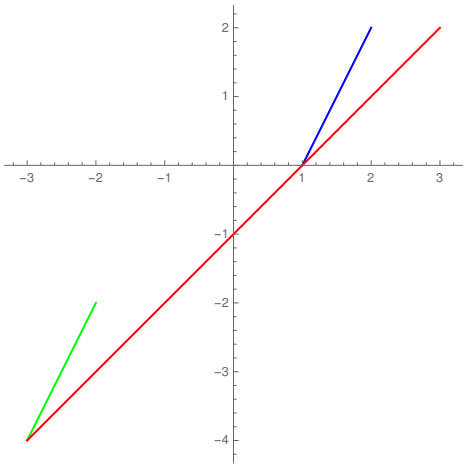


Figure 7: Linear preimages $\mathbf{q}_1$ a $\tilde{\mathbf{q}}_2$.

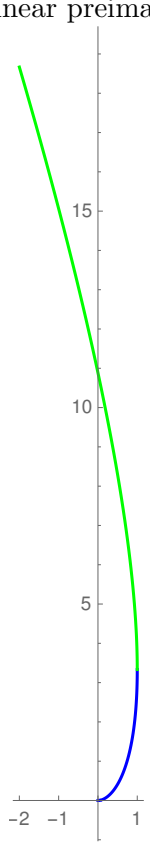


Figure 8: PH curves $\mathbf{p}_1$ and $\tilde{\mathbf{p}}_2$ connected in the $\mathcal{C}^2$ way.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

A Probabilistic Evolution Theoretical (PREVTH) Approach to Quantum Evolver Dynamical Equations for Singular Hamiltonians: Fluctuationlessness Approximation

Berfin Kalay[1] **and Metin Demiralp**[1]

[1] *Computational Science and Engineering Department, Informatics Institute, İstanbul Technical University*

emails: berfinkalay@gmail.com, metin.demiralp@gmail.com

Abstract

This work is devoted to the investigation of the Evolver Dynamics of a singular quantum system via Probabilistic Evolution Theoretical tools recently developed in our group studies. Even though the purpose is rather conceptuality we have also developed an approximation which is based on the fluctuationlessness theorem of multivariate functions. We have taken hydrogen-like systems as the target for constructing evolver dynamical equations. The system vector which is composed of operators involves either the position and momentum operator or position, momentum and Hamilton operator related operators as the basic entities and this gives expressions which are singular in position operator. On the other hand, the use of Hamiltonian powers instead of the inverse powers of position operator has facilitated the analysis especially for the use of ground state eigenfunction of the Hamiltonian as the initial wave function. We do not present any implementation because of the conceptual structure of the paper while we have reported how to use fluctuationlessness theorem to evaluate the expectation value evolutions.
Key words: PREVTH, Evolver Dynamics, Expectation Values, Mathematical Fluctuations, Uncertainty.

1 Introduction

Probabilistic Evolution Theory (PREVTH) has been recently developed in our group studies, [1–16] and has been quite well investigated for explicit autonomous ODE(s) with initial impositions. As we have done in our various works whose findings are reported until now,

we have attempted to use PREVTH in quantum expectation value dynamical equations arising from one of our main interests: quantum dynamical motions. We have employed Kronecker power series we proposed in last decade and Poisson Bracket algebra over the basic operators of the quantum system under consideration and obtained an analytical solution after using so-called space extensions, where the solution's vector components are determined thru rather simple recursions. Yet practical efficiency has not been provided and rather complicated cures are proposed in our group studies. All these urged us to develop the Evolver Dynamics which is somehow related to Heisenberg Picture of Quantum Mechanics. The coming sections detail this and related issues.

2 Introducing the Evolver Dynamics

In Schrödinger Picture, a quantum dynamical system can be described by the following partial differential equation which is known as time dependent Schrödinger equation and the accompanying initial imposition

$$i\hbar\frac{\partial\psi\left(\mathbf{x},t\right)}{\partial t}=\widehat{H}\left(\widehat{p},\widehat{q}\,\right)\psi\left(\mathbf{x},t\right),\qquad\psi\left(\mathbf{x},t_{0}\right)=\psi_{0}\left(\mathbf{x}\right)\tag{1}$$

where ψ denotes the wave function of the system, which is a function of position variable denoted by $\mathbf{x}$ and time is symbolized by t while $\hbar$ is the "Reduced Planck Constant" which is in fact the ratio of the universal Planck Constant to 2π. The initial moment of the system's evolution has been denoted by t_0 instead of 0 to give the possibility of using any initialization instant and therefore any desired subinterval for the evolution. Even though ψ_0 can be anything from the space of wave functions, we are going to confine ourselves to rather minimum uncertainty wave packets here for conciseness and brevity.

The expectation value dynamics avoids the use of wave function explicitly except its initial form and therefore the solution of the parabolic Schrödinger equations. In these type approaches, the transition operators are also frequently used and their investigation can be considered as "Transition Operator Picture" (even though this term is rather infrequently employed). This operator is defined as follows

$$\psi\left(\mathbf{x},t_{1}\right)=\widehat{T}\left(\widehat{p},\widehat{q},t_{1},t_{0}\right)\psi\left(\mathbf{x},t_{0}\right)\tag{2}$$

and connects the system's wave function values at two different time instants, t_0 and t_1. Transition operators for isolated systems (which do not interact with their environment) and interacted systems (having interactions with their environments) have of course different structures and the former case is analytically expressible. We do not intend to use different symbols or subscripts to this end but the Hamiltonian will reveal the difference.

 ISBN: 978-84-617-8694-7

The use of (2) in (1) implies that the transition operator $\widehat{T}$ will satisfy the following partial differential equation

$$i\hbar\frac{\partial\widehat{T}\left(\widehat{p},\widehat{q},t,t_0\right)}{\partial t}=\widehat{H}\left(\widehat{p},\widehat{q},t\right)\widehat{T}\left(\widehat{p},\widehat{q},t,t_0\right),\qquad\widehat{T}\left(\widehat{p},\widehat{q},t_0,t_0\right)=\widehat{I}\tag{3}$$

Transition Operator definition does not pretty much facilitate the investigation of quantum system evolution. However it helps us to construct the evolver dynamics. Beside Schrödinger Picture there is another and useful point of view which does not use wave function concept. It is a very well known theory and called "Heisenberg Picture". This picture uses an operator evolving in time for each given operator. This operator is produced by using a general superoperator.

We can now define

$$\mathbb{E}\left(\widehat{o}\right)(t)\equiv\widehat{T}\left(t,t_0\right)^{\dagger}\widehat{o}\,\widehat{T}\left(t,t_0\right)\tag{4}$$

where we call $\mathbb{E}\left(\widehat{o}\right)(t)$ "Evolver Operator of the Operator $\widehat{o}$" or "$\widehat{o}$'s Evolver Operator" while $\mathbb{E}$ which is in fact a superoperator mapping from the space of linear operators to the same space of linear operators since it defines a unitary transformation, is called "Evolver Superoperator".

Since $\widehat{o}$'s evolver operator is a function of t, we may investigate its time derivative and write

$$\begin{aligned}\frac{d\mathbb{E}\left(\widehat{o}\right)(t)}{dt}&=\frac{d\widehat{T}\left(t,t_0\right)^{\dagger}}{dt}\widehat{o}\,\widehat{T}\left(t,t_0\right)+\widehat{T}\left(t,t_0\right)^{\dagger}\widehat{o}\,\frac{d\widehat{T}\left(t,t_0\right)}{dt}\\&=\widehat{T}\left(t,t_0\right)^{\dagger}\left\{\widehat{H},\widehat{o}\right\}\widehat{T}\left(t,t_0\right)\end{aligned}\tag{5}$$

This result shows that time derivative of the image of a target operator $\widehat{o}$ under "Evolver Superoperator" takes us to the image of Poisson Bracket of that operator with the system Hamiltonian, under the same "Evolver Superoperator".

$$\frac{d\mathbb{E}\left(\widehat{o}\right)(t)}{dt}=\mathbb{E}\left(\left\{\widehat{H},\widehat{o}\right\}\right)\tag{6}$$

which is a general conclusion for any time independent operator $\widehat{o}$. The temporal variation of the operator $\widehat{o}$ changes the right hand side of (6) by adding the temporal derivative of $\widehat{o}$ to $\left\{\widehat{H},\widehat{o}\right\}$.

3 Constructing the PREVTH System Vector With No Hamiltonian Dependence

Now to explain how to construct evolver dynamics for a quantum system we are going to focus on a rather simple singular system, hydrogen-like system whose Hamiltonian can be

 ISBN: 978-84-617-8694-7

explicitly given below

$$\widehat{H} \equiv \frac{1}{2\mu}\widehat{p}^{\,2} - \alpha\widehat{q}^{\,-1} \tag{7}$$

where μ denotes the reduced mass parameter while $\widehat{p}$ and $\widehat{q}$ stand for momentum and position operators respectively whereas the scalar α is proportional to the atomic number of the system and contains also certain universal physical constants. We can naturally focus on the momentum and position operators and try to evaluate their Poisson Brackets with the system Hamiltonian. To this end we can use the following well-known properties of the Poission Bracket: (1) it is linear with respect to its both operands, (2) It is distributive on the products in one operand, (3) The leftmost or rightmost Hamilton operator factors in the operands can be factored out the Poisson Bracket. By using these features we can write the following equalities without giving the intermediate details

$$\begin{aligned} \left\{\widehat{H},\widehat{p}\right\} &= \alpha\left\{\widehat{p},\widehat{q}^{\,-1}\right\} = -\alpha\widehat{q}^{\,-2}, \\ \left\{\widehat{H},\widehat{q}\right\} &= \frac{1}{2\mu}\left\{\widehat{p}^{\,2},\widehat{q}\right\} - \alpha\left\{\widehat{q}^{\,-1},\widehat{q}\right\} = \frac{1}{\mu}\widehat{p} \end{aligned} \tag{8}$$

These equalities show that the set composed of momentum and position operators is not multinomially closed under the Poisson Bracket operation. Hence these are not appropriate for the PREVTH equations construction. To get at least multinomiality and beyond that conicality (second degree multinomial right hand sides) we need to consider the inverse of the position operator as a new basic operator. This is a space extension since the dimensionality of the basic operators is increased by one. However, momentum and position operator together with the reciprocal of the position operator does not provide the conicality and enforces us to increase the dimensionality. Until the conicality is obtained the basic operator set remains open under the action of the Poisson Bracket and we are further enforced to extend the space. After our all efforts we could have been able to show that the following system vector of operators provides the conicality.

$$\widehat{\mathbf{s}} \equiv [\,\widehat{s}_1 \;\; ... \;\; \widehat{s}_6\,] \equiv \left[\,\widehat{p} \;\; \widehat{q} \;\; \widehat{q}^{\,-1} \;\; \widehat{p}\,\widehat{q}^{\,-1} \;\; \widehat{q}^{\,-2} \;\; \widehat{q}^{\,-1}\widehat{p}\,\right] \tag{9}$$

The Poisson Bracket evaluations based on this structure can be given as follows

$$\begin{aligned} \left\{\widehat{H},\widehat{s}_1\right\} &= -\alpha\widehat{s}_3^2 & \left\{\widehat{H},\widehat{s}_2\right\} &= \tfrac{1}{\mu}\widehat{s}_1 \\ \left\{\widehat{H},\widehat{s}_3\right\} &= -\tfrac{1}{\mu}\widehat{s}_3\widehat{s}_4 & \left\{\widehat{H},\widehat{s}_4\right\} &= -\alpha\widehat{s}_3\widehat{s}_5 - \tfrac{1}{\mu}\widehat{s}_4^{\,2} \\ \left\{\widehat{H},\widehat{s}_5\right\} &= -\tfrac{1}{\mu}\widehat{s}_5\widehat{s}_4 - \tfrac{1}{\mu}\widehat{s}_6\widehat{s}_5 & \left\{\widehat{H},\widehat{s}_6\right\} &= -\tfrac{1}{\mu}\widehat{s}_6^{\,2} - \alpha\widehat{s}_3\widehat{s}_5 \end{aligned} \tag{10}$$

These equalities can be rewritten in the following concise format

$$\left\{\widehat{H},\widehat{\mathbf{s}}\right\} = \mathbf{F_1}\widehat{\mathbf{s}} + \mathbf{F_2}\widehat{\mathbf{s}}^{\otimes 2} \tag{11}$$

 ISBN: 978-84-617-8694-7

where $\mathbf{F}_1$ is desired to be proportional to 6×6 type identity matrix for getting analytical PREVTH solution while $\mathbf{F}_2$ stands for 6×36 type rectangular matrix. However $\mathbf{F}_1$ is not at that format which can be achieved by using an extra space extension by adding the operator element $\widehat{s}_7$ thru $\widehat{s}_7 \equiv a\widehat{I}$ where $\widehat{I}$ stands for the identity operator while a is arbitrary at this moment. This space extension is also known Constancy Adding Space Extension (CASE) whose original form was more comprehensive. After this CASE we can rewrite the above relevant equalities as follows

$$\widehat{\mathbf{s}} \equiv [\, \widehat{s}_1 \;\; ... \;\; \widehat{s}_7 \,] \equiv \left[\widehat{p} \;\; \widehat{q} \;\; \widehat{q}^{\,-1} \;\; \widehat{p}\widehat{q}^{\,-1} \;\; \widehat{q}^{\,-2} \;\; \widehat{q}^{\,-1}\widehat{p} \;\; a\widehat{I} \right] \tag{12}$$

$$\begin{aligned}
&\left\{\widehat{H}, \widehat{s}_1\right\} = \beta\widehat{s}_1 - \tfrac{\beta}{a}\widehat{s}_1\widehat{s}_7 - \alpha\widehat{s}_3^2 && \left\{\widehat{H}, \widehat{s}_2\right\} = \beta\widehat{s}_2 - \tfrac{\beta}{a}\widehat{s}_2\widehat{s}_7 + \tfrac{1}{a\mu}\widehat{s}_1\widehat{s}_7 \\
&\left\{\widehat{H}, \widehat{s}_3\right\} = \beta\widehat{s}_3 - \tfrac{\beta}{a}\widehat{s}_3\widehat{s}_7 - \tfrac{1}{\mu}\widehat{s}_3\widehat{s}_4 && \left\{\widehat{H}, \widehat{s}_4\right\} = \beta\widehat{s}_4 - \tfrac{\beta}{a}\widehat{s}_4\widehat{s}_7 - \alpha\widehat{s}_3\widehat{s}_5 - \tfrac{1}{\mu}\widehat{s}_4^{\,2} \\
&\left\{\widehat{H}, \widehat{s}_5\right\} = \beta\widehat{s}_5 - \tfrac{\beta}{a}\widehat{s}_5\widehat{s}_7 - \tfrac{1}{\mu}\widehat{s}_5\widehat{s}_4 - \tfrac{1}{\mu}\widehat{s}_6\widehat{s}_5 && \left\{\widehat{H}, \widehat{s}_6\right\} = \beta\widehat{s}_6 - \tfrac{\beta}{a}\widehat{s}_6\widehat{s}_7 - \tfrac{1}{\mu}\widehat{s}_6^{\,2} - \alpha\widehat{s}_3\widehat{s}_5 \\
&\left\{\widehat{H}, \widehat{s}_7\right\} = \beta\widehat{s}_7 - \tfrac{\beta}{a}\widehat{s}_7^{\,2}
\end{aligned} \tag{13}$$

where first six equalities of last formula turn out to be matching previous equalities while the seventh one becomes having zero right hand side when $\widehat{s}_7$ is replaced with $a\widehat{I}$. However we can write these 7 equalities in the following concise format

$$\left\{\widehat{H}, \widehat{\mathbf{s}}\right\} = \beta\,\widehat{\mathbf{s}} + \mathbf{F}\,\widehat{\mathbf{s}}^{\otimes 2} \tag{14}$$

where $\mathbf{F}$ is a rectangular matrix of 7×49 type. The nonzero elements of $\mathbf{F}$ are explicitly given below

$$\begin{aligned}
&F_{1,\,7} = -\tfrac{\beta}{a} && F_{1,\,17} = -\alpha && F_{2,14} = -\tfrac{\beta}{a} && F_{2,\,7} = \tfrac{1}{a\mu} \\
&F_{3,18} = -\tfrac{1}{\mu} && F_{3,21} = -\tfrac{\beta}{a} && F_{4,19} = -\alpha && F_{4,25} = -\tfrac{1}{\mu} \\
&F_{4,28} = -\tfrac{\beta}{a} && F_{5,32} = -\tfrac{1}{\mu} && F_{5,35} = -\tfrac{\beta}{a} && F_{5,40} = -\tfrac{1}{\mu} \\
&F_{6,19} = -\alpha && F_{6,41} = -\tfrac{1}{\mu} && F_{6,42} = -\tfrac{\beta}{a} && F_{7,49} = -\tfrac{\beta}{a}
\end{aligned} \tag{15}$$

The rectangular matrix ultimately obtained above is not unique because of two important reasons: (1) the employed space extension is based on the addition of new functions of original unknowns such that the additions are generally chosen linearly independent even though they are chosen functionally dependent on the existing unknowns, (2) the use of CASE brings very extensive flexibilites for the structuring of the matrix $\mathbf{F}$.

On the other hand the selection of the basic operators may change the $\mathbf{F}$ structuring depending on the preferences.

4 Constructing the PREVTH System Vector With Hamiltonian Dependence

The operator basis set we have used in the previous section contains the powers of finite region singularity in position operator explicitly. This may not be always desired and it preferred to hi de them in Hamiltonian and its powers for certain conveniences. In this section we are going to use Hamilton operator and its certain powers instead of position operator's reciprocal powers. Our investigations have shown that the following system vector's operator elements set is conically closed under the Poisson Bracket operation with the system Hamiltonian and also fits PREVTH format as CASE has been used to get that format

$$\widehat{\mathbf{s}} \equiv [\,\widehat{s}_1 \;\; ... \;\; \widehat{s}_{10}\,] \equiv \left[\,\widehat{p} \;\; \widehat{p}^{\,2} \;\; \widehat{H} \;\; \widehat{H}\,\widehat{p} \;\; \widehat{p}\,\widehat{H} \;\; \widehat{H}^{\,2} \;\; \widehat{p}^{\,3} \;\; \widehat{p}^{\,2}\,\widehat{H} \;\; \widehat{H}\,\widehat{p}^{\,2} \;\; a\widehat{I}\,\right] \tag{16}$$

This enables us to write the following equalities

$$\begin{aligned}
\left\{\widehat{H},\widehat{s}_1\right\} &= -\tfrac{1}{4\mu^2\alpha}\widehat{s}_2^2 + \tfrac{1}{2\mu\alpha}\widehat{s}_2\widehat{s}_3 + \tfrac{1}{2\mu\alpha}\widehat{s}_3\widehat{s}_2 \\
&\quad - \tfrac{1}{\alpha}\widehat{s}_3^2 + \beta\widehat{s}_1 - \tfrac{\beta}{a}\widehat{s}_1\widehat{s}_{10} \\
\left\{\widehat{H},\widehat{s}_2\right\} &= -\tfrac{1}{2\mu^2\alpha}\widehat{s}_2\widehat{s}_7 + \tfrac{1}{2\mu\alpha}\left(\widehat{s}_2\widehat{s}_4 + \widehat{s}_4\widehat{s}_2\right) \\
&\quad - \tfrac{1}{\alpha}\widehat{s}_3\widehat{s}_4 + \beta\widehat{s}_2 - \tfrac{\beta}{a}\widehat{s}_2\widehat{s}_{10} \\
\left\{\widehat{H},\widehat{s}_3\right\} &= \beta\widehat{s}_3 - \tfrac{\beta}{a}\widehat{s}_3\widehat{s}_{10} \\
\left\{\widehat{H},\widehat{s}_4\right\} &= -\tfrac{1}{4\mu^2\alpha}\widehat{s}_9\widehat{s}_2 + \tfrac{1}{2\mu\alpha}\left(\widehat{s}_4\widehat{s}_5 + \widehat{s}_6\widehat{s}_2\right) \\
&\quad - \tfrac{1}{\alpha}\widehat{s}_3\widehat{s}_6 + \beta\widehat{s}_4 - \tfrac{\beta}{a}\widehat{s}_4\widehat{s}_{10} \\
\left\{\widehat{H},\widehat{s}_5\right\} &= -\tfrac{1}{4\mu^2\alpha}\widehat{s}_2\widehat{s}_8 + \tfrac{1}{2\mu\alpha}\widehat{s}_2\widehat{s}_6 + \tfrac{1}{2\mu\alpha}\widehat{s}_4\widehat{s}_5 \\
&\quad - \tfrac{1}{\alpha}\widehat{s}_3\widehat{s}_6 + \beta\widehat{s}_5 - \tfrac{\beta}{a}\widehat{s}_5\widehat{s}_{10} \\
\left\{\widehat{H},\widehat{s}_6\right\} &= \beta\widehat{s}_6 - \tfrac{\beta}{a}\widehat{s}_6\widehat{s}_{10} \\
\left\{\widehat{H},\widehat{s}_7\right\} &= -\tfrac{3}{4\mu^2\alpha}\widehat{s}_7^2 + \tfrac{1}{\mu\alpha}\widehat{s}_8\widehat{s}_2 + \tfrac{1}{2\mu\alpha}\widehat{s}_9\widehat{s}_2 \\
&\quad - \tfrac{1}{\alpha}\widehat{s}_3\widehat{s}_9 + \tfrac{1}{2\mu\alpha}\widehat{s}_7\widehat{s}_4 + \tfrac{1}{2\mu\alpha}\widehat{s}_5\widehat{s}_7 \\
&\quad - \tfrac{1}{\alpha}\widehat{s}_5\widehat{s}_4 + \tfrac{1}{2\mu\alpha}\widehat{s}_2\widehat{s}_8 - \tfrac{1}{\alpha}\widehat{s}_8\widehat{s}_3 \\
&\quad + \beta\widehat{s}_7 - \tfrac{\beta}{a}\widehat{s}_7\widehat{s}_{10} \\
\left\{\widehat{H},\widehat{s}_8\right\} &= -\tfrac{1}{2\mu^2\alpha}\widehat{s}_7\widehat{s}_8 + \tfrac{1}{2\mu\alpha}\left(\widehat{s}_8\widehat{s}_5 + \widehat{s}_4\widehat{s}_8\right) \\
&\quad - \tfrac{1}{\alpha}\widehat{s}_6\widehat{s}_5 + \tfrac{1}{2\mu\alpha}\widehat{s}_7\widehat{s}_6 + \tfrac{1}{2\mu\alpha}\widehat{s}_5\widehat{s}_8 \\
&\quad - \tfrac{1}{\alpha}\widehat{s}_5\widehat{s}_6 + \beta\widehat{s}_8 - \tfrac{\beta}{a}\widehat{s}_8\widehat{s}_{10} \\
\left\{\widehat{H},\widehat{s}_9\right\} &= -\tfrac{1}{2\mu^2\alpha}\widehat{s}_9\widehat{s}_7 + \tfrac{1}{2\mu\alpha}\widehat{s}_9\widehat{s}_4 + \tfrac{1}{2\mu\alpha}\widehat{s}_6\widehat{s}_7 \\
&\quad - \tfrac{1}{\alpha}\widehat{s}_6\widehat{s}_4 + \tfrac{1}{2\mu\alpha}\widehat{s}_4\widehat{s}_8 + \tfrac{1}{2\mu\alpha}\widehat{s}_4\widehat{s}_9 \\
&\quad - \tfrac{1}{\alpha}\widehat{s}_4\widehat{s}_6 + \beta\widehat{s}_9 - \tfrac{\beta}{a}\widehat{s}_9\widehat{s}_{10} \\
\left\{\widehat{H},\widehat{s}_{10}\right\} &= \beta\widehat{s}_{10} - \tfrac{\beta}{a}\widehat{s}_{10}^2
\end{aligned} \tag{17}$$

This structure can be rewritten in the following much more concise formula

$$\dot{\widehat{\mathbf{s}}}(t) = \beta\,\widehat{\mathbf{s}}(t) + \mathbf{F}\,\widehat{\mathbf{s}}(t)^{\otimes 2} \tag{18}$$

where $\mathbf{F}$ is an 11×121 type rectangular matrix with constant elements which may depend on certain system parameters beyond β and a. 54 non-zero elements of 1331 total elements

 ISBN: 978-84-617-8694-7

of $\mathbf{F}$ are explicitly given below

- $F_{1,10} = -\frac{\beta}{a}$, $F_{1,12} = -\frac{1}{4\alpha\mu^2}$, $F_{1,13} = \frac{1}{2\alpha\mu}$, $F_{1,22} = \frac{1}{2\alpha\mu}$, $F_{1,23} = -\frac{1}{\alpha}$;
- $F_{2,14} = \frac{1}{2\alpha\mu}$, $F_{2,17} = -\frac{1}{2\alpha\mu^2}$, $F_{2,20} = -\frac{\beta}{a}$, $F_{2,24} = -\frac{1}{\alpha}$ $F_{2,32} = \frac{1}{2\alpha\mu}$;
- $F_{3,30} = -\frac{\beta}{a}$;
- $F_{4,26} = -\frac{1}{\alpha}$, $F_{4,35} = \frac{1}{2\alpha\mu}$, $F_{4,40} = -\frac{\beta}{a}$, $F_{4,52} = \frac{1}{2\alpha\mu}$, $F_{4,82} = -\frac{1}{4\alpha\mu^2}$;
- $F_{5,16} = \frac{1}{2\alpha\mu}$, $F_{5,18} = -\frac{1}{4\alpha\mu^2}$, $F_{5,26} = -\frac{1}{\alpha}$, $F_{5,35} = \frac{1}{2\alpha\mu}$, $F_{5,50} = -\frac{\beta}{a}$;
- $F_{6,60} = -\frac{\beta}{a}$;
- $F_{7,18} = \frac{1}{2\alpha\mu}$, $F_{7,29} = -\frac{1}{\alpha}$, $F_{7,44} = -\frac{1}{\alpha}$, $F_{7,47} = \frac{1}{2\alpha\mu}$, $F_{7,64} = \frac{1}{2\alpha\mu}$, $F_{7,67} = -\frac{3}{4\alpha\mu^2}$, $F_{7,70} = -\frac{\beta}{a}$ $F_{7,72} = \frac{1}{\alpha\mu}$ $F_{7,73} = -\frac{1}{\alpha}$, $F_{7,82} = \frac{1}{2\alpha\mu}$;
- $F_{8,38} = \frac{1}{2\alpha\mu}$, $F_{8,46} = -\frac{1}{\alpha}$, $F_{8,48} = \frac{1}{2\alpha\mu}$, $F_{8,55} = -\frac{1}{\alpha}$, $F_{8,66} = \frac{1}{2\alpha\mu}$, $F_{8,68} = -\frac{1}{2\alpha\mu^2}$, $F_{8,75} = \frac{1}{2\alpha\mu}$, $F_{8,80} = -\frac{\beta}{a}$;
- $F_{9,36} = -\frac{1}{\alpha}$, $F_{9,38} = \frac{1}{2\alpha\mu}$, $F_{9,39} = \frac{1}{2\alpha\mu}$, $F_{9,54} = -\frac{1}{\alpha}$, $F_{9,57} = \frac{1}{2\alpha\mu}$, $F_{9,84} = \frac{1}{2\alpha\mu}$, $F_{9,87} = -\frac{1}{2\alpha\mu^2}$, $F_{9,90} = -\frac{\beta}{a}$;
- $F_{10,100} = -\frac{\beta}{a}$. (19)

where we have used the bullet symbol to mark rows of $\mathbf{F}$.

These equations apparently urge us to decompose $\mathbf{F}$ as follows

$$\mathbf{F} = \frac{\beta}{a}\mathbf{F}_{\frac{\beta}{a}} + \frac{1}{a}\mathbf{F}_{\frac{1}{a}} + F_R \tag{20}$$

where right hand side $\mathbf{F}$ matrices with subscripts do not depend on either β or a even though they vary with α and μ. We will use this decomposition in our companion paper [17] of this conference for optimization purposes. In the above analysis we have not focused on the position operator. However, it can be expressed in terms of the integral over the momentum evolver and since it is in principle evaluated through PREVTH solution there is no problem except the evaluation of an integral which is rather a straightforward task.

 ISBN: 978-84-617-8694-7

5 Probabilistic Evolution Theoretical Solution to the Equations of Evolver Dynamics

If both sides of (14) are premultiplied by the Hermitian conjugate of the transition operator and postmultiplied by the transition operator's itself then we can write

$$\dot{\widehat{\mathbf{S}}}(t) = \beta\,\widehat{\mathbf{S}}(t) + \mathbf{F}\,\widehat{\mathbf{S}}(t)^{\otimes 2}, \qquad \widehat{\mathbf{S}}\,(t_0) = \widehat{\mathbf{s}}, \qquad \widehat{\mathbf{S}}(t) \equiv \widehat{T}\,(t,t_0)^{\dagger}\,\widehat{\mathbf{s}}\,\widehat{T}\,(t,t_0) \tag{21}$$

where the first equation can be simplified through the following definitions and equations

$$\widehat{\mathbf{S}}(t) = \mathrm{e}^{\beta(t-t_0)}\boldsymbol{\xi}(t), \qquad u(t) \equiv \frac{\mathrm{e}^{\beta(t-t_0)}-1}{t}, \qquad \frac{d\boldsymbol{\xi}(u)}{du} = \mathbf{F}\boldsymbol{\xi}(u)^{\otimes 2}, \qquad \boldsymbol{\xi}\,(t_0) = \widehat{\mathbf{s}}. \tag{22}$$

The solution of this problem can be provided by PREVTH and can be written as follows

$$\boldsymbol{\xi}(u) = \sum_{j=0}^{\infty} \frac{u^j}{j!}\mathbf{T}_j\widehat{\mathbf{s}}^{\otimes(j+1)}, \qquad \widehat{\mathbf{S}}(t) = \mathrm{e}^{\beta(t-t_0)}\sum_{j=1}^{\infty}\frac{\mathrm{e}^{\beta(t-t_0)}-1}{\beta^j j!}\mathbf{T}_j\widehat{\mathbf{s}}^{\otimes(j+1)} \tag{23}$$

where $\widehat{\mathbf{s}}$ stands for the n-operator-element initial vector while $\mathbf{T}_j$'s (Telescope Matrices) are of $n \times n^{(j+1)}$ type and can be expressed in Monocular matrices defined below.

$$\mathbf{T}_j \equiv \prod_{k=1}^{j}\mathbf{M}_k \qquad j = 0,1,2,... \qquad \mathbf{M}_k \equiv \sum_{\ell=0}^{k-1}\mathbf{I}_n^{\otimes\ell}\otimes\mathbf{F}\otimes\mathbf{I}_n^{\otimes(k-1-\ell)}, \qquad k = 1,2,... \tag{24}$$

where $\mathbf{M}_k$ (Monocular Matrix) is of $n^k \times n^{(k+1)}$ type. Telescope matrices are cascaded form of them.

Telescope matrices are very sparse because of their above-given structures. This sparsity can be suppressed as much as possible by using the following "Squarified Telescope Matrices (SquTelMats)", $\mathbf{\Sigma}_j$s, and beyond them, more suppressed terms, $\mathbf{v}_j$ vectors. This takes us to the following vector recursion also.

$$\mathbf{v}_j = \sum_{k=0}^{j-1}\binom{j-1}{k}\lfloor\,\mathbf{F},\mathbf{v}_k\,\rceil\,\mathbf{v}_{j-k-1}, \quad j = 1,2,... \quad \mathbf{v}_0 = \widehat{\mathbf{s}}, \quad \mathbf{v}_j = \mathbf{\Sigma}_j(\widehat{\mathbf{s}})\widehat{\mathbf{s}} = \mathbf{T}_j\widehat{\mathbf{s}}^{\otimes j+1} \tag{25}$$

The squtelmat between $n \times n^2$ type $\mathbf{F}$ matrix and n element $\mathbf{y}$ vector can be explicitly defined as follows

$$\lfloor\mathbf{F},\mathbf{y}\rceil \equiv \sum_{j=1}^{n}\mathbf{y}_j\mathbf{F}_j, \qquad \mathbf{y} \equiv \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, \qquad \mathbf{F} = [\mathbf{F}_1\cdots\mathbf{F}_n] \tag{26}$$

The recursive equation of the squarification and the calculation of the squtelmats are detailed in [12–16].

 ISBN: 978-84-617-8694-7

$\mathbf{v}_k$ vectors depend on operator elements, basically the elements of $\widehat{\mathbf{s}}$ as can be noticed from (25). Hence, their expressions contain 11 operator elements and form kth degree multinomials in those operators. Since these operators are not totally commutative their expressions are quite cumbersome and the structural complications increases very rapidly as k grows. Hence, we do not give these structures explicitly even for a few small k values. Beyond that their evaluations via computer algebra systems is not easy because noncommutative algebra or very specific algorithms are needed for calculations.

6 Evaluating the Expectation Values at the Fluctuationlessness Limit

The expectation value of a given operator $\widehat{o}$ and its evolver $\mathbb{E}\,(\widehat{o})$ are defined as follows

$$\langle\, \widehat{o}\, \rangle \equiv \int_0^\infty dx \psi(x,t)^*\, \widehat{o}\, \psi(x,t), \qquad \langle\, \mathbb{E}\,(\, \widehat{o}\,)\, \rangle\,(t) \equiv \int_0^\infty dx \psi_0(x)^* \mathbb{E}\,(\widehat{o})\,(t) \psi_0(x) \tag{27}$$

where we have taken the domain of the position variable nonnegative real values consistently with our target system in this work. In our present case the system vector is composed of 11 operator elements when Hamiltonian involving operators are used instead of the reciprocal powers of the position operator. For each of these system operators we can write

$$\widehat{s}_j = \langle\, \widehat{s}_j\, \rangle\, \widehat{I} + \widehat{\phi}_j, \qquad j = 1, 2, ..., 11 \tag{28}$$

each of which is in fact the definition of the relevant fluctuation operator at the same time. As can be noticed immediately each fluctuation operator has vanishing expectation value. Now the expectation value of a multivariate analytic function depending on these system operators can be approximated by replacing its each argument operator with its expectation value. This is known as the "Fluctuationlessness Theorem" and enables us to get the conclusion "The expectation value of an operators product is equal to the product of relevant operator's expectation values at the fluctuationlessness limit". By using this fact we can rewrite the rightmost equation of (23) after using the rightmost equality of (25) and then evaluating the expectation values of both sides in the resulting equality as follows

$$\left\langle \widehat{\mathbf{S}}(t) \right\rangle = \mathrm{e}^{\beta(t-t_0)} \sum_{j=1}^{\infty} \frac{\mathrm{e}^{\beta(t-t_0)} - 1}{\beta^j j!} \langle \mathbf{v}_j \rangle \tag{29}$$

which needs to be combined with the following equality which can be produced from (25).

$$\langle \mathbf{v}_j \rangle = \sum_{k=0}^{j-1} \binom{j-1}{k} \lfloor\, \mathbf{F}, \langle \mathbf{v}_k \rangle\, \rceil\, \langle \mathbf{v}_{j-k-1} \rangle\,, \quad j = 1, 2, ... \quad \langle \mathbf{v}_0 \rangle = \langle \widehat{\mathbf{s}} \rangle \tag{30}$$

 ISBN: 978-84-617-8694-7

This recursion exactly matches the recursion obtained via PREVTH for 11 ODEs where the unknown vector is the expectation value of the system Evolver Vector, $\widehat{\mathbf{S}}(t)$, and the initial vector is the expectation value of the system vector $\widehat{\mathbf{s}}$ composed of the system operator expectation values. We do not intend to get into details of this issues since they are available through some of our publications.

7 Focusing on the Initial Wave Packet

We have not specifically consider the structure of the initial wave function until now even though the use of Hamiltonian instead of reciprocal powers of the position operator has been implicitly or, more or less, explicitly implying the Hamiltonian eigenfunctions as the initial wave function. If the initial wave function matches one of the eigenfunction of the considered system's Hamiltonian then the evolver of a considered temporally invariant operator does not change, remains temporally constant. Our specific Hamiltonian has both discrete (bound states) and continuous (scattering states) spectra and the union of eigenfunctions for these spectra forms a complete basis set for representing any given initial wave function. Hence, the eigenfunction case for initial wave function is not so interesting in this sense.

If the initial wave function is not expressible in linear combinations of finitely many eigenfunctions of the Hamilton operator then the expectation value of an operator may not be expanded to a temporal Maclaurin series unless the initial wave function's images under all nonnegative integer powers of the Hamilton operator remains in the function space where the wave functions lie. This however may require an essential singularity in the initial wave function at the Hamiltonian's singular point with respect to the position operator. For example an exponential function of a linear combination of x and $1/x$ with negative combination coefficients can be considered to this end.

What we have told above also imply that the the convergence properties of the PREVTH series quite strongly depends on the initial wave function structure. This convergence is also affected by the level of the uncertainty in the system operators expectation values and therefore the uncertainty in the initial wave function. We do not intend to get into the details of these issues here. We are conducting studies on these issues and planning to report them when done.

8 Concluding Remarks

This work is devoted the first time construction of what we call Evolver Dynamical Equations. We have used hydrogen-like systems as the sample target systems and constructed two different set of equations by using two different operator basis set which is conically closed under the Poisson Bracket operation with the system Hamiltonian with the special PREVTH conicality. We have also given certain details of the fluctuation free expectation

 ISBN: 978-84-617-8694-7

value evaluation. Many important part of this study is based on the Probabilistic Evolution Theory (PREVTH) recently developed in our group studies whose relevant publications can be found in the scientific literature.

We have introduced two arbitrary constant in our formulations. These parameters can be optimised to extend the convergence properties of PREVTH solution to larger domain. This is given in our another paper, where the introduction possibilities of some other flexibilities for optimisation will be mentioned, in this conference.

References

[1] M. Demiralp, *Quantum Expected Value Dynamics in Probabilistic Evolution Perspective*, Proceedings of the 12th International Conference on Computational and Mathematical Methods in Science and Engineering (CMMSE), ISBN: 978-84-615-5392-1, Murcia, Spain (2012) 449-459.

[2] M. Demiralp, Journal of Mathematical Chemistry, URL doi:http://dx.doi. org/10.1007/s10910-012-0079-6 **51**(4) (2012) 1170.

[3] M. Demiralp and B. Tunga, Journal of Mathematical Chemistry, URL doi:http:// dx.doi.org/10.1007/s10910-012-0081-z **51**(4) (2012) 1198.

[4] M. Demiralp, E. Demiralp and L. Hernandez-Garcia, Journal of Mathematical Chemistry, URL doi:http://dx.doi.org/10.1007/s10910-011-9929-x **50** (2012) 850.

[5] M. Demiralp and E. Demiralp, Journal of Mathematical Chemistry, URL doi: http://dx.doi.org/10.1007/s10910-012-0070-2 **51**(1) (2012) 58.

[6] M. Demiralp and E. Demiralp, Journal of Mathematical Chemistry, URL doi: http://dx.doi.org/10.1007/s10910-012-0064-0 **51**(19) (2012) 38.

[7] S. Tuna and M. Demiralp, *Certain Validations of Probabilistic Evolution Approach for Initial Value Problems*, Proceedings of the 12th WSEAS International Conference on Systems Theory and Scientific Computation (ISTASC'12), ISBN: 978-1-61804-115-9, İstanbul, Türkiye (2012) 246-249.

[8] N. A. Baykara, E. Gürvit and M. Demiralp, *Univariate single quantum harmonic oscillator from probabilistic evolution perspective*, Proceedings of the 13th WSEAS International Conference on Mathematics and Computers in Biology and Chemistry (MCBC'12), Wisconsin, ABD (2012) 27-32.

[9] M. Ayvaz and M. Demiralp, *Getting Triangularity and Conicality in the Probabilistic Evolutionary Expectation Dynamics of the Purely Quartic Quantum Anharmonic Oscillator*, Proceedings of the 12th WSEAS International Conference on Systems Theory and Scientific Computation (ISTASC'12), ISBN: 978-1-61804-115-9, İstanbul, Türkiye (2012) 268-271.

[10] E. Demiralp, M. Demiralp and L. Hernandez-Garcia, *Journal of Mathematical Chemistry*, URL doi:http://dx.doi.org/10.1007/s10910-011-9930-4 **50** (2012) 870.

[11] B. Tunga and M. Demiralp, *Probabilistic Evolution of the State Variable Expected Values in Liouville Equation Perspective, for a Many Particle System Interacting Via Elastic Forces*, Proceedings of the 12th International Conference on Computational and Mathematical Methods in Science and Engineering (ICCMSE), ISBN: 978-84-615-5392-1, Murcia, Spain (2012) 1186-1197.

[12] M. Demiralp, *Squarificating the Telescope Matrix Images of Initial Value Vector in Probabilistic Evolution Theory (PET)*, Proceedings of the 19th International Conference on Applied Mathematics (AMATH'14), ISBN: 978-1-61804-258-3, İstanbul, Türkiye (2014) 99104.

[13] M. E. Kirkin and C. Gözükirmizi, *Probabilistic Evolution Theory for ODE Sets with Second Degree Multinomial Right Hand Side Functions: Certain Reductive Cases* ICCMSE, Athens, Greece (2015).

[14] C. Gözükirmizi, *Probabilistic Evolution Theory for ODE Sets with Second Degree Multinomial Right Hand Side Functions: Implementation*, ICCMSE, Athens, Greece (2015).

[15] C. Gözükirmizi and M. E. Kirkin, *Classical Symmetric Fourth Degree Potential Systems in Probabilistic Evolution Theoretical Perspective: Most Facilitative Conicalization and Squarification of Telescope Matrices*, International Conference in Nonlineer Problems in Aviation and Aerospace (ICNPAA), La Rochelle, France (2016).

[16] C. Gözükirmizi, M. E. Kirkin and M. Demiralp, Journal of Mathematical Chemistry, URL doi:http://dx.doi.org/10.1007/s10910-016-0678-8 (2017) 1-20.

[17] B. Kalay, M. Demiralp, The proceedings of CMMSE 2017, Initial Wavefunction Construction for Probabilistic Evolution Theoretical (PREVTH) Evolver Dynamics via PREVTH Parameters and Initial Wave Function Optimization.

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

Initial Wavefunction Construction for Probabilistic Evolution Theoretical (PREVTH) Evolver Dynamics via PREVTH Parameters and Initial Wave Function Optimization

Berfin Kalay[1] **and Metin Demiralp**[1]

[1] *Informatics Institute, Computational Science and Engineering Department, İstanbul Technical University*

emails: berfinkalay@gmail.com, metin.demiralp@gmail.com

Abstract

This is somehow a companion paper to our another contribution relevant to the evolver dynamical equations construction for a quantum system whose Hamiltonian has a singular point in finite domain of the position operator, in this conference. We focus on again the hydrogen-like systems as sample target and take the operator basis set which involves Hamilton operator's positive integer powers instead of position operator's integer reciprocal powers. We optimise β and a parameters appearing in the PREVTH format to extend the convergence properties validity region to larger domains. We propose to use exponential weight function which has essential singularities in both 0 and infinite values of the position variable. This weight's parameters are also optimised to suppress the uncertainty such that the smoothest element of the basis function set to represent the initial wave function becomes the least oscillating basis function. Even though the fluctuation expansion is also possible we do not involve discussions to this end since it complicates the presentation unnecessarily too much.

Key words: Evolver dynamics, Expectation values, Singular Hamiltonian, Mathematical fluctuation theory.

1 Probabilistic Evolution Theoretical Introduction

In our another contribution [1] to this conference, which is somehow companion to this paper, we have developed a new method based on the Heisenberg picture of quantum mechanics and combined with the Probabilistic Evolution Theory (PREVTH) [2–15], recently-developed in our group studies. We called the method "Evolver Dynamics in Probabilistic

Evolution Theoretical Perspective". Evolver is a unitary superoperator producing a linear operator from another but given operator and is the unknown of the Heisenberg Picture equations. It uses the interaction representation operator connecting the wave function of the considered system at a time instant to the same function but at a different time instant. It satisfies the Schrödinger equation with the unit operator initial imposition.

In the companion paper we have focused on hydrogen-like systems Which are composed of two electrically charged particles (mathematically point objects) mutually interacting through a Coulomb force. The Hamiltonian of such a system can be expressed as follows

$$\widehat{H} \equiv \frac{1}{2\mu}\widehat{p}^{\,2} - \alpha\widehat{q}^{\,-1} \tag{1}$$

where μ denotes the system's reduced mass whose reciprocal is equal to the sum of particles' mass reciprocals. This is in fact a reduced Hamiltonian which is obtained after (i) separating the mass center coordinates and related Hamiltonian part, (ii) then passing through the spherical coordinates constructed on the relative coordinates system which assumes that one of the particles is as if motionless. The operators $\widehat{p}$ and $\widehat{q}$ respectively stand for the momentum and position operators which are defined through the following equalities.

$$\widehat{p}f(x) \equiv -i\hbar\frac{\partial f(x)}{\partial x}, \qquad \widehat{q}f(x) \equiv xf(x), \qquad x \in [0,\infty) \tag{2}$$

where we call x position variable (note that we have used different symbols for position operator and position variable). Hamilton operator for this system has singularity at $x = 0$ as can be seen immediately.

PREVTH aims to use a system operator vector $\widehat{\mathbf{s}}$ which has finite number, say n, operator elements such that each element's Poisson Bracket with the system Hamiltonian is a second degree multinomial. To this end various types of space extensions are used as long as the set of these elements is multinomially closed under the Poisson Bracketing operation. The conicality (second degree multinomiality) is in such a way that we can write

$$\left\{\widehat{H}, \widehat{\mathbf{s}}\right\} = \beta\widehat{\mathbf{s}} + \mathbf{F}\widehat{\mathbf{s}}^{\otimes 2} \tag{3}$$

where β is an arbitrary scalar at the moment while $\widehat{\mathbf{s}}$ may have some finite number of operator elements. This finite number may be varying depending on how the space extensions are used. In other words this form is not unique in dimension even though there are a lot of signals for the existence of minimum dimensional space extension. In the companion of this paper we have shown that 7 operator elements can remain conically closed under Poisson Bracketing such that the linear term coefficient is proportional to identity matrix (PREVTH format) when we have used momentum and position operators together with certain integer powers of position operator reciprocal. We have also constructed another similar operator basis set which contains certain positive integer powers of momentum and

 ISBN: 978-84-617-8694-7

Hamilton operators together with the position operator. In this case the system vector of operator elements has 11 elements whose explicit structures are given below (we prefer to focus on this case since the singularity is somehow hidden in the Hamiltonian)

$$\widehat{\mathbf{s}} \equiv [\, \widehat{s}_1 \;\; ... \;\; \widehat{s}_{10} \,] \equiv \left[\widehat{p} \;\; \widehat{p}^{\,2} \;\; \widehat{H} \;\; \widehat{H}\,\widehat{p} \;\; \widehat{p}\,\widehat{H} \;\; \widehat{H}^{\,2} \;\; \widehat{p}^{\,3} \;\; \widehat{p}^{\,2}\,\widehat{H} \;\; \widehat{H}\,\widehat{p}^{\,2} \;\; a\widehat{I} \right] \tag{4}$$

This leads us to construct the following equalities

$$\begin{aligned}
\left\{\widehat{H},\widehat{s}_1\right\} &= -\tfrac{1}{4\mu^2\alpha}\widehat{s}_2^2 + \tfrac{1}{2\mu\alpha}\widehat{s}_2\widehat{s}_3 + \tfrac{1}{2\mu\alpha}\widehat{s}_3\widehat{s}_2 \\
&\quad - \tfrac{1}{\alpha}\widehat{s}_3^2 + \beta\widehat{s}_1 - \tfrac{\beta}{a}\widehat{s}_1\widehat{s}_{10} \\
\left\{\widehat{H},\widehat{s}_2\right\} &= -\tfrac{1}{2\mu^2\alpha}\widehat{s}_2\widehat{s}_7 + \tfrac{1}{2\mu\alpha}\left(\widehat{s}_2\widehat{s}_4 + \widehat{s}_4\widehat{s}_2\right) \\
&\quad - \tfrac{1}{\alpha}\widehat{s}_3\widehat{s}_4 + \beta\widehat{s}_2 - \tfrac{\beta}{a}\widehat{s}_2\widehat{s}_{10} \\
\left\{\widehat{H},\widehat{s}_3\right\} &= \beta\widehat{s}_3 - \tfrac{\beta}{a}\widehat{s}_3\widehat{s}_{10} \\
\left\{\widehat{H},\widehat{s}_4\right\} &= -\tfrac{1}{4\mu^2\alpha}\widehat{s}_9\widehat{s}_2 + \tfrac{1}{2\mu\alpha}\left(\widehat{s}_4\widehat{s}_5 + \widehat{s}_6\widehat{s}_2\right) \\
&\quad - \tfrac{1}{\alpha}\widehat{s}_3\widehat{s}_6 + \beta\widehat{s}_4 - \tfrac{\beta}{a}\widehat{s}_4\widehat{s}_{10} \\
\left\{\widehat{H},\widehat{s}_5\right\} &= -\tfrac{1}{4\mu^2\alpha}\widehat{s}_2\widehat{s}_8 + \tfrac{1}{2\mu\alpha}\widehat{s}_2\widehat{s}_6 + \tfrac{1}{2\mu\alpha}\widehat{s}_4\widehat{s}_5 \\
&\quad - \tfrac{1}{\alpha}\widehat{s}_3\widehat{s}_6 + \beta\widehat{s}_5 - \tfrac{\beta}{a}\widehat{s}_5\widehat{s}_{10} \\
\left\{\widehat{H},\widehat{s}_6\right\} &= \beta\widehat{s}_6 - \tfrac{\beta}{a}\widehat{s}_6\widehat{s}_{10} \\
\left\{\widehat{H},\widehat{s}_7\right\} &= -\tfrac{3}{4\mu^2\alpha}\widehat{s}_7^2 + \tfrac{1}{\mu\alpha}\widehat{s}_8\widehat{s}_2 + \tfrac{1}{2\mu\alpha}\widehat{s}_9\widehat{s}_2 \\
&\quad - \tfrac{1}{\alpha}\widehat{s}_3\widehat{s}_9 + \tfrac{1}{2\mu\alpha}\widehat{s}_7\widehat{s}_4 + \tfrac{1}{2\mu\alpha}\widehat{s}_5\widehat{s}_7 \\
&\quad - \tfrac{1}{\alpha}\widehat{s}_5\widehat{s}_4 + \tfrac{1}{2\mu\alpha}\widehat{s}_2\widehat{s}_8 - \tfrac{1}{\alpha}\widehat{s}_8\widehat{s}_3 \\
&\quad + \beta\widehat{s}_7 - \tfrac{\beta}{a}\widehat{s}_7\widehat{s}_{10} \\
\left\{\widehat{H},\widehat{s}_8\right\} &= -\tfrac{1}{2\mu^2\alpha}\widehat{s}_7\widehat{s}_8 + \tfrac{1}{2\mu\alpha}\left(\widehat{s}_8\widehat{s}_5 + \widehat{s}_4\widehat{s}_8\right) \\
&\quad - \tfrac{1}{\alpha}\widehat{s}_6\widehat{s}_5 + \tfrac{1}{2\mu\alpha}\widehat{s}_7\widehat{s}_6 + \tfrac{1}{2\mu\alpha}\widehat{s}_5\widehat{s}_8 \\
&\quad - \tfrac{1}{\alpha}\widehat{s}_5\widehat{s}_6 + \beta\widehat{s}_8 - \tfrac{\beta}{a}\widehat{s}_8\widehat{s}_{10} \\
\left\{\widehat{H},\widehat{s}_9\right\} &= -\tfrac{1}{2\mu^2\alpha}\widehat{s}_9\widehat{s}_7 + \tfrac{1}{2\mu\alpha}\widehat{s}_9\widehat{s}_4 + \tfrac{1}{2\mu\alpha}\widehat{s}_6\widehat{s}_7 \\
&\quad - \tfrac{1}{\alpha}\widehat{s}_6\widehat{s}_4 + \tfrac{1}{2\mu\alpha}\widehat{s}_4\widehat{s}_8 + \tfrac{1}{2\mu\alpha}\widehat{s}_4\widehat{s}_9 \\
&\quad - \tfrac{1}{\alpha}\widehat{s}_4\widehat{s}_6 + \beta\widehat{s}_9 - \tfrac{\beta}{a}\widehat{s}_9\widehat{s}_{10} \\
\left\{\widehat{H},\widehat{s}_{10}\right\} &= \beta\widehat{s}_{10} - \tfrac{\beta}{a}\widehat{s}_{10}^2
\end{aligned} \tag{5}$$

which can be rewritten in much more concise formula as follows

$$\dot{\widehat{\mathbf{s}}}(t) = \beta\,\widehat{\mathbf{s}}(t) + \mathbf{F}\,\widehat{\mathbf{s}}(t)^{\otimes 2} \tag{6}$$

where $\mathbf{F}$ is an 11×121 type rectangular matrix whose 54 non-zero elements of 1331 total

 ISBN: 978-84-617-8694-7

elements are explicitly given below

$$
\begin{aligned}
&\bullet\ F_{1,10}=-\frac{\beta}{a},\quad F_{1,12}=-\frac{1}{4\alpha\mu^2},\quad F_{1,13}=\frac{1}{2\alpha\mu},\quad F_{1,22}=\frac{1}{2\alpha\mu},\quad F_{1,23}=-\frac{1}{\alpha};\\
&\bullet\ F_{2,14}=\frac{1}{2\alpha\mu},\quad F_{2,17}=-\frac{1}{2\alpha\mu^2},\quad F_{2,20}=-\frac{\beta}{a},\quad F_{2,24}=-\frac{1}{\alpha}\quad F_{2,32}=\frac{1}{2\alpha\mu};\\
&\bullet\ F_{3,30}=-\frac{\beta}{a};\\
&\bullet\ F_{4,26}=-\frac{1}{\alpha},\quad F_{4,35}=\frac{1}{2\alpha\mu},\quad F_{4,40}=-\frac{\beta}{a},\quad F_{4,52}=\frac{1}{2\alpha\mu},\quad F_{4,82}=-\frac{1}{4\alpha\mu^2};\\
&\bullet\ F_{5,16}=\frac{1}{2\alpha\mu},\quad F_{5,18}=-\frac{1}{4\alpha\mu^2},\quad F_{5,26}=-\frac{1}{\alpha},\quad F_{5,35}=\frac{1}{2\alpha\mu},\quad F_{5,50}=-\frac{\beta}{a};\\
&\bullet\ F_{6,60}=-\frac{\beta}{a};\\
&\bullet\ F_{7,18}=\frac{1}{2\alpha\mu},\quad F_{7,29}=-\frac{1}{\alpha},\quad F_{7,44}=-\frac{1}{\alpha},\quad F_{7,47}=\frac{1}{2\alpha\mu},\quad F_{7,64}=\frac{1}{2\alpha\mu},\\
&\quad F_{7,67}=-\frac{3}{4\alpha\mu^2},\quad F_{7,70}=-\frac{\beta}{a}\quad F_{7,72}=\frac{1}{\alpha\mu}\quad F_{7,73}=-\frac{1}{\alpha},\quad F_{7,82}=\frac{1}{2\alpha\mu};\\
&\bullet\ F_{8,38}=\frac{1}{2\alpha\mu},\quad F_{8,46}=-\frac{1}{\alpha},\quad F_{8,48}=\frac{1}{2\alpha\mu},\quad F_{8,55}=-\frac{1}{\alpha},\quad F_{8,66}=\frac{1}{2\alpha\mu},\\
&\quad F_{8,68}=-\frac{1}{2\alpha\mu^2},\quad F_{8,75}=\frac{1}{2\alpha\mu},\quad F_{8,80}=-\frac{\beta}{a};\\
&\bullet\ F_{9,36}=-\frac{1}{\alpha},\quad F_{9,38}=\frac{1}{2\alpha\mu},\quad F_{9,39}=\frac{1}{2\alpha\mu},\quad F_{9,54}=-\frac{1}{\alpha},\quad F_{9,57}=\frac{1}{2\alpha\mu},\\
&\quad F_{9,84}=\frac{1}{2\alpha\mu},\quad F_{9,87}=-\frac{1}{2\alpha\mu^2},\quad F_{9,90}=-\frac{\beta}{a};\\
&\bullet\ F_{10,100}=-\frac{\beta}{a}.
\end{aligned}
\tag{7}
$$

where the bullet symbol has been used to mark rows of $\mathbf{F}$.

These equations imply the following decomposition

$$\mathbf{F}=\frac{\beta}{a}\mathbf{F}_{\frac{\beta}{a}}+\frac{1}{a}\mathbf{F}_{\frac{1}{a}}+\mathbf{F}_R \tag{8}$$

where right hand side subscripted $\mathbf{F}$ matrices do not depend on PREVTH parameters, either β or a. Dependence on given α and μ parameters are existing as expected.

In the last formula here, a is an arbitrary constant at this moment while $\widehat{s}_{11}$ definition has been introduced to the operator definitions to provide the very specific linear structure in (3). This is based on the facts that (i) the degree of any operator pre or post multiplied by $\widehat{s}_{11}$ is increased by one. This enables us to push any desired first degree operator to the second degree operators, (ii) the square of $\widehat{s}_{11}$ is proportional to itself, (iii) any operator

 ISBN: 978-84-617-8694-7

inserted to first degree operators can be balanced by adding its product with an appropriate multiple of $\widehat{s}_{11}$. All these enable us to write

$$\widehat{T}\left(t,t_0\right)^{\dagger}\left\{\widehat{H},\widehat{\mathbf{s}}\right\}\widehat{T}\left(t,t_0\right)=\beta\,\widehat{\mathbf{S}}(t)+\mathbf{F}\,\widehat{\mathbf{S}}(t)^{\otimes 2},\qquad \widehat{\mathbf{S}}(t)\equiv\widehat{T}\left(t,t_0\right)^{\dagger}\,\widehat{\mathbf{s}}\,\widehat{T}\left(t,t_0\right). \tag{9}$$

This implies

$$\dot{\widehat{\mathbf{S}}}(t)=\beta\,\widehat{\mathbf{S}}(t)+\mathbf{F}\,\widehat{\mathbf{S}}(t)^{\otimes 2},\qquad \widehat{\mathbf{S}}(t)\equiv\widehat{T}\left(t,t_0\right)^{\dagger}\,\widehat{\mathbf{s}}\,\widehat{T}\left(t,t_0\right) \tag{10}$$

where $\psi\left(\mathbf{x},t\right)=\widehat{T}\left(\widehat{p},\widehat{q},t,t_0\right)\psi\left(\mathbf{x},t_0\right)$ and

$$i\hbar\frac{\partial\widehat{T}\left(\widehat{p},\widehat{q},t,t_0\right)}{\partial t}=\widehat{H}\left(\widehat{p},\widehat{q},t\right)\widehat{T}\left(\widehat{p},\widehat{q},t,t_0\right),\qquad \widehat{T}\left(\widehat{p},\widehat{q},t_0,t_0\right)=\widehat{I}. \tag{11}$$

The last equation has been obtained from the Schrödinger equation which is a linear parabolic partial differential equation in temporal and spatial coordinates. However, we do not want to use the wave function to determine the expectation values of the observables related to the system under consideration and this is the main reason why we develop evolver dynamics.

The solutions of the PREVTH equations for evolver dynamics in (10) is the basic goal of PREVTH and we can write the solution as follows

$$\widehat{\mathbf{S}}\left(u(t)\right)=\left(1+\beta u(t)\right)\sum_{j=0}^{\infty}\frac{1}{j!}u(t)^{j}\mathbf{v}_j\left(\widehat{\mathbf{S}}_0\right),\qquad u(t)\equiv\frac{\mathrm{e}^{\beta(t-t_0)}-1}{\beta} \tag{12}$$

where $\mathbf{v}_j$ vectors satisfy the following recursion

$$\mathbf{v}_j\left(\widehat{\mathbf{S}}_0\right)=\sum_{k=0}^{j-1}\binom{j-1}{k}\left\lfloor\mathbf{F},\mathbf{v}_k\left(\widehat{\mathbf{S}}_0\right)\right\rceil\mathbf{v}_{j-1-k}\left(\widehat{\mathbf{S}}_0\right),\quad j=1,2,3,...\qquad \mathbf{v}_0\left(\widehat{\mathbf{S}}_0\right)\equiv\widehat{\mathbf{S}}_0 \tag{13}$$

and we have used the squarified telescope matrix (SquTelMat) notation which is defined as follows

$$\left\lfloor\mathbf{F},\mathbf{v}_k\left(\widehat{\mathbf{S}}_0\right)\right\rceil\equiv\sum_{\ell=1}^{7}v_{\ell}^{(k)}\mathbf{F}_{\ell},\qquad \mathbf{F}\equiv\left[\,\mathbf{F}_1\;\;...\;\;\mathbf{F}_7\,\right]^{T},\qquad \mathbf{v}_k\equiv\left[\,v_1^{(k)}\;\;...\;\;v_7^{(k)}\,\right]^{T} \tag{14}$$

On the other hand, $\widehat{\mathbf{S}}_0$ is equal to $\widehat{\mathbf{S}}(0)$ and therefore to $\widehat{\mathbf{s}}$.

Now each $\widehat{s}_i$ $(i=1,2,...,11)$ operator can be written in terms of relevant fluctuations as follows $\widehat{s}_i\equiv\langle\widehat{s}_i\rangle\,\widehat{I}+\widehat{\phi}_i,\qquad i=1,2,3,...,7$ which are definitions of the related fluctuation operators at the same time and imply that the expectation values of all fluctuation operators vanish. The fluctuationlessness theorem dictates us that the expectation value of a multivariate function depending on certain operators is equal to the image of all those

 ISBN: 978-84-617-8694-7

operator's expectation values under that multivariate function at the zero fluctuation limit. This enables us to replace (12) with its following fluctuationless counterpart.

$$\mathbf{v}_j\left(\langle\widehat{\mathbf{s}}\rangle\right)=\sum_{k=0}^{j-1}\binom{j-1}{k}\left\lfloor\mathbf{F},\mathbf{v}_k\left(\langle\widehat{\mathbf{s}}\rangle\right)\right\rceil\mathbf{v}_{j-1-k}\left(\langle\widehat{\mathbf{s}}\rangle\right),\quad j=1,2,3,...;\quad \mathbf{v}_0\left(\langle\widehat{\mathbf{s}}\rangle\right)\equiv\langle\widehat{\mathbf{s}}\rangle \tag{15}$$

where we have replaced $\widehat{\mathbf{S}}_0$ with its equivalent, $\widehat{\mathbf{s}}$.

2 Optimization of Fundamental PREVTH Parameters

As we have seen in the previous section, β and a constants are arbitrary parameters appearing in the PREVTH equations relevant to evolver dynamics. Hence the determination of these parameters under certain conditions gains a lot of importance. We consider the convergence of the PREVTH solution as most important issue and we can try to choose β and a in such a way that the convergence rules in a largest domain. This can be accomplished via optimization. We can write the following inequality from (12) to this end

$$\left\|\sum_{j=0}^{\infty}\frac{1}{j!}u(t)^j\mathbf{v}_j\left(\widehat{\mathbf{s}}\right)\right\|\leq\sum_{j=0}^{\infty}\frac{1}{j!}\left(u(t)^*u(t)\right)^{\frac{j}{2}}\left\|\mathbf{v}_j\right\| \tag{16}$$

where we have used Frobenious vector norm definition which is submultiplicative. The right hand side can be enlarged by replacing the norm of the vector $\mathbf{v}_j$ with a bound b_j which is valid for all nonnegative integer j values. That is, $\|\mathbf{v}_j\|\leq b_j,(j=0,1,2,...)$ On the other hand we can write the following formulae from (13)

$$\begin{aligned}\left\|\left\lfloor\mathbf{F},\left\langle\,\mathbf{v}_k\left(\widehat{\mathbf{s}}\right)\right\rangle\right\rceil\left\langle\,\mathbf{v}_{j-1-k}\left(\widehat{\mathbf{s}}\right)\right\rangle\right\|&=\left\|\mathbf{F}\left(\left\langle\,\mathbf{v}_k\left(\widehat{\mathbf{s}}\right)\right\rangle\otimes\left\langle\,\mathbf{v}_{j-1-k}\left(\widehat{\mathbf{s}}\right)\right\rangle\right)\right\|\\&\leq\|\mathbf{F}\|\left\|\mathbf{v}_j\right\|\left\|\mathbf{v}_{j-k-1}\right\|\leq\rho\left(\mathbf{F}\right)b_kb_{j-k-1},\qquad j=1,2,3,...\;\;k=0,1,2,...,j-1\end{aligned} \tag{17}$$

where $\rho\left(\mathbf{B}\right)$ stands for the spectral radius (that is, maximum singular value) of $\mathbf{F}$. Since b_j is defined via an inequality its choice is somehow at our disposal. So it is quite reasonable to impose the following equality which can be derived from (15)

$$\frac{b_j}{(j-1)!}=\rho\left(\mathbf{F}\right)\sum_{k=0}^{j-1}\frac{b_k}{k!}\frac{b_{j-k-1}}{(j-k-1)!},\qquad j=1,2,3,... \tag{18}$$

which is in fact a recursion to determine b_js. For the solution we can use the generating function method. If we multiply the both sides of this equation by z^j where z is a scalar variable and then sum over all positive integer j values we can write

$$\sum_{j=1}^{\infty}z^j\frac{b_j}{(j-1)!}=\rho\left(\mathbf{F}\right)\sum_{j=1}^{\infty}\sum_{k=0}^{j-1}z^j\frac{b_k}{k!}\frac{b_{j-k-1}}{(j-k-1)!},\quad j=1,2,3,...;\quad b_0=\left\|\left\langle\widehat{\mathbf{s}}\right\rangle\right\| \tag{19}$$

 ISBN: 978-84-617-8694-7

which can also be rewritten in the following concise form

$$\frac{dG(z)}{dz} = \rho\left(\mathbf{F}\right) G(z)^2, \qquad G(0) = \| \langle \widehat{\mathbf{s}} \rangle \|, \qquad G(z) \equiv \sum_{j=0}^{\infty} \frac{b_j}{j!} z^j \tag{20}$$

whose unique solution is as follows

$$G(z) = \frac{\| \langle \widehat{\mathbf{s}} \rangle \|}{1 - \rho\left(\mathbf{F}\right) \| \langle \widehat{\mathbf{s}} \rangle \| z}. \tag{21}$$

If we now use bound for v_j in (16) and replace z by $\sqrt{(u(t)^* u(t))}$ then (16) can be rewritten as follows

$$\left\| \sum_{j=0}^{\infty} \frac{1}{j!} u(t)^j \mathbf{v}_j\left(\widehat{\mathbf{s}}\right) \right\| \leq G\left(\sqrt{u(t)^* u(t)}\right) = \frac{\| \langle \widehat{\mathbf{s}} \rangle \|}{1 - \rho\left(\mathbf{F}\right) \| \langle \widehat{\mathbf{s}} \rangle \| \sqrt{u(t)^* u(t)}}. \tag{22}$$

which brings the following condition on time

$$\rho\left(\mathbf{F}\right) \| \langle \widehat{\mathbf{s}} \rangle \| \sqrt{u(t)^* u(t)} < 1. \tag{23}$$

The left hand side expression at the left of this inequality depends on β and a and therefore can be optimised to diminish this expression as much as possible. To this end, we need to find the functional dependences of this expression on these parameters to get results precise as much as possible. We can write first

$$| \langle \widehat{\mathbf{s}} \rangle | = \sqrt{\nu^2 + |a|^2} < \nu + |a|, \qquad \nu^2 \equiv \sum_{j=1}^{10} | \langle \widehat{s}_j \rangle |^2 \tag{24}$$

where the entities between single vertical pipe symbols denote the complex modulus of the relevant scalar entity. In fact all $\widehat{s}_j$ operators which are self-adjoint have real valued expectation values. All these expectation values will depend on which structure is used for the initial wave function. a does not appear in the first 10 elements of system vector. In this formula we have assumed that the unknown scalar a can take complex values also.

On the other hand, the spectral radius of the $\mathbf{F}$ matrix is also a norm (induced, spectral norm) and therefore urges us to write

$$\rho\left(\mathbf{F}\right) = \frac{|\beta|}{|a|} \rho\left(\mathbf{F}_{\frac{\beta}{a}}\right) + \frac{1}{|a|} \rho\left(\mathbf{F}_{\frac{1}{a}}\right) + \rho\left(\mathbf{F}_R\right), \qquad \rho\left(\mathbf{F}_{\frac{\beta}{a}}\right) = 1, \qquad \rho\left(\mathbf{F}_{\frac{1}{a}}\right) = \frac{1}{\mu} \tag{25}$$

where β and a are unknown scalars as we stated previously and we have skipped the intermediate details of rightmost two equalities.

 ISBN: 978-84-617-8694-7

We can also write the following formulae for the temporal entity $u(t)$

$$\begin{aligned}\sqrt{(u(t)^{*}u(t))} &= |u(t) = \left|\frac{\mathrm{e}^{\beta(t-t_0)}-1}{\beta}\right| = \left|\sum_{j=1}^{\infty}\frac{\beta^{j-1}}{j!}(t-t_0)^j\right| < \sum_{j=1}^{\infty}\frac{|\beta|^{j-1}}{j!}T^j \\ &< \frac{\mathrm{e}^{|\beta|}-1}{|\beta|}, \qquad t\in[\,t_0,t_0+T\,] \end{aligned} \tag{26}$$

where the positive constant T denotes the evolution time duration.

Entire analysis of this section until now urges us to define the following cost functional

$$J(\beta,a)\equiv\left[\left(|\beta|+\frac{1}{\mu}\right)\frac{1}{|a|}+\rho\left(\mathbf{F}_R\right)\right](\nu+|a|)\,\frac{\mathrm{e}^{|\beta|}-1}{|\beta|}, \qquad t\in[\,t_0,t_0+T\,] \tag{27}$$

whose less than one values for some T values correspond to the convergence of PREVTH solution. However, this depends on β and a and can be minimized with respect to these parameters. For this minimization J must be separately differentiated with respect to these parameters and then each resulted equation should be set equal to zero. This gives two coupled algebraic equations which are quite nonlinear and therefore can not be solved analytically even though various numerical procedures to get the solution can be used to this end.

Despite we have assumed that the parameters β and a can take complex values, the above analysis involves only complex moduli of these parameters. This is because of the pessimism in the construction of the above cost functional. We have used in fact quite loose bounds in this construction such that only the complex moduli appeared in the results and therefore all phase related terms have been somehow rounded off. The utilization of much more tight bounds through rigorous analyses will bring the dependences on phases also, beside the moduli. However, this is a quite cumbersome procedure and we do not intend to get into further details of this issue.

3 Further Convergence Maximization Possibilities

The analysis in the previous section is based on the PREVTH parameters since they are the only arbitrary entities to be optimised in the PREVTH solution. However, it is also possible to introduce further arbitrarinesses into the rectangular matrix $\mathbf{F}$. One important thing to this end is the commutativity of the system operators. A careful look at (4) reveals that, the first, second, and, third powers of the momentum operator are mutually commutative. At the same time the first and second powers of the Hamilton operator are also commutative. Beyond that $\widehat{s}_{11}$ is commutative with all remaining operators. All these mean the existence of totally 14 commutativity relations. The commutativity between any given two operators, $\widehat{o}_1$ and $\widehat{o}_2$, enables us to write $c\,(\widehat{o}_1\widehat{o}_2-\widehat{o}_2\widehat{o}_1)=0$ where c is an arbitrary

 ISBN: 978-84-617-8694-7

coefficient. This permits us to add this kind vanishing entities with different c coefficients and different operator pairs to the right hand side of each equation in (5). This enters totally 154 arbitrary c coefficients in the structure of $\mathbf{F}$, and therefore, of $\mathbf{F}_R$. Thus, there is a possibility to suppress the spectral norm of this matrix. This optimisation is based on the optimisation of a quadratic form in cs. However, we do not intend to get into details of this issue here furthermore.

4 Basis Function Set for Representing Initial Wave Function

In last two sections we have not specified the structure of the initial wave function. One way to choose initial wave function is to use eigenfunctions of the system Hamiltonian. However, the expectation values of integer reciprocal powers can not be evaluated through these eigenfunctions as long as the integer power of the position operator remains greater than or equal to -2. On the other hand, the cases where the initial wave function is one of the eigenfunctions corresponds to the separation of the temporal and spatial dependences and is very well-known. Hence, the profits coming from PREVTH utilisation may not be considered so much important and therefore is not so interesting in the PREVTH point of view.

The other way to use a different basis set is to take the singularity in the Hamiltonian to the consideration. Since the Hamiltonian is singular at both 0 and ∞ for x, the function whose image under Hamiltonian must have some specific behavior at these singularities. Otherwise the image function becomes having singularities at these values of the position variable. As told above, Hamiltonian eigenfunctions do not give singularities at these points. However, they are insufficient to evaluate the expectation values of the greater than two powers of the position operator reciprocal. So, in this case, we intend to construct such a basis set that the expectation values of all powers (positive and nonpositive) can be evaluated through these basis functions. This enforces us to use a weight function and insert such a behaviour to this weight that all negative and nonnegative powers in expectation values can be suppressed by this weight function. We can write the following structure which is composed of the weight function alone for the initial wave function as a first step to be taken for the set construction.

$$\psi_0(x) \equiv cx^{-\frac{1}{4}} \mathrm{e}^{-\left(\frac{\nu_1}{2x}+\frac{\nu_2}{2}x\right)+i\nu_3 x}, \qquad x \in [0,\infty) \tag{28}$$

where c and νs are positive parameters and we need to find the values of these parameters in such a way that this function should have unit norm and can be able to produce the given expectation values of the position and momentum operators. This leaves one parameter unknown and we can connect that parameter to the standard deviation to get uniqueness. For normalization we need to evaluate the modulus square integral of this function's except

 ISBN: 978-84-617-8694-7

the c factor. We can write

$$\int_0^\infty dx x^{-\frac{1}{2}} \mathrm{e}^{-\left(\frac{\nu_1}{x}+\nu_2 x\right)} = 2\mathrm{e}^{-2\sqrt{\nu_1}\sqrt{\nu_2}} \int_0^\infty dx \mathrm{e}^{-\left(\frac{\sqrt{\nu_1}}{x}-\sqrt{\nu_2}x\right)^2} = 2\sqrt{\nu_1}\sqrt{\nu_2}\mathrm{e}^{-2\sqrt{\nu_1}\sqrt{\nu_2}}$$
$$\times \int_0^\infty dx \mathrm{e}^{-\sqrt{\nu_1}\sqrt{\nu_2}\left(\frac{1}{x}-x\right)^2} = 2\sqrt{\nu_1}\sqrt{\nu_2}\mathrm{e}^{-2\sqrt{\nu_1}\sqrt{\nu_2}} \int_0^\infty dx \frac{1}{x^2} \mathrm{e}^{-\sqrt{\nu_1}\sqrt{\nu_2}\left(\frac{1}{x}-x\right)^2} \quad (29)$$

where the last equality implies

$$\int_0^\infty dx \mathrm{e}^{-\sqrt{\nu_1}\sqrt{\nu_2}\left(\frac{1}{x}-x\right)^2} = \frac{1}{2}\int_0^\infty d\left(x-\frac{1}{x}\right) \mathrm{e}^{-\sqrt{\nu_1}\sqrt{\nu_2}\left(\frac{1}{x}-x\right)^2} = \frac{\sqrt{\pi}}{4\left(\nu_1\nu_2\right)^{\frac{1}{4}}} \quad (30)$$

which can be combined with (29) to give

$$\int_0^\infty dx x^{-\frac{1}{2}} \mathrm{e}^{-\left(\frac{\nu_1}{x}+\nu_2 x\right)} = \frac{\sqrt{\pi}}{2}\left(\nu_1\nu_2\right)^{\frac{1}{4}} \mathrm{e}^{-2\sqrt{\nu_1}\sqrt{\nu_2}} \quad (31)$$

This integral therefore can be considered as a generating function and enables us to write the following formulae

$$\int_0^\infty dx x^{j-\frac{1}{2}} \mathrm{e}^{-\left(\frac{\nu_1}{x}+\nu_2 x\right)} = (-1)^j \frac{\partial^j}{\partial \nu_1^j}\left\{\frac{\sqrt{\pi}}{2}\left(\nu_1\nu_2\right)^{\frac{1}{4}} \mathrm{e}^{-2\sqrt{\nu_1}\sqrt{\nu_2}}\right\}, \qquad j=0,1,2,... \quad (32)$$

$$\int_0^\infty dx x^{-j-\frac{1}{2}} \mathrm{e}^{-\left(\frac{\nu_1}{x}+\nu_2 x\right)} = (-1)^j \frac{\partial^j}{\partial \nu_2^j}\left\{\frac{\sqrt{\pi}}{2}\left(\nu_1\nu_2\right)^{\frac{1}{4}} \mathrm{e}^{-2\sqrt{\nu_1}\sqrt{\nu_2}}\right\}, \qquad j=0,1,2,... \quad (33)$$

These equalities are sufficient to evaluate expectation values of all power involving terms. This means that we can also construct a set of polynomials in the position variable's non-negative powers or in its reciprocal powers as a basis set to represent initial wave function. There are of course certain properties facilitating this construction. This issue is under an intense study of us even though we do not report certain further informations. We are planning to tell more issues in the presentations and the possible extended publication.

5 Concluding Remarks

In our two contrubutions to this conference we have constructed evolver dynamics somehow based on the Heisenberg Picture of Quantum Mechanics. We have obtained the super ODEs on operators, the right hand sides of which has a Probabilistic Evolution Theory (PREVTH) specific conicality by using certain type space extensions. This has been accomplished in the previous paper.

This paper is basically devoted to the PREVTH parameter optimisation and the convergence of PREVTH solution has also been revisited. The initial function representation

 ISBN: 978-84-617-8694-7

which is a very important issue in our analysis has also been considered. A specific basis set has been proposed such that its weight function's modulus square can be evaluated analytically. Beyond that all polynomial factors can also be imported to the analytic integration such that integrations can be performed by using partial differentiation on a generating function appearing in the modulus square integral of the weight function.

We have proposed the fluctuation free evaluations for approximation. However, quantum mechanics of our singular systems depend on fluctuation terms beyond that the fluctuationlessness terms. Howeever we have not intended to deal fluctuation expansion here even though that extension seems to be quite straightforward and we are planning to report our future results on this issue as soon as possible.

References

[1] B. Kalay and M. Demiralp, *A Probabilistic Evolution Theoretical (PREVTH) Approach to Quantum Evolver Dynamical Equations for Singular Hamiltonians: Fluctuationlessness Approximation*, Proceedings of the 17th International Conference on Computational and Mathematical Methods in Science and Engineering (CMMSE 2017), Rota, Cadiz, Spain (2017), Companion Paper.

[2] M. Demiralp, *Quantum Expected Value Dynamics in Probabilistic Evolution Perspective*, Proceedings of the 12th International Conference on Computational and Mathematical Methods in Science and Engineering (CMMSE), ISBN: 978-84-615-5392-1, Murcia, Spain (2012) 449-459.

[3] M. Demiralp, Journal of Mathematical Chemistry, URL doi:http://dx.doi. org/10.1007/s10910-012-0079-6 **51**(4) (2012) 1170.

[4] M. Demiralp and B. Tunga, Journal of Mathematical Chemistry, URL doi:http:// dx.doi.org/10.1007/s10910-012-0081-z **51**(4) (2012) 1198.

[5] M. Demiralp, E. Demiralp and L. Hernandez-Garcia, Journal of Mathematical Chemistry, URL doi:http://dx.doi.org/10.1007/s10910-011-9929-x **50** (2012) 850.

[6] M. Demiralp and E. Demiralp, Journal of Mathematical Chemistry, URL doi: http://dx.doi.org/10.1007/s10910-012-0070-2 **51**(1) (2012) 58.

[7] M. Demiralp and E. Demiralp, Journal of Mathematical Chemistry, URL doi: http://dx.doi.org/10.1007/s10910-012-0064-0 **51**(19) (2012) 38.

[8] S. Tuna and M. Demiralp, *Certain Validations of Probabilistic Evolution Approach for Initial Value Problems*, Proceedings of the 12th WSEAS International Conference

on Systems Theory and Scientific Computation (ISTASC'12), ISBN: 978-1-61804-115-9, İstanbul, Türkiye (2012) 246-249.

[9] N. A. Baykara, E. Gürvit and M. Demiralp, *Univariate single quantum harmonic oscillator from probabilistic evolution perspective*, Proceedings of the 13th WSEAS International Conference on Mathematics and Computers in Biology and Chemistry (MCBC'12), Wisconsin, ABD (2012) 27-32.

[10] M. Ayvaz and M. Demiralp, *Getting Triangularity and Conicality in the Probabilistic Evolutionary Expectation Dynamics of the Purely Quartic Quantum Anharmonic Oscillator*, Proceedings of the 12th WSEAS International Conference on Systems Theory and Scientific Computation (ISTASC'12), ISBN: 978-1-61804-115-9, İstanbul, Türkiye (2012) 268-271.

[11] F. Hunutlu, N. A. Baykara and M. Demiralp, *Truncation Approximants to Probabilistic Evolution for ODEs Having Two Diagonal Banded Evolution Matrices Under Initial Conditions: Simple Case*, Proceedings of the 12th International Conference on Computational and Mathematical Methods in Science and Engineering (ICCMSE), ISBN: 978-84-615-5392-1, Murcia, Spain (2012) 720-731.

[12] E. Demiralp, M. Demiralp and L. Hernandez-Garcia, *Journal of Mathematical Chemistry*, URL doi:http://dx.doi.org/10.1007/s10910-011-9930-4 **50** (2012) 870.

[13] B. Tunga and M. Demiralp, *Probabilistic Evolution of the State Variable Expected Values in Liouville Equation Perspective, for a Many Particle System Interacting Via Elastic Forces*, Proceedings of the 12th International Conference on Computational and Mathematical Methods in Science and Engineering (ICCMSE), ISBN: 978-84-615-5392-1, Murcia, Spain (2012) 1186-1197.

[14] M. Demiralp, *Squarificating the Telescope Matrix Images of Initial Value Vector in Probabilistic Evolution Theory (PET)*, Proceedings of the 19th International Conference on Applied Mathematics (AMATH'14), ISBN: 978-1-61804-258-3, İstanbul, Türkiye (2014) 99104.

[15] C. Gözükirmizi, M. E. Kirkin and M. Demiralp, Journal of Mathematical Chemistry, URL doi:http://dx.doi.org/10.1007/s10910-016-0678-8 (2017) 1-20.

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

Efficient local smoothed particle hydrodynamics with precomputed patches

Yasutomo Kanetsuki[1], John C. Wells[2] and Susumu Nakata[3]

[1] *Graduate School of Information Science and Engineering, Ritsumeikan University, Japan*

[2] *College of Science and Engineering, Ritsumeikan University, Japan*

[3] *College of Information Science and Engineering, Ritsumeikan University, Japan*

emails: is0061ee@ed.ritsumei.ac.jp, jwells@se.ritsumei.ac.jp,
snakata@is.ritsumei.ac.jp

Abstract

This paper presents an improved method for smoothed particle hydrodynamics (SPH) within a nested Lagrangian domain of fluid particles. In the previous method, ghost particles, generated around fluid particles using Poisson-disk sampling method, transfer necessary physical quantities of the outer domain to the fluid. Using this technique, the local fluid motion is reproduced which agrees with the results simulated with full SPH. However, the computational cost for generation of ghost particles are non-negligible. In order to reduce the cost, we develop a patch-based sampling method to generate ghost particles. In our approach, the ghost particles are generated locally around each fluid particle for computational efficiency and corresponding physical quantities are determined using the local ghost particles. Furthermore, we introduce a new technique that determines physical quantities of ghost particles to transfer information from the outer domain to the local fluid particles.

Key words: smoothed particle hydrodynamics, fluid simulation

1 Introduction

In this paper, we propose an efficient method of local smoothed particle hydrodynamics (SPH). Particle tracking is a typical and important problem in environmental fluid engineering [1] and the local SPH is designed specifically to the problem [2]. The formulation of the local SPH is based on the conventional SPH which we refer to as full SPH (See Figure

 ISBN: 978-84-617-8694-7

1 for comparison of full SPH and local SPH) in this paper and works under an assumption that global physical quantities of outer fluid flow are given *a priori*. In this formulation, the global quantities are transferred to local fluid particles of the nested Lagrangian domain, using so called ghost particles that are temporally allocated around the local fluid particles. The ghost particles are generated at every time step of SPH based on the technique of Poisson-disk sampling (PDS) so that they surround the fluid particles. This ghost particle generation process is time consuming and is often a bottleneck of speed in practical computation. In order to address this problem, we present a new method to transfer the physical quantities from the outer domain to the fluid particles. In our approach, the ghost particles are generated around each fluid particle individually. This method does not require globally consistent ghost particles and can be performed with lower computational cost.

PDS is a method to generate randomly distributed points following some statistical properties [3]. Many fast PDS techniques have been developed for CPU [4] and GPU [5], and also adopted to fluid simulation with SPH [6]. In our local SPH context, although PDS can generate appropriate ghost particles around the fluid, it is not efficient to operate on the fly. In order to avoid inefficient generation of the particles, we develop patch-based ghost particle sampling.

SPH is a particle-based fluid simulation method, and a review of recent studies is found in [7]. In this paper, we adopt δ-SPH [8, 9] to simulate weakly compressible flow based on the Euler equations and continuity equation. Since the physical quantities evaluated in SPH are generally not equal to those obtained by analytical solutions or other methods, if the physical quantities obtained with such ways are used as outer flow, local SPH does not work correctly. In order to avoid this situation, we introduce a new technique that determines physical quantities of ghost particles appropriately.

2 Improved ghost particles

The previous method uses PDS to generate ghost particles around local fluid particles. This appropriately reproduces local fluid motion and reduces computational cost compared to full SPH. However the computational time of ghost particle generation is not negligible. For more efficient computation, we propose here a new sampling method.

We generate ghost particles around each fluid particle by applying a precomputed patch to the fluid particle, then removing the samples that are too close to fluid particles. The effects of the ghosts at the fluid particle is determined individually using only the ghost particles around the fluid particle after removal. This approach is more efficient than the previous method [2] in the sense that globally consistent ghost particles are not required.

As a precomputation process, many patches are prepared using PDS around one dummy fluid particle. We distribute sampled particles within the kernel effective radius from the dummy, yet outside a minimal distance of any particles of αl (α is a user defined parameter).

 ISBN: 978-84-617-8694-7

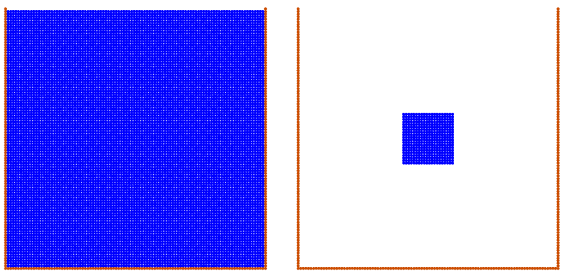


Figure 1: Full SPH (Left) and local SPH (Right). The blue particles indicate the fluid and the blown particles are the wall. The particles are arranged with initial spacing l.

Since the distances from the fluid particle to the samples are known before actual simulation, we avoid the distance evaluation during the simulation.

In order to transfer outer physical quantities to the fluids, we copy the velocity and pressure of the outer domain to each ghost particle and evaluate the i-th ghost's density as $\rho_i^t = \frac{p_i^t}{c_0^2} + \rho_0$, where ρ is the density, p the pressure, c_0 the sound speed and t is time. The density ρ_i^t is then used for the density diffusive term in δ-SPH while the other terms are evaluated with rest density for weak compressibility. This technique naturally passes the outer information to the fluids using only δ-SPH formulation.

3 Results

We compare the proposed method and the previous method [2] with two dimensional static water test case as shown in right hand side of Figure 1. The fluid particles are arranged in square and 20 particles are used for each edge (totally 400 fluid particles).

In this test case, the previous method takes 5.99ms per time step, while the proposed one takes 1.58ms. As shown in Figure 2, the errors are less than initial spacing l in both methods even after 1M time steps, and our new method has less errors for long time simulation. This means that our method works appropriately and faster than the previous one.

Acknowledgements

This work was supported by JSPS KAKENHI Grant Numbers JP00351320 and JP17J00443.

References

[1] L. Postma, J. K. L. van Beek, H. F. P. van den Boogaard and G. S. Stelling,

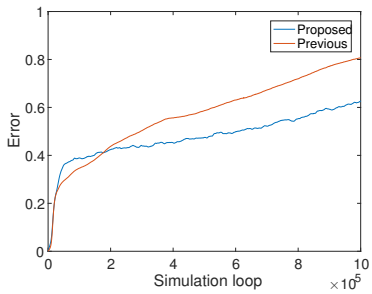


Figure 2: Comparison of average relative errors (average distance between analytical and simulated positions divided by l) obtained by the proposed and previous methods.

Consistent and efficient particle tracking on curvilinear grids for environmental problems, International Journal for Numerical Methods in Fluids **71**(10) (2013) 1226–1237.

[2] Y. Kanetsuki, J. C. Wells and S. Nakata, *Smoothed particle hydrodynamics method with partially defined fluid particles*, Mathematical Methods in the Applied Sciences, Early View (2016).

[3] R. L. Cook, *Stochastic sampling in computer graphics*, ACM Transactions on Graphics **5**(1) (1986) 51–72.

[4] R. Bridson, *Fast Poisson disk sampling in arbitrary dimensions*, In ACM SIGGRAPH 2007 sketches (2007) 22:1.

[5] C. Y. Ip, M. A. Yalçin, D. Luebke and A. Varshney, *PixelPie: Maximal Poisson-disk sampling with rasterization*, In Proceedings of the 5th High-Performance Graphics Conference (2013) 17–26.

[6] H. Schechter and R. Bridson, *Ghost SPH for animating water*, ACM Transactions on Graphics **31**(4) (2012) 61:1–61:8.

[7] M. B. Liu and G. R. Liu, *Smoothed particle hydrodynamics (SPH): an overview and recent developments*, Archives of Computational Methods in Engineering **17**(1) (2010) 25–76.

[8] D. Molteni and A. Colagrossi, *A simple procedure to improve the pressure evaluation in hydrodynamic context using the SPH*, Computer Physics Communications **180**(6) (2009) 861–872.

[9] M. Antuono, A. Colagrossi, S. Marrone and D. Molteni, *Free-surface flows solved by means of SPH schemes with numerical diffusive terms*, Computer Physics Communications **181**(3) (2010) 532–549.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

Dynamics of a Four-Dimensional Hypothalamic-Pituitary-Adrenal Axis Model with Distributed Delays

Eva Kaslik[1,2] and Mihaela Neamţu[3]

[1] *Dept. of Mathematics and Computer Science, West University of Timişoara, Romania*

[2] *Institute e-Austria Timisoara, Romania*

[3] *Dept. of Economics and Modelling, West University of Timişoara, Romania*

emails: ekaslik@gmail.com, mihaela.neamtu@e-uvt.ro

Abstract

A four-dimensional mathematical model is presented and analyzed, which describes the hypothalamus-pituitary-adrenal (HPA) axis with the influence of the GR concentration and includes general feedback functions. Due to the fact that the involved processes are not instantaneous, distributed time delays are incorporated, providing a more realistic modeling approach, since the whole past history of the variables is taken into account. Sufficient conditions for the local asymptotic stability of the equilibrium points are obtained and the occurrence of Hopf bifurcations is investigated, accounting for the appearance of limit cycles which successfully model the ultradian rhythm of the HPA axis. Numerical simulations reflect the importance of the theoretical results.

Key words: HPA axismathematical model distributed time delay stability bifurcation numerical simulation

1 Introduction

One of the most important self-regulated dynamic feedback neuroendocrine systems, which helps the body respond to stress, is the hypothalamuspituitaryadrenal (HPA) axis [6]. It consists of three regions: the hypothalamus, pituitary and adrenal glands, along with a set of direct influences and positive and negative feedback interactions. Both physical and psychological stressors (e.g. infection, dehydration, anticipation, fear) activate the hypothalamus

 ISBN: 978-84-617-8694-7

to release corticotropin-releasing hormone (CRH), which induces the corticotropin (ACTH) production in the pituitary. Then, ACTH is transported by the blood to the adrenal cortex, where it stimulates the production of cortisol (CORT), which in turn suppresses the production of both CRH and ACTH.

Mathematical modeling has been successfully applied in the study of metabolic and endocrine processes [3]. Several mathematical models of the HPA axis have been recently explored [2, 3, 6, 9, 10, 13, 14, 16, 17, 18, 20, 21, 22]. Oscillatory solutions of the mathematical models of the HPA axis should reflect the circadian as well as ultradian rhythm of hormone levels [5]. The ultradian rhythm is seen as an inherent behavior of the HPA axis, while the circadian rhythm is regarded as an external input to the axis [2]. Additionally, it is important to emphasize that time delays unavoidably exist in the HPA axis, due to the transportation of the hormones among the three glands, therefore, it is mandatory to incorporate them in the mathematical model.

The "minimal model" of the HPA axis, consisting of a system of three coupled, nonlinear differential equations, with the hormones CRH, ACTH and cortisol as variables, has been developed in [22]. No oscillatory behavior has been observed in this minimal model if time delays are not taken into consideration [2, 22]. This model has been recently generalized in [13], including memory terms in the form of distributed delays and fractional-order derivatives, which are linked with generating oscillatory solutions.

In this paper, we investigate a four-dimensional model of the HPA axis which includes distributed time delays. General distributed delays are helpful to reflect the whole past history of the variables, proving to be more realistic and more accurate in real world applications than discrete time delays [7]. Distributed delay models appear in a wide range of applications such as hematopoiesis [1], population biology [8], neural networks [11].

2 Mathematical model of HPA with distributed delays

In formulating the mathematical model which describes the variation in time of the concentrations of the three hormones CRH, ACTH and CORT, the following sequence of typical events is considered. CRH is secreted from the hypothalamus and released into the portal blood vessel of the hypophyseal stalk, and then transported to the anterior pituitary where it stimulates the secretion of ACTH, with an average time delay τ_1. Then, in the cortex of the adrenal glands, ACTH stimulates the secretion of the stress hormone cortisol with the average time delay τ_2. Cortisol has a negative feedback effect on the hypothalamus and the pituitary, expressed by two feedback functions f_1 and f_2, affecting the synthesis and release of CRH and ACTH, respectively. On one hand, cortisol inhibits the secretion of CRH through glucocorticoid receptors (GRs) situated in the hypothalamus [15], with an average time delay τ_{31}. On the other hand, cortisol also performs a negative feedback on the secretion of ACTH through GRs situated in the pituitary, with an average time delay

 ISBN: 978-84-617-8694-7

τ_{32}. The hormone concentrations of CRH, ACTH and cortisol are depleted through the rate constants w_1, w_2 and w_3, respectively.

Denoting the hormone concentrations, for simplicity, by $CRH(t) = x_1(t)$, $ACTH(t) = x_2(t)$, $CORT(t) = x_3(t)$, $GR(t) = x_4(t)$, the following system of differential equations with distributed delays is considered:

$$\begin{cases} \dot{x}_1(t) = f_1\left(\int_{-\infty}^{t} x_3(s)h_{31}(t-s)ds\right) - w_1x_1(t), \\ \dot{x}_2(t) = f_2\left(x_4(t)\int_{-\infty}^{t} x_3(s)h_{32}(t-s)ds\right)\int_{-\infty}^{t} x_1(s)h_1(t-s)ds - w_2x_2(t), \\ \dot{x}_3(t) = k_3\int_{-\infty}^{t} x_2(s)h_2(t-s)ds - w_3x_3(t), \\ \dot{x}_4(t) = f_3\left(x_4(t)\int_{-\infty}^{t} x_3(s)h_{34}(t-s)ds\right) - w_4x_4(t), \end{cases} \tag{1}$$

where all the first terms on the right hand side represent production and all the second terms represent depletion of hormones. The constant k_3 as well as the elimination constants w_1, w_2, w_3, w_4 are positive.

The functions $f_1, f_2 : [0,\infty) \to (0,\infty)$, which represent the negative feedback from CORT on CRH and ACTH, respectively, are assumed to be strictly decreasing, smooth and bounded on $[0,\infty)$. In particular, the results presented in this paper are also applicable when Hill functions are being used in the expression of the feedback functions [2, 22]:

$$f_1(u) = k_1\left(1 - \eta\frac{u^{\alpha_1}}{c^{\alpha_1}+u^{\alpha_1}}\right) \quad , \quad f_2(u) = k_2\left(1 - \mu\frac{u^{\alpha_2}}{c^{\alpha_2}+u^{\alpha_2}}\right), \tag{2}$$

with $\alpha_1, \alpha_2 \geq 1$, $k_1, k_2 > 0$, $\eta, \mu \in (0,1)$, $c > 0$. It is easy to verify that functions (2) satisfy all the properties mentioned above. However, it may be possible to model the negative feedback using different types of functions f_1 and f_2. In this paper, our aim is to obtain general results which will also be applicable to other choices of negative feedback functions, besides functions (2), often used in the literature.

The function $f_3 : [0,\infty) \to (0,\infty)$, which represents the positive feedback from CORT on the GR production is strictly increasing, smooth and bounded on $[0,\infty)$. In particular, it can also be considered as a Hill function of the form:

$$f_3(u) = k_4\left(1 + \xi\frac{u^{\alpha_3}}{c^{\alpha_3}+u^{\alpha_3}}\right), \tag{3}$$

with $\alpha_3 \geq 1$, $k_4 > 0$, $\xi \in (0,1)$, $c > 0$.

In system (1), the delay kernels $h_1, h_2, h_{31}, h_{32}, h_{34} : [0,\infty) \to [0,\infty)$ are probability density functions representing the probability that a particular time delay occurs. They are assumed to be bounded, piecewise continuous and satisfy

$$\int_0^{\infty} h(s)ds = 1. \tag{4}$$

 ISBN: 978-84-617-8694-7

The average delay of a delay kernel $h(t)$ is given by

$$\tau = \int_0^\infty sh(s)ds < \infty.$$

Two important classes of delay kernels often used in the literature, are worth mentioning:

- Dirac kernels: $h(s) = \delta(s - \tau)$, where $\tau \geq 0$, equivalent to a discrete time delay:

$$\int_{-\infty}^{t} x(s)h(t-s)ds = \int_0^\infty x(t-s)\delta(s-\tau)ds = x(t-\tau).$$

- Gamma kernels: $h(s) = \dfrac{s^{p-1}e^{-s/\beta}}{\beta^p\Gamma(p)}$, where $p, \beta > 0$, with the average delay $\tau = p\beta$.

In the mathematical modeling of real world phenomena, the exact distribution of time delays is generally unavailable, and hence, general kernels may provide better results [4, 23]. The analysis of models which include particular classes of delay kernels (e.g. weak Gamma kernels with $p = 1$ or strong Gamma kernels with $p = 2$) may reveal the more realistic effect of distributed delays on the system's dynamics, compared to discrete delays.

Initial conditions associated with system (1) are of the form:

$$x_i(s) = \varphi_i(s), \quad \forall\, s \in (-\infty, 0],\ i = 1, 2, 3, 4,$$

where φ_i are bounded continuous functions defined on $(-\infty, 0]$, with values in $[0, \infty)$.

3 Local stability analysis

The existence of an equilibrium point of system (1) is provided by the following:

Proposition 1. *The equilibrium states of system (1) are of the form*

$$E = \left(\frac{f_1(x_0)}{w_1}, \frac{w_3x_0}{k_3}, x_0, \frac{1}{x_0}f_2^{-1}\left(\frac{w_1w_2w_3}{k_3}\frac{x_0}{f_1(x_0)}\right)\right). \tag{5}$$

where $x_0 \in \left[0, \dfrac{k_1k_2k_3}{w_1w_2w_3}\right]$ *with* $k_i = f_i(0)$, $i = 1, 2$, *is a solution of the equation*

$$f_3\left(f_2^{-1}\left(\frac{w_1w_2w_3}{k_3}\frac{x}{f_1(x)}\right)\right) = \frac{w_4}{x}f_2^{-1}\left(\frac{w_1w_2w_3}{k_3}\frac{x}{f_1(x)}\right). \tag{6}$$

In what follows, we provide necessary and sufficient conditions for the local asymptotic stability of an equilibrium point E and the occurrence of limit cycles in a neighborhood of E (due to Hopf bifurcations) that can explain the ultradian rhythm. Considering general

 ISBN: 978-84-617-8694-7

delay kernels, we first obtain delay independent sufficient conditions for the local asymptotic stability of the equilibrium point E, which may prove to be useful if the time delays in system (1) cannot be accurately estimated.

The characteristic equation of the linearized system at the equilibrium point E is:

$$\begin{aligned}&(z+w_1)(z+w_2)(z+w_3)(z+\tilde{w_4})+a(w_4-\tilde{w_4})(z+w_1)H_2(z)H_{34}(z)+\\&+b(z+\tilde{w_4})H_1(z)H_2(z)H_{31}(z)+a(z+w_1)(z+\tilde{w_4})H_2(z)H_{32}(z)=0,\end{aligned}\tag{7}$$

where $H_i(z)=\int_0^\infty e^{-zs}h_i(s)ds$ represent the Laplace transforms of the delay kernels h_i, $i\in\{1,2,31,32,34\}$ and

$$a=-\frac{k_3}{w_1}f_1(x_0)f_2'(x_0r_0)r_0=-w_2w_3\frac{x_0r_0f_2'(x_0r_0)}{f_2(x_0r_0)}>0,\tag{8}$$

$$b=-k_3f_1'(x_0)f_2(x_0r_0)=-w_1w_2w_3\frac{x_0f_1'(x_0)}{f_1(x_0)}>0,\tag{9}$$

$$\tilde{w_4}=w_4-x_0f_3'(x_0r_0)<w_4.\tag{10}$$

The following inequalities will be useful for the theoretical analysis:

(I_0) $\tilde{w_4}>0$ and
$(w_1+\tilde{w_4})(w_2+\tilde{w_4})(w_3+\tilde{w_4})\geq(\tilde{w_4}-w_1)(\tilde{w_4}-w_4)(w_1+w_2+w_3+\tilde{w_4})$;

(I_1) $a(w_1+w_4)+b\leq(w_1+w_2)(w_2+w_3)(w_1+w_3)$;

(I_2) $\dfrac{aw_4}{\tilde{w_4}}+\dfrac{b}{w_1}<w_2w_3$;

$(\overline{I_2})$ $\dfrac{aw_4}{\tilde{w_4}}+\dfrac{b}{w_1}\geq w_2w_3$.

Theorem 1 (Local asymptotic stability)**.**

1. *In the non-delayed case, if inequality (I_1) is satisfied, then the equilibrium point E of system (1) is locally asymptotically stable.*

2. *For any delay kernels $h_i(t)$, $i\in\{1,2,31,32,34\}$, if inequality (I_2) holds, then the equilibrium point E of system (1) is locally asymptotically stable.*

Corollary 1. *For any delay kernels $h_i(t)$, $i\in\{1,2,31,32,34\}$, if the equilibrium point E of system (1) is unstable, then inequality $(\overline{I_2})$ holds. In other words, inequality $(\overline{I_2})$ is a necessary condition for the occurrence of bifurcations in system (1).*

 ISBN: 978-84-617-8694-7

4 Bifurcation analysis

For simplicity, we further assume that

$$H_{32}(z) = H_{34}(z) = H_1(z)H_{31}(z),$$

and we denote

$$H(z) = H_2(z)H_{32}(z) = H_2(z)H_{34}(z) = H_1(z)H_2(z)H_{31}(z).$$

In fact, $H(z)$ is the Laplace transform of the convolution of the delay kernels h_2 and h_{32} defined as

$$h(t) = \int_0^t h_2(s)h_{32}(t-s)ds,$$

with the mean

$$\tau = \int_0^\infty sh(s)ds = \tau_2 + \tau_{32}, \tag{11}$$

where τ_2 and τ_{32} represent the average delays of the kernels h_2 and h_{32} respectively.

Therefore, the characteristic equation (7) becomes

$$(z+w_1)(z+w_2)(z+w_3)(z+\tilde{w_4}) + [a(z+w_1)(z+w_4) + b(z+\tilde{w_4})]H(z) = 0,$$

which can be rewritten as:

$$H(z)^{-1} = Q(z), \tag{12}$$

where

$$Q(z) = -\frac{a(z+w_1)(z+w_4) + b(z+\tilde{w_4})}{(z+w_1)(z+w_2)(z+w_3)(z+\tilde{w_4})}.$$

The properties of the function $Q(z)$ are given in the following Lemma.

Lemma 1. *The function*

$$\omega \mapsto |Q(i\omega)| = \sqrt{\frac{(b\tilde{w_4} + aw_1w_4 - a\omega^2)^2 + \omega^2(a(w_1+w_4)+b)^2}{(\omega^2+w_1^2)(\omega^2+w_2^2)(\omega^2+w_3^2)(\omega^2+\tilde{w_4}^2)}}$$

is strictly decreasing on $[0,\infty)$ *and the equation*

$$|Q(i\omega)| = 1$$

has a unique positive real root ω_0 *if and only if inequality* $(\overline{I2})$ *is satisfied.*

Moreover, the following inequality holds:

$$\Im\left(\frac{Q'(i\omega)}{Q(i\omega)}\right) > 0 \qquad \forall\, \omega > 0.$$

For the bifurcation analysis, we focus our attention on the following two cases: (1) all delay kernels are Dirac kernels; (2) all delay kernels are Gamma kernels.

 ISBN: 978-84-617-8694-7

4.1 Dirac kernels

If all the delay kernels are Dirac kernels: $h_1(t) = \delta(t-\tau_1)$, $h_2(t) = \delta(t-\tau_2)$, $h_{31}(t) = \delta(t-\tau_{31})$, $h_{32}(t) = \delta(t-\tau_{32})$, $h_{34}(t) = \delta(t-\tau_{34})$ where $\tau_1, \tau_2, \tau_{31}, \tau_{32}, \tau_{34} \geq 0$ satisfy the property

$$\tau_2 + \tau_{32} = \tau_2 + \tau_{34} = \tau_1 + \tau_2 + \tau_{31} = \tau > 0, \tag{13}$$

then, the characteristic equation (12) becomes:

$$e^{\tau z} = Q(z). \tag{14}$$

Choosing τ as bifurcation parameter and following the same proof as in [13], we have:

Theorem 2 (Hopf bifurcations in the case of Dirac kernels)**.** *Assume that inequalities* (I_0), (I_1) *and* $(\overline{I_2})$ *are satisfied. For any* $p \in \mathbb{Z}^+$*, consider*

$$\tau_p = \frac{\arccos\left[\Re(Q(i\omega_0))\right] + 2p\pi}{\omega_0}, \tag{15}$$

where $\omega_0 > 0$ *is given by Lemma 1. The equilibrium point* E *is asymptotically stable if any only if* $\tau \in [0, \tau_0)$*. For any* $p \in \mathbb{Z}^+$*, at* $\tau = \tau_p$*, system (1) undergoes a Hopf bifurcation at the equilibrium point* E*.*

4.2 Gamma kernels

If the delay kernels are Gamma kernels: $h_1(t) = \dfrac{t^{n_1-1}e^{-t/\beta}}{\beta^{n_1}(n_1-1)!}$, $h_2(t) = \dfrac{t^{n_2-1}e^{-t/\beta}}{\beta^{n_2}(n_2-1)!}$, $h_{31}(t) = \dfrac{t^{n_{31}-1}e^{-t/\beta}}{\beta^{n_{31}}(n_{31}-1)!}$, $h_{32}(t) = \dfrac{t^{n_{32}-1}e^{-t/\beta}}{\beta^{n_{32}}(n_{32}-1)!}$, $h_{34}(t) = \dfrac{t^{n_{34}-1}e^{-t/\beta}}{\beta^{n_{34}}(n_{34}-1)!}$, where $\beta > 0$ and $n_1, n_2, n_{31}, n_{32}, n_{34} \in \mathbb{Z}^+ \setminus \{0\}$ satisfy:

$$n_2 + n_{32} = n_2 + n_{34} = n_1 + n_2 + n_{31} = n \geq 2,$$

the characteristic equation (7) is:

$$(\beta z + 1)^n = Q(z). \tag{16}$$

Choosing β as bifurcation parameter, as in [13], the following result holds:

Theorem 3 (Hopf bifurcations in the case of Gamma kernels)**.** *Assume that inequalities* (I_0), (I_1) *and* $(\overline{I_2})$ *are satisfied. Let* ω_n *denote the largest real root of the equation*

$$T_n\left(\frac{1}{|Q(i\omega)|^{1/n}}\right) = \frac{\Re(Q(i\omega))}{|Q(i\omega)|} \tag{17}$$

 ISBN: 978-84-617-8694-7

from the interval $(0, \omega_0)$*, where* T_n *is the Chebyshev polynomial of the first kind of order* n*, and consider*

$$\beta_n = \frac{1}{\omega_n}\sqrt{|Q(i\omega_n)|^{2/n} - 1}. \tag{18}$$

The equilibrium point E *is asymptotically stable if* $\beta \in (0, \beta_n)$*. At* $\beta = \beta_n$*, system (1) undergoes a Hopf bifurcation at the equilibrium point* E*.*

5 Numerical simulations

For numerical simulations, the literature values of the elimination constants are considered: $w_1 = 0.17329$ min^{-1}, $w_2 = 0.034831$ min^{-1}, $w_3 = 0.0090726$ min^{-1} and $w_4 = 0.01$ min^{-1} [21].

The equilibrium point E of the system consists of the 24-h mean values of the hormones: $\bar{x}_1 = 7.659$ pg/ml, $\bar{x}_2 = 21$ pg/ml, $\bar{x}_3 = 3.055$ ng/ml [5] and $\bar{x}_4 = 1$ pg/ml.

The feedback functions f_1, f_2 and f_3 are considered as in eqs. (2) and (3), with $\alpha_1 = \alpha_2 = 5$, $\alpha_3 = 7$, $\eta = \mu = 1$, $\xi = 0.75$, $k_1 = 12.364$ pg/(ml $\cdot$ min), $k_2 = 0.88969$ min^{-1}, $k_4 = 0.0058368$ min^{-1} and $c = 2$ ng/ml. Moreover, $k_3 = 1.31985$ min^{-1}.

According to [3], we assume the mean delay $\tau_1 = 0$. Recently, it has been shown that humans show fast HPA negative feedback [19], suggesting that both GR (glucocorticoid receptors) and MR (mineralocorticoid receptors) are involved in this mechanism, with GR effecting a rapid nongenomic feedback at the level of the anterior pituitary and MR sensing higher glucocorticoid levels while levels are still rising [12]. This is in accordance with our numerical simulations (Figs. 1,2), which show that for small values (less than 10 min) of the average time delays $\tau_2, \tau_{31}, \tau_{32}, \tau_{34}$, oscillatory behavior occurs in system (1), with different types of distributed delays.

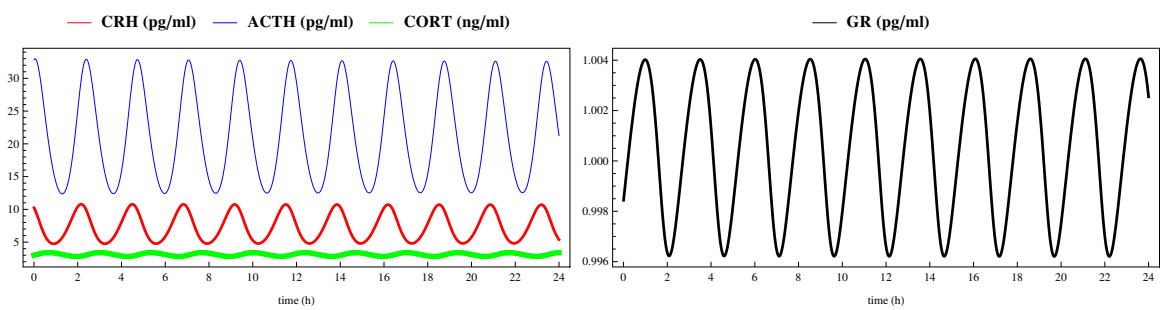


Figure 1: Stable periodic orbit of system (1) with Dirac kernels (discrete delays $\tau_1 = 0$, $\tau_2 = 7.5$ (min), $\tau_{31} = \tau_{32} = \tau_{34} = 8$ (min)) due to the Hopf bifurcation taking place when the bifurcation parameter τ exceeds the critical value $\tau_0 = 14.81$ (min).

 ISBN: 978-84-617-8694-7

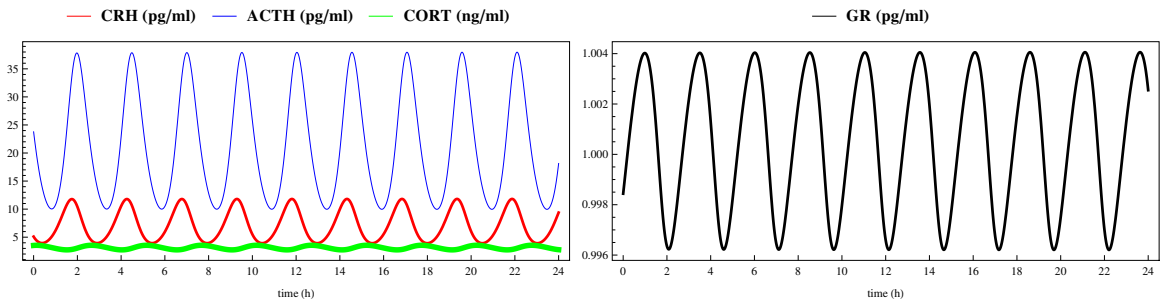


Figure 2: Stable periodic orbit of system (1) with strong Gamma kernels ($n_1 = 0$, $n_2 = n_{31} = n_{32} = n_{34} = 2$, $\beta = 4.5$ (min), mean delays: $\tau_2 = \tau_{31} = \tau_{32} = \tau_{34} = 9$ (min)) due to the Hopf bifurcation taking place when the bifurcation parameter β exceeds the critical value $\beta_4 = 4.07056$ (min).

6 Conclusions

This paper presents an analysis of a four-dimensional mathematical model describing the hypothalamus-pituitary-adrenal axis with the influence of the GR concentration, considering general feedback functions to account for the interactions within the HPA axis. Due to the fact that the involved processes are not instantaneous, distributed delays have been included. This is a more realistic approach to the modeling of the biological processes, as it takes into account the whole past history of the variables, efficiently capturing the vital mechanisms of the HPA system. Sufficient conditions have been obtained for the local asymptotic stability of the equilibrium point and the occurrence of Hopf bifurcations has been investigated. Numerical simulations reflect the importance of the theoretical results and are in accordance with experimental findings.

As a direction for future research, a fractional-order formulation of the mathematical model will be analyzed.

Acknowledgements

This work was supported by a grant of the Romanian National Authority for Scientific Research and Innovation, CNCS-UEFISCDI, project no. PN-II-RU-TE-2014-4-0270.

References

[1] M. Adimy, F. Crauste, M. Halanay, A. Neamţu, and D. Opriş, *Stability of limit cycles in a pluripotent stem cell dynamics model*, Chaos, Solitons & Fractals, 27

(2006), pp. 1091–1107.

[2] M. Andersen, F. Vinther, and J. T. Ottesen, *Mathematical modeling of the hypothalamic–pituitary–adrenal gland (hpa) axis, including hippocampal mechanisms*, Mathematical Biosciences, 246 (2013), pp. 122–138.

[3] N. Bairagi, S. Chatterjee, and J. Chattopadhyay, *Variability in the secretion of corticotropin-releasing hormone, adrenocorticotropic hormone and cortisol and understandability of the hypothalamic-pituitary-adrenal axis dynamicsa mathematical study based on clinical evidence*, Mathematical Medicine and Biology, (2008), pp. 1–27.

[4] S. Campbell and R. Jessop, *Approximating the stability region for a differential equation with a distributed delay*, Mathematical Modelling of Natural Phenomena, 4 (2009), pp. 1–27.

[5] B. Carroll, F. Cassidy, D. Naftolowitz, N. Tatham, W. Wilson, A. Iranmanesh, P. Liu, and J. Veldhuis, *Pathophysiology of hypercortisolism in depression*, Acta Psychiatrica Scandinavica, 115 (2007), pp. 90–103.

[6] M. Conrad, C. Hubold, B. Fischer, and A. Peters, *Modeling the hypothalamus–pituitary–adrenal system: homeostasis by interacting positive and negative feedback*, Journal of Biological Physics, 35 (2009), pp. 149–162.

[7] J. M. Cushing, *Integrodifferential equations and delay models in population dynamics*, vol. 20, Springer Science & Business Media, 2013.

[8] T. Faria and J. J. Oliveira, *Local and global stability for lotka–volterra systems with distributed delays and instantaneous negative feedbacks*, Journal of Differential Equations, 244 (2008), pp. 1049–1079.

[9] S. Gupta, E. Aslakson, B. M. Gurbaxani, and S. D. Vernon, *Inclusion of the glucocorticoid receptor in a hypothalamic pituitary adrenal axis model reveals bistability*, Theoretical Biology and Medical Modelling, 4 (2007), p. 8.

[10] S. Jelić, Ž. Čupić, and L. Kolar-Anić, *Mathematical modeling of the hypothalamic–pituitary–adrenal system activity*, Mathematical Biosciences, 197 (2005), pp. 173–187.

[11] R. Jessop and S. A. Campbell, *Approximating the stability region of a neural network with a general distribution of delays*, Neural Networks, 23 (2010), pp. 1187–1201.

[12] H. Karst, S. Berger, M. Turiault, F. Tronche, G. Schütz, and M. Joëls, *Mineralocorticoid receptors are indispensable for nongenomic modulation of hippocampal glutamate transmission by corticosterone*, Proceedings of the National Academy of Sciences of the United States of America, 102 (2005), pp. 19204–19207.

 ISBN: 978-84-617-8694-7

[13] E. Kaslik and M. Neamtu, *Stability and hopf bifurcation analysis for the hypothalamic-pituitary-adrenal axis model with memory*, Mathematical Medicine and Biology, (2017).

[14] V. Kyrylov, L. Severyanov, and A. Vieira, *Modeling robust oscillatory behavior of the hypothalamic-pituitary-adrenal axis*, Biomedical Engineering, IEEE Transactions on, 52 (2005), pp. 1977–1983.

[15] L. Landsberg, J. Young, J. Wilson, and D. Foster, *Williams Textbook of Endocrinology*, Prentice Hall International, New Jersey, 1992.

[16] Y. Lenbury and P. Pornsawad, *A delay-differential equation model of the feedback-controlled hypothalamus–pituitary–adrenal axis in humans*, Mathematical Medicine and Biology, 22 (2005), pp. 15–33.

[17] V. M. Markovic, Z. Cupic, V. Vukojevic, and L. Kolar-Anic, *Predictive modeling of the hypothalamic-pituitary-adrenal (hpa) axis response to acute and chronic stress*, Endocrine Journal, 58 (2011), pp. 889–904.

[18] P. Pornsawad, *The feedforward-feedback system of the hypothalamus-pituitary-adrenal axis*, in Advances in Computing, Communications and Informatics (ICACCI), 2013 International Conference on, IEEE, 2013, pp. 1374–1379.

[19] G. M. Russell, D. E. Henley, J. Leendertz, J. A. Douthwaite, S. A. Wood, A. Stevens, W. W. Woltersdorf, B. W. Peeters, G. S. Ruigt, A. White, et al., *Rapid glucocorticoid receptor-mediated inhibition of hypothalamic–pituitary–adrenal ultradian activity in healthy males*, The Journal of Neuroscience, 30 (2010), pp. 6106–6115.

[20] D. Savić, S. Jelić, and N. Burić, *Stability of a general delay differential model of the hypothalamo-pituitary-adrenocortical system*, International Journal of Bifurcation and Chaos, 16 (2006), pp. 3079–3085.

[21] K. Sriram, M. Rodriguez-Fernandez, and F. J. Doyle III, *Modeling cortisol dynamics in the neuro-endocrine axis distinguishes normal, depression, and post-traumatic stress disorder (ptsd) in humans*, PLoS Comput Biol, 8 (2012), p. e1002379.

[22] F. Vinther, M. Andersen, and J. T. Ottesen, *The minimal model of the hypothalamic–pituitary–adrenal axis*, Journal of Mathematical Biology, 63 (2011), pp. 663–690.

[23] Y. Yuan and J. Bélair, *Stability and hopf bifurcation analysis for functional differential equation with distributed delay*, SIAM Journal on Applied Dynamical Systems, 10 (2011), pp. 551–581.

 ISBN: 978-84-617-8694-7

Volume IV

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

On quasi-contractive multi-valued mappings' open problem in complete metric spaces

Farshid Khojasteh[1], Antonio Francisco Roldan Lopez de Hierro[2] and Sirous Moradi[3]

[1] *Young Researcher and Elite Club, Arak-Branch, Islamic Azad, University, Arak, Iran.*

[2] *Department of Quantitative Methods for Economics and Business, University of Granada, Granada, Spain.*

[2] *Department of Mathematics Faculty of Science, Arak University, Arak 38156-8-8349, Iran.*

emails: f-khojaste@iau-arak.ac.ir, aroldan@ugr.es, afroldan@ujaen.es, sirousmoradi@gmail.com

Abstract

In recent years, many authors have tried to find at least a fixed point for multi-valued quasi-contractions whose contractivity constants α belonged to the interval $(0, 1)$. Up to now, efforts in this direction when $\alpha \in (\frac{1}{2}, 1)$ either have failed or have been lessened to a lighter version. The main result of current research gives a partial positive answer to the above-mentioned problem by adding a necessary and sufficient condition in order to guarantee the existence of strict fixed points for quasi-contractive multi-valued mappings. Ultimately, some examples and results obtain which have a closed relation to quasi-contractions mappings.

Key words: Strict fixed point, fixed point, Quasi-contraction, Pompeiu-Hausdorff metric, Multi valued mapping
MSC 2000: AMS 47H10, 47H08, 54C60.

1 Introduction

In 2011, Wardowski published a paper [11] where he introduced fixed point results for multi-valued contractive mappings in normal cone metric spaces. In 2011, Amini-Harandi

managed to prove a result on the existence of fixed points in the set of multi-valued quasi-contractive mappings in metric spaces by using Rezapour *et al.*'s technique which had been given in [8]. But, like Kadelburg *et al.* [5], he could only prove it for $\alpha \in (0, \frac{1}{2})$ [1]. In 2012, Rezapour *et al.* [4] introduced quasi-contractive type multi-valued mappings and they demonstrated that the main result of Amini-Harandi also held in the set of quasi-contractive type multi-valued mappings.

In what follows, the following notions are needed to achieve the goal. From now on, let (X, d) be a metric space and let $\mathcal{CB}(X)$ be the family of all nonempty, closed, bounded subsets of X. Let $T : X \to \mathcal{CB}(X)$ be a multi-valued mapping on X. A point $x \in X$ is called a *fixed point* of T if $x \in Tx$, and it is called a *strict fixed point of* T if $Tx = \{x\}$. We denote by $Fix(T)$ (respectively, by $SFix(T)$) the family of all fixed points (respectively, the family of all strict fixed points) of T. Obviously, $SFix(T) \subseteq Fix(T)$. Some authors call *endpoints* to strict fixed points (for instance, see [9] and [10]). From our point of view, the nomenclature "*strict fixed point*" is more appropriate, so we will use it throughout this paper.

A multi-valued mapping $T : X \to \mathcal{CB}(X)$ is said to be a *multi-valued quasi-contractive mapping* whenever there exists $\alpha \in (0, 1)$ such that

$$\mathcal{H}(Tx, Ty) \leq \alpha\, M(x, y) \quad \textit{for all } x, y \in X,$$

where

$$M(x, y) = \max\{\, d(x, y),\, d(x, Tx),\, d(y, Ty),\, d(x, Ty),\, d(y, Tx)\,\}.$$

In recent years, many authors (such as Amini-Harandi, Rezapour *et al.*, and Kadelburg *et al.*) have introduced some fixed point theorems for multi-valued quasi-contractions whose contractivity constants α belonged to the interval $\left(\frac{1}{2}, 1\right)$. However, Amini-Harandi pointed out, it is not clear whether such results also hold when $\alpha \in \left(0, \frac{1}{2}\right)$, Up to now, efforts in this direction either have failed or they have been lessened to a lighter version. The main result of the current research gives a partial positive answer to the above-mentioned problem by adding a necessary and sufficient condition in order to guarantee existence of strict fixed points of quasi-contractive multi-valued mappings. This problem has remained open for many years. Moreover, some important results have obtained in this direction.

Theorem 1.1. *Let (X, d) be a complete metric space and let $T : X \to \mathcal{CB}(X)$ be a multi-valued quasi-contraction for some $\alpha \in (0, \frac{1}{2})$. Then T has a fixed point.*

Immediately, he proposed the following question.

Question 1.2. *[**Amini Harandi's Conjecture**] Does the conclusion of Theorem 1.1 remain true for any $\alpha \in [\frac{1}{2}, 1)$*

In this manuscript, we give a partial positive answer.

 ISBN: 978-84-617-8694-7

2 A Partial Positive Answer

In this section, we introduce a necessary and sufficient condition for guaranteeing existence of strict fixed points of quasi-contractive multi-valued mappings, which gives a partial positive answer to Question 1.2.

The following definition plays a crucial role to continue.

Definition 2.1. *([6, Moradi, Khojasteh]) A multi-valued mapping $T : X \to \mathcal{CB}(X)$ has* approximate strict fixed point property *if*

$$\inf \{ \mathcal{H}(\{x\}, Tx) : x \in X \} = 0. \tag{1}$$

Theorem 2.2. *Let (X, d) be a complete metric space and let $T : X \to \mathcal{CB}(X)$ be a multi-valued mapping such that*

$$\mathcal{H}(Tx, Ty) \leq \alpha \max \{ d(x, y),\, d(x, Tx),\, d(y, Ty),\, d(x, Ty),\, d(y, Tx) \}, \tag{2}$$

for all $x, y \in X$, where $0 \leq \alpha < 1$. Then T has a strict fixed point in X, if and only if, T has the approximate strict fixed point property. In such case, $SFix(T) = Fix(T)$ and T has a unique (strict) fixed point.

Corollary 2.3. *Let (X, d) be a complete metric space and let $T : X \to \mathcal{CB}(X)$ be a multi-valued mapping such that*

$$\mathcal{H}(Tx, Ty) \leq \alpha \max \left\{ d(x, y),\, d(x, Tx),\, d(y, Ty),\, \frac{d^2(y, Tx) + d^2(x, Ty)}{d(y, Tx) + d(x, Ty)} \right\},$$

for all $x, y \in X$, where $0 \leq \alpha < 1$. Then T has a unique strict fixed point in X, if and only if, T has the approximate strict fixed point property. In such a case, $SFix(T) = Fix(T)$.

Corollary 2.4. *Let (X, d) be a complete metric space and let $f : X \to X$ be a mapping such that*

$$d(fx, fy) \leq \alpha \max \left\{ d(x, y),\, d(x, fx),\, d(y, fy),\, \frac{d^2(y, fx) + d^2(x, fy)}{d(y, fx) + d(x, fy)} \right\},$$

for all $x, y \in X$, and $0 \leq \alpha < \frac{1}{2}$. Then T has the approximate fixed point property in X, that is,

$$\inf \{ d(x, fx) : x \in X \} = 0$$

Corollary 2.5. *Let (X, d) be a complete metric space and let $f : X \to X$ be a mapping such that*

$$d(fx, fy) \leq \alpha \max \left\{ d(x, y),\, d(x, fx),\, d(y, fy),\, \frac{d^2(y, fx) + d^2(x, fy)}{d(y, fx) + d(x, fy)} \right\}, \tag{3}$$

for all $x, y \in X$, $0 \leq \alpha < \frac{1}{2}$. Then T has a unique fixed point.

 ISBN: 978-84-617-8694-7

3 Some Consequences

In this section, we present some new fixed point results in the set of multi-valued mappings which generalize the results introduced by Nadler [7], Ćirić [2], and Daffer and Kaneko [3]. The following theorem is our first main result.

Theorem 3.1. *Let (X,d) be a complete metric space and let $T: X \to \mathcal{CB}(X)$ be a multi-valued mapping such that*

$$\mathcal{H}(Tx,Ty) \leq \alpha \max\left\{ d(x,y),\, d(x,Tx),\, d(y,Ty),\, c\cdot\frac{d^2(y,Tx)+d^2(x,Ty)}{d(y,Tx)+d(x,Ty)} \right\}, \qquad (4)$$

for all $x,y \in X$, where $\frac{1}{2} \leq c < 1$ and $0 \leq \alpha < \frac{1}{4c^2}$. Then at least one of the following conditions holds:

(i). T has a fixed point

(ii). T^2 has a fixed point (that is, there exists $z \in X$ such that $z \in T^2 z$, where $T^2 z = \bigcup_{\omega \in Tz} T\omega$).

The following consequences are obtained by replacing the contractivity condition by a stronger one.

Corollary 3.2. *Let (X,d) be a complete metric space and let $T: X \to \mathcal{CB}(X)$ be a multi-valued mapping such that*

$$\mathcal{H}(Tx,Ty) \leq \alpha \max\left\{ d(x,y),\, d(x,Tx),\, d(y,Ty),\, c\sqrt{\frac{d^2(y,Tx)+d^2(x,Ty)}{2}} \right\}, \qquad (5)$$

for all $x,y \in X$ where $\frac{1}{2} \leq c < 1$ and $0 \leq \alpha < \frac{1}{4c^2}$. Then at least one of the following conditions holds:

(i). T has a fixed point,

(ii). T^2 has a fixed point.

Corollary 3.3. *Let (X,d) be a complete metric space and let $T: X \to \mathcal{CB}(X)$ be two multi-valued mapping such that*

$$\mathcal{H}(Tx,Ty) \leq \alpha \max\left\{ d(x,y),\, d(x,Tx),\, d(y,Ty),\, \frac{d^2(y,Tx)+d^2(x,Ty)}{d(y,Tx)+d(x,Ty)} \right\}, \qquad (6)$$

for all $x,y \in X$, where $0 \leq \alpha < \frac{1}{2}$. Then T has a fixed point.

We also have found two suitable examples to show that the new class of contractions which have been introduced in Theorem 3.1 or Corollary 2.5 is a non-empty set.

 ISBN: 978-84-617-8694-7

References

[1] A. Amini-Harandi, *Fixed point theory for set-valued quasi-contraction maps in metric spaces*, Appl. Math. Lett. **24** (2011) 1791–1794.

[2] Lj. B. Ćirić, *Generalized contraction and fixed point theorems*, Publ. Inst. Math (Beograd) **12** (1971) 19–26.

[3] P. Z. Daffer, H. Kaneko, *Fixed points of generalized contractive multi-valued mappings*, *J. Math. Anal. Appl.* **192** (1995) 655–666.

[4] R.H. Haghi, Sh. Rezapour, N. Shahzad, *On fixed points of quasi-contraction type multifunctions*, Appl. Math. Lett. **25** (2012) 843–846.

[5] Z. Kadelburg, S. Radenović, V. Rakočević, *Remarks on "quasi-contraction on a cone metric space"*, Appl. Math. Lett. **22** (2009) 1674–1679.

[6] F. Khojasteh, V. Rakočević, *Some new common fixed point results for generalized contractive multi-valued non-self-mappings*, Appl. Math. Lett. **25** (2012) 287–293.

[7] S. B. Nadler, *Multi-valued contraction mappings*, Pacific J. Math. **30** (1969)475–488.

[8] Sh. Rezapour, R.H. Haghi, N. Shahzad, *Some notes on fixed points of quasi-contraction maps*, Appl. Math. Lett. **23** (2010) 498–502.

[9] I. A. Rus, *Generalized Contractions and Applications*, Cluj University Press, Cluj-Napoca, 2001.

[10] I. A. Rus, A. Petruşel and G. Petruşel, *Fixed Point Theory*, Cluj University Press, Cluj-Napoca, 2008.

[11] D. Wardowski, *On set-valued contractions of Nadler type in cone metric spaces*, Appl. Math. Lett. **24** (2011) 275–278.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

Recursion Based Sensitivity Coefficient Determination for Probabilistic Evolution Theoretical (PREVTH) Solutions to Explicit Autonomous Ordinary Differential Equations

Melike Ebru KIRKIN[1] and Metin DEMİRALP[2]

[1] *Computational Science and Engineering Department, Informatics Institute, İstanbul Technical University, Maslak, 34469, İstanbul, Türkiye (Turkey)*

emails: ebrkirkin@itu.edu.tr, metin.demiralp@gmail.com

Abstract

Probabilistic Evolution Theory (PREVTH) has been recently developed in our group studies [1–17] and it stands as one of most effective methods used for the solution of explicit autonomous ordinary differential equations. Its most recent form is based on a rather simple recursion between so-called Squarified Telescope Matrices or between the images of initial vector under these matrices. The initial vector elements can be considered as parameters to the solutions and the sensitivities of the solution vector to these parameters can be utilized for robustness investigations. The partial derivatives can be used to this end, and for each derivative a different recursion can be constructed simply partial differentiating the PREVTH vector recursion accordingly. This work is devoted to this issue and a sample target system (Henon-Heiles system) has been taken to the focus.

Key words: Ordinary Differential Equation, Probabilistic Evolution Theory, Constancy Adding Space Extension, Telescope Matrices, Squarification, Henon Heiles, Recursion between Vectors.

1 Introduction

Even though the Probabilistic Evolution Theory (PREVTH) has been developed for the first order autonomous explicit ODEs, this is not a great limitation since (1) all higher-than-first order ODEs can be converted to a set of first order ODEs whose number of unknown is higher than the original ODEs, (2) any nonautonomy can be converted to autonomy by defining a new unknown function which is identical to the time variable, t. This leaves us

 ISBN: 978-84-617-8694-7

with no loss of generality. Under this situation we can write the ODEs to be solved via following equations

$$\dot{\mathbf{x}}(t) = \mathbf{f}\left(x_1(t), ..., x_n(t)\right), \qquad \mathbf{x}(0) = \mathbf{a} \tag{1}$$

where t stands for "time" and the right hand side does not explicitly depend on time because of autonomy. The unknown functions are the elements of the vector $\mathbf{x}(t)$ while $\mathbf{a}$ vector's elements denote the initial values of the unknown functions.

The right hand side of first equation in (1) is assumed to be analytic in unknown functions. Hence it can be expanded to a multivariate Taylor series which is not preferable to use because of its quite complicated multiindex structure. Instead we can use the Kronecker power series defined as follows in our group studies [5–7].

$$\mathbf{f}\left(\mathbf{x}(t)\right) = \sum_{j=0}^{\infty} \mathbf{F}_j \mathbf{x}(t)^{\otimes j} \tag{2}$$

where $\mathbf{F}_j$ stands for a constant matrix of $n \times n^j$, which can be constructed from the vector function $\mathbf{f}$. The Kronecker product of two entities ($\mathbf{A}$ and $\mathbf{B}$), matrix and/or vector, and, the Kronecker power are defined as follows

$$\left[\, \mathbf{A} \otimes \mathbf{B} \,\right]_{i,j} \equiv \left[\, \mathbf{A} \,\right]_{i,j} \mathbf{B}, \qquad \mathbf{x}(t)^{\otimes j} \equiv \mathbf{x}(t) \otimes \cdots \otimes \mathbf{x}(t) \tag{3}$$

If the right hand side function vector $\mathbf{f}$ has a special structure such that the addition of new functions depending on the unknown functions as new elements to unknown function set then the new right hand side vector function can take a multinomial form which can be reduced to a second degree right hand side function by new unknown additions (space extension) [12, 13]. So under these conditions one can write

$$\dot{\mathbf{x}}(t) = \mathbf{F}_0 + \mathbf{F}_1 \mathbf{x}(t) + \mathbf{F}_2 \mathbf{x}(t)^{\otimes 2} \tag{4}$$

where $\mathbf{F}_0$ is an $n \times 1$ type vector while $\mathbf{F}_1$ and $\mathbf{F}_2$ are $n \times n$ and $n \times n^2$ type matrices. Beyond that, $\mathbf{x}$ is a vector of n elements and is composed of unknown functions. If we attempt to extend the space by adding an unknown constant as a new additional function (this procedure is called "Constancy Adding Space Extension" CASE) then the constant term $\mathbf{F}_0$ in (4) can be removed while $\mathbf{F}_1$ can be converted to $\beta \mathbf{I}_n$ where β is an arbitrary parameter. All these mean that the following vector ODE and accompanying initial condition become valid for our original equation

$$\dot{\mathbf{x}}(t) = \beta \mathbf{I}_n \mathbf{x}(t) + \mathbf{F} \mathbf{x}(t)^{\otimes 2}, \qquad \mathbf{x}(0) = \mathbf{a} \tag{5}$$

This can be further simplified by using the following definitions

$$\mathbf{x}(t) = e^{\beta t} \boldsymbol{\xi}(t), \qquad u(t) \equiv \frac{\mathrm{e}^{\beta t} - 1}{\beta} \tag{6}$$

 ISBN: 978-84-617-8694-7

to get

$$\frac{d\boldsymbol{\xi}(u)}{du} = \mathbf{F}\boldsymbol{\xi}(u)^{\otimes 2}, \qquad \boldsymbol{\xi}(0) = \mathbf{a} \tag{7}$$

whose analytic solution and the related form of $\mathbf{x}(t)$ can be written as follows in accordance with PREVTH.

$$\boldsymbol{\xi}(u) = \sum_{j=0}^{\infty} \frac{u^j}{j!} \mathbf{T}_j \mathbf{a}^{\otimes j+1}, \qquad \mathbf{x}(t) = e^{\beta t} \sum_{j=0}^{\infty} \frac{1}{j!} \left(\frac{\mathrm{e}^{\beta t} - 1}{\beta} \right)^j \mathbf{T}_j \mathbf{a}^{\otimes j+1} \tag{8}$$

where $\mathbf{a}$ stands for the initial vector with n given elements as we stated before and this equation obtained after the use of "Constancy Additional Space Extension" which is detailed in [18, 21]. In this formula $\mathbf{T}_j$'s are called as "Telescope Matrices" which are of $n \times n^{(j+1)}$ type. Their explicit structures are given thru below equalities

$$\mathbf{T}_j \equiv \prod_{k=1}^{j} \mathbf{M}_k, \qquad j = 0, 1, 2, ... \qquad \mathbf{M}_k \equiv \sum_{\ell=0}^{k-1} \mathbf{I}_n^{\otimes \ell} \otimes \mathbf{F} \otimes \mathbf{I}_n^{\otimes (k-1-\ell)}, \qquad k = 1, 2, ... \tag{9}$$

where the matrix factor, $\mathbf{M}_k$ is called "Monocular Matrix" and is of $n^k \times n^{(k+1)}$ type. They are cascaded to produce telescope matrices as expressed above.

Telescope matrices have abundantly many of zero elements and hence are very sparse. In order to suppress the sparsity as much as possible, the following "Squarified Telescope Matrices (SquTelMats)" denoted by $\mathbf{S}_j$'s can be brought to scene as follows

$$\mathbf{T}_j \mathbf{a}^{\otimes j+1} \equiv \mathbf{S}_j(\mathbf{a})\mathbf{a} \qquad j = 0, 1, 2, ... \tag{10}$$

The following recursion between squarified telescope matrices hold

$$\mathbf{S}_j = \sum_{k=0}^{j-1} \binom{j-1}{k} \lfloor\, \mathbf{F}, \mathbf{S}_k \mathbf{a} \,\rceil\, \mathbf{S}_{j-k-1}, \qquad j = 1, 2, ... \qquad \mathbf{S}_0 = \mathbf{I} \tag{11}$$

where the squtelmat between $n \times n^2$ type $\mathbf{F}$ matrix and n element $\mathbf{y}$ vector is explicitly defined below

$$\lfloor \mathbf{F}, \mathbf{y} \rceil \equiv \sum_{j=1}^{n} \mathbf{y}_j \mathbf{F}_j, \qquad \mathbf{y} \equiv \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, \qquad \mathbf{F} = [\mathbf{F}_1 \cdots \mathbf{F}_n] \tag{12}$$

The recursion equation of the squarification and the calculation of the squtelmats are detailed in [18–23]. The above recursion amongst squtelmats has a computational complexity which is much more than the computational complexity of the following vector recursion

 ISBN: 978-84-617-8694-7

which can be obtained from the postmultiplication of the both sides in the above squtelmat recursion with the initial vector.

$$\mathbf{v}_j = \mathbf{S}_j(\mathbf{a})\mathbf{a}, \qquad \mathbf{v}_j = \sum_{k=0}^{j-1} \binom{j-1}{k} \lfloor\, \mathbf{F}, \mathbf{v}_k \,\rceil\, \mathbf{v}_{j-k-1}, \qquad j = 1, 2, ... \qquad \mathbf{v}_0 = \mathbf{a} \tag{13}$$

This recursion is expected to be implemented much more rapidly than the relevant matrix recursion and our implementations confirm this point.

2 Sensitivity Coefficients

The use of the above $\mathbf{v}$ vectors enable us to rewrite the PREVTH solution as follows

$$\mathbf{x}(t) = \mathrm{e}^{\beta t} \sum_{j=0}^{\infty} \frac{1}{j!} \left(\frac{\mathrm{e}^{\beta t} - 1}{\beta} \right)^j \mathbf{v}_j(\mathbf{a}) \tag{14}$$

We have nothing further to do with this formula and the above recursion but just to use. However, there is an important question about the parameterization of the PREVTH ODEs and relevant solutions together with the above recursion: "How do the solution vector elements of PREVTH solution changes when the parameters in the equations change?". Parameters can come from the right hand function or initial vector structure. We can investigate these changes simply by evaluate the first partial derivative of the solution with respect to those parameters. We call these derivatives "Sensitivity Coefficients". For simplicity we are going to focus on the sensitivity coefficients with respect to the initial vector elements in this work.

If we denote the parameter, with respect to which we are going to evalaute sensitivity coefficient by σ then we can write

$$\mathbf{F} = \mathbf{F}(\sigma), \qquad \mathbf{v}_j = \mathbf{v}_j(\sigma), \qquad \mathbf{a} = \mathbf{a}(\sigma) \tag{15}$$

where σ dependence of $\mathbf{v}$ vectors are in fact thru a few channel, from F and $\mathbf{a}$ dependencies. The partial differentiation of above vector recursion with respect to σ produces

$$\begin{aligned} \frac{\partial \mathbf{v}_j}{\partial \sigma} \;=\; & \sum_{k=0}^{j-1} \binom{j-1}{k} \left(\left\lfloor \mathbf{F}, \frac{\partial \mathbf{v}_k}{\partial \sigma} \right\rceil \mathbf{v}_{j-k-1} + \lfloor \mathbf{F}, \mathbf{v}_k \rceil \frac{\partial \mathbf{v}_{j-k-1}}{\partial \sigma} \right) \\ & + \sum_{k=0}^{j-1} \binom{j-1}{k} \left\lfloor \frac{\partial \mathbf{F}}{\partial \sigma}, \mathbf{v}_k \right\rceil \mathbf{v}_{j-k-1}, \qquad \frac{\partial \mathbf{v}_0}{\partial \sigma} = \frac{\partial \mathbf{a}}{\partial \sigma} \end{aligned} \tag{16}$$

$$\frac{\partial \mathbf{x}}{\partial \sigma} = \mathrm{e}^{\beta t} \sum_{j=0}^{\infty} \frac{1}{j!} \left(\frac{\mathrm{e}^{\beta t} - 1}{\beta} \right)^j \frac{\partial \mathbf{v}_j}{\partial \sigma} \tag{17}$$

 ISBN: 978-84-617-8694-7

In the case where σ is taken as a_ℓ we can write

$$\frac{\partial \mathbf{V}_0}{\partial \mathbf{a}_1} = \mathbf{e}_\ell, \qquad \ell = 1, 2, 3, ..., n \tag{18}$$

where $\mathbf{e}_\ell$ stands for the ℓth standard unit vector whose only nonzero element is 1 and is positioned on the ℓth position.

3 Implementations

The following four ODEs can be written for Henon-Heiles system which is given by [24, 25]

$$\dot{x}_1 = x_2, \qquad \dot{x}_2 = -x_1 - 2x_1x_3, \qquad \dot{x}_3 = x_4, \qquad \dot{x}_4 = -x_3 - x_1^2 + x_3^2 \tag{19}$$

which are conical at the right hand side genuinely. However, the first degree terms do not have a scalar matrix (identity matrix scaled by a scalar) coefficient. On the other hand, the use of CASE (Constancy Additional Space Extension) adds a constant function with an arbitrary value to the unknowns and therefore makes a space extension by a constant. This creates 5-unknown ODEs whose right hand side $\mathbf{F}_1^{(aug)}$ matrix is now $\beta\mathbf{I}_5$ and the other matrix, 5×25 type $\mathbf{F}_2^{(aug)}$ has unknown parameters like β and the constant coming from CASE. Those parameters are determined to get a minimum norm form the 5×25 type rectangular. After this optimization β and the CASE constant are found as zero while the new form of 5×25 type $\mathbf{F}_2^{(aug)}$ becomes having only 4 nonzero elements which are explicitly given below

$$\left[F_2^{(aug)} \right]_{2,3} = -1, \qquad \left[F_2^{(aug)} \right]_{2,9} = -1, \qquad \left[F_2^{(aug)} \right]_{4,1} = -1, \qquad \left[F_2^{(aug)} \right]_{4,11} = 1. \tag{20}$$

The accompanying initial conditions in the implementations of this section have been taken as follows

$$\mathbf{x}(0) = \mathbf{a} \equiv [\ 0.1 \quad 0.2 \quad 0.3 \quad 0.4\ ]^T \tag{21}$$

In the first implementation the general σ variable has been taken as a_1 and approximants formed by 2, 3, 4, 5, 6 and 10 terms truncations are plotted in the same graphics separately for the first, second, third, and, fourth components of the 5 element PREVTH solution. These are evaluated by using the vector recursion above and the fifth element which in fact corresponds to zero function has not been shown. The variation of the approximants are given throughout the interval $[\,0, 1\,]$. The convergence is at easily noticable level as seen from the plots.

 ISBN: 978-84-617-8694-7

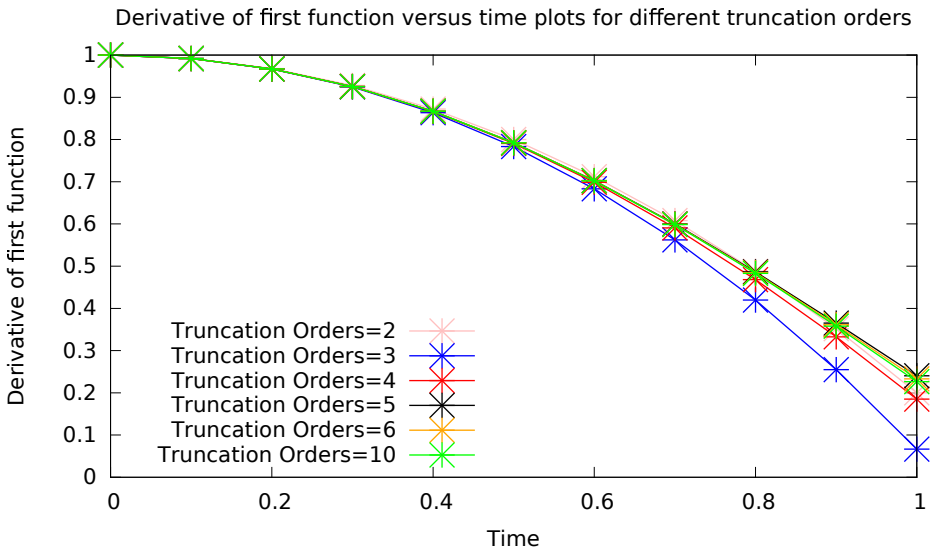


Figure 1: Derivative of the first PREVTH solution vector component with respect to the first element of the initial vector

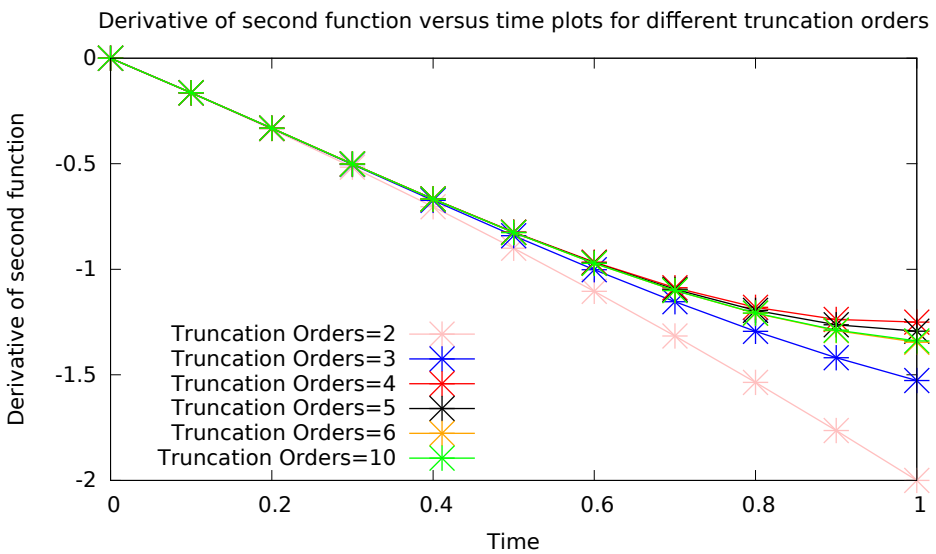


Figure 2: Derivative of the second PREVTH solution vector component with respect to the first element of the initial vector

 ISBN: 978-84-617-8694-7

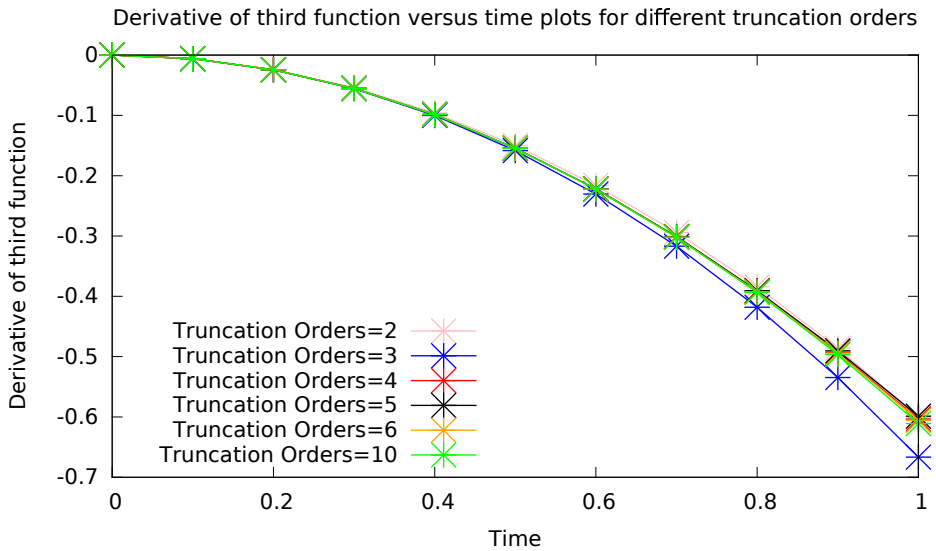


Figure 3: Derivative of the third PREVTH solution vector component with respect to the first element of the initial vector

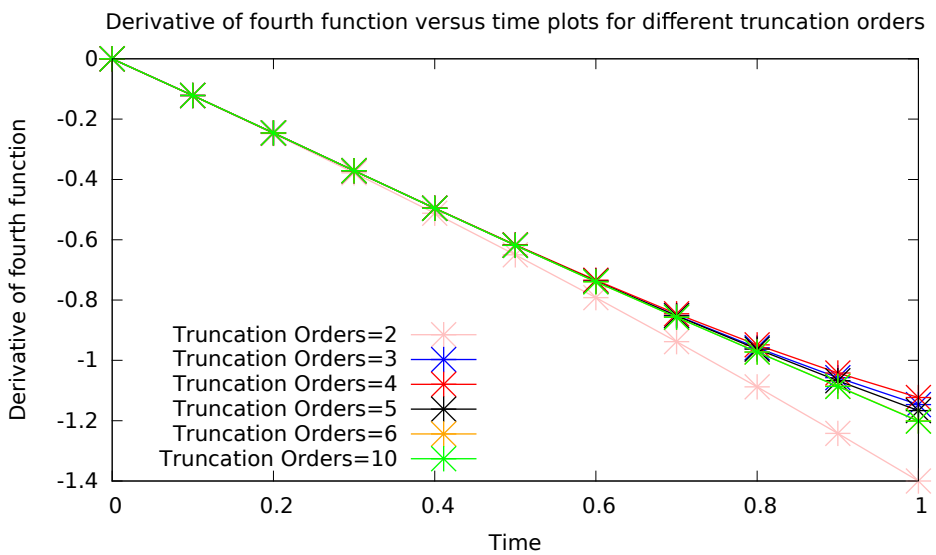


Figure 4: Derivative of the fourth PREVTH solution vector component with respect to the first element of the initial vector

For these truncations the absolute errors between certain truncation approximants can

 ISBN: 978-84-617-8694-7

also be given in the following graphics.

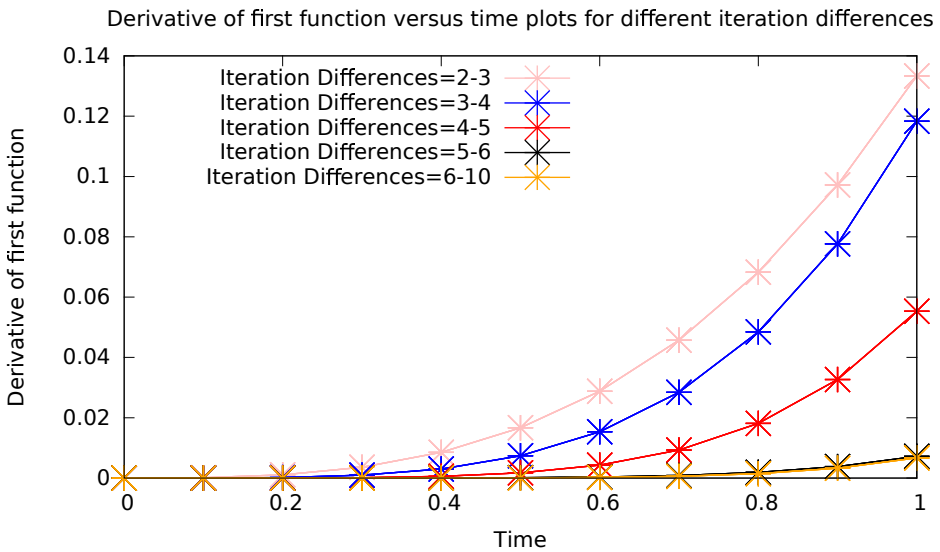


Figure 5: Truncation approximant comparisons for the derivative of the first PREVTH solution vector component with respect to the first element of the initial vector

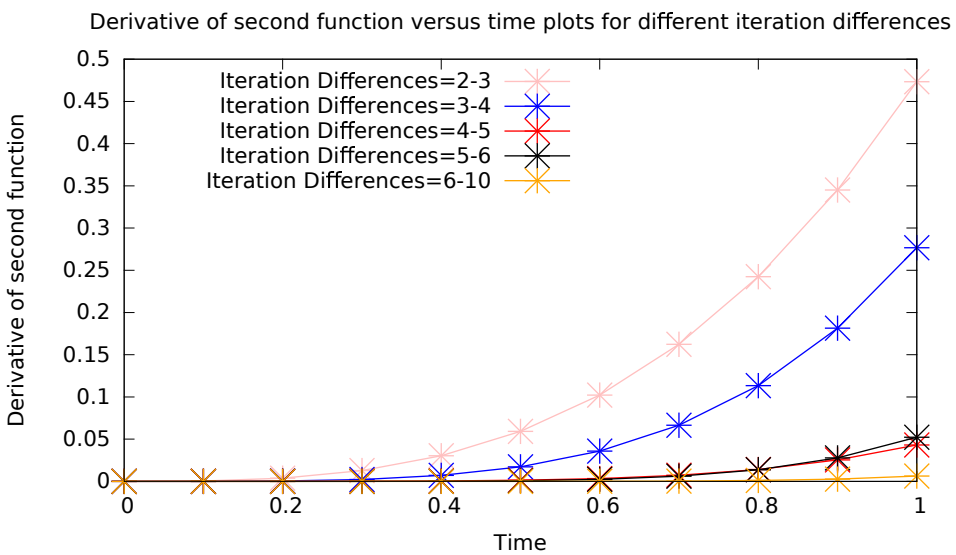


Figure 6: Truncation approximant comparisons for the derivative of the second PREVTH solution vector component with respect to the first element of the initial vector

 ISBN: 978-84-617-8694-7

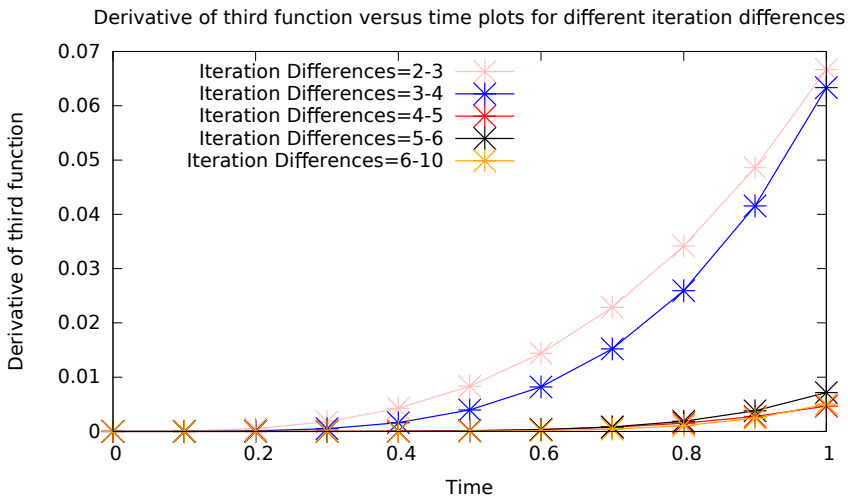


Figure 7: Truncation approximant comparisons for the derivative of the third PREVTH solution vector component with respect to the first element of the initial vector

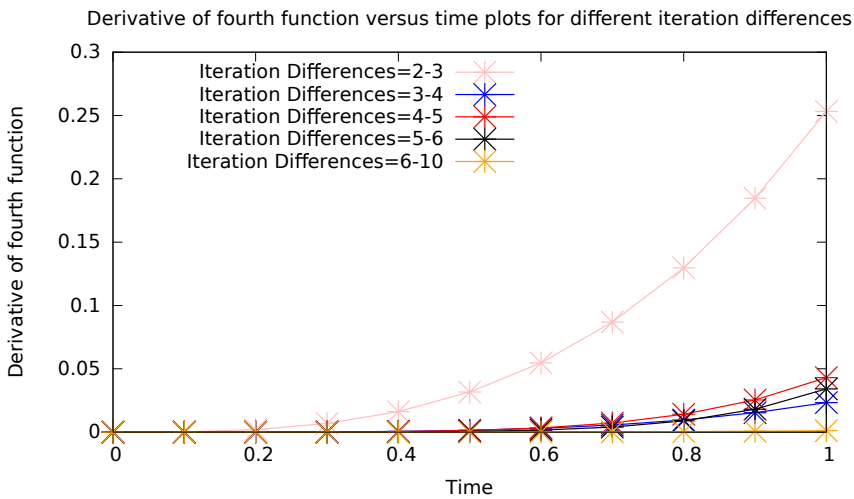


Figure 8: Truncation approximant comparisons for the derivative of the fourth PREVTH solution vector component with respect to the first element of the initial vector

 ISBN: 978-84-617-8694-7

Even though we have evaluated the derivatives of the PREVTH solution vector elements with respect to all elements of the initial vector, we have reported only the partial derivatives with respect to the first element of the initial vector because of the typographical room insufficency. However, the convergence behavior of all cases seem to be almost same and the rather lower truncation order approximants seem to be sufficent for moderate qualities.

Conclusion

In this work we have attempted to evaluate certain sensitivity coefficients by using PREVTH solutions accompanied by a recently developed recursion. Our all implementations support what we expect theoretical aspects of PREVTH. The sensitivity equations are in fact linear and facilitates the solution techniques. We intend to continue our studies towards this direction in our future works.

References

[1] M. Demiralp, *Quantum Expected Value Dynamics in Probabilistic Evolution Perspective*, Proceedings of the 12th International Conference on Computational and Mathematical Methods in Science and Engineering (CMMSE), ISBN: 978-84-615-5392-1, Murcia, Spain (2012) 449-459.

[2] M. Demiralp, *Singing in the Magic Empire of Stars: Probabilistic Evolution Approach to Celestial Mechanical Problems*, Proceedings of the 12th WSEAS International Conference on Systems Theory and Scientific Computation (ISTASC'12), ISBN: 978-1-61804-115-9, İstanbul, Türkiye (2012) 14.

[3] M. Demiralp, Journal of Mathematical Chemistry, URL doi:http://dx.doi. org/10.1007/s10910-012-0079-6 **51**(4) (2012) 1170.

[4] M. Demiralp and B. Tunga, Journal of Mathematical Chemistry, URL doi:http:// dx.doi.org/10.1007/s10910-012-0081-z **51**(4) (2012) 1198.

[5] M. Demiralp, E. Demiralp and L. Hernandez-Garcia, Journal of Mathematical Chemistry, URL doi:http://dx.doi.org/10.1007/s10910-011-9929-x **50** (2012) 850.

[6] M. Demiralp and E. Demiralp, Journal of Mathematical Chemistry, URL doi: http://dx.doi.org/10.1007/s10910-012-0070-2 **51**(1) (2012) 58.

[7] M. Demiralp and E. Demiralp, Journal of Mathematical Chemistry, URL doi: http://dx.doi.org/10.1007/s10910-012-0064-0 **51**(19) (2012) 38.

[8] T. Öztürk and M. Demiralp, *Classical Dynamics of Isolated Univariate Quartic Anharmonic Oscillator via Probabilistic Evolution*, Proceedings of the 12th WSEAS International Conference on Systems Theory and Scientific Computation (ISTASC'12), ISBN: 978-1-61804-115-9, İstanbul, Türkiye (2012) 224-228.

[9] S. Tuna and M. Demiralp, *Certain Validations of Probabilistic Evolution Approach for Initial Value Problems*, Proceedings of the 12th WSEAS International Conference on Systems Theory and Scientific Computation (ISTASC'12), ISBN: 978-1-61804-115-9, İstanbul, Türkiye (2012) 246-249.

[10] S. Bayat and M. Demiralp, *Quantum Optimal Control Theoretical Observable Transitions Between State and Costate in Probabilistic Evolution Perspective*, Proceedings of the 12th WSEAS International Conference on Systems Theory and Scientific Computation (ISTASC'12), ISBN: 978-1-61804-115-9, İstanbul, Türkiye (2012) 272-277.

[11] N. A. Baykara, E. Gürvit and M. Demiralp, *Univariate single quantum harmonic oscillator from probabilistic evolution perspective*, Proceedings of the 13th WSEAS International Conference on Mathematics and Computers in Biology and Chemistry (MCBC'12), Wisconsin, ABD (2012) 27-32.

[12] M. Ayvaz and M. Demiralp, *Getting Triangularity and Conicality in the Probabilistic Evolutionary Expectation Dynamics of the Purely Quartic Quantum Anharmonic Oscillator*, Proceedings of the 12th WSEAS International Conference on Systems Theory and Scientific Computation (ISTASC'12), ISBN: 978-1-61804-115-9, İstanbul, Türkiye (2012) 268-271.

[13] F. Hunutlu, N. A. Baykara and M. Demiralp, *Truncation Approximants to Probabilistic Evolution for ODEs Having Two Diagonal Banded Evolution Matrices Under Initial Conditions: Simple Case*, Proceedings of the 12th International Conference on Computational and Mathematical Methods in Science and Engineering (CMMSE), ISBN: 978-84-615-5392-1, Murcia, Spain (2012) 720-731.

[14] E. Demiralp, M. Demiralp and L. Hernandez-Garcia, *Journal of Mathematical Chemistry*, URL doi:http://dx.doi.org/10.1007/s10910-011-9930-4 **50** (2012) 870.

[15] E. Gürvit and M. Demiralp, *Enhanced Multivariate Product Representation at Constancy Level in Probabilistic Evolution Approach to First Order Explicit ODEs*, Proceedings of the 12th WSEAS International Conference on Systems Theory and Scientific Computation (ISTASC'12), ISBN: 978-1-61804-115-9, İstanbul, Türkiye (2012) 229-234.

[16] B. Tunga and M. Demiralp, *Probabilistic Evolution of the State Variable Expected Values in Liouville Equation Perspective, for a Many Particle System Interacting Via Elastic Forces*, Proceedings of the 12th International Conference on Computational and Mathematical Methods in Science and Engineering (CMMSE), ISBN: 978-84-615-5392-1, Murcia, Spain (2012) 1186-1197.

[17] B. Tunga and M. Demiralp, *Probabilistic evolutions in classical dynamics: Conicalization and block triangularization of Lennard-Jones systems*, AIP Conference Proceedings, **1479**(1) (2012) 1986-1989.

[18] M. Demiralp, *Squarificating the Telescope Matrix Images of Initial Value Vector in Probabilistic Evolution Theory (PET)*, Proceedings of the 19th International Conference on Applied Mathematics (AMATH'14), ISBN: 978-1-61804-258-3, İstanbul, Türkiye (2014) 99104.

[19] M. E. Kirkin and C. Gözükirmizi, *Probabilistic Evolution Theory for ODE Sets with Second Degree Multinomial Right Hand Side Functions: Certain Reductive Cases* ICCMSE, Athens, Greece (2015).

[20] C. Gözükirmizi, *Probabilistic Evolution Theory for ODE Sets with Second Degree Multinomial Right Hand Side Functions: Implementation*, ICCMSE, Athens, Greece (2015).

[21] M. E. Kirkin and M. Demiralp, *A Case Study on Squarification in Probabilistic Evolution Theory (PREVTH) for Henon-Heiles Systems*, 10th International Conference on Applied Mathematics, Simulation, Modelling (ASM'16), International Journal of Computers, İstanbul, Türkiye (2016) 158-165.

[22] C. Gözükirmizi and M. E. Kirkin, *Classical Symmetric Fourth Degree Potential Systems in Probabilistic Evolution Theoretical Perspective: Most Facilitative Conicalization and Squarification of Telescope Matrices*, International Conference in Nonlineer Problems in Aviation and Aerospace (ICNPAA), La Rochelle, France (2016).

[23] C. Gözükirmizi, M. E. Kirkin and M. Demiralp, Journal of Mathematical Chemistry, URL doi:http://dx.doi.org/10.1007/s10910-016-0678-8 (2017) 1-20.

[24] M. Henon and C. Heiles, *The Applicability of the Third Integral of Motion: Some Numerical Experiments*, The Astrophysical Journal, URL doi:http://dx.doi.org/10.1086/109234 **69** (1964) 73-79.

[25] C. H. Skiadas and C. Skiadas, *Chaotic Modeling and Simulation Analysis of Chaotic Models, Attractors and Forms*, ISBN: 9781420079005, CRC Press (2009).

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

Fast Numerical Method for Solving Delta Greek for a Class of Non-linear Option Pricing Models

Miglena N. Koleva[1] **and Lubin G. Vulkov**[1]

[1] *Department of Mathematics and Applied Mathematics and Statistics, University of Ruse*

emails: koleva@uni-ruse.bg, lvalkov@uni-ruse.bg

Abstract

In this paper, we are concentrated on solving Delta equation for a class of non-linear option pricing models. The unknown solution is the first spatial derivative of the option value - Greek Delta. We develop and analyze monotone finite difference method for the model problem. Some basic properties of the numerical scheme are established. To improve the efficiency of the computations, two-grid method is used.

Key words: Delta Greek, Delta equation, finite difference scheme, monotonicity, convergence, two-grid method

1 Introduction and possing the problem

We consider Black-Scholes equation with non-linear volatility term

$$\frac{\partial V}{\partial t} + \frac{1}{2}\sigma^2\left(S, t, \frac{\partial^2 V}{\partial S^2}\right) S^2 \frac{\partial^2 V}{\partial S^2} + (r-q)S\frac{\partial V}{\partial S} - rV, \quad S > 0, \quad 0 \le t \le T, \tag{1}$$

where the solution $V = V(S,t)$ depends on the time variable t and the underlying asset price S. The other parameters are: $r > 1$ is the interest rate, $q \ge 0$ is the dividend yield rate, T is the maturity. Such modifications of the classical option pricing equation, model for instance, presence of constant transaction costs (Leland [4]), non-constant transactions costs (see e.g. Ševčovič and Žitňanská [5]), imperfect replication and investor's preferences (cf. Barles and Soner [2]), etc.

The so called Greeks, representing the sensitivity of the price of derivatives, are very important tools in mathematical engineering. In particular Greek Delta - the first spatial

 ISBN: 978-84-617-8694-7

derivative of the option value, measure the rate of change of the option value with respect to changes in the underlying asset's price.

Let $W = \frac{\partial V}{\partial S}$. Applying time inversion $t = T - t$ and differentiating (1) with respect to S, we derive Delta equation

$$\frac{\partial W}{\partial t} - \frac{1}{2}\frac{\partial}{\partial S}\left[\sigma^2\left(S, t, \frac{\partial W}{\partial S}\right) S^2 \frac{\partial W}{\partial S}\right] + qW - (r-q)S\frac{\partial W}{\partial S} = 0, \ (S,t) \in (0, S_{\max}) \times (0,T), \quad (2)$$

The semi-infinite domain is truncated by large enough computational interval $[0, S_{\max}]$. The problem is completed with initial and boundary conditions, which depends on the type of the option. Let $H(x) = 1_{[0,\infty)}(x)$ stands for the Heaviside function and g_1', g_2^W, g_3^W are known functions. The initial conditions for some popular options are

$$g_1'(S) = \begin{cases} H(S-K) & \text{for Vanilla call,} \\ -H(K-S) & \text{for Vanilla put,} \\ H(S-K_1) - 2H(S-K) + H(S-K_2) & \text{for Butterfly Spread,} \end{cases} \quad (3)$$

The corresponding boundary conditions for (2): g_2^W at $S = 0$ and g_3^W at $S = S_{\max}$, are

$$g_2^W = \begin{cases} 0 & \text{for Vanilla call,} \\ -1 & \text{for Vanilla put,} \\ 0 & \text{for Butterfly Spread,} \end{cases} \qquad g_3^W = \begin{cases} 1 & \text{for Vanilla call,} \\ 0 & \text{for Vanilla put,} \\ 0 & \text{Butterfly Spread.} \end{cases} \quad (4)$$

The aim of this work is to construct and investigate appropriate and efficient numerical method for solving the model problem (2)-(4).

In the papers [6, 7] a numerical approach for computing the Delta Greek and the option price of the Black-Scholes-Barenblatt equation is developed. In our previous paper [3] a first order upwind numerical method is constructed for solving (2)-(4).

2 Numerical method

Finite difference discretization. We define non-uniform meshes in space and time. The space step is $h_i = S_{i+1} - S_i$, $i = 0, 1, \ldots, M-1$, $S_0 = 0$, $S_M = S_{\max}$ and time step $\triangle t_n = t_{n+1} - t_n$, $n = 0, 1, \ldots, N-1$. Let $\hbar_0 = h_0/2$, $\hbar_i = (h_{i-1} + h_i)/2$, $i = 1, 2, \ldots, M-1$, $\hbar_M = h_{M-1}/2$ and the numerical solution at point (S_i, t^n) is denoted by W_i^n.

Using upwind scheme, combined with 'maximal use of central differencing' [8] for the convection term, we construct the weighted ($\theta \in [0,1]$) discretization of (2). For $i = 1, 2, \ldots, M-1$, $n = 0, 1, \ldots, N-1$ we have

$$\begin{aligned} \frac{W_i^{n+1} - W_i^n}{\triangle t_n} - \frac{\theta}{2\hbar_i}\Big[& S_{i+1/2}^2 \widehat{\sigma}_{i+1/2}^{2,n+1} (W_S)_i^{n+1} - S_{i-1/2}^2 \widehat{\sigma}_{i-1/2}^{2,n+1} (W_{\bar{S}})_i^{n+1}\Big] \\ & - \theta(r-q)^+ S_i [\chi_i^+ (W_S)_i^{n+1} + (1-\chi_i^+)(W_{\mathring{S}})_i^{n+1}] \\ & + \theta(r-q)^- S_i [\chi_i^- (W_{\bar{S}})_i^{n+1} + (1-\chi_i^-)(W_{\mathring{S}})_i^{n+1}] + q\theta W_i^{n+1} \end{aligned} \quad (5)$$

 ISBN: 978-84-617-8694-7

$$= \frac{(1-\theta)}{2\hbar_i}\left[S^2_{i+1/2}\widehat{\sigma}^{2,n}_{i+1/2}(W_S)^n_i - S^2_{i-1/2}\widehat{\sigma}^{2,n}_{i-1/2}(W_{\bar{S}})^n_i\right]$$
$$+(1-\theta)(r-q)^+S_i[\chi^+_i(W_S)^n_i + (1-\chi^+_i)(W_{\mathring{S}})^n_i]$$
$$-(1-\theta)(r-q)^-S_i[\chi^-_i(W_{\bar{S}})^n_i + (1-\chi^-_i)(W_{\mathring{S}})^{n+1}_i] + q(1-\theta)W^n_i,$$

where $(W_S)^n_i = (W^n_{i+1} - W^n_i)/h_i$, $(W_{\bar{S}})^n_i = (W_S)^n_{i-1}$, $(W_{\mathring{S}})^n_i = (h_{i-1}(W_S)^n_i + h_i(W_{\bar{S}})^n_i)/2\hbar_i$,

$$\widehat{\sigma}^{2,n+1}_{i+1/2} := \sigma^2\left(S_{i+1/2}, t_{n+1}, (W_S)^{n+1}_i\right), \quad \widehat{\sigma}^{2,n+1}_{i-1/2} := \sigma^2\left(S_{i-1/2}, t_{n+1}, (W_{\bar{S}})^{n+1}_i\right),$$

$$\chi^+_i = \begin{cases} 0, & h_i < \dfrac{S^2_{i-1/2}\widehat{\sigma}^{2,n+1}_{i-1/2}}{(r-q)^+S_i}, \\ 1, & \text{otherwise}, \end{cases} \quad \chi^-_i = \begin{cases} 0, & h_{i-1} < \dfrac{S_{i+1/2}\widehat{\sigma}^{2,n+1}_{i+1/2}}{(r-q)^-}, \\ 1, & \text{otherwise}. \end{cases}$$

The finite difference scheme is completed with boundary and initial conditions

$$W^n_0 = g^W_2(t_n),\ W^n_M = g^W_3(t_n),\ n = 1, \ldots, N, \quad W^0_i = g'_1(S_i),\ i = 0, \ldots, M. \tag{6}$$

Convergence. On the base of the results in [1] for a second order non-linear PDE, from stability, monotonicity and consistency of the numerical discretization follows convergence of the numerical solution to the viscosity solution.

Lemma 1. (Stability) *If the following restriction is fulfilled*

$$\begin{aligned} \triangle t_n \leq & \frac{1}{1-\theta}\left(\frac{1}{2\hbar_i h_{i-1}}S^2_{i-1/2}\widehat{\sigma}^{2,n}_{i-1/2} + \frac{1}{2\hbar_i h_i}S^2_{i+1/2}\widehat{\sigma}^{2,n}_{i+1/2}\right. \\ & +(r-q)^-S_i\left[\frac{\chi^-_i}{h_{i-1}} + \frac{(1-\chi^-_i)h_i}{2\hbar_i}\left(\frac{h_i}{h_{i-1}} - \frac{h_{i-1}}{h_i}\right)\right] \\ & \left.+(r-q)^+S_i\left[\frac{\chi^+_i}{h_i} + \frac{(1-\chi^+_i)h_{i-1}}{2\hbar_i}\left(\frac{h_{i-1}}{h_i} - \frac{h_i}{h_{i-1}}\right)\right] + q\right)^{-1}, \end{aligned} \tag{7}$$

then the solution of the discretization (5), (6) *satisfies the estimate*

$$\|W^{n+1}\|_\infty \leq \max\{\|g'_1\|_\infty, \|g^W_3\|_\infty, \|g^W_3\|_\infty\}.$$

Lemma 2. (Monotonicity) The discretization (5), (6), $\theta = 1$ is monotone, if

$$h_i \leq \frac{S^2_{i-1/2}}{S_i|r-q|}\min\left\{\frac{1}{(1-\chi^-_{i+1})}\frac{\partial(\widehat{\sigma}^{2,n+1}_{i+3/2}U^{n+1}_{i+3/2})}{\partial U^{n+1}_{i+3/2}}, \frac{1}{(1-\chi^+_i)}\frac{\partial(\widehat{\sigma}^{2,n+1}_{i-1/2}U^{n+1}_{i-1/2})}{\partial U^{n+1}_{i-1/2}}\right\}. \tag{8}$$

Lemma 3. *The discretization* (5), (6) *is consistent.*

Theorem 1. *Let the conditions of Lemmas 1 - 3 are fulfilled. Then the solution of* (5), (6), $\theta = 1$, *converges to the viscosity solution* as $(|h|, \triangle t) \to (0^+, 0^+)$, where $|h| = \max\limits_{0\leq i\leq M} h_i$ and $\triangle t = \max\limits_{0\leq n\leq N-1} \triangle t_n$.

 ISBN: 978-84-617-8694-7

Numerical implementation. Let define a new - fine space mesh with step size $h_i^f << h_i$. To improve the computational efficiency of the numerical method (5), (6) we apply the two-grid idea. At each time level, we solve the non-linear system (5), (6) by Picard or Newton iteration process on the coarse mesh (with step size h_i). Then, on the fine mesh, we perform only one Newton/Picard iteration.

Various numerical experiments were performed and they confirm the efficiency of the proposed method.

Acknowledgements

This research has been supported by the Bulgarian National Fund of Science under Project I02/20-2014.

References

[1] G. Barles, *Convergence of numerical schemes for degenerate parabolic equations arising in finance*, in: Numerical Methods in Finance, L. Rogers, D. Talay (Eds.), Cambridge University Press, Cambridge, 1997.

[2] G. Barles, M. H. Soner, *Option pricing with transaction costs and a nonlinear Black-Scholes equation.*, Finance Stoch. **2** (1998) 369–397.

[3] M. N.Koleva, L. G. Vulkov, *Computation of Delta Greek for non-linear models in mathematical finance*, LNCS **10187** (2017) 430–438.

[4] H. E.Leland, *Option pricing and replication with transaction costs*, J. Finance **40**, (1985) 1283–1301.

[5] D. Ševčovič, M. Žitňanská , *Analysis of the nonlinear option pricing model under variable transaction costs*, Asia-Pacific Financial Markets **23**(2) (2016) 153–174.

[6] R. Valkov, *Fitted strong stability-preserving schemes for the Black-Scholes-Barenblatt equation*, Int. J. of Computer Math. **92**(12) (2015) 2475–2497.

[7] R. Valkov, *Predictor-Corrector balance method for the worst-case 1D option pricing*, Comput. Meth. in Appl. Math. **16**(1) (2015) 175–186.

[8] J. Wang, P. Forsyth, *Maximal use of central differencing for Hamilton-Jacobi-Bellman PDEs in finance*, SIAM J. Numer. Anal. **46**(3) (2008) 1580–1601.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

A Numerical Study of a Semilinear Parabolic System of Optimal Regime-Switching

Miglena N. Koleva[1] **and Lubin G. Vulkov**[1]

[1] *Department of Mathematics and Applied Mathematics and Statistics, University of Ruse*

emails: koleva@uni-ruse.bg, lvalkov@uni-ruse.bg

Abstract

In this work we consider a system of weakly coupled semi-linear parabolic equations of optimal portfolio in a regime-switching model in the case of exponential utility function, proposed by A.R. Valdez and T. Vargiolu [8]. We develop efficient finite difference method for 2D case. For the approximation of the convection term we implement van Leer flux limiter technique and for the discretization of the mixed derivatives we apply different stencils, depending on the sign of the correlation. We prove comparison principle and convergence for the approximated solution.

Key words: Regime-switching model, system of semi-linear parabolic PDE, exponential non-linearity, finite difference scheme, flux limiter, convergence

1 Introduction and model formulation

Following the regime-switching model in [8], we consider the PDE system

$$C_t^k + rSC_S^k + \tfrac{1}{2}\mathrm{tr}(\overline{S}\sum_k \sum_k^T \overline{S} C_{SS}^k) + \frac{e^{-r(T-t)}}{\alpha}\left[\sum_{j=1}^m (e^{-\alpha\phi(C^k - C^j)} - 1)\lambda^{kj} - \tfrac{1}{2}z_k^2\right] = rC^k, \quad C^k(T) = 0. \tag{1}$$

Here $C(t,S) := (C^k(t,S))_{k=1,\dots,m} \in C_b^{1,2}([0,T]\times B, \mathbb{R}^m)$, $B = (0,\infty)^d$, r is the risk-free interest rate, $\phi(t) := e^{r(T-t)}$, $\alpha > 0$, $\mu_k = \mu_k(t,S) : [0,T]\times B \subset \mathbb{R}^d \to \mathbb{R}^d$, $\sum_k \sum_k^T$ are locally Lipschitz and bounded, $\sum_k = \sum_k(t,S) : [0,T]\times B \subset \mathbb{R}^d \to \mathbb{R}^{d\times d}$ is nonsingular for all (t,S), $\sum_k^{-1}\mu_k$ and $\lambda^{kj} : [0,T]\times B \to [0,\infty)$, $\lambda^{kj} \in C_b^1([0,T]\times B)$ are bounded on $\Omega_T = [0,T]\times B$, for all $k,j = 1,\dots,m$, $\mathbf{1}$ is the d-dimensional unit column vector,

 ISBN: 978-84-617-8694-7

$\overline{S} := \mathrm{diag}(S)$, $S = (S_1, S_2, \ldots, S_d)$, C^k_S is the gradient with respect to S, C^k_{SS} is the Hessian matrix with entries $C^k_{S_iS_j}$, $i, j = 1, 2, \ldots, d$ and functions z^2_k are defined by

$$z^2_k(t, S) := (\mu_k(t, S) - r\mathbf{1})^T \left(\sum_k(t, S) \sum^T_k(t, S)\right)^{-1} (\mu_k(t, S) - r\mathbf{1}).$$

A simpler regime-switching model (represented by a parabolic-ODE system) in the case of exponential utility is derived in [6]. For solving this market model, in [1, 2, 7] are constructed and analyzed efficient finite difference schemes. In our previous work [3] a scalar case of problem (1) is studied.

We consider two-dimensional case of (1) with more general terminal conditions $C^k(T, S) = C^k_0(S)$, $k = 1, 2$. The semi-infinite domain B is truncated by large enough computational region $[0, S_{1\max}] \times [0, S_{2\max}]$, imposing natural boundary conditions. Next, we apply the logarithmic change of the space variables $x_i = \ln S_i$. As a result $[0, S_{i\max}]$ transforms to the semi-infinite domain $(-\infty, \ln S_{i\max}]$. Then, this domain is truncated by large enough computational domain $D = D_1 \times D_2$, $D_i = [L^-_i, L^+_i]$, $L^+_i = \ln S_{i\max}$ and $L^-_i < 0$, $L^+_i > 0$ are real numbers. For $i = 1, 2$ we define the function δ_i

$$\delta_i = \begin{cases} 0, & x_i = L^{\pm}_i, \\ 1, & \text{otherwise}, \end{cases} \qquad \delta^2_i = \delta_i.$$

In order to obtain an initial-value problem we invert the time, setting $\tau := T - t$. Let $\sum_k \sum^T_k = \{\sigma^k_{il}\}^{2,2}_{i,l=1} \geq 0$ and denote $2\rho^k_{il} = \sigma^k_{il} + \sigma^k_{li}$, $k = 1, 2$. Finally, the resulting problem for $(\tau, x) \in Q_T = (0, T] \times D_1 \times D_2$, is

$$C^k_\tau - \delta_1 \left(r - \frac{1}{2}\sigma^k_{11}\right) C^k_{x_1} - \delta_2 \left(r - \frac{1}{2}\sigma^k_{22}\right) C^k_{x_2} + rC^k$$

$$- \frac{1}{2}\left(\delta_1\sigma^k_{11}C^k_{x_1x_1} + 2\delta_1\delta_2\rho^k_{12}C^k_{x_1x_2} + \delta_2\sigma^k_{22}C^k_{x_2x_2}\right) = \frac{e^{-r\tau}}{\alpha}\left(\sum_{j=1}^{m}(e^{-\alpha\phi(C^k - C^j)} - 1)\lambda^{kj} - \frac{1}{2}z^2_k\right) \quad (2)$$

$$C^k(0, x) = C^k_0(e^x), \quad x \in D_1 \times D_2.$$

We prove comparison principle for the differential problem (2).

2 Numerical method

We consider uniform mesh in space $\overline{\omega}_h = \overline{\omega}_{h_1} \times \overline{\omega}_{h_2}$ with mesh step size h_i in x_i direction

$$\overline{\omega}_{h_i} = \left\{x_{i,j_i} : x_{i,j_i} = L^-_i + (j_i - 1)h_i,\ j_i = 1, \ldots, N_i, \quad h_i = (L^+_i - L^-_i)/(N_i - 1)\right\}$$

and non-uniform mesh ω_τ in time with time increments $\triangle\tau^n$, i.e. $\tau^{n+1} = \tau^n + \triangle\tau^n$, $n = 0, 1, \ldots, N_\tau$. The numerical solution at grid nodes from $Q^h_T = \omega_\tau \times \overline{\omega}_h$ is denoted by

 ISBN: 978-84-617-8694-7

$C^k_{j_1,j_2} := C^k(\tau^n, x_{1j_1}, x_{2j_2})$ and $\widehat{C}^k_{j_1,j_2} := C^k(\tau^{n+1}, x_{1j_1}, x_{2j_2})$. Further, we use the following notations for the derivative approximations

$$C^k_{t_{j_1,j_2}} = \frac{\widehat{C}^k_{j_1,j_2} - C^k_{j_1,j_2}}{\triangle\tau^n}, \quad C^k_{\overline{x}_1 j_1,j_2} = \frac{C^k_{j_1,j_2} - C^k_{j_1-1,j_2}}{h_1}, \quad C^k_{x_1 j_1,j_2} = C^k_{\overline{x}_1 j_1+1,j_2},$$

$$C^k_{\overline{x}_2 j_1,j_2} = \frac{C^k_{j_1,j_2} - C^k_{j_1,j_2-1}}{h_2}, \quad C^k_{x_2 j_1,j_2} = C^k_{\overline{x}_2 j_1,j_2+1}, \quad C^k_{\overline{x}_s x_p} = (C^k_{\overline{x}_s})_{x_p},\ s,p = \{1,2\},$$

$$(C^k)^-_{x_1x_2} = \frac{1}{2}[C^k_{\overline{x}_1 x_2} + C^k_{x_1 \overline{x}_2}], \quad (C^k)^+_{x_1x_2} = \frac{1}{2}[C^k_{x_1x_2} + C^k_{\overline{x}_1\,\overline{x}_2}], \quad (C^k)^\pm = \max\{0, \pm C^k\}.$$

In order to approximate the model problem (2), we use different stencils for the approximation of the mixed derivative, depending on the sign of ρ^k_{12} and for convection term we implement van Leer flux limiter technique [5]. For the exponential term we apply Newton-like linearization. The resulting weighted ($\theta_1, \theta_2, \theta_3 \in [0,1]$) numerical scheme is

$$\begin{aligned}
&C^k_t - \theta_1\delta_1[(A^k_1)^+(\widehat{\Lambda}^k_1)^+\widehat{C}^k_{x_1} - (A^k_1)^-(\widehat{\Lambda}^k_1)^-\widehat{C}^k_{\overline{x}_1}] \\
&\quad - \theta_1\delta_2[(A^k_2)^+(\widehat{\Lambda}^k_2)^+\widehat{C}^k_{x_2} - (A^k_2)^-(\widehat{\Lambda}^k_2)^-\widehat{C}^k_{\overline{x}_2}] + r\theta_2\widehat{C}^k + \theta_3\sum_{j=1}^m e^{-\alpha\widehat{\phi}(C^k-C^j)}(\widehat{C}^k - \widehat{C}^j)\widehat{\lambda}^{kj} \\
&\quad - \frac{1}{2}\theta_2\left(\delta_1\sigma^k_{11}\widehat{C}^k_{\overline{x}_1x_1} + 2\delta_1\delta_2(\rho^k_{12})^+(\widehat{C}^k)^+_{x_1x_2} - 2\delta_1\delta_2(\rho^k_{12})^-(\widehat{C}^k)^-_{x_1x_2} + \delta_2\sigma^k_{22}\widehat{C}^k_{\overline{x}_2x_2}\right) \qquad (3) \\
&= (1-\theta_1)\delta_1[(A^k_1)^+(\Lambda^k_1)^+C^k_{x_1} - (A^k_1)^-(\Lambda^k_1)^-C^k_{\overline{x}_1}] \\
&\quad + (1-\theta_1)\delta_2[\left((A^k_2)^+(\Lambda^k_2)^+C^k_{x_2} - (A^k_2)^-(\Lambda^k_2)^-C^k_{\overline{x}_2}\right)] + \mathcal{F}^m_1 + \mathcal{F}^m_2 - r(1-\theta_2)C^k \\
&\quad + \frac{1-\theta_2}{2}\left(\delta_1\sigma^k_{11}C^k_{\overline{x}_1x_1} + 2\delta_1\delta_2(\rho^k_{12})^+(C^k)^+_{x_1x_2} - 2\delta_1\delta_2(\rho^k_{12})^-(C^k)^-_{x_1x_2} + \delta_2\sigma^k_{22}C^k_{\overline{x}_2x_2}\right),
\end{aligned}$$

coupled with initial conditions $C^k(0,x_1,x_2) = C^k_0(e^{x_1}, e^{x_2})$, $(x_1,x_2) \in \overline{\omega}_h$. The following notations are used in (3)

$$(\Lambda^k_i)^+ = 1 + \frac{1}{2}\Phi((\theta^k_{j_i+1/2})^{-1}) - \frac{1}{2}\Phi(\theta^k_{j_i+3/2}), \quad (\Lambda^k_i)^- = 1 + \frac{1}{2}\Phi(\theta^k_{j_i+1/2}) - \frac{1}{2}\Phi((\theta^k_{j_i-1/2})^{-1}),$$

$$\Phi(\theta^k) = \frac{|\theta^k| + \theta^k}{1 + |\theta^k|}, \quad \theta^k_{j_i+1/2} = \frac{C^k_{x_i j_1,j_2}}{C^k_{\overline{x}_i j_1,j_2}}, \quad i = 1,2,$$

$$\mathcal{F}^m_1 = (1-\theta_3)\frac{e^{-r\tau^n}}{\alpha}\left[\sum_{j=1}^m (e^{-\alpha\phi(C^k-C^j)} - 1)\lambda^{kj} - \frac{1}{2}z^2_k\right],$$

$$\mathcal{F}^m_2 = \theta_3\frac{e^{-r\tau^{n+1}}}{\alpha}\left[\sum_{j=1}^m [(1 + \alpha\widehat{\phi}(C^k - C^j))e^{-\alpha\widehat{\phi}(C^k-C^j)} - 1]\widehat{\lambda}^{kj} - \frac{1}{2}\widehat{z}^2_k\right].$$

Now, if $|\rho^k_{12}|(\sigma^k_{22})^{-1} \le h_1h_2^{-1} \le \sigma^k_{11}|\rho^k_{12}|^{-1}$, then the coefficient matrix of the system (3) is an M-matrix.

 ISBN: 978-84-617-8694-7

Further, for a mild time step restriction, we prove negativity preserving property and convergence in maximal discrete norm of the solution of (3).

Various numerical experiments confirm computational efficiency of the proposed method and validate the theoretical statements.

Acknowledgements

This research has been supported by the Bulgarian National Fund of Science under Project I02/20-2014.

References

[1] M. N. Koleva, W. Mudzimbabwe, L. G. Vulkov, *Fourth-order compact schemes for a parabolic-ordinary system of European option pricing liquidity shocks model*, Numerical Algorithms **74**(1) (2016), 59–75.

[2] M. N. Koleva, L. G. Vulkov, *Fully implicit time-stepping schemes for a parabolic-ODE system of European options with liquidity shocks*, LNCS **9374** (2015) 360–368.

[3] M. N. Koleva, L. G. Vulkov, *A numerical study for optimal portfolio regime-switching model I. 2D BlackScholes equation with an exponential non-linear term*, J. Comp. Appl. Math. **318** (2017) 358–349.

[4] D. C. Lesmana, S. Wang, *An upwind finite difference method for a nonlinear BlackScholes equation governing European option valuation under transaction costs*, Appl. Math. and Comp. **219**(16) (2013) 8811 – 8828.

[5] R. J. LeVeque, *Numerical Methods for Conservation Laws*, Birkhäuser, 1992.

[6] M. Ludkovski and Q. Shen, *European option pricing with liquidity shocks*, Int. J. Theor. Appl. Finance **16**(7) (2013) 135–143.

[7] W. Mudzimbabwe, L. G. Vulkov, *IMEX schemes for a parabolic-ODE system of European options with liquidity shock*, J. Comp. Appl. Math. **299** (2016) 245–256.

[8] A. R. L. Valdez, T. Vargiolu, *Optimal portfolio in a regime-switching model*, In: Proceedings of the Ascona '11 Seminar on Stochastic Analysis, Random Fields and Applications, R. C. Dalang, M. Dozzi, F. Russo (Eds.), (2013) 435 – 449.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

Common Random Fixed Point Theorems for Weakly Compatible Mapping via Implicit Relation in Cone Random Metric Spaces

Chayut Kongban[1] and Poom Kumam[2]

[1] *KMUTTFixed Point Research Laboratory, Department of Mathematics, Room SCL 802 Fixed Point Laboratory,Science Laboratory Building, Faculty of Science, King Mongkuts University of Technology Thonburi (KMUTT),126 Pracha-Uthit Road, Bang Mod, Thrung Khru, Bangkok 10140, Thailand.*

[2] *KMUTTFixed Point Research Laboratory, Department of Mathematics, Room SCL 802 Fixed Point Laboratory,Science Laboratory Building, Faculty of Science, King Mongkuts University of Technology Thonburi (KMUTT),126 Pracha-Uthit Road, Bang Mod, Thrung Khru, Bangkok 10140, Thailand.*

emails: chayut_kb@hotmail.com, poom.kum@kmutt.ac.th

Abstract

In this paper, we investigate and prove common random fixed point theorems in cone random metric spaces for weakly compatible mappings satisfying implicit relations.

Key words: Random Fixed Point, Cone Random Metric Spaces, Implicit Relation
MSC 2000: 47H10, 54H25

1 Introduction

Random fixed point theorems is stochastic generalizations of classical fixed point theorems. The study of random fixed points form a central topic in this area. Random fixed point theorems for random contraction mappings on separable complete metric spaces were first proved by Špaček [1] and Hanš [2]. Subsequently, Bharucha-Reid [3] proved the stochastic version of the well-known Banachs and Schauders fixed point theorem and hence random fixed point theory and applications have been developed rapidly in recent years, see [4, 5, 6, 7]. In 2007, Huang and Zhang [8] defined the cone metric spaces. They also described the

 ISBN: 978-84-617-8694-7

convergence of sequences and introduced the notion of completeness in cone metric spaces and proved some fixed point theorems of contractive mappings on complete cone metric space. In 2008, Rezapour and Hamlbarani [9], showed that there are no normal cones with normal constant $M < 1$, and for each $k > 1$ there are cones with normal constant $M > k$, by providing non-normal cones and omitting the assumption of normality. Random fixed point results in cone random metric spaces are stochastic generalization of deterministic fixed point results in cone metric space. Several other authors see [10, 11, 12] studied the existence of random fixed points and common random fixed points of mappings satisfying contractive type conditions in setting cone random metric spaces. Recently, Rashwan and Hammad [12] a common random fixed point for four weakly compatible mappings on a nonempty separable closed subset of cone random metric spaces. In this paper, we prove common random fixed point theorems for weakly compatible mapping vai implicit relation.

2 Preliminaries

Definition 2.1. *[10] Let $(\mathcal{E}, \tau)$ be a topological vector space. A subset p of $\mathcal{E}$ is called a cone if the following conditions satisfied:*

(c_1) p is closed, nonempty and $p \neq \{0\}$;

(c_2) $a, b \in \mathbb{R}$, $a, b \geq 0$ and $x, y \in p \Rightarrow ax + by \in p$;

(c_3) If $x \in p$ and $-x \in p \Rightarrow x = 0$.

For a given cone $p \subset \mathcal{E}$ we define a partial ordering $\preceq$ with respect to p by $x \preceq y$ if $y - x \in p$. We shall write $x \prec y$ to indicate that $x \preceq y$ but $x \neq y$, while $x \ll y$ will stand for $y - x \in p^{\circ}$, where p° indicate to the interior of p.

Definition 2.2. *[8, 14] Let X be a nonempty set and the mapping the mapping $d : X \times X \to \mathcal{E}$ satisfies:*

(d_1) $0 \leq d(x, y)$ for all $x, y \in X$ and $d(x, y) = 0 \Leftrightarrow x = y$;

(d_2) $d(x, y) = d(y, x)$ for all $x, y \in X$;

(d_3) $d(x, y) \leq d(x, z) + d(z, y)$; $x, y, z \in X$.

Then d is called a cone metric [8] or K-metric [14] on X and (X, d) is called a cone metric space [8].
The concept of a cone metric space is more general than that of a metric space, because each metric space is a cone metric space where $\mathcal{E} = \mathbb{R}$ and $p = [0, +\infty)$.

 ISBN: 978-84-617-8694-7

Definition 2.3. *Let (X, d) be a cone metric space. We say that $\{x_n\}$ is:*

(i) a Cauchy sequence if for every ε in $\mathcal{E}$ with $0 \ll \varepsilon$, then there is an $\mathbb{N}$ such that for all $n, m > \mathbb{N}$, $d(x_n, x_m) \ll \varepsilon$;

(ii) a convergent sequence if for every ε in $\mathcal{E}$ with $0 \ll \varepsilon$, then there is an $\mathbb{N}$ such that for all $n > \mathbb{N}$, $d(x_n, x) \ll \varepsilon$ for some fixed x in X.

A cone metric space X is said to be complete if every Cauchy sequence in X is convergent in X.

Definition 2.4. *Let (Ω, Σ) be a measurable space with Σ- a sigma algebra of subsets of Ω and $\mathcal{M}$ be a nonempty subset of a metric space $X = (X, d)$. Let $2^{\mathcal{M}}$ be the family of nonempty subsets of $\mathcal{M}$ and $C(\mathcal{M})$ the family of all nonempty closed subsets of $\mathcal{M}$. A mapping $G : \Omega \to 2^{\mathcal{M}}$ is called measurable if for each open subset U of $\mathcal{M}$, $G^{-1}(U) \in \Sigma$, where $G^{-1}(U) = \{\omega \in \Omega : G(\omega) \cap U \neq \emptyset\}$*

Definition 2.5. *A mapping $\xi : \Omega \to \mathcal{M}$ is called measurable selector of a measurable mappings $G : \Omega \to 2^{\mathcal{M}}$ if ξ is measurable and $\xi(\omega) \in G(\omega)$ for each $\omega \in \Omega$.*

Definition 2.6. *The mapping $T : \Omega \times \mathcal{M} \to X$ is called a random operator if for each fixed $x \in \mathcal{M}$, the mapping $T(\cdot, x) : \Omega \to X$ is measurable.*

Definition 2.7. *A random operator $T : \Omega \times \mathcal{M} \to X$ is called continuous random operator if for each fixed $x \in \mathcal{M}$ and $\omega \in \Omega$, the mapping $T(\omega, \cdot) : \Omega \to X$ is continuous.*

Definition 2.8. *A measurable mappings $\xi : \Omega \to \mathcal{M}$ is a random fixed point of a random operator $T : \Omega \times \mathcal{M} \to X$ if $T(\omega, \xi(\omega)) = \xi(\omega)$ for each $\omega \in \Omega$.*

Definition 2.9. *Let $\mathcal{M}$ be a nonempty set and the mapping $d : \Omega \times \mathcal{M} \to p$, where p is a cone, $\omega \in \Omega$ be a selector, satisfy the following conditions:*

(i) $d(x(\omega), y(\omega)) \geq 0$ and $d(x(\omega), y(\omega)) = 0 \Leftrightarrow x(\omega) = y(\omega)$ for all $x(\omega), y(\omega) \in \Omega \times \mathcal{M}$,

(ii) $d(x(\omega), y(\omega)) = d(y(\omega), x(\omega))$ for all $x, y \in \mathcal{M}$, $\omega \in \Omega$ and $x(\omega), y(\omega) \in \Omega \times \mathcal{M}$,

(iii) $d(x(\omega), y(\omega)) \leq d(x(\omega), z(\omega)) + d(z(\omega), y(\omega))$ for all $x, y, z \in \mathcal{M}$ and $\omega \in \Omega$ be a selector,

(iv) for any $x, y \in \mathcal{M}, \omega \in \Omega, d(x(\omega), y(\omega))$ is nonincreasing and left continuous.

Then d is called cone random metric on $\mathcal{M}$ and $(\mathcal{M}, d)$ is called a cone random metric space.

 ISBN: 978-84-617-8694-7

Definition 2.10. *Random operators $T, S : \Omega \times X \to X$ are weakly compatible if $T(S(\xi(\omega))) = S(T(\xi(\omega)))$ provided that $T(\xi(\omega)) = S(\xi(\omega))$ for every $\omega \in \Omega$.*

Definition 2.11. *[15] Let F_6 be the family of all continuous mappings $F(t_1, t_2, t_3, t_4, t_5, t_6) : \mathbb{R}_+^6 \to \mathbb{R}$ with $t_3 + t_4 \neq 0$ satisfying the following condition:*
(F_1) there exists $0 \leq h < 1$ such that for all $u, v, w \geq 0$ with

(F_a) $F(u, v, v, u, w, 0) \leq 0$ or

(F_b) $F(u, v, u, v, 0, 0) \leq 0$
we have $u \leq hv$.

3 Main result

Theorem 3.1. *Let (X, d) be a complete cone random metric space with respect to a cone $\mathcal{P}$ and let $\mathcal{M}$ be nonempty separable closed subset of X. Assume that S, T, f, and g be four continuous random operators defined on $\mathcal{M}$ such that $\omega \in \Omega$, $S(\omega, \cdot), T(\omega, \cdot), f(\omega, \cdot), g(\omega, \cdot) : \Omega \times \mathcal{M} \to \mathcal{M}$ satisfying the following conditions:*

(i) $S(\omega, X) \subseteq g(\omega, X)$ and $T(\omega, X) \subseteq f(\omega, X)$,

(ii) the pairs $\{S, f\}$ and $\{T, g\}$ are random weakly compatible mappings,

(iii)

$$\begin{aligned} &F(d(S(x(\omega)), T(y(\omega))), d(f(x(\omega)), g(y(\omega))), d(f(x(\omega)), S(x(\omega))), \\ &d(g(y(\omega)), T(y(\omega))), d(f(x(\omega)), T(y(\omega))), d(S(x(\omega)), g(y(\omega)))) \leq 0 \end{aligned} \tag{1}$$

for all $x(\omega), y(\omega) \in \Omega \times X$ and $F \in F_6$.
Satifies (F_1) if $d(f(x(\omega)), S(x(\omega))) + d(g(y(\omega)), T(y(\omega))) \neq 0$, or $d(S(x(\omega)), T(x(\omega))) = 0$ if $d(f(x(\omega)), S(x(\omega))) + d(g(y(\omega)), T(y(\omega))) = 0$.
Then the four random mapping have unique common random fixed point in X.

Proof. For each $x_0(\omega), x_1(\omega) \in \Omega \times X$ and $n = 1, 2, 3, \ldots$, we choose $y_1(\omega), y_2(\omega) \in \Omega \times X$ such that

$$y_1(\omega) = S(x_0(\omega)) = g(x_1(\omega))$$

and

$$y_2(\omega) = T(x_1(\omega)) = f(x_2(\omega))$$

In general we construct a sequence of measurable mappings $x_n(\omega), y_n(\omega) : \Omega \to X$ defined by

$$y_{2n+1}(\omega) = S(x_{2n}(\omega)) = g(x_{2n+1}(\omega)) \tag{2}$$

 ISBN: 978-84-617-8694-7

$$y_{2n+2}(\omega) = T(x_{2n+1}(\omega)) = f(x_{2n+2}(\omega)) \tag{3}$$

If $d(f(x_{2n}(\omega)), S(x_{2n}(\omega))) + d(g(x_{2n+1}(\omega)), T(x_{2n+1}(\omega))) \neq 0$. Then from (1), (2), and (3), we get

$$\begin{aligned}
&F(d(S(x_{2n}(\omega)), T(x_{2n+1}(\omega))), d(f(x_{2n}(\omega)), g(x_{2n+1}(\omega))), d(f(x_{2n}(\omega)), S(x_{2n}(\omega))),\\
&d(g(x_{2n+1}(\omega)), T(x_{2n+1}(\omega))), d(f(x_{2n}(\omega)), T(x_{2n+1}(\omega))), d(S(x_{2n}(\omega)), g(x_{2n+1}(\omega))))\\
&= F(d(y_{2n+1}(\omega), y_{2n+2}(\omega)), d(y_{2n}(\omega), y_{2n+1}(\omega)), d(y_{2n}(\omega), y_{2n+1}(\omega)),\\
&d(y_{2n+1}(\omega), y_{2n+2}(\omega)), d(y_{2n}(\omega), y_{2n+2}(\omega)), d(y_{2n+1}(\omega), y_{2n+1}(\omega)))\\
&\leq 0.
\end{aligned}$$

By (F_a), we get

$$d(y_{2n+1}(\omega), y_{2n+2}(\omega)) \leq h d(y_{2n}(\omega), y_{2n+1}(\omega))$$

Similarly, if

$$d(f(x_{2n+2}(\omega)), S(x_{2n+2}(\omega))) + d(g(x_{2n+1}(\omega)), T(x_{2n+1}(\omega))) \neq 0.$$

we obtain

$$d(y_{2n}(\omega), y_{2n+1}(\omega)) \leq h d(y_{2n-1}(\omega), y_{2n}(\omega)),$$

hence

$$d(y_{2n+1}(\omega), y_{2n+2}(\omega)) \leq h^2 d(y_{2n-1}(\omega), y_{2n}(\omega)).$$

On continuing this process, we have

$$d(y_{2n+1}(\omega), y_{2n+2}(\omega)) \leq h^{2n} d(y_0(\omega), y_1(\omega)).$$

Also, for $n > m$, we get

$$\begin{aligned}
d(y_n(\omega), y_m(\omega)) &\leq d(y_n(\omega), y_{n-1}(\omega)) + d(y_{n-1}(\omega), y_{n-2}(\omega)) + \cdots + d(y_{m+1}(\omega), y_m(\omega))\\
&\leq (h^{n-1} + h^{n-2} + \cdots + h^m) d(y_0(\omega), y_1(\omega))\\
&\leq (\frac{h^m}{1-h}) d(y_0(\omega), y_1(\omega)).
\end{aligned}$$

Let $0 \ll \varepsilon$ is given. Choose a natural number N such that $(\frac{h^m}{1-h}) d(y_0(\omega), y_1(\omega)) \ll \varepsilon$ for every $m \geq N$, hence

$$d(y_n(\omega), y_m(\omega)) \leq (\frac{h^m}{1-h}) d(y_0(\omega), y_1(\omega)) \ll \varepsilon,$$

this implies that $\{y_n(\omega)\}$ is Cauchy sequence in $\Omega \times X$. Since (X, d) is complete, then there exists $z(\omega) \in \Omega \times X$ such that $y_n(\omega) \to z(\omega)$ as $n \to \infty$. Then from (2) and (3), we get

$$\lim_{n\to\infty} S(x_{2n}(\omega)) = \lim_{n\to\infty} g(x_{2n+1}(\omega)) = z(\omega)$$

 ISBN: 978-84-617-8694-7

and

$$\lim_{n\to\infty} T(x_{2n+1}(\omega)) = \lim_{n\to\infty} f(x_{2n+2}(\omega)) = z(\omega).$$

Therefore

$$\lim_{n\to\infty} S(x_{2n}(\omega)) = \lim_{n\to\infty} g(x_{2n+1}(\omega)) = \lim_{n\to\infty} T(x_{2n+1}(\omega)) = \lim_{n\to\infty} f(x_{2n+2}(\omega)) = z(\omega) \quad (4)$$

Since $T(\omega, X) \subseteq f(\omega, X)$, then there exists $v(\omega) \in \Omega \times X$ such that

$$z(\omega) = f(v(\omega)). \quad (5)$$

From (1), we obtain

$$F(d(S(x_{2n}(\omega)), T(x_{2n+1}(\omega))), d(f(x_{2n}(\omega)), g(x_{2n+1}(\omega))), d(f(x_{2n}(\omega)), S(x_{2n}(\omega))),$$
$$d(g(x_{2n+1}(\omega)), T(x_{2n+1}(\omega))), d(f(x_{2n}(\omega)), T(x_{2n+1}(\omega))), d(S(x_{2n}(\omega)), g(x_{2n+1}(\omega)))) \leq 0,$$

this implies that $d(z(\omega), T(v(\omega))) \leq 0$, thus $-d(z(\omega), T(v(\omega))) \in p$. But $d(z(\omega), T(v(\omega))) \in p$, therefore by Definition 2.1 (c_3), we have $d(z(\omega), T(v(\omega))) = 0$ and so $d(z(\omega) = T(v(\omega)))$. Form (5) we get

$$z(\omega) = f(v(\omega)) = T(v(\omega)).$$

Hence $v(\omega)$ is a random coincidence point of T and f. Since the pair T and f are random weakly compatible, i.e. $T(f(v(\omega))) = f(T(v(\omega)))$ this implies that

$$T(z(\omega)) = f(z(\omega)).$$

Now we show that $z(\omega)$ is a random fixed point of S, we have from (1) that

$$\begin{aligned} &F(d(S(u(\omega)), T(v(\omega))), d(f(u(\omega)), g(v(\omega))), d(f(u(\omega)), S(u(\omega))), \\ &d(g(v(\omega)), T(v(\omega))), d(f(u(\omega)), T(v(\omega))), d(S(u(\omega)), g(v(\omega)))) \\ &= F(d(S(u(\omega)), (z(\omega))), 0, d(z(\omega), S(u(\omega))), 0, 0, d(S(u(\omega)), z(\omega))) \\ &\leq 0. \end{aligned}$$

By (F_b) we get $f(z(\omega)) = S(z(\omega))$ and $g(z(\omega)) = T(z(\omega))$. Since $d(f(z(\omega)), S(z(\omega))) + d(g(v(\omega)), T(v(\omega))) = 0$, it follows that $d(S(z(\omega)), T(v(\omega))) = 0$ i.e.

$$z(\omega) = S(z(\omega)) = f(z(\omega)). \quad (6)$$

By a similar way and using (6), we can prove that for all $\omega \in \Omega$,

$$z(\omega) = T(z(\omega)) = g(z(\omega)). \quad (7)$$

 ISBN: 978-84-617-8694-7

The equations (6) and (7) show that $z(\omega)$ is common random fixed point of T, S, f, g. For uniqueness. Let $z(\omega) \neq q(\omega)$ be another common random fixed point of four mappings, then from (1), one can write

$$\begin{aligned}
&F(d(S(u(\omega)), T(q(\omega))), d(f(u(\omega)), g(q(\omega))), d(f(u(\omega)), S(u(\omega))),\\
&d(g(q(\omega)), T(q(\omega))), d(f(u(\omega)), T(q(\omega))), d(S(u(\omega)), g(q(\omega))))\\
&= F(d(q(\omega), (z(\omega))), 0, d(z(\omega), q(\omega)), 0, 0, d(q(\omega), z(\omega)))\\
&\leq 0.
\end{aligned}$$

a contradiction. Hence $z(\omega) = q(\omega)$ and so $z(\omega)$ is a unique common random fixed point of T, S, f, g. □

If we take, $f = g$ in above theorem we obtain the following corollary.

Corollary 3.2. *Let (X, d) be a complete cone random metric space with respect to a cone $\mathcal{P}$ and let $\mathcal{M}$ be nonempty separable closed subset of X. Assume that S, T and f be three continuous random operators defined on $\mathcal{M}$ such that $\omega \in \Omega$, $S(\omega, \cdot), T(\omega, \cdot), f(\omega, \cdot) : \Omega \times \mathcal{M} \to \mathcal{M}$ satisfying the following conditions:*

(i) $S(\omega, X) \subseteq f(\omega, X)$ and $T(\omega, X) \subseteq f(\omega, X)$,

(ii) the pairs $\{S, f\}$ and $\{T, f\}$ are random weakly compatible mappings,

(iii)

$$\begin{aligned}
&F(d(S(x(\omega)), T(y(\omega))), d(f(x(\omega)), f(y(\omega))), d(f(x(\omega)), S(x(\omega))),\\
&d(f(y(\omega)), T(y(\omega))), d(f(x(\omega)), T(y(\omega))), d(S(x(\omega)), f(y(\omega)))) \leq 0
\end{aligned}$$

for all $x(\omega), y(\omega) \in \Omega \times X$ and $F \in F_6$ satifies (F_1) if $d(f(x(\omega)), S(x(\omega))) + d(f(y(\omega)), T(y(\omega))) \neq 0$, or $d(S(x(\omega)), T(x(\omega))) = 0$ if $d(f(x(\omega)), S(x(\omega))) + d(f(y(\omega)), T(y(\omega))) = 0$.
Then the three random mapping have unique common random fixed point in X.

If we take, $f = g$ and $S = T$ in above theorem we obtain the following corollary.

Corollary 3.3. *Let (X, d) be a complete cone random metric space with respect to a cone $\mathcal{P}$ and let $\mathcal{M}$ be nonempty separable closed subset of X. Assume that S and f be two continuous random operators defined on $\mathcal{M}$ such that $\omega \in \Omega$, $S(\omega, \cdot), f(\omega, \cdot) : \Omega \times \mathcal{M} \to \mathcal{M}$ satisfying the following conditions:*

(i) $S(\omega, X) \subseteq f(\omega, X)$,

(ii) the pairs $\{S, f\}$ is a random weakly compatible mappings,

 ISBN: 978-84-617-8694-7

(iii)

$$F(d(S(x(\omega)),S(y(\omega))),d(f(x(\omega)),f(y(\omega))),d(f(x(\omega)),S(x(\omega))),$$
$$d(f(y(\omega)),S(y(\omega))),d(f(x(\omega)),S(y(\omega))),d(S(x(\omega)),f(y(\omega)))) \leq 0$$

for all $x(\omega), y(\omega) \in \Omega \times X$ *and* $F \in F_6$ *satifies* (F_1) *if* $d(f(x(\omega)),S(x(\omega)))+d(f(y(\omega)),S(y(\omega))) \neq 0$, *or* $d(S(x(\omega)),S(x(\omega))) = 0$ *if* $d(f(x(\omega)),S(x(\omega))) + d(f(y(\omega)),S(y(\omega))) = 0$.
Then the two random mapping have unique common random fixed point in X.

4 Acknowledgments

The first author thanks for the support of Petchra Pra Jom Klao Doctoral Scholarship for Ph.d. student of King Mongkut's University of Technology Thonburi (KMUTT).

References

[1] A. Špacek, *Zufllige Gleichungen. Czechoslov.*, Czechoslov. Math. J. **5(80)** (1955) 462–466.

[2] O. Hanš, *Reduzierende zufällige transformationen.*, Czechoslov. Math. J.. **7(82)** (1957) 154–158.

[3] A. T. Bharucha-Reid, *Fixed point theorems in probabilistic analysis.*, Bull. Amer. Math. Soc. **82** (19764) 641–657.

[4] S. Itoh, *A random fixed point theorem for a multivalued contraction mapping.*, Pacific Journal of Mathematics., **68(1)** (1977) 85–90.

[5] I. Beg, and N. Shahzad, *An application of a random fixed point theorem to random best approximation.*, Archiv der Mathematik. **74(4)** (2000) 298–301.

[6] P. Kumam, and P. Plubtieng, *Random coincidence and random common fixed points of nonlinear multivalued random operators.*, Thai J. Math. **5(3)** (2007) 155–163.

[7] G. S. Saluja, and M. P. Tripathi, *Random common fixed point theorems for a pair of multi-valued and single-valued nonexpansive random operators in a separable Banach space.*, Indian J. Math. **51(1)** (2009) 101–115.

[8] L. G. Huang, and X. Zhang, *Cone metric spaces and fixed point theorems of contractive mappings.*, J. Math. Anal. Appl. **332** (2007) 1468–1476.

 ISBN: 978-84-617-8694-7

[9] Sh. Rezapour, and R. Hamlbarani, *Some notes on the paper One metric spaces and fixed point theorems of contractive mappings.*, Math. Anal. Appl.. **345** (2008) 719–724.

[10] S. Mehta, A. D. Singh, and V. B. Dhagat, *Fixed point theorems for weak contraction in cone random metric spaces.*, Bull. Math. Soc. **103** (2011) 303–310.

[11] S. Mehta, and A. D. Singh, *On common random fixed point results in cone random metric spaces.*, South Asian J. Math. **2(3)** (2012) 248–254.

[12] R. A. Rashwan, and H. A. Hammad, *A common random fixed point theorem for weakly compatible mappings in cone random metric spaces.*, Universal Journal of Computational Mathematics. **4(4)** (2016) 67–74.

[13] R. A. Rashwan, and H. A. Hammad, *Random fixed point theorem for weakly compatible mappings under implicit relation in cone random metric spaces.*, Universal Journal of Computational Mathematics. **5(1)** (2017) 8–16.

[14] P. P. Zabrejko, *K-metric and K-normed linear spaces survey.*, Collectanea Math. **48** (199) 825–859.

[15] A. Aliouche, *Common fixed point theorems via implicit relation.*, Miskolc Mathematical Notes. **11(1)** (2010) 3–12.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

Transmutation operators: construction and applications

Vladislav V. Kravchenko[1], Sergii M. Torba[1] and Kira V. Khmelnytskaya[2]

[1] *Department of Mathematics, CINVESTAV del IPN, Unidad Querétaro, México*

[2] *Faculty of Engineering, Autonomous University of Querétaro, México*

emails: vkravchenko@math.cinvestav.edu.mx, storba@math.cinvestav.edu.mx, khmel@uaq.edu.mx

Abstract

Recent results on the construction and applications of the transmutation (transformation) operators are discussed. Three new representations for solutions of the one-dimensional Schrödinger equation are considered. Due to the fact that they are obtained with the aid of the transmutation operator all the representations possess an important for practice feature. The accuracy of the approximate solution is independent of the real part of the spectral parameter. This makes the representations especially useful in problems requiring computation of large sets of eigendata with a nondeteriorating accuracy.

Applications of the exact representations for the transmutation operators to partial differential equations are discussed as well. In particular, it is shown how the methods based on complete families of solutions can be extended onto equations with variable coefficients.

Key words: Sturm-Liouville equation, Transmutation operator, Neumann series of Bessel functions, Spectral problem, Complete family of solutions, Method of fundamental solutions

MSC 2000: AMS codes (optional)

1 Transmutation operators

Transmutation operators also called transformation operators are a widely used tool in the theory of linear differential equations (see, e.g., [2], [4], [5] [18], [19], [20] and many other publications). In particular, let $q \in C[-b, b]$ be a complex valued function. Consider the Sturm-Liouville equation

$$Ay := y'' - q(x)y = -\omega^2 y. \tag{1}$$

It is well known (see, e.g., [19]) that there exists a Volterra integral operator T called the transmutation (or transformation) operator defined on $C[-b,b]$ by the formula

$$Tu(x) = u(x) + \int_{-x}^{x} K(x,t)u(t)dt$$

such that for any $u \in C^2[-b,b]$ the following equality is valid

$$ATu = Tu''$$

and hence any solution of (1) can be written as $y = T[u]$ where $u(x) = c_1 \cos \omega x + c_2 \sin \omega x$ with c_1 and c_2 being arbitrary constants.

The transmutation kernel K is a solution of a certain Goursat problem for the hyperbolic equation

$$\left(\frac{\partial^2}{\partial x^2} - q(x)\right) K(x,t) = \frac{\partial^2}{\partial t^2} K(x,t).$$

2 Construction of the transmutation kernel and new representations for solutions of the Sturm-Liouville equation

In spite of fundamental importance of the transmutation kernel K in the theory of linear differential equations, besides the method of successive approximations derived directly from the Goursat problem (see, e.g., [5]) very few attempts of its practical construction have been reported. In this relation we mention the paper [3] where analytic approximation formulas for the integral kernel were obtained and the recent publications [14], [15] where another procedure of analytical approximation was proposed.

To the difference of those previous results, in the recent paper [13] an exact representation for K in the form of a Fourier-Legendre series with explicit formulas for the coefficients was obtained. Suppose that $q \in W_2^{-1}[-b,b]$ (that is, q can be a piecewise continuous function, may have a singularity, e.g., $q(x) \sim c/x$, etc.). In this case $K(x,t)$ is an L_2-function. Under these conditions the following theorem was proved in [13].

Theorem 1 *The kernel K has the form*

$$K(x,t) = \sum_{n=0}^{\infty} \frac{\beta_n(x)}{x} P_n\left(\frac{t}{x}\right) \tag{2}$$

where for every $x \in [-b,b]$ the series converges with respect to t in the L_2-norm (if $q \in C[-b,b]$ the series converges uniformly),

$$\beta_n(x) = \frac{2n+1}{2}\left(\sum_{k=0}^{n} \frac{l_{k,n}\varphi_k(x)}{x^k} - 1\right),$$

 ISBN: 978-84-617-8694-7

with $l_{k,n}$ being the coefficient at x^k of the Legendre polynomial P_n, and φ_k being the so-called formal powers constructed as follows (see [8], [12]).

Definition 2 (Formal powers φ_k) *Let f be a solution of*

$$f'' - q(x)f = 0, \quad x \in [-b, b], \tag{3}$$

$$f(0) = 1, \quad f'(0) = 0.$$

Then $\{\varphi_k\}_{k=0}^{\infty}$ are defined by the equalities

$$\varphi_k = \begin{cases} fX^{(k)}, & k \text{ odd}, \\ f\widetilde{X}^{(k)}, & k \text{ even}, \end{cases}$$

where

$$X^{(0)} \equiv 1, \qquad X^{(n)}(x) = n \int_0^x X^{(n-1)}(s) \left(f^2(s)\right)^{(-1)^n} \mathrm{d}s,$$

and

$$\widetilde{X}^{(0)} \equiv 1, \qquad \widetilde{X}^{(n)}(x) = n \int_0^x \widetilde{X}^{(n-1)}(s) \left(f^2(s)\right)^{(-1)^{n-1}} \mathrm{d}s.$$

It is worth mentioning that φ_k are easily computable (at least numerically) in practice (see, e.g., [13] for additional details).

A representation for the kernel K leads to a representation for the solution of (1). Let $u(\omega, x)$ denote the solution of (1) satisfying the initial conditions

$$u(\omega, 0) = 1, \qquad u'(\omega, 0) = i\omega.$$

Then we have

$$u(\omega, x) = e^{i\omega x} + \int_{-x}^{x} K(x, t) e^{i\omega t} dt.$$

Substitution of (2) into the last integral gives us the equality [13]

$$u(\omega, x) = e^{i\omega x} + \sum_{n=0}^{\infty} \beta_n(x) \int_{-1}^{1} P_n(y)\, e^{i\omega x y} dy = e^{i\omega x} + \sum_{n=0}^{\infty} i^n \beta_n(x) j_n(\omega x)$$

where $j_n(z) = \sqrt{\frac{\pi}{2z}} J_{n+1/2}(z)$ are spherical Bessel functions. The series converges uniformly with respect to x.

Moreover, take $\omega \in \mathbb{R}$. Consider

$$K_N(x, t) = \sum_{n=0}^{N} \frac{\beta_n(x)}{x} P_n\left(\frac{t}{x}\right)$$

 ISBN: 978-84-617-8694-7

and

$$u_N(\omega,x)=e^{i\omega x}+\sum_{n=0}^{N}i^n\beta_n(x)j_n(\omega x),\quad x>0.$$

We have [13]

$$\begin{aligned}|u(\omega,x)-u_N(\omega,x)|&=\left|\int_{-x}^{x}\left(K(x,t)-K_N(x,t)\right)e^{i\omega t}dt\right|\\&\le\|K(x,\cdot)-K_N(x,\cdot)\|_{L_2(-x,x)}\left\|e^{i\omega t}\right\|_{L_2(-x,x)}\\&=\varepsilon_N(x)\sqrt{2x}\end{aligned}$$

—independent of ω. More generally, for any $\omega\in\mathbb{C}$, $\omega\neq 0$ belonging to the strip $|\mathrm{Im}\,\omega|\le C$, $C\ge 0$,

$$|u(\omega,x)-u_N(\omega,x)|\le\varepsilon_N(x)\frac{\sinh(Cx)}{C}.$$

This ω-independence of the approximation accuracy was shown in [13] to give a very fast and efficient method for computing large sets of eigendata with a nondeteriorating accuracy. In [17] it was generalized onto perturbed Bessel equations, and in [16] onto Sturm-Liouville equations.

Another representation for the kernel K and as a corollary for the solutions of (1) was obtained in [9]. Consider the following extension of the transmutation kernel K,

$$\widetilde{K}(x,t)=\begin{cases}K(x,t), & -x\le t\le x,\\ 0, & -\infty<t<-x.\end{cases}$$

Then

$$u(\omega,x)=e^{i\omega x}+\int_{-\infty}^{x}\widetilde{K}(x,y)e^{i\omega y}dy=e^{i\omega x}\left(1+\int_0^{\infty}\widetilde{K}(x,x-t)e^{-i\omega t}dt\right).$$

Consider

$$\widetilde{K}(x,x-t)=\mathbf{k}(x,t)e^{-t}.$$

The function $\mathbf{k}(x,\cdot)$ then belongs to the space $L_2\left(0,\infty;e^{-t}\right)$ equipped with the scalar product $\langle u,v\rangle:=\int_0^{\infty}u(t)\overline{v}(t)e^{-t}dt$. Thus, $\mathbf{k}(x,\cdot)$ admits a Fourier-Laguerre expansion convergent in the corresponding norm,

$$\mathbf{k}(x,t)=\sum_{n=0}^{\infty}a_n(x)L_n(t).$$

The kernel has the form [9]

$$\widetilde{K}(x,y)=\sum_{n=0}^{\infty}a_n(x)L_n(x-y)e^{-(x-y)},$$

 ISBN: 978-84-617-8694-7

with the coefficients a_n defined by

$$a_n(x) = \sum_{j=0}^{n} (-1)^j \left(\varphi_j(x) - x^j\right) \sum_{k=j}^{n} (-1)^k \frac{n!}{(n-k)!k!(k-j)!j!} x^{k-j}. \tag{4}$$

The solution $u(\omega, x)$ has the form [9]

$$u(\omega, x) = e^{i\omega x} \left(1 + \sum_{n=0}^{\infty} a_n(x) \frac{(i\omega)^n}{(1+i\omega)^{n+1}}\right). \tag{5}$$

The following estimate is valid for any $\omega \in \mathbb{R}$,

$$|u(\omega, x) - u_N(\omega, x)| \le \varepsilon_N(x), \tag{6}$$

where

$$u_N(\omega, x) := e^{i\omega x} \left(1 + \sum_{n=0}^{N} a_n(x) \frac{(i\omega)^n}{(1+i\omega)^{n+1}}\right),$$

and $\varepsilon_N(x)$ is a nonnegative function independent of ω and such that $\varepsilon_N(x) \to 0$ for all $x \in [-b, b]$ when $N \to \infty$. More generally,

$$|u(\omega, x) - u_N(\omega, x)| \le \frac{\varepsilon_N(x) e^{-\operatorname{Im}\omega x}}{\sqrt{1 - 2\operatorname{Im}\omega}}, \qquad \text{when } \operatorname{Im}\omega < 1/2.$$

Consideration of another extension of the transmutation kernel defined by

$$\widetilde{K}(x, y) := \begin{cases} K(x, y) & \text{when } x \in [-b, b] \text{ and } y \in [-x, x] \\ 0 & \text{otherwise.} \end{cases}$$

leads to the following series expansion

$$\widetilde{K}(x, y) = \sum_{n=0}^{\infty} c_n(x) H_n(y) e^{-y^2}$$

where H_n stands for an Hermite polynomial of order n and the coefficients c_n are to be found. Note that

$$\int_{-\infty}^{\infty} \widetilde{K}(x, y) H_n(y) dy = \sqrt{\pi} n! 2^n c_n(x).$$

Hence

$$c_n(x) = \frac{1}{\sqrt{\pi} n! 2^n} \int_{-x}^{x} K(x, y) H_n(y) dy = \frac{1}{\sqrt{\pi} n! 2^n} \sum_{k=0}^{n} h_{k,n} \left(\varphi_k(x) - x^k\right) \tag{7}$$

 ISBN: 978-84-617-8694-7

where $h_{k,n}$ denotes the coefficient of x^k from the Hermite polynomial $H_n(x)$.

This leads to another representation for the solution of (1),

$$\begin{aligned} u(\omega,x) &= e^{i\omega x} + \int_{-\infty}^{\infty} \widetilde{K}(x,y) e^{i\omega y} dy \\ &= e^{i\omega x} + \sum_{n=0}^{\infty} c_n(x) \int_{-\infty}^{\infty} H_n(y) e^{i\omega y} e^{-y^2} dy \\ &= e^{i\omega x} + \sqrt{\pi} e^{-\frac{\omega^2}{4}} \sum_{n=0}^{\infty} c_n(x) \, (i\omega)^n \, . \end{aligned}$$

Consider the partial sum

$$u_N(\omega,x) = e^{i\omega x} + \sqrt{\pi} e^{-\frac{\omega^2}{4}} \sum_{n=0}^{N} c_n(x) \, (i\omega)^n \, .$$

Then it is easy to see that

$$|u(\omega,x) - u_N(\omega,x)| \le \pi^{\frac{1}{4}} e^{\frac{(\operatorname{Im}\omega)^2}{2}} \varepsilon_N(x)$$

which means that the truncation error is uniformly bounded in any strip $|\operatorname{Im}\omega| \le C$.

3 Applications to PDEs

Exact representations of the transmutation kernel lead to numerous applications for partial differential equations admitting certain symmetry. In particular, let us consider the possibility to obtain complete systems of solutions. For example, application of the transmutation operator T to a complete system of harmonic functions leads to a complete system of solutions of the equation

$$(\Delta - q(x)) \, u(x,y) = 0. \tag{8}$$

Indeed,

$$(\Delta - q(x)) \, T = T \Delta$$

whenever the domain of interest is such that the integration in T is well defined.

Example Harmonic polynomials ($\operatorname{Re} z^n$ and $\operatorname{Im} z^n$) can be written in the form

$$\begin{aligned} p_0(x,y) &= 1, \\ p_{2m+1}(x,y) = \operatorname{Re} z^{m+1} &= \sum_{\text{even } k=0}^{m} (-1)^{\frac{k}{2}} \binom{m+1}{k} x^{m+1-k} y^k, \qquad m \ge 0, \\ p_{2m}(x,y) = \operatorname{Re}(iz^m) &= \sum_{\text{odd } k=1}^{m} (-1)^{\frac{k+1}{2}} \binom{m}{k} x^{m-k} y^k, \qquad m \ge 1. \end{aligned}$$

 ISBN: 978-84-617-8694-7

Since

$$T : x^k \mapsto \varphi_k(x),$$

the following functions are the images of p_m under the action of T and represent a complete system of solutions of (8)

$$\begin{aligned} u_0(x,y) &= f(x), \\ u_{2m+1}(x,y) &= \sum_{\text{even } k=0}^{m+1} (-1)^{\frac{k}{2}} \binom{m+1}{k} \varphi_{m+1-k}(x) y^k, \qquad m \geq 0, \\ u_{2m}(x,y) &= \sum_{\text{odd } k=1}^{m} (-1)^{\frac{k+1}{2}} \binom{m}{k} \varphi_{m-k}(x) y^k, \qquad m \geq 1. \end{aligned}$$

Other complete systems of solutions can be obtained.

Example The method of fundamental solutions (discrete sources) (see, e.g., [1], [6]) can be extended onto equations with variable coefficients. Consider the fundamental solution for the Laplace operator on the plane

$$\log|x + iy - (\eta + i\xi)| = \log|x - Z|\,.$$

Application of T leads to the following integrals

$$T\left[\log|x - Z|\right] = \log|x - Z| + \sum_{n=0}^{\infty} \frac{\beta_n(x)}{x} \int_{-x}^{x} P_n\left(\frac{t}{x}\right) \log|t - Z|\; dt.$$

Their calculation gives us the image of the fundamental solution in the form

$$\begin{aligned} T\left[\log|x - Z|\right] = \log|x - Z| + \beta_0(x)\,\mathrm{Re}\left(\log\left((Z + x)\,(Z - x)\right) + 2Q_1\left(\frac{Z}{x}\right)\right) \\ + 2\sum_{n=1}^{\infty} \frac{\beta_n(x)}{2n+1}\,\mathrm{Re}\left(Q_{n+1}\left(\frac{Z}{x}\right) - Q_{n-1}\left(\frac{Z}{x}\right)\right) \end{aligned}$$

where Q_n are Legendre functions of the second kind.

Similar considerations can be applied to systems arising in hypercomplex analysis (see, e.g., [7]) and may lead to extensions of well known methods based on monogenic polynomials or other complete systems of solutions onto systems with variable coefficients [10], [11] important, e.g., in electromagnetic theory and quantum physics.

Acknowledgements

This work has been supported by CONACYT, Mexico via the projects 166141 and 222478.

References

[1] M. A. Alexidze, *Fundamental Functions in Approximate Solutions of Boundary Value Problems* (in Russian), Nauka, Moscow, 1991.

[2] H. Begehr and R. Gilbert, *Transformations, transmutations and kernel functions, vol. 1–2*, Longman Scientific & Technical, Harlow, 1992.

[3] A. Boumenir, *The approximation of the transmutation kernel*, J. Math. Phys. **47** (2006), 013505.

[4] R. W. Carroll, *Transmutation theory and applications, Mathematics Studies, Vol. 117*, North-Holland, 1985.

[5] D. Colton, *Solution of boundary value problems by the method of integral operators*, Pitman, London, 1976.

[6] A. Doicu, Yu. Eremin and Th. Wriedt, *Acoustic and Electromagnetic Scattering Analysis,* Acad. Press, London, 2000.

[7] K. Gürlebeck and W. Sprössig, *Quaternionic and Clifford Calculus for Physicists and Engineers*, John Wiley & Sons, Chichester, 1997.

[8] V. V. Kravchenko, *A representation for solutions of the Sturm-Liouville equation*, Complex Var. Elliptic Equ. **53** (2008) 775–789.

[9] V. V. Kravchenko, *Construction of a transmutation for the one-dimensional Schrödinger operator and a representation for solutions*, submitted, available from arXiv: 1612.09577.

[10] V. V. Kravchenko, *Applied quaternionic analysis*, Heldermann Verlag, Lemgo, 2003.

[11] V. V. Kravchenko, *Applied pseudoanalytic function theory, Series: Frontiers in Mathematics*, Birkhäuser, Basel, 2009.

[12] V. V. Kravchenko and R. M. Porter, *Spectral parameter power series for Sturm-Liouville problems*, Math. Methods Appl. Sci. **33** (2010) 459–468.

[13] V. V. Kravchenko, L. J. Navarro and S. M. Torba, *Representation of solutions to the one-dimensional Schrödinger equation in terms of Neumann series of Bessel functions*, submitted, available from arXiv:1508.02738.

[14] V. V. Kravchenko and S. M. Torba, *Construction of transmutation operators and hyperbolic pseudoanalytic functions*, Complex Anal. Oper. Theory **9** (2015) 389–429.

 ISBN: 978-84-617-8694-7

[15] V. V. Kravchenko and S. M. Torba, *Analytic approximation of transmutation operators and applications to highly accurate solution of spectral problems*, J. Comput. Appl. Math. **275** (2015) 1–26.

[16] V. V. Kravchenko and S. M. Torba, *A Neumann series of Bessel functions representation for solutions of Sturm-Liouville equations*, submitted, available from: arXiv:1612.08803.

[17] V. V. Kravchenko, S. M. Torba and R. Castillo-Perez, *A Neumann series of Bessel functions representation for solutions of perturbed Bessel equations*, Appl. Analysis **2017**, published online, doi:10.1080/00036811.2017.1284313.

[18] B. M. Levitan, *Inverse Sturm-Liouville problems*, VSP, Zeist, 1987.

[19] V. A. Marchenko, *Sturm-Liouville operators and applications*, Birkhäuser, Basel, 1986.

[20] S. M. Sitnik, *Transmutations and applications: a survey*, arXiv:1012.3741v1, originally published in the book: *Advances in Modern Analysis and Mathematical Modeling*, Editors: Yu. F. Korobeinik and A. G. Kusraev, Vladikavkaz Scientific Center of the Russian Academy of Sciences and Republic of North Ossetia–Alania, Vladikavkaz, 2008, 226–293.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

Fixed Point Approach to Solution Existence of Differential Equations*

Wiyada Kumam[1], Parin Chaipunya[2], Poom Kumam[3] and Phatiphat Thounthong[4]

[1] *Program in Applied Statistics, Department of Mathematics and Computer Science, Faculty of Science and Technology, Rajamangala University of Technology Thanyaburi (RMUTT), Rungsit-Nakorn Nayok Rd., Klong 6, Thanyaburi, Pathumthani 12110, Thailand*

[2] *Department of Teacher Training in Electrical Engineering, Faculty of Technical Education, King Mongkut's University of Technology North Bangkok (KMUTNB), Wongsawang, Bangsue, Bangkok 10800, Thailand*

[3] *KMUTT-Fixed Point Theory and ApplicationsResearch Group, Theoretical and Computational Science Center (TaCS), Science Laboratory Building, Faculty of Science, King Mongkut's University of Technology Thonburi (KMUTT), 126 Pracha Uthit Rd., Bang Mod, Thung Khru, Bangkok 10140, Thailand*

[4] *KMUTTFixed PointResearch Laboratory,Department of Mathematics, Room SCL 802 Fixed Point Laboratory, Science Laboratory Building, Faculty of Science, King Mongkut's University of Technology Thonburi (KMUTT), 126 Pracha Uthit Rd., Bang Mod, Thung Khru, Bangkok 10140, Thailand*

emails: wiyada.kum@rmutt.ac.th, parin.cha@mail.kmutt.ac.th, poom.kum@kmutt.ac.th, phtt@kmutnb.ac.th

Abstract

*This project was partially supported by the Theoretical and Computational Science (TaCS) Center under Computational and Applied Science for Smart Innovation Cluster (CLASSIC), Faculty of Science, KMUTT. Furthermore, this research work was financially supported by King Mongkut's University of Technology North Bangkok. Contract No. KMUTNB-60-ART-084.

In this my talk, we consider mainly on the question about existence and uniqueness of solution of fixed point equation for several nonlinear mappings and several kinds of domain spaces. We then give a brief examples of practical applications to illustrate the usability of our theoretical results.

Key words: Fixed Point; Contraction; Generalized Contraction; Differential Equation; Partial Differential Equation.
MSC 2000: 47*H*10*,* 54*H*25*.*

1 Introduction

The question about solvability of particular forms of ordinary or partial differential equations is naturally the first to ask. But the answer is unfortunately not easily acquired. By transforming a particular differential equation into a fixed point equation, we can reduce the difficulty of the question and we can apply various fixed point theorems to the equation and obtain the desired affirmative answer.

In discrete time domain, major considerations turn to the difference equations and generating functions. While in the latter one, under which we shall be considering mainly for this chapter, the system is usually represented by differential equations. It might be more influential to talk about the inclusion problems if a set-valued system is to be analyzed.

The very first and fundamental dynamical system is known now a days under the term Cauchy problem. It is represented with the following C^1 initial-valued problem:

$$\begin{cases} u'(t) = f(t, u(t)), \\ u(0) = u_0. \end{cases}$$

In this case, we assume that $f : [0,T] \times R \to R$ is continuous and $u \in C^1([0,T])$. From simple calculus, we may see that this system is equivalent to the following integral equation:

$$u(t) = u_0 + \int_{[0,t]} f(s, u(s))ds. \tag{1}$$

This is where Banach got the idea to solve the problem. He proposed his famous fixed point theorem known today as the contraction principle in 1922 [4], mainly to solve this Cauchy problem effectively. Recall that the contraction principle states that if X is a complete metric space and $T : X \to X$ is Lipschitz continuous with constant $0 < L < 1$, then T has a unique fixed point.

Let us consider a map $\Lambda : C^1([0,T]) \to C^1([0,T])$ given by

$$\Lambda(u)(t) := u_0 + \int_{[0,t]} f(s, u(s))ds, \quad \forall u \in C^1([0,T]),\ \forall t \in [0,T].$$

One can notice that $u \in C^1([0,T])$ solves Cauchy problem (1) if and only if it is a fixed point of Λ. With this approach, by considering $C^1([0,T])$ with the supremum norm $\|\cdot\|_\infty$, we end up with the local solvability of the Cauchy problem.

 ISBN: 978-84-617-8694-7

2 Example of Fixed Point Theory

Let X be a nonempty set and $\lambda \in (0, \infty)$. Due to the disparity of the arguments, a function $\omega : (0, \infty) \times X \times X \to [0, \infty]$ will be written as $\omega_\lambda(x, y) = \omega(\lambda, x, y)$ for all $\lambda > 0$ and $x, y \in X$. [8] Let X be a nonempty set. A function $\omega : (0, \infty) \times X \times X \to [0, \infty]$ is called a *metric modular* on X is satisfies the following condition: for all $x, y, z \in X$,

(i) $\omega_\lambda(x, y) = 0$ for all $\lambda > 0$ if and only if $x = y$;

(ii) $\omega_\lambda(x, y) = \omega_\lambda(y, x)$ for all $\lambda > 0$;

(iii) $\omega_{\lambda+\mu}(x, y) \leq \omega_\lambda(x, z) + \omega_\mu(z, y)$ for all $\lambda, \mu > 0$.

If, instead of (i), we have only the following condition:

(i′) $\omega_\lambda(x, x) = 0$ for all $\lambda > 0$, then ω called a *(metric) pseudomodular* on X.

If $\omega_\lambda(x, y) = \omega(x, y)$ does not depend on $\lambda > 0$ and has only finite values, then the axioms (i)-(iii) mean that ω is a metric on X if (i) is replaced by (i′).

We next give another route of investigation of fixed point inclusion in modular metric spaces. This time, we shall apply more on analytical assumptions. Briefly said, we shall use the contractivity assumptions (see [16]).

2.1 Fixed point theorems in modular metric spaces

In this section, we prove new existence theorems of fixed points for contraction mappings in modular metric spaces. Let ω be a metric modular on X, X_ω be a modular metric space induced by ω and $T : X_\omega \to X_\omega$ be an arbitrary mapping. A mapping T is called a *contraction* if, for all $x, y \in X_\omega$ and $\lambda > 0$, there exists $0 \leq k < 1$ such that

$$\omega_\lambda(Tx, Ty) \leq k\omega_\lambda(x, y). \tag{2}$$

Let X_ω be a complete modular metric space and $T : X_\omega \to X_\omega$ be a contraction mapping. Assume that there exists $x_0 \in X$ such that $\omega_\lambda(x_0, Tx_0) < \infty$ for all $\lambda > 0$. Then T has a fixed point in $x_* \in X_\omega$ and the sequence $\{T^n x_0\}$ converges to x_*. Moreover, if, $z \in F(X_\omega)$, where $F(X_\omega)$ is a set of fixed point of T such that $\omega_\lambda(x_*, z) < \infty$ for all $\lambda > 0$, then $x_* = z$. Let x_0 be an element in X_ω such that $\omega_\lambda(x_0, Tx_0) < \infty$ for all $\lambda > 0$ and we write $x_1 = Tx_0$, $x_2 = Tx_1 = T^2 x_0$ and, in general, $x_n = Tx_{n-1} = T^n x_0$ for all $n \geq 1$. Observe that

$$\omega_\lambda(T^n x_0, T^{n+1} x_0) \leq k\omega_\lambda(T^{n-1} x_0, T^n x_0) \leq \cdots \leq k^n \omega_\lambda(x_0, Tx_0) < \infty$$

 ISBN: 978-84-617-8694-7

for all $n \geq 1$. Assume that n and m are two positive integers with $m > n$. Then we have

$$\begin{aligned}
\omega_\lambda(T^n x_0, T^m x_0) &\leq \omega_{\frac{\lambda}{m-n}}(T^n x_0, T^{n+1} x_0) + \omega_{\frac{\lambda}{m-n}}(T^{n+1} x_0, T^{n+2} x_0) \\
&\quad + \cdots + \omega_{\frac{\lambda}{m-n}}(T^{m-1} x_0, T^m x_0) \\
&\leq (k^n + k^{n+1} + \cdots + k^{m-1}) \omega_{\frac{\lambda}{m-n}}(x_0, Tx_0) \\
&\leq (k^n + k^{n+1} + \cdots) \omega_{\frac{\lambda}{m}}(x_0, Tx_0) \\
&= k^n 1 - k\omega_{\frac{\lambda}{m}}(x_0, Tx_0).
\end{aligned}$$

Since $\omega_\lambda(x_0, Tx_0) < \infty$ for all $\lambda > 0$, we deduce that, for any $\epsilon > 0$, $\omega_\lambda(T^n x_0, T^m x_0) < \epsilon$ for all $m > n > N$ with sufficiently large. Thus $\{T^n x_0\}$ is a Cauchy sequence and hence it converges to some $x_* \in X_\omega$ by the completeness of X_ω. Observe further that

$$\omega_\lambda(x_*, Tx_*) \leq \omega_{\frac{\lambda}{2}}(x_*, T^n x_0) + k\omega_{\frac{\lambda}{2}}(T^{n-1} x_0, x_*).$$

Letting $n \to \infty$, we have $\omega_\lambda(x_*, Tx_*) = 0$ for all $\lambda > 0$. Therefore, x_* is a fixed point of f.

Let z be another fixed points of T such that $\omega_\lambda(x_*, z) < \infty$ for all $\lambda > 0$, then we get

$$\omega_\lambda(x_*, z) \quad = \omega_\lambda(Tx_*, Tz) \leq \quad k\omega_\lambda(x_*, z)$$

for all $\lambda > 0$. Since $0 \leq k < 1$, we get $\omega_\lambda(x, z) = 0$ for all $\lambda > 0$, which implies that $x_* = z$. This completes the proof.

Theorem 2.1 *Let X_ω be a complete modular metric space and $T : X_\omega \to X_\omega$ be a contraction mapping. Suppose that $x^* \in X_\omega$ is a fixed point of T, $\{\varepsilon_n\}$ is a sequence of positive numbers for which $\lim_{n\to\infty} \varepsilon_n = 0$ and $\{y_n\} \subseteq X_\omega$ satisfies*

$$\omega_\lambda(y_{n+1}, Ty_n) \leq \varepsilon_n$$

for all $\lambda > 0$. Then $n \to \infty \lim y_n = x^*$.

Let $y_0 = x \in X_\omega$. Then we observe that, for all $m \geq 1$,

$$\begin{aligned}
\omega_\lambda(T^{m+1} x, y_{m+1}) &= \omega_{\frac{\lambda \cdot m}{m}}(T^{m+1} x, y_{m+1}) \\
&\leq \omega_{\frac{\lambda \cdot (m-1)}{m}}(T^{m+1} x, Ty_m) + \omega_{\frac{\lambda}{m}}(Ty_m, y_{m+1}) \\
&\leq k\omega_{\frac{\lambda \cdot (m-1)}{m}}(T^m x, y_m) + \varepsilon_m \\
&\leq k\omega_{\frac{\lambda \cdot (m-2)}{m}}(T^m x, Ty_{m-1}) + kw_{\frac{\lambda}{m}}(Ty_{m-1} x, y_m) + \varepsilon_m \\
&\leq k^2 \omega_{\frac{\lambda \cdot (m-2)}{m}}(T^{m-1} x, y_{m-1}) + k\varepsilon_{m-1} + \varepsilon_m \\
&\quad \vdots \\
&\leq \sum_{i=0}^{m} k^{m-i} \varepsilon_i
\end{aligned} \tag{3}$$

 ISBN: 978-84-617-8694-7

for all $\lambda > 0$. Thus we get

$$\begin{aligned} \omega_\lambda(y_{m+1}, x^*) &\leq \omega_{\frac{\lambda}{2}}(y_{m+1}, T^{m+1}x) + \omega_{\frac{\lambda}{2}}(T^{m+1}x, x^*) \\ &\leq \sum_{i=0}^{m} k^{m-i}\varepsilon_i + \omega_{\frac{\lambda}{2}}(T^{m+1}x, x^*). \end{aligned} \tag{4}$$

Next, we claim that $m \to \infty \lim \omega_\lambda(y_{m+1}, x^*) = 0$ for all $\lambda > 0$. Now, let $\varepsilon > 0$. Since $n \to \infty \lim \varepsilon_n = 0$, there exists a positive integer N such that, for all $m \geq N$, $\varepsilon_m \leq \varepsilon$. Thus we have

$$\begin{aligned} \sum_{i=0}^{m} k^{m-i}\varepsilon_i &= \sum_{i=0}^{N} k^{m-i}\varepsilon_i + \sum_{i=N+1}^{m} k^{m-i}\varepsilon_i \\ &\leq k^{m-N} \sum_{i=0}^{N} k^{N-i}\varepsilon_i + \varepsilon \sum_{i=N+1}^{m} k^{m-i}. \end{aligned} \tag{5}$$

Taking limit as $m \to \infty$ in (5), we have

$$\lim_{m\to\infty} \sum_{i=0}^{m} k^{m-i}\varepsilon_i = 0. \tag{6}$$

Since x^* is a fixed point of T, using Theorem 2.1, it follows that the sequence $\{T^n x\}$ converge to x^*. This implies that

$$\lim_{m\to\infty} \omega_{\frac{\lambda}{2}}(T^{m+1}x, x^*) = 0 \tag{7}$$

for all $\lambda > 0$. Therefore, from (4), (6) and (7), we have

$$\lim_{m\to\infty} \omega_\lambda(y_{m+1}, x^*) = 0 \tag{8}$$

for all $\lambda > 0$, which implies that $\lim_{n\to\infty} y_n = x^*$. This completes the proof.

Theorem 2.2 *Let X_ω be a complete modular metric space and, for any $x^* \in X_\omega$, define*

$$B_\omega(x^*, \gamma) := \{x \in X_\omega : \omega_\lambda(x, x^*) \leq \gamma, \ \forall \lambda > 0\}.$$

If $T : B_\omega(x^*, \gamma) \to X_\omega$ is a contraction mapping with

$$\omega_{\frac{\lambda}{2}}(Tx^*, x^*) \leq (1 - k)\gamma \tag{9}$$

for all $\lambda > 0$, where $0 \leq k < 1$, then T has a fixed point in $B_\omega(x^*, \gamma)$.

By Theorem 2.1, we only prove that $B_\omega(x^*, \gamma)$ is complete and $Tx \in B_\omega(x^*, \gamma)$ for all $x \in B_\omega(x^*, \gamma)$. Suppose that $\{x_n\}$ is a Cauchy sequence in $B_\omega(x^*, \gamma)$, and then also $\{x_n\}$ is a Cauchy sequence in X_ω. Since X_ω is complete, there exists $x \in X_\omega$ such that

$$\lim_{n\to\infty} \omega_{\frac{\lambda}{2}}(x_n, x) = 0 \tag{10}$$

 ISBN: 978-84-617-8694-7

for all $\lambda > 0$. Since, for each $n \geq 1$, $x_n \in B_\omega(x^*, \gamma)$, using the property of a metric modular, we get

$$\begin{aligned} \omega_\lambda(x^*, x) &\leq \omega_{\frac{\lambda}{2}}(x^*, x_n) + \omega_{\frac{\lambda}{2}}(x_n, x) \\ &\leq \gamma + \omega_{\frac{\lambda}{2}}(x_n, x^*) \end{aligned} \tag{11}$$

for all $\lambda > 0$. It follows the inequalities (10) and (11) that $w_\lambda(x^*, x) \leq \gamma$, which implies that $x \in B_\omega(x^*, \gamma)$. Therefore, $\{x_n\}$ is a convergent sequence in $B_\omega(x^*, \gamma)$ and also $B_\omega(x^*, \gamma)$ is complete.

Next, we prove that $Tx \in B_\omega(x^*, \gamma)$ for all $x \in B_\omega(x^*, \gamma)$. Let $x \in B_\omega(x^*, \gamma)$. From the inequalities (9), the contraction of T and the notion of a metric modular, we have

$$\begin{aligned} \omega_\lambda(x^*, Tx) &\leq \omega_{\frac{\lambda}{2}}(x^*, Tx^*) + \omega_{\frac{\lambda}{2}}(Tx^*, Tx) \\ &\leq (1-k)\gamma + k\omega_{\frac{\lambda}{2}}(x^*, x) \\ &\leq (1-k)\gamma + k\gamma \\ &= \gamma. \end{aligned}$$

Therefore, $Tx \in B_\omega(x^*, \gamma)$. This completes the proof.

Theorem 2.3 *Let X_ω be a complete modular metric space and T be a self-mapping on X_ω satisfying*

$$\omega_\lambda(Tx, Ty) \leq k(\omega_{2\lambda}(Tx, x) + \omega_{2\lambda}(Ty, y)) \tag{12}$$

for all $x, y \in X_\omega$, where $k \in \left[0, \frac{1}{2}\right)$. Assume that there exists $x_0 \in X$ such that $\omega_\lambda(x_0, Tx_0) < \infty$ for all $\lambda > 0$. Then T has a fixed point in $x \in X_\omega$ and the sequence $\{T^n x_0\}$ converges to x. Moreover, if, $z \in F(X_\omega)$, where $F(X_\omega)$ is a set of fixed point of T such that $\omega_\lambda(x_*, z) < \infty$ for all $\lambda > 0$, then $x_* = z$.

Let x_0 be an element in X_ω such that $\omega_\lambda(x_0, Tx_0) < \infty$ for all $\lambda > 0$. We write $x_1 = Tx_0, x_2 = Tx_1 = T^2x_0$ and, in general, $x_n = Tx_{n-1} = T^n x_0$ for all $n \geq 1$. If $Tx_{n_0-1} = Tx_{n_0}$ for some $n_0 \geq 1$, then $Tx_{n_0} = x_{n_0}$. Thus x_{n_0} is a fixed point of T.

Suppose that $Tx_{n-1} \neq Tx_n$ for all $n \geq 1$. For any $k \in [0, \frac{1}{2})$, we have

$$\begin{aligned} \omega_\lambda(x_{n+1}, x_n) &= \omega_\lambda(Tx_n, Tx_{n-1}) \\ &\leq k(\omega_{2\lambda}(Tx_n, x_n) + \omega_{2\lambda}(Tx_{n-1}, x_{n-1})) \\ &\leq k(\omega_\lambda(x_{n+1}, x_n) + \omega_\lambda(x_n, x_{n-1})) \end{aligned} \tag{13}$$

for all $\lambda > 0$ and $n \geq 1$. Hence we have

$$\omega_\lambda(x_{n+1}, x_n) \leq \tfrac{k}{1-k}\omega_\lambda(x_n, x_{n-1}) \tag{14}$$

for all $\lambda > 0$ and $n \geq 1$. Put $\beta := \frac{k}{1-k}$. Since $k \in (0, \frac{1}{2})$, we get $\beta \in (0, 1)$ and hence

$$\begin{aligned} \omega_\lambda(x_{n+1}, x_n) &\leq \beta\omega_\lambda(x_n, x_{n-1}) \\ &\leq \beta^2\omega_\lambda(x_{n-1}, x_{n-2}) \\ &\vdots \\ &\leq \beta^n\omega_\lambda(x_1, x_0) \end{aligned} \tag{15}$$

 ISBN: 978-84-617-8694-7

for all $\lambda > 0$ and $n \geq 1$. Similar to the proof of Theorem 2.1, we can conclude that $\{x_n\}$ is a Cauchy sequence and, by the completeness of X_ω there exists a point $x \in X_\omega$ such that $x_n \to x$ as $n \to \infty$. By the property of a metric modular and the inequality (12), we have

$$\begin{aligned} \omega_\lambda(Tx, x) &\leq \omega_{\frac{\lambda}{2}}(Tx, Tx_n) + \omega_{\frac{\lambda}{2}}(Tx_n, x) \\ &\leq k(\omega_\lambda(Tx, x) + \omega_\lambda(Tx_n, x_n)) + \omega_{\frac{\lambda}{2}}(Tx_n, x) \\ &\leq k(\omega_\lambda(Tx, x) + \omega_{\frac{\lambda}{2}}(Tx_n, x) + \omega_{\frac{\lambda}{2}}(x, x_n)) + \omega_{\frac{\lambda}{2}}(Tx_n, x) \\ &= k(\omega_\lambda(Tx, x) + \omega_{\frac{\lambda}{2}}(x_{n+1}, x) + \omega_{\frac{\lambda}{2}}(x, x_n)) + \omega_{\frac{\lambda}{2}}(x_{n+1}, x) \end{aligned} \tag{16}$$

for all $\lambda > 0$ and $n \geq 1$. Taking $n \to \infty$ in the inequality (16), we obtain

$$\omega_\lambda(Tx, x) \leq k\omega_\lambda(Tx, x). \tag{17}$$

Since $k \in [0, \frac{1}{2})$, we have $Tx = x$. Thus x is a fixed point of T.

Let z be another fixed points of T such that $\omega_\lambda(x_*, z) < \infty$ for all $\lambda > 0$, then we get

$$\begin{aligned} \omega_\lambda(x, z) &= \omega_\lambda(Tx, Tz) \\ &\leq k(\omega_{2\lambda}(Tx, x) + \omega_{2\lambda}(Tz, z)) \\ &= 0 \end{aligned}$$

for all $\lambda > 0$, which implies that $x = z$. This completes the proof.

2.2 Kannan's set valued contraction mappings

Before we could stomp into the main exploration, we need the following knowledge of metric modular of sets.

We write (X) to denote the set of all nonempty closed subsets of X. For any subset $A \subset X_w$ and point $x \in X$, we denote $w_t(x, A) := \inf_{y \in A} w_t(x, y)$.

Given two subsets $A, B \in (X)$, define $w_t(A, B) := \sup_{x \in A} w_t(x, B)$. Most importantly, the Hausdorff-Pompieu metric modular $W_t(A, B) := \max\{w_t(A, B), w_t(B, A)\}$.

Let (X, w) be a modular metric space, $A \in (X)$, and $x \in X$. Then,

$\mathrm{w}_t(x, A) = 0$ for all $t > 0 \iff x \in A$.

Given a modular metric space (X, w) and an arbitrary point $x \in X$. A subset $Y \subset X$ is said to be *reachable* from x if

$$\inf_{y \in Y} \sup_{t>0} w_t(x, y) = \sup_{t>0} w_t(x, Y) < \infty.$$

This lemma gives a simple criterion of when the reachability holds. Let (X, w) be a modular metric space with w being l.s.c., $Y \subset X$ a nonempty compact subset. For a point

 ISBN: 978-84-617-8694-7

$x \in X$, if either $\inf_{y\in Y} \sup_{t>0} w_t(x,y) < \infty$ or $\sup_{t>0} w_t(x,Y) < \infty$, then Y is reachable from x.

The following lemma is essential in showing the solvability of fixed point inclusion for contractivity condition. Suppose that $Y, Z \in (X)$ are nonempty and $z \in Z$. If Y is reachable from z, then for each $\varepsilon > 0$, there exists a point $y_\varepsilon \in Y$ such that $\sup_{t>0} w_t(z, y_\varepsilon) \leq \sup_{t>0} W_t(X,Y) + \varepsilon$.

Now, we state the notion of the contraction and the Kannan's contraction. Make note that these two concepts are not generalizations of one another.

Let (X, w) be a modular metric space. A set-valued operator $F : X \rightrightarrows X$ is said to be a *contraction* if there exists a constant $k \in [0, 1)$ such that

$$W_t(Fx, Fy) \leq k w_t(x, y), \tag{18}$$

for all $t > 0$ and $x, y \in X$. If k is restricted in $[0, \frac{1}{2})$ and the equation above is replaced with the following inequality:

$$W_t(F(x), F(y)) \leq k[w_t(x, F(x)) + w_t(y, F(y))].$$

Then, we call F a *Kannan's contraction.*

Now, we present the main existence theorems (see [7]).

Theorem 2.4 *Let (X, w) be a complete modular metric space with w being l.s.c., and F a contraction on X having compact values with contraction constant k. Suppose that there exists a pair of points $x_0 \in X$ and $x_1 \in F(x_0)$ with the following properties:*

1. *the set $\{x_0, x_1\}$ is bounded,*
2. *$F(x_1)$ is reachable from x_1.*

Then, F has at least one fixed point.

3 Fractional Integral Inclusion

In this section, we shall illustrate some use of the theorems presented in the previous section. We focus on applications to integral inclusion.

For another application, we investigate the solvability of the fractional inclusion problem in which we include also the delayed behaviors. Fractional integral is used in describing various natural phenomena, and also in electric transmission, travelling waves, fluid dynamics, and in control engineering

It is natural to raise the situation of set-valued integral, which proved itself for its importance in practical applications especially in engineering. In 1965, Aumann [2] introduced

 ISBN: 978-84-617-8694-7

the concept of definite set-valued integral on real line and Euclidean spaces. Suppose that Ψ is an interval $[0,T]$, where $T>0$. Let $F:\Psi\to 2^{\mathbb{R}}$ be a set-valued operator. A selection of F is the function $f:\Psi\to\mathbb{R}\cup\{\pm\infty\}$ such that $f(t)\in F(t)$ a.e. $t\in\Psi$. We write F to denote the set containing all integrable selections of F. According to Aumann [2], the set-valued integral is determined by the operator in the following:

$$ {}_{\Psi}F(t)dt := \left\{ \int_{\Psi} f(t)dt \ ; \ f\in F \right\}, $$

i.e., the set of the integrals of integrable selections of F.

On the other hand, in elementary calculus, one deals with derivatives and integrals, including the higher-integer-order iterations. Here, in fractional integral, one looks at a broader concept where the real-order iteration is taken into account. There are many approaches to study this kind of extensions. In our context, we shall use the classical notion introduced by Riemann and Liouville, the latter of which is the first one to point out the possibility of fractional calculus in 1832. Given a function $f\in L^1(\Psi,\mu)$, the fractional integral of order $\alpha>0$ is given by

$$ {}^{\alpha}_{\Psi}f(t)dt := \frac{1}{\Gamma(\alpha)}\int_{\Psi}(t-\tau)^{\alpha-1}f(\tau)d\tau. $$

Naturally, we may further consider the following fractional integral:

$$ {}^{\alpha}_{\Psi}F(t)dt := \left\{ {}^{\alpha}_{\Psi}f(t)dt \ ; \ f\in F \right\}. $$

In this particular subsection, we shall use notations a bit differently than those of earlier sections. This is due to conventional uses of variables and functions that is common to integral and differential equations.

Suppose that Ψ is the interval mentioned in the previous section. Let us assume throughout the section that the real line $\mathbb{R}$ is equipped with the metric modular

$$ \omega^{\mathbb{R}}_{\lambda}(x,y) := \frac{1}{1+\lambda}\left|x-y\right|, $$

for $\lambda>0$ and $x,y\in\mathbb{R}$. Thus, for the space $C(\Psi)$ of all continuous (in $\omega^{\mathbb{R}}$-topology) real-valued functions on Ψ, we shall use the metric modular

$$ \omega^{C(\Psi)}_{\lambda}(\varphi,\psi) := \sup_{t\in\Psi}\omega^{\mathbb{R}}_{\lambda}(\varphi(t),\psi(t)), $$

for $\lambda>0$ and $\varphi,\psi\in C(\Psi)$. Note that both $\omega^{\mathbb{R}}$ and $\omega^{C(\Psi)}$ satisfy the Fatou's property. Also note that the set $\mathbb{R}$ is second countable, i.e., it has a countable base, w.r.t. $\omega^{\mathbb{R}}$-topology. Moreover, it is clear that the set $\{\varphi,\psi\}$ is bounded w.r.t. $\omega^{C(\Psi)}$, for any $\varphi,\psi\in C(\Psi)$. Suppose that $F:\Psi\times\mathbb{R}\to 2^{\mathbb{R}}$ is a set-valued operator with nonempty compact values,

 ISBN: 978-84-617-8694-7

and $u \in C(\Psi)$. We shall use the following notation to explain the collection of integrable selections:

$$S_F(u) := \left\{ f \in L^1(\Psi, \mu) \; ; \; f(t) \in F(t, u(t)) a.e.t \in \Psi \right\}.$$

It is clear that $S_F(u)$ is closed. Next, for each $i \in \{0, 1, \cdots, N\}$, $N \in \mathbb{N}$, assume that $\beta_i : \Psi \to \mathbb{R}$ is continuous, and $\tau_i : \Psi \to \mathbb{R}_+$ is a function with $\tau_i(t) \leq t$. We write $B := \max_{0 \leq i \leq N} \sup_{t \in \Psi} \beta_i(t)$. The main aim of this section is to consider the fractional integral inclusion:

$$u(t) - \sum_{i=0}^{N} \beta_i(t) u(t - \tau_i(t)) \in {}^{\alpha}_{\Psi} F(t, u(t)) dt, \quad \alpha \in (0, 1]. \quad (FII)$$

In the above inclusion, the summation here is interpreted to be the delay term.

We shall define a set-valued operator $\Lambda : C(\Psi) \to 2^{C(\Psi)}$ by

$$\Lambda(u) := \left\{ w \in C(\Psi) \; ; \; w(t) = \sum_{i=0}^{N} \beta_i(t) u(t - \tau_i(t)) + {}^{\alpha}_{\Psi} f(t, u(t)) dt, \quad f \in S_F(u) \right\}.$$

Note here that for any $\varphi \in C(\Psi)$, we have $\Lambda(\varphi)$ is reachable from φ w.r.t. $\omega^{C(\Psi)}$.
To restrict the operator Λ with some nice property, we assume that $S_F(u)$ is nonempty.

Lemma 3.1 *The operator Λ given above is compact valued if $S_F(u)$ is nonempty.*

For the proof, we shall show the compactness by its sequential characterization. Suppose that $u \in C(\Psi)$, and (w_n) is an arbitrary sequence in $\Lambda(u)$. By definition, there corresponds a convergent sequence (f_n) in $S_F(u) \subset F(\cdot, u(\cdot))$ satisfying

$$w_n(t) = \sum_{i=0}^{N} \beta_i(t) u(t - \tau_i(t)) + {}^{\alpha}_{\Psi} f_n(t, u(t)) dt.$$

The conclusion is then followed.

Now, we shall state now the solvability result for the problem (FII). It is clear that $u \in C(\Psi)$ solves (FII) if and only if u is a fixed point of Λ.

Theorem 3.2 *Suppose that F defined above is compact-valued, and $S_F(u)$ is nonempty. Assume further that*

1. *for any given $u, v \in C(\Psi)$ and a selection $f \in S_F(u)$ of F, there corresponds a function $f' \in S_F(v)$ such that*

$$\begin{cases} \omega^{\mathbb{R}}_{\lambda}(f(t, u(t)), f'(t, v(t))) = \omega^{\mathbb{R}}_{\lambda}(f_1(t, u(t)), F(t, v(t))), \\ \omega^{\mathbb{R}}_{\lambda}(f(t, u(t)), f'(t, v(t))) \leq L \omega^{C(\Psi)}_{\lambda}(u, v), \end{cases}$$

for all $t \in \Psi$;

 ISBN: 978-84-617-8694-7

2. $\frac{(N+1)B\Gamma(\alpha)+LT^{\alpha}}{\Gamma(\alpha)} < 1.$

Then, Λ has a fixed point.

For each $u, v \in C(\Psi)$, we may choose, from the assumption, functions f_1, f_2 such that

$$\begin{cases} f_1 \in S_F(u), \\ f_2 \in S_F(v), \\ \omega_{\lambda}^{\mathbb{R}}(f_1(t,u(t)), f_2(t,v(t))) = \omega_{\lambda}^{\mathbb{R}}(f_1(t,u(t)), F(t,v(t))), \\ \omega_{\lambda}^{\mathbb{R}}(f_1(t,u(t)), f_2(t,v(t))) \leq L\omega_{\lambda}^{C(\Psi)}(u,v), \end{cases}$$

for each $t \in \Psi$. Consider the two functions $w_1 \in \Lambda(u)$ and $w_2 \in \Lambda(v)$, respectively as follows:

$$\begin{cases} w_1(t) := \sum_{i=0}^{N} \beta_i(t)u(t-\tau_i(t)) +_{\Psi}^{\alpha} f_1(t,u(t))dt, \\ w_2(t) := \sum_{i=0}^{N} \beta_i(t)v(t-\tau_i(t)) +_{\Psi}^{\alpha} f_2(t,v(t))dt. \end{cases}$$

Now, consider the following computation:

$$\begin{aligned} &\omega_{\lambda}^{\mathbb{R}}(w_1(t), w_2(t)) \\ &\leq \sum_{i=0}^{N} \beta_i(t)\omega_{\lambda}^{\mathbb{R}}(u(t-\tau_i(t)), v(t-\tau_i(t)) \\ &\quad +\omega_{\lambda}^{C(\Psi)}({}_{\Psi}^{\alpha}f_1(t,u(t))dt, {}_{\Psi}^{\alpha} f_2(t,u(t))dt) \\ &\leq (N+1)B\omega_{\lambda}^{C(\Psi)}(u,v) +_{\Psi}^{\alpha} \omega_{\lambda}^{\mathbb{R}}(f_1(t,u(t)), f_2(t,v(t))) \\ &\leq (N+1)B\omega_{\lambda}^{C(\Psi)}(u,v) + \frac{LT^{\alpha}}{\Gamma(\alpha)}\omega_{\lambda}^{C(\Psi)}(u,v) \\ &= \left[\frac{(N+1)B\Gamma(\alpha)+LT^{\alpha}}{\Gamma(\alpha)}\right]\omega_{\lambda}^{C(\Psi)}(u,v). \end{aligned}$$

It follows that

$$\Omega_{\lambda}^{C(\Psi)}(\Lambda(u), \Lambda(v)) \leq \left[\frac{(N+1)B\Gamma(\alpha)+LT^{\alpha}}{\Gamma(\alpha)}\right]\omega_{\lambda}^{C(\Psi)}(u,v).$$

The proof ends here by applying Theorem 2.4.

4 Conclusion

Taking into account several classes of nonlinear differential equations, either ordinary or partial, we may design the corresponding nonlinear operator to suit a particular class of the original problem with the most important property; the fixed point of the corresponded

 ISBN: 978-84-617-8694-7

operator solves the original problem. By imposing some conditions on the involved differential equations, we can show equivalence between the original differential equation and the transformed fixed point equation. Moreover, we can also derive the solution existence from appropriate fixed point theorems and finally obtain the desired results.

Using simple conditions on the underlying space and operators, we show existence of fixed point for such operators. We emphasize results in nonlinear spaces and on generalized contractive operators. Finally, we deliver an equivalence existence theorems for certain classes of nonlinear system of differential equations. Our results cover both initial valued and boundary value problems.

Acknowledgements
The first author was supported by Rajamangala University of Technology Thanyaburi (RMUTT) for financial support. This project was partially supported by the Theoretical and Computational Science (TaCS) Center under Computational and Applied Science for Smart Innovation Cluster (CLASSIC), Faculty of Science, KMUTT. Moreover, this research work was financially supported by King Mongkut's University of Technology North Bangkok. Contract No. KMUTNB-60-ART-084.

References

[1] A. H. Ansari, P. Kumam, and B. Samet. A fixed point problem with constraint inequalities via an implicit contraction. *Journal of Fixed Point Theory and Applications*, pages 1–19, 2016.

[2] R. J. Aumann. Integrals of set-valued functions. *Journal of Mathematical Analysis and Applications*, 12(1):1 – 12, 1965.

[3] H. Aydi, C. Vetro, W. Sintunavarat, and P. Kumam. Coincidence and fixed points for contractions and cyclical contractions in partial metric spaces. *Fixed Point Theory and Applications*, 2012(1):124, 2012.

[4] S. Banach. Sur les opérations dans les ensembles abstraits et leur application aux équations intégrales. *Fundamenta math.*, 3:133–181, 1922.

[5] P. Chaipunya, Y. Je Cho, and P. Kumam. Geraghty-type theorems in modular metric spaces with an application to partial differential equation. *Advances in Difference Equations*, 2012(1):83, 2012.

[6] P. Chaipunya and P. Kumam. An observation on set-valued contraction mappings in modular metric spaces. *Thai Journal of Mathematics*, 13(1):9–17, 2015.

 ISBN: 978-84-617-8694-7

[7] P. Chaipunya, C. Mongkolkeha, W. Sintunavarat, and P. Kumam. Fixed-point theorems for multivalued mappings in modular metric spaces. *Abstr. Appl. Anal.*, 2012:14 p., 2012.

[8] V. V. Chistyakov. Modular metric spaces. I: Basic concepts. *Nonlinear Anal., Theory Methods Appl., Ser. A, Theory Methods*, 72(1):A, 1–14, 2010.

[9] D. Delbosco and L. Rodino. Existence and uniqueness for a nonlinear fractional differential equation. *J. Math. Anal. Appl.*, 204(2):609–625, 1996.

[10] A. El-Sayed and A. Ibrahim. Multivalued fractional differential equations. *Applied Mathematics and Computation*, 68(1):15 – 25, 1995.

[11] A. M. El-Sayed and A.-G. Ibrahim. Set-valued integral equations of fractional-orders. *Applied Mathematics and Computation*, 118(1):113 – 121, 2001.

[12] A.-G. Ibrahim and A. M. El-Sayed. Definite integral of fractional order for set-valued functions. *J. Fractional Calc.*, 11:81–87, 1997.

[13] A. Kilbas and J. Trujillo. Differential equations of fractional order: Methods, results and problems. I. *Appl. Anal.*, 78(1-2):153–192, 2001.

[14] W. Kirk, P. Srinivasan, and P. Veeramani. Fixed points for mappings satisfying cyclical contractive conditions. *Fixed Point Theory*, 4(1):79–89, 2003.

[15] Y. Ling and S. Ding. A class of analytic functions defined by fractional derivation. *J. Math. Anal. Appl.*, 186(2):504–513, 1994.

[16] C. Mongkolkeha, Geometric properties of some Banach spaces and fixed point theorems for generalized contraction mappings, A Ph. D. Dissertation (Applied Mathematics), Faculty of Science, King Mongkut's University of Technology Thonburi (KMUTT), Thailand, (2013).

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

Algorithms for accretive operators with applications to convex minimization problem

Wiyada Kumam[1], Anantachai Padcharoen[2], Duangkamon Kitkuan[2] and Poom Kumam[3]

[1] *Program in Applied Statistics, Department of Mathematics and Computer Science, Faculty of Science and Technology, Rajamangala University of Technology Thanyaburi, Thanyaburi, Pathumthani 12110, Thailand.*

[2] *KMUTTFixed PointResearch Laboratory, Department of Mathematics, Room SCL 802 Fixed Point Laboratory, Science Laboratory Building, Faculty of Science, King Mongkut's University of Technology Thonburi (KMUTT), 126 Pracha Uthit Rd., Bang Mod, Thung Khru, Bangkok 10140, Thailand.*

[3] *KMUTT-Fixed Point Theory and ApplicationsResearch Group, Theoretical and Computational Science Center (TaCS), Science Laboratory Building, Faculty of Science, King Mongkuts University of Technology Thonburi (KMUTT), 126 Pracha-Uthit Road, Bang Mod, Thrung Khru, Bangkok 10140, Thailand*

emails: wiyada.kum@rmutt.ac.th, apadcharoen@yahoo.com, or_duangkamon@hotmail.com, poom.kum@kmutt.ac.th

Abstract

In this paper, we introduce a new iterative method for finding the common zeros of two accretive operators in the framework of uniformly smooth Banach spaces. Also, we will prove the strong convergence theorems for the iterative algorithms and give the example of the main theorems. The results of this paper are improvements and extensions of the corresponding ones announced by many others.

Key words: common zeros, accretive operator
MSC 2000: 47H10, 54H25.

ISBN: 978-84-617-8694-7

1 Introduction

Let E be a real Banach space, C a nonempty closed convex subset of E, and Let $f : C \to C$ be a mapping. Recall that f is said to be contractive if there exists a constant $\sigma \in (0, 1)$ such that $\|fx - fy\| \leq \sigma\|x - y\|$ for all $x, y \in C$, f is said to be nonexpansive if $\sigma = 1$.

A point $x \in C$ is a fixed point of T provided $Tx = x$. Denote by $F(T)$ the set of fixed points of T; that is, $F(T) = \{x \in C : Tx = x\}$. It is assumed throughout the paper that T is a nonexpansive mapping such that $F(T) \neq \emptyset$.

For an operator $A : E \to 2^E$, we denote its domain, range and graph as follows:

$$D(A) = \{x \in E : Ax \neq \emptyset\},$$
$$\mathcal{R}(A) = \cup\{Az : z \in D(A)\},$$

and

$$G(A) = \{(x, y) \in E \times E : x \in D(A), y \in Ax\},$$

respectively. The inverse A^{-1} ofA is defined by $x \in A^{-1}y$, if and only if $y \in Ax$.

Let E be a real Banach space with norm $\|\cdot\|$ and let E^* be its dual. The value of $f \in E^*$ at $x \in E$ will be denoted by $\langle x, f\rangle$. A Banach space E is said to be strictly convex if $\frac{\|x+y\|}{2} < 1$ for all $x, y \in E$ with $\|x\| = \|y\| = 1$. It is also said to be uniformly convex if $\lim_{n\to\infty} \|x_n - y_n\| = 0$. for any two sequences $\{x_n\}, \{y_n\}$ in E such that $\|x\| = \|y\| = 1$. and $\lim_{n\to\infty} \frac{\|x+y\|}{2} = 1$.

The (normalized) duality mapping J from E into the family of nonempty (by Hahn Banach theorem) weak-star compact subsets of its dual E is defined by

$$J(x) = \{f \in E^* : \langle x, f\rangle = \|x\|^2 = \|f\|^2\}$$

for each $x \in E$, where $\langle\cdot, \cdot\rangle$ denotes the generalized duality pairing.

The norm of E is said to be Gâteaux differentiable if

$$\lim_{t\to 0} \frac{\|x + ty\| - \|x\|}{t}$$

exists for each x, y in its unit sphere $U = \{x \in E : \|x\| = 1\}$.

A closed convex subset C of a Banach space E is said to have the fixed point property for nonexpansive mappings if every nonexpansive mapping of a nonempty closed convex subset D of C into itself has a fixed point in D.

A subset C of Banach space E is called a retract of E if there is a continuous mapping Q from E onto C such that $Qx = x$ for all $x \in C$. We call such Q a retraction of E onto C. It follows that if a mapping Q is a retraction, then $Qy = y$ for all y in the range of Q. A retraction Q is said to be sunny if $Q(Qx + t(xQx)) = Qx$ for all $x \in E$ and $t \geq 0$. If a sunny

 ISBN: 978-84-617-8694-7

retraction Q is also non expansive, then C is said to be a sunny non expansive retract of E [1]. In a smooth Banach space E, it is known (cf. [1], p. 48) that $Q : C \to D$ is a sunny nonexpansive retraction if and only if the following condition holds:

$$\langle x - Q(x), J(z - Q(x))\rangle \leq 0,\ x \in C,\ z \in D.$$

We know that the sunny non expansive retract mapping form Hilbert space H onto a closed convex subset $C \subset H$ is metric projection and we use P_C to denote the metric projection from H onto C.

An accretive operator A in a Banach space X is said to satisfy the range condition if $\overline{D(A)} \subset \mathcal{R}(I + \lambda A)$ for all $\lambda > 0$, where $\overline{D(A)}$ denotes the closure of the domain of A. We know that for an accretive operator A which satisfies the range condition, $A^{-1}0 = F(J_\lambda^A)$ for all $\lambda > 0$.

2 Preliminaries

Lemma 2.1. *[2] A Banach space E is uniformly smooth if and only if the duality map J is the single-valued and norm-to-norm uniformly continuous on bounded sets of E.*

Lemma 2.2. *[3] Let E be a Banach space. Then for every $x, y \in E$, we have*

$$\|x + y\|2 \leq \|x\|^2 + 2\langle y, j(x + y)\rangle,$$

for all $j(x + y) \in J(x + y)$.

Lemma 2.3. *[4] Let E be a Banach space and let Let $A : D(A) \to 2^E$ be an accretive operator. For $\lambda > 0, \mu > 0$ and $x \in E$, we have $J_\lambda x = J_\mu(\frac{\mu}{\lambda}x + (1 - \frac{\mu}{\lambda})J_\lambda x)$, where $J_\lambda = (I + \lambda A)^{-1}$ and $J_\mu = (I + \mu A)^{-1}$.*

Lemma 2.4. *[6] Let $\{a_n\}$ be a sequence of nonnegative real numbers such that $a_{n+1} \leq (1 - t_n)a_n + b_n + c_n$, where $\{c_n\}$ is a sequence of nonnegative real numbers, $\{t_n\} \subset (0, 1)$ and $\{b_n\}$ is a number sequence. Assume that $\sum_{n=0}^{\infty} t_n = \infty, \limsup_{n\to\infty} \frac{b_n}{t_n} \leq 0$, and $\sum_{n=0}^{\infty} c_n < \infty$. Then $\lim_{n\to\infty} a_n = 0$.*

Theorem 2.5. *[7] Let E be a reflexive Banach space with a uniformly Gâteaux differentiable norm such that every weakly compact convex subset of E has fixed point property for nonexpansive mappings. Let C be a closed convex subset of E and T non expansive mapping from C into itself with $F(T) \neq \emptyset$. Then $\{x_t\}$ defined by $x_t = tfx_t + (1 - t)Tx_t$ where $f : C \to C$ is a contraction mapping and $t \in (0, 1)$, converges strongly to a point in $x^* \in F(T)$ which satisfies $Q_{F(T)}f(x^*) = x^*$.*

 ISBN: 978-84-617-8694-7

3 Main results

Theorem 3.1. *Let C be a nonempty closed convex subset of a uniformly convex and uniformly smooth Banach space X. Let $A : D(A) \subseteq C \to 2^X$ and $B : D(B) \subseteq C \to 2^X$ be accretive operators such that $\Omega = A^{-1}0 \cap B^{-1}0 \neq \emptyset$, $\overline{D(A)} \subset C \subset \bigcap_{r>0} \mathcal{R}(I + rA)$ and $\overline{D(B)} \subset C \subset \bigcap_{r>0} \mathcal{R}(I + rB)$. Let $f : C \to C$ be a contrative mapping with contractive constant $\sigma \in (0, 1)$. Let $\{x_n\}$ be a sequence generated by $x_0 \in C$ and*

$$\begin{cases} y_n = \beta_n x_n + (1 - \beta_n) J^A_{\lambda_n} x_n \\ x_{n+1} = \alpha_n f(x_n) + (1 - \alpha_n) J^B_{\gamma_n} y_n, \end{cases} \tag{1}$$

where the sequence $\{\alpha_n\}, \{\beta_n\}$, and $\{\gamma_n\}$ satisfy the following restrictions:

(i) $\lim_{n\to\infty} \alpha_n = 0, \sum_{n=1}^{\infty} \alpha_n = \infty$;

(ii) $\lambda_n \geq \epsilon, \gamma_n \geq \epsilon$ *for some* $\epsilon > 0$ *and for all* n,
$\sum_{n=0}^{\infty} |\alpha_{n+1} - \alpha_n| < \infty$, $\sum_{n=0}^{\infty} |\beta_{n+1} - \beta_n| < \infty$, $\sum_{n=0}^{\infty} |\lambda_{n+1} - \lambda_n| < \infty$,
and $\sum_{n=0}^{\infty} |\gamma_{n+1} - \gamma_n| < \infty$.

Then $\{x_n\}$ converges strongly to $w = Qf(w)$, where Q is a sunny nonexpansive retraction of E onto Ω.

Theorem 3.2. *Let H be a Hilbert space. Let $A : H \to 2^H$ and $B : H \to 2^H$ be maximal monotone operators such that $S = A^{-1}0 \cap B^{-1}0 \neq \emptyset$. Let $f : C \to C$ be a contrative mapping with contractive constant $\sigma \in (0, 1)$. Let $\{x_n\}$ be a sequence generated by* (1)*, where the sequence $\{\alpha_n\}, \{\beta_n\}$, and $\{\gamma_n\}$ satisfy the following restrictions:*

(i) $\lim_{n\to\infty} \alpha_n = 0, \sum_{n=1}^{\infty} \alpha_n = \infty$;

(ii) $\lambda_n \geq \epsilon, \gamma_n \geq \epsilon$ *for some* $\epsilon > 0$ *and for all* n,
$\sum_{n=0}^{\infty} |\alpha_{n+1} - \alpha_n| < \infty$, $\sum_{n=0}^{\infty} |\beta_{n+1} - \beta_n| < \infty$, $\sum_{n=0}^{\infty} |\lambda_{n+1} - \lambda_n| < \infty$,
and $\sum_{n=0}^{\infty} |\gamma_{n+1} - \gamma_n| < \infty$.

Then $\{x_n\}$ converges strongly to $P_S f(w^*) = w^*$, where $P_S : H \to S$ is a metric projection from H onto S.

Acknowledgments

The first author thanks for the support of Petchra Pra Jom Klao Doctoral Scholarship for Ph.D. student of King Mongkut's University of Technology Thonburi. The second author thanks for Theoretical and Computational Science Center (TaCS), Furthermore, this work was supported by the Higher Education Research Promotion and National Research University Project of Thailand, Office of the Higher Education Commission (NRU59 Grant No.59000399).

References

[1] K. Goebel , S. Reich , Uniform convexity, hyperbolic geometry and non expansive mappings, Marcel Dekker, New York and Basel, 1984 .

[2] X. Qin, Y. Su, Approximation of a zero point of accretive operator in Banach spaces, J. Math. Anal. Appl. 329 (2007) 415-424.

[3] W.V. Petryshn , A characterization of strictly convexity of banach spaces and other uses of duality mappings, J. Funct. Anal. 6 (1970) 282-291.

[4] V. Barbu, Nonlinear semigroups and differential equations in Banach spaces, Noordhoff, Groningen, 1976.

[5] R.P. Agarwal , D. ORegan , D.R. Sahu , Fixed Point Theory for Lipschitzian-type Mappings with Applications, Springer, 2009 .

[6] L.S. Liu, Ishikawa and Mann iterative process with errors for nonlinear strongly accretive mappings in Banach spaces. J. Math. Anal. Appl. 194, 114-125 (1995).

[7] J.S. Jung , Viscosity approximation methods for family of finite non expansive mappings in banach spaces, Nonlinear Anal. Appl. 64 (2006) 2536-2552.

[8] G. López,, V. Martín-Márquez,F. Wang, H.K. Xu, Forward-Backward splittingmethods for accretive operators in Banach spaces. Abstr. Appl. Anal. 2012, 25 (2012).

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

Adaptive Steffensen-like with memory methods for solving nonlinear equations with highest efficiency indices

Mohammad Javad Lalehchini[1] and Taher Lotfi[1]

[1] *Department of Applied Mathematics, Hamedan Branch,,*
Islamic Azad University, Hamadan, Iran

emails: mj_lalehchini@yahoo.com, Corresponding author: lotfitaher@yahoo.com

Abstract

The primary goal of this work is to introduce two Steffensen-like adaptive with memory methods with the highest efficiency indices. In the existing methods, to improve the convergence order applying the with memory concept, it has only been focused on the current and previous iterations. However, it is possible to improve the accelerators, considering date from the first to the current iterations. Therefore, we achieve superior convergence orders and obtain as high as possible efficiency indices. These are the main contributions of this work. *Key words: Nonlinear equations, iterative methods, Steffensen-like method, with memory methods, adaptive methods.*

1 Introduction

One of the most important subjects in developing numerical algorithms is to establish optimal algorithms with economic complexity. For example, developing iterative methods for approximating zero(s) of a given nonlinear equation falls within this matter, and many studies have been devoted to it [3, 2]. Inspired by this, we will set up two adaptive Steffensen-like with memory methods in which they are improvement of existing methods. To our knowledge, these kinds of adaptive methods have not been studied in the literature.

Truab developed the first method with memory from Steffensen's method [4] as following [2]:

$$\begin{cases} w_k = x_k + \gamma_k\, f(x_k), \\ x_{k+1} = x_k - \frac{f(x_k)}{f[x_k, w_k]}, \qquad k = 0, 1, 2, \cdots, \\ \gamma_{k+1} = -\frac{1}{N_1'(x_{k+1})}, \end{cases} \tag{1.1}$$

 ISBN: 978-84-617-8694-7

where x_0 and γ_0 are given initially suitable, and $N_1(t) = f(x_{k+1}) + (t - x_{k+1})f[x_{k+1}, x_k]$ is the linear Newton's interpolation. The convergence order of the with memory method (1.1) is $1 + \sqrt{2} \approx 2.414$. Also, Džunnić and Petković improved Traub's idea, introducing a better accelerator [3]:

$$\begin{cases} w_k = x_k + \gamma_k f(x_k), \\ x_{k+1} = x_k - \frac{f(x_k)}{f[x_k, w_k]}, \qquad k = 0, 1, 2, \cdots, \\ \gamma_{k+1} = -\frac{1}{N_2'(x_{k+1})}, \end{cases} \tag{1.2}$$

where x_0 and γ_0 are given initially suitable, and $N_2(t)$ is the Netwon's interpolation polynomial given by

$$N_2(t) = f(x_{k+1}) + (t - x_{k+1})f[x_{k+1}, w_k] + (t - x_{k+1})(t - w_k)f[x_{k+1}, w_k, x_k].$$

The convergence order of the with memory method (1.2) is $1 + \sqrt{4} = 3$. Moreover, Džunić added another parameter to the Steffensen's method and obtained a more efficient with memory method [3]:

$$\begin{cases} x_{k+1} = x_k - \frac{f(x_k)}{f[x_k, w_k] + \lambda_k f(w_k)}, \qquad k = 0, 1, 2, \cdots, \\ \gamma_{k+1} = -\frac{1}{N_2'(x_{k+1})}, \\ w_{k+1} = x_{k+1} + \gamma_{k+1} f(x_{k+1}), \\ \lambda_{k+1} = \frac{-N_3''(w_{k+1})}{2N_3'(w_{k+1})}, \end{cases} \tag{1.3}$$

where x_0, γ_0, and λ_0 are given initially suitable. This method has convergence order $\frac{3+\sqrt{17}}{2} \approx 3.56$.

Remark 1.1. *If γ and λ are constants, then the methods (1.2) and (1.3) are without memory methods with convergence order two with the following error equations, respectively:*

$$e_{k+1} = c_2\big(1 + \gamma f'(\alpha)\big)e_k^2 + O(e_k^3), \tag{1.4}$$

and

$$e_{k+1} = \big(c_2 + \lambda\big)\big(1 + \gamma f'(\alpha)\big)e_k^2 + O(e_k^3). \tag{1.5}$$

In this work, we will attempt to carry two adaptive with memory methods out from (1.2) and (1.3) which are superior to them. To this purpose, we first update the accelerator γ_k in each iteration using available data not only from the current and previous iterations, but also from the current and *all* previous iterations (adaptive concept). We prove that this method has convergence order 3.4 using the same function evaluations as (1.2), so its efficiency index is much better as well. Similarly, we derive another adaptive with memory for (1.3) which acquires convergence order 3.9 using the same functional evaluations. Therefore, this method is better not only than our adaptive method with one accelerator, but also than all the existing methods.

 ISBN: 978-84-617-8694-7

2 Developing adaptive with memory methods

This section deals with two new adaptive methods with memory. To this end, we modify and extend methods (1.2) and (1.3) in such a way that they consider all previous information to attain as high as possible convergence order without any new functional evaluation. In this manner we use the adaptive idea which has not been considered to our best knowledge.

2.1 Mono accelerator adaptive with memory method

In (1.2), to update accelerator γ_k in each iteration, we only use the information from the current and previous iterations and reach the convergence order 3. However, as the procedure goes ahead it is possible to use the old information not only from the current and previous step, but also from the current and all the previous steps. In other words, we wish to apply the adaptive idea to construct the new methods with memory. Accordingly, we introduce the following new adaptive method with memory

$$\begin{cases} w_k = x_k + \gamma_k f(x_k), \\ x_{k+1} = x_k - \frac{f(x_k)}{f[x_k, w_k]}, \\ \gamma_{k+1} = -\frac{1}{N'_{2k+2}(x_{k+1})}, \end{cases} \qquad k = 0, 1, 2, \cdots, \tag{2.1}$$

where x_0 and γ_0 are given initially suitable, and $N_{2k+2}(t)$ is Netwon's interpolation polynomial of degree $2k+2$ at the points $x_{k+1}, w_k, x_k, \ldots, w_0, x_0$.

Referring to the Error Equation (1.4), it is observed that if $1 + \gamma f'(\alpha) = 0$, then convergence order of the method without memory (2.1), for a moment suppose γ_k is fix, increases. Since α is unknown, we cannot suppose $\gamma = -1/f'(\alpha)$. Even if we assumed that α was known, we could not use it to evaluate $f'(\alpha)$, since it increases the functional evaluation, and optimality of the methods is destroyed. It is assumed that the sequence $\{x_k\}$ converges to α. Moreover, f' is at least continuous, so $\lim f'(x_k) = f'(\alpha)$ as $k \to \infty$. Thus, we can use $N'_{2k+2}(x_k)$ instead of $f'(x_k)$ to our mission, i.e., $\gamma_k = -1/N'_{2k+2}(x_k)$. To discuss the convergence order of (2.1), we need:

Lemma 2.1. *If* $\gamma_k = -1/N'_{2k+2}(x_k)$, *and* $\lambda_{k+1} = \frac{-N''_{2k+2}(w_{k+1})}{2N'_{2k+2}(w_{k+1})}$ *then*

$$1 + \gamma f'(\alpha) \sim \prod_{i=0}^{k-1} e_{w,i} e_i, \quad c_2 + \lambda \sim \prod_{i=0}^{k-1} e_{w,i} e_i, \tag{2.2}$$

where $e_i = x_i - \alpha$ *and* $e_{w,i} = w_i - \alpha$.

Theorem 2.2. *Let the initial approximation* x_0 *be sufficiently close to the zero* α *of* f, R *and* p *denote the convergence order of the sequences* $\{x_k\}$ *and* $\{w_k\}$, *respectively, generated*

 ISBN: 978-84-617-8694-7

by the adoptive method with memory (2.1). Then, we have

$$\begin{cases} R^k p - R^k - (p+1)\sum_{i=0}^{k-1} R^i = 0, \\ R^{k+1} - R^k - (p+1)\sum_{i=0}^{k-1} R^i = 0. \end{cases} \tag{2.3}$$

Proof. We can assume

$$e_{k+1} \sim e_k^R. \tag{2.4}$$

Hence,

$$e_{k+1} \sim (e_{k-1}^R)^R = e_{k-1}^{R^2}. \tag{2.5}$$

Inductively,

$$e_{k+1} \sim e_0^{R^{k+1}}. \tag{2.6}$$

Similarly, we have

$$e_{w,k} \sim e_k^p = (e_{k-1}^R)^p = e_{k-1}^{Rp}. \tag{2.7}$$

Thus,

$$e_{w,k} \sim e_0^{R^k p}. \tag{2.8}$$

By (2.4) and (2.7), Lemma 2.1 results

$$1 + \gamma f'(\alpha) \sim e_0^{(p+1)\sum_{i=0}^{k-1} R^i}. \tag{2.9}$$

On the other hand, since $e_{w,k} \sim (1 + \gamma f'(\alpha))e_k$ and $e_{k+1} \sim (1 + \gamma f'(\alpha))e_k^2$, taking into account (2.9), then

$$e_{w,k} \sim e_0^{R^k + (p+1)\sum_{i=0}^{k-1} R^i}, \tag{2.10}$$

and

$$e_{k+1} \sim e_0^{2R^k + (p+1)\sum_{i=0}^{k-1} R^i}. \tag{2.11}$$

From (2.8)-(2.10) and (2.6)-(2.11), we conclude that

$$e_0^{R^k p} \sim e_0^{R^k + (p+1)\sum_{i=0}^{k-1} R^i}, \tag{2.12}$$

$$e_0^{R^{k+1}} \sim e_0^{R^k + (p+1)\sum_{i=0}^{k-1} R^i}. \tag{2.13}$$

Consequently,

$$\begin{cases} R^k p - R^k - (p+1)\sum_{i=0}^{k-1} R^i = 0, \\ R^{k+1} - R^k - (p+1)\sum_{i=0}^{k-1} R^i = 0. \end{cases} \tag{2.14}$$

 ISBN: 978-84-617-8694-7

2.2 Bi accelerators adaptive with memory method

We now introduce bi accelerators adaptive with memory method. Since much of the details are similar to the descriptions of (2.1) we confine ourselves to repeat them.

$$\begin{cases} x_{k+1} = x_k - \frac{f(x_k)}{f[x_k,w_k]+\lambda_k f(w_k)}, & k = 0,1,2,\cdots, \\ \gamma_{k+1} = -\frac{1}{N'_{2k+2}(x_{k+1})}, \\ w_{k+1} = x_{k+1} + \gamma_{k+1} f(x_{k+1}), \\ \lambda_{k+1} = \frac{-N''_{2k+3}(w_{k+1})}{2N'_{2k+3}(w_{k+1})}, \end{cases} \tag{2.15}$$

Also, we have

Theorem 2.3. *Let the initial approximation x_0 be sufficiently close to the zero α of f, R and p denote the convergence order of the sequences $\{x_k\}$ and $\{w_k\}$, respectively, generated by the adoptive method with memory (2.15). Then, we have*

$$\begin{cases} R^k p - R^k - (p+1)\sum_{i=0}^{k-1} R^i = 0, \\ R^{k+1} - 2R^k - 2(p+1)\sum_{i=0}^{k-1} R^i = 0. \end{cases} \tag{2.16}$$

The proof is similar in spirit to the Theorem 2.2.

References

[1] A.M. Ostrowski, *Solution of equations and systems of equations*, , Prentice-Hall, Englewood Cliffs, NJ, USA, 1964.

[2] J.F. Traub, *Iterative Methods for the Solution of Equations*, Prentice Hall, New York, 1964.

[3] J. Dzunic and M.S. Petkovic, *A cubically convergent steffensen-like method for solving nonlinear equations*, ,Applied Mathematics Letters, **25** (2012) 1881–1886.

[4] I.F. Steffensen, *Remarks on iteration*, Aktuarietidskr, **16** (1933) 64–79.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

Double-pendulum with both-sided stops simulation analysis

Marek Lampart[1,2] and Jaroslav Zapoměl[3,4]

[1] *IT4Innovations, VŠB - Technical University of Ostrava,*
17. listopadu 15/2172, 708 33 Ostrava, Czech Republic.

[2] *Department of Applied Mathematics, VŠB - Technical University of Ostrava,*
17. listopadu 15/2172, 708 33 Ostrava, Czech Republic.

[3] *Department of Applied Mechanics, VŠB - Technical University of Ostrava,*
17. listopadu 15/2172, 708 33 Ostrava, Czech Republic.

[4] *Department of Dynamics and Vibrations,*
Institute of Thermomechanics of the CAS, v.v.i.,
Dolejškova 1402/5, 182 00 Prague 8, Czech Republic.

emails: marek.lampart@vsb.cz, jaroslav.zapomel@vsb.cz

Abstract

This research was motivated by a real technological problem of vibrations of bodies hanging on chains or ropes in tubes or spaces limited by walls or other bodies. The system has two degrees of freedom. Its movement is governed by a set of nonlinear ordinary differential equations. As the main result it is shown that the system exhibits regular, irregular and chaotic patterns for suitable choice of parameters.

Key words: mechanical model, chaos tests, bifurcation, vibration
MSC 2000: 34H20, 34H10, 37N30

1 Introduction

The research area of multi-pendulum systems are widely studied by many authors form different point of view under several motivations. In [3] a control system combining input shaping and feedback is developed for double-pendulum systems subjected to external disturbances. The authors in [2] proposed a parametrically excited pendulum with irrational

nonlinearity which comprises a simple pendulum linked by a linear spring under base excitation. The dynamics of hybrid systems and analytical dynamics of discrete material particle system containing creep elements described by fractional order derivatives, is presented in [4] (see references therein for more literature about dynamics of multi-pendulum systems and solutions of the system equations).

The main aim of the present paper is to derive equations of motions of the double-pendulum with both sided walls where collisions of the pendulums occur. The pendulums movement analysis is performed in dependence of the sliding motion excitation frequency; movements are showing periodic and also chaotic patterns that are supported by bifurcation diagrams and 0-1 test for chaos.

2 Mathematical model of the system

The investigated system consists of a driving body and of the upper and lower pendulums see Fig. 1). The driving body performs sliding motion in the horizontal direction which excites the movement of both pendulums. The motion of the pendulums is limited by two vertical walls. All bodies included the walls are considered as absolutely rigid, some elasticity is taken into account in the contact area between the pendulums and the walls.

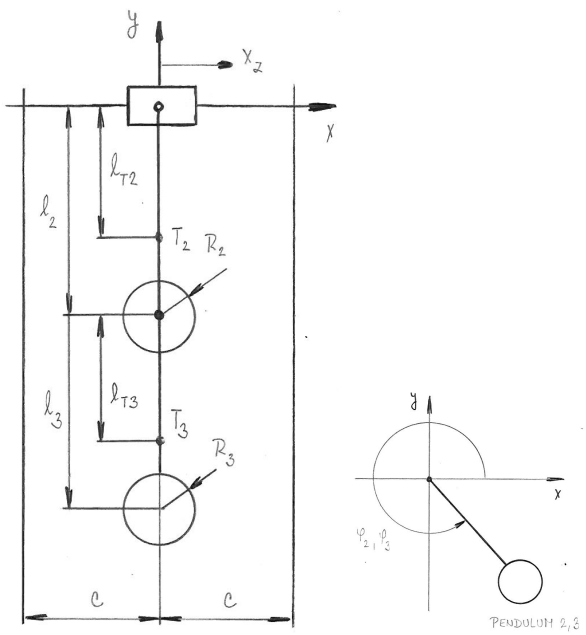


Figure 1: Model of investigated system

 ISBN: 978-84-617-8694-7

The task of the present research is to investigate the influence of the systems parameters, especially the frequency of kinematic excitation on a character of the double-pendulum movement. The investigated system has two degrees of freedom. Its instantaneous position is defined by two generalized coordinates:

φ_2 - angel of rotation of the upper pendulum,

φ_3 - angel of rotation of the lower pendulum.

The equations of motion have been derived using the Lagrange equations of the second kind:

$$\begin{aligned}(J_{T2}+m_2l_{T2}^2+m_3l_2^2)\ddot{\varphi}_2+A\ddot{\varphi}_3 &= Q_2+B\dot{\varphi}_3^2-m_2gl_{T2}\cos\varphi_2-m_3gl_2\cos\varphi_3+\\ &\quad +m_2\ddot{x}_zl_{T2}\sin\varphi_2+m_3\ddot{x}_zl_2\sin\varphi_2\\ (J_{T3}+m_3l_{T3}^2)\ddot{\varphi}_3+A\ddot{\varphi}_2 &= Q_3-B\dot{\varphi}_2^2+\\ &\quad +m_3\ddot{x}_zl_{T3}\sin\varphi_3-m_3gl_{T3}\cos\varphi_3\end{aligned} \tag{1}$$

where

$$A = m_3l_2l_{T3}(\sin\varphi_2\sin\varphi_3+\cos\varphi_2\cos\varphi_3) \tag{2}$$

$$B = m_3l_2l_{T3}(\cos\varphi_2\sin\varphi_3-\sin\varphi_2\cos\varphi_3) \tag{3}$$

$$Q_2 = -(F_{R2}+F_{R3})l_2\sin\varphi_2+Q_{D2} \tag{4}$$

$$Q_3 = -F_{R3}l_3\sin\varphi_3+Q_{D3} \tag{5}$$

$$Q_{D2} = (b_2+b_3)\dot{x}_zl_2\sin(\varphi_2)-(b_2+b_3)l_2^2\dot{\varphi}_2-b_3l_2l_3(\sin\varphi_2\sin\varphi_3+\cos\varphi_2\cos\varphi_3)\dot{\varphi}_2 \tag{6}$$

$$Q_{D3} = b_3\dot{x}_zl_3\sin(\varphi_3)-b_3l_3^2\dot{\varphi}_3-b_3l_2l_3(\sin\varphi_2\sin\varphi_3+\cos\varphi_2\cos\varphi_3)\dot{\varphi}_2 \tag{7}$$

here $(\dot{})$ and $(\ddot{})$ denote the first and second derivative with respect to time, respectively.

The impact forces F_{R2} and F_{R3} are defined as follows:

$$F_{R2}=\begin{cases}\text{if } x_{R2M}<-c & \text{then } F_{R2}=-k_c(x_{R2M}+c),\\ \text{if } x_{R2P}>c & \text{then } F_{R2}=-k_c(x_{R2P}-c),\\ \text{else } F_{R2}=0,\end{cases} \tag{8}$$

$$F_{R3}=\begin{cases}\text{if } x_{R3M}<-c & \text{then } F_{R3}=-k_c(x_{R3M}+c),\\ \text{if } x_{R3P}>c & \text{then } F_{R3}=-k_c(x_{R3P}-c),\\ \text{else } F_{R3}=0,\end{cases} \tag{9}$$

where

$$x_{R2M} = x_z-r_2+l_2\cos\varphi_2, \tag{10}$$

$$x_{R2P} = x_z+r_2+l_2\cos\varphi_2, \tag{11}$$

$$x_{R3M} = x_z-r_3+l_2\cos\varphi_2+l_3\cos\varphi_3, \tag{12}$$

$$x_{R3P} = x_z+r_3+l_2\cos\varphi_2+l_3\cos\varphi_3. \tag{13}$$

 ISBN: 978-84-617-8694-7

The horizontal position position of the driving body is given by

$$x_z(t) = A\,(1 - \mathrm{e}^{-\alpha t})\sin(\omega t) \tag{14}$$

where A is the amplitude of the kinematic excitation of the upper pendulum, α the run up coefficient and ω stands for the excitation frequency. All system parameters are summarized in Table 1.

At the beginning the system takes the equilibrium position, which corresponds to the initial conditions:

$$\begin{aligned}
\varphi_2(0) &= 3/2\,\pi, && (15)\\
\varphi_3(0) &= 3/2\,\pi, && (16)\\
\dot{\varphi}_2(0) &= 0, && (17)\\
\dot{\varphi}_3(0) &= 0. && (18)\\
& && (19)
\end{aligned}$$

3 Main results

Main results were reached by numerical simulations of (1) for system parameters summarized in Table 1 where the excitation frequency ω was changed form 1 rad s^{-1} to 100 rad s^{-1}.

The character of movement of the double-pendulum system is analyzed in detail. The first observations are shown in bifurcation diagrams Figs. 2, where periodic as well as chaotic movements are visible for suitable choices of ω. Here, x_2 and x_3 stand for the horizontal displacements of the upper and lower pendulum ball centers, respectively. In this figures it is also marked that for the excitation frequencies $\omega \in [1, 10] \cup \{25,\ 28,\ 29\}$ there are no contacts of the upper pendulum with the wall (marked green) and for $\omega \in [11, 24] \cup \{26,\ 27\} \cup [30, 100]$ collision of the upper pendulum with the wall occurs; and for the excitation frequencies $\omega \in [1, 10]$ there are no contacts of the lower pendulum with the wall (marked green) and for $\omega \in [11, 100]$ collision of the upper pendulum with the wall occurs.

It is visible form phase portraits (see Figs. 3, 4 and 5) that the movement is periodic and also chaotic and its character is changing with increasing excitation frequency.

The output parameter of the 0-1 test for chaos can acquire only one of the values 0 or 1 which correspond to the regular and chaotic motions, respectively. More details can be found in [1]. The results of the 0-1 test for the range of investigated frequencies are shown in Fig. 6 where chaotic and non-chaotic movements were detected. Note, that the result of 0-1 test coincides with bifurcation diagrams in Fig. 2 and with phase diagrams, for special choices of the excitation frequencies shown in Figs. 3, 4 and 5.

 ISBN: 978-84-617-8694-7

Table 1: Parameters of the system (1).

quantity	value	description
m_2	1 kg	mass of the upper pendulum
m_3	1 kg	mass of the lower pendulum
J_{T2}	0.1 kg m^2	moment of inertia of the upper pendulum referred to its centre of gravity
J_{T3}	0.2 kg m^2	moment of inertia of the lower pendulum referred to its centre of gravity
l_2	1 m	length of the upper pendulum
l_{T2}	0.8 m	length of the upper pendulum (rotational joint - centre of gravity)
r_2	0.1 m	radius of the ball of the upper pendulum
l_3	1.2 m	length of the lower pendulum
l_{T3}	0.9 m	length of the lower pendulum (rotational joint - centre of gravity)
r_3	0.1 m	radius of the ball of the lower pendulum
k_c	2×10^6 N m^{-1}	the contact stiffness
c	0.4 m	the half width of the clearance
g	9.80665 m s^{-2}	the gravity acceleration
A	0.1 m	the amplitude of the kinematic excitation of the upper pendulum
α	500 s^{-1}	the run up coefficient
b_2	1 Nms rad^{-1}	the damping coefficient of the upper pendulum
b_3	1.5 Nms rad^{-1}	the damping coefficient of the lower pendulum

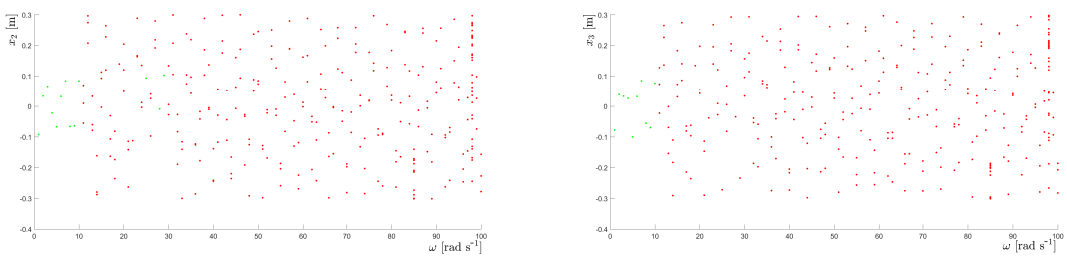


Figure 2: Bifurcation diagrams of x_2 (Left) and x_3 (Right) in dependence on ω.

 ISBN: 978-84-617-8694-7

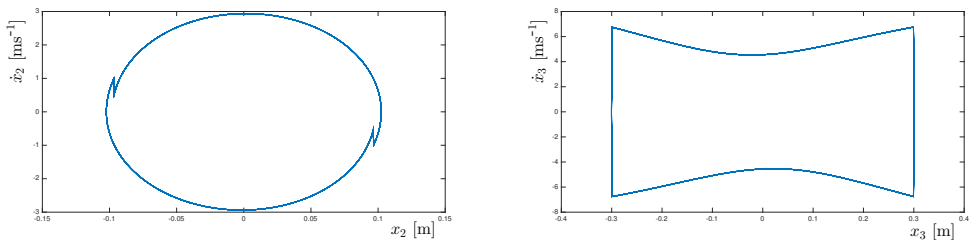


Figure 3: Phase portraits x_2 versus $\dot{x}_2$ (Left) and x_3 versus $\dot{x}_3$ (Right) for $\omega = 28$ rad s^{-1}.

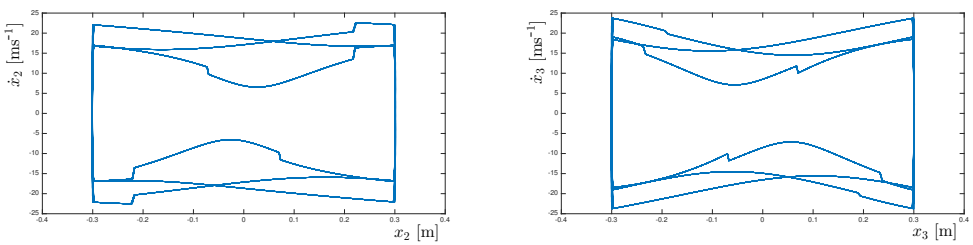


Figure 4: Phase portraits x_2 versus $\dot{x}_2$ (Left) and x_3 versus $\dot{x}_3$ (Right) for $\omega = 78$ rad s^{-1}.

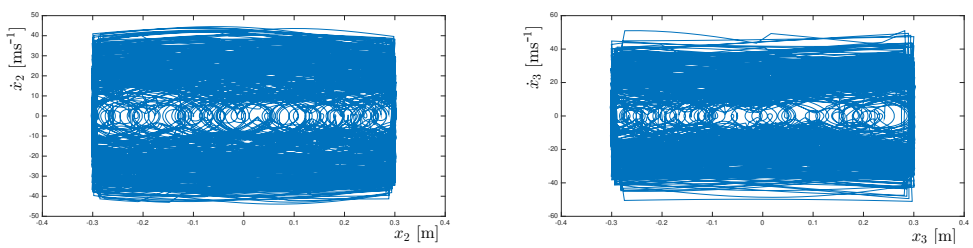


Figure 5: Phase portraits x_2 versus $\dot{x}_2$ (Left) and x_3 versus $\dot{x}_3$ (Right) for $\omega = 98$ rad s^{-1}.

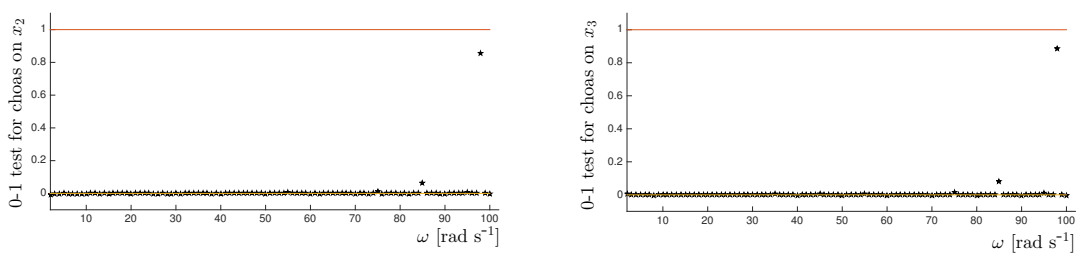


Figure 6: Output of the 0-1 test for chaos of x_2 (Left) and x_3 (Right) in dependence on ω.

 ISBN: 978-84-617-8694-7

4 Conclusions

In this paper, a mechanical system with two degrees of freedom has been investigated and the movement of the double-pendulum with both sides walls was analyzed. This model was inspired by a real problem of bodies hanging on chains or ropes in tubes or spaces limited by walls or stops.

The equation of movement was solved numerically using Runge-Kutta method implemented as *ode45* solver in Matlab.

It was observed that the movement is showing regular and also chaotic patterns for suitable choice of parameters, mainly excitation frequency of the driving body played a key role here. For this purpose 0-1 test for chaos and bifurcation diagrams were used.

Acknowledgements

This work was supported by The Ministry of Education, Youth and Sports from the National Programme of Sustainability (NPU II) project "IT4Innovations excellence in science - LQ1602"; by The Ministry of Education, Youth and Sports from the Large Infrastructures for Research, Experimental Development and Innovations project IT4Innovations National Supercomputing Center LM2015070; by grant SGS No. SP2017/122, VŠB - Technical University of Ostrava, Czech Republic; and by the Czech Science Foundation, Grant No. 15-06621S.

References

[1] G. A Gottwald, I. Melbourne, *A new test for chaos in deterministic systems*, Proc. R. Soc. London A **460** (2004) 603–611.

[2] N. Han, Q. Cao, *A parametrically excited pendulum with irrational nonlinearity*, International Journal of NonLinear Mechanics **88** (2017) 122–134.

[3] R. Mar, A. Goyal, V. Nguyen, T. Yang, W. Singhose, *Combined input shaping and feedback control for double-pendulum systems*, Mechanical Systemsand Signal Processing **85** (2017) 267–277.

[4] Katica R. (Stevanovi) Hedrih, *Dynamics of multi-pendulum systems with fractional order creep elements*, Journal of Theoretical and Applied Mechanics **43** (2018) 483–509.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

Branching pieces of rational skins from polynomial MOS patches

Miroslav Lávička[1,2] and Michal Bizzarri[1]

[1] *New Technologies for the Information Society, Faculty of Applied Sciences, University of West Bohemia, Univerzitní 8, 306 14 Plzeň, Czech Republic*

[2] *Department of mathematics, Faculty of Applied Sciences, University of West Bohemia, Univerzitní 8, 306 14 Plzeň, Czech Republic*

emails: lavicka@kma.zcu.cz, bizzarri@ntis.zcu.cz

Abstract

In this paper we will investigate one certain application of polynomial 2-surfaces possessing the polynomial area element in the Minkowski space $\mathbb{R}^{3,1}$, where they coincide with the so called MOS surfaces (i.e., medial surface transforms with rational domain boundaries). We formulate an efficient algorithm for Hermite interpolation by MOS surfaces and apply the developed method to the construction of branching pieces which occur during the operation of rational skinning. We recall that when branched skins of systems of spheres are constructed then the envelopes of suitable two-parametric systems of spheres must be considered. MOS surfaces are presented as especially suitable candidates for modelling these shapes because they provide not only rational envelopes but also all offsets of these envelopes are rational.

Key words: Medial surface transforms; MOS surfaces; rational envelopes; skinning

1 Introduction

We continue with the investigation of modelling parts of branched skins using special medial surface transforms yielding rational envelope surfaces. We recall that the operation of skinning can be considered as a certain analogy to the well-known interpolation of point data sets, which is a problem often solved in geometric modelling, see [7]. For us it is a construction of a G^k/C^k continuous interpolation surface of an ordered sequence of spatial shapes, in particular of spheres/balls. Due to its technical importance, skinning has

 ISBN: 978-84-617-8694-7

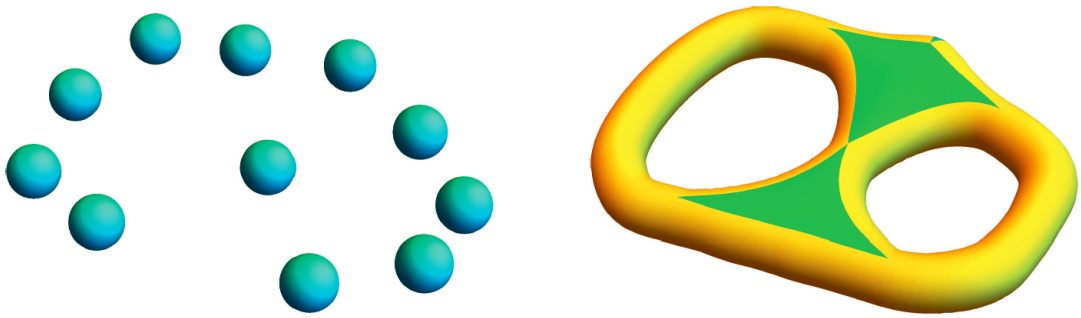


Figure 1: Branched skin of a system of spheres Σ in 3D containing a triangular and quadrangular branching element (green).

attracted the researchers in recent years and one can find a growing number of articles on this subject, see e.g. [14, 9, 1].

The problem of skinning can be identified in various situations, maybe the best known is the application in computer animation. Computing skins of a system of spheres/balls in 3D appears further for instance in the area of computational chemistry or molecular biology when surface meshes for molecular models are supposed to be generated. In a discrete sense, skinning can be considered as a part of the problem of computing envelopes of families of spheres using the cyclographic mapping [12, 13]. Skinning is closely related to representing shapes with the help of the associated medial axis/surface transforms [4, 11] and the theory of canal surfaces [10, 6].

When constructing the skin of spheres in the linear configuration then finding the solution is relatively mastered. However the problem is more complicated when branching of skins is allowed. Branching in this sense means that there exists a sphere which is connected via the skin with more than two neighbouring spheres. In these scenarios, the rational envelopes of suitable two-parameter systems of spheres must be thoroughly investigated and functional algorithms for interpolations with them must be designed. By utilizing polynomial MOS surfaces, we present a simple method for computing pieces of rational branched skins of a system of spheres/balls in 3D. We present MOS surfaces as very suitable candidates for modelling these parts as they give not only rational envelopes but also rational offsets of these envelopes.

2 Branched skins of systems of spheres

Using the method from [9], we consider a given sequence of spheres $\Sigma = \{S_1, S_2, \ldots, S_n\}$, admissible as defined in [1]. The goal is to construct a smooth spline surface $\mathcal{S}(\Sigma)$ skinning

 ISBN: 978-84-617-8694-7

this system. We will admit not only linear sequences of input spheres but also more complicated cases. In particular, we plan to focus on configurations when branched skins are designed and when special branching elements shall be constructed, cf. Fig. 1.

If we construct the skin of spheres in some linear configuration, $\mathcal{S}(\Sigma)$ consists of the following elements: (i) parts of S_i obtained as the differences of S_i and the spherical caps determined by the contact circles; (ii) surfaces smoothly joining two consecutive spheres S_{i-1} and S_i along prescribed contact circles, e.g. parts of canal surfaces, see Fig. 1 and we refer also to [2] for more details. These scenarios are quite sufficiently solved in the literature. However, the problem is significantly more complicated when branching of skins is allowed and new challenges dealing with new types of elements must be discussed.

When branching of skins occurs then two types of situations must be discussed. Either, branching takes place on a particular sphere and then the skin consists again of the elements of type (i) and (ii) only. Or, a joining part between more than two spheres is needed then an element of a new type must be constructed. In this case, the rational envelopes of suitable two-parameter systems of spheres must be investigated in more detail. We recall that this problem was suitably solved in [2] devoted to study of rational envelope (RE) surfaces, where formulated efficient algorithms for G^1 data interpolation using RE surfaces were applied also to rational skinning. Using the recent results from [3] we present polynomial MOS surfaces as shapes especially suitable for modelling rational branching elements. The extra-feature of these newly constructed elements is that they guarantee also the rationality of their offsets. This is a main contribution of the designed method.

3 Medial surface transforms yielding rational envelopes

Let be given a spatial domain $\Omega \subset \mathbb{R}^3$ and the family of all inscribed spheres partially ordered with respect to inclusion of the associated balls. The *medial surface* (MS) of Ω is the set of all centers $(x,y,z)^\top \in \mathbb{R}^3$ of maximal inscribed spheres and the *medial surface transform* (MST) of Ω is obtained by appending the corresponding sphere radius r to the MS, i.e., it consists of the points $(x,y,z,r)^\top$ in the four-dimensional Minkowski space $\mathbb{R}^{3,1}$.

Let a medial surface transform $\mathbf{x}(u,v) = (x,y,z,r)^\top \subset \mathbb{R}^{3,1}$ be given. If we denote by $\hat{\mathbf{x}}(u,v) = (x,y,z)^\top$ the corresponding medial surface in $\mathbb{R}^3$ then the *envelope formula* is

$$\mathbf{b}^{\pm}(u,v) = \hat{\mathbf{x}}(u,v) - r\mathbf{n}^{\pm}(u,v), \tag{1}$$

with

$$\mathbf{n}^{\pm} = \frac{1}{\hat{E}\hat{G}-\hat{F}^2}\left[\left(\frac{\partial r}{\partial u}\hat{G}-\frac{\partial r}{\partial v}\hat{F}\right)\hat{\mathbf{x}}_u + \left(\frac{\partial r}{\partial v}\hat{E}-\frac{\partial r}{\partial u}\hat{F}\right)\hat{\mathbf{x}}_v \mp \sqrt{EG-F^2}(\hat{\mathbf{x}}_u\times\hat{\mathbf{x}}_v)\right], \tag{2}$$

where $\mathbf{n}^{\pm}$ is a unit vector perpendicular to $\mathbf{b}^{\pm}$. The components E,F,G of the first fundamental form of $\mathbf{x}(u,v)$ are computed using the indefinite Minkowski inner product $\langle \cdot,\cdot \rangle$

 ISBN: 978-84-617-8694-7

with the signature (3,1), whereas the components $\hat{E}, \hat{F}, \hat{G}$ of the first fundamental form of $\hat{\mathbf{x}}(u,v)$ are determined using the standard Euclidean inner product in $\mathbb{R}^3$.

MOS surfaces, i.e., Medial surfaces Obeying the Sum of squares condition, were introduced in [8] as the shapes satisfying the distinguishing property that if considered as an MST of a spatial domain, then the associated envelope and its offsets admit exact rational parameterization. So, they are given by the condition

$$EG - F^2 = \sigma^2(u,v), \qquad \text{where } \sigma(u,v) \in \mathbb{R}(u,v), \tag{3}$$

which guarantees the rationality of (2) and thus of the envelope $\mathbf{b}^{\pm}(u,v)$.

From this it is evident that MOS surfaces are simultaneously surfaces with rational *area element* in $\mathbb{R}^{3,1}$ as the squared area element of the parametric surface $\mathbf{x}(u,v)$ has the form

$$\mathrm{d}A^2 = \begin{vmatrix} \langle \mathbf{x}_u, \mathbf{x}_u \rangle & \langle \mathbf{x}_u, \mathbf{x}_v \rangle \\ \langle \mathbf{x}_u, \mathbf{x}_v \rangle & \langle \mathbf{x}_v, \mathbf{x}_v \rangle \end{vmatrix} \mathrm{d}u^2 \mathrm{d}v^2 = (EG - F^2)\, \mathrm{d}u^2 \mathrm{d}v^2. \tag{4}$$

Moreover, all polynomial MOS surfaces possess polynomial surface area $A(u,v) = \iint \sqrt{EG - F^2}\, \mathrm{d}u\mathrm{d}v$. This offers many useful computational advantages compared to other surfaces.

In what follows, we will study suitable normal vector fields that help us to construct parameterizations of polynomial MOS surfaces. We recall a result from [3] expressing that for a polynomial surface $\mathbf{x}(u,v)$ in $\mathbb{R}^{3,1}$ it holds

$$\Gamma(\mathbf{x}_u, \mathbf{x}_v) = f^2\, \Gamma(\mathbf{n}_1, \mathbf{n}_2), \tag{5}$$

where $\mathbf{n}_1, \mathbf{n}_2$ are the normal vectors generating the normal space $\mathrm{span}\{\mathbf{n}_1(u,v), \mathbf{n}_2(u,v)\}$, Γ is the the Grammian of the considered vectors and $f(u,v) \in \mathbb{R}(u,v)$ is a non-zero factor. This means that it is possible to start with suitable normal vectors when constructing parameterizations of polynomial MOS surfaces as the distinguishing condition (3) depends on $\Gamma(\mathbf{n}_1, \mathbf{n}_2)$ equally as on $\Gamma(\mathbf{x}_u, \mathbf{x}_v) = EG - F^2$. Moreover, when at least one of the normal vectors $\mathbf{n}_1$, or $\mathbf{n}_2$ is isotropic, i.e., its squared norm is zero, then $\Gamma(\mathbf{n}_1, \mathbf{n}_2)$ is automatically a perfect square.

Hence we can recall the main ideas from the approach discussed in [3]. We start with the normal space $\mathrm{span}\{\mathbf{n}^+(u,v), \mathbf{n}^-(u,v)\}$ given by the polynomial isotropic vectors of degree k, i.e., $\langle \mathbf{n}^{\pm}, \mathbf{n}^{\pm} \rangle \equiv 0$. Their parameterizations can be obtained e.g. from polynomial Pythagorean quadruples, see [5]. To find an associated polynomial MOS surface of degree $\ell + 1$, it is necessary to find suitable polynomial vector fields

$$\begin{aligned} \mathbf{q}(u,v) &= \left(\sum_{i+j \le \ell} q_{1ij} u^i v^j, \sum_{i+j \le \ell} q_{2ij} u^i v^j, \sum_{i+j \le \ell} q_{3ij} u^i v^j, \sum_{i+j \le \ell} q_{4ij} u^i v^j \right)^{\top}, \\ \mathbf{r}(u,v) &= \left(\sum_{i+j \le \ell} r_{1ij} u^i v^j, \sum_{i+j \le \ell} r_{2ij} u^i v^j, \sum_{i+j \le \ell} r_{3ij} u^i v^j \sum_{i+j \le \ell} r_{4ij} u^i v^j \right)^{\top}, \end{aligned} \tag{6}$$

 ISBN: 978-84-617-8694-7

which will play the role of $\mathbf{x}_u$, $\mathbf{x}_v$, respectively. Thus, $\mathbf{q}$, $\mathbf{r}$ must satisfy the following conditions

$$\begin{aligned} \langle \mathbf{q}, \mathbf{n}^{\pm} \rangle &\equiv 0, \\ \langle \mathbf{r}, \mathbf{n}^{\pm} \rangle &\equiv 0, \\ \frac{\partial \mathbf{q}}{\partial v} - \frac{\partial \mathbf{r}}{\partial u} &\equiv 0, \end{aligned} \tag{7}$$

where the last equation expresses the condition of integrability. For ℓ large enough, system of linear equations (7) with unknowns $q_{1ij}, q_{2ij}, q_{3ij}, q_{4ij}, r_{1ij}, r_{2ij}, r_{3ij}, r_{4ij}$ possesses a solution and thus we arrive at a MOS parameterization $\mathbf{x}(u,v)$, for which it holds $EG - F^2 = f(u,v)^2 \Gamma(\mathbf{n}^+, \mathbf{n}^-)$, where $f(u,v)$ is a factor relates suitably the degrees of $\mathbf{n}^{\pm}$ and $\mathbf{x}$.

4 MOS branching elements for skinning of spheres

The G^1 Hermite interpolations by polynomial MOS surfaces can be efficiently used when branching elements of rational skinning surfaces shall be constructed as the rational envelopes of two-parameter families of spheres. We present the approach at least for triangular and quadrilateral patches.

Consider three or four points $\mathbf{p}_{ij} \in \mathbb{R}^{3,1}$, $i,j = 0,1$ (and $i+j < 2$ for the case of three points), and three or four associated tangent planes τ_{ij} determined by the vectors $\mathbf{t}_{ij,1}$ and $\mathbf{t}_{ij,2}$. A choice of these points and associated tangent planes reflects a particular situation obtained when the branching skins shall be constructed. Following the previously presented ideas, we can formulate the whole algorithm consisting of three subparts:

(i) We find the isotropic vectors $\mathbf{n}_{ij}^{\pm}$ lying in the normal planes orthogonal to the given tangent planes τ_{ij}

(ii) We construct a normal vector field $\mathbf{n}^+(u,v)$ interpolating data $\lambda_{ij}\, \mathbf{n}_{ij}^+$ and having the polynomial norm.

(iii) We compute a polynomial patch interpolating the points $\mathbf{p}_{ij}$, possessing the isotropic normal vector field $\mathbf{n}^+(u,v)$ and in addition having $\mathbf{n}_{ij}^-$ as normal vectors at $\mathbf{p}_{ij}$.

As concerns step (i), we consider the projective closure of $\mathbb{R}^{3,1}$ consisting of the points with the homogeneous coordinates $(x_0 : x_1 : x_2 : x_3 : x_4)$, and the equation $x_0 = 0$ describes the points at infinity (the *ideal hyperplane*). We find the ideal lines of τ_{ij}, compute the conjugated lines with respect to Σ (i.e., the ideal lines of the normal planes at $\mathbf{p}_{ij}$), and by intersecting them with the absolute quadric $\Omega: \; x_1^2 + x_2^2 + x_3^2 - x_4^2 = x_0 = 0$, we arrive at the isotropic vectors $\mathbf{n}_{ij}^{\pm}$ associated to $\mathbf{p}_{ij}$. The isotropic normals $\mathbf{n}^{\pm}$ can be identified with points of the oval quadric Ω considered as the unit sphere in $\mathbb{R}^3$. Hence we consider the associated normals $\mathbf{N}_{ij} = (n_1, n_2, n_3)/n_4$ on the unit sphere $\mathcal{S}^2$.

 ISBN: 978-84-617-8694-7

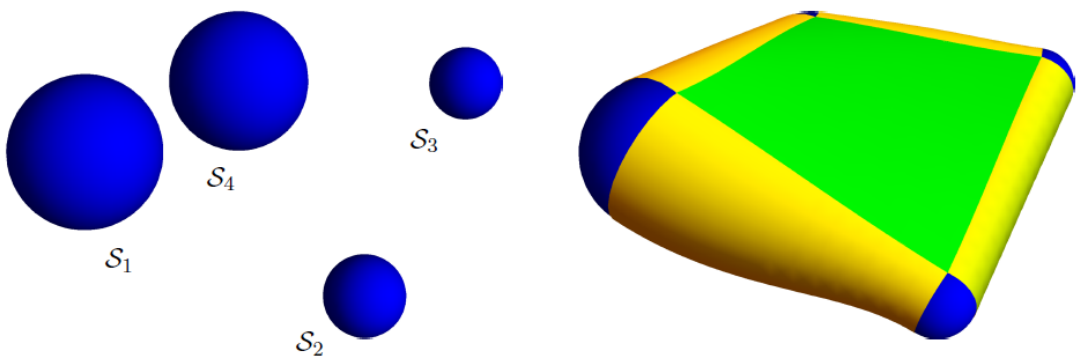


Figure 2.

Next we continue with step (ii) and interpolate the so called *isotropic Gauss image* $\mathcal{G}^+$ by a suitable rational patch $\mathbf{N}(u,v) = (N_1/N_4, N_2/N_4, N_3/N_4)$ on $\mathcal{S}^2$. For this purpose we apply a standard method based on using the stereographic projection $\pi_{\mathbf{w}}$ with the center $\mathbf{w}$, the consequent construction of a suitable polynomial patch in $\mathbb{R}^2$ interpolating $\pi_{\mathbf{w}}(\mathbf{N}_{ij})$ and applying the inverse stereographic projection on the obtained planar patch. Thus we arrive at $\mathbf{n}^+(u,v) = (N_1, N_2, N_3, N_4)$ interpolating data $\lambda_{ij}\,\mathbf{n}^+_{ij}$.

Finally in step (iii), we take $\mathbf{n}^+(u,v)$ as the input for (7) and add to this system the equations

$$\mathbf{x}(i,j) = \mathbf{p}_{ij}, \quad \langle \mathbf{x}_u(i,j), \mathbf{n}^-_{ij} \rangle = 0, \quad \langle \mathbf{x}_v(i,j), \mathbf{n}^-_{ij} \rangle = 0, \tag{8}$$

which guarantee satisfying the interpolation conditions. Solving the system of linear equations we arrive at an MOS patch interpolating given Hermite data $\{\mathbf{p}_{ij}, \tau_{ij}\}$.

The envelopes determined by these polynomial MOS patches. considered as MSTs in $\mathbb{R}^{3,1}$, can be taken as suitable branching elements of the constructed skins. Let us emphasize that not only they are rational, but the properties of MOS surfaces guarantee that also their offsets are rational, as they belong among the so called surfaces with Pythagorean normals (i.e., rational surfaces with rational offsets)

5 Computed examples

In this section we present the designed method on several particular examples. It is assumed that the connectivity (topology) of the skin is given.

Example 5.1 Consider 4 spheres given by ...

 ISBN: 978-84-617-8694-7

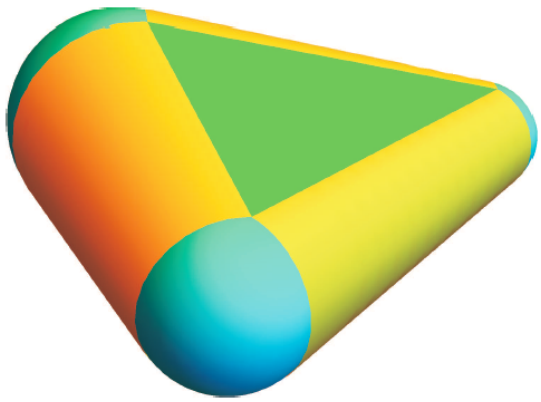


Figure 3.

Example 5.2 Consider 3 spheres given by ...

6 Conclusion

This paper was devoted to the construction of branching elements of rational skins of systems of spheres with the prescribed connectivity. We have shown how the so-called MOS surfaces can be used for this operation. The functionality of the presented method was illustrated on some examples. We would like to emphasize that the designed approach is not limited to constructing branching elements joining only 3 or 4 spheres. In the case when we have more than 4 spheres, the principles of the construction remain the same, but the goal is to construct an n-sided MOS patch which interpolates, at its corners, n points and their associated tangent planes. In our future work we would like to focus on further constructions, important in technical practice, in which MOS patches can be efficiently used – as e.g. when branching blends (i.e., smooth joins of several given shapes) shall be constructed, see Fig. 4.

Acknowledgements

The authors are supported by the project LO1506 of the Czech Ministry of Education, Youth and Sports.

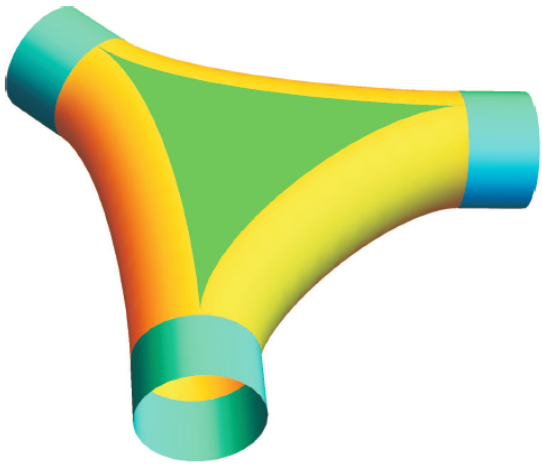


Figure 4: A blending surface between more than two canal surfaces containing the parts which are defined as envelopes of special two-parameter families of spheres.

References

[1] B. Bastl, J. Kosinka, and M. Lávička, *Simple and branched skins of systems of circles and convex shapes*, Graphical Models, 78 (2015), pp. 1 – 9.

[2] M. Bizzarri, M. Lávička, and J. Kosinka, *Skinning and blending with rational envelope surfaces*, Computer-Aided Design, 87 (2017), pp. 41–51.

[3] M. Bizzarri, M. Lávička, Z. Šír, and J. Vršek, *Hermite interpolation by piecewise polynomial surfaces with polynomial area element*, Computer Aided Geometric Design, 51 (2017), pp. 30–47.

[4] H. Choi, C. Han, H. Moon, K. Roh, and N. S. Wee, *Medial axis transform and offset curves by Minkowski Pythagorean hodograph curves*, Computer-Aided Design, 31 (1999), pp. 59–72.

[5] R. Dietz, J. Hoschek, and B. Jüttler, *An algebraic approach to curves and surfaces on the sphere and on other quadrics*, Computer Aided Geometric Design, 10 (1993), pp. 211–229.

[6] M. Dohm and S. Zube, *The implicit equation of a canal surface*, J. Symb. Comput., 44 (2009), pp. 111–130.

 ISBN: 978-84-617-8694-7

[7] G. Farin, *Curves and surfaces for CAGD: A practical guide*, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2002.

[8] J. Kosinka and B. Jüttler, *MOS surfaces: Medial surface transforms with rational domain boundaries*, in The Mathematics of Surfaces XII, vol. 4647 of Lecture Notes in Computer Science, Springer, 2007, pp. 245–262.

[9] R. Kunkli and M. Hoffmann, *Skinning of circles and spheres*, Computer Aided Geometric Design, 27 (2010), pp. 611–621.

[10] G. Landsmann, J. Schicho, and F. Winkler, *The parametrization of canal surfaces and the decomposition of polynomials into a sum of two squares*, Journal of Symbolic Computation, 32 (2001), pp. 119–132.

[11] H. Moon, *Minkowski Pythagorean hodographs*, Computer Aided Geometric Design, 16 (1999), pp. 739–753.

[12] M. Peternell and H. Pottmann, *A Laguerre geometric approach to rational offsets*, Computer Aided Geometric Design, 15 (1998), pp. 223–249.

[13] H. Pottmann and M. Peternell, *Applications of Laguerre geometry in CAGD*, Computer Aided Geometric Design, 15 (1998), pp. 165–186.

[14] G. Slabaugh, B. Whited, J. Rossignac, T. Fang, and G. Unal, *3D ball skinning using PDEs for generation of smooth tubular surfaces*, Computer-Aided Design, 42 (2010), pp. 18–26.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference on Computational and Mathematical Methods in Science and Engineering, CMMSE 2017 4–8 July, 2017.

Unlocking datasets by calibrating populations of models to data density: a study in atrial electrophysiology

Brodie A. J. Lawson[1], Christopher C. Drovandi[1], Nicole Cusimano[2], Pamela Burrage[3], Blanca Rodriguez[5] and Kevin Burrage[1,2,4,6]

[1] *Mathematical Sciences School, Queensland University of Technology, Brisbane, QLD 4000, Australia*

[2] *ARC Centre of Excellence for Mathematical and Statistical Frontiers, Queensland University of Technology, Gardens Point campus, 2 George Street, Brisbane, QLD 40 0 0, Australia*

[3] *Electrical Engineering and Computer Science School, Queensland University of Technology, Brisbane 4001, Australia*

[4] *Institute for Future Environments, Queensland University of Technology, Brisbane, QLD 4000, Australia*

[5] *Department of Computer Science, University of Oxford, Oxford OX13QD, UK*

[6] *Visiting Professor, Department of Computer Science, University of Oxford, Oxford OX13QD, UK*

emails: brodie.lawson86@gmail.com, c.drovandi@qut.edu.au, nicole.cusimano@outlook.com, pamela.burrage@qut.edu.au, blanca@cs.ox.ac.uk, kevin.burrage@qut.edu.au

Abstract

The understanding of complex physical or biological systems nearly always requires a characterisation of the variability that underpins these processes. In addition, the data used to calibrate such models may also often exhibit considerable variability. A recent approach to deal with these issues has been to calibrate populations of models (POMs), that is multiple copies of a single mathematical model but with different parameter values. To date this calibration has been limited to selecting models that produce outputs that fall within the ranges of the dataset, ignoring any trends that might be present in the data.

ISBN: 978-84-617-8694-7

We present here a novel and general methodology for calibrating POMs to the distributions of a set of measured values in a dataset. We demonstrate the benefits of our technique using a dataset from a cardiac atrial electrophysiology study based on the differences in atrial action potential readings between patients exhibiting sinus rhythm (SR) or chronic atrial fibrillation (cAF) and the Courtemanche-Ramirez-Nattel model for human atrial action potentials.

Our approach accurately captures the variability inherent in the experimental population, allows for uncertainty quantification and also allows us to identify the differences underlying stratified data as well as the effects of drug block.

Key words: analysis of complex systems, populations of models, calibration

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

The data-driven COS method

Álvaro Leitao[1,2], Cornelis W. Oosterlee[1,2], Luis Ortiz-Gracia[3] and Sander M. Bohte[4]

[1] *Delft Institute of Applied Mathematics, TU Delft, The Netherlands*

[2] *Scientific Computing group, Centrum Wiskunde & Informatica, The Netherlands*

[3] *Department of Econometrics, University of Barcelona, Spain*

[4] *Machine Learning group, Centrum Wiskunde & Informatica, The Netherlands*

emails: A.LeitaoRodriguez@tudelft.nl, c.w.oosterlee@cwi.nl, luis.ortiz-gracia@ub.edu, S.M.Bohte@cwi.nl

Abstract

In this paper, we present the data-driven COS method, ddCOS. It is a Fourier-based financial option valuation method which assumes the availability of asset data samples: a characteristic function of the underlying asset probability density function is not required. As such, the method represents a generalization of the well-known COS method [1]. Convergence with respect to the number of asset samples is according the convergence of Monte Carlo methods for pricing financial derivatives. The ddCOS method is particularly interesting for density recovery and also for the efficient computation of the option's sensitivities Delta and Gamma. This paper is a short version of an already available preprint [6].

Key words: The COS method, density estimation, data-driven approach, Greeks, the SABR model

1 Introduction

In quantitative finance, statistical distributions are commonly used for the valuation of financial derivatives and within risk management. The underlying assets are often modeled by means of stochastic differential equations (SDEs). Except for the classical and most simple asset models, the corresponding *probability density function* (PDF) and *cumulative distribution function* (CDF) are typically not known and need to be approximated.

In order to compute option prices, and to approximate statistical distributions, Fourier-based methods are commonly used numerical techniques. They are based on the connection between the PDF and the *characteristic function* (ChF), which is the Fourier transform of the probability density. The ChF is often available, and sometimes even in closed form, for the broad class of regular diffusions and also for Lévy processes. Here, we focus on the COS method from [1], which is based on an approximation of the PDF by means of a cosine series expansion.

Still, however, the asset dynamics for which the ChF are known is not exhaustive, and for many relevant asset price processes we do not have such information to recover the density. In this work, we extend the applicability of the COS method to the situation where only data (like samples from an unknown underlying risk neutral asset distribution) are available. We will focus on the framework of *statistical learning*, see [7].

The use of the COS method gives us expressions for option prices and, in particular, for the *option sensitivities or Greeks*. These option Greeks are the derivatives of option price with respect to a variable or parameter. The efficient computation of the Greeks is a challenging problem when only asset samples are available. Existing approaches are based on Monte Carlo (MC)-based techniques, like on finite-differences (bump and revalue), pathwise or likelihood ratio techniques, for which details can be found in [2], chapter 7. The ddCOS method is not directly superior to Monte Carlo methods for option valuation, but it is competitive for the computation of the corresponding sensitivities. We derive simple expressions for the Greeks Delta and Gamma.

This paper is organized as follows. The ddCOS method, and the origins in statistical learning and Fourier-based option pricing, are presented in Section 2. Numerical experiments, with a focus on the option Greeks, are presented in Section 3. We conclude in Section 4.

2 The data-driven COS method

In this section we will discuss the ddCOS method, in which aspects of the MC method, density estimators and the COS method are combined to approximate, in particular, the option Greeks Delta and Gamma for European options. The connection with the COS method is found in the fact that the data-driven PDF appears as a cosine series expansion.

2.1 The COS method

The starting point for the well-known COS method is the risk-neutral option valuation formula, where the value of a European option at time t, $v(x,t)$, is an expectation under the risk neutral pricing measure, i.e.,

$$v(x,t) = \mathrm{e}^{-r(T-t)}\mathbb{E}\left[v(y,T)|x\right] = \mathrm{e}^{-r(T-t)}\int_{\mathbb{R}} v(y,T)f(y|x)\mathrm{d}y, \tag{1}$$

 ISBN: 978-84-617-8694-7

with r the risk-free rate, T the maturity time, and $f(y|x)$ the PDF of the underlying process, and $v(y,T)$ represents the option value at maturity time, being the payoff function. Typically, x and y are chosen to be scaled variables,

$$x := \log\left(\frac{S(0)}{K}\right) \quad \text{and} \quad y := \log\left(\frac{S(T)}{K}\right),$$

where $S(t)$ is the underlying asset process at time t, and K is the strike price.

Density $f(y|x)$ is unknown in most cases and in the COS method it is approximated, on a finite interval $[a,b]$, by a cosine series expansions, i.e.,

$$f(y|x) = \frac{1}{b-a}\left(A_0 + 2\sum_{k=1}^{\infty} A_k(x)\cdot\cos\left(k\pi\frac{y-a}{b-a}\right)\right),$$
$$A_0 = 1, \quad A_k(x) = \int_a^b f(y|x)\cos\left(k\pi\frac{y-a}{b-a}\right)\mathrm{d}y, \quad k = 1,2,\ldots.$$

By substituting this expression in Equation (1), interchanging the summation and integration operators using Fubini's Theorem, and introducing the following definition,

$$V_k := \frac{2}{b-a}\int_a^b v(y,T)\cos\left(k\pi\frac{y-a}{b-a}\right)\mathrm{d}y,$$

we find that the option value is given by

$$v(x,t) \approx \mathrm{e}^{-r(T-t)}\sum_{k=0}^{\infty}{}' A_k(x)V_k, \tag{2}$$

where $'$ indicates that the first term is divided by two. So, the product of two real-valued functions in Equation (1) is transformed into the product of their cosine expansion coefficients, A_k and V_k. Density coefficients A_k can be computed by the ChF and V_k is known analytically (for many types of options).

Closed-form expressions for the option Greeks can also be derived. From the COS option value formula, Δ and Γ are obtained by

$$\begin{aligned}\Delta &= \frac{\partial v(x,t)}{\partial S} = \frac{1}{S(0)}\frac{\partial v(x,t)}{\partial x} \approx \exp(-r(T-t))\sum_{k=0}^{\infty}{}' \frac{\partial A_k(x)}{\partial x}\frac{V_k}{S(0)},\\ \Gamma &= \frac{\partial^2 v(x,t)}{\partial S^2} = \frac{1}{S^2(0)}\left(-\frac{\partial v(x,t)}{\partial x} + \frac{\partial^2 v(x,t)}{\partial x^2}\right)\\ &\approx \exp(-r(T-t))\sum_{k=0}^{\infty}{}'\left(-\frac{\partial A_k(x)}{\partial x)} + \frac{\partial^2 A_k(x)}{\partial x^2)}\right)\frac{V_k}{S^2(0)}\end{aligned} \tag{3}$$

Due to the rapid decay of the coefficients, $v(x,t)$, Δ and Γ can be approximated with high accuracy by truncating the infinite summation in Equations (2) and (3) to N terms. Under suitable assumptions, exponential convergence is proved and numerically observed.

 ISBN: 978-84-617-8694-7

2.2 Statistical learning theory for density estimation

In the setting of this paper, we assume a vector of n independent and identically distributed (i.i.d.) samples, $X_1, X_2, \ldots, X_n$. Based on these samples, we wish to find an accurate approximation of the PDF estimator, $f_n(x)$, which should approximate density $f(x)$.

By definition, the PDF is related to its CDF $F(x)$,

$$\int_{-\infty}^{x} f(y)\mathrm{d}y = F(x). \tag{4}$$

Function $F(x)$ is approximated by the empirical approximation,

$$F_n(x) = \frac{1}{n}\sum_{j=1}^{n} \eta(x - X_j), \tag{5}$$

where $\eta(\cdot)$ is a step function. This approximation converges to the "true CDF" with rate $\mathcal{O}(1/\sqrt{n})$. Rewriting Equation (4) as a linear operator equation, gives us,

$$Cf = F \approx F_n,$$

where the operator $Ch := \int_{-\infty}^{x} h(z)\mathrm{d}z$.

As explained in [7], this matrix equation represents an ill-posed problem, and therefore a *risk functional* should be constructed, with a regularization term, as follows

$$R_{\gamma_n}(f, F_n) = L^2_{\mathcal{H}}(Cf, F_n) + \gamma_n W(f), \tag{6}$$

where $L_{\mathcal{H}}$ is a metric of the space $\mathcal{H}$, $\gamma_n > 0$ is a parameter which gives a weight to the regularization term $W(f)$, with $\gamma_n \to 0$ as $n \to \infty$. The solution of $Cf = F_n$ belongs to $\mathcal{D}$, the domain of definition of $W(f)$. Functional $W(f)$ takes real non-negative values in $\mathcal{D}$. Furthermore, $\mathcal{M}_c = \{f : W(f) \leq c\}$ is a compact set in $\mathcal{H}$ (the space where the solution exists and is unique).

The solution f_n, minimizing the functional in Equation (6), converges almost surely to the desired density. By considering particular choices of $L_{\mathcal{H}} = L_2(0, \pi)$ and W (see [7] for details), the the risk functional in Equation (6) becomes

$$R_{\gamma_n}(f, F_n) = \int_0^{\pi} \left(\int_0^{x} f(y)\mathrm{d}y - F_n(x) \right)^2 \mathrm{d}x + \gamma_n \int_0^{\pi} \left(f^{(p)}(x) \right)^2 \mathrm{d}x. \tag{7}$$

Given a cosine series expansion, $\psi_1(\theta) = \cos(\theta), \ldots, \psi_k(\theta) = \cos(k\theta), \ldots$, the approximation to the unknown PDF will be of the form

$$f_n(\theta) = \frac{1}{\pi} + \frac{2}{\pi}\sum_{k=1}^{\infty} \tilde{A}_k \cos(k\theta), \tag{8}$$

 ISBN: 978-84-617-8694-7

with $\tilde{A}_0, \tilde{A}_1, \ldots, \tilde{A}_k, \ldots$ expansion coefficients, defined as $\tilde{A}_k =< f_n, \psi_k >$. We need to compute the expansion coefficients so that the functional in Equation (7) is minimized. The coefficients $\tilde{A}_k$ cannot be directly computed from the definition since the unknown PDF, f_n, is implicitly involved in the expression, but from the equivalent expression $\tilde{A}_k =< \hat{f}_n, \hat{\psi}_k >$. Thus, it can be proved that the minimum of the functional using cosine series expansions (see [7]) is obtained when

$$\tilde{A}_k = \frac{1}{1+\gamma_n k^{2(p+1)}} \frac{1}{n} \sum_{j=1}^{n} \cos(k\theta_j), \tag{9}$$

where $\theta_j \in (0, \pi)$ are given samples of the unknown distribution.

Assuming that the samples are given, the solution contains two free parameters: *regularization parameter* γ_n, and *smoothing parameter* p. The parameter p is selected as $p = 0$, implying that the actual function enters in the regularization term. The choice of parameter γ_n impacts the efficiency of the data-driven COS method, since it is related to the required number of data samples, and by reducing the number of samples, the overall computational cost can be reduced. The parameter γ can be obtained by the following rule,

$$\gamma_n = \frac{\log \log n}{n}. \tag{10}$$

As proved in [7], this rule provides a robust asymptotic rate of convergence under the assumption of a compactly supported density. It implies, with probability one, uniformly converging approximations f_n to the unknown density.

2.3 The ddCOS method

We are now ready to present the ddCOS method, where we employ the series expansion coefficients from the regularization approach. We replace the A_k-coefficients from Equation (2) by those coefficients based on data, $\tilde{A}_k$ in Equation (9).

So, suppose we have risk neutral samples (or values) from an underlying asset at a future time t, i.e., $S_1(t), S_2(t), \ldots, S_n(t)$. We compute the value of a European option with maturity time T and strike price K, and require therefore the samples $S_j(T)$. With a logarithmic transformation, we have

$$Y_j := \log\left(\frac{S_j(T)}{K}\right).$$

Before employing these samples in the regularization approach and because the solution is defined in $(0, \pi)$, we need to transform the samples by the following change of variables,

$$\theta_j = \pi \frac{Y_j - a}{b - a},$$

 ISBN: 978-84-617-8694-7

where the boundaries a and b are defined as

$$a := \min_{1\leq j\leq n}(Y_j), \quad b := \max_{1\leq j\leq n}(Y_j).$$

The A_k coefficients in Equation (2) are replaced by the data-driven $\tilde{A}_k$ in Equation (9),

$$A_k \approx \tilde{A}_k = \frac{\frac{1}{n}\sum_{j=1}^{n}\cos\left(k\pi\frac{Y_j-a}{b-a}\right)}{1+\gamma_n k^{2(p+1)}}.$$

The ddCOS pricing formula for European options based on risk neutral data is now obtained as

$$\begin{aligned}\tilde{v}(x,t) &= \mathrm{e}^{-r(T-t)}\sum_{k=0}^{\infty}{}'\frac{\frac{1}{n}\sum_{j=1}^{n}\cos\left(k\pi\frac{Y_j-a}{b-a}\right)}{1+\gamma_n k^{2(p+1)}}\cdot V_k \\ &= \mathrm{e}^{-r(T-t)}\sum_{k=0}^{\infty}{}'\tilde{A}_k V_k.\end{aligned} \tag{11}$$

The samples Y_j should originate from one initial state, i.e. the dependency on the state x is implicitly assumed. In the case of European options this is typically fulfilled. In the Monte Carlo method, for example, all simulated asset paths depart from the same point $S(0)$, so that $x := \log\left(\frac{S(0)}{K}\right)$.

Regarding the Greeks, we can also derive data-driven expressions for the Δ and Γ sensitivities. We first define the corresponding sine coefficients as

$$\tilde{B}_k := \frac{\frac{1}{n}\sum_{j=1}^{n}\sin\left(k\pi\frac{Y_j-a}{b-a}\right)}{1+\gamma_n k^{2(p+1)}}.$$

Taking derivatives in Equation (11) w.r.t the samples, Y_j, and following the COS expression for the sensitivities in Equation (3), the data-driven Greeks, $\tilde{\Delta}$ and $\tilde{\Gamma}$, can be obtained by

$$\begin{aligned}\tilde{\Delta} &= \mathrm{e}^{-r(T-t)}\sum_{k=0}^{\infty}{}'\tilde{B}_k\cdot\left(-\frac{k\pi}{b-a}\right)\cdot\frac{V_k}{S(0)}, \\ \tilde{\Gamma} &= \mathrm{e}^{-r(T-t)}\sum_{k=0}^{\infty}{}'\left(\tilde{B}_k\cdot\frac{k\pi}{b-a}-\tilde{A}_k\cdot\left(\frac{k\pi}{b-a}\right)^2\right)\cdot\frac{V_k}{S^2(0)}.\end{aligned}$$

As in the original COS method, we must truncate the infinite sum to a finite number of terms N.

 ISBN: 978-84-617-8694-7

2.3.1 Application of variance reduction

Because of the focus on asset path data, the ddCOS method is related to the Monte Carlo method. Variance reduction in Monte Carlo methods is typically achieved by the use of variance reduction techniques. The ddCOS method also admits an additional variance reduction, in this case, for the computation of the expansion coefficients, $\tilde{A}_k$. We show how to introduce *antithetic variates (AV)* to our method. Since one of the assumptions for the regularization approach is that the samples are i.i.d., an immediate application of AV is not possible. Therefore, if we assume that antithetic samples, Y_i', to the original samples Y_i, can be computed without any serious computational effort, a new estimator for the coefficients can be defined as

$$\bar{A}_k := \frac{1}{2}\left(\tilde{A}_k + \tilde{A}_k'\right),$$

where we denote by $\tilde{A}_k'$ the corresponding "antithetic coefficients", obtained by Y_i'. By a similar derivation as for the standard AV technique, it can be proved that the use of coefficients $\bar{A}_k$ will give us a variance reduction compared to using the $\tilde{A}_k$ coefficients. Other variance reduction techniques may also be considered for the ddCOS method under the assumption of i.i.d. samples.

3 Applications of the ddCOS method

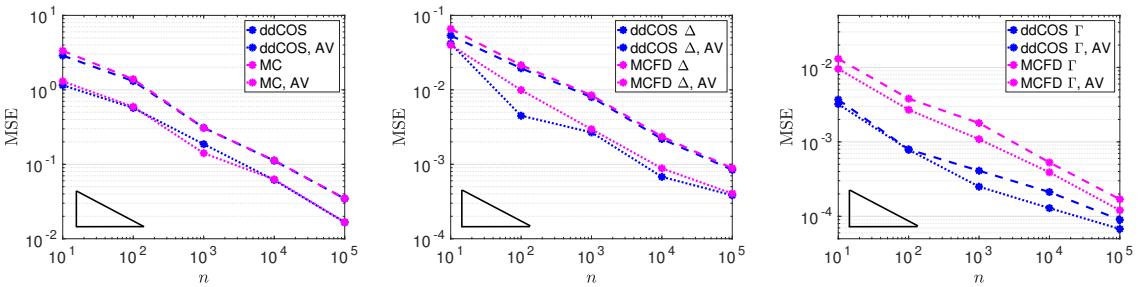
In this section, we present some applications of the ddCOS method. The first application is an option pricing experiment, where we show the method's convergence. Subsequently, we present the performance regarding the computation of the Greeks, where ddCOS exhibits a stable convergence and can be employed with involved models, as we only need asset samples. We also compute the Greeks under Merton and SABR models. The experiments have been carried out on a computer system with the following characteristics: CPU Intel Core i7-4720HQ 2.6GHz and RAM memory of 16GB RAM. The employed software package is Matlab R2016b.

3.1 Option valuation

First of all, we numerically test the convergence of the ddCOS method in an option valuation experiment. The *Geometric Brownian Motion* (GBM) asset dynamics are employed, since a reference value for the option value is available by the Black-Scholes formula. The regularization parameter γ_n is set as in Equation (10). As is common in MC experiments, the *Mean Squared Error* (MSE) is considered as the error measure. In the convergence tests, the reported values are computed as the average of 50 experiments.

The expected order of convergence for the option values is $\mathcal{O}(1/\sqrt{n})$, according to the convergence of the empirical CDF towards the true CDF in Equation (5). In Section

 ISBN: 978-84-617-8694-7



(a) Option price. (b) Greek Δ. (c) Greek Γ.

Figure 1: Convergence in prices and Greeks of the ddCOS method: Antithetic Variates (AV); GBM, $S(0) = 100$, $r = 0.1$, $\sigma = 0.3$, $K = S(0)$ and $T = 2$.

2.3.1, the application of antithetic variates in the ddCOS framework has been presented. In Figure 1a, we confirm that this variance reduction technique provides a similar improvement in terms of precision as when it is applied to the plain MC method.

3.2 Greeks

We have empirically shown in Figure 1a that the ddCOS method converges to the true price with the expected convergence rate $\mathcal{O}(1/\sqrt{n})$, which resembles the plain MC convergence. However, by the ddCOS method, not only the option value but also the sensitivities can readily be obtained. This is an advantage w.r.t MC-based methods for estimating sensitivities, where often, additional simulations, intermediate time-steps or prior knowledge are required. Thus, a similar convergence test is performed for the Δ and Γ sensitivities, see Figures 1b and 1c. As MC-based method for the Greeks calculation we consider the *Finite Difference* method (bump and revalue, denoted as MCFD). We have chosen MCFD for the comparison because it is flexible and it does not require prior knowledge. MCFD may require one or two extra simulations, and the choice of optimal *shift* parameter may not be trivial. The reference Delta and Gamma are given by the Black-Scholes formula. In both experiments, while the Δ is very well approximated by the ddCOS and MCFD methods, the second derivative, Γ, appears more complicated for the MCFD method. This fact was already pointed out by Glasserman in [2]. The ddCOS estimator, however, is accurate and stable as it is based on the data-driven PDF and the ddCOS machinery.

Using $n = 10^5$, in Table 1 we now compare the Δ and Γ estimations obtained under the GBM dynamics for several strikes. The performance of the ddCOS method is very satisfactory as it is accurate, with small *Relative Error* (RE, averaged over K) and reproduces the reference values very well. The difficulties of the MCFD estimating Γ are more clearly visible.

 ISBN: 978-84-617-8694-7

K (% of $S(0)$)	80%	90%	100%	110%	120%
			Δ		
Ref.	0.8868	0.8243	0.7529	0.6768	0.6002
ddCOS	0.8867	0.8240	0.7528	0.6769	0.6002
RE	1.1012×10^{-4}				
MCFD	0.8876	0.8247	0.7534	0.6773	0.6006
RE	7.5168×10^{-4}				
			Γ		
Ref.	0.0045	0.0061	0.0074	0.0085	0.0091
ddCOS	0.0045	0.0062	0.0075	0.0084	0.0090
RE	8.5423×10^{-3}				
MCFD	0.0045	0.0059	0.0071	0.0079	0.0083
RE	4.9554×10^{-2}				

Table 1: GBM option Greeks: Call, $S(0) = 100$, $r = 0.1$, $\sigma = 0.3$ and $T = 2$.

We wish to test the ddCOS method in a more complex situation, by adding jumps in the form of a Merton jump-diffusion asset price process. To accurately compute the option sensitivities in this case gives rise to difficulties for MC-based methods. We perform a similar experiment as before, where now the underlying asset follows the Merton jump-diffusion model, and the obtained Δ and Γ are presented in Table 2. In this case, the reference value is provided by the COS method at a high accuracy.

The SABR model [3] is interesting within the ddCOS framework since the ChF is not known and, furthermore, the asset path MC simulation is not trivial. The authors provided a closed-form approximation formula for the *implied volatility* under the SABR dynamics, which is often used within the calibration. However, the closed-form expression is derived by perturbation theory, and therefore the formula is not accurate for small strike values, for long time to maturity options or for high volatilities (see, for example, [4, 5]). Therefore, the calculation of the Greeks under the SABR model becomes challenging but can be addressed by the ddCOS method. To employ the method, we need samples of the underlying asset at time T. Here, we make use of the one time-step SABR MC simulation introduced by Leitao et al. in [4]. Thus, the ddCOS method will be combined with the one time-step SABR simulation to efficiently compute Δ and Γ under the SABR dynamics.

For the numerical experiments, we consider two parameter settings. First of all, a basic parameter set is taken, where the SABR formula is valid and can be used as a reference. The results are presented in Table 3. For the second test we use a more difficult set of parameters (i.e., Set III in [4]), where the SABR formula does not provide accurate results anymore. In Table 4, we observe that the ddCOS provides accurate Δ-values in this case, without any problems. The reference value has been computed by the MCFD in combination with the

 ISBN: 978-84-617-8694-7

K (% of $S(0)$)	80%	90%	100%	110%	120%
			Δ		
Ref.	0.8385	0.8114	0.7847	0.7584	0.7328
ddCOS	0.8383	0.8113	0.7846	0.7585	0.7333
RE	2.7155×10^{-4}				
MCFD	0.8387	0.8118	0.7850	0.7586	0.7330
RE	3.1265×10^{-4}				
			Γ		
Ref.	0.0022	0.0024	0.0027	0.0029	0.0030
ddCOS	0.0022	0.0024	0.0027	0.0029	0.0030
RE	8.2711×10^{-3}				
MCFD	0.0023	0.0026	0.0028	0.0031	0.0033
RE	6.118×10^{-2}				

Table 2: Merton jump-diffusion option Greeks: Call, $S(0) = 100$, $r = 0.1$, $\sigma = 0.3$, $\mu_j = -0.2$, $\sigma_j = 0.2$ and $\lambda = 8$ and $T = 2$.

K (% of $S(0)$)	80%	90%	100%	110%	120%
			Δ		
Ref.	0.9914	0.9284	0.5371	0.0720	0.0058
ddCOS	0.9916	0.9282	0.5363	0.0732	0.0058
RE	5.2775×10^{-3}				
MCFD	0.9911	0.9279	0.5368	0.0737	0.0058
RE	5.5039×10^{-3}				

Table 3: Greek Δ under the SABR model: Call, $S(0) = 100$, $r = 0$, $\sigma_0 = 0.3$, $\alpha = 0.4$, $\beta = 0.6$, $\rho = -0.25$ and $T = 2$.

SABR MC simulation in [5], with a large number of MC paths ($n = 10,000,000$) and time steps ($4T$).

4 Conclusions

In this work, the ddCOS method has been introduced. The method extends the COS method applicability to cases when only data samples of the underlying asset are available. The method exploits a closed-form solution, in terms of Fourier cosine expansions, of a density. The use of the COS machinery in combination with density estimation allowed us to develop a data-driven method which can be employed for option pricing and risk management. The ddCOS method particularly results in an efficient method for the Δ and

 ISBN: 978-84-617-8694-7

K (% of $S(0)$)	80%	90%	100%	110%	120%
			Δ		
Ref.	0.8384	0.7728	0.6931	0.6027	0.5086
ddCOS	0.8364	0.7703	0.6902	0.6006	0.5084
RE	2.7855×10^{-3}				
Hagan	0.8577	0.7955	0.7170	0.6249	0.5265
RE	3.1751×10^{-2}				

Table 4: Greek Δ under SABR model. Setting: Call, $S(0) = 0.04$, $r = 0.0$, $\sigma_0 = 0.4$, $\alpha = 0.8$, $\beta = 1.0$, $\rho = -0.5$ and $T = 2$.

Γ sensitivities computation, based solely on the samples.

References

[1] F. Fang and C. W. Oosterlee, *A novel pricing method for European options based on Fourier-cosine series expansions*, SIAM Journal on Scientific Computing **31** (2008) 826–848.

[2] P. Glasserman, *Monte Carlo Methods in Financial Engineering*, Springer, 2004.

[3] P. S. Hagan and D. Kumar and A. S. Lesniewski and D. E. Woodward, *Managing smile risk*, Wilmott Magazine (2002) 84–108.

[4] Á. Leitao and L. A. Grzelak and C. W. Oosterlee, *On a one time-step Monte Carlo simulation approach of the SABR model: application to European options*, Applied Mathematics and Computation **293** (2017) 461–479.

[5] Á. Leitao and L. A. Grzelak and C. W. Oosterlee, *On an efficient multiple time step Monte Carlo simulation of the SABR model*, Quantitative Finance (2017).

[6] Á. Leitao and C. W. Oosterlee and L. Ortiz-Gracia and S. M. Bohte, *On the data-driven COS method*, Available at SSRN: http://ssrn.com/abstract=291753 (2017).

[7] V. N. Vapnik, *Statistical learning theory*, Wiley-Interscience, 1998.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

Energy-efficient QR Factorization on FPGAs

Germán León[1], Carlos González[2], Rafael Mayo[1], Enrique S. Quintana-Ortí[1] and Daniel Mozos[2]

[1] *Depto. de Ingeniería y Ciencia de Computadores, Univ. Jaume I, Castellón (Spain)*

[2] *Depto. de Arquitectura de Computadores y Automática, Universidad Complutense de Madrid (Spain)*

emails: leon@uji.es, carlosgo@ucm.es, mayo@uji.es, quintana@uji.es, mozos@ucm.es

Abstract

We analyze the implementation, performance and energy efficiency of an algoritm to compute the QR factorization of a dense matrix on a Field Programmable Gate Array (FPGA). Our implementation is based on a simple level-2 BLAS formulation of the factorization algorithm that relies on Householder reflectors. The results on a Xilinx Virtex-7 show a dissipation rate that is slightly above 3 Watts.

Key words: QR factorization, high performance, energy consumption, field programmable gate array (FPGA), linear algebra.

1 Introduction

The QR factorization is a key numerical algorithm for the solution of linear least squares problems arising, among others, in statistics, geodetics, signal processing, control and, in general, in any scenario where it is necessary to fit a model to observations containing errors [3]. As a consequence, over the past decades there has been a continuous effort to accelerate the computation and improve the stability of this factorization, via efficient algorithms and computational kernels for a wide variety of computer architectures. In particular, from the mathematical and algorithmic perspectives one of the most remarkable advances is the formulation of the QR factorization via Householder reflectors [5]. On the other hand, from the point of view of high performance on modern computer architectures, the introduction of the WY transform and its compact variant [6], which unleashed the utilization of compute-intensive kernels in the factorization, are two significant advances.

ISBN: 978-84-617-8694-7

As computer architectures progress on the road to Exascale systems, energy has arisen as a primary challenge on par with performance [4, 8]. A distinguisable milestone in high performance computing (HPC) system have been the shift towards heterogeneous systems, equipped with some sort of accelerator, in order to improve the performance-per-energy unit ratio [1]. As we move along this line, we can expect an eventual adoption of customizable technologies such as field programmable gate arrays (FPGAs) due to their flexibility, performance and energy efficiency, as recent movements from companies such as IBM, Intel and Microsoft hint.

In this paper we assess the performance and energy efficiency of an algorithm to compute the QR factorization on FPGA based on Householder reflectors and implemented on a Xilinx Virtex-7 board.

2 The QR Factorization

Given a nonzero vector $x \in \mathbb{R}^n$, the corresponding Householder reflector $H := \text{HOUSE}(x) = I_n - \tau v v^T$, where $v = x + \alpha e_1$, $\tau = 2/(v^T v)$, $\alpha = \pm ||x||_2$, I_n denotes the square identity matrix of order n and e_1 is the first column of I_n, satisfies $y := Hx = \mp ||x||_2 e_1$ [6]; that is, all entries of x are annihilated by the application (from the left) of the Housedoher reflector H, except the first entry of x which, after the application, becomes $\mp ||x||_2$.

For a matrix $A \in \mathbb{R}^{m \times n}$, this type of orthogonal reflectors can be applied to compute the QR factorization $A = QR$, where $Q \in \mathbb{R}^{m \times m}$ is orthogonal and $R \in \mathbb{R}^{m \times n}$ is upper triangular. For this purpose, the procedure commences with $A^{(0)} = A$ and consecutively applies a sequence of Householder reflectors $H_1, H_2, \ldots, H_n$ such that H_j annihilates the (subdiagonal) entries in the j-th column of $A^{(j-1)} = (H_{j-1} \ldots H_1)A$. Thus, upon completion, $A^{(n)} = R$, and $Q = H_1 H_2 \cdots H_n$ [6]. Figure 1 illustrates the calculation of this factorization, using the FLAME notation [7], via an unblocked algorithm that is implemented as routine GEQR2 in the *Linear Algebra Package* (LAPACK) [2]. Internally, the algorithm relies on routine LARFG to generate a Householder reflector for the vector consisting of α_{11} and a_{21}. Routine LARF then applies this reflector (from the left) to the trailing submatrix composed of a_{12}^T and A_{22}. The Householder reflector H is not explicitly built, but applied implicitly using the parameters v, τ. In particular, note that the application of the Householder reflector to a matrix $\hat{A}$ can be performed as $H\hat{A} = (I - \tau v v^T)\hat{A} = \hat{A} - \tau v (v^T \hat{A})$, which boils down to a matrix-vector product, $w^T := w^T \hat{A}$, followed by a (scaled) rank-1 update, $\hat{A} := \hat{A} - \tau v w^T$.

In practice, the upper triangular factor R overwrites the corresponding entries of A. Furthermore, the parameters v_j and τ_j that define the Householder reflector H_j (which annihilates the subdiagonal entries in the j-th column of $A^{(j-1)}$) are respectively stored using the annhilated entries of the column plus the j-th entry of an additional vector of size n. (Here, the first entry of v equals 1 and does not need to be explicitly stored.)

 ISBN: 978-84-617-8694-7

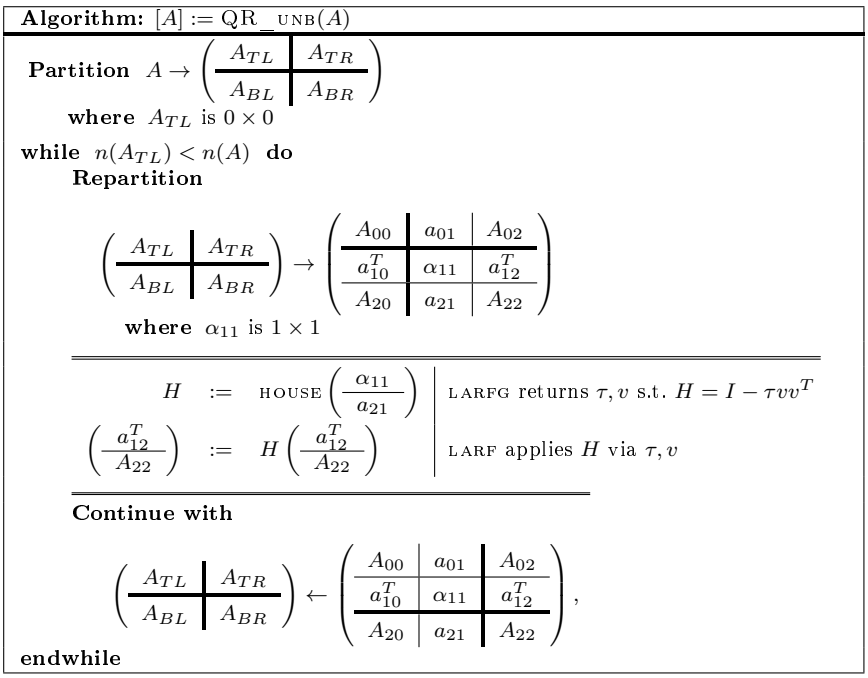


Figure 1: Unblocked algorithm for the QR factorization using Householder reflectors. $n(\cdot)$ is a function that returns the number of columns of its input argument.

3 Implementation on FPGAs

Figure 2 illustrates the hardware architecture that implements the algorithm for the QR factorization. The *memory* module provides the input data; the *House* module calculates the Householder reflector; and the *Row house* module applies the Householder reflector to the appropriate blocks of the data matrix.

Matrix $A \in \mathbb{R}^{m \times n}$ is partitioned, passsed to and processed by the FPGA by blocks of dimension $n \times n$, starting from the bottom and proceedings upwards. While processing two of these blocks, the block immediately above them is transferred to the FPGA in order to overlap communication with computation and avoid idle periods. Module *Memory* is composed of three banks: input, upper and lower. The input bank stores the block in transference while the remaining two banks provide the information to the rest of the processing system. When the calculation of a submatrix is completed, the roles of the banks are rotated so that the upper bank becomes the input bank, the receiver becomes the lower bank, and the lower becomes becomes the upper one.

The calculation modules (*House* and *Row house*) work in parallel. As soon as the first

 ISBN: 978-84-617-8694-7

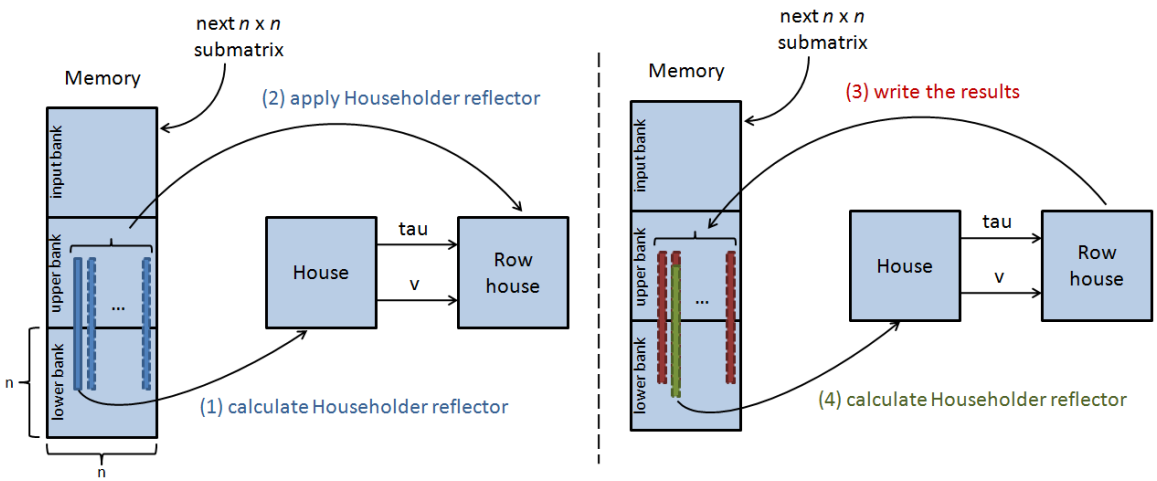


Figure 2: Hardware architecture used to implement QR factorization and steps involved in an intermediate iteration

column is updated, the Householder reflector for the next iteration can be calculated. Thus, steps (3) and (4) in Figure 2 are executed simultaneously. The generation of the Householder reflector exhibits a reduced degree of parallelism and does not have a contiguous source of data. Therefore, the design of module *House* aims to offer low latency and consume a low number of DSP resources. The majority of DSPs are dedicated to the implementation of the *Row house* module to take full advantage of the parallelism of this operation.

There are two versions for the *Row house* module. The first one utilizes the maximum number of floating–point units (following a non-blocking pipeline architecture). The second version adapts the design depending on the number of resources, implementing operators of smaller size and multiplexing the data entry (following a blocking pipeline architecture) in time.

4 Experimental Results

Table 1: Summary of resource utilization

Resource	LUT	LUTRAM	BRAM	DSP
Available	433,200	174,200	1,470	3,600
Utilization (%)	251,042 (57.95%)	4,604 (2.64%)	617 (41.97%)	2,649 (73.58%)

We study three metrics of our FPGA implementation: Board usage, computational performance and power consumption. The hardware architecture was implemented on a Xilinx

Virtex-7 XC7VX690T. Table 1 shows the resources necessary for our hardware implementation for a problem of size $n = 256$. In reference to performance, in order to factorize a block of dimension 256×256, the implementation requires a total of 175,205 clock cycles and proceeds at 60 MHz. Looking at power consumption, the static power consumption is 0.58 Watts while the dynamic power consumption is 2.42 Watts, for a total of 3 Watts.

Acknowledgements

The researchers from Universidad Jaime I (UJI) were supported by the CICYT project TIN2014-53495-R of MINECO and FEDER. The researchers from Universidad Complutense de Madrid (UCM) were supported by the Spanish Ministry of Science and Innovation under reference READAR (TIN2013-40968-P).

References

[1] The Green500 list (2017). Available at http://www.green500.org

[2] Anderson, E., Bai, Z., Bischof, C., Blackford, L.S., Demmel, J., Dongarra, J.J., Croz, J.D., Hammarling, S., Greenbaum, A., McKenney, A., Sorensen, D.: LAPACK Users' guide, 3rd edn. SIAM (1999)

[3] Björck, A.: Numerical Methods for Least Squares Problems. Society for Industrial and Applied Mathematics (1996). DOI 10.1137/1.9781611971484. URL http://epubs.siam.org/doi/abs/10.1137/1.9781611971484

[4] Duranton *et al*, M.: HiPEAC vision 2015. High performance and embedded architecture and compilation (2015). http://www.hipeac.net/vision

[5] Golub, G.: Numerical methods for solving linear least squares problems. Numerische Mathematik **7**(3), 206–216 (1965). DOI 10.1007/BF01436075. URL http://dx.doi.org/10.1007/BF01436075

[6] Golub, G.H., Loan, C.F.V.: Matrix Computations, 3rd edn. The Johns Hopkins University Press, Baltimore (1996)

[7] Gunnels, J.A., Gustavson, F.G., Henry, G.M., van de Geijn, R.A.: FLAME: Formal linear algebra methods environment. ACM Trans. Math. Soft. **27**(4), 422–455 (2001). URL http://doi.acm.org/10.1145/504210.504213

[8] Lucas, R.: Top ten Exascale research challenges (2014). http://science.energy.gov/~/media/ascr/ascac/pdf/meetings/20140210/Top10reportFEB14.pdf

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

Mathematical Modelling the Spread of Zika and Microcephaly in Brazil.

Yanfeng Liang[1] **and David Greenhalgh**[1]

[1] *Department of Mathematics and Statistics, University of Strathclyde, Glasgow, UK*

emails: yanfeng.liang@strath.ac.uk, david.greenhalgh@strath.ac.uk

Abstract

In this paper we look at a non-age-structured model for the spread of the Zika Virus and Microcephaly in Brazil. We first outline the non-seasonal differential equation model, and discuss parameter values and their estimation. Then we talk about the basic reproduction number and details of the calculation of the number of Microcephaly cases. Next we estimate how the model can be made more realistic by introducing seasonality into the mosquito population. Finally we consider sensitivity of the results to the mosquito biting rate.

Key words: Zika, Brazil, Microcephaly, Aedes Aegypti mosquito, basic reproduction number, differential equation model, seasonality.
MSC 2000: AMS codes 92B15, 92C60, 92D30.

1 Extended Abstract

1.1 Non-seasonal Model

The Zika virus is spread by the same species of mosquito, namely the *Aedes Aegypti* (*A. Aegypti*), as Dengue. Zika is a member of the virus family *Flaviviridae*. The first discovery of the Zika virus was in 1947, however despite being around for a while, Zika has not received much attention until recently when it has been discovered that it is associated with Microcephaly which is a serious birth defect in newborns, caused if women are infected with Zika during pregnancy. Most importantly there is still no vaccine to prevent the Zika virus. Apart from causing severe birth defects to newborn babies, infected individuals can also experience fever, rash and joint pain. As a result, in this paper we will use an existing

 ISBN: 978-84-617-8694-7

Parameter values	Biological meanings	Values
a	*A. Aegypti* biting rate	0.5×7/week [1, 5]
b	Probability of transmission of Zika when an infectious mosquito bites a susceptible human	$0.10 - 0.75$ [1]
c	Probability of transmission of Zika when a susceptible mosquito bites an infectious human	$0.30 - 0.75$ [3]
N_H	Human population in Brazil in 2015	$207,848,000$ [12]
μ_H	Per capita human mortality rate in Brazil	$1/(75 \times 52)$/week [12]
γ	Per capita human recovery rate	7/6/week [5]
μ_v	Per capita mortality rate for *A. Aegypti*	0.025×7/week [9]
N_v	*A. Aegypti* population	$1.5 \times N_H$ [4, 6]
τ	Zika extrinsic incubation period	8.2/7 weeks [5]

Table 1: Parameter values given in Equation (1).

time-delayed mathematical model for Dengue mentioned in [9] to analyse the dynamical behaviour for the Zika virus, in Brazil, as well as estimating the future expected number of cases of Microcephaly due to Zika.

The definitions of the parameter values used in the differential equation model and their corresponding values are given in Table 1. The model that we are working with is given as follows:

$$\begin{aligned}
\frac{dS_H(t)}{dt} &= -abI_v(t)\frac{S_H(t)}{N_H} - \mu_H S_H(t) + \mu_H N_H, \\
\frac{dI_H(t)}{dt} &= abI_v(t)\frac{S_H(t)}{N_H} - (\mu_H + \gamma)I_H(t), \quad \frac{dR_H(t)}{dt} = \gamma I_H(t) - \mu_H R_H(t), \qquad (1) \\
\frac{dS_v(t)}{dt} &= -acS_v(t)\frac{I_H(t)}{N_H} - \mu_v S_v(t) + \mu_v N_v, \\
\frac{dL_v(t)}{dt} &= acS_v(t)\frac{I_H(t)}{N_H} - \mu_v L_v(t) - acS_v(t-\tau)\frac{I_H(t-\tau)}{N_H}e^{-\mu_v \tau}, \\
\frac{dI_v(t)}{dt} &= acS_v(t-\tau)\frac{I_H(t-\tau)}{N_H}e^{-\mu_v \tau} - \mu_v I_v(t),
\end{aligned}$$

with initial conditions $S_H(0), I_H(0), R_H(0), S_v(0), L_v(0)$ and $I_v(0)$, where $S_H(t), I_H(t)$ and $R_H(t)$ respectively represent the susceptible, infected and recovered individuals for humans, while $S_v(t), L_v(t)$ and $I_v(t)$ respectively represent the susceptible, latent and infected mosquitoes. Note that $N_H = S_H + I_H + R_H$ denotes the total human population size and $N_v = S_v + L_v + I_v$ represents the total *A. Aegypti* population size where both populations

 ISBN: 978-84-617-8694-7

are constant. Although the Zika virus and Dengue are spread by the same transmission route and thus some parameter values would remain the same, parameters such as the transmission probabilities between humans and *A. Aegypti* mosquitoes which are defined as b and c in Equation (1) may vary. Therefore one of the aims in this project is to use the least squares estimation technique and the real Zika virus data from Brazil given in [5] to estimate these two values.

We assume that a single Zika infected human enters the disease free population at some time t_0, where $t_0 < t_1$ and t_1 is the first time when we have available Zika data values in Brazil obtained from [5] as the first week in 2015. We have estimated t_0 and hence $S_H(t_1), I_H(t_1), R_H(t_1), S_v(t_1), L_v(t_1)$ and $I_v(t_1)$, used as simulation starting values, by least squares. The basic reproduction number for our delayed Zika model given in [10] is defined as

$$R_0 = \frac{ma^2bce^{-\mu_v\tau}}{\mu_v(\mu_H+\gamma)}, \tag{2}$$

where all the parameter values are defined as in Table 1. For a in the range 0.7 – 3.5/week we get R_0 in the range 1.27 – 11.01/week.

1.2 Numerical Solutions

Once all the parameter values are obtained and estimated, we use R to solve the differential equations given in Equation (1) and produced simulations which illustrate the number of susceptible, infected and recovered individuals over both a short time period, to represent the immediate future, and over a long time period, to represent what happens when the endemic equilibrium has been reached.

We focus on analysing the effect of pregnant women infected with Zika virus during their first trimester as various reports (e.g. [2, 8]) suggest that pregnant women who are infected with the Zika virus during the first trimester have a much higher risk of their babies developing Microcephaly as opposed to those who are infected with Zika in their second or third trimesters. We have obtained an estimated expected future number of cases of Microcephaly due to pregnant women infected with Zika during their first trimester both in the short and in the long term.

1.3 Model With Seasonality

It is well-known that the life cycle of *A. Aegypti* is influenced by many environmental factors such as rainfall and temperature (e.g. [7, 11, 13]). As a result, in order to fully capture the behaviour of the *A. Aegypti* mosquitoes under the influence of environmental factors and its effect on the number of Microcephaly cases, later on we decided to improve on our model by adding seasonality into the birth function of *A. Aegypti* mosquitoes. Similarly, with this seasonality model, we use the least squares estimation technique to parameter

 ISBN: 978-84-617-8694-7

estimate new values of b and c, and thus calculate the future expected number of cases of Microcephaly both in the short and long term due to pregnant women being infected in their first trimester.

1.4 Results

For both models, numerical simulations are produced to illustrate the spread of the Zika virus over a period of time and the future expected number of cases of Microcephaly as a result of the Zika virus are calculated. The suggested value of $a = 3.5$/week for the parameter a, the *A. Aegypti* biting rate, is high compared to the other values in the literature. So we discuss the sensitivity of our results to different values of this parameter. We will later extend the results to an age-structured model.

Acknowledgements

The authors are grateful to the EPSRC and the University of Strathclyde for support for this work under the EPSRC Global Challenges Research Fund Institutional Award 2016 (EPSRC grant reference number EP/P511055/1) and the British Council, Malaysia for funding from the Dengue Tech Challenge (Application Reference DTC 16022). DG is grateful to the Science Without Borders Program, Brazil, for a Special Visiting Fellowship (CNPq grant 30098/2014-7), with Professor E. Massad, Department of Legal Medicine, University of Sao Paulo, Sao Paulo, Brazil, and to the Leverhulme Trust for support from a Leverhulme Research Fellowship (RF-2015-88).

References

[1] M. Andraud, N. Hens, C. Marais and P. Beutels, *Dynamic epidemiological models for Dengue transmission: a systematic review of structural approaches*, PLoS one **7(11)** (2012) e49085.

[2] S. Cauchemez, M. Besnard, P. Bompard, T. Dub, P. Guillemette-Artur, D. Eyrolle-Guignot, H. Salje, M. D. V. Kerkhove, V. Abadie, C. Garel, A. Fontanet and H. Mallet, *Association between Zika Virus and Microcephaly in French Polynesia, 2013-2015: A retrospective study*, The Lancet **387(10033)** (2016) 2125–2132.

[3] E. Chikak and H. Ishikawa, *A Dengue transmission model in Thailand considering sequential infections with all four serotypes*, J. Inf. Develop. Countries **3(09)** (2009) 711–722.

 ISBN: 978-84-617-8694-7

[4] Centers for Disease Control and Prevention, *Surveillance and control of Aedes Aegypti and Aedes Albopictus in the United States*, 16 pages. Vancouver, 2017. Retrieved from https://www.cdc.gov/chikungunya/resources/vector-control.html, last accessed on 20th March 2017.

[5] N. M. Ferguson, Z. M. Cucunub, I. Dorigatti, G. L. Nedjati-Gilan, C. A. Donnelly, M. G. Basez, P. Nouvellet and J. Lessler, *Countering the Zika epidemic in Latin America*, Science **353(6297)** (2016) 353–354.

[6] D. Focks, N. Alexander and E. Villegas, *Multicountry study of Aedes Aegypti pupal productivity survey methodology: findings and recommendations*, UNICEF/UNDP/World Bank/WHO Special Programme for Research and Training in Tropical Disease (TDR), (2006), retrieved from http://www.who.int/tdr/publications/documents/aedes_aegypti.pdf, last accessed on 20th March 2017.

[7] J. Liu-Helmersson, H. Stenlund, A. Wilder-Smith and J. Rocklv, *Vectorial capacity of Aedes Aegypti: effects of temperature and implications for global Dengue epidemic potential*, PLoS one **9(3)** (2014) e89783.

[8] M. A. Johansson, L. Mier-y-Teran-Romero, J. Reefhuis, S. M. Gilboa and S. L. Hills, *Zika and the risk of of Microcephaly*, New Eng. J. Med. **375(1)** (2016) 1–3.

[9] E. Massad, F. A. B. Coutinho, M. N. Burattini and M. Amaku, *Estimation of R_0 from the initial phase of an outbreak of a vector-borne infection*, Trop. Med. Int. Health **15(1)** (2010) 120–126.

[10] E. Massad, F. A. B. Coutinho, M. N. Burattini and L. R. Lopez, *The risk of Yellow Fever in a Dengue-infested area*, Trans. Roy. Soc. Trop. Med. Hyg. **95** (2001) 370–374.

[11] H. S. Rodrigues, M. T. Monteiro and D. M. Torres, *Seasonality effects on Dengue: basic reproduction number, sensitivity analysis and optimal control*, Math. Meth. Appl. Sci. **39(16)** (2016) 4671–4679.

[12] World Health Organization, *Countries: Brazil*, 2017 Retrieved from http://www.who.int/countries/bra/en/, last accessed on 20th March 2017.

[13] S. Wiwanitkit and V. Wiwanitkit, *Predicted pattern of Zika Virus infection distribution with reference to rainfall in Thailand*, Asian Pacif. J. Trop. Med. **9(7)** (2016) 719–720.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

Efficient Consistency Library for Multiple Sequence Alignment Tools

Jordi Lladós[1], Fernando Cores[1] and Fernando Guirado[1]

[1] *INSPIRES Research Center, Universitat de Lleida, Jaume II. 69, 25001 Lleida, Spain*

emails: jordi.llados@diei.udl.cat, fcores@diei.udl.cat, f.guirado@diei.udl.cat

Abstract

With the advent of new high-throughput next generation sequencing technologies, the volume of genetic data processed has increased significantly. It is becoming essential for these applications to achieve large-scale alignments with thousands of sequences or even whole genomes. However, all current MSA tools have exhibited scalability issues when the number of sequences increases. The main drawback of these methods is that errors made in early pairwise alignments are propagated to the final result, affecting the accuracy of the global alignment. The use of consistency information allows the final result to be improved and makes it more stable from the accuracy point of view. However, such methods are severely limited by the memory required to store the consistency information. In the present paper, we use evolutionary algorithms to analyze and determine the optimal consistency data that must be stored with the dual aim of maximizing the quality of the resulting alignment, while also reducing the memory requirements.

Key words: Multiple Sequence Alignment, Memory Efficiency, T-Coffee, Consistency, Accuracy

1 Introduction

The Multiple Sequence Alignment (MSA) is gaining importance in the analysis of biological sequence data. Phylogenetic tree reconstruction ([3]), structure prediction ([8]) or hidden Markov modeling ([2]) require MSA to infer residue-level homology or structural or functional identity.

The alignment of two sequences can be done optimally using Dynamic Programming. However, for a greater number of sequences, the alignment was shown to be a non-deterministic

 ISBN: 978-84-617-8694-7

polynomial (NP)-complete problem [15], requiring the utilization of heuristics algorithms. In this case, the goal of MSA is to find an alignment that maximizes its accuracy, approximated by the sum of similarities for all pairs of sequences (SP score, [5]).

Among the different MSA approaches, progressive alignment is the most prevalent for large data sets. Progressive alignment builds up a final MSA by combining pairwise alignments beginning with the most similar pair and progressing, following a guide tree, to the most distantly related. The main drawback of these methods is that errors made in early pairwise alignments are propagated to the final result, thus affecting the accuracy of the global alignment. To lessen the early-error propagation, consistency-based methods were proposed.

Consistency-based methods use consistency information from different pairwise alignments to improve the final result. However, such methods are severely limited by the memory requirements needed to store the consistency information. In T-Coffee (TC) [9], the most representative method in this category, the consistency-library size is in the order of $O(N^2L^2)$, N being the number of sequences and L, the length of the sequence, which limits its performance and scalability considerably.

Our final goal is to develop an algorithm able to select the best consistency information for aligning the sequences, with the aim of reducing memory requirements but maintaining the accuracy of the original method. To this end, in this paper we analyzed the individual impact of consistency on the alignment accuracy.

Achieving these objectives requires a deeper analysis of the consistency, identifying the scores that compose the optimal consistency library, which is very useful in the alignment process. Finally, based on these, a new method is proposed for selecting/filtering where consistency information must be used, with the dual aim of maximizing the quality of the resulting alignment and reducing the memory requirements.

The paper is organized as follows: Section 2 presents a brief state of the art of consistency-based MSA tools. In Section 3, we define the problem statement. In Section 4, we explore the consistency distribution and significance. From these analyses, two new methods are derived, and these are presented in Section 5. The performance and accuracy evaluation are shown in Section 6 and finally, the main conclusions are presented in Section 7.

2 State of art

The consistency based MSA has been shown to be able to increase final alignment accuracy. In [4], O. Gotoh first introduced consistency to identify anchor points for reducing the search space of an MSA. Since then, some MSA tools based on consistency have appeared in the literature.

Do et al. presented ProbCons in [1]. This was a modification of the traditional sum-of-pairs scoring system that incorporates Hidden Markov Models to specify the probability

 ISBN: 978-84-617-8694-7

distribution over all alignments between a pair of sequences. Furthermore, Subramanian et al. developed a new tool, DIALIGN-T, in [13], which formulated consistency based on finding ungapped local alignments via segment-to-segment comparisons that determine new weights using consistency. Notredame et al. presented T-Coffee. This improves the alignment accuracy by seeking consistency from a set of global and local pairwise alignments. The scoring function for aligning two sequences or two pre-aligned groups is determined by the whole set of sequences via two processes called library generation and library extension. Although T-Coffee can produce high alignment accuracy, the consistency library is time and memory consuming when the number of sequences is large. Another method based on consistency, MAFFT, was presented by Katoh et al. in [6]. This uses a new objective function combining the WSP score from Gotoh and the COFFEE-like score ([10]), which evaluates the consistency between a multiple and pairwise alignments.

However, it is known that when the number of sequences to be aligned increases, there is a degradation of accuracy [12]. Other studies focusing on phylogeny estimation from nucleotide datasets, have confirmed this hypothesis [7]. This situation can be mitigated by the use of consistency. However, consistency-based methods do not scale well because of the computational resources required to calculate and store the consistency information, which grows quadratically.

3 Consistency stage

Consistency-based methods use the point of view that prevention is the best way to avoid errors in early stages of the alignment. However, next-generation sequencing applications are unable to take advantage of this improvement due to the bottleneck their huge memory requirements represent. This problem is evident in the analysis of homologous sequences. This is especially troublesome for marker genes, like the ribosomal RNA (rRNA) where millions of sequences are already publicly available and individual studies can easily produce hundreds of thousands of new sequences [11].

Newer methodologies to use consistency have been developed since COFFEE. Although some are faster, the consistency data is not as tractable as in TC. It allows input/output of libraries and also generate the library without aligning the sequences. In this paper, we use TC to generate the information needed to complete our analysis.

3.1 Consistency calculation

The consistency library information is a collection of pairwise alignments obtained from computing all-against-all pairwise alignments. It can be represented by an NxN matrix (see Figure 1), where each cell $S_i - S_j$ when $i \neq j$ contains a list of residue matches between those sequences. Each residue match is represented by a constraint/entry $\{x, y, W_{(x,y)}\}$, x being a residue of S_i matched with y a residue of S_j and a weight $W_{(x,y)}$ representing its

 ISBN: 978-84-617-8694-7

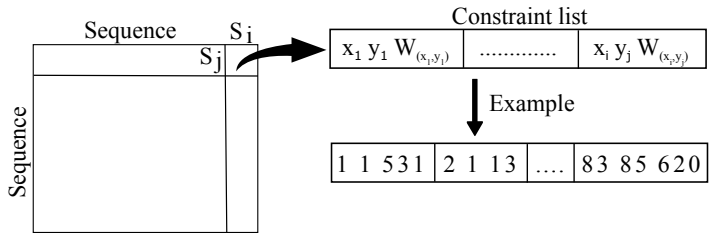


Figure 1: Library structure.

correctness. Each constraint list is used in the progressive alignment stage in order to fill the dynamic programming matrix.

The size of the consistency library is in the order of $O(N^2L^2)$, where N^2 is given by all the possible combinations of sequences without repetition and L^2 by the worst scenario in one pairwise (there is no matches between both sequences). In TC, this process is implemented in parallel as there are no dependencies between each position in the matrix.

3.2 Consistency analysis

It is impossible to calculate the optimal library. Thus, there are two alternatives for dealing with consistency: 1) Retain all the constraint information as in T-Coffee, which requires high computational resources; 2) Select only a subset of the library in order to reduce these requirements and improve efficiency. We explore this second approach, searching for a method capable of selecting a representative subset of data for a given library.

To make this possible, we need to analyze the consistency and discover patterns that allow us to discard information that does not generate added value for the alignment, preserving the consistency and ensuring the final quality. With this in mind, we designed a set of experiments to explore several consistency features. Firstly, we show the importance of the constraint selection policy. Secondly, we use evolutionary algorithms as an approximation to the optimal consistency library and then carry out a deeper study of the consistency from different points of view.

In order to carry out this analysis, a reliable database is needed. The choice is BAliBASE [14], a database of high-quality documented and manually-refined reference alignments based on 3D structural superpositions. The accuracy of the alignments is measured using two accuracy metrics: the Sum-of-Pairs (SP) and the Total Column Score (TCS), which are obtained by comparing the user alignment with a reference alignment.

We show all the results with the smallest data set included in BAliBASE, BB11001, so we are able to show more readable graphics, as the other datasets contains up to hundred sequences and longer residues per sequence. BB11001 is composed of 4 sequences and the original consistency method from TC generates 1540 constraints.

 ISBN: 978-84-617-8694-7

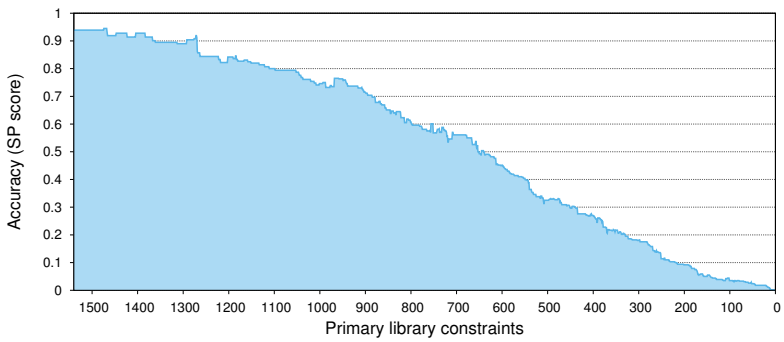


Figure 2: Random discarding policy when solving the BB11001 data set.

3.2.1 Influence of the constraints selection

First of all, we wished to demonstrate the effect of discarding the constraints of an alignment. In order to generate this test case, we developed a script which (1) executes TC and generates the output library for the input dataset and scores it, (2) removes a constraints from the library, (3) executes TC with the input library and generates a new output library scoring it and finally (4) repeats step (2) until the library is empty. In Figure 2, the alignment quality is plotted as the library size is decreased. It can be seen how a randomly discarding method behaves. Thus, it seems to be following a linear fall. When more constraints are discarded, less accuracy is achieved. For this reason, it is mandatory to have a policy capable of deciding which residue matches have to be discarded, because if not, the accuracy of the alignment is highly compromised.

The next important point is whether the residues (library information) are evenly distributed between sequences and along the alignment. From the point of view of the residue distribution in the sequences, we can observe that the number of constraints generated for each one is similar, all four of them having almost the same number of constraints, these being 853, 743, 701 and 783 respectively.

However, if we analyze the distribution of all the library residues over the alignment, we can observe more variability. Figure 3 shows that all the residue matches are grouped by their position regardless of the pairwise. A maximum (40 residues around positions 73-82) and a relative maximum (22 residues around positions 30-37) can be distinguished. These regions tend to be the most problematic areas of the alignment, because the more information you have, the more choices there are where the alignment may split. It must be decided if the histogram shape needs to be maintained when the library is reduced (maintaining more consistency in the maximums), or it could be better to trim the peaks and balance the histogram to the same level.

Finally, the distribution of the BB11001 library regarding the weighting score of the

 ISBN: 978-84-617-8694-7

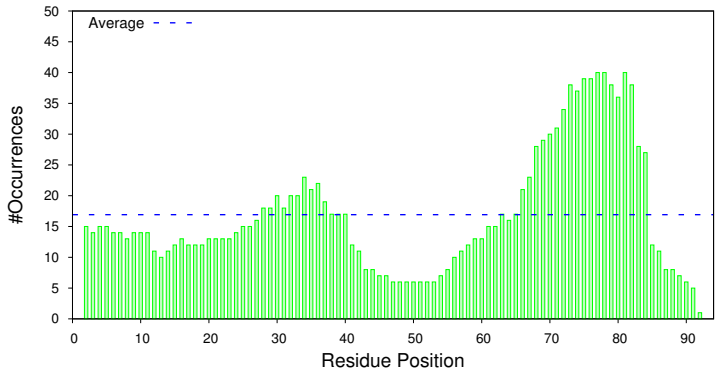


Figure 3: Constraints distribution in the Library when solving the BB11001 data set.

Table 1: Constraint weight distribution when solving the BB11001 data set.

Weight	*Occurrences*	*Percentage*	*Weight*	*Occurrences*	*Percentage*
999-900	247	16.04	499-400	42	2.73%
899-800	34	2.21	399-300	66	4.29%
799-700	47	3.05	299-200	72	4.68%
699-600	22	1.43	199-100	156	10.13%
599-500	33	2.14	99-1	821	53.31%

constrains is analyzed in Table 1. We can observe that the most outstanding weights seem to be the higher and the lower ones, while the intermediate values have a similar number of occurrences. Although the lower weighted constraints are the least valuable, regarding their correctness, half the library is filled with these. It has to be decided whether a range of weights must maintain a higher proportion of values, or if they need to be distributed equally.

3.2.2 Analyzing optimal libraries

For the following test, we need to obtain a representative approximation of the optimal library. It was decided to design a genetic algorithm (GA) to look for the best library constraints[1]. Our evolutionary gene defines which constrains are selected in the library. The genetic algorithm developed is capable of obtaining the best possible subset of constraints for a given size of the library.

The first execution showed that the GA was able to obtain an accuracy of 1 (perfect) with a subset size of 1540, meaning that some constraints add noise to the alignment (some

[1]The GA sources and its installation instructions can be found at: github.com/jllados/CL-GA.

 ISBN: 978-84-617-8694-7

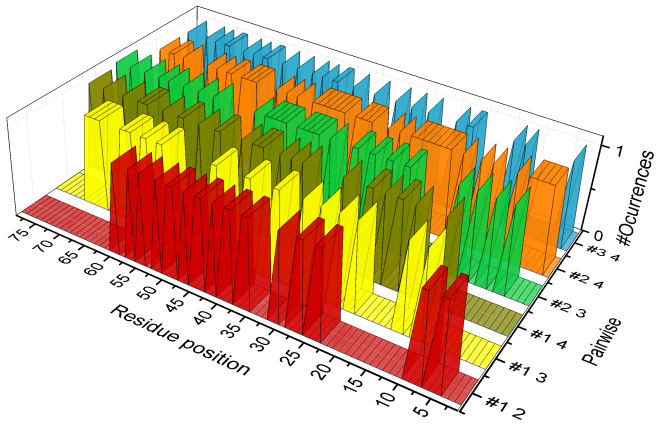


Figure 4: Residues selected with GA in the library for each pairwise in the data set BB11001 with 100 constraints (Accuracy 0.436).

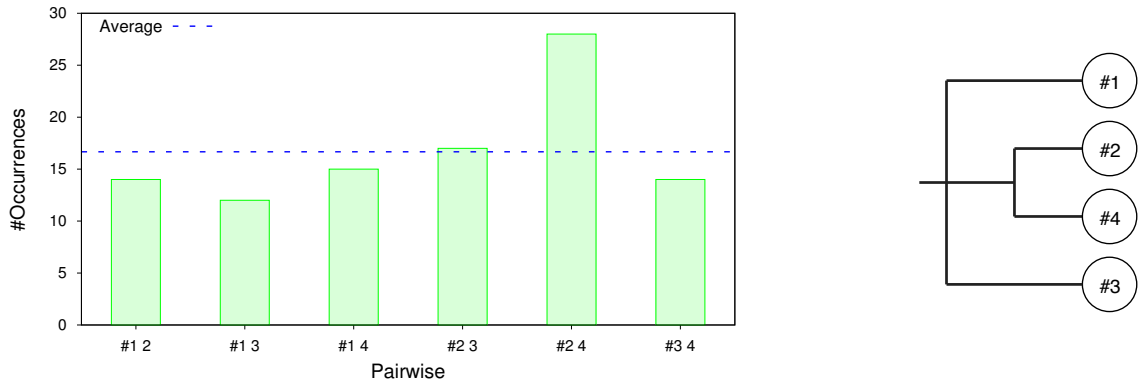
constraints where repeated). This accuracy was consistent until the gene size was set below 250, so just 250 constraints were actually needed to obtain a perfect alignment.

Next, we decided to go further. We wanted to observe the behavior of the library under different situations. The following test consisted of generating 96 possible libraries with a given size of 100 constraints. On the other hand, we decided to use one hundred constraints, a few more than the length of each sequence in the data set. The reason was that we wanted to observe how well the selected residues from each constraint were distributed over the alignment length.

Figure 4 was generated using the BB11001 library as input and a given maximum library size of 100 constraints. It shows us that all the pairwise seem to have a rather similar number of constraints, #2 4 being the one with the most. This fact was interesting as in almost every data set analyzed, some pairwise always had a little more information than the others. Furthermore, there are no repeated occurrences in the data. Another fact is that we cannot see blocks of data. The constraints seem to be distributed over the length of the alignment. This is an indicator that the selected constraints may have a variety of weights (frequently the most heavily weighted constraints seem to be concentrated in the same region). This solves the previous doubt about Figure 3, we should balance the histogram among the residues.

Then we need to look at the constraints weight distribution. They were separated into different groups. Each group represents the weight range and a the average number of occurrences over the 96 test: 999-900: 64, 899-800: 4, 799-700: 8, 699-600: 1, 599-500: 3, 499-400: 2, 399-300: 3, 299-200: 2, 199-100: 3 and 99-1: 7. In this particular case, 64%

 ISBN: 978-84-617-8694-7



(a) Pairwise constrains distribution.

(b) Guide tree.

Figure 5: BB11001: Pairwise occurrences and guide tree.

of the library is composed of the most weighted ones, while the rest is well distributed. In comparison with Table 1, were the library contained the 1540 constraints, the lower ranges are drastically reduced and a bigger proportion is maintained for the highest ones.

Regarding the fact of having more data in one pairwise, Figure 5a shows that #2 4 almost doubles the occurrences of the rest. The reason can be observed in Figure 5b, as it is known that MSA aligners follow the order of a guide tree to generate the alignment. It is reasonable that the first couple of sequences to be aligned (the most closely related ones) contain more information. We must ensure its accuracy to avoid propagating an error to the rest of profile alignments.

3.3 Memory efficient consistency library

The following premises were extracted from the previous analysis, in order to implement the new method for build de consistency library:

- The higher the weight, the better.
- The most closely related leaves of the guide trees must have more constraints.
- The constraints have to cover all the domain of the alignment.

These premises were used to define the pattern that we followed in the design of the method, named Memory Efficient consistency Library (MEL), to generate an algorithm capable of deciding whether or not a constraint should be maintained.

The implementation of MEL, Algorithm 1, is based on a temporary queue structure, in which all the library constraints are stored sorted by their weight. After this, each constraint is evaluated in order to determine if it must be pushed into the final list. The evaluation

for *each sequence* $S_i \in S_1..S_N$ *and* $S_i \neq S_j$ **do**
  for *each sequence* $S_j \in S_i..S_N$ *where* $S_i \neq S_N$ **do**
    PA_{ij}=Pairwaise_Alignment_Fork(S_i, S_j);
    for *each residue* $x \in S_i, y \in S_j|$ *are aligned in* $PA_{i,j}$ **do**
      $W_{(x,y)} = \frac{\sum OCCURRENCE(PA_{i,j})}{RESIDUES(PA_{i,j})}$;
      Q_Push_Sorted($x, y, W_{(x,y)}$));
    end
    MAX_Constraints_$PA_{ij} = \frac{MAX_Library}{C_{N,2}}$;
    Bound=Ceil(MAX_Constraints_PA_{ij},MAX_Length(S_i, S_j));
    if *isleaf*(S_i, S_j) **then**
      MAX_Constraints_PA_{ij}+=remainder;
    end
    for *each constraint* $x_i, y_j, W_{(xi,yi)} \in Q$ **do**
      if *balanced*($x_i, y_j, Bound$)$|$ *not full MAX_Constraints_*PA_{ij} **then**
        $L(S_i^x, S_j^y) = W_{(S_i^x, S_j^y)}$;
      end
    end
  end
end

Algorithm 1: Memory Efficient consistency Library (MEL) construction

function has to check a pair of parameters, the first being the maximum bound previously generated. With this value, the method prunes the number of residues matching the same region of the alignment. The second parameter indicates whether or not a new entry has enough memory to be allocated. Also, if the evaluated pairwise PA_{ij} is a leaf node, the remainder lost allocating each pairwise its added to the maximum number of constraints. As the queue is sorted by weight, the higher ones will be the first to be evaluated, thus ensuring that they have more chances of surviving than the lower ones.

Our functions and data structures were adapted to make use of the current parallelism implemented in TC.

4 Experimentation

In this section, we evaluate MEL. This experimental study evaluates (1) the effectiveness of the discarding method, (2) the accuracy obtained from our proposal decreasing the amount

Table 2: BAliBASE accuracy results with MEL.

Entries	%	*Average*	*RV*11	*RV*12	*RV*20	*RV*30	*RV*40	*RV*50	*Time*
157M	100.00	**0.746**	0.534	**0.879**	**0.827**	0.718	0.758	**0.759**	45689.91
134M	84.89	0.745	0.534	0.878	0.826	0.715	0.760	0.755	40673.9
125M	79.09	0.745	**0.535**	0.877	0.826	0.716	0.760	0.756	39721.54
114M	72.62	0.744	0.532	0.878	0.826	0.717	0.758	0.755	38926.47
103M	65.41	0.745	0.528	0.875	0.825	0.717	0.767	0.757	37202.72
90M	57.42	0.745	0.527	0.875	0.826	**0.723**	0.767	0.753	35834.48
77M	48.73	0.745	0.525	0.872	0.822	0.721	**0.775**	0.752	34543.23
62M	39.46	0.740	0.521	0.865	0.817	0.719	0.773	0.747	34369.66
47M	29.74	0.725	0.511	0.833	0.794	0.715	0.766	0.730	**32731.14**
31M	19.86	0.672	0.471	0.734	0.728	0.678	0.731	0.688	33726.12
16M	9.93	0.512	0.349	0.494	0.553	0.552	0.579	0.546	46617.86

of library/pairwise and finally (3) the global performance of T-Coffee[2].

The tests consisted of solving BAliBASE. The figures are total Sum-of-Pairs (SP) produced using bali score. The first column indicates the number of entries used in the run and the second the percentage reduction. The average score over all families is given in the third column. The results for BAliBASE subgroupings are in columns 4–9. Finally, the last column refers to the total execution time.

First, we present the results produced by reducing the total numbers of entries on BAliBASE with MEL. These are shown in Table 2. We must highlight the results between 84.89% and 48.73%. In comparison with the full library accuracy, all these cases behave quite similarly, losing on average 0.002 at most. Despite this, the library is reduced by 50%, which is a far greater improvement.

In general, we must say that the reduced scenarios attain a better accuracy with RV11 (equidistant sequences of similar length with very divergent conservation <20% identity) and a better gain is achieved with RV30 (equidistant divergent families with <25% residue identity between groups) and RV40 (sequences with N/C-terminal extensions). Furthermore, concerning time, we can see an inverted bell curve, fewer entries implies less execution time, until the minimum is achieved (47M entries), where time increases.

Next, we evaluated the above-mentioned random discard policy and the best possible scenario against the proposal presented. Figure 6 shows the accuracy of aligning as the number of entries decreases. The behavior is clear, the greater the reduction applied, the more pronounced the impact becomes. As can be seen, MEL is notably superior to the random one, so it is clear enough that a discarding policy is worth it. Even though there is a gap between our method and the GA, because GA uses the reference alignment and none algorithm will be able to reach that result. Despite this, both shapes are quite similar.

[2]T-Coffee sources and its installation instructions can be found at: github.com/jllados/TCoffee-MEL

 ISBN: 978-84-617-8694-7

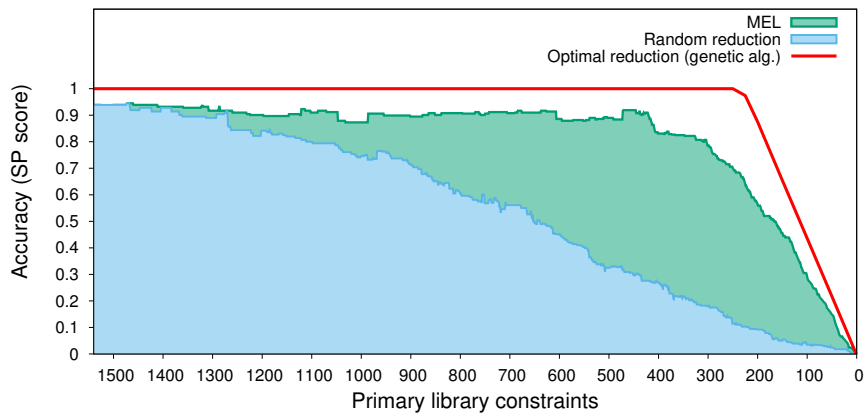


Figure 6: Comparison of MEL with the Optimal and a random policy for BB11001 dataset.

5 Conclusions

In this paper, the authors present a method to build the consistency library of a Multiple Sequence Aligner. The approach is applied during the process of building the library. Its goal is to maintain the best consistency information while reducing the size of the memory that the library may use without affecting the accuracy.

We proved that MEL is able to filter the consistency library more efficiently than the current methods. This has a positive effect on an MSA aligner, because the fewer the constraints used, the smaller the amount of memory needed. One of the best scenarios occurs when the library maintains around 50% of the entries. This means that we are able to reduce the memory requirements by half while obtaining almost the same results.

In the future, we must integrate consistency with a more basic MSA tool in order to avoid errors in the early stages of the alignment and maintain the accuracy.

Acknowledgements

This work has been supported by the MEyC-Spain under contract TIN2014-53234-C2-2-R and TIN2016-81840-REDT.

References

[1] Do C, Brudno M and Batzoglou S. PROBCONS: Probabilistic Consistency-based Multiple Alignment of Amino Acid Sequences. *Proceedings Nineteenth National Conference on Artificial Intelligence*, pages 703-708, 2004.

[2] Eddy SR. A new generation of homology search tools based on probabilistic inference. In *Genome Inform*, volume 23, pages 205-211, 2009.

[3] Gouy M, Guindon S and Gascuel O. Seaview version 4: A multiplatform graphical user interface for sequence alignment and phylogenetic tree building. *Molecular Biology and Evolution*, 27(2):221-224, 2010.

[4] Gotoh O. Consistency of optimal sequence alignments. *Bulletin of Mathematical Biology*, 52(4):509-525, 1990.

[5] Just W. Computational complexity of multiple sequence alignment with sp-score. *Journal of computational biology*, 8(6):615-623, 2001.

[6] Katoh K, Misawa K, Kuma K and Miyata T. Mafft: a novel method for rapid multiple sequence alignment based on fast fourier transform. *Nucleic Acids Research*, 30(14):3059-3066, 2002.

[7] Liu K, Linder CR and Warnow T. Multiple sequence alignment: a major challenge to large-scale phylogenetics. *PLoS currents*, 2, 2010.

[8] Marks, D.S., Hopf, T.A. and Sander C. Protein structure prediction from sequence variation. *Nat Biotech*, 30(11):1072-1080, 11 2012.

[9] Notredame C, Higgins DG and Heringa J. T-coffee: A novel method for fast and accurate multiple sequence alignment. *Journal of molecular biology*, 302(1):205-217, 2000.

[10] Notredame C, Holm L and Higgins DG. Coffee: an objective function for multiple sequence alignments. *Bioinformatics*, 14(5):407-422, 1998.

[11] Pruesse E, Peplies J and Glöckner FO. SINA: accurate high throughput multiple sequence alignment of ribosomal RNA genes. *Bioinformatics*, 28(14):1823-1829, 2012.

[12] Sievers F, Dineen D, Wilm A and Higgins DG. Making automated multiple alignments of very large numbers of protein sequences. *Bioinformatics*, 29(8):989-995, 2013.

[13] Subramanian AR, Weyer-Menkhoff J, Kaufmann M, et al. Dialign-T: an improved algorithm for segment-based multiple sequence alignment. *BMC Bioinformatics.*, 66(6), 2005.

[14] Thompson J.D., Plewniak F. and Poch O. Balibase: a benchmark alignment database for the evaluation of multiple alignment programs. *Bioinformatics*, 15(1):87-88, 1999.

[15] Wang L. and Jiang T. On the complexity of multiple sequence alignment. *Journal of Computational Biology*, 1(4):337-348, 1994.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

Finite-time consensus of uncertain multi-agent systems

Vincenzo Loia[1] and Stefania Tomasiello[2]

[1] *DISA-MIS, University of Salerno, Italy*

[2] *CORISA, University of Salerno, Italy*

emails: loia@unisa.it, stomasiello@unisa.it

Abstract

In this paper we investigate the consensus of a new class of uncertain multi-agent systems, according to the Liu's uncertainty theory, as a counterpart of stochastic multi-agent systems. We revise the concept of finite-time stability in the context of uncertainty theory.

Key words: uncertain differential equation, stability, Liu process

1 Introduction

In the last decades, analysis and control design of multi-agent systems (MAS) have become very popular, because of their applications including traffic control, sensor networks, mobile robots and social networks.

In particular, stochastic MAS attracted attention due to the fact that in real-world applications, the behaviour of dynamical systems is concretely affected by disturbances and uncertainties because of the unpredictable environmental conditions (e.g. see [1]).

Stability problems, also known as consensus issues, for such class of MAS have been widely discussed. Consensus means the states of all the agents converge to a common value according to some control schemes. It is usually referred to the asymptotic behaviour of the trajectories of the system as time goes to infinity. The concept of finite-time stability for stochastic systems was introduced in [2] and adapted to the consensus problem in MAS (e.g. see [3]), in order to get a faster convergence rate.

Recently, an uncertainty theory has been proposed by Liu [4]. In particular, Liu [5] introduced the concept of uncertain process, in order to describe the evolution of an uncertain phenomenon, and designed a canonical Liu process [6], which is an uncertain process

 ISBN: 978-84-617-8694-7

with stationary and independent normal uncertain increments. Liu [6] founded the uncertain calculus as a counterpart of Ito calculus, by introducing a type of uncertain differential equation (UDE) driven by canonical Liu process.

In this work, we aim to introduce a new class of uncertain MAS, according to the Liu's theory, as a counterpart of stochastic MAS.

2 Preliminaries

Definition 1 *Let $\mathcal{L}$ be a σ-algebra on a nonempty set Γ. The uncertain measure $\mathcal{M}$ is a set function $\mathcal{M}: \mathcal{L} \to [0,1]$, satisfying three axioms:*

- $\mathcal{M}\{\Gamma\} = 1$ *for the universal set Γ (normality axiom);*
- $\mathcal{M}\{\Lambda\} + M\{\Lambda^c\} = 1$ *for any event Λ (duality axiom);*
- *for every countable sequence of events $\Lambda_1, \Lambda_2, \ldots$, we have* $\mathcal{M}\bigcup_{i=1}^{\infty}\Lambda_i \leq \sum_{i=1}^{\infty}\mathcal{M}\{\Lambda_i\}$ *(subadditivity axiom).*

The triplet $(\Gamma, \mathcal{L}, \mathcal{M})$ is called an uncertainty space.

Definition 2 *[6] An uncertain process C_t is said to be a canonical Liu process if*

- $C_0 = 0$ *and almost all sample paths are Lipschitz continuous;*
- C_t *has stationary and independent increments;*
- *every increment $C_{s+t} - C_s$ is a normal uncertain variable with expected value 0 and variance t^2, whose uncertainty distribution is*

$$\Phi_t(x) = \left(1 + \exp(-\frac{\pi x}{\sqrt{3}t})\right)^{-1} \tag{1}$$

Definition 3 *[6] Let C_t be a canonical Liu process, f and g two given functions. Then*

$$dX = f(t,X)dt + g(t,X)dC_t \tag{2}$$

is called an uncertain differential equation.

The UDE (2) has a unique solution if the functions $f(t,X)$ and $g(t,X)$ satisfy [7]

 ISBN: 978-84-617-8694-7

- the linear growth condition

$$|f(t,X)| + |g(t,X)| \leq K(1+|X|), \forall X \in \mathbb{R}, t \geq 0 \tag{3}$$

- the Lipschitz condition

$$|f(t,X) - f(t,Y)| + |g(t,X) - g(t,Y)| \leq K|X-Y|, \quad \forall X \in \mathbb{R}, t \geq 0 \tag{4}$$

  for any constant K.

Obviously, the UDE (2) admits a trivial solution.

Definition 4 *[6] An uncertain differential equation is said to be stable if for any solution X and Y, with initial values X_0 and Y_0 respectively, any given (though arbitrarily small) $\xi, \epsilon > 0$, there exists $\delta > 0$ s.t.*

$$\mathcal{M}\{|X-Y| > \xi\} < \epsilon, \qquad \forall t > 0 \tag{5}$$

whenever $|X_0 - Y_0| < \delta$.

In the following, we will revise some definitions in the context of the uncertainty theory.

We first introduce a new definition of finite-time stability for the UDE (2). This definition represents the counterpart of the one for stochastic nonlinear systems [2].

Definition 5 *The trivial solution of UDE (2) is said to be finite–time stable, if the equation admits a unique solution, say $X(t;X_0)$, for any initial value $X_0 \in \mathbb{R}^n$ s.t.*

- *the corresponding settling time τ_{X_0} is finite, that is $\mathcal{M}\{\tau < \infty\} = 0$;*
- *the UDE is stable.*

All the definitions above can be easily extended to the multi-dimensional case [8].

3 Problem formulation

Assumption 6 *Let C be a canonical Liu process. There exists a continuous function $\eta(t)$ s.t. $dC{=}\eta(t)dt$.*

Remark 7 *Let us consider, for instance, the continuously differentiable function $h(t,C) = tC$. By fixing $dh = Cdt + tdC = 0$, then the assumption is trivially true.*

 ISBN: 978-84-617-8694-7

Let the dynamics of a first-order multi-agent system along an undirected graph V be described as follows

$$\dot{x}_i = f(x_i) + u_i + g(x_i)\eta(t), \qquad x_i(0) = x_{0i}, \qquad i \in V \tag{6}$$

where $x_i \in \mathbb{R}$ is the state of the ith agent, $X^T = (x_1, x_2, \ldots, x_n)$ is the state of the whole network, u_i is the local control input and x_{0i} are the initial conditions. Henceforth it is assumed that the functions $f(x_i)$ and $g(x_i)$ satisfy the conditions (3) and (4).

Due to the Assumption 1, Eq. (6) can be written as follows

$$dx_i = (f(x_i) + u_i)dt + g(x_i)dC, \qquad x_i(0) = x_{0i}, \qquad i \in V \tag{7}$$

Now we introduce the finite-time quasi–consensus.

Definition 8 *The first-order finite-time consensus is achieved if there exists a settling time T, satisfying $\mathcal{M}\{T < \infty\} = 0$, such that*

$$\mathcal{M}\{x_i = x_j, \forall t \geq T\} = 0, \qquad \forall i \neq j = 1, \ldots, n. \tag{8}$$

In order to discuss the finite-time consensus, one has to prove that the trivial solution of the UDE (7) in terms of errors $e_i = x_i - \frac{1}{n}\sum_{j=1}^{n} x_j$ is finite–time stable.

References

[1] P. Ming et al., *Consensus stabilization in stochastic multi-agent systems with Markovian switching topology, noises and delay*, Neurocomputing **200** (2016) 1-10.

[2] J. Yin, S. Khoo, Z. Man, X. Yu *Finite-time stability and instability of stochastic nonlinear systems* Automatica **47(12)** (2011) 2671-2677.

[3] Y.Zheng, W.Chen, L.Wang, *Finite-time consensus for stochastic multi-agent systems*, Int J Control **84(10)** (2011) 1644-1652.

[4] B. Liu, *Uncertainty theory*, Springer, Berlin, 2007.

[5] B. Liu, *Fuzzy process, hybrid process and uncertain process*, J Uncertain Syst **2(1)** (2008) 3-16.

[6] B. Liu, *Some research problems in uncertainty theory*, J Uncertain Syst **3(1)**(2009) 3-10.

[7] X.W. Chen, B. Liu *Existence and uniqueness theorem for uncertain differential equations* Fuzzy Optim Decis Mak **9(1)** (2010) 69-81.

[8] T. Su, H. Wu, J. Zhou, *Stability of multi-dimensional uncertain differential equation*, Soft Comput, in press, DOI 10.1007/s00500-015-1788-0

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference on Computational and Mathematical Methods in Science and Engineering, CMMSE 2017 4–8 July, 2017.

Electron-nucleus cusp dressing in single-determinant wave functions

Pierre-François Loos[1], Anthony Scemama[1], Yann Garniron[1] and Michel Caffarel[1]

[1] *Laboratoire de Chimie et Physique Quantiques,, Université de Toulouse, CNRS, UPS, France*

emails: loos@irsamc.tps-tlse.fr, scemama@irsamc.tps-tlse.fr, garniron@irsamc.tps-tlse.fr, caffarel@irsamc.tps-tlse.fr

Abstract

Universal features of the many-electron wave function Φ are of continued interest to physicists and chemists, as they guide the construction of highly accurate wave functions [1, 2], explicitly correlated ansätze within F12 theory [3, 4] and accurate Jastrow factors for quantum Monte Carlo (QMC) calculations [5]. The Coulombic singularity at short interparticule distances dominates all other terms and, near the two-particle coalescence point, the behavior of Φ becomes independent of other details of the system.

Early work by Kato [6], and elaborations by Pack and Byers-Brown [7], showed that, as one electron at $\boldsymbol{r}_i$ approaches a nucleus of charge Z_A at $\boldsymbol{r}_A$, we have

$$\left.\frac{\partial \langle \Phi \rangle}{\partial r_i}\right|_{r_i=r_A} = -Z_A \, \langle \Phi \rangle|_{r_i=r_A} \, , \tag{1}$$

where $\langle \Phi \rangle$ is the spherical average of the n-electron wave function $\Phi(\boldsymbol{r}_1, \ldots, \boldsymbol{r}_n)$ about $\boldsymbol{r}_i = \boldsymbol{r}_A$.

To remove divergences in the local energy $\Phi^{-1}\hat{H}\Phi$ at the electron-nucleus coalesence points, cusp conditions such as (1) must be satisfied. These divergences are especially harmful in diffusion QMC calculations, where they can lead to a large increase of the statistical variance, population-control problems and significant biases [5].

In this talk, we propose to show how to introduce the correct electron-nucleus cusp within single-determinant wave functions (such as Hartree-Fock wavefunction) via a dressing of the Fock matrix. This method, involving effective Hamiltonian theory, has been shown to be also successful in other scenario [8, 9]. Illustrative examples will be given for atomic and molecular systems [10].

ISBN: 978-84-617-8694-7

References

[1] K. Frankowski and C. L. Pekeris. *Phys. Rev.*, **146** (1984) 46.

[2] D. E. Freund, B. D. Huxtable, and J. D. Morgan III. *Phys. Rev. A*, **29** (1984) 980.

[3] C. Hattig, W. Klopper, A. Kohn, and D. P. Tew. *Chem. Rev.*, **112** (2012) 4.

[4] L. Kong, F. A. Bischo, and E. F. Valeev. *Chem. Rev.*, **112** (2012) 75.

[5] N. D. Drummond, M. D. Towler, and R. J. Needs. *Phys. Rev. B*, **70** (2004) 235119.

[6] T. Kato. *Commun. Pure Appl. Math.*, **10** (1957) 151.

[7] R. T. Pack and W. Byers Brown. *J. Chem. Phys.*, **45** (1966) 556.

[8] J.-L Heuilly and J.-P. Malrieu. *Chem. Phys. Lett.*, **199** (1992) 545.

[9] J.-P. Daudey, J.-L. Heully, and J.-P. Malrieu. *J. Chem. Phys.*, **99** (1993) 1240.

[10] P. F. Loos, A. Scemama, Y. Garniron, and M. Caffarel. *in preparation.*

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

A consistent second order theory about the equilibrium figures of rotating celestial bodies

José Antonio López Ortí[1], Manuel Forner Gumbau[1] and Miguel Barreda Rochera[1]

[1] *Departmento de Matemáticas, Universidad Jaume I de Castellón*

emails: lopez@mat.uji.es, fornerm@mat.uji.es, barreda@mat.uji.es

Abstract

This paper is addressed to the study of equilibrium figures of uniform rotating celestial bodies. The study of this topic in the classical theory is based on the Laplace desideratum which, unfortunately cannot be proved.

After proving in a previous paper a first order amplitudes theory, the authors extend now their work up to second order. So, the main achievement of this work is its capacity to obtain the results up to second order in amplitudes without using the unproved Laplace desideratum needed by the classical theory.

The paper is based on the Clairaut method and the results about the deformation amplitudes obtained in the first order theory are used in our developments. From these results, by using the asymptotic properties of numerical quadrature, a consistent second order theory is obtained.

Key words: Celestial Mechanics. Figures of Celestial Bodies. Spherical Harmonics. Potential Theory.
MSC 2000: 70F15, 74Gxx.

1 Introduction

The main objective of this work is to develop a consistent second order amplitudes theory to evaluate the potential of a rotating deformable celestial body when the hydrostatic equilibrium of the system has been achieved. This case can be modeled as:

$$\overrightarrow{\nabla} P = \rho \overrightarrow{\nabla} \Psi$$
$$\triangle \Psi = -4\pi G\rho + 2\omega^2$$

 ISBN: 978-84-617-8694-7

where P is the pressure, ρ is the density, Ψ is the total potential, $\triangle$ is *Laplace* operator, G is the gravitational constant, and $\overrightarrow{\omega}$ is the system's angular velocity.

To integrate these equations in a general case of mass distribution a state equation relating pressure and density is needed.

To assess the full potential Ψ to calculate the self-gravitational potential Ω and the centrifugal potential V_c it is needed. The equilibrium configuration involves the hydrostatic equilibrium that is to say, the rigid rotation of the system corresponding to the minimum potential and, according with Kopal [4], this state involves the identification of equipotential, isobaric, isothermal and isopycnic surfaces.

To study the structure of the body, a coordinate system $OXYZ$ will be defined where O is the center of mass of the component, OX is an axis fixed in an arbitrary point of the body equator, OZ the axis parallel to angular velocity $\vec{\omega}$ and finally OY defines a direct trihedrom $OXYZ$.

For an arbitrary point P in the primary component the Clairaut coordinates are given by (a, θ, λ) where a is the radius of the sphere containing the same mass that the equipotential surface containing P and (θ, λ) the angular spherical coordinates of P.

Classical theory of this problem can see in Finlay [1], Kopal [3], [4].

To achieve our main objective two different methods are proposed in this paper: the first one, which we will call analytical method, is similar to that used by Laplace to develop the inverse of the distance between two planets, and the second one, which we have called numerical quadrature method, will be based on the asymptotic properties of numerical quadrature formulas.

The main problem to develop the total potential is the development of the self-gravitational potential. For this purpose we will proceed by using the classical development of the potential as

$$\Omega = U + V, \qquad U = G\int_{r_0}^{r_1}\int_0^{2\pi}\int_0^{\pi}\frac{dm'}{\Delta}, \quad V = G\int_0^{r_0}\int_0^{2\pi}\int_0^{\pi}\frac{dm'}{\Delta}$$

where Δ is the distance between the positions of element of mass dm' placed in P' and the point P, r_0 is the radius the sphere centered in O containing P and r_1 is the radius of the minor sphere centered in O containing the primary component. The mas element is given by $dm' = \rho r \cos\theta' d\theta' d\lambda' dr'$

2 Classical theory about the self-gravitational potential

The classical theory is based on the development of the inverse of the distance:

$$\frac{1}{\Delta} = \begin{cases} \frac{1}{r}\sum\limits_{n=0}^{\infty}\left(\frac{r'}{r}\right)P_n(\cos\gamma) & r > r' \\ \frac{1}{r'}\sum\limits_{n=0}^{\infty}\left(\frac{r}{r'}\right)P_n(\cos\gamma) & r < r' \end{cases}, \tag{1}$$

 ISBN: 978-84-617-8694-7

where γ is the angle between $\overrightarrow{OP}$ and $\overrightarrow{OP'}$. From this development we can write in the form $\Omega = \sum_{n=0}^{\infty} U_n r^n + \sum_{n=1}^{\infty} V_n r^{-n-1}$, where

$$U_n = \int_{r_0}^{r_1}\int_0^{\pi}\int_0^{2\pi} r'^{1-n} P_n(\cos\theta')dr'd\theta'd\lambda',\ V_n = \int_0^{r_0}\int_0^{\pi}\int_0^{2\pi} r'^{2+n} P_n(\cos\theta')dr'd\theta'd\lambda'. \tag{2}$$

By symmetry reasons the coordinate r is connected with the Clairaut coordinates (a,θ,λ) by $r = a(1+\sum_{n=0}^{\infty} f_{2n}(a)P_{2n}(\cos\theta))$, where $f_{2n}(a)$ are the amplitudes and $P_{2n}(\cos\theta)$ the Legendre polinomials. The classical theory assumes that

$$U_n = \frac{G}{2-n}\int_{a_0}^{a_1}\rho'\frac{\partial}{\partial a}\left[\int_0^{\pi}\int_0^{2\pi} r'^{1-n}P_n(\cos\theta')d\theta'd\lambda'\right]da',\quad if n\neq 2 \tag{3}$$

$$U_2 = G\int_{a_0}^{a_1}\rho'\frac{\partial}{\partial a}\left[\int_0^{\pi}\int_0^{2\pi}\ln(r')P_n(\cos\theta')d\theta'd\lambda'\right]da' \tag{4}$$

$$V_n = \frac{G}{n+3}\int_0^{a}\rho'\frac{\partial}{\partial a}\left[\int_0^{\pi}\int_0^{2\pi} r'^{n+3}P_n(\cos\theta')d\theta'd\lambda'\right]da' \tag{5}$$

Notice that these functions are defined in the internal and external region as the equipotential surface that contains P and it is necessary to assume these series converge in this region [2], [6]. This assumption is known as Laplace desideratum.

From these hypothesis and approaching r'^k and $\ln r'$ up to the required order in amplitudes we obtain $\Omega = \sum_{n=0}^{\infty}\sum_{m=-n}^{n}\left[F_{2n}(a)r^n + E_{2m}r^{-n-1}\right]P_{2n}(\cos\theta)$.

3 A second order consistent theory of autogravitatory potential

To solve the inconvenience to make use of the Laplace desideratum assumption, the authors have developed [5] two different methods to obtain the self-gravitational potential up to first order in amplitudes without using the referred non demonstrated hypotesis.

In this same paper the potential obtained in the classical theory has been shown to be right up to first order in ω^2.Therefore, from this it is derived that the amplitudes [4] it is only $f_2(a)$ are of first order in ω^2. From this result it easy to probe that up to second order in ω^2.

Being D the equipotential surface containing P and let $r = r_0$ the sphere containing P then equation of this sphere in Clairaut coordinates will be given up to second order in amplitudes by

$$a' = a(1 + D - D' + D'^2 - DD'),$$

 ISBN: 978-84-617-8694-7

where $D = \sum_{n=0}^{\infty} f_{2n}(a)P_{2n}(\cos\theta)$ and $D' = \sum_{n=0}^{\infty} f_{2n}(a')P_{2n}(\cos\theta')$. Evaluating the integrals (2) with a numerical quadrature method [5] it is easy to show that the obtained results up to second order (3), (4) and (5) are false.

As it has been discussed up to now, the classical theory developments of the inner and outer potentials of the auto gravitational potential up to second order do not coincide with those obtained in this work. However, the results up to the second order obtained in this work, without using Laplace's desiderata, prove that the auto gravitational potential up to the second order obtained in classical theory is correct.

4 Acknowledgments

This research has been partially supported by Grant 16I358.01/1 from the Jaume I of Castellón University.

References

[1] E. Finlay-Frendulich, *Celestial Mechanics*, Pergamon Press Inc., New York 1958.

[2] W. Jardetzky, *Theorie of figures of Celestial Bodies*, Interscience Publishers, Inc., New York, 1958.

[3] Z. Kopal, *Figures of Celestial Bodies*, Univ Wisconsin Press, Madison, 1960.

[4] Z. Kopal, *Dynamic of Close Binary Systems*, Kluwer, Dordrecht, Holland 1978.

[5] J.A. López Ortí, M. Forner Gumbau & M. Barreda Rochera, *A note on the first order theories of equilibrium figures of celestial bodies*, International Journal of Computer Mathematics Vol. 88, No. 9, 1969-1978, 2011.

[6] F. F. Tisserand, *Traité de Mécanique Celeste*, Ed Gauthier-Villars, Paris, 1896.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

An Active Attack on CLIQUES

J.A. López-Ramos[1], J. Rosenthal[2], D. Schipani[2] and R. Schnyder[2]

[1] *Department of Mathematics, University of Almeria*

[2] *Department of Mathematics, University of Zurich*

emails: jlopez@ual.es, rosenthal@math.uzh.ch, davide.schipani@math.uzh.ch, reto.schnyder@math.uzh.ch

Abstract

An active attack is presented which targets the well-known multiparty key exchange protocol CLIQUES and assumes malicious control of the communications of an arbitrary user for the duration of the key exchange only.

Key words: Group Key Management, Active Attack
MSC 2000: 95A60

1 Introduction

Group Key Management is a major concern in many applications nowadays, mainly due to the so called Internet of Things getting more and more widespread. One of the most important approaches to solve this matter is using a distributed solution, where users collaborate to build a common key. Examples of this type of protocols can be found in [2] and [6] and their references.

One of the best known protocols for distributed Group Key Management is known as CLIQUES and was proposed by Steiner et al. in [5]. In [4] the authors provided an active attack on this protocol based on the well-known man-in-the middle attack for the case of an attacker that is able to control communications of a user with a special role in the group, namely the last user during the set up phase, i.e., when the group is collaborating to agree on the first common key. They also show that it is possible that the attacker leaves the group without letting the group members know about the attack.

Our aim in this work is to show the possibility of extending the attack to any member of the group. The work is structured as follows. In the second section we describe the original protocol. In the third section we show the attack and in the last section we conclude the paper.

ISBN: 978-84-617-8694-7

2 The Group Key Agreement

The following protocol describes the Initial Key Agreement of CLIQUES named as IKA.2 in [5]. In [3], the authors generalize these schemes considering a general action on a semigroup, and this is how IKA.2 is presented below.

Suppose we have n users $\mathcal{U}_1, \ldots, \mathcal{U}_n$ who wish to agree upon a common key.

Let G be an abelian group, written multiplicatively. Let S be a set, and suppose we have a group action

$$\begin{aligned} G \times S &\to S \\ (g, s) &\mapsto g \cdot s. \end{aligned}$$

The users publicly agree on a common element $C_0 = s \in S$, and for each $i = 1, \ldots, n$, the user $\mathcal{U}_i$ selects a secret group element $g_i \in G$.

The protocol proceeds as follows:

(1) For $i = 1, \ldots, n-2$, $\mathcal{U}_i$ sends to $\mathcal{U}_{i+1}$ the message $C_i = g_i \cdot C_{i-1}$.

(2) $\mathcal{U}_{n-1}$ broadcasts $C_{n-1} = g_{n-1} \cdot C_{n-2}$ to the other users $\mathcal{U}_1, \ldots, \mathcal{U}_{n-2}, \mathcal{U}_n$.

(3) $\mathcal{U}_n$ computes the shared key $K = g_n \cdot C_{n-1}$.

(4) For $i = 1, \ldots, n-1$, $\mathcal{U}_i$ sends $D_i = g_i^{-1} \cdot C_{n-1}$ to $\mathcal{U}_n$.

(5) $\mathcal{U}_n$ broadcasts $\{g_n \cdot D_1, g_n \cdot D_2, \ldots, g_n \cdot D_{n-1}, C_{n-1}\}$ to $\mathcal{U}_i$, $i = 1, \ldots, n-1$.

(6) For $i = 1, \ldots, n-1$, $\mathcal{U}_i$ computes the shared key $K = g_i \cdot (g_n \cdot D_i)$.

3 The active attack

In this section we give an active attack on the protocol described in the preceding section. The aim of the attacker, $\mathcal{M}$, is that the users $\mathcal{U}_1, \ldots, \mathcal{U}_n$ agree on a shared key as it is obtained after running the protocol, but she will know the key as well, so she can listen and send messages using that common key as if she was a legal member of the group.

As in a usual man-in-the-middle attack, $\mathcal{M}$ needs to have full control over the communication of a user $\mathcal{U}_i$, $i \in \{1, \ldots, n\}$, but only during the key exchange. Then, unlike in a regular man-in-the-middle attack, she does not need to maintain this control after the key exchange is completed. In this paper we will assume that $i < n-1$. The case $i = \{n-1, n\}$ is given in [4].

In the beginning, $\mathcal{M}$ chooses her own secret group element $g \in G$. She then proceeds as follows:

 ISBN: 978-84-617-8694-7

(a) Step (1) is carried out as usual until $\mathcal{U}_i$ sends $C_i = g_i \cdot C_{i-1}$ to $\mathcal{U}_{i+1}$. At this point he is sitting in step (1) waiting for the broadcast of $\mathcal{U}_{n-1}$.

(b) $\mathcal{M}$ stops this message and sends $C_i' = g \cdot C_{i-1}$ to $\mathcal{U}_{i+1}$ as if she was $\mathcal{U}_i$.

(c) $\mathcal{U}_{n-1}$ broadcasts $C_{n-1} = g_{n-1} \cdot C_{n-2}$ to the other users $\mathcal{U}_1, \ldots, \mathcal{U}_{n-2}, \mathcal{U}_n$. At this point $\mathcal{U}_n$ computes the key $K = g_n \cdot C_{n-1}$.

(d) $\mathcal{M}$ stops this message for $\mathcal{U}_i$ and sends $D_i = g^{-1} C_{n-1} = \left(\prod_{j=1, j\neq i}^{n-1} g_j \right)$ to $\mathcal{U}_n$.

(e) $\mathcal{U}_n$ broadcasts $\{g_n \cdot D_1, g_n \cdot D_2, \ldots, g_n \cdot D_{n-1}, C_{n-1}\}$ to $\mathcal{U}_i$, $i = 1, \ldots, n-1$. $\mathcal{M}$ stops this message for $\mathcal{U}_i$.

(f) $\mathcal{M}$ computes $K = g \cdot (g_n \cdot D_i)$, which is the common key computed by all the remaining users.

(g) $\mathcal{M}$ chooses $b \in G$ and sends $b \cdot K$ to $\mathcal{U}_i$ as if it was the message that $\mathcal{U}_{n-1}$ broadcasts in (2).

(h) $\mathcal{U}_i$ sends $g_i^{-1} \cdot (b \cdot K)$ to $\mathcal{U}_n$.

(i) $\mathcal{M}$ stops this message, computes $g_i^{-1} \cdot K$ using b^{-1}, and sends back $\{g_n \cdot D_1, g_n \cdot D_2, \ldots, g_i^{-1} \cdot K, \ldots, g_n \cdot D_{n-1}, C_{n-1}\}$ to $\mathcal{U}_i$.

(j) $\mathcal{U}_i$ recovers $g_i \cdot (g_i^{-1} \cdot K) = K$.

By the end of the attack all users $\mathcal{U}_1, \ldots, \mathcal{U}_n$ and $\mathcal{M}$ share a common key and $\mathcal{M}$ stays in the group as another "invisible" legal member.

She can also leave the group without leaving any evidence of her stay in the group. To do so, before any rekeying takes place, $\mathcal{M}$ carries out the following simple strategy.

(i) $\mathcal{M}$ chooses g' and computes a new key $g' \cdot K = \left(g' g \prod_{j=1, j\neq i}^{n} g_j \right) \cdot s$.

(ii) $\mathcal{M}$ broadcasts $\{g' \cdot (g_n \cdot D_1), g' \cdot (g_n \cdot D_2), \ldots, g' \cdot (g_i^{-1} \cdot K), \ldots, g' \cdot (g_n \cdot D_{n-1}), g' \cdot C_{n-1}\}$ to $\mathcal{U}_1, \ldots, \mathcal{U}_n$.

Thus $\mathcal{U}_1, \ldots, \mathcal{U}_n$ and $\mathcal{M}$ will share the new key $g' \cdot K$ and she will keep being an "invisible" legal member of the group until a new rekeying is done by any user. Then they all will be rekeyed but they will never know about the attack.

We point out that $\mathcal{M}$ may take control of the group by carrying out the leaving strategy for as long as she wishes before any other member rekeys.

 ISBN: 978-84-617-8694-7

4 Conclusions

We have shown that IKA.2 implementation of CLIQUES is insecure unless the Initial Key Agreement takes place under an authenticated environment. In particular, we have presented an active attack that allows the attacker to share a common key with all the legal members in a communication group and a strategy to take control of the communications or leave the group without letting the legal members know about the attack.

In [1] the authors propose an authenticated version of IKA.2 for CLIQUES. An analogous attack can be also carried out for this, but a desynchronizing of users may take place and the attack will be discovered. However the attacker may also be part of the group without letting the legal members know about her presence during the attack.

Acknowledgements

The Research was supported in part by the Swiss National Science Foundation under grant No. 169510. First author is partially suppported by Ministerio de Economia y Competitividad grant MTM2014-54439 and Junta de Andalucia (FQM0211). The last author is supported by Armasuisse.

References

[1] G. Ateniese, M. Steiner, G. Tsudik, *New Multiparty Authentication Services and Key Agreement Protocols*, IEEE J. Sel. Areas Commun.**18(4)** (2000) 1–13.

[2] P. P. C. Lee, J. C. S. Lui, D. K. Y. Yau, *Distributed Collaborative Key Agreement and Authentication Protocols for Dynamic Peer Groups*, IEEE/ACM Trans. Networking **14(2)** (2006) 263–276.

[3] J. A. López-Ramos, J. Rosenthal, D. Schipani, R. Schnyder, *Group key management based on semigroup actions*, J. Algebra Appl. doi:10.1142/S0219498817501481 (2016).

[4] R. Schnyder, J. A. Lopez-Ramos, J. Rosenthal, D. Schipani, *An active attack on a multiparty key exchange protocol*, J. Algebra Comb. Discrete Appl. **3(1)** (2016) 31–36.

[5] M. Steiner, G. Tsudik, M. Waidner, *Key agreement in dynamic peer groups*, IEEE Trans. Parallel Distrib. Syst., **11(8)** (2000) 769–780.

[6] J. Van der Merwe, D. Dawoud, S. McDonald, *A survey on peer-to-peer key management for mobile ad hoc networks*, ACM Computing Surveys **39 (1)** 2007.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

Existence of unbounded solutions of IVPs with ϕ-Laplacian

Lucía López-Somoza[1]

[1] *Institute of Mathematics, University of Santiago de Compostela*

(Joint work with J. Burkotová, I. Rachunková, M. Rohleder, J. Stryja)

emails: lucia.lopez.somoza@usc.es

Abstract

We will study the existence of unbounded solutions of the problem

$$(p(t)\phi(u'(t)))' + p(t)f(\phi(u(t))) = 0, \quad u(0) = u_0 \in [L_0, L], \quad u'(0) = 0.$$

Key words: second order ODE, ϕ-Laplacian, unbounded solution, time singularity.

1 Introduction

Our aim is to analyze the singular nonlinear equation

$$(p(t)\phi(u'(t)))' + p(t)f(\phi(u(t))) = 0, \tag{1}$$

coupled with the initial conditions

$$u(0) = u_0, \quad u'(0) = 0, \quad u_0 \in [L_0, L]. \tag{2}$$

Problem (1)-(2) is investigated under the basic assumptions

$$\phi \in C^1(\mathbb{R}), \quad \phi'(x) > 0 \text{ for } x \in (\mathbb{R} \setminus \{0\}), \quad \phi(\mathbb{R}) = \mathbb{R}, \quad \phi(0) = 0, \tag{3}$$

$$L_0 < 0 < L, \quad f(\phi(L_0)) = f(0) = f(\phi(L)) = 0, \tag{4}$$

$$f \in C[\phi(L_0), \infty), \quad xf(x) > 0 \text{ for } x \in ((\phi(L_0), \phi(L)) \setminus \{0\}), \quad f(x) \leq 0 \text{ for } x > \phi(L), \tag{5}$$

$$p \in C[0, \infty) \cap C^1(0, \infty), \quad p'(t) > 0 \text{ for } t \in (0, \infty), \quad p(0) = 0. \tag{6}$$

Our problem can be singular in the sense that $p(0) = 0$ and $1/p(t)$ may not be integrable.

Definition 1.1. Let $[0, b) \subset [0, \infty)$ be a maximal interval such that a function $u \in C^1[0, b)$ with $\phi(u') \in C^1(0, b)$ satisfies equation (1) for every $t \in (0, b)$ and let u satisfy the initial conditions (2). Then u is called a *solution* of problem (1)-(2) on $[0, b)$.

 ISBN: 978-84-617-8694-7

Definition 1.2. Consider a solution of problem (1)-(2) on $[0,\infty)$ with $u_0 \in [L_0, L]$ and denote $u_{sup} = \sup\{u(t)\colon t \in [0,\infty)\}$.
If $u_{sup} = L$, then u is called a *homoclinic solution* of problem (1)-(2).
If $u_{sup} < L$, then u is called a *damped solution* of problem (1)-(2).

Definition 1.3. Let u be a solution of (1)-(2) on $[0,b)$. If there exists $c \in (0,b)$ such that $u(c) = L$, $u'(c) > 0$, then u is called an *escape solution* of problem (1)-(2) on $[0,b)$.

Analytical properties of solutions of problem (1)-(2) with a ϕ-Laplacian have already been studied in [1] with a focus on the existence of bounded solutions. The goal of this work (which can be found in [2]) is to find conditions which guarantee the existence of unbounded solutions of (1)-(2). Since in general an escape solution does not need to be unbounded, criteria for an escape solution to tend to infinity are necessary. We distinguish two cases:

- Case I: If functions ϕ^{-1} and f are Lipschitz continuous, the uniqueness of solution of problem (1)-(2) is guaranteed.
- Case II: If ϕ^{-1} and f are not Lipschitz continuous, the lack of uniqueness causes difficulties. Problems are overcome by means of the lower and upper function method.

2 Auxiliary Problem

To simplify our considerations we introduce the auxiliary equation

$$(p(t)\,\phi(u'(t)))' + p(t)\,\tilde{f}(\phi(u(t))) = 0, \quad t \in (0,\infty), \tag{7}$$

where

$$\tilde{f}(x) = \begin{cases} f(x) & \text{for } x \in [\phi(L_0), \phi(L)], \\ 0 & \text{for } x < \phi(L_0), \quad x > \phi(L). \end{cases}$$

Two important hypothesis that we will assume are

$$\lim_{t\to\infty} \frac{p'(t)}{p(t)} = 0, \tag{8}$$

$$\exists \bar{B} \in (L_0, 0)\colon \tilde{F}\left(\bar{B}\right) = \tilde{F}(L), \quad \text{where } \tilde{F}(x) = \int_0^x \tilde{f}(\phi(s))\,\mathrm{d}s, \quad x \in \mathbb{R}. \tag{9}$$

Under the previous hypothesis, the two following results are proved in [1].

Theorem 2.1. *If* (3)–(6) *hold, then for each* $u_0 \in [L_0, L]$, *there exists a solution of* (7)-(2). *Moreover, if*

$$f \in \mathrm{Lip}[\phi(L_0), \phi(L)], \quad \text{and} \quad \phi^{-1} \in \mathrm{Lip}_{\mathrm{loc}}(\mathbb{R}), \tag{10}$$

then any solution of (7)-(2) *with* $u_0 \in [L_0, L]$ *is unique on* $[0,\infty)$.

Theorem 2.2. *Assume* (3)–(6), (8) *and* (9). *Then, for each* $u_0 \in \left[\bar{B}, L\right)$, *problem* (1)-(2) *has a solution. If* $u_0 \in \left[\bar{B}, L\right)$, *every solution is damped.*

 ISBN: 978-84-617-8694-7

Now we will formulate the results of existence of escape solutions for the auxiliary problem. We note that, despite the results are quite similar, their proofs are quite different.

Theorem 2.3 (Existence of escape solutions of (7)-(2), Case I)**.** *If* (3)–(6), (8)–(10) *hold, then there exist infinitely many escape solutions of* (7)-(2) *with starting values in* $(L_0, \bar{B})$.

Theorem 2.4 (Existence of escape solutions of (7)-(2), Case II)**.** *If* (3)–(6), (8)–(9) *hold, then there exist infinitely many escape solutions of* (7)-(2) *with starting values in* $\left[L_0, \bar{B}\right)$.

3 Unbounded solutions

Now we will provide conditions for p and f to ensure the existence of unbounded solutions of (1)-(2). Assume that (3)–(6), (8) and (9) hold and let u be a solution of (7)-(2). Then

$$\exists\, c \in (0,\infty)\colon u(t) \in [L_0, L),\ t \in [0, c), \quad u(c) = L, \quad u'(c) > 0. \tag{11}$$

Clearly, u fulfils (1) on $[0, c]$ and can be extended as a solution of the original problem (1)-(2) on some maximal interval $[0, b)$, where $c < b \leq \infty$. So, unbounded solutions of the original problem (1)-(2) can be searched just as prolongations of escape solutions of (7)-(2).

Lemma 3.1. *Assume that* (3)–(6) *hold. Let u be an escape solution of* (1)-(2) *on* $[0, b)$. *Then* $u(t) > L$, $u'(t) > 0$, $t \in (c, b)$, *where c is from* (11). *If* $b < \infty$, *then* $\lim_{t\to b^-} u(t) = \infty$.

So, it is enough to investigate the boundedness of escape solutions having $[0, b) = [0, \infty)$.

Theorem 3.2. *Assume* (3)–(6) *hold and let*

$$\lim_{t\to\infty} p(t) < \infty. \tag{12}$$

Let u be an escape solution of problem (1)-(2). *Then* $\lim_{t\to\infty} u(t) = \infty$.

Theorem 3.3. *Assume* (3)–(6), (8) *and*

$$f(x) < 0 \quad \text{for } x > \phi(L). \tag{13}$$

Let u be an escape solution of problem (1)-(2). *Then* $\lim_{t\to\infty} u(t) = \infty$.

Theorem 3.4. *Assume* (3)–(6),

$$f(x) \equiv 0 \quad \text{for } x > \phi(L), \tag{14}$$

$$\phi(ab) = \phi(a)\,\phi(b), \quad a,\, b \in (0,\infty). \tag{15}$$

Let u be an escape solution of problem (1)-(2). *Then*

$$\lim_{t\to\infty} u(t) = \infty \quad \Longleftrightarrow \quad \int_1^\infty \phi^{-1}\left(\frac{1}{p(s)}\right) \mathrm{d}s = \infty.$$

If we replace condition (15) *by*

$$\phi(ab) \leq \phi(a)\,\phi(b), \quad a,\, b \in (0,\infty), \tag{16}$$

then $\lim_{t\to\infty} u(t) = \infty$ *holds if*

$$\int_1^\infty \phi^{-1}\left(\frac{1}{p(s)}\right) \mathrm{d}s = \infty. \tag{17}$$

 ISBN: 978-84-617-8694-7

4 Conclusions and example

If we combine the theorems about existence of escape solutions of the auxiliary problem (7)-(2) with one of the theorems from Section 3, we get the main existence results about unbounded solutions of (1)-(2). We will write them for Case II, being analogous for Case I.

Theorem 4.1 (Existence of unbounded solutions of problem (1)-(2), Case II, 1)**.** *Let* (3)–(6), (8), (9) *and* (12) *hold. Then there exist infinitely many unbounded solutions* u_n *of problem* (1)-(2) *on* $[0, b_n)$ *with starting values in* $\big[L_0, \bar{B}\big)$, $n \in \mathbb{N}$.

Theorem 4.2 (Existence of unbounded solutions of problem (1)-(2), Case II, 2)**.** *Assume that* (3)–(6), (8), (9) *and* (13) *hold. Then there exist infinitely many unbounded solutions* u_n *of problem* (1)-(2) *on* $[0, b_n)$ *with starting values in* $\big[L_0, \bar{B}\big)$, $n \in \mathbb{N}$.

Theorem 4.3 (Existence of unbounded solutions of problem (1)-(2), Case II, 3)**.** *Let* (3)–(6), (8), (9) *and* (14)–(17) *hold. Then there exist infinitely many unbounded solutions* u_n *of problem* (1)-(2) *on* $[0, b_n)$ *with starting values in* $\big[L_0, \bar{B}\big)$, $n \in \mathbb{N}$.

Example 4.4. Consider problem (1)-(2) with $\phi(x) = |x|^\alpha \operatorname{sgn} x$, $\alpha > 1$, $p(t) = \arctan t$ and

$$f(x) = \begin{cases} \sqrt{|x|}\, \operatorname{sgn} x(x - \phi(L_0))(\phi(L) - x) & \text{for } x \in [\phi(L_0), \phi(L)], \\ \cos(x - \phi(L)) - 1 & \text{for } x > \phi(L), \end{cases} \qquad 0 < L < -L_0.$$

Previous functions satisfy (3)–(6) and (8). As f is continuous, $0 < L < -L_0$ and ϕ is a continuous and odd function, (9) holds. From Theorem 4.1, problem (1)-(2) has infinitely many unbounded solutions with starting values in $\big[L_0, \bar{B}\big)$.

Acknowledgements

Supported by FPU scholarship, Ministerio de Educación, Cultura y Deporte, Spain, and Fundación Barrié Scholarship. Partially supported by Ministerio de Economía y Competitividad, Spain and FEDER, project MTM2013-43014-P. Partially supported by the Agencia Estatal de Investigación (AIE) Spain, under grant MTM2016-75140-P, co-financed by the European Community fund FEDER.

References

[1] J. Burkotová, I. Rachůnková M. Rohleder, J. Stryja, Existence and uniqueness of damped solutions of singular IVPs with ϕ-Laplacian, *Electron. J. Qual. Theory Differ. Equ.* **121** (2016), 1–28.

[2] J. Burkotová, L. López-Somoza, M. Rohleder, J. Stryja, On unbounded solutions os singular IVPs with ϕ-Laplacian. (Submitted)

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

An energy method for nonlinear Riesz space-fractional wave equations

J. E. Macías-Díaz[1]

[1] *Departamento de Matemáticas y Física, Centro de Ciencias Básicas, Universidad Autónoma de Aguascalientes, Mexico*

emails: jemacias@correo.uaa.mx

Abstract

In this work, we will investigate a Riesz space-fractional wave equation with a nonlinear potential term, and approximate its solutions following an implicit finite-difference approach. Classical wave equations are susceptible of being investigated through Lagrangian and Hamiltonian formulations; however, the lack of a physically meaningful theory has been a major drawback in the investigation of fractional systems. Recently, some energy-like operators have been introduced for fractional models. Using those results, we propose a discretization for nonlinear fractional wave equations that conserves the total energy of the system. Amongst other properties, we establish that the method is a convergent and stable technique. Some comparisons against results available in the literature are provided in the way.

Key words: wave equation, Riesz space-fractional equation, Lagrangian formulation, Hamiltonian formulation, energy-preserving method, implicit finite-difference scheme

1 Introduction

Nonlinear supratransmission is a phenomenon that was thoroughly investigated in arrays of harmonic oscillators and in many other hyperbolic systems [2]. Historically, the study of energy transmission in nonlinear wave equations has been an interesting topic of investigation. These models have applications in the description of data transmission in optical fibers and in the study of the self-induced transparency of systems subject to a high-energy incident laser pulse. More generally, the behavior of continuous media subject to a continuous wave radiation is a fundamental problem that has potential applications in many nonlinear systems [6].

 ISBN: 978-84-617-8694-7

Additionally, there are many reports which investigate mathematically the occurrence of supratransmission. However, still many questions remain unanswered while other avenues of research open up with the development of new mathematical tools. For instance, models that consider fractional derivatives have attracted the attention of many researchers in recent years. Classical systems that include derivatives of integral order have been extended in this way using various inequivalent approaches [5], and interesting results have been derived in the way. Motivated by these facts, we will provide a continuous Riesz fractional extension of the model investigated in [1] for derivative orders in $(1, 2)$, and design an energy-preserving and convergent method to approximate the solutions of Riesz space-fractional wave equations with nonlinear potentials.

2 Preliminaries

The point of departure and the motivation of this work are the $(1 + 1)$-dimensional sine-Gordon and nonlinear Klein-Gordon equations from relativistic quantum mechanics and field theory, which are described by the hyperbolic partial differential equation

$$\frac{\partial^2 u}{\partial t^2}(x,t) - \frac{\partial^2 u}{\partial x^2}(x,t) + V'(u(x,t)) = 0, \tag{1}$$

where

$$V(u) = 1 - \cos u \tag{2}$$

in the case of the classical sine-Gordon system. In the case of the nonlinear Klein-Gordon equation, the potential function V takes on the form

$$V(u) = \frac{1}{2!}u^2 - \frac{1}{4!}u^4 + \frac{1}{6!}u^6. \tag{3}$$

Both equations possess a Lagrangian functional and an associated Hamiltonian which are defined for each $(x,t) \in I \times \mathbb{R}^+$, and that are given respectively by

$$\mathcal{L}(x,t) = \frac{1}{2}\left[\frac{\partial u}{\partial t}(x,t)\right]^2 - \frac{1}{2}\left[\frac{\partial u}{\partial x}(x,t)\right]^2 - V(u(x,t)), \tag{4}$$

$$\mathcal{H}(x,t) = \frac{1}{2}\left[\frac{\partial u}{\partial t}(x,t)\right]^2 + \frac{1}{2}\left[\frac{\partial u}{\partial x}(x,t)\right]^2 + V(u(x,t)). \tag{5}$$

The total energy of the system (1) at the time t is provided by

$$\mathcal{E}(t) = \int_I \mathcal{H}(x,t)dx. \tag{6}$$

 ISBN: 978-84-617-8694-7

3 Mathematical problem

In this work, we will consider a space-fractional extension of the sine-Gordon equation (1) using the Riesz fractional approach [4]. More precisely, let us define the Riesz space-fractional derivative of u of order $0 \leq \alpha \leq 2$ at the point (x,t) by

$$\frac{\partial^\alpha u}{\partial |x|^\alpha}(x,t) = \begin{cases} -\dfrac{\mathcal{D}_+^\alpha + \mathcal{D}_-^\alpha}{2\cos(\alpha\pi/2)} u(x,t), & \alpha \neq 1, \\ \left(\dfrac{d}{dx} H\right) u(x,t), & \alpha = 1. \end{cases} \tag{7}$$

Here H is the Hilbert transform in the first variable, that is,

$$Hu(x,t) = \text{p.v.} \frac{1}{\pi} \int_{-\infty}^{\infty} \frac{u(\xi,t)}{\xi - x} d\xi, \tag{8}$$

where the principal value of the integral is understood in the sense of Cauchy.

For the sake of briefness, the Riesz space-fractional derivative of order α of $u(x,t)$ is sometimes denoted by $\mathcal{D}^\alpha u(x,t)$. In (7), $\mathcal{D}_\pm^\alpha$ denotes a Weyl fractional derivative operator of order α, which is given in terms of the Weyl fractional integrals by

$$\mathcal{D}_\pm^\alpha u(x,t) = \begin{cases} \pm \left(\dfrac{d}{dx} I_\pm^{1-\alpha}\right) u(x,t), & 0 < \alpha < 1, \\ \left(\dfrac{d^2}{dx^2} I_\pm^{2-\alpha}\right) u(x,t), & 1 < \alpha < 2. \end{cases} \tag{9}$$

For each $\beta > 0$,

$$I_+^\beta u(x,t) = \frac{1}{\Gamma(\beta)} \int_{-\infty}^{x} (x-\xi)^{\beta-1} u(\xi,t) d\xi, \tag{10}$$

$$I_-^\beta u(x,t) = \frac{1}{\Gamma(\beta)} \int_{x}^{\infty} (\xi-x)^{\beta-1} u(\xi,t) d\xi. \tag{11}$$

Let A and Ω be positive numbers, and let $I = (0,L)$ for some $L > 0$. Using the conventions introduced above, this manuscript investigates the following initial-boundary-value problem:

$$\frac{\partial^2 u}{\partial t^2}(x,t) - \gamma(x)\frac{\partial u}{\partial t}(x,t) - \frac{\partial^\alpha u}{\partial |x|^\alpha}(x,t) + V'(u(x,t)) = 0, \quad (x,t) \in I \times \mathbb{R}^+,$$
$$\text{subject to } \begin{cases} u(x,0) = u_t(x,0) = 0, & \forall x \in I, \\ u(0,t) = A\sin(\Omega t), & \forall t \in \mathbb{R}^+, \\ u_x(L,t) = 0, & \forall t \in \mathbb{R}^+. \end{cases} \tag{12}$$

Here, the function $\gamma : I \to \mathbb{R}$ represents damping and it will be used to account for an absorbing boundary at the right end of I.

 ISBN: 978-84-617-8694-7

The problem considered in this work is a non-local version of the system studied in different media [2, 3] in the context of the investigation of nonlinear supratransmission. It is important to point out that the undamped form of the fractional model (12) has the generalized fractional Hamiltonian functional

$$\mathcal{H}(x,t) = \frac{1}{2}\left[\frac{\partial u}{\partial t}(x,t)\right]^2 + \frac{1}{2}\left[\frac{\partial u}{\partial x}(x,t)\right]\left[\frac{\partial^{\alpha-1}u}{\partial|x|^{\alpha-1}}(x,t)\right] + V(u(x,t)), \tag{13}$$

for each $(x,t) \in I \times \mathbb{R}^+$ (see [5] for details). With these conventions, an energy integral associated to the undamped sine-Gordon equation of the problem (12) is defined by (6). Moreover, if we consider a finite period of time T, the total energy of the system in that periods will be defined as

$$\mathcal{E} = \int_0^T \mathcal{E}(t)dt. \tag{14}$$

4 Aim of this work

The purpose of this work is to provide a consistent finite-difference discretization of (12), together with consistent discrete forms of (13) and (14). These and more properties of our numerical techniques will be established thoroughly, and suitable applications to the investigation of nonlinear supratransmission will be proposed in the way.

References

[1] D Chevriaux, R Khomeriki, and J Leon. Theory of a Josephson junction parallel array detector sensitive to very weak signals. *Physical Review B*, 73(21):214516, 2006.

[2] F Geniet and J Leon. Energy transmission in the forbidden band gap of a nonlinear chain. *Physical review letters*, 89(13):134102, 2002.

[3] F Geniet and J Leon. Nonlinear supratransmission. *Journal of Physics: Condensed Matter*, 15(17):2933, 2003.

[4] Stefan G Samko, Anatoly A Kilbas, Oleg I Marichev, et al. Fractional integrals and derivatives. *Theory and Applications, Gordon and Breach, Yverdon*, 1993, 1993.

[5] Vasily E Tarasov and George M Zaslavsky. Conservation laws and hamiltons equations for systems with long-range interaction and memory. *Communications in Nonlinear Science and Numerical Simulation*, 13(9):1860–1878, 2008.

[6] AV Ustinov. Solitons in Josephson junctions. *Physica D: Nonlinear Phenomena*, 123(1-4):315–329, 1998.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

A structure-preserving computational method in the simulation of the dynamics of cancer growth with radiotherapy

J. E. Macías-Díaz[1] and Armando Gallegos[2]

[1] *Departamento de Matemáticas y Física, Centro de Ciencias Básicas, Universidad Autónoma de Aguascalientes, Mexico*

[2] *Departamento de Ciencias Exactas y Tecnología, Centro Universitario de los Lagos, Universidad de Guadalajara, Mexico*

emails: jemacias@correo.uaa.mx, gallegos@culagos.udg.mx

Abstract

In this work, we provide a discretization of a nonlinear diffusion-reaction system that models the growth of cancer with radiotherapy. Only positive and bounded solutions are physically relevant in this context, and the discretization that we provide in this manuscript is able to preserve both properties. The method is computationally economic; some qualitative and quantitative comparisons are carried out in support of the advantages of our scheme. Moreover, the technique used in the present manuscript has the advantage over other similar methodologies that it yields no singularities. In addition, the preservation of the properties of non-negativity and boundedness of both the solution and the total mass are distinctive features which are established analytically in this work. The numerical simulations on cancer growth obtained with the exponential method are found to be in good agreement with the experimental results available in the literature.

Key words: cancer growth model, two-dimensional diffusion-reaction equation, preservation of structure, exponential finite-difference method, fast computational method

1 Introduction

The present work is motivated by various systems of partial differential equations that describe the growth of cancer. More precisely, the present work is motivated by partial

 ISBN: 978-84-617-8694-7

differential equations in the dynamics of brain cancer [1] and the effects of proliferation and motility of transforming growth factor (TGF) β on cancer cells [5]. In all these cases, the models under consideration are extensions of the classical Fisher's equation of population dynamics [2], and all of them are physically interesting in the two-dimensional scenario in view of the applications to the dynamics of growth of tumors on human tissues. Moreover, the variable of interest in either case is the density of tumor cells at each point of the tissue, whence the properties of non-negativity and boundedness naturally arise. In view of these remarks, one is immediately led to ask whether it is possible to design an efficient and dynamically consistent technique to approximate the solutions of a generalized two-dimensional form of the equations investigated in [5]. More precisely, we are interested in developing numerical methods to approximate the solutions of those equations with the following characteristics:

1. The non-negativity and the boundedness of numerical approximations is preserved.
2. The method preserves the non-negativity and the boundedness of the total mass of the cancer tumor.
3. The technique is computationally fast.
4. The method is easy to implement in any computer language.
5. The computational implementation allows to employ fine grid meshes.

2 Mathematical model

Throughout this manuscript, we assume that α, β and D_p are positive numbers and that D_m is nonnegative. Let Ω be a domain of $\mathbb{R}^2$ which is open, bounded, connected and measurable, with area given by $\mathcal{A}(\Omega)$. We let $u = u(x, y, t)$ be a real function defined on the closure of $\Omega \times \mathbb{R}^+$, which is twice differentiable in the interior of its domain and which satisfies the initial-value problem with homogeneous Neumann boundary conditions

$$\begin{gathered}\frac{\partial u}{\partial t}(x,y,t) = \nabla \cdot (D(\mu(t))\nabla u(x,y,t)) + \alpha u(x,y,t)\left(1-\beta u(x,y,t)\right), \quad \forall (x,y) \in \Omega, \forall t \in \mathbb{R}^+, \\ \begin{cases} u(x,y,0) = \phi(x,y), & \forall (x,y) \in \overline{\Omega}, \\ \dfrac{\partial u}{\partial \mathbf{n}}(x,y,t) = 0, & \forall (x,y) \in \partial\Omega, \forall t \in \mathbb{R}^+ \cup \{0\}, \end{cases}\end{gathered} \tag{1}$$

for some continuous function $\phi : \overline{\Omega} \to \mathbb{R}$ which satisfies the condition $0 \leq \phi(x, y) \leq \frac{1}{\beta}$ at each $(x, y) \in \overline{\Omega}$. Here, D is the real function defined by

$$D(z) = D_m \left(1 - \frac{z}{\mathcal{A}(\Omega)}\right) + D_p \tag{2}$$

 ISBN: 978-84-617-8694-7

for each $z \in \mathbb{R}$, and the function μ will be the function of *total mass* of the system at the time t. More precisely, for each $t \geq 0$ we let

$$\mu(t) = \mu(u(x, y, t)) = \beta \iint_{\Omega} u(x, y, t) dx dy. \tag{3}$$

3 Physical description

Following the physical context of [5], the value $u(x, y, t)$ represents the spatial density of tumor cells at the time t and at the point (x, y) on a flat surface of interior equal to Ω. Clearly, u has units $length^{-2}$. The parameter α denotes the intrinsic proliferation rate of the cells, with units given in $time^{-1}$. Meanwhile, β is given in units of $length^2$ and represents the (positive) area occupied by a single cell in average, whence the constant $1/\beta$ represents the number of cells per unit of area. Here, we impose the constraint

$$0 \leq u(x, y, t) \leq \frac{1}{\beta}, \tag{4}$$

for each $(x, y) \in \Omega$ and each $t \in \mathbb{R}^+$. Moreover, the expression

$$N(t) = \frac{\mu(t)}{\beta} \tag{5}$$

denotes the total number of cells in the region Ω.

Obviously, the maximum value of N is achieved when u is the constant $1/\beta$, in which case the number of cells is given by the value $N_{\max} = \mathcal{A}(\Omega)/\beta$. The role of the constant D_p is to account for the spatial expansion of proliferating cells, while the parameter D_m accounts for the facts that the cells undergo a random walk which follows the laws of Brownian motion, and that the cells of a cluster may break lose from the cluster [4]. Finally, it is worthwhile to notice that (1) is also a model of the growth of brain cancer when the diffusion coefficient is constant and when radiotherapy is not considered [1, 3].

The following property is recorded as a lemma for the sake of future reference.

Lemma 1 *The inequalities*

$$0 \leq \mu(t) \leq \mathcal{A}(\Omega), \tag{6}$$

$$D_p \leq D(\mu(t)) \leq D_m + D_p, \tag{7}$$

are satisfied when (4) *holds.*

 ISBN: 978-84-617-8694-7

4 Aim of this work

The purpose of this work is to provide a structure-preserving finite-difference discretization of the following initial-boundary-value problem:

$$\begin{aligned}\frac{\partial u}{\partial t}(x,y,t) &= \nabla\cdot(D(\mu(t))\nabla u(x,y,t)) + \alpha u(x,y,t)\,(1-\beta u(x,y,t)) + f(u,x,y,t),\\ &\begin{cases} u(x,y,0) = \phi(x,y), & \forall (x,y)\in\overline{\Omega},\\ \dfrac{\partial u}{\partial \mathbf{n}}(x,y,t) = 0, & \forall (x,y)\in\partial\Omega, \forall t\in\mathbb{R}^+\cup\{0\},\end{cases}\end{aligned} \tag{8}$$

This problem describes, among other physical situations, the dynamics of growth of cancer subject to a therapy function f, whence its importance is pragmatically justified.

References

[1] Juan Belmonte-Beitia, Gabriel F Calvo, and Víctor M Pérez-García. Effective particle methods for Fisher–Kolmogorov equations: Theory and applications to brain tumor dynamics. *Communications in Nonlinear Science and Numerical Simulation*, 19(9):3267–3283, 2014.

[2] Ronald Aylmer Fisher. The wave of advance of advantageous genes. *Annals of Eugenics*, 7(4):355–369, 1937.

[3] S Nawrocki and B Zubik-Kowal. Clinical study and numerical simulation of brain cancer dynamics under radiotherapy. *Communications in Nonlinear Science and Numerical Simulation*, 22(1):564–573, 2015.

[4] Ali Nawshad, Damian LaGamba, Ahmad Polad, and Elizabeth D Hay. Transforming growth factor-β signaling during epithelial-mesenchymal transformation: implications for embryogenesis and tumor metastasis. *Cells Tissues Organs*, 179(1-2):11–23, 2005.

[5] Shizhen Emily Wang, Peter Hinow, Nicole Bryce, Alissa M Weaver, Lourdes Estrada, Carlos L Arteaga, and Glenn F Webb. A mathematical model quantifies proliferation and motility effects of TGF-β on cancer cells. *Computational and mathematical methods in medicine*, 10(1):71–83, 2009.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

Traveling-wave solutions of a generalized damped wave equation with time-dependent coefficients through the trial equation method

J. E. Macías-Díaz[1] and Héctor Vargas-Rodríguez[2]

[1] *Departamento de Matemáticas y Física, Centro de Ciencias Básicas, Universidad Autónoma de Aguascalientes, Mexico*

[2] *Departamento de Ciencias Exactas y Tecnología, Centro Universitario de los Lagos, Universidad de Guadalajara, Mexico*

emails: jemacias@correo.uaa.mx, hvargas@culagos.udg.mx

Abstract

In this note, we investigate the existence of exact solutions of a nonlinear partial differential equation with time-dependent coefficients that generalizes the well-known nonlinear wave model with external and internal damping. The model under consideration generalizes other classical models from physics, like the nonlinear Klein–Gordon equation, the $(1+1)$-dimensional ϕ^4-theory, the Fisher–Kolmogorov–Petrovsky–Piscounov equation from population dynamics and the Hodgkin–Huxley model used in the description of the propagation of electric signals through the nervous system. An extension of the trial equation method (also known as the direct integral method) for partial differential equations with non-constant coefficients is used in this work in order to derive traveling-wave solutions in exact form.

Key words: generalized wave equation, time-dependent coefficients, traveling-wave solutions, trial equation method, external and internal damping, direct integral method

1 Introduction

In this note, we use $\mathbb{R}^+$ to represent the set of positive numbers, and we let a, b, c, d, e, f and g be real-valued functions defined on $\mathbb{R}^+\cup\{0\}$. Throughout we suppose that u is a real function defined on $\mathbb{R}\times\mathbb{R}^+\cup\{0\}$ which has derivatives up to the third order in the interior

 ISBN: 978-84-617-8694-7

of its domain, and which satisfies the nonlinear wave equation with external damping and time-dependent coefficients

$$\frac{\partial^2 u}{\partial t^2}(x,t) - a^2(t)\frac{\partial^2 u}{\partial x^2}(x,t) = b(t)\frac{\partial^3 u}{\partial x^2 \partial t}(x,t) + c(t)\frac{\partial u}{\partial t}(x,t) + F(u(x,t),t), \tag{1}$$

for all $(x,t) \in \mathbb{R} \times \mathbb{R}^+$. Here the reaction function is given by

$$F(u,t) = d(t) + e(t)u + f(t)u^2 + g(t)u^3, \quad \forall (x,t) \in \mathbb{R} \times \mathbb{R}^+. \tag{2}$$

It is worth noting that this model is a generalization of various models appearing mathematical physics. Amongst those equations, we may quote the following:

- The classical wave equation results from (1) when the coefficient a is a constant, and when b, c and F are identically equal to zero. So (1) is a generalization of the wave equation with power-law nonlinearities.

- The classical linear Klein–Gordon equation from relativistic quantum mechanics is obtained when a is a constant, e is a positive real number, and all the other coefficients are identically equal to zero. The nonlinear Klein–Gordon equation with constant coefficients is also a particular case of Equation (1).

- Our model is also a generalization of the classical Fisher's equation investigated simultaneously and independently by R. A. Fisher and A. N. Kolmogorov, I. G. Petrovsky and N. S. Piscounov in the context of the dynamics of some populations.

- Equation (1) is an extension of the Hogdkin–Huxley model describing the propagation of electric signals through nerves.

- Finally, our model is also a continuous and damped generalization of both the α- and β-Fermi–Pasta–Ulam systems.

The presence of time-dependent coefficients in (1) accounts for the possible effects of an inhomogeneous medium. From that perspective, the model under investigation in this work may provide a more realistic description of phenomena described by constant-coefficient wave equations in the homogeneous regime. This is particularly interesting in view of the fact that wave-like equation with constant coefficients have been proposed to describe a wide range of physical phenomena, including the process of nonlinear supratransmission. Supratransmission which is a nonlinear phenomenon which was discovered in mechanical chains of oscillators described by coupled Klein-Gordon equations [2], and it was quickly studied in other nonlinear models [2, 3].

 ISBN: 978-84-617-8694-7

2 Methodology

In this work, we will derive traveling-wave solutions of the generalized damped wave equation (1) using an extension of the well-known trial equation method for partial differential equations with time-dependent coefficients. To that end, we start considering a solution of (1) of the form

$$u(x,t) = u(\xi(x,t)), \qquad \xi(x,t) = \kappa(t)x + \omega(t), \tag{3}$$

where both κ and ω are real functions defined on $\mathbb{R}^+ \cup \{0\}$. After differentiating and using these assumptions, we obtain the equations

$$\begin{aligned}
\frac{\partial u}{\partial t}(x,t) &= \left(\kappa'(t)x + \omega'(t)\right) u'(\xi), &(4)\\
\frac{\partial^2 u}{\partial t^2}(x,t) &= \left(\kappa''(t)x + \omega''(t)\right) u'(\xi) + \left(\kappa'(t)x + \omega'(t)\right)^2 u''(\xi), &(5)\\
\frac{\partial^2 u}{\partial x^2}(x,t) &= \kappa^2(t) u''(\xi), &(6)\\
\frac{\partial^3 u}{\partial x^2 \partial t}(x,t) &= 2\kappa(t)\kappa'(t)u''(\xi) + \kappa^2(t)\left(\kappa'(t)x + \omega'(t)\right) u'''(\xi). &(7)
\end{aligned}$$

In order to simplify the notation, for the remainder of this work we will convey that $u = u(\xi)$. Substituting the derivatives (4)–(7) into (1), we readily obtain the following nonlinear ordinary differential equation:

$$\Delta(x,t)u' + \Theta(x,t)u'' - \Lambda(x,t)u''' - F(u,t) = 0. \tag{8}$$

where we have employed the notation

$$\begin{aligned}
\Delta(x,t) &= \left[\kappa''(t) - c(t)\kappa'(t)\right] x + \left[\omega''(t) - c(t)\kappa'(t)\right], &(9)\\
\Theta(x,t) &= \left[\kappa'(t)x + \omega'(t)\right]^2 - a^2(t)\kappa^2(t) - 2\kappa(t)\kappa'(t)b(t), &(10)\\
\Lambda(x,t) &= \kappa^2(t)\left[\kappa'(t)x + \omega'(t)\right] b(t). &(11)
\end{aligned}$$

3 Aim of this work

In work, we will employ the trial equation method to obtain solutions of the ordinary differential equation (8). We obtain some conditions to guarantee that the parameters of the trial equation are constant real numbers. The results of our calculations will be summarized in a proposition at the end of our work. We will also derive some traveling-wave solutions of the damped wave equation (1) in exact form. To that end, we will make use of some results reported in [1]. Various cases will be considered in terms of a complete discriminant system for the polynomial resulting from the trial equation.

 ISBN: 978-84-617-8694-7

References

[1] Liu Cheng-Shi. Exact travelling wave solutions for (1+ 1)-dimensional dispersive long wave equation. *Chinese Physics*, 14(9):1710, 2005.

[2] F Geniet and J Leon. Energy transmission in the forbidden band gap of a nonlinear chain. *Physical Review Letters*, 89(13):134102, 2002.

[3] Ramaz Khomeriki. Nonlinear band gap transmission in optical waveguide arrays. *Physical Review Letters*, 92(6):063905, 2004.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference on Computational and Mathematical Methods in Science and Engineering, CMMSE 2017 4–8 July, 2017.

A numerical method to simulate the dynamics of nonlinear hysteresis in a fractional β-Fermi–Pasta–Ulam lattice

J. E. Macías-Díaz[1] **and L. E. Piña**[1]

[1] *Departamento de Matemáticas y Física, Centro de Ciencias Básicas, Universidad Autónoma de Aguascalientes, Mexico*

emails: jemacias@correo.uaa.mx, lepina@correo.uaa.mx

Abstract

In this work, we introduce a spatially discrete model that generalizes the well-known α-Fermi–Pasta–Ulam chain with damping. The system is perturbed at one end by a harmonic disturbance irradiating at a frequency in the forbidden band-gap of the classical regime, and a nonlocal coupling between the oscillators is considered using discrete Riesz fractional derivatives. We propose fully discrete expressions to approximate a pseudo-energy functional of the system, and we use them to calculate the total energy of fractional chains over a relatively long period of time. As an application, we provide evidence that the process of supratransmission is present in spatially discrete Fermi–Pasta–Ulam lattices with Riesz fractional derivatives in space.

Key words: wave equation, fractional nonlinear chains, energy-preserving method, explicit finite-difference scheme

1 Introduction

In this work, we investigate the following continuous form of the classical chain investigated by Fermi, Pasta and Ulam in [1]:

$$\frac{\partial^2 u}{\partial t^2}(x,t) = \frac{\partial^\alpha u}{\partial |x|^\alpha}(x,t)\left[1+\epsilon\left(\frac{\partial u}{\partial x}(x,t)\right)^p\right] - \gamma\frac{\partial u}{\partial t}(x,t), \quad \forall (x,t)\in\Omega\times[0,\infty),$$
$$\begin{cases} u(x,0)=\phi(x), & \forall x\in\Omega,\\ \dfrac{\partial u}{\partial t}(x,0)=\psi(x), & \forall x\in\Omega,\\ u(a,t)=f(t), & \forall t\in[0,\infty),\\ u(b,t)=g(t), & \forall t\in[0,\infty), \end{cases} \tag{1}$$

 ISBN: 978-84-617-8694-7

where $\phi, \psi : [a, b] \to \mathbb{R}$ and $f, g : [0, \infty) \to \mathbb{R}$ are sufficiently smooth functions. In this expression, we employ the Riesz fractional derivative of order $\alpha \in (1, 2)$, which is given by

$$\frac{\partial^\alpha u}{\partial |x|^\alpha} = \frac{-1}{2\cos(\frac{\pi\alpha}{2})\Gamma(2-\alpha)} \frac{d^2}{dx^2} \int_{-\infty}^{\infty} \frac{u(\xi, t)}{|x-\xi|^{\alpha-1}} d\xi, \quad \forall (x, t) \in \Omega \times [0, \infty). \tag{2}$$

Here $\Gamma(z)$ represents the gamma function, for each $z \in \mathbb{R} \setminus \{n : -n \in \mathbb{N} \text{ or } n = 0\}$. In the absence of a physically meaningful formulation of the Euler–Lagrange formality for fractional systems, we propose the following pseudo-Hamiltonian for the undamped scenario:

$$H(x, t) = \frac{1}{2} \left[\left(\frac{\partial u}{\partial t}(x, t) \right)^2 + u(x, t) \frac{\partial^\alpha u}{\partial |x|^\alpha}(x, t) \right] + \frac{\epsilon}{(p+2)(p+1)} \left(\frac{\partial u}{\partial x}(x, t) \right)^{p+2}. \tag{3}$$

Under these circumstances, a functional for the total energy of the system would be

$$E(t) = \int_a^b H(x, t) dx, \quad \forall t \geq 0. \tag{4}$$

2 Discrete model

Motivated by the continuous space-fractional equation (1), we propose now a spatially discrete fractional array that extends the original FPU chains. Recall that for each $f : \mathbb{R} \to \mathbb{R}$, each $h > 0$ and each $\alpha > -1$ the *fractional centered difference* of order α of f at the point x and step h is

$$\Delta_h^\alpha f(x) = \sum_{k=-\infty}^{\infty} g_k^\alpha f(x - kh), \quad \forall x \in \mathbb{R}, \tag{5}$$

where

$$g_k^\alpha = \frac{(-1)^k \Gamma(\alpha + 1)}{\Gamma(\frac{\alpha}{2} - k + 1)\Gamma(\frac{\alpha}{2} + k + 1)}, \quad \forall k \in \mathbb{Z}. \tag{6}$$

In the case that $1 < \alpha < 2$ the fractional centered differences satisfy (see [3])

$$\lim_{h \to 0} \frac{-1}{h^\alpha} \Delta_h^\alpha f(x) = \frac{\partial^\alpha f}{\partial |x|^\alpha}(x), \quad \forall x \in \mathbb{R}. \tag{7}$$

In the present work, we let $h = 1/c$ for the spatially discrete scenario. For simplification purposes, we convey that $\Delta_x^\alpha = \Delta_1^\alpha$. The Riesz space-fractional FPU chain under

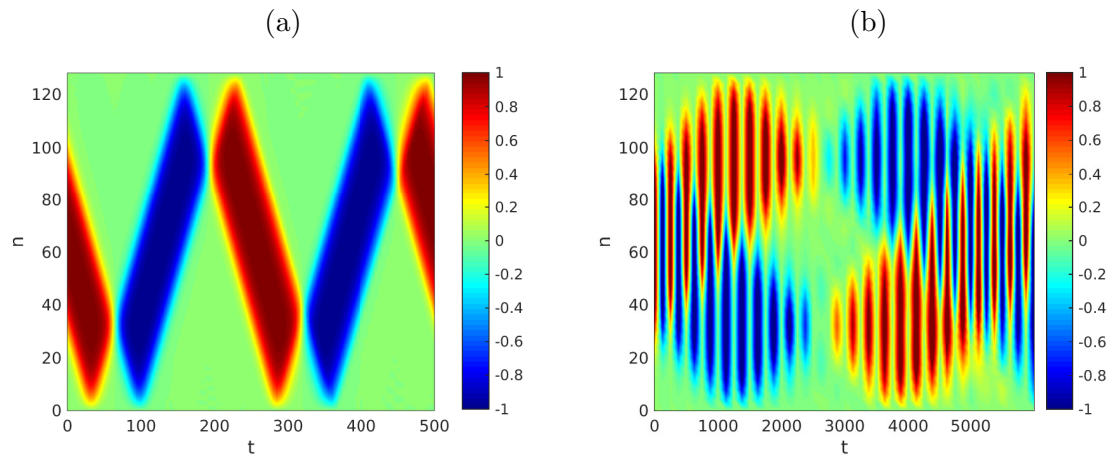


Figure 1: (Color online). Graphs of the approximate solution of a system of $N = 128$ anharmonic oscillators governed by (8) versus node number n and time t for periods of time (a) $T = 500$ and (b) $T = 6000$. The model parameters are $\alpha = 2$, $p = c = 1$, $\epsilon = 0.01$ and $\gamma = 0$, with initial-boundary data $\psi = f = g \equiv 0$ and $\phi(n) = \chi(n, 0)$ for each $n \in \{0, 1, \ldots, N\}$. The function χ is given by (12) with $\kappa = 0.1$. Computationally, we used $\tau = 0.02$.

consideration in this work is given by the system

$$\frac{d^2u_n}{dt^2}(t) = c^\alpha \Delta_x^\alpha u_n(t) + \delta_x V'(\delta_x u_{n-1}(t)) - \gamma \frac{du_n}{dt}(t), \quad \forall n \in \{0, 1, \ldots, N\}, \forall t \in [0, \infty),$$
$$\begin{cases} u_n(0) = \phi(n), & \forall n \in \{0, 1, \ldots, N\}, \\ \dfrac{du_n}{dt}(0) = \psi(n), & \forall n \in \{0, 1, \ldots, N\}, \\ u_0(t) = f(t), & \forall t \in [0, \infty), \\ u_N(t) = g(t), & \forall t \in [0, \infty). \end{cases} \tag{8}$$

Obviously, in this case $\phi, \psi : \{0, 1, \ldots, N\} \to \infty$. In the practice, we will consider a discrete domain of the form $\{0, 1, \ldots, N\}$ for some $N \in \mathbb{N}$, so

$$\Delta_x^\alpha u_n = -\frac{1}{h^\alpha} \sum_{k=n-N}^{n} g_k^\alpha u_{n-k} - \frac{1}{h^\alpha} \sum_{k=0}^{N} g_{n-k}^\alpha u_k, \tag{9}$$

for each $n \in \{1, \ldots, N-1\}$. Under these circumstances the pseudo-Hamiltonian and the

 ISBN: 978-84-617-8694-7

pseudo-energy of the undamped system will be defined respectively as

$$H_n(t) = \frac{1}{2}\left[\left(\frac{du_n}{dt}(t)\right)^2 + c^\alpha u_n(t)\Delta_x^\alpha u_n(t)\right] + V(\delta_x u_n(t)), \tag{10}$$

$$E(t) = \sum_{n=1}^{\infty} H_n(t) + \frac{c^\alpha}{2} u_0(t)\Delta_x^\alpha u_0(t) + V(\delta_x u_0(t)). \tag{11}$$

3 Numerical results

Our last example is motivated by the Toda lattice used to describe the motion of particles in one-dimensional nonlinear crystals [4]. Recall that the damped Toda lattice is described by (8) with $c = 0$, $\epsilon = 1$ and $V(u) = e^{-u} + u - 1$ for $u \in \mathbb{R}$. This system has soliton/anti-soliton solutions. In our simulations, we will use a superposition of those solutions of the form

$$\chi(n,t) = 5\ln\left\{\left(\frac{1+\exp\left[2(\kappa(n-97)+t\sinh\kappa)\right]}{1+\exp\left[2(\kappa(n-96)+t\sinh\kappa)\right]}\right)\left(\frac{1+\exp\left[2(\kappa(n-32)+t\sinh\kappa)\right]}{1+\exp\left[2(\kappa(n-33)+t\sinh\kappa)\right]}\right)\right\}, \tag{12}$$

for each $n \in \{0, 1, \ldots, N\}$ and $t \geq 0$, and $\kappa \in \mathbb{R}$.

Consider $N = 128$ nodes satisfying (8) with $\epsilon = 0.01$, $\gamma = 0$ and $\kappa = 0.1$, and let $\phi(n) = \chi(n, 0)$ for each $n \in \{0, 1, \ldots, N\}$. The approximate solutions of this chain are shown in Figure 1 for (a) $T = 500$ and (b) $T = 6000$. Again, the results are in agreement with those reported in [2].

References

[1] Enrico Fermi, J Pasta, and S Ulam. Studies of nonlinear problems. *Los Alamos Report LA-1940*, 978, 1955.

[2] Jorge Eduardo Macías-Díaz. An explicit finite-difference method for the approximate solutions of a generic class of anharmonic dissipative nonlinear media. *Numerical Methods for Partial Differential Equations*, 26(6):1351–1376, 2010.

[3] Manuel Duarte Ortigueira. Riesz potential operators and inverses via fractional centred derivatives. *International Journal of Mathematics and Mathematical Sciences*, 2006, 2006.

[4] Morikazu Toda. Waves in nonlinear lattice. *Progress of Theoretical Physics Supplement*, 45:174–200, 1970.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

A finite-difference method that preserves the dissipation of energy of a fractional sine-Gordon equation

J. E. Macías-Díaz[1] and L. F. Martínez-Álvarez[2]

[1] *Departamento de Matemáticas y Física, Centro de Ciencias Básicas, Universidad Autónoma de Aguascalientes, Mexico*

[2] *Centro de Ciencias Básicas, Universidad Autónoma de Aguascalientes, Mexico*

emails: jemacias@correo.uaa.mx, lfmartinez@correo.uaa.mx

Abstract

We consider an initial-boundary-value problem governed by a $(1+1)$-dimensional hyperbolic partial differential equation with constant damping that generalizes many nonlinear wave equations from mathematical physics. The model contemplates the presence of a spatial Laplacian of fractional order which is defined in terms of Riesz fractional derivatives, as well as the inclusion of a generic continuously differentiable potential. It is known that the undamped regime has an energy functional that is preserved throughout time under suitable boundary conditions. To approximate the solutions of this model, we propose a finite-difference discretization of our model based on fractional centered differences. Some discrete quantities are proposed here to estimate the energy functionals, and we show that the numerical method is capable of conserving the discrete energy under the same boundary conditions for which the continuous model conserves it. The method is is both stable and convergent.

Key words: sine-Gordon equation, Riesz space-fractional equation, energy-preserving method, stability and convergence analyses

1 Introduction

In this manuscript we let $T > 0$ and $\gamma \in \mathbb{R}$, and suppose that $a < b$. Throughout we will assume that $\alpha \in (1, 2)$ and $\Omega = (a, b) \times (0, T) \subset \mathbb{R}^2$. We will employ $\overline{\Omega}$ to represent the closure of Ω in $\mathbb{R}^2$. Suppose that $G : \mathbb{R} \to \mathbb{R}$, that $\phi, \psi : [a, b] \to \mathbb{R}$ and that $f, g : [0, T] \to \mathbb{R}$ are all continuously differentiable and satisfy the conditions $\phi(a) = f(0)$, $\phi(b) = g(0)$,

 ISBN: 978-84-617-8694-7

$\psi(a) = f'(0)$ and $\psi(b) = g'(0)$. Moreover, we will suppose that $u : \overline{\Omega} \to \mathbb{R}$ is a sufficiently smooth function that satisfies the initial-boundary-value problem

$$\frac{\partial^2 u}{\partial t^2}(x,t) - \frac{\partial^\alpha u}{\partial |x|^\alpha}(x,t) + \gamma \frac{\partial u}{\partial t}(x,t) + G'(u(x,t)) = 0, \quad \forall (x,t) \in \Omega,$$
$$\text{such that } \begin{cases} u(x,0) = \phi(x), & \forall x \in (a,b), \\ \dfrac{\partial u}{\partial t}(x,0) = \psi(x), & \forall x \in (a,b), \\ u(a,t) = f(t), & \forall t \in (0,T), \\ u(b,t) = g(t), & \forall t \in (0,T). \end{cases} \tag{1}$$

For convenience, we let $u(x,t) = 0$ for each $x \in (\mathbb{R} \setminus [a,b]) \times [0,T]$ and define (see [1])

$$\frac{\partial^\alpha u}{\partial |x|^\alpha}(x,t) = \frac{-1}{2\cos(\frac{\pi\alpha}{2})\Gamma(2-\alpha)} \frac{d^2}{dx^2} \int_{-\infty}^{\infty} \frac{u(\xi,t)}{|x-\xi|^{\alpha-1}} d\xi, \quad \forall (x,t) \in \Omega. \tag{2}$$

2 Numerical method

Let $h > 0$ and τ be step-sizes in space and time, respectively, and assume $N = T/\tau$ and $M = (b-a)/h$ are positive integers. Consider uniform partitions of $[a,b]$ and $[0,T]$, respectively, given by $x_j = jh$ and $t_n = n\tau$. In this work v_j^n will represent a numerical approximation to $u_j^n = u(x_j, t_n)$, for each $0 \le j \le M$ and each $0 \le n \le N$. Define

$$\mu_t u_j^n = \frac{u_j^{n+1} + u_j^n}{2} = u(x_j,t_n) + \mathcal{O}(\tau), \tag{3}$$
$$\mu_t^{(1)} u_j^n = \frac{u_j^{n+1} + u_j^{n-1}}{2} = u(x_j,t_n) + \mathcal{O}(\tau^2), \tag{4}$$
$$\delta_t u_j^n = \frac{u_j^{n+1} - u_j^n}{\tau} = \frac{\partial u}{\partial t}(x_j,t_n) + \mathcal{O}(\tau), \tag{5}$$
$$\delta_t^{(1)} u_j^n = \frac{u_j^{n+1} - u_j^{n-1}}{2\tau} = \frac{\partial u}{\partial t}(x_j,t_n) + \mathcal{O}(\tau^2), \tag{6}$$
$$\delta_t^{(2)} u_j^n = \frac{u_j^{n+1} - 2u_j^n + u_j^{n-1}}{\tau^2} = \frac{\partial^2 u}{\partial t^2}(x_j,t_n) + \mathcal{O}(\tau^2), \tag{7}$$

for each $1 \le j \le K-1$ and $1 \le n \le N-1$. Obviously, the right-hand sides of (3)–(7) denote the consistency properties of each of discrete operator. In addition, the following operator estimates $G'(u(x_j,t_n))$ with an order of consistency equal to $\mathcal{O}(\tau^2)$:

$$\delta_{u,t}^{(1)} G(u_j^n) = \frac{G(u_j^{n+1}) - G(u_j^{n-1})}{u_j^{n+1} - u_j^{n-1}}, \tag{8}$$

 ISBN: 978-84-617-8694-7

Definition 1. For any function $f : \mathbb{R} \to \mathbb{R}$, any $h > 0$ and any $\alpha > -1$ we define the *fractional centered difference* of order α of f at the point x as (see [2])

$$\Delta_h^\alpha f(x) = \sum_{k=-\infty}^{\infty} g_k^\alpha f(x - kh), \quad \forall x \in \mathbb{R}, \tag{9}$$

where

$$g_k^\alpha = \frac{(-1)^k \Gamma(\alpha + 1)}{\Gamma(\frac{\alpha}{2} - k + 1)\Gamma(\frac{\alpha}{2} + k + 1)}, \quad \forall k \in \mathbb{Z}. \tag{10}$$

Note that the series in the right-hand side of (9) converges absolutely for any bounded function $f \in L_1(\mathbb{R})$. It is easy to see that any function $f \in \mathcal{C}^5(\mathbb{R})$ for which all of its derivatives up to order five belong to $L_1(\mathbb{R})$, if $1 \leq j \leq M-1$ and $1 \leq n \leq N-1$ then

$$\frac{\partial^\alpha u}{\partial |x|^\alpha}(x_j, t_n) = -\frac{1}{h^\alpha} \sum_{k=(b-x_j)/h}^{(x_j-a)/h} g_k^\alpha u(x_j - kh, t_n) + \mathcal{O}(h^2) = \delta_h^\alpha u_j^n + \mathcal{O}(h^2), \tag{11}$$

where

$$\delta_h^\alpha u_j^n = -\frac{1}{h^\alpha} \sum_{k=0}^{M} g_{j-k}^\alpha u_k^n. \tag{12}$$

The finite-difference method to approximate the solution of (1) on Ω is given by

$$\begin{gathered} \delta_t^{(2)} v_j^n - \delta_h^\alpha v_j^n + \gamma \delta_t^{(1)} v_j^n + \delta_{v,t}^{(1)} G(v_j^n) = 0, \quad \forall j \in \{1, \ldots, M\}, \forall n \in \{1, \ldots, N\}, \\ \text{such that } \begin{cases} v_j^0 = \phi(x_j), & \forall j \in \{1, \ldots, M-1\}, \\ \delta_t v_j^0 = \psi(x_j), & \forall j \in \{1, \ldots, M-1\}, \\ v_0^n = f(t_n), & \forall n \in \{1, \ldots, N-1\}, \\ v_M^n = g(t_n), & \forall n \in \{1, \ldots, N-1\}. \end{cases} \end{gathered} \tag{13}$$

3 Energy invariants

In this section we would like to show that the finite-difference method (13) satisfies physical properties similar to those satisfied by (1) (see [3], for instance). For that reason we will suppose that the initial-boundary conditions satisfy the constraints

$$\begin{cases} f(t) = g(t) = 0, & \forall t \in [0, T], \\ \phi(x) = \psi(x) = 0, & \text{for } x = a, b. \end{cases} \tag{14}$$

Throughout this section, we will employ the spatial mesh

$$R_h = \left\{ (x_j)_{j=1}^{M-1} \in \mathbb{R}^{M-1} | x_j = a + jh \text{ for each } 1 \leq j \leq M-1 \right\}. \tag{15}$$

 ISBN: 978-84-617-8694-7

Let $\mathcal{V}_h$ be the vector space of all real grid functions on R_h. For any $u \in \mathcal{V}_h$ and $j \in \{1, \ldots, M-1\}$ convey that $u_j = u(x_j)$. Moreover, define respectively the inner product $\langle \cdot, \cdot \rangle : \mathcal{V}_h \times \mathcal{V}_h \to \mathbb{R}$ and the norm $\| \cdot \|_1 : \mathcal{V}_h \to \mathbb{R}$ by

$$\langle u, v \rangle = h \sum_{j=1}^{M-1} u_j v_j, \quad \|u\|_1 = h \sum_{j=1}^{M-1} |u_j|, \tag{16}$$

for any $u, v \in \mathcal{V}_h$. The Euclidean norm induced by $\langle \cdot, \cdot \rangle$ will be denoted by $\| \cdot \|_2$.

In the following, we will represent the solutions of the finite-difference method (13) by $(v^n)_{n=0}^N$, where we convey that $v^n = (v_1^n, \ldots, v_{M-1}^n)$ for each $0 \leq n \leq N$. We will also need the following real matrix of size $(M-2) \times (M-2)$:

$$A = \begin{pmatrix} g_0^\alpha & g_{-1}^\alpha & \cdots & g_{2-M}^\alpha \\ g_1^\alpha & g_0^\alpha & \cdots & g_{3-M}^\alpha \\ \vdots & \vdots & \ddots & \vdots \\ g_{M-2}^\alpha & g_{M-3}^\alpha & \cdots & g_0^\alpha \end{pmatrix}. \tag{17}$$

We will require the following lemma.

Lemma 2. *There exists a unique linear positive operator $\Lambda^\alpha : \mathcal{V}_h \to \mathcal{V}_h$ such that*

$$\langle -\delta_h^\alpha u, v \rangle = \langle \Lambda_h^\alpha u, \Lambda_h^\alpha v \rangle, \quad \forall u, v \in \mathcal{V}_h. \tag{18}$$

The next theorem establishes the existence of invariants for the discrete system (13).

Theorem 3. *Let $(v^n)_{n=0}^N$ be solution of the system* (13) *under the conditions* (14)*. Let*

$$E^n = \frac{1}{2} \|\delta_t v^n\|_2^2 + \frac{1}{2} \left\langle \Lambda^\alpha v^n, \Lambda^\alpha v^{n+1} \right\rangle + \|\mu_t G(v^n)\|_1, \quad \forall n \in \{1, \ldots, N-1\}. \tag{19}$$

Then for each $1 \leq n \leq N-1$ the following identity holds: $\delta_t E^n = -\gamma \|\delta_t^{(1)} v^{n+1}\|_2^2$. In particular, the quantities E^n are invariants of (13) *if $\gamma = 0$ and the conditions* (14) *hold.* □

References

[1] Cem Çelik and Melda Duman. Crank–Nicolson method for the fractional diffusion equation with the Riesz fractional derivative. *Journal of Computational Physics*, 231(4):1743–1750, 2012.

[2] Manuel Duarte Ortigueira. Riesz potential operators and inverses via fractional centred derivatives. *International Journal of Mathematics and Mathematical Sciences*, 2006, 2006.

[3] Pengde Wang and Chengming Huang. An energy conservative difference scheme for the nonlinear fractional Schrödinger equations. *Journal of Computational Physics*, 293:238–251, 2015.

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

A positive and linear approach to solve some nonlinear fractional diffusion-reaction equations

J. E. Macías-Díaz[1] and Axel Chávez-Guzmán[2]

[1] *Departamento de Matemáticas y Física, Centro de Ciencias Básicas, Universidad Autónoma de Aguascalientes, Mexico*

[2] *Centro de Ciencias Básicas, Universidad Autónoma de Aguascalientes, Mexico*

emails: jemacias@correo.uaa.mx, faxelchavezgusman@gmail.com

Abstract

In this work, we depart from the well-known one-dimensional Fisher's equation from population dynamics, and consider an extension of this model using Riesz fractional derivatives in space. Positive and bounded initial-boundary data are imposed on a closed and bounded domain, and a fully discrete form of this fractional initial-boundary-value problem is provided next using fractional centered differences. The fully discrete population model is implicit and linear, so a convenient vector representation is readily derived. Under suitable conditions, the matrix representing the implicit problem is an inverse-positive matrix. Using this fact, we establish that the discrete population model is capable of preserving the positivity and the boundedness of the discrete initial-boundary conditions. Moreover, the computational solubility of the discrete model is tackled in the closing remarks.

Key words: discrete fractional Huxley's equation, fractional centered differences, discrete fractional population model, inverse-positive matrices, positivity preservation, boundedness preservation

1 Introduction

Throughout we will assume that $a, b \in \mathbb{R}$ satisfy $a < b$, and that κ, λ, K and T are positive numbers. Let $\phi : [a, b] \to \mathbb{R}$ and $\psi_1, \psi_2 : [0, T] \to \mathbb{R}$ be continuously differentiable functions whose ranges are subsets of $[0, K]$. Assume additionally that the compatibility conditions $\phi(a) = \psi_1(0)$ and $\phi(b) = \psi_2(0)$ are satisfied. Let $1 < \alpha < 2$ and define $\Omega = (a, b) \times (0, T)$.

ISBN: 978-84-617-8694-7

In this work, we will suppose that $u : \overline{\Omega} \to \mathbb{R}$ is a sufficiently smooth function that satisfies the initial-boundary-value problem

$$\begin{gathered}\frac{\partial u}{\partial t}(x,t) - \kappa \frac{\partial^\alpha u}{\partial |x|^\alpha}(x,t) = \lambda u(x,t)(1-u(x,t)/K), \\ \text{such that } \begin{cases} u(x,0) = \phi(x), & \forall x \in (a,b), \\ u(a,t) = \psi_1(t), & \forall t \in (0,T), \\ u(b,t) = \psi_2(t), & \forall t \in (0,T), \end{cases}\end{gathered} \tag{1}$$

for each $(x,t) \in \Omega$.

For the sake of convenience, we will define $u(x,t) = 0$ for each $x \in (-\infty, a) \cup (b, \infty)$ and each $t \in [0,T]$. Using this convention, the space-fractional operator in (1) is the Riesz fractional derivative of order α, which is given by

$$\frac{\partial^\alpha u}{\partial |x|^\alpha}(x,t) = \frac{-1}{2\cos(\frac{\pi\alpha}{2})\Gamma(2-\alpha)} \frac{d^2}{dx^2} \int_{-\infty}^{\infty} \frac{u(\xi,t)}{|x-\xi|^{\alpha-1}} d\xi, \tag{2}$$

for each $(x,t) \in \Omega$. Here Γ is the gamma function, which is defined in $\mathbb{R} \setminus \{n : -n \in \mathbb{N} \text{ or } n = 0\}$ by

$$\Gamma(z) = \int_0^\infty s^{z-1} e^{-s} ds. \tag{3}$$

Note that the partial differential equation of (1) is actually a space-fractional extension of the model from population dynamics investigated simultaneously and independently by R. A. Fisher [2] and A. N. Kolmogorov, I. G. Petrovskii and N. Piskunov [3] in 1937.

2 Preliminary results

In this work we follow a fractional difference approach to approximate the solutions of (1), and use fractional centered differences to approximate Riesz space-fractional derivatives. For any function $f : \mathbb{R} \to \mathbb{R}$, any $h > 0$ and any $\alpha > -1$, the fractional centered difference of order α of f at the point x is defined as

$$\Delta_h^\alpha f(x) = \sum_{k=-\infty}^{\infty} g_k^\alpha f(x - kh), \quad \forall x \in \mathbb{R}, \tag{4}$$

where

$$g_k^\alpha = \frac{(-1)^k \Gamma(\alpha+1)}{\Gamma(\frac{\alpha}{2} - k + 1)\Gamma(\frac{\alpha}{2} + k + 1)}, \quad \forall k \in \mathbb{Z}. \tag{5}$$

In the case that $1 < \alpha \leq 2$, the fractional centered differences satisfy (see [4])

$$\lim_{h \to 0} \frac{-1}{h^\alpha} \Delta_h^\alpha f(x) = \frac{\partial^\alpha f}{\partial |x|^\alpha}(x), \quad \forall x \in \mathbb{R}. \tag{6}$$

 ISBN: 978-84-617-8694-7

Lemma 1 (See [1]). *If $1 < \alpha < 2$ then the coefficients $(g_k^\alpha)_{k=-\infty}^{\infty}$ satisfy:*

(a) $g_0^\alpha \geq 0$,

(b) $g_k^\alpha = g_{-k}^\alpha < 0$ *for all* $k \geq 1$, *and*

(c) $\sum_{k=-\infty}^{\infty} g_k^\alpha = 0$. □

Lemma 2 (See [1]). *Let $f \in \mathcal{C}^5(\mathbb{R})$, and assume that all its derivatives up to order five belong to $L_1(\mathbb{R})$. If $1 < \alpha < 2$ then*

$$-\frac{\Delta_h^\alpha f(x)}{h^\alpha} = \frac{\partial^\alpha f(x)}{\partial |x|^\alpha} + \mathcal{O}(h^2). \tag{7}$$

□

3 Numerical methodology

For the remainder of this work, we assume that $M, N \in \mathbb{N}$ and use partition norms $h = (b-a)/M$ and $\tau = T/N$ for the intervals $[a, b]$ and $[0, T]$, respectively. For each $j \in \{0, 1, \ldots, M\}$ and each $n \in \{0, 1, \ldots, N\}$ we define

$$x_j = a + jh, \tag{8}$$
$$t_n = n\tau, \tag{9}$$
$$u_j^n = u(x_j, t_n). \tag{10}$$

Under these circumstances, note that

$$\begin{aligned} \frac{\partial^\alpha u}{\partial |x|^\alpha}(x_j, t_n) &= -\frac{1}{h^\alpha} \sum_{k=j-M}^{j} g_k^\alpha u_{j-k}^n + \mathcal{O}(h^2) \\ &= -\frac{1}{h^\alpha} \sum_{k=0}^{M} g_{j-k}^\alpha u_k^n + \mathcal{O}(h^2), \end{aligned} \tag{11}$$

for each $j \in \{1, \ldots, M-1\}$ and each $n \in \{1, \ldots, N-1\}$. In the following, we will consider the discrete operators

$$\delta_t u_j^n = \frac{u_j^{n+1} - u_j^n}{\tau}, \tag{12}$$
$$\Delta_h^\alpha u_j^n = -\frac{1}{h^\alpha} \sum_{k=0}^{M} g_{j-k}^\alpha u_k^n. \tag{13}$$

 ISBN: 978-84-617-8694-7

Using this nomenclature, the fully discrete population model used in this work to approximate the solutions of (1) is given by the discrete system of equations

$$\begin{gathered}\delta_t u_j^n - \kappa \Delta_h^\alpha u_j^{n+1} = \lambda u_j^{n+1}(1 - u_j^n/K), \\ \text{such that } \left\{ \begin{array}{ll} u_j^0 = \phi(x_j), & \forall j \in \{0, 1, \ldots, M\}, \\ u_0^n = f(a), & \forall n \in \{0, 1, \ldots, N\}, \\ u_M^n = g(b), & \forall n \in \{0, 1, \ldots, N\}, \end{array} \right.\end{gathered} \tag{14}$$

for each $j \in \{1, \ldots, M-1\}$ and $n \in \{1, \ldots, N-1\}$. In this work, we will derive various analytical and numerical properties of (14).

References

[1] Cem Çelik and Melda Duman. Crank–Nicolson method for the fractional diffusion equation with the Riesz fractional derivative. *Journal of Computational Physics*, 231(4):1743–1750, 2012.

[2] Ronald Aylmer Fisher. The wave of advance of advantageous genes. *Annals of eugenics*, 7(4):355–369, 1937.

[3] Andrei Nikolaevitch Kolmogorov, Ivan G Petrovskii, and Nikolaï S Piskunov. A study of the equation of diffusion with increase in the quantity of matter, and its application to a biological problem. *Bjul. Moskovskogo Gos. Univ*, 1(7):1–26, 1937.

[4] Manuel Duarte Ortigueira. Riesz potential operators and inverses via fractional centred derivatives. *International Journal of Mathematics and Mathematical Sciences*, 2006, 2006.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

A Mathematical Model for the Propagation of Bovine Tuberculosis in Wild Animals

Luciana Mafalda Elias de Assis[1], Eduardo Massad[2], Silvia Raimundo Martorano[2], Raul Abreu de Assis[1] and Ezio Venturino[3]

[1] *Faculdade de Ciências Exatas e Tecnológicas,, Universidade do Estado de Mato Grosso, BRAZIL*

[2] *Faculdade de Medicina, Universidade de São Paulo, BRAZIL*

[3] *Dipartimento di Matematica "Giuseppe Peano", Università di Torino, ITALY*

emails: lucianam@unemat-net.br, edmassad@dim.fm.usp.br, silviamr@dim.fm.usp.br, raulaassis@gmail.com, ezio.venturino@unito.it

Abstract

In this paper a model for the spread of TB between wild herbivores and their predators is presented and analysed. The most important system parameters are identified: vertical and horizontal disease transmission among the prey, the influence that intraspecific competition between healthy and diseased prey has on the infected prey population. Removal of diseased-prey may be the most effective strategy to free the ecosystem from the disease.

Key words: herd behavior, disease transmission
MSC 2000: AMS 92D25, 92D30, 92D40

1 Introduction

Bovine Tuberculosis (TB) is a threat to wildlife health. The transmission of TB between herd individuals occurs most frequently by aerosol, [4] and predators contract the disease mostly by ingestion of tuberculoses tissues [16]. It directly impacts animal productivity and fitness and can lead to an increase in the mortality rate.

Wild animals appear to be able to harbour mycobacteria for months to years. As infection progresses, there is evidence that TB may decrease reproductive and other fitness

 ISBN: 978-84-617-8694-7

parameters. However, it may not significantly affect them unless they experience other stressors, such as drought or concurrent disease, suggesting that infected animals may remain in the population for prolonged periods [16, 17]. For instance, buffalo herds with a higher prevalence of bovine TB had worse body condition in the dry season than those with lower bovine TB prevalence. Consequently, affected buffalo might be more susceptible to predation by lions, [7]. In addition the loss of these prey animals may influence the predators. A simulation over 50 years of the the impact of bovine TB on lions shows a scenario suggesting a serious threat to the species survival, [17].

To study the impact of TB on wild animals, we present a predator-prey model involving buffalos and lions, both subject to the disease, which is horizontally as well as vertically transmitted. Because the wild buffalo congregates in huge herds, the border individuals are usually captured. Mathematically, this can be modeled via a square root response function, [1, 2, 25, 6]. Indeed since the total population of prey, say P, occupies a certain area A on the ground, the number of the individuals who are found at the border is proportional to the length of the perimeter of the patch A, which in turn is proportional to $\sqrt{P}$.

The paper is organized as follows. In the next section, we formulate the model, show that trajectories are bounded and adimensionalize it. The feasibility and stability analysis are in Section 3. Section 4 contains a study about the basic reproduction number $\mathcal{R}_0$ where we focus the discussion on conditions for the eradication of the disease. Transcritical and Hopf bifurcations are investigated in Section 5. Section 6 shows that the parameter m is crucial for the feasibility and stability of points E_1 and E_2, respectively. The establishment of the disease is directly related to the value of $\mathcal{R}_0$ at both points. To investigate the behavior of model (4) in relation to the parameters, we first separate the simulations in two specific cases. In the first case, E_1 is stable and E_2 is unstable, that is, $m > a$. The second case is obtained when $a/\sqrt{3} < m < a$, in which E_1 is unstable and E_2 is feasible. A final discussion concludes the paper.

2 Model Formulation

We consider a model for species interactions, subject to a TB that can infect both the buffalos and lions. Once infected, an animal remains infected for its life, so that the disease is of type SI. Infected prey $\widehat{I}$ become weak and are left behind the herd. Let further $\widehat{R}$ denote the susceptible prey population, $\widehat{F}$ and $\widehat{W}$ respectively the susceptible and infected predators. The prey have a highly socialized "herd behavior", in that they live together wandering in search of pastures. They are generally followed in their wanderings by the predators. If the hunt is successful for the predator, generally the prey individuals residing on the boundary of herd are harmed. The model, accounting for all the possible population

 ISBN: 978-84-617-8694-7

interactions described below, reads as follows:

$$\frac{d\widehat{R}}{d\tau} = \tilde{r}\widehat{R} + \bar{r}\,(1-\widehat{\alpha})\,\widehat{I} - k\widehat{R}\left(\widehat{R}+\widehat{g}\widehat{I}\right) - \widehat{\lambda}\widehat{R}\widehat{I} - \widehat{cp}\sqrt{\widehat{R}}\widehat{W} - \widehat{a}\sqrt{\widehat{R}}\widehat{F} - \widehat{\theta}\,(1-\widehat{p})\,\widehat{c}\sqrt{\widehat{R}}\widehat{W}, \quad (1)$$

$$\frac{d\widehat{I}}{d\tau} = \bar{r}\widehat{\alpha}\widehat{I} + \widehat{I}\left(\widehat{\lambda}\widehat{R} - \widehat{b}\widehat{F} - \widehat{\mu} - \widehat{\ell}\widehat{W}\right) + \widehat{\theta}(1-\widehat{p})\widehat{c}\sqrt{\widehat{R}}\widehat{W} - k\widehat{I}(\widehat{u}\widehat{R}+\widehat{q}\widehat{I}),$$

$$\frac{d\widehat{F}}{d\tau} = \widehat{ae}\sqrt{\widehat{R}}\widehat{F} + (1-\widehat{\sigma})\widehat{b}\widehat{e}\widehat{I}\widehat{F} + (1-\widehat{\gamma})\widehat{W}(\widehat{ce}\widehat{p}\sqrt{\widehat{R}} + \widehat{\ell}\widehat{e}\widehat{I}) - \widehat{m}\widehat{F} - \widehat{\beta}\widehat{F}\widehat{W},$$

$$\frac{d\widehat{W}}{d\tau} = \widehat{W}\left(\widehat{\gamma}\widehat{ce}\widehat{p}\sqrt{\widehat{R}} + \widehat{\gamma}\widehat{\ell}\widehat{e}\widehat{I} - \widehat{\nu}\right) + \widehat{\sigma}\widehat{b}\widehat{e}\widehat{I}\widehat{F} + \widehat{\beta}\widehat{F}\widehat{W}.$$

The first equation describes the dynamics of $\widehat{R}$ (susceptible prey). The term $\tilde{r}\widehat{R}$ expresses growth rate of $\widehat{R}$ due to their own reproduction and $\bar{r}(1-\widehat{\alpha})\widehat{I}$ is the fraction of $\widehat{I}$ that are born healthy (vertical transmission). The term $k\widehat{R}\left(\widehat{R}+\widehat{g}\widehat{I}\right)$ is the mortality by intraspecific competition between $\widehat{R}$ individuals among themselves and with $\widehat{I}$ individuals. The term $\widehat{\lambda}\widehat{R}\widehat{I}$ represents the susceptible prey, that become infected (horizontal transmission). The term $\widehat{cp}\sqrt{\widehat{R}}\widehat{W}$ represents the prey individuals that are captured by infected predators and $\widehat{a}\sqrt{\widehat{R}}\widehat{F}$ is the capture rate of $\widehat{R}$ by $\widehat{F}$. Finally, $\widehat{\theta}\,(1-\widehat{p})\,\widehat{c}\sqrt{\widehat{R}}\widehat{W}$ are the new infections by an unsuccessful attack of infected predator $\widehat{W}$ on healthy prey $\widehat{R}$, and then latter gets disease. Note that, we have a fraction $(1-\widehat{\theta})(1-\widehat{p})$ of healthy prey that are not captured, but do not get the disease from predators.

The second equation describes the dynamics of $\widehat{I}$ (infected prey). The term $\bar{r}\widehat{\alpha}\widehat{I}$ is the fraction of the reproduction rate of $\widehat{I}$ that is born infected (possibility of vertical transmission), $\widehat{\lambda}\widehat{R}\widehat{I}$ represents the susceptibles prey, that become infected (horizontal transmission), $\widehat{b}\widehat{I}\widehat{F}$ is the predation of $\widehat{I}$ by $\widehat{F}$, $\widehat{\mu}\widehat{I}$ is the mortality of $\widehat{I}$ (disease-related) and $\widehat{\ell}\widehat{I}\widehat{W}$ is the predation of $\widehat{I}$ by $\widehat{W}$. Finally, $\widehat{\theta}\,(1-\widehat{p})\,\widehat{c}\sqrt{\widehat{R}}\widehat{W}$ are the new infections by an unsuccessful attack of infected predator $\widehat{W}$ on healthy prey $\widehat{R}$, and then latter gets disease and $k\widehat{I}(\widehat{u}\widehat{R}+\widehat{q}\widehat{I})$ is the mortality by intraspecific competition of $\widehat{I}$ among themselves and with $\widehat{R}$. Note that, in the model, infected prey do not benefit from the effects of herd behavior.

The third equation describes the dynamics of $\widehat{F}$ (susceptible predator). The term $\widehat{ae}\sqrt{\widehat{R}}\widehat{F}$ expresses the increase of $\widehat{F}$ due to the consumption of $\widehat{R}$ on the boundary with conversion factor $0 < \widehat{e} < 1$. The term $(1-\widehat{\sigma})\widehat{b}\widehat{e}\widehat{I}\widehat{F}$ represents the growth rate of $\widehat{F}$ due to the consumption of $\widehat{I}$, that is, predators that consume infected prey but do not become infected. The term $(1-\widehat{\gamma})\widehat{W}(\widehat{ce}\widehat{p}\sqrt{\widehat{R}})$ expresses the fraction of the reproduction rate of $\widehat{W}$ that is born as a healthy predator due to the consumption $\widehat{R}$ on boundary. Besides that, $(1-\widehat{\gamma})\widehat{W}(\widehat{\ell}\widehat{e}\widehat{I})$ is the analogous term, but due to the consumption of infected prey $\widehat{I}$. Now, the term $\widehat{m}\widehat{F}$ is the mortality of $\widehat{F}$, $\widehat{\beta}\widehat{F}\widehat{W}$ represents the predator who moves from one class to another, that is, susceptibles become infected by contact with another infected predator $\widehat{W}$ (horizontal transmission).

 ISBN: 978-84-617-8694-7

The fourth equation describes the dynamics of $\widehat{W}$ (infected predator). The term $\widehat{\gamma}\widehat{c}\widehat{e}\widehat{p}\sqrt{\widehat{R}}\widehat{W}$ represents the vertical transmission of $\widehat{W}$ in converting captured $\widehat{R}$ on boundary in infected predators $\widehat{W}$. The term $\widehat{\gamma}\widehat{\ell}\widehat{e}\widehat{I}\widehat{W}$ is the vertical transmission of $\widehat{W}$ due to the consumption of $\widehat{I}$, $\widehat{\nu}\widehat{W}$ is the mortality of $\widehat{W}$ (disease-related) and $\widehat{\sigma}\widehat{b}\widehat{e}\widehat{I}\widehat{F}$ is the fraction of healthy predators that gives birth to infected offspring by eating $\widehat{I}$ (vertical transmission). Finally, $\widehat{\beta}\widehat{F}\widehat{W}$ represents the predator who moves from one class to another, that is, susceptibles become infected by contact with another infected predator $\widehat{W}$ (horizontal transmission). All parameters are non-negative and listed in Table 2.

2.1 Boundedness

Following closely [25], the system's trajectories are confined within a compact set. For the total environment population $\varphi(\tau) = \widehat{R}(\tau) + \widehat{I}(\tau) + \widehat{F}(\tau) + \widehat{W}(\tau)$, summing equations (1),

$$\begin{aligned}\frac{d\varphi(\tau)}{dt} = & \left(\tilde{r}\widehat{R} + \bar{r}\widehat{I}\right) - k\widehat{R}\left(\widehat{R} + \widehat{g}\widehat{I}\right) - k\widehat{I}\left(\widehat{u}\widehat{R} + \widehat{q}\widehat{I}\right) \\ & + (\widehat{e} - 1)\left(\widehat{c}\widehat{p}\sqrt{\widehat{R}}\widehat{W} + \widehat{a}\sqrt{\widehat{R}}\widehat{F} + \widehat{\ell}\widehat{I}\widehat{W}\right) - \left(\widehat{\mu}\widehat{I} + \widehat{m}\widehat{F} + \widehat{\nu}\widehat{W}\right).\end{aligned} \tag{2}$$

Using $(\widehat{e} - 1) \leq 0$ we can drop the term that contains it, to get

$$\frac{d\varphi(\tau)}{d\tau} \leq (\tilde{r}\widehat{R} + \bar{r}\widehat{I}) - k\widehat{R}(\widehat{R} + \widehat{g}\widehat{I}) - k\widehat{I}(\widehat{u}\widehat{R} + \widehat{q}\widehat{I}) - (\widehat{\mu}\widehat{I} + \widehat{m}\widehat{F} + \widehat{\nu}\widehat{W})$$

Letting M be the maximum value of the parabola $\Phi(\widehat{R}+\widehat{I}) = -k\eta_2(\widehat{R}+\widehat{I})^2+(\eta_1+\eta_2)(\widehat{R}+\widehat{I})$, with $\eta_1 = \max\{\tilde{r}, \bar{r}\}$, $\eta_2 = \min\{1, \widehat{g}, \widehat{u}, \widehat{q}\}$, $\eta_3 = \min\{\widehat{\mu}, \widehat{m}, \widehat{\nu}\}$, we find the estimate

$$\frac{d\varphi(t)}{dt} + \eta_3\varphi(t) \leq \Phi(\widehat{R} + \widehat{I}) \leq \bar{M} = \frac{(\eta_1 + \eta_3)^2}{4\eta_2 K},$$

from which we establish the boundedness result

$$\varphi(t) \leq (\widehat{e})^{\eta_3 t}\frac{M}{\eta_3} + \varphi(0)(\widehat{e})^{-\eta_3 t} \leq \frac{M}{\eta_3} + \varphi(0) = \bar{M}$$

Note also that from below, the coordinate hyperplanes cannot be crossed toward negative values, although the system is not homogeneous. For instance, the first equation for $\widehat{R} = 0$ gives $\widehat{R}'(\tau) = \bar{r}\,(1 - \widehat{\alpha})\,\widehat{I} \geq 0$ because the parameters are nonnegative and $1 - \widehat{\alpha} \geq 0$.

The model (1) can be nondimensionalized via $R(t) = \frac{k}{\tilde{r}}\widehat{R}(\tau)$, $I(t) = \frac{k}{\tilde{r}}\widehat{I}(\tau)$, $F(t) = \frac{k}{e\tilde{r}}\widehat{F}(\tau)$, $W(t) = \frac{k}{e\tilde{r}}\widehat{W}(\tau)$, $t = \tilde{r}\tau$, and

$$\begin{aligned} r = \frac{\bar{r}}{\tilde{r}}, \quad g = \widehat{g}, \quad \lambda = \frac{\widehat{\lambda}}{k}, \quad c = \frac{\widehat{c}\widehat{e}}{\sqrt{\tilde{r}k}}, \quad a = \frac{\widehat{a}\widehat{e}}{\sqrt{\tilde{r}k}}, \quad b = \frac{\widehat{b}\widehat{e}}{k}, \\ \mu = \frac{\widehat{\mu}}{\tilde{r}}, \quad \ell = \frac{\widehat{\ell}\widehat{e}}{k}, \quad u = \widehat{u}, \quad q = \widehat{q}, \quad m = \frac{\widehat{m}}{\tilde{r}}, \quad \frac{\widehat{\beta}\widehat{e}}{k}, \end{aligned} \tag{3}$$

 ISBN: 978-84-617-8694-7

Table 1: Parameters of model (1) and their biological meanings

Parameter	Biological meaning
$\tilde{r}$	Specific growth rate of healthy prey $\widehat{R}$
$\bar{r}$	Specific growth rate of infected prey I ($\bar{r} \leq \tilde{r}$)
$\widehat{\alpha}$	Non-dimensional parameter that represents the fraction of prey $\widehat{R}$ that born infected (vertical transmission)
$1-\widehat{\alpha}$	Non-dimensional parameter that represents the fraction of prey $\widehat{R}$ that born no infected
k	Mortality of healthy prey $\widehat{R}$ due intraspecific competition ($k = \frac{r}{K}$)
K	Carrying capacity of prey $\widehat{R}$ in absence of predator
$\widehat{g}$	Non-dimensional parameter that regulates competition among infected prey $\widehat{I}$ with healthy prey $\widehat{R}$
$\widehat{\lambda}$	Infection rate in healty prey $\widehat{R}$ (horizontal transmission)
$\widehat{\theta}$	Probability that healthy prey $\widehat{R}$, not captured by infected predator $\widehat{W}$, becomes infected
$\widehat{p}$	Probability that represents the fraction of captured prey $\widehat{R}$
$1-\widehat{p}$	Probability that represents the fraction of no captured prey $\widehat{R}$
$\widehat{a}$	Predation rate of healthy prey $\widehat{R}$ by healthy predator $\widehat{F}$
$\widehat{c}$	Predation rate of healthy prey $\widehat{R}$ by infected predator $\widehat{W}$
$\widehat{b}$	Predation rate of infected prey $\widehat{I}$ by healthy predator $\widehat{F}$
$\widehat{\mu}$	Mortality rate of infected prey $\widehat{I}$
$\widehat{\ell}$	Predation rate of infected prey $\widehat{I}$ by infected predator $\widehat{W}$
$\widehat{u}$	Non-dimensional parameter that regulates how infected prey $\widehat{I}$ compete with healthy prey $\widehat{R}$
$\widehat{q}$	Non-dimensional parameter that regulates how infected prey $\widehat{I}$ compete with infected prey $\widehat{I}$
$\widehat{e}$	Efficiency of the predator in converting captured prey into reproductive success
$\widehat{\sigma}$	Non-dimensional parameter that regulates the infection rate in predator $\widehat{F}$ (vertical transmission)
$1-\widehat{\sigma}$	Non-dimensional parameter that represents the fraction of no infection of $\widehat{F}$ in the vertical transmission for the case where there is consumption of infected prey
$\widehat{m}$	Mortality rate of healthy predator $\widehat{F}$
$\widehat{\beta}$	Infection rate in healty predator $\widehat{F}$ (horizontal transmission)
$\widehat{\gamma}$	Probalility of vertical transmission in predator $\widehat{F}$
$1-\widehat{\gamma}$	Non-dimensional parameter that describes the fraction that rate of no infection in $\widehat{F}$ (no infected individual)
$\widehat{\nu}$	Mortality rate of infected predator $\widehat{W}$

 ISBN: 978-84-617-8694-7

to get re-scaled system:

$$\frac{dX}{dt} = f(X), \quad X = (R, I, F, W)^T, \quad f = (f_1, f_2, f_3, f_4)^T, \tag{4}$$

$$\begin{aligned} f_1 &= R + r\,(1-\alpha)\,I - R\,(R + gI) - \lambda RI - cp\sqrt{R}W - a\sqrt{R}F - \theta\,(1-p)\,c\sqrt{R}W, \\ f_2 &= \alpha I + I\,(\lambda R - bF - \mu - \ell W) + \theta(1-p)c\sqrt{R}W - I(uR + qI), \\ f_3 &= a\sqrt{R}F + (1-\sigma)bIF + (1-\gamma)cp\sqrt{R}W - mF + (1-\gamma)\ell IW - \beta FW, \\ f_4 &= \gamma cp\sqrt{R}W + \gamma\ell IW - \nu W + \sigma bIF + \beta FW. \end{aligned}$$

3 System's equilibria

Model (4) has five equilibria. The origin E_0, two disease-free equilibria, $E_1 = (1, 0, 0, 0)$, and $E_2 = (R_2, 0, F_2, 0)$, the predator-free $E_3 = (R_3, I_3, 0, 0)$ and coexistence $E_4 = (R_4, I_4, F_4, W_4)$ which is studied numerically. Their components are:

$$R_2 = \frac{m^2}{a^2}, \quad F_2 = \frac{m^2}{a^2}\left(\frac{1}{m} - \frac{m}{a^2}\right), \quad I_3 = \frac{\alpha - \mu}{q} + \frac{\lambda - u}{q}R_3 \tag{5}$$

and R_3 given by the roots of the quadratic equation

$$\Phi(R_3) = \alpha_2 R_3^2 + \alpha_1 R_3 + \alpha_0 = 0 \tag{6}$$

with

$$\alpha_2 = \frac{(g+\lambda)(u-\lambda) - q}{q}, \quad \alpha_0 = \frac{r(\alpha-1)(\mu-\alpha)}{q}, \quad \alpha_1 = \frac{(1-\alpha)r(\lambda-u) + (\mu-\alpha)(\lambda+g) + q}{q}.$$

3.1 Feasibility

Feasibility for E_2 is ensured by

$$m \le a \tag{7}$$

and in case of E_3 the feasibility conditions are for $I_3 \ge 0$

$$\alpha + \lambda R_3 \ge \mu + uR_3 \tag{8}$$

and for $R_3 \ge 0$ two positive roots exist if

$$\Delta = \alpha_1^2 - 4\alpha_2\alpha_0 > 0, \quad -\alpha_1\alpha_2^{-1} > 0, \quad \alpha_0\alpha_2^{-1} > 0$$

while at least one positive root is ensured by

$$\Delta = \alpha_1^2 - 4\alpha_2\alpha_0 > 0, \quad \alpha_0\alpha_2^{-1} < 0.$$

 ISBN: 978-84-617-8694-7

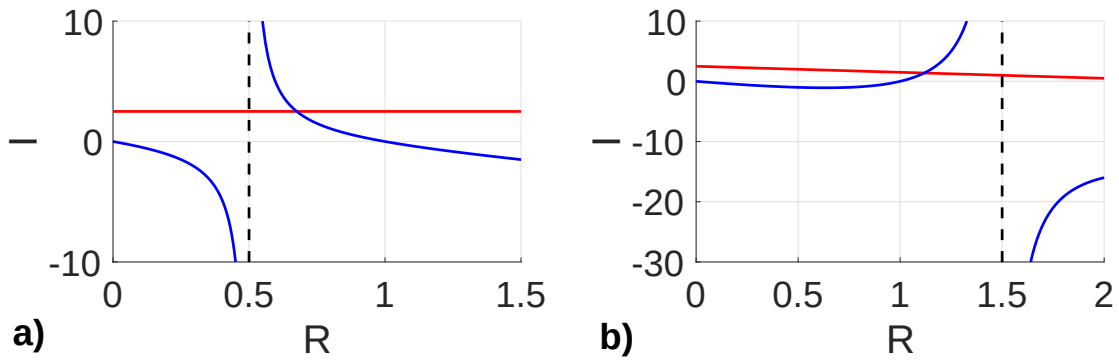


Figure 1: Nullclines in the plane $R-I$, considering $F=W=0$. If $\lambda+\alpha>u+\mu$ ($\lambda_1^{E_1}>1$), there is a unique feasible equilibrium point in the form $(R_3, I_3, 0, 0)$. a) Parameter values: $\alpha=0.5$, $q=g=0.1$, $r=0.75$ and $u=\mu=0.25$ we obtain $p^*=0.5$. b) Parameter values: $\alpha=0.5$, $q=g=0.1$, $r=0.75$ and $u=\mu=0.25$ we obtain $p^*=0.5$.

A feasible equilibrium solution for system (4) with $F=0$ and $W=0$ is an intersection in the first quadrant of the two curves:

$$f(R)=\frac{(R-1)R}{(1-\alpha)r-(g+\lambda)R}, \quad h(R)=\frac{(\alpha-\mu)+(\lambda-u)R}{q}. \tag{9}$$

Note that the function $f(R)$ has an asymptote in $R=p^*=r(1-\alpha)/(g+\lambda)$, thus there are two cases: $p^*>1$ and $p^*<1$, which respectively give:

$$f(R):\begin{cases}<0 & \text{if } 0<R<p^*\\ >0 & \text{if } p^*<R<1\\ <0 & \text{if } 1<R\end{cases} \qquad f(R):\begin{cases}<0 & \text{if } 0<R<1\\ >0 & \text{if } 1<R<p^*\\ <0 & \text{if } p^*<R\end{cases}. \tag{10}$$

Assume

$$\lambda+\alpha>\mu+u. \tag{11}$$

It follows thay $f(1)=0$ and f is unbounded around $(p^*-\epsilon, p^*+\epsilon)$, $\epsilon>0$, since $h(R)$ is bounded in any closed interval in $(0,\infty)$ it follows that $f(R)$ and $h(R)$ must intersect at a point in $(p^*,1)$ if $p^*<1$ or in the interval $(1,p^*)$ if $p^*>1$. Since $f(R)$ is convex in such intervals and negative outside them, uniqueness of the intersection is assured. In Figure 1 we present a sketch with the two cases.

3.2 Local Stability Analysis

The Jacobian of the system (4) is given by

$$J = \begin{pmatrix} J_{11} & -\lambda R - gR + (1-\alpha)r & -a\sqrt{R} & -\theta(1-p)c\sqrt{R} - cp\sqrt{R} \\ J_{21} & J_{22} & -bI & \theta(1-p)c\sqrt{R} - \ell I \\ J_{31} & J_{32} & J_{33} & (1-\gamma)(\ell I + cp\sqrt{R}) - \beta F \\ \frac{c\gamma pW}{2\sqrt{R}} & \ell\gamma W + b\sigma F & \beta W + b\sigma I & cp\gamma\sqrt{R} + \gamma\ell I + \beta F - \nu \end{pmatrix} \tag{12}$$

with

$$J_{11} = -\lambda I - \frac{\theta(1-p)cW}{2\sqrt{R}} - \frac{cpW}{2\sqrt{R}} - 2R - \frac{aF}{2\sqrt{R}} - gI + 1, \quad J_{21} = \lambda I + \frac{\theta(1-p)cW}{2\sqrt{R}} - uI,$$

$$J_{22} = \lambda R - \ell W - 2qI - uR - bF + \alpha - \mu, \quad J_{31} = \frac{c(1-\gamma)pW}{2\sqrt{R}} + \frac{aF}{2\sqrt{R}},$$

$$J_{32} = (1-\gamma)\ell W + b(1-\sigma)F, \quad J_{33} = -\beta W + a\sqrt{R} + b(1-\sigma)I - m,$$

3.2.1 Stability Analysis of E_0

The origin for this model presents a particular behavior. Although unstable in the Lyapunov sense [19], it is still capable of attracting trajectories over a set of initial condition with positive measure in $\mathbb{R}^4$. The instability of the origin can be seen, by observing that any trajectory starting in the line defined by $I = F = W = 0$ remains in it, since $\dot{I} = \dot{F} = \dot{W} = 0$. The equation for $\dot{R}$ on the line is simply $\dot{R} = R(1-R)$ which implies that the origin is unstable, since any trajectory with initial condition $0 < \epsilon = R_0 < 1$ moves away from the origin.

This particular behavior of the origin is caused by the predation term which is proportional to the square root of the healthy prey population. When $R \to 0$, such term has a higher order when compared to the reproduction term (R). In fact such proportionality to the square root of R is adequate for "large" population, since the group defense effect is negligible when the population is "small". Also, in the present model we seek to analyze the biological situations which are relevant to the analysis to the spread of the disease, that is, the scenarios where at least one of the populations is present.

However, a particular phenomenon has been observed in similar models in these conditions, [26, 11, 12, 3]. The right hand side of the system is not Lipschitz-continuous because of the presence of the square root in every component, so that the uniqueness theorem does not hold. In [26] this has been investigated, showing that trajectories lying in a narrow stripe may well end up on the prey axis in finite time, and from there they move toward the origin, with ecosytem collapse. This has been further investigated in [11, 12, 3], showing that it entails a wealth of bifurcation phenomena. For a generalization to an arbitrary power instead of the square root, see [5].

 ISBN: 978-84-617-8694-7

At E_1 the eigenvalues are easily found and given by $\lambda_1 = cp\gamma - \nu$, $\lambda_2 = a - m$, $\lambda_3 = -1$ and $\lambda_4 = \lambda + \alpha - \mu - u$. E_1 is stable if and only if

$$cp\gamma < \nu, \quad a < m, \quad \lambda + \alpha < \mu + u. \tag{13}$$

For E_2 the product of two quadratic equations is obtained. The first one has the Routh-Hurwitz conditions $\mathrm{tr}(\overline{J}^1_{E_2}) = (a^2 - 3m^2)(2a)^{-2} < 0$, $\det(\overline{J}^1_{E_2}) = (ma^2 - m^3)(2a)^{-2} > 0$, the stability conditions

$$\frac{a}{\sqrt{3}} < m < a. \tag{14}$$

The second quadratic has more complicated Routh-Hurwitz conditions that provide the second set of stability conditions

$$a^4\nu + a^2m^2u + a^4\mu + \beta m^3 + bma^2 > am^2\lambda + cmp\gamma a^3 + bm^3 + \beta ma^2 + \alpha a^4 \tag{15}$$

and

$$\begin{aligned} a^5c\gamma m^3p\lambda + a^4\beta m^3\lambda + a^6m^2u\nu + a^8\mu\nu + a^6bm\nu + a^2\beta m^5u + a^5bcpm^2\theta\sigma + a^3bcm^4\theta\sigma \\ +a^3bcpm^4\gamma + a^7cmp\alpha\gamma + a^4\beta m^3\mu + a^2 2b\beta m^4 + a^6\beta m\alpha > a^6m^2\nu\lambda + a^2\beta m^5\lambda \\ +a^4bm^3\nu + a^8\alpha\nu + a^5cpu\gamma m^3 + a^3bcpm^4\theta\sigma + a^4\beta m^3u + a^5bcm^2\theta\sigma + a^7cmp\mu\gamma \\ +a^5bcpm^2\gamma + a^6\beta m\mu + \beta bm^6 + a^4\beta m^3\alpha + a^4\beta bm^2. \end{aligned} \tag{16}$$

At E_3 again the characteristic equation factorizes into the product of two quadratic equations, that have the Routh-Hurwitz conditions

$$\begin{aligned} \mathrm{tr}(\overline{J}^1_{E_3}) &= (\lambda R_3 - \lambda I_3 - 2qI_3 - gI_3 - uR_3 - 2R_3 + \alpha + 1 - \mu < 0, \\ \det(\overline{J}^1_{E_3}) &= -2\lambda {R_3}^2 + \lambda R_3 + 2q\lambda {I_3}^2 + 2gq{I_3}^2 - 2r\lambda I_3 - g\mu I_3 - 2u{R_3}^2 + 4qR_3I_3 - 2qI_3 \\ &\quad -rukR_3 - r\alpha uI_3 - ruI_3 - 2\mu R_3 - r\lambda I_3 + \mu\lambda I_3 - \alpha\lambda I_3 - g\alpha I_3 - 2\alpha r - \mu + \alpha > 0, \end{aligned}$$

from which the stability conditions follow

$$\begin{aligned} \lambda I_3 + 2qI_3 + gI_3 + uR_3 + 2R_3 + \mu > \lambda R_3 + \alpha + 1, \\ 2\lambda {R_3}^2 + \lambda R_3 + 2q\lambda {I_3}^2 + 2gq{I_3}^2 + 4qR_3I_3 + \mu\lambda I_3 + \alpha \\ > 2r\lambda I_3 + g\mu I_3 + 2u{R_3}^2 + 2qI_3 + rukR_3 + r\alpha uI_3 \\ +ruI_3 + 2\mu R_3 + r\lambda I_3 + \alpha\lambda I_3 + g\alpha I_3 + 2\alpha r + \mu. \end{aligned} \tag{17}$$

The second quadratic instead gives the Routh-Hurwitz conditions

$$\begin{aligned} \mathrm{tr}(\overline{J}^2_{E_3}) &= cp\gamma\sqrt{R_3} + a\sqrt{R_3} - b\sigma I_3 + \gamma\ell I_3 + bI_3 - \nu - m < 0, \\ \det(\overline{J}^2_{E_3}) &= acp\gamma R_3 - bcp\sigma\sqrt{R_3}I_3 + bc\gamma p\sqrt{R_3}I_3 + a\gamma\ell\sqrt{R_3}I_3 - a\nu\sqrt{R_3} \\ &\quad -cmp\gamma\sqrt{R_3} - b\sigma\ell {I_3}^2 + b\ell\gamma {I_3}^2 + b\sigma\nu I_3 - b\nu I_3 - m\ell\gamma I_3 + m\nu > 0, \end{aligned}$$

 ISBN: 978-84-617-8694-7

once again providing the stability conditions

$$\begin{aligned} b\sigma I_3+\nu+m > cp\gamma\sqrt{R_3}+a\sqrt{R_3}+\gamma\ell I_3+bI_3 \\ acp\gamma R_3+bc\gamma p\sqrt{R_3}I+a\gamma\ell\sqrt{R_3}I_3+b\ell\gamma {I_3}^2+b\sigma\nu I_3+m\nu \\ > bcp\sigma\sqrt{R_3}+a\nu\sqrt{R_3}I_3+cmp\gamma\sqrt{R_3}+b\sigma\ell {I_3}^2+b\nu I_3+m\ell\gamma I_3. \end{aligned} \tag{18}$$

4 The basic reproduction number

Conditions for the eradication of the disease can be obtained from the basic reproduction number $\mathcal{R}_0$, the spectral radius of the next generation matrix at each disease-free equilibrium [24]. Let F_I, F_W be the corresponding new infectious rates and V_I, V_W the analogous flows, the dynamics of the infectious classes I and W can be written as:

$$\frac{dI}{dt}=F_I-V_I=F_I-(V_I^- - V_I^+), \quad \frac{dW}{dt}=F_W-V_W=(V_W^- - V_W^+), \tag{19}$$

where

$$\begin{aligned} F_I=\bar{r}\alpha I+\lambda RI+\theta(1-p)c\sqrt{R}W, \quad F_W=be\sigma IF+\beta FW+cep\gamma\sqrt{R}W+el\gamma IW, \\ {V_I}^-=bIF+\mu I+lIW+ukRI+qkI^2, \quad {V_I}^+=0, \quad {V_W}^-=vW, \quad {V_W}^+=0. \end{aligned}$$

Letting

$$F=\begin{pmatrix} \frac{\partial F_I}{\partial I} & \frac{\partial F_W}{\partial I} \\ \frac{\partial F_I}{\partial W} & \frac{\partial F_W}{\partial W} \end{pmatrix}=\begin{pmatrix} \alpha+\lambda R & \gamma\ell W+b\sigma F \\ \theta(1-p)c\sqrt{R} & pc\gamma\sqrt{R}+\ell\gamma I+\beta F \end{pmatrix}$$

$$V=\begin{pmatrix} \frac{\partial V_I}{\partial I} & \frac{\partial V_W}{\partial I} \\ \frac{\partial V_I}{\partial W} & \frac{\partial V_W}{\partial W} \end{pmatrix}=\begin{pmatrix} bF+\mu+\ell W+uR+2qI & 0 \\ \ell I & \nu \end{pmatrix}.$$

the next generation matrix is

$$G=FV^{-1}=\begin{pmatrix} \frac{\alpha\nu+\lambda\nu R-\ell\gamma IW-b\ell\sigma IF}{\nu(\ell W+uR+2qI+bF+\mu)} & \frac{\ell\gamma W+b\sigma F}{\nu} \\ \frac{\theta(1-p)c\nu\sqrt{R}-pc\ell\gamma I\sqrt{R}-\ell^2\gamma I^2-\beta\ell IF}{\nu(\ell W+uR+2qI+bF+\mu)} & \frac{\beta F+\ell\gamma I+cp\gamma\sqrt{R}}{\nu} \end{pmatrix}.$$

$\mathcal{R}_0$ is defined in each disease-free equilibrium as the spectral radius of G. For model (4) the only feasible disease-free equilibria are E_0, E_1 and E_2. Therefore, we proceed the analysis of the disease-free equilibria.

 ISBN: 978-84-617-8694-7

4.1 Stability analysis of disease-free equilibria

The methodology requires five conditions to be applied [24], one of them is that disease-free equilbrium should be stable if the number of new cases are set to zero. Considering $F_I = 0$ and $F_W = 0$ in model (4) we obtain:

$$\begin{aligned}
\frac{dR}{dt} = \quad & rR + \bar{r}\,(1-\alpha)\,I - kR\,(R+gI) - \lambda RI - cp\sqrt{R}W \\
& -a\sqrt{R}F - \theta\,(1-p)\,c\sqrt{R}W, \\
\frac{dI}{dt} = \quad & -bIF - \mu I - lIW - ukRI - qkI^2, \\
\frac{dF}{dt} = \quad & ae\sqrt{R}F + (1-\sigma)beIF + (1-\gamma)W(cep\sqrt{R} + \ell eI) \\
& -mF - \beta FW, \\
\frac{dW}{dt} = \quad & -vW.
\end{aligned} \tag{20}$$

The Jacobian of the system (20) is given by

$$J = \begin{pmatrix} J_{11} & -\lambda R - gR + (1-\alpha)r & -a\sqrt{R} & J_{14} \\ -uI & J_{22} & -bI & -\ell I \\ \frac{cp(1-\gamma)W + aF}{2\sqrt{R}} & J_{32} & J_{33} & J_{34} \\ 0 & 0 & 0 & -v \end{pmatrix} \tag{21}$$

with

$$J_{11} = -\lambda I - \frac{\theta(1-p)cW}{2\sqrt{R}} - \frac{cpW}{2\sqrt{R}} - gI - 2R - \frac{aF}{2\sqrt{R}} + 1,$$

$$J_{14} = -\theta(1-p)c\sqrt{R} - cp\sqrt{R}, \quad J_{22} = -bF - \mu - \ell W - uR - 2qI,$$

$$J_{32} = (1-\gamma)\ell W + b(1-\sigma)F,$$

$$J_{33} = a\sqrt{R} + (1-\sigma)bI - m - \beta W, \quad J_{34} = (1-\gamma)(\ell I + cp\sqrt{R}) - \beta F,$$

4.1.1 Disease-free equilibrium E_1

For this point the stability condition under $F_I = F_W = 0$ is

$$m > a. \tag{22}$$

The eigenvalues of G in E_1 are:

$$\lambda_1^{E_1} = \frac{\alpha + \lambda}{u + \mu}, \quad \lambda_2^{E_1} = \frac{cp\gamma}{\nu}. \tag{23}$$

 ISBN: 978-84-617-8694-7

Since $\lambda_1^{E_1}$ and $\lambda_2^{E_1}$ are both positive, the value of $\mathcal{R}_0$ in E_1 is simply $\mathcal{R}_0^{E_1} = \max\{\lambda_1^{E_1}, \lambda_2^{E_1}\}$. Disease-induced instability occurs if $\mathcal{R}_0 > 1$.

If $\nu \geq m$, that is, the rate of mortality of infected predators is greater than non-infected ones, then we can write:

$$\nu \geq m > a \geq c \geq cp\gamma,$$

because we consider $a \geq c$ (healthy predators are more efficient in hunting than infected ones) and $p, \gamma \leq 1$. Therefore, $\lambda_2^{E_1} < 1$ if E_1 is stable in the absence of disease.

The condition for $\mathcal{R}_0 < 1$ coming from $\lambda_1^{E_1}$ can be written as

$$\alpha + \lambda < u + \mu.$$

The left side of the inequality represents rates that are favorable to the permanence of the disease, i.e, reproduction rate of infectious prey and generation of new infectious cases at the equilibrium. In the right side of inequality are the factors that contribute to the eradication of the disease, i.e, the mortality rates of infected prey and mortality due to competition with healthy prey at the equilibrium. In this case stability in the absence of the disease does not imply $\lambda_1^{E_1} < 1$. For instance, even if $\alpha = 0$ it is sufficient to take $\lambda > u + \mu$ to obtain $\mathcal{R}_0 > 1$.

4.1.2 Disease-free equilibrium E_2

Considering $F_I = F_W = 0$, the viability condition for the equilibrium E_2 is given by

$$m \leq a \tag{24}$$

and the stability conditions for it are

$$\frac{a}{\sqrt{3}} < m < a. \tag{25}$$

The matrix G in E_2 is

$$G(E_2) = \begin{pmatrix} \dfrac{\alpha a^4 + \lambda m^2 a^2}{um^2a^2 + bma^2 - bm^3 + \mu a^4} & \dfrac{\sigma ba^2 m - \sigma bm^3}{\nu a^4} \\ \dfrac{(-cpma^3 + cma^3)\theta}{a^2m^2u + \mu a^4 - bm^3 + bma^2} & \dfrac{cpm\gamma a^3 - \beta m^3 + \beta ma^2}{\nu a^4} \end{pmatrix} \tag{26}$$

The formulas for eigenvalues of $G(E_2)$ do not provide any immediate insight on the behavior of $\mathcal{R}_0$, but through numerical simulations we can state that both $\lambda_1^{E_2}$ and $\lambda_2^{E_2}$ can have absolute values greater than one.

 ISBN: 978-84-617-8694-7

5 Bifurcations

In this section we will use Sotomayor theorem [19, 18].

5.1 Transcritical bifurcation

Note that the general second order term of the Taylor expansion of f, recall (4), is given by

$$D^2 f(X,m)(U,U) = (D_{11}, D_{21}, D_{31}, D_{41})^T, \tag{27}$$

taking m as any bifurcation parameter and $U = (\xi_1, \xi_2, \xi_3, \xi_4)^T$ being the vector of variations in R, I, F and W, with

$$\begin{aligned} D_{11} &= \frac{\theta(1-p)cW{\xi_1}^2 + cpW{\xi_1}^2 + aF{\xi_1}^2}{4R\sqrt{R}} - 2{\xi_1}^2 - 2g\xi_1\xi_2 - 2\lambda\xi_1\xi_2 \\ &- \frac{a}{\sqrt{R}}\xi_1\xi_3 - \frac{cp}{\sqrt{R}}\xi_1\xi_4 - \frac{\theta(1-p)c}{\sqrt{R}}\xi_1\xi_4, \end{aligned}$$

$$D_{21} = -\frac{\theta(1-p)cW}{4R\sqrt{R}}{\xi_1}^2 + 2(\lambda-u)\xi_1\xi_2 - 2b\xi_2\xi_3 - 2\ell\xi_2\xi_4 - 2q{\xi_2}^2 + \frac{\theta(1-p)c}{\sqrt{R}}\xi_1\xi_4,$$

$$\begin{aligned} D_{31} &= -\frac{c(1-\gamma)pW{\xi_1}^2 - aF{\xi_1}^2}{4R\sqrt{R}} + \frac{a}{\sqrt{R}}\xi_1\xi_3 + \frac{c(1-\gamma)p}{\sqrt{R}}\xi_1 l\xi_4 \\ &+ 2b(1-\sigma)\xi_2\xi_3 + 2(1-\gamma)\ell\xi_2\xi_4 - 2\beta\xi_3\xi_4, \end{aligned}$$

$$D_{41} = -\frac{cpW}{4R\sqrt{R}}{\xi_1}^2 + \frac{cp\gamma}{\sqrt{R}}\xi_1\xi_4 + 2b\sigma\xi_2\xi_3 + 2\gamma\ell\xi_2\xi_4 + 2\beta\xi_3\xi_4.$$

5.1.1 Bifurcation between E_1 and E_2

Comparing the second inequality in (13) given by $m > a$ for the equilibrium point E_1 and, the inequality in (7) and the right inequality in (14) for equilibrium point E_2, we find that E_2 becomes feasible and stable exactly when E_1 becomes unstable. There is thus a transcritical bifurcation for which E_2 emanates from E_1 when the bifurcation parameter m crosses the critical value $m^\dagger$ given by

$$m^\dagger = a. \tag{28}$$

The simulation presented in Figure 2 (a) shows it explicitly for the chosen parameter values $m^\dagger = a = 1$. We prove that there is a transcritical bifurcation between points E_1 and E_2, according to the following proposition

 ISBN: 978-84-617-8694-7

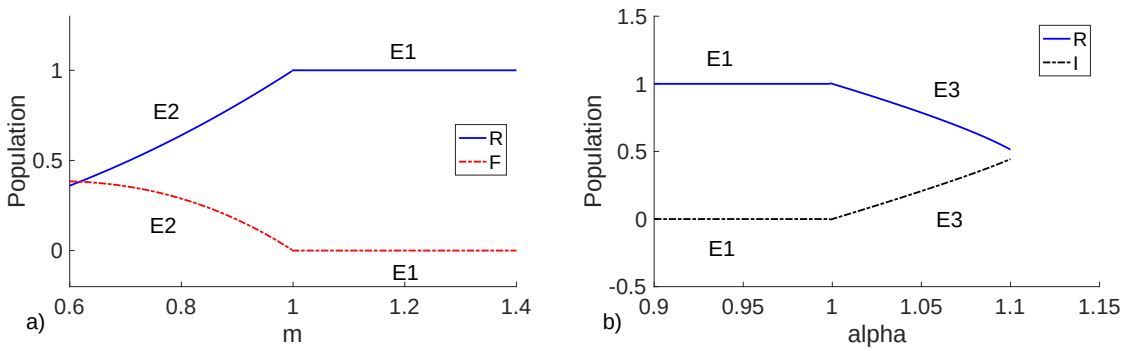


Figure 2: a) Transcritical bifurcation between E_1 and E_2 for the parameter values $\lambda = \beta = \ell = \alpha = \sigma = \gamma = \theta = q = c = p = r = g = b = 0.5$, $a = 1$, $u = \mu = 0.75$, $\nu = 2m$. Initial conditions $R_0 = I_0 = F_0 = W_0 = 0.1$. b) Transcritical bifurcation between E_2 and E_3. In this case, we have same parameter value and initial conditions, except for $m = 0.5$.

Proposition 1 Assuming that $\alpha + \lambda < u + \mu$ and $cp\gamma < \nu$, when m passes through the value $m^\dagger = a$, model (4) near the disease-free equilibrium $E_1 = (1, 0, 0, 0)$ has:

- no saddle-node bifurcation;
- a transcritical bifurcation;
- no pitchfork bifurcation.

Proof Since $\alpha + \lambda < u + \mu$, $cp\gamma < \nu$ and $m > a$, the equilibrium point E_1 is stable. The Jacobian matrix of model (4) evaluated at E_1 with $m^\dagger = a$, is given by

$$J_{E_1}(m^\dagger) = \begin{pmatrix} -1 & -\lambda - g + (1-\alpha)r & -a & -\theta(1-p)c - cp \\ 0 & \lambda - u - \mu + \alpha & 0 & \theta(1-p)c \\ 0 & 0 & 0 & c(1-\gamma)p \\ 0 & 0 & 0 & cp\gamma - \nu \end{pmatrix}. \tag{29}$$

In this case, we have one eigenvalue equal zero in (29), in which the corresponding eigenvector is $V_1 = \varphi_1(1, 0, -\frac{1}{a}, 0)^T$, where φ_1 is any nonzero real number.

Similarly, $Z_1 = \omega_1(0, 0, 1, -p(c\gamma - c)(\nu - c\gamma p)^{-1})^T$ represents the eigenvector corresponding to eigenvalue equal zero of $(J_{E_1}(m^\dagger))^T$, where ω_1 is any nonzero real number. Differentiating partially the right hand sides of the equations of system (4) with respect to $m^\dagger = a$, we find

$$\frac{df}{dm} = f_m(E_1, a) = (0, 0, 0, 0)^T, \tag{30}$$

 ISBN: 978-84-617-8694-7

which gives $Z_1^T f_m(E_1, a) = 0$. Thus, according to Sotomayor's theorem for local bifurcation, model (4) has no saddle-node bifurcation near disease-free equilibrium at $m^\dagger = a$. Besides that,

$$Df_m(E_1, a) = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \tag{31}$$

then, $Z_1^T[Df_m(E_1, a).V_1] = \varphi_1\omega_1/a \neq 0$. Now, considering E_1, $m^\dagger = a$ and V_1 in (27) we get

$$D^2 f(E_1, a).(V_1, V_1) = \varphi_1^2 (-1, 0, -1, 0)^T . \tag{32}$$

Therefore,

$$Z_1^T[D^2 f(E_1, a).(V_1, V_1)] = -\omega_1\varphi_1^2 \neq 0. \tag{33}$$

According to Sotomayor's theorem model (4) has a transcritical bifurcation at E_1 with parameter $m^\dagger = a$, while the pitchfork bifurcation cannot occur.

5.1.2 Bifurcation between E_1 and E_3

Comparing the third inequality in (13) for the equilibrium point E_1 and, the inequality in (11) for equilibrium point E_3, we find that E_3 becomes feasible exactly when E_1 becomes unstable. There is thus a transcritical bifurcation for which E_3 emanates from E_1 when the bifurcation parameter α crosses the critical value $\alpha^\dagger$ given by

$$\alpha^\dagger = u + \mu - \lambda \tag{34}$$

The simulation presented in Figure 2 (b) shows it explicitly for the chosen parameter values $\alpha^\dagger = u + \mu - \lambda = 1$. We prove that there is a transcritical bifurcation between points E_1 and E_3, according to the following proposition

Proposition 2 Assuming that $\alpha + \lambda < u + \mu$ and $cp\gamma < \nu$, when α passes through the value $\alpha^\dagger = u + \mu - \lambda$, model (4) near the disease-free equilibrium E_1 has:

- no saddle-node bifurcation;
- a transcritical bifurcation;
- no pitchfork bifurcation.

Proof

Since $\alpha + \lambda < u + \mu$, $cp\gamma < \nu$ and $m > a$, the equilibrium point E_1 is stable. The Jacobian matrix of model (4) evaluated at E_1, with $\alpha^\dagger = u + \mu - \lambda$ is

 ISBN: 978-84-617-8694-7

$$J_{E_1}(\alpha^\dagger) = \begin{pmatrix} -1 & \lambda(r-1)+r(1-u-\mu)-g & -a & -\theta(1-p)c-cp \\ 0 & 0 & 0 & \theta(1-p)c \\ 0 & 0 & a-m & c(1-\gamma)p \\ 0 & 0 & 0 & cp\gamma-\nu \end{pmatrix}. \tag{35}$$

in which has one eigenvalue equal zero and the corresponding eigenvector is $V_2 = \varphi_2(1, (\lambda - ru+r-r\mu-g)^{-1}, 0, 0)^T$. For $J_{E_1}(\alpha^\dagger))^T$, the eingenvector is $Z_2 = \omega_2(0, 1, 0, ((cp-c)\theta)(c\gamma p-\nu)^{-1})^T$. We have φ_2 and ω_2 are any nonzero real number.

Differentiating partially the right hand sides of the equations of system (4) with respect to $\alpha^\dagger = u + \mu - \lambda$, we find

$$\frac{df}{d\alpha} = f_\alpha(E_1, \alpha^\dagger) = (0, 0, 0, 0)^T, \tag{36}$$

which gives $Z_2^T.f_\alpha(E_1, \alpha^\dagger) = 0$. Thus, according to Sotomayor's theorem for local bifurcation, model (4) has no saddle-node bifurcation near disease-free equilibrium at $\alpha^\dagger = u+\mu-\lambda$.

Moreover,

$$Df_\alpha(E_1, \alpha^\dagger) = \begin{pmatrix} 0 & -r & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0, \end{pmatrix} \tag{37}$$

then, $Z_2^T[Df_m(E_1, \alpha^\dagger)V_2] = \varphi_2\omega_2(\lambda - ru + r - r\mu - g)^{-1} \neq 0$. Now, considering E_1, $\alpha^\dagger$ and V_2 in (27) we get with $\Lambda = \lambda - ru + r - r\mu - g$

$$D^2 f(E_1, \alpha^\dagger).(V_2, V_2) = \varphi_2^2 \left(\frac{-4\lambda - 2r + 2ru - 2r\mu}{\Lambda}, \frac{2(\lambda - u)}{\Lambda} - \frac{2q}{\Lambda^2}, 0, 0 \right)^T. \tag{38}$$

Therefore,

$$Z_2^T[D^2 f(E_1, \alpha^\dagger).(V_2, V_2)] = \frac{2\omega_2\varphi_2^2(\lambda - u)(\lambda - ru + r - r\mu - g)}{(\lambda - ru + r - r\mu - g)^2} \neq 0. \tag{39}$$

So, according to Sotomayor's theorem model (4) has a transcritical bifurcation at E_1 with parameter $\alpha^\dagger = u + \mu - \lambda$, while the pitchfork bifurcation cannot occur.

5.2 Hopf bifurcations

We now try to establish whether there are parameter combinations giving sustained population oscillations. For E_1 it is not the case, since the eigenvalues are all real.

At E_2, see Figure 3, we have

 ISBN: 978-84-617-8694-7

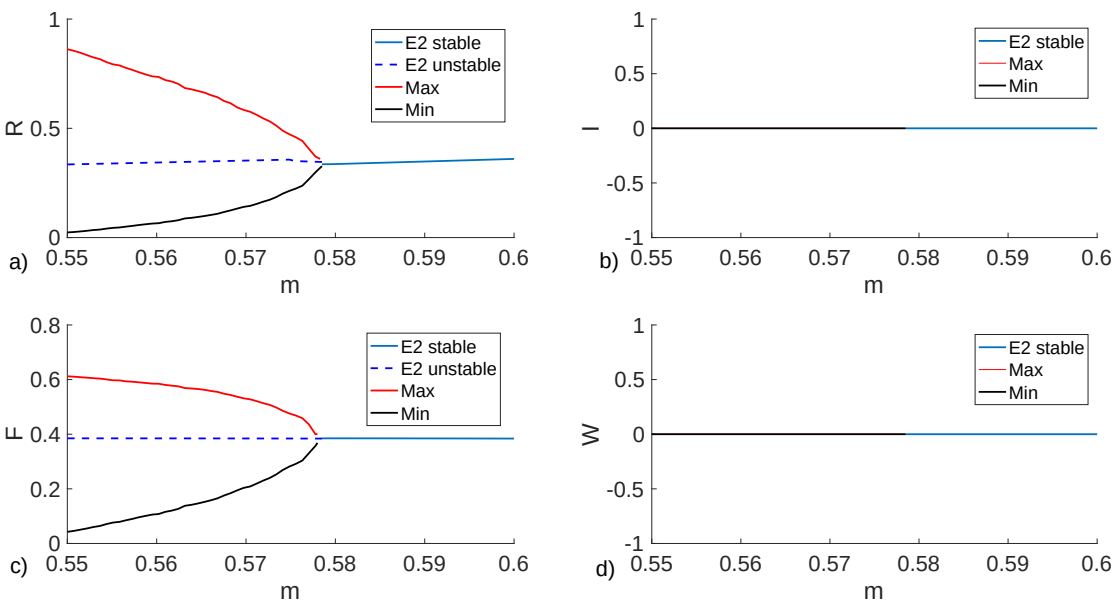


Figure 3: Hopf bifurcation for point E_2 for parameter values $\lambda = \beta = \ell = \alpha = \sigma = \gamma = \theta = q = c = p = r = g = b = 0.5$, $a = 1$, $u = \mu = 0.75$, $\nu = 1.157$. Initial conditions $R_0 = I_0 = F_0 = W_0 = 0.1$.

 ISBN: 978-84-617-8694-7

Proposition 3 Assuming that conditions (15) and (16) hold, then model (4) undergoes Hopf bifurcation around the equilibrium point E_2 when parameter m crosses the critical positive value $m^* = a/\sqrt{3}$.

Proof For systems in four-dimensional spaces, for a Hopf bifurcation to occur, the following conditions should be satisfied [10, 27, 18]:

- The characteristic equation at E_2 has two real and negative eigenvalues and two complex eigenvalues;
- $\tau_1(m^*) = 0$;
- $(\frac{d}{dm}\tau_1(m))|_{m=m^*}$ (The transversality condition).

The stability analysis of E_2 showed that we obtain two real and negative eigenvalues and another two given by:

$$\Lambda_\pm = \tau_1 \pm \sqrt{P(m)}, \quad \tau_1 = 3m^2 - a^2, \quad P(m) = 9m^4 + 8a^2m^3 - 6a^2m^2 - 8a^4m + a^4.$$

Thus, since $P(m)$ is continuous, and $P(m^*) = -16a^5(3\sqrt{3})^{-1} < 0$, there is an interval $T = (m^* - \epsilon, m^* + \epsilon)$ around m^* , such that, $P(x) < 0$ whenever $x \in T$ and further $\tau_1(m^*) = 0$. Finally, the transversality condition is satisfied because

$$\frac{d}{dm}\tau_1(m^*) = \frac{3}{2a\sqrt{3}} \neq 0.$$

At E_4 we can only perform numerical simulations. We present two different bifurcation scenarios. In the first, when parameter m crosses from above the critical value $m_1^\star \approx 0.71462$ there is a transcritical bifurcation between E_4 losing feasibility and E_2 becoming stable. Then when $m_2^\star \approx 0.578$ we find a Hopf bifurcation in E_2. With a further decrease of m a four-dimensional limit cycle arises when $m_3^\star \approx 0.56884$, Figure 4. The second situation is simply a Hopf bifurcation in E_4 when m crosses from above the critical value of $m_4^\star \approx 0.4626875$, Figure 5.

6 Numerical results

The analysis of model (4), shows that the parameter m is crucial for the feasibility and stability of points E_1 and E_2, respectively. The establishment of the disease is directly related to the value of $\mathcal{R}_0$ at both points. Thus, we conduct an exploration of the parameter space with relation to those fundamental quantities. As it will be shown in this section, the results of the majority of the simulations can be predicted simply by the analysis of m, $\mathcal{R}_0^{E_1}$ and $\mathcal{R}_0^{E_2}$.

 ISBN: 978-84-617-8694-7

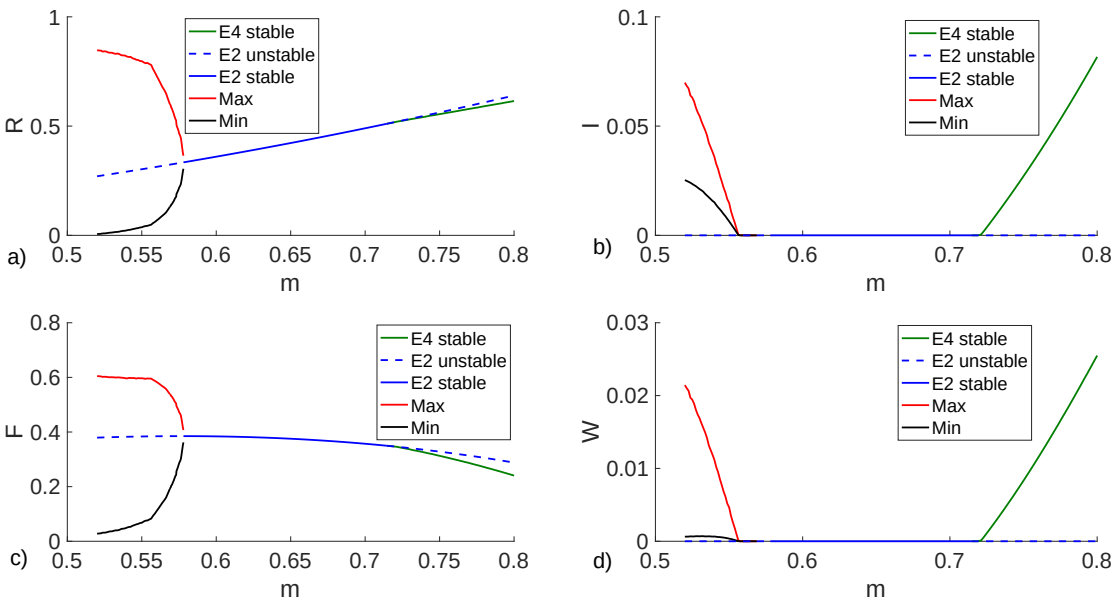


Figure 4: a), b), c) and d) illustrate a transcritical bifurcation between E_4 and E_2 for $m_1^\star \approx 0.71462$; Hopf bifurcation for E_2 when $m_2^\star \approx 0.578$; Loss of stability of the two-dimensional limit cycle and creation of a four-dimensional limit cycle when $m_3^\star \approx 0.56884$. The parameter values are: $\lambda = \sigma = \theta = r = q = g = \mu = 0.5$, $a = 1$, $\alpha = 0.6$, $c = 0.8289$, $\beta = 0.2056$, $\gamma = \ell = 0.99$, $b = 0.5066$, $\nu = 0.8$, $p = 0.7389$ and $u = 0.4$. Initial conditions $R_0 = I_0 = F_0 = W_0 = 0.1$.

 ISBN: 978-84-617-8694-7

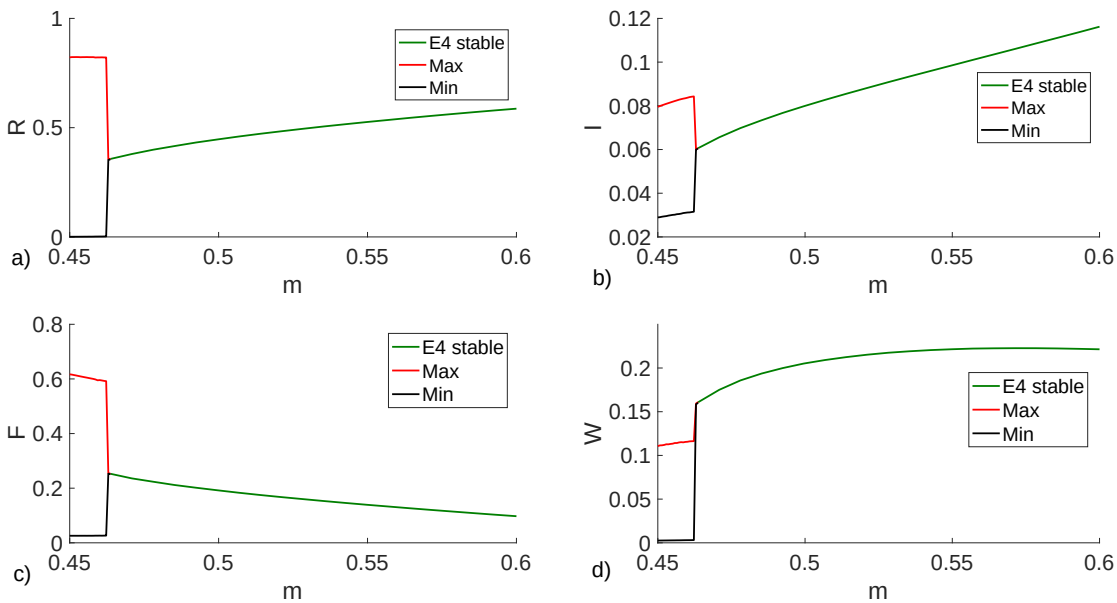


Figure 5: Hopf bifurcation for point E_4 for parameter values $\lambda = \sigma = \theta = r = q = g = \mu = 0.5$, $a = 1$, $\alpha = 0.61$, $c = 0.8289$, $\beta = 0.9433$, $\gamma = \ell = 0.99$, $b = 0.5066$, $\nu = 0.6833$, $p = 0.7389$ and $u = 0.4$.

 ISBN: 978-84-617-8694-7

6.1 Details about the numerical implementation

To investigate the behavior of model (4) in relation to the parameters, we first separate the simulations in two specific cases. In the first case, E_1 is stable and E_2 is unstable, that is, $m > a$. The second case, is obtained when $a/\sqrt{3} < m < a$, in which E_1 is unstable and E_2 is feasible. The reason for the adoption of such thresholds comes from the analysis of the analogous predator-prey model without the presence of disease [5]. In our model a transcritical bifurcation E_1 and E_2 is observed when $m = a$ and Hopf bifurcation occurs when $m = a/\sqrt{3}$ as we can see in subsection 5.

Given the high number of parameters (18), we opt for a random exploration of the space. In Table 2 represent the distributions adopted for each parameter.

Table 2: Distribution for the parameters in simulations using Matlab. Here, we make a random exploration of the space to each parameter. $\mathcal{U}(x, y)$ stands for an uniform distribution between x and y.

Parameters	Distribution
$u, \quad q, \quad g$	50% $\mathcal{U}(0.5, 1)$; 50% $\mathcal{U}(1, 2)$
$\mu, \quad \beta, \quad \lambda$	50% $\mathcal{U}(0.1, 1)$; 50% $\mathcal{U}(1, 10)$
$a, \quad b$	50% $\mathcal{U}(0.1, 1)$; 50% $\mathcal{U}(1, 2)$
c	$\mathcal{U}(0.1, a)$
$m > a$ (case 1)	$\mathcal{U}(1.1a, (2 - 1.1/\sqrt{3})a)$
$a/\sqrt{3} < m < a$ (case 2)	$\mathcal{U}(1.1a/\sqrt{3}, 0.9a)$
ν	$\mathcal{U}(m, 3m)$
ℓ	$\mathcal{U}(0.1, b)$
$\alpha, \quad \sigma, \quad \gamma, \quad \theta, \quad p$	$\mathcal{U}(0.05, 0.95)$
r	$\mathcal{U}(0.05, 1)$

When non-dimensional parameters cross the threshold 1, it usually means a transition between two qualitatively distinct scenarios. For this reason, the random sampling is chosen to be half of the time in each situation. For biological reasons, some parameters are linked. For instance, the mortality of diseased predators (ν) is greater than or equal to the mortality of healthy predators (m).

For the numerical simulation of the system of differential equations we use the Matlab ode45 routine. For each random combination of parameters, a random initial condition was chosen with the distributions: $R_0 \sim \mathcal{U}(0.2, 1)$, $I_0 \sim \mathcal{U}(0.2, 0.6)$, $F_0 \sim \mathcal{U}(0.05, 1.05)$ and $W_0 \sim \mathcal{U}(0.05, 0.1)$ ($\mathcal{U}(x, y)$ stands for an uniform distribution between x and y). The choice of initial conditions with smaller predator populations is made in order to avoid trajectories that converge to the origin, in which the approximation of the predation term by $\sqrt{R}$ is not valid. Given the initial conditions, the system is simulated in the time interval

 ISBN: 978-84-617-8694-7

$I_{t_1} = [0, 200]$, if such interval is not enough to find an equilibrium, another try is attempted with $I_{t_2} = [0, 2000]$.

Also, for each combination of parameters the equilibrium points E_1, E_2, E_3 and E_4 are estimated. For E_1 and E_2 the analytical formulae of section (3) are used. For the equilibrium point E_3 the quadratic equation (6) is numerically solved using the routine ROOTS of Matlab. In sequence, the equation for I_3 (5) is employed to establish if there was any feasible equilibrium solution E_3.

For the equilibrium point E_4 we do not have analytic formulae. Thus, we used the routine FMINCON of Matlab, to minimize the sum of the squares of the derivatives subject to the conditions: $R_4 > 2^{-8}$, $I_4 > 2^{-8}$, $F_4 > 2^{-8}$, e $W_4 > 2^{-8}$. Since the results of the minimization process depend on the initial guess, 10 starting points are taken for each parameter combination. The distributions of the initial guesses are taken as: $R_g \sim \mathcal{U}(0, 1)$, $I_g \sim \mathcal{U}(0, 1)$, $F_g \sim \mathcal{U}(0, 1)$ and $W_g \sim \mathcal{U}(0, 1)$. Each time, if the routine obtained with success a solution for the minimization problem, it is stored as a candidate for equilibrium point E_4. After the 10 executions of the routine FMINCON, redundant solutions are removed from the list of equilibrium candidates. Two solutions x_1, $x_2 \in \mathbb{R}^4$ are considered redundant if

$$\frac{\|x_1 - x_2\|}{\|x_1\|} < 0.01.$$

After obtaining the list of all equilibrium candidates (from E_1 to E_4), again all redundant solutions are removed, using the same criterion. For the equilibrium points E_1 and E_2 the values of $\lambda_1^{E_1}$, $\lambda_2^{E_1}$, $\lambda_1^{E_2}$ and $\lambda_2^{E_2}$ are computed. For points E_3 and E_4 (possibly multiple) the stability (eigenvalues of the Jacobian) are computed numerically.

The remaining list of candidates is then used to be compared with the result of the numerical simulation. The relative error between all the candidate solutions and the result of the numerical simulation is calculated. If the smallest error between the simulated solution and the candidate solutions is smaller than 0.001, then the simulation is classified as a success and said to converge to the candidate solution closest to the simulated solution. In figure 6 we present a scheme on how each simulation is conducted.

As we shall show, from the results of the numerical simulations, the majority of the results of the simulations can be predicted only by analyzing the values of m (which is a critical ecological parameter) and the values related to the spread of the epidemy: $\lambda_1^{E_1}$, $\lambda_2^{E_1}$, $\lambda_1^{E_2}$ and $\lambda_2^{E_2}$. We divide the discussion according to the values of those variables.

6.2 Case 1: m > a

We ran 10,000 simulations with random parameters as in Table 2, in which $m > a$. Of this total, 9870 (98.7%) were concluded with success in the sense defined in section 6.1 and Figure 6. The results can be subdivided in two main cases, one when $\mathcal{R}_0^{E_1} < 1$ and the

 ISBN: 978-84-617-8694-7

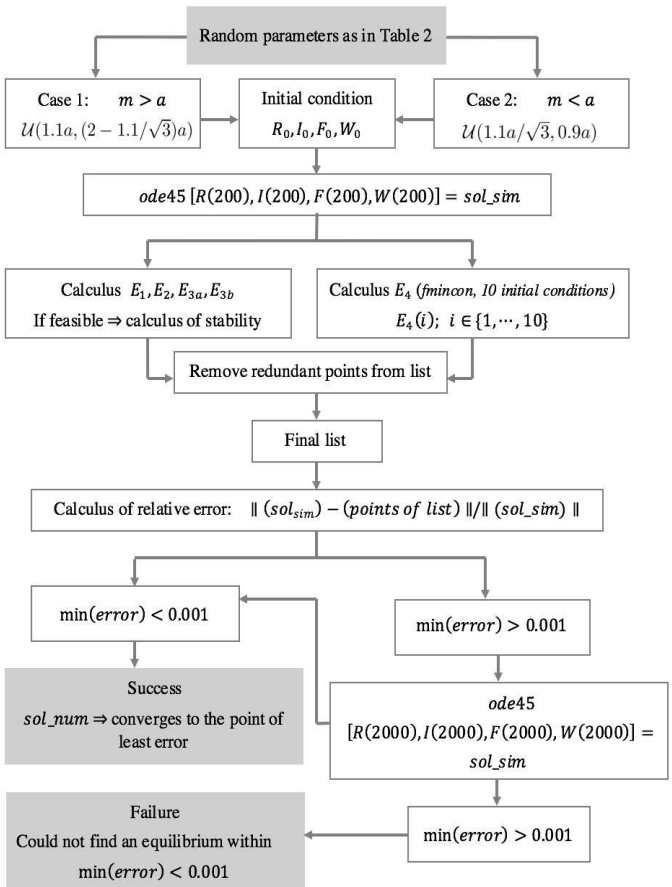


Figure 6: Scheme for a numerical simulation of the system. The equilibrium points, their stability and a simulated solution is computed. If the numerical solution converges to any of the computed candidate equilibrium points, the simulation is classified as a success.

 ISBN: 978-84-617-8694-7

other when $\mathcal{R}_0^{E_1} > 1$.

6.2.1 $\mathcal{R}_0^{E_1} < 1$:

In 6155 (63,26%) of the total 9870 successful simulations, $\mathcal{R}_0^{E_1} < 1$. Of those, in 6143 (99.81%) the system converged to E_1 while in 12 (0.19%) cases it converged to E_3. To understand these results, we analyzed also the feasibility and stability of the other equilibrium points under these conditions.

We have shown that if $m > a$, E_2 is unfeasible, therefore, besides E_1, the system could converge to E_3 or E_4. We map the behavior of those other two points in the simulations. For each simulation, we classify three possible states for the points E_3 and E_4:

- State 1: There are no feasible points of this type of equilibrium.
- State 2: There is at least one feasible point of this type of equilibrium, but it is unstable.
- State 3: There is at least a feasible and stable point of this type of equilibrium.

In table 3 we present the distribution of the results for points E_3 and E_4. It is easy to note that, for the vast majority of the simulations, the only feasible and stable point is E_1, in agreement with the result that 99.81 % of the simulations with $\mathcal{R}_0^{E_1} < 1$ converged to E_1.

Table 3: Distribution of states of points E_3 and E_4 when $m > a$ and $\mathcal{R}_0^{E_1} < 1$. The rows and columns represent the number of times the points E_3 and E_4, respectively, were in the states 1, 2 or 3. State 1: No feasible points of this type are found in the simulation. State 2: there is at least one feasible point of this type, but it is unstable. State 3: There is at least a feasible and stable point of this type.

E_3 \ E_4	1	2	3
1	5285 (85.87%)	652 (10.59%)	199 (3.23 %)
2	0	0	2 (0.03 %)
3	16 (0.26%)	1 (0.02%)	0

The results for point E_3 can be summarized as follows: in 6136 (99.69%) we have two unfeasible points, in two (0.02%) we have two feasible and unstable points and in 17 (0.28 %) we have two feasible points of which one is stable.

The behavior of E_4 can be summarized as follows. The 6155 simulations with $\mathcal{R}_0^{E_1} < 1$ are distributed as follows: 5301 (86.13 %) no feasible points are found, 628 (10.2%) one feasible and unstable point, 25 (0.41%) two feasible and unstable points, 130 (2.11%) one

 ISBN: 978-84-617-8694-7

feasible and stable point, 59 (0.96%) two feasible points with one stable and 12 (0.19%) three feasible points with one stable. Thus, in the vast majority of the simulations, E_4 is either unfeasible or unstable. We may also observe that, since E_4 is calculated numerically it can be very close to other equilibria, but not enough to be eliminated as redundant from the candidate list.

6.2.2 $\mathcal{R}_0^{E_1} > 1$:

In this particular case it is possible to show that there exists a unique feasible point E_3 in the form $(R_3, I_3, 0, 0)$ (see section 3.2).

In 3715 (37.64%) of the total 9870 successful simulations, $\mathcal{R}_0^{E_1} > 1$. Of those, in 3567 (96.02%), the system converges to E_3 and in 148 (3.98%) it converges to E_4. In this case, E_1 is unstable, E_2 is unfeasible and the equilibria for which the solution could converge are only E_3 or E_4.

In Table 4 we present the distribution of the results for points E_3 and E_4, observing that E_3 can never be unfeasible in this case. It is easy to note that, for the vast majority of the simulations, the only feasible and stable point is E_3, in agreement with the result that 96.02 % of the simulations with $\mathcal{R}_0^{E_1} > 1$ converges to E_3. The only cases in which we have convergence for E_4 is in when E_3 is unstable.

Table 4: Distribution of states of points E_3 and E_4 when $m > a$ and $\mathcal{R}_0^{E_1} > 1$.

E_3 \ E_4	1	2	3
2	0	0	148 (3.98%)
3	3050 (82.10%)	302 (8.13%)	215 (5.79%)

The results for point E_3 can be summarized as follows: in 148 (3.98%) simulations we have one feasible and unstable point and in the other 3567 (96.02%) we obtain one feasible and stable point.

The behavior of E_4 can be summarized as follows. Of the total 3715 simulations with $\mathcal{R}_0^{E_1} > 1$ we have: 3050 (82.10%) no feasible point found, 282 (7.59 %) one feasible and unstable point, 14 (0.38 %) two feasible and unstable points, 6 (0.16%) three feasible and unstable points, 306 (8.24%) one feasible and stable point, 50 (1.35%) two feasible points with one stable, 2 (0.05%) three feasible points with one stable, 3 (0.08 %) four feasible points with one stable and 2 (0.05%) two stable and feasible points.

6.2.3 Sensitivity analysis

Since $\mathcal{R}_0^{E_1}$ has a fundamental role in the determination of the behavior of the system, we can discuss its sensibility in relation to the parameters. In the first place, it is worth to

 ISBN: 978-84-617-8694-7

note that, as shown in section 4.1 in this case, $\lambda_2^{E_1} < 1$. Therefore, point E_1 can only be destabilized through $\lambda_1^{E_1}$. The explicit relation of equation (23) indicates that $\lambda_1^{E_1}$ should be sensible to parameters α and λ with a positive correlation and to parameters u and μ with a negative correlation.

For each of the 9870 successful simulations, $\lambda_1^{E_1}$ is computed. Using this collection of values, we calculate the slope of the linear regression of $\lambda_1^{E_1}$ with each of the parameters. In table 5 we present the coefficients. As expected, the strongest correlations are those of parameters α, λ, u and μ.

Table 5: Regression slopes of $\lambda_1^{E_1}$ for each parameter.

Parameter	α	σ	γ	q	u	g
Slope	0.4576	0.0642	0.1734	-0.0188	-0.6930	-0.0080
Parameter	p	a	c	θ	r	m
Slope	-0.0013	0.0899	0.0503	0.0202	0.0720	0.0643
Parameter	λ	β	ν	μ	ℓ	b
Slope	0.4156	0.0045	0.0283	-0.2584	-0.0103	-0.0153

6.3 Case 2: $\mathbf{a/\sqrt{3} < m < a}$

We ran 10,000 simulations with random parameters as in table 2, in which $m < a$. Of this total, 9001 (90.1%) are successul in the sense defined in section 6.1 and Figure 6. In this case, E_1 is always unstable, the stability of E_2 hinges on $\mathcal{R}_0^{E_2}$ and the distribution of the convergences is more complex than when $m > a$. Again, we discuss separately the two cases $\mathcal{R}_0^{E_2} < 1$ and $\mathcal{R}_0^{E_2} > 1$.

6.3.1 $\mathcal{R}_0^{E_2} < 1$:

In 4598 (51,08%) of the total 9001 successful simulations, $\mathcal{R}_0^{E_2} < 1$. Of those, in 4581 (99.63%) the system converges to E_2, in 16 (0.35%) cases it converges to E_3 and in 1 (0.02%) case it converges to E_4. To understand these results, we analyze also the feasibility and stability of the other equilibrium points under these conditions.

We have shown that if $m < a$, E_1 is unstable, therefore, besides E_2, the system could converge to E_3 or E_4. Just as in the examples above, we map the behavior of those other two points in the simulations. In table 6 we present the distribution of the results for points E_3 and E_4. It is easy to note that, for the vast majority of the simulations, the only feasible and stable point is E_2, in agreement with the result that 99.63 % of the simulations with $\mathcal{R}_0^{E_2} < 1$ converged to E_2.

The results for point E_3 can be summarized as follows. Of the total 4598 simulations with $\mathcal{R}_0^{E_2} < 1$ we obtain: 4028 (87.60%) two unfeasible points, 544 (11.83%) one feasible

 ISBN: 978-84-617-8694-7

Table 6: Distribution of states of points E_3 and E_4 when $m < a$ and $\mathcal{R}_0^{E_2} < 1$.

E_3 \ E_4	1	2	3
1	3600 (78.29%)	273 (5.94%)	155 (3.37 %)
2	406 (8.83 %)	63 (1.37 %)	78 (1.70 %)
3	2 (0.04%)	20 (0.43%)	1 (0.02%)

and unstable point, 12 (0.26 %) one feasible and stable point, 3 (0.07%) two feasible and unstable points and 11 (0.24%) two feasible points with one stable.

The behavior of E_4 can be summarized as follows. The 4598 simulations with $\mathcal{R}_0^{E_2} < 1$ are distributed in this form: 4008 (87.17 %) no feasible points are found, 346 (7.53%) one feasible and unstable point, 9 (0.2%) two feasible and unstable points, 1 (0.02%) three feasible and unstable points, 207 (4.5%) one feasible and stable point, 26 (0.57%) two feasible points with one stable and 1 (0.02%) two feasible and stable points. Thus, in the vast majority of the simulations, E_4 is either unfeasible or unstable.

6.3.2 $\mathcal{R}_0^{E_2} > 1$:

In 4403 (48.92%) of the total 9001 successful simulations, $\mathcal{R}_0^{E_2} > 1$. Of those, in 2255 (51.22%) the system converges to E_4, in 2137 (48.54%) cases it converges to E_3 and in 12 (0.25%) case it converges to E_2. In the cases where the solution converge to E_2, despite its instability, were due to the fact that $\mathcal{R}_0^{E_2}$ is close to one, so the numerical solution remains quasi-stationary close to E_2 for a long period of time, in that time the relative error between E_2 and the numerical solution is estimated and found to be smaller than 0.001. Below, we present the analysis of the stability of the other equilibrium points in those simulations.

In table 7 we present the distribution of the results for points E_3 and E_4.

Table 7: Distribution of states of points E_3 and E_4 when $m < a$ and $\mathcal{R}_0^{E_2} > 1$.

E_3 \ E_4	1	2	3
1	0	8 (0.18%)	1518 (34.48 %)
2	0	0	622 (14.13 %)
3	1483 (33.68%)	346 (7.86%)	426 (9.68%)

The results for point E_3 can be summarized as follows. Of the total 4403 simulations with $\mathcal{R}_0^{E_2} > 1$ we obtain: 1526 (34.66%) two unfeasible points, in 620 (14.08%) one feasible and unstable point, 2248 (51.06%) one feasible and stable point, 2 (0.05%) two feasible and unstable points and 7 (0.16%) two feasible points with one stable.

 ISBN: 978-84-617-8694-7

The behavior of E_4 can be summarized as follows. The 4403 simulations with $\mathcal{R}_0^{E_2} > 1$ are distributed in this form: 1483 (33.68 %) no feasible points are found, 331 (7.52%) one feasible and unstable point, 20 (0.45) two feasible and unstable points, 2 (0.05%) three feasible and unstable points, 1 (0.02%) 5 feasible and unstable points, 2032 (46.15%) one feasible and stable point, 493 (11.20%) two feasible points with one stable, 28 (0.64%) three feasible points with one stable, 4 (0.09%) four feasible points with one stable, 1 (0.02%) five feasible points with one stable, 4 (0.09%) two feasible and stable points, 3 (0.07%) three feasible points with two stable ones and 1 (0.02%) four feasible points with two stable ones.

6.3.3 Sensitivity analysis

In this case, there is not a clear behavior of convergence dependent only on the value of $\mathcal{R}_0^{E_2}$. In fact, it is possible to predict the vast majority of the results (see section 6.4) if we analyze the three eigenvalues $\lambda_1^{E_1}$, $\lambda_1^{E_2}$ and $\lambda_2^{E_2}$. Therefore, we present in Tables 8 and 9, the analysis of sensitivity for $\lambda_1^{E_2}$, $\lambda_2^{E_2}$. The sensitivity results for $\lambda_1^{E_1}$ are very similar to the case were $m > a$ (case 1).

Table 8: Regression slopes of $\lambda_1^{E_2}$ for each parameter.

Parameter	α	σ	γ	q	u	g
Slope	0.4764	0.2535	-1.0134	0.1274	-0.1382	0.1062
Parameter	p	a	c	θ	r	m
Slope	- 0.6685	-4.8558	-3.8989	-0.1514	0.0643	-6.1289
Parameter	λ	β	ν	μ	ℓ	b
Slope	0.0841	0.6735	-2.2916	-0.0819	0.0761	-0.0622

Table 9: Regression slopes of $\lambda_2^{E_2}$ for each parameter.

Parameter	α	σ	γ	q	u	g
Slope	0.5313	0.0484	-0.0187	-0.0313	-0.3421	-0.0486
Parameter	p	a	c	θ	r	m
Slope	-0.0726	0.2256	0.1760	0.0122	0.0105	0.3220
Parameter	λ	β	ν	μ	ℓ	b
Slope	0.2590	0.0036	0.1006	-0.1795	-0.2026	-0.2385

6.4 Behavior based on m, $\lambda_1^{E_1}$, $\lambda_1^{E_2}$ and $\lambda_2^{E_2}$

Based on the results of the simulations and the stability analysis of the equilibrium points it is possible to suggest a prediction rule based on the values of $\lambda_1^{E_1}$, $\lambda_1^{E_2}$ and $\lambda_2^{E_2}$. In Figure

 ISBN: 978-84-617-8694-7

(7) we present the way in which the behavior of the system can be classified. We show that, for the vast majority of the parameter space that was explored in this work, a simple analysis of the values of m, $\lambda_1^{E_1}$, $\lambda_1^{E_2}$ and $\lambda_2^{E_2}$ is enough to predict the model's behavior. The scheme was successful in predicting the outcome of the numerical simulations in 96.35% of the simulations.

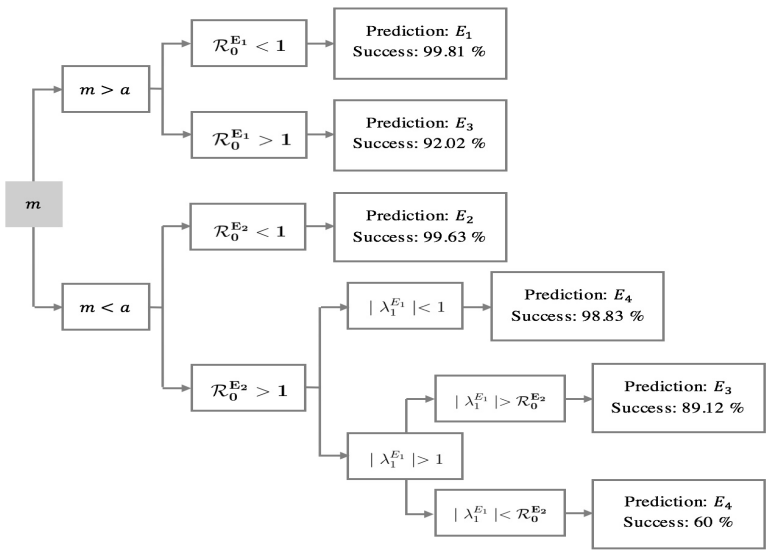


Figure 7: Scheme for a numerical simulation of the system. The equilibrium points, their stability and a simulated solution is computed. If the numerical solution converges to any of the computed candidate equilibrium points, the simulation is classified as a success.

7 Conclusions

In this work we presented an model for the study of prey-predator dynamics with the presence of disease and herd behavior. The theoretical analysis and the numerical simulations suggest that, in the majority of the parameter combinations studied, the behavior of the model can be predicted by the analysis of just four fundamental quantities in the system (m, $\lambda_1^{E_1}$, $\lambda_1^{E_2}$ and $\lambda_2^{E_2}$).

The first fundamental quantity is the natural mortality rate of predators, which is crucial to define the survival of the predator species. Since the mortality of diseased predators is supposed to be equal or higher than the the mortality of healthy ones, parameter m plays a determinant role in the dynamics of the system.

 ISBN: 978-84-617-8694-7

The second, third and fourth fundamental quantities are the values of $\mathcal{R}_0$ (basic reproduction number) calculated in the two disease-free equilibria of the system (E_1 and E_2): $\mathcal{R}_0^{E_1}$ and $\mathcal{R}_0^{E_2}$. Sensitivity analysis through linear regression between the parameters has shown that the parameter with the strongest influence is $\mathcal{R}_0^{E_1}$.

Parameters α and λ are related to the vertical and horizontal transmission rates, respectively, and have a positive correlation with $\mathcal{R}_0^{E_1}$. Parameters u (influence intra-specific competition between healthy and diseased prey on the infected prey population) has a negative effect on the spread of the disease. Therefore, it is clear that a species with the behavior of marginalizing or being hostile to the diseased individuals reduces the chance of permanence of an epidemics in the population.

The analysis of sensitivity for $\mathcal{R}_0^{E_2}$, involves two eigenvalues $\lambda_1^{E_2}$ and $\lambda_2^{E_2}$. The strongest positive correlation between the first one and the parameters occurs for parameters α and β. Thus, again, vertical transmission in prey plays an important role in the maintenance of the disease. Interestingly, the horizontal transmission in the predator population (β) plays a more important role in the destabilization of the disease-free coexistence than the horizontal transmission between prey (λ). Strong negative correlations were observed for parameters a and m. For the sensitivity of $\lambda_2^{E_2}$, again, α displayed a strong positive correlation, followed by parameters m and λ (horizontal transmission). Parameters u and b (mortality of diseased predators) displayed the strongest negative correlation.

Given the importance of parameters α, λ and u, our results suggest that the removal of diseased-prey may be the most effective strategy to lead the system to a disease-free equilibrium.

Acknowledgements

The research has been partially supported by the project "Metodi numerici nelle scienze applicate" of the Dipartimento di Matematica "Giuseppe Peano".

References

[1] V. Ajraldi, E. Venturino, *Mimicking spatial effects in predator-prey models with group defenses*, J. Vigo Aguiar, P. Alonso, S. Oharu, E. Venturino, B. Wade (Eds), Proceedings of the International Conference CMMSE 2009. (2009), 57–66.

[2] V. Ajraldi, M. Pittavino, E. Venturino, *Modelling herd behavior in population systems*, Nonlinear Analysis Real World Applications, **12** (2011) 2319–2338.

[3] M. Banerjee, B. W. Kooi, E. Venturino, *An ecoepidemic model with prey herd behavior and predator feeding saturation response on both healthy and dis-*

 ISBN: 978-84-617-8694-7

eased prey, Mathematical Models in Natural Phenomena, **12**(2), (2017) 133–161. https://doi.org/10.1051/mmnp/201712208

[4] R. G. Bengis, N. P. J. Kriek, D. F. Keet, J. P. Raath, V. De Vos, H. F. A. K. Huchezermeyer, *An outbreak of bovine tuberculosis in a free-ranging buffalo in the Kruger National Park*, Journal of Veterinary Research. **63** (1996), 15–18.

[5] I. M. Bulai, E. Venturino, *Shape effects on herd behavior in ecological interacting population models*, to appear in Mathematics and Computers in Simulation.

[6] E. Cagliero, E. Venturino, *Ecoepidemics with infected prey in herd defence: the harmless and toxic cases*, International Journal of Computer Mathematics. **93**(1) (2016), 108–127.

[7] A. Caron, P. C. Cross, J. T. Du Toit, *Ecological implications of bovine tuberculosis in African buffalo herds*, Ecol. Appl. **13** (2003), 1338–1345.

[8] J. Chattopadhyay, S. Chatterjee, E. Venturino, *Patchy agglomeration as a transition from monospecies to recurrent plankton blooms*, J Theor Biol. **253** (2008), 289–295.

[9] O. Diekmann, J. A. P. Heesterbeek, J. A. J. Metz, *On the definition and the computation of the basic reproduction ratio R_0 in models for infectious diseases in heterogeneous populations*, Journal of Mathematical Biology. **28**(4) (1990), 356–382.

[10] M. M. A. El-Sheikh, S. A. A. El-Marouf, *On stability and bifurcation of solutions of an SEIR epidemic model with vertical transmission*, International Journal of Mathematics and Mathematical Sciences. **56** (2004), 2971–2987.

[11] G. Gimmelli, B. W. Kooi, E. Venturino, *Ecoepidemic models with prey group defense and feeding saturation*, Ecological Complexity, **22**, (2015) 50–58.

[12] B. W. Kooi, E. Venturino, *Ecoepidemic predator-prey model with feeding satiation, prey herd behavior and abandoned infected prey*, Math. Biosc **274** (2016) 58–72.

[13] J. M. Heffernan, R. J. Smith, L. M. Wahl, *Perspectives on the Basic Reproductive Ratio*, Journal of the Royal Society Interface. **2**(4) (2005), 281–293, doi:10.1098/rsif.2005.0042. PMC 1578275Freely accessible. PMID 16849186.

[14] D. F. Keet, H. Davies-Mostert, R. G. Bengis, P. Buss, M. Hofmeyr, S. Ferreira, E. Lane, P. Miller, B. G. Daly (eds.), *Disease risk assessment workshop report: African lion (Panthera leo) bovine tuberculosis*, Conservation Breeding Specialist Group (CBSG SSC/UICN)/CBSG Southern Africa. Endangered Wildlife Trust, 2009.

 ISBN: 978-84-617-8694-7

[15] M. Kretzschmar, P. F. Teunis, R. G. Pebody, *Incidence and reproduction numbers of pertussis: estimates from serological and social contact data in five European countries*, PLOS Medicine **7**(6) (2010), e1000291. doi:10.1371/journal.pmed.1000291. PMC 2889930Freely accessible. PMID 20585374.

[16] A.L. Michel, R.G. Begins, D.F. Keet, M. Hofmeyr, L.M. de Klerk, P.C. Cross, A.E. Jolles, D. Cooper, I.J. Whyte, P. Buss, J. Godfroid, *Wildlife tuberculosis in South African conservation areas: Implications and challenges*, Veterinary Microbiology. **112** (2006), 91–100.

[17] M. A. Miller, P. C. White, R. G. Bengis, *Tuberculosis in South African Wildlife:Why is it important?*, SU Language Centre, SUN MeDIA, ISBN: 978-0-7972-1552-8, 2015.

[18] R. K. Naji, R. M. Hussien, *The Dynamics of Epidemic Model with Two Types of Infectious Diseases and Vertical Transmission*, Journal of Applied Mathematics. (2016), article ID 4907964, 16 pages http://dx.doi.org/10.1155/2016/4907964

[19] L. Perko, *Differential equations and dynamical systems*, Springer, New York, NY, USA, 3rd edition, 2001.

[20] A. F. Renwick, P. C. White, R. G. Bengis, *Bovine tuberculosis in southern African wildlife: a multi-species pathogen system*, Epidemiol. Infect. **135** (2007), 529–540.

[21] R. P. Sanches, *Análise do número de reprodutibilidade basal na fase inicial de doenças causadas por vetores (Analysis of the basic reproduction number in the initial phase of vector-borne diseases)*, Tese de Doutorado, Universidade de São Paulo, São Paulo, 2015.

[22] S. Sarwardi, M. Haque, E. Venturino, *A Leslie-Gower Holling-type II ecoepidemic model*, J Appl Math Comput. **35** (2011), 263–280.

[23] A. A. Soares, *Tuberculose afecta milhares de animais selvagens em parque da África do Sul*, http://www.publico.pt/j193597, accessed on 11/29/2016.

[24] P. van den Driessche, J. Watmough, *Reproduction numbers and sub-threshold endemic equilibria for compartmental models of disease transmission*, Mathematical Biosciences. **180** (2002), 29–48.

[25] E. Venturino, *A minimal model for ecoepidemics with group defense*, Biological Systems. **19** (2011), 763–781.

 ISBN: 978-84-617-8694-7

[26] E. Venturino, S. Petrovskii, *Spatiotemporal behavior of a prey-predator system with a group defense for prey*, Ecological Complexity **14** (2013) 37–47. http://dx.doi.org/10.1016/j.ecocom.2013.01.004

[27] X. Zhou, J. Cui, *Analysis of stability and bifurcation for an SEIV epidemic model with vaccination and nonlinear incidence rate*, Nonlinear Dynamics. **63**(4) (2011), 639–653.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

Ball convergence of a sixth-order Newton-like method based on means under weak conditions

Á. A. Magreñán[1], I. K. Argyros[2], J. J. Rainer[1] and J. A. Sicilia[1]

[1] *Escuela Superior de Ingeniería y Tecnología, Universidad Internacional de La Rioja*

[2] *Department of Mathematics Sciences, Cameron University*

emails: alberto.magrenan@unir.net, iargyros@cameron.edu,
javier.rainer@unir.net, juanantonio.sicilia@unir.net

Abstract

We study the local convergence of a Newton-like method of convergence order six to approximate a locally unique solution of a nonlinear equation. Earlier studies show convergence under hypotheses on the seventh derivative or even higher. The convergence in this study is shown under hypotheses on the first derivative although only the first derivative appears in these methods. Hence, the applicability of the method is expanded.

Key words: Newton-like method, local convergence, Stolarky means, Gini means, efficiency index
MSC 2000: 65D10, 65D99, 65G99, 90C30

1 Introduction

In this study we are concerned with the problem of approximating a locally unique solution ξ of equation

$$F(x) = 0, \tag{1.1}$$

where F is a differentiable function defined on a convex subset D of S with values in S, where S is $\mathbb{R}$ or $\mathbb{C}$.

Many problems from Applied Sciences including engineering can be solved by means of finding the solutions of equations in a form like (1.1) using Mathematical Modelling [4, 5]. For example, dynamic systems are mathematically modeled by difference or differential

 ISBN: 978-84-617-8694-7

equations, and their solutions usually represent the states of the systems. Except in special cases, the solutions of these equations can be found in closed form. This is the main reason why the most commonly used solution methods are usually iterative. The convergence analysis of iterative methods is usually divided into two categories: semilocal and local convergence analysis. The semilocal convergence matter is, based on the information around an initial point, to give criteria ensuring the convergence of iteration procedures. A very important problem in the study of iterative procedures is the convergence domain. In general the convergence domain is small. Therefore, it is important to enlarge the convergence domain without additional hypothesis. Another important problem is to find more precise error estimates on the distances $\|x_{n+1}-x_n\|$, $\|x_n-\xi\|$. These are with the study of the dynamical behavior our objectives in this paper.

The most popular method for approximating a simple solution ξ of equation (1.1) is undoubtedly Newton's method defined for all $n=0,1,2,\ldots$ by

$$x_{n+1} = x_n - \frac{F(x_n)}{F'(x_n)}, \tag{1.2}$$

where x_0 is an initial point. Newton's method converges quadratically to ξ [4, 5] provided that F' does not vanish in D and x_0 is close enough to ξ. To obtain higher order of convergence many third order methods have been proposed [1]– [6].

These methods look like

$$x_{n+1} = x_n - \frac{F(x_n)}{F'(x_n)}\Gamma(s(x_n)), \tag{1.3}$$

where

$$s=s(x)=\frac{F'\left(x-\frac{F(x)}{F'(x)}\right)}{F'(x)}$$

and only differ in the choice of function Γ.

Recently the method defined for all $n=0,1,2,\ldots$ by

$$x_{n+1} = x_n - B(x_n)\tfrac{F(x_n)}{F'(x_n)}, \tag{1.4}$$

where

$$B(x_0)=\Gamma\left(s(x_0)\right)+\frac{F\left(x_0-\frac{F(x_0)}{F'(x_0)}\Gamma\left(s(x_0)\right)\right)\left[F'\left(x_0-\frac{F(x_0)}{F'(x_0)}\right)+F'(x_0)\right]}{F(x_0)\left[3F'\left(x_0-\frac{F(x_0)}{F'(x_0)}\right)-F'(x_0)\right]}$$

was studied in [7]. The sixth order of convergence was shown under hypotheses reauhing up to the seventh derivative of the function F. In terms of the computational cost, method (1.4) require two function evaluations and two first derivative evaluations per iterations.

 ISBN: 978-84-617-8694-7

Therefore, the efficiency index is $6^{1/4} \approx 1.56508$ provided that evaluations of functions such as exp, ln or mth root are negleited.

In the present study we first study the local convergence of method (1.4) using hypotheses up to the first derivative of function F. We also provide the radius of the convergence ball, computable error bounds on the distances involved and a uniqueness of the solution result. Such results were not given in [7] or the earlier related studies [8], [9]. This way we expand the applicability of method (1.4). It is convenient for us to simplify method (1.4) and study the equivalent method defined for all $n = 0, 1, 2, \ldots$ by

$$\begin{aligned} y_n &= x_n - \frac{F(x_n)}{F'(x_n)}, \\ z_n &= x_n - \frac{F(x_n)}{F'(x_n)}\Gamma\left(s(x_n)\right), \\ x_{n+1} &= z_n - \frac{s(x_n)+1}{3s(x_n)-1}\frac{F(z_n)}{F'(x_n)}, \end{aligned} \tag{1.5}$$

where $s(x_n)$ is defined in method (1.3).

2 Main result

Let $F : D \subset X \to Y$ be a continuously Fréchet-differentiable operator, X, Y be Banach spaces and D an open, convex subset of X. Consider (1.1) rewritten

$$\begin{aligned} y_n &= x_n - F'(x_n)^{-1}F(x_n), \\ z_n &= x_n - \Gamma(F'(x_n)^{-1}(F'(x_n - F'(x_n)^{-1}F(x_n))F'(x_n)^{-1}F(x_n), \\ x_{n+1} &= z_n - (3s(x_n) - I)^{-1}(s(x_n) + I)F'(x_n)^{-1}F(z_n), \end{aligned} \tag{2.1}$$

where $s(x) = F'(x)^{-1}F'(x - F'(x)^{-1}F(x))$ and $\Gamma(\cdot) : X \to \mathbb{L}(Y, X)$ is a linear operator.

Next, using the preceding notation we can show the local convergence result for method (2.1).

Theorem 1 *Let F: $U(\xi, R_0) \subseteq X \to Y$ be Fréchet-differentiable. Suppose that there exist $\xi \in D$, $L_0 > 0$ such that for each $x \in U(\xi, R_0)$:*

$$F(\xi) = 0 \tag{2.2}$$

$$F'(\xi) \neq 0, \tag{2.3}$$

$$\left\| F'(\xi)^{-1}\left(F'(x) - F'(\xi)\right)\right\| \leqslant L_0 \|x - \xi\|. \tag{2.4}$$

Moreover, suppose that there exist $L > 0$, $M \geq 1$, $\{s_n\}$ and functions $\Gamma(\cdot) : X \to L(Y, X)$, A:$\left[0, \frac{1}{L_0}\right) \to \left[0, \frac{1}{M}\right)$ continuous and non-decreasing such that for each $x, y \in U(\xi, \frac{1}{L_0}) \cap U(\xi, R_0)$

$$\left\| F'(\xi)^{-1}\left(F'(x) - F'(y)\right)\right\| \leqslant L \|x - y\|, \tag{2.5}$$

 ISBN: 978-84-617-8694-7

$$\left\|F'(\xi)^{-1}F'(x)\right\| \leqslant M, \tag{2.6}$$

and

$$|1-\Gamma\left(s(x)\right)| \leqslant A(\|x-\xi\|)\|x-\xi\|. \tag{2.7}$$

Then, sequence $\{x_n\}$ *generated by method* (1.5) *for* $x_0 \in U\left(\xi,r\right)\backslash\{\xi\}$ *remains in* $U\left(\xi,r\right)$ *for each* $n=0,1,2,\ldots$ *and converges to* ξ *where* r *is defined in* (2.1). *Moreover, the following estimates for each* $n=0,1,2,\ldots$

$$\|y_n-\xi\| \leqslant g_1\left(\|x_n-\xi\|\right)\|x_n-\xi\| \leq \|x_n-\xi\| < r, \tag{2.8}$$

$$\|z_n-\xi\| \leqslant g_2\left(\|x_n-\xi\|\right)\|x_n-\xi\| \leq \|x_n-\xi\| \tag{2.9}$$

$$\left\|\left(3F'(y_n)-F'(x_n)\right)^{-1}\left(2F'\left(\xi\right)\right)\right\| \leqslant \tfrac{1}{1-g_0(\|x_n-\xi\|)} \tag{2.10}$$

and

$$\|x_{n+1}-\xi\| \leqslant g_3\left(\|x_n-\xi\|\right)\|x_n-\xi\| \leq \|x_n-\xi\|, \tag{2.11}$$

where

$$g_1(t)=\frac{Lt}{2\left(1-L_0t\right)},\quad g_2(t)=\frac{Lt+2MA(t)}{2\left(1-L_0t\right)},\quad h_2(t)=g_2(t)-1$$

$$r_1=\frac{2}{2L_0+L}<\frac{1}{L_0}g_0(t)=\frac{L_0}{2}\left(\frac{3}{2}Lt+1\right)th_0(t)=g_0(t)-1$$

$$r_0=\frac{4L_0L}{L_0+\sqrt{L_0^2+12L_0L}}$$

$$g_3(t)=\left[1+\frac{ML_0\left(1+g_1(t)\right)}{2(1-L_0t)(1-g_0(t))}\right]g_2(t)=\frac{1}{2(1-L_0t)}\left[1+\frac{ML_0\left(1+g_1(t)\right)t}{2(1-L_0t)(1-g_0(t))}\right](Lt+2MA(t))$$

$$h_3(t)=g_3(t)-1.$$

$$r=\min\left\{r_i\right\},i=0,1,2,3.$$

and

$$R_0:=sup\{t\in[0,R):U(\xi,t)\subseteq D\}.$$

Furthermore for $T\in[r,\frac{2}{L_0})$, *the vector point* ξ *is the only solution of equation* $F(x)=0$ *in* $U\left(\xi,R_0\right)\cap\overline{U}(\xi,T)$.

 ISBN: 978-84-617-8694-7

References

[1] S. Amat, S. Busquier, S. Plaza, Dynamics of the King and Jarratt iterations, Aequationes Math. 69 (2005) 3, 212–223.

[2] S. Amat, S. Busquier, S. Plaza, Chaotic dynamics of a third-order Newton-type method, J. Math. Anal. Appl. 366 (2010) 1, 24–32.

[3] S. Amat, M.A. Hernández, N. Romero, A modified Chebyshev's iterative method with at least sixth order of convergence, Appl. Math. Comput. 206(1) (2008), 164–174.

[4] I.K. Argyros, "Convergence and Application of Newton-type Iterations," Springer, 2008.

[5] I.K. Argyros, S. Hilout, Numerical methods in Nonlinear Analysis, World Scientific Publ. Comp. New Jersey, 2013.

[6] D.D. Bruns, J.E. Bailey, Nonlinear feedback control for operating a nonisothermal CSTR near an unstable steady state, Chem. Eng. Sci. 32, (1977), 257–264.

[7] D. Herceg, Dj. Herceg, Sixth order modifications on Newton's method based on Stolarsky and Gini means, J. Comput Appl. Math. 267 (2014), 244–253.

[8] M.A. Hernández, Chebyshev's approximation algorithms and applications, Computers Math. Applic. 41(3-4) (2001),433-455.

[9] M.A. Hernández, M.A. Salanova, Sufficient conditions for semilocal convergence of a fourth order multipoint iterative method for solving equations in Banach spaces. Southwest J. Pure Appl. Math, 1 (1999), 29–40.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

An efficient optimal family of sixteenth order methods for nonlinear equations

Á. A. Magreñán[1], I. K. Argyros[2], R. Behl[3] and S. S. Motsa[3]

[1] *Escuela Superior de Ingeniería y Tecnología, Universidad Internacional de La Rioja*

[2] *Departamento de Matemática Aplicada y Estadística, Universidad Politécnica de Cartagena, Spain*

[3] *School of Mathematics, Statistics and Computer Sciences, University of KwaZulu-Natal*

emails: alberto.magrenan@unir.net, sergio.amat@upct.es, ramanbehl87@yahoo.in, sandilemotsa@gmail.com

Abstract

The principle aim of this manuscript is to propose a scheme that can be applied to any optimal iteration function of order eight or a family of eighth-order methods to further develop new interesting optimal method/family of iterative methods of order sixteen. The proposed scheme requires four evaluations of the involved function and one evaluation of its first-order derivative, being optimally consistent with the conjecture of Kung-Traub. In addition, theoretical and computational properties are fully investigated along with a main theorem describing the order of convergence of the proposed scheme. Moreover, the conjugacy maps and the strange fixed points of some iterative methods are discussed, their basins of attractions are also given to show their dynamical behavior around the simple roots. From the numerical experiments, we find that our proposed schemes perform better than the existing ones when we checked the performance of the proposed methods on a variety of nonlinear equations.

Key words: Efficient method, high order, nonlinear equations,
MSC 2000: 65D10, 65D99, 65G99, 90C30

1 Introduction

The conceptualization and construction of multi-point solution techniques have always been a paramount importance in the field of numerical analysis that provide more accurate

and efficient approximate solution ξ of nonlinear equation of the form

$$f(x) = 0, \tag{1.1}$$

where $f : \mathbb{D} \subseteq \mathbb{R} \to \mathbb{R}$ is a nonlinear sufficiently differentiable function in an interval $\mathbb{D}$. This topic has attracted the attention of many researchers from the worldwide, when Traub [9] initiated the qualitative as well as the quantitative analysis of iterative methods.

In the past, Ostrowski was the first mathematician who proposed an optimal fourth-order multi-point method which requires only three functional evaluations. On the other hand, Jarratt [3,4], in 1966 and King [6], in 1975, had also proposed various optimal fourth-order multi-point methods. King further showed that Ostrowski's method a particular case of his proposed family.

Due to the advancement of digital computer, advanced computer arithmetics and symbolic computation, with in the last two decade a large number of optimal eighth-order methods have been proposed by various authors. Most of them are the extension of Newton's method or Newton like method at the expense of additional functional evaluations or increase the sub step of the original methods.

In recent years, some researchers like Guem and Kim [1, 2], Sharma et al. [8], Ullah et al. [10], have also proposed optimal sixteen order extension of Newton's method. Kung and Traub [7], proposed two general classes of m-point iterative methods with optimal convergence order of 2^{m-1} with first-order derivative and without derivative. Nowadays, obtaining new four-step optimal methods of order sixteen not requiring the computation of second derivative, is very important and interesting task from the computational point of view because the efficiency indices of all the proposed methods is $E = \sqrt[5]{16} \approx 1.741$, far better than the classical Newton's method $E \approx 1.414$ and they converge very fast towards the required root. Sharma et al. [8] and Ullah et al. [10], have proposed sixteen-order modification of fourth-order King's method. On the other hand, Guem and Kim [1, 2] have given sixteen-order modification of of a particular fourth and eight-order iterative methods. However, Kung-Traub [7] also given two general classes of m-point iterative methods with optimal convergence order of 2^{m1}. But that is the modification of Newton's method at every step. According to our knowledge, no one given a general scheme which will be applicable on any optimal eight-order methods to further extend sixteenth-order optimal method whose first sub step should be Newton's method.

In order to develop a new scheme, it is quite often to approximate functions. Several types of approximations are available in the literature for e.g. Functional approach, Sampling approach, Geometric approach, Weight function approach, Adomain approach, Composition approach and Rational function approach. Every approach have some advantageous and disadvantageous because it's dependent on the considered problem. The choice of suitable approximation approach can save considerable amount of computation. Rational function approach is one of the most important techniques in numerical analysis for approximating the function or to find the next approximation.

 ISBN: 978-84-617-8694-7

According to our knowledge, Jarratt and Nudds [5] were the first persons who used this approach for developing new schemes with higher-order convergence. The rational function operator have the property of being unitary. The reason behind this, we have to find the same number of undermined constants in the same number of linear independent equations. These linear independent equations can be formed with the help of tangency conditions. In general, the number of tangency conditions are equal to number of undetermined constants. Further, we will get an improved method with higher-order convergence as we increase the number of undetermined constants in the rational function.

Therefore, we are interested to develop a scheme that can be applied on any eighth-order optimal method/family of methods whose first sub -step should be Newton's method to get further optimal family of sixteen-order method instead of applying that approach only on any particular iterative method. The derivation of the proposed scheme is based on the concept of the rational approximations. The beauty of the proposed scheme is that it is applicable to every optimal scheme of order eight whose first sub step should be Newton's method.

2 Development of sixteenth-order optimal scheme

We propose an optimal sixteenth-order family of iterative methods. Therefore, we consider a eighth-order scheme in the following way

$$\begin{aligned} y_n =& x_n - \frac{f(x_n)}{f'(x_n)}, \\ z_n =& \psi_4(x_n,\ y_n), \\ t_n =& \phi_8(x_n,\ y_n,\ z_n). \end{aligned} \tag{2.1}$$

For getting the next approximation x_{n+1} to a root, we consider the following rational function

$$w(x) = \frac{(x - x_n) + \alpha_1}{\alpha_2(x - x_n)^3 + \alpha_3(x - x_n)^2 + \alpha_4(x - x_n) + \alpha_5}, \tag{2.2}$$

where $\alpha_1,\ \alpha_2,\ \alpha_3,\ \alpha_4$ and α_5 can be determined by imposing the following tangency conditions

$$w(x_n) = f(x_n),\ \ w'(x_n) = f'(x_n),\ \ w(y_n) = f(y_n),\ \ w(z_n) = f(z_n),\ \ w(t_n) = f(t_n). \tag{2.3}$$

Now, we shall find the next approximation x_{n+1} by using the above rational function (2.2). Further, let us assume that this function meets the x – axis at $x = x_{n+1}$, we have

$$w(x_{n+1}) = 0, \tag{2.4}$$

which further yields

$$x_{n+1} = x_n - \alpha_1. \tag{2.5}$$

 ISBN: 978-84-617-8694-7

It is straight forward to say that from the above equation (2.5), that if we get the value of disposable parameter α_1 then our main target is achieved. In order to achieve this, we impose the first two tangency conditions defined in equation (2.3), to obtain

$$\alpha_1 = \alpha_5 f(x_n), \;\; \alpha_4 = \frac{1-\alpha_5 f'(x_n)}{f(x_n)}. \tag{2.6}$$

From the last three tangency conditions, we get

$$\begin{aligned}
&f(y_n)\big[f'(x_n)\left(f'(x_n)\left(2\alpha_5 f'(x_n)-1\right)+\alpha_3 f(x_n)^2\right)-\alpha_2 f(x_n)^3\big] = f'(x_n)^2 f(x_n)\left(\alpha_5 f'(x_n)-1\right),\\
&f(z_n)\left[\frac{(1-\alpha_5 f'(x_n))(z_n-x_n)}{f(x_n)}+\alpha_2(z_n-x_n)^3+\alpha_3(x_n-z_n)^2+\alpha_5\right] = \alpha_5 f(x_n)+z_n-x_n,\\
&f(t_n)\left[\frac{(1-\alpha_5 f'(x_n))(t_n-x_n)}{f(x_n)}+\alpha_2(t_n-x_n)^3+\alpha_3(t_n-x_n)^2+\alpha_5\right] = \alpha_5 f(x_n)+t_n-x_n.
\end{aligned} \tag{2.7}$$

By eliminating α_2 and α_3 from the above equations, we obtain the following value of α_5

$$\alpha_5 = \frac{a_n b_n\big(u_1 f(x_n)^2 f(y_n)+u_2 f'(x_n) f(t_n) f(z_n)\big)}{v_1 f(x_n)^3+v_2 f'(x_n) f(t_n) f(z_n)} \tag{2.8}$$

where
$u_1 = f(t_n)\big(b_n^2 f'(x_n)+b_n f(x_n)-c_n f(z_n)\big)+a_n\big(f(x_n)-a_n f'(x_n)\big)f(z_n)$, $u_2 = a_n b_n c_n f'(x_n)\big(f(y_n)-f(x_n)\big)+c_n f(y_n) f(x_n)(a_n-b_n)$, $v_1 = f(y_n)\big[b_n f(t_n)\left(b_n^2 f'(x_n)+b_n f(x_n)-c_n f(z_n)\right)+\big(a_n^3 f'(x_n)+c_n a_n f(t_n)-a_n^2 f(x_n)\big)f(z_n)\big]$, $v_2 = a_n^2 b_n^2 c_n f'(x_n)^2\big(2f(y_n)-f(x_n)\big)+a_n b_n c_n(2a_n-c_n)f'(x_n) f(y_n) f(x_n)+c_n\big(a_n b_n-a_n c_n-b_n^2\big)f(y_n) f(x_n)^2$, $a_n = x_n - z_n$, $b_n = t_n - x_n$, $c_n = t_n - z_n$.
Finally, we obtain the new optimal scheme of order sixteen by using the equation (2.8), we yield

$$\begin{aligned}
y_n &= x_n - \frac{f(x_n)}{f'(x_n)},\\
z_n &= \psi_4(x_n,\, y_n),\\
t_n &= \phi_8(x_n,\, y_n,\, z_n),\\
x_{n+1} &= x_n - \alpha_5 f(x_n).
\end{aligned} \tag{2.9}$$

where α_5 is previously defined in equation (2.8). The following theorem 1 demonstrates that the order of convergence will reach at the optimal eighth-order without using any more functional evaluations.

Theorem 1 *Let $f : \mathbb{R} \to \mathbb{R}$ has a simple zero ξ and is a sufficiently differentiable function an interval containing ξ. Further, assume that $\phi_4(x_n,\, y_n)$ and $\psi_8(x_n,\, y_n,\, z_n)$ are any optimal scheme of order four and eight-order, respectively. In addition, we also consider that initial guess $x = x_0$ is sufficiently close to ξ for the guaranteed convergence. Then, the*

 ISBN: 978-84-617-8694-7

iterative scheme defined by (2.9) *has an optimal sixteenth-order convergence and satisfy the following error equation*

$$\begin{aligned} e_{n+1} =& c_2\left(c_2^4 - 3c_3c_2^2 + 2c_4c_2 + c_3^2 - c_5\right) A_0B_0e_n^{16} + \left[c_2^5(c_1B_0 + c_0B_1) - 4c_2^6c_0B_0 + 16c_3c_2^4c_0B_0\right. \\ & - c_2^3\{10c_4c_0B_0 + 3c_3(c_1B_0 + c_0B_1)\} + 2c_2^2\{(-7c_3^2c_0B_0 + 3c_5c_0B_0 + c_4(c_1B_0 + c_0B_1)\} \\ & + c_2\{c_3^2(c_1B_0 + c_0B_1) + 8c_4c_3c_0B_0 - 2c_6c_0B_0 - c_5(c_1B_0 + c_0B_1)\} + 2c_3(c_3^2 - c_5)c_0B_0\big]e^{17} \\ & + O(e_n^{18}), \end{aligned} \tag{2.10}$$

where $e_n = x_n - \xi$ *and* $c_j = \frac{f^{(j)}(\xi)}{j!f'(\xi)}$ *for* $j = 2, 3, \ldots, 16$.

References

[1] Y.H. Geum, Y.I. Kim, A family of optimal sixteenth-order multipoint methods with a linear fraction plus a trivariate polynomial as the fourth-step weighting function, Comput. Math. Appli. 61 (2011), 3278–3287.

[2] Y.H. Geum, Y.I. Kim, A biparametric family of optimally convergent sixteenth-order multipoint methods with their fourth-step weighting function as a sum of a rational and a generic two-variable function, J. Comput. Appl. Math. 235 (2011), 3178–3188.

[3] P. Jarratt, Some fourth-order multipoint iterative methods for solving equations, Math. Comput. 20 (1966), 434–437.

[4] P. Jarratt, Some efficient fourth-order multipoint methods for solving equations, BIT 9 (1969), 119–124.

[5] P. Jarratt, D. Nudds, The use of rational functions in the iterative solution of equations on a digital computer, The Comput. J. 8(1) (1965), 62–65.

[6] R.F. King, A family of fourth order methods for nonlinear equations, SIAM J. Numer. Anal. 10 (1973), 876–879.

[7] H.T. Kung and J.F. Traub, Optimal order of one-point and multi-point iteration, J. ACM 21 (1974), 643–651.

[8] J. R. Sharma, R.K. Guha, Puneet Gupta, Improved King's methods with optimal order of convergence based on rational approximations, Appl. Math. Lett. 26 (2013), 473–480.

[9] J.F. Traub, Iterative methods for the solution of equations, Prentice-Hall, Englewood Cliffs, 1964.

[10] M.Z. Ullah, A.S. Al-Fhaid, F. Ahmad, Four-Point Optimal Sixteenth-Order Iterative Method for Solving Nonlinear Equations, J. Appl. Math. 2013 (2013), Article ID 850365.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

Expansions of ratios of gamma functions – an application to the distribution of the likelihood ratio test statistic used to test the equality of several covariance matrices

Filipe J. Marques[1]

[1] *Centro de Matemática e Aplicações (CMA) and Departamento de Matemática, Faculdade de Ciências e Tecnologia da Universidade NOVA de Lisboa*

emails: fjm@fct.unl.pt

Abstract

In this work the problem of testing the equality of several covariance matrices is addressed. Using an expansion of the ratio of two gamma functions, new approximations for the distribution of the test statistic are derived. These approximations are obtained using a new representation of the Fourier transform of the density function the logarithm of the test statistic which is based on products of ratios of gamma functions. The approximations developed are simple, precise and easy to implement.

Key words: asymptotic expansions, Fourier transforms, gamma function, generalized Bernoulli polynomials, likelihood ratio test, mixtures, ratio of gamma functions

1 Introduction

The ratio of two gamma functions appears in different problems and applications of mathematics. In statistics, among other possible applications, we show that it is possible to represent the Fourier transform of the density of the logarithm of the likelihood ratio test statistic used to test the equality of several covariance matrices in the form of products of ratios of gamma functions. Using an expansion for the ratio of two gamma functions given in [10] it is possible obtain a representation for the product of ratios of gamma functions which will enable us to derive simple and precise approximations for the distribution of the likelihood ratio test statistic. A similar procedure was also used in related problems in [3] and [7]. For other interesting results on the ratio of two gamma functions please see [1, 4, 5, 9]. A simple illustration of the procedure used will be given in the next sections.

 ISBN: 978-84-617-8694-7

2 Expansion for the ration of two gamma functions

In 1951, Tricomi and A. Erdélyi, obtained a representation, for the ratio of two gamma functions, in terms of generalized Bernoulli polynomials, which is given by (see [6] and [10])

$$\frac{\Gamma(z+a)}{\Gamma(z+b)} = z^{a-b}\sum_{k=0}^{M-1}\frac{(-1)^k B_k^{a-b+1}(a)(b-a)_k}{k!}z^{-k} + z^{a-b}O(z^{-M}) \tag{1}$$

with $|arg(z+a)| \leq \pi - \epsilon$, $\epsilon > 0$ and where $B_n^y(x)$ are the generalized Bernoulli polynomials which can be defined through the equality (see [6])

$$\frac{t^y \mathrm{e}^{xt}}{(\mathrm{e}^t-1)^y} = \sum_{k=0}^{\infty}\frac{t^k}{k!}B_k^y(x), \quad |t| < 2\pi\,.$$

The expansion in (1) may be used to developed a representation for the product of ratios of gamma functions which will enable the development of approximations for the likelihood ratio test statistic used to test the equality of several covariance matrices. The procedure is briefly illustrate in Section 3.

3 Asymptotic approximations for the likelihood ratio test statistic

Let us assume we have q independent samples, $N_1, \ldots, N_q$, from q multivariate Normal populations, $N_p\big(\underline{\mu}, \Sigma\big)$. As it is referred in [2] the multivariate normal distribution assumption can be extended to a more general assumption of elliptically contoured distributions. The null hypothesis is

$$H_0 : \Sigma_1 = \ldots = \Sigma_q\,.$$

The modified likelihood ratio test statistic is given by (see [8])

$$\Lambda = \frac{N^{Np/2}}{\prod_{k=1}^{q} n_k^{n_k p/2}}\,\frac{\prod_{k=1}^{q}|A_k|^{n_k/2}}{|A|^{N/2}} \quad \text{with} \quad N = \sum_{k=1}^{q} n_k\,, \quad n_k = N_k - 1\,, \quad k = 1, \ldots, q\,, \tag{2}$$

and where A_k is the matrix of corrected sums of squares and products formed from the k-th sample and $A = A_1 + \ldots + A_q$. For our purposes it is useful to consider the random variable $W = -\log\Lambda$, and it can be shown that its characteristic function can be written in the following form

$$\Phi_W(t) = \prod_{j=1}^{p}\prod_{k=1}^{q}\prod_{\ell=1}^{n_k}\frac{\Gamma\left(\alpha_{jk\ell}+\beta_{jk\ell}\right)\Gamma\left(\alpha_{jk\ell}-\frac{\mathrm{i}t}{2}\right)}{\Gamma\left(\alpha_{jk\ell}\right)\Gamma\left(\alpha_{jk\ell}+\beta_{jk\ell}-\frac{\mathrm{i}t}{2}\right)} \tag{3}$$

 ISBN: 978-84-617-8694-7

for some parameters $\alpha_{jk\ell}$ and $\beta_{jk\ell}$. If these parameters were all positive the expression in (3) would correspond to the characteristic function of the sum of independent logbeta distributions, however these parameters are not all positive. Nevertheless, it is possible to observe that the representation in (3) is based on products of ratios of gamma functions and thus, using the expansion in (1), it is possible to obtain a new representation for the characteristic function which is a mixture of gamma distributions with the same rate parameter. This feature together with the well known matching moments technique, will enable the derivation of asymptotic approximations in the form of mixture gamma of distributions, all with the same rate parameter, λ, with shape parameters r_j and with weights π_j, $j = 1, \ldots, m$, that is

$$\sum_{j=0}^{m} \pi_j \, \lambda^{r_j} \left(\lambda - \frac{\mathrm{i}t}{2}\right)^{-r_j} . \tag{4}$$

The matching moments technique will be used to determine the parameters in (4). One should note that in [2] the authors already developed very impressive and precise near-exact distributions for the test statistic. However, in the unbalanced case, the approximations obtained using the procedure described in this paper are much simpler and even more precise.

4 Conclusion

The approximations obtained using procedure described in this paper are simple and of easy computational implementation, since they are standard mixtures of gamma distributions and, in addition, they are extremely precise. The main contribution of these results is for the unbalanced case (the case where the samples may not have the same size) where is possible to obtain a significant improvement of the results already available in the literature, however even in the balanced case these approximations are very precise.

Acknowledgements

This work was partially supported by the Fundação para a Ciência e a Tecnologia (Portuguese Foundation for Science and Technology) through the project UID/MAT/00297/2013 (Centro de Matemática e Aplicações).

References

[1] J. Abad and J. Sesma, *Two new asymptotic expansions of the ratio of two gamma functions*, Journal of Computational and Applied Mathematics 173 (2005) 359–363.

[2] C. A. Coelho and F. J. Marques, *Near-exact distributions for the likelihood ratio test statistic to test equality of several variance-covariance matrices in elliptically contoured distributions*, Computational Statistics **27** (2012) 627–659.

[3] C. A. Coelho and R. P. Alberto, *On the Distribution of the Product of Independent Beta Random Variables Applications*, Technical Report, CMA (2012) 12.

[4] J. L. Fields, *A note on the asymptotic expansion of a ratio of gamma functions*, Proc. Edinburgh Math. Soc. **15** (1966) 43–45.

[5] A. Laforgia and P. Natalini, *On the asymptotic expansion of a ratio of gamma functions*, Journal of Mathematical Analysis and Applications **389** (2012) 833–837.

[6] Y. L. Luke, *The special functions and their approximations*, Academic Press, Inc., London 1969.

[7] P. G. Moschopoulos, *New Representations for the Distribution Function of a Class of Likelihood Ratio Criteria*, Journal of Statistical Research **20** (1986) 13–20.

[8] R. J. Muirhead, *Aspects of Multivariate Statistical Theory*, Wiley, New York, 1982.

[9] N. E. Norlund, *Sur les valeurs asymptotics des nombres et des polynômes de Bernoulli*, Rendiconti del Circolo Matematico di Palermo **10** B (1961) 27–44.

[10] F. G. Tricomi and A. Erdélyi, *The asymptotic expansion of a ratio of Gamma functions*, Pacific Journal of Mathematics **1** (1951) 133–142.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

Lie symmetries and Conservation laws for the viscous Cahn-Hilliard equation

A.P. Márquez[1], M.S. Bruzón[1], T.M. Garrido[1] and E. Recio[1]

[1] *Department of Mathematics, University of Cádiz*

emails: almudena.marquez@uca.es, m.bruzon@uca.es, tamara.garrido@uca.es, elena.recio@uca.es

Abstract

In this work we study a viscous Cahn-Hilliard equation from the point of view of Lie symmetries in partial differential equations. A classification of Lie symmetries is found applying the classical Lie method. These symmetries allow the obtaining of travelling wave solutions. On the other hand, we analyze conservation laws using multipliers method proposed by Anco and Bluman. Afterwards, it is possible to calculate potential symmetries of the equation from the corresponding conserved systems.

Key words: nonlinear partial differential equation, Lie symmetries, Conservation laws, travelling wave solutons, multipliers method

In [8] the authors investigated different singular limits of an initial–boundary value problem of the viscous Cahn–Hilliard equation

$$\nu u_t = \Delta(\varphi(u) - \alpha\Delta u + \beta u_t), \tag{1}$$

where u is the solute concentration at point x, t is the time, $\varphi(u)$ is a "homogeneous free energy" and $\alpha/2$ is the gradient energy coefficient describing the contribution of the diffuse interface to the decomposition, βu_t is the viscous term and $\nu = 1 - \beta$.
If $\nu = 1$ and $\beta = 0$, then the Cahn–Hilliard equation is

$$u_t = \Delta(\varphi(u) - \alpha\Delta u). \tag{2}$$

This equation was introduced to study phase separation in binary alloys glasses and polymers and is a good approach to spinodal decomposition [7]. In [6], the authors obtained

 ISBN: 978-84-617-8694-7

classical and nonclassical symmetries of the equation and reduced the equation to ordinary differential equations.

In this paper we solve a group classification problem for equation (1), by studying those diffusion coefficients $f(u)$ which admit classical symmetries. Both the symmetry group and the diffusion coefficients will be found through consistent applications of the Lie-group formalism. We determine conservation laws of equation (1) by using the general multipliers method for this equation.

1 Lie symmetries

Lie group analysis [1, 2, 3, 4, 5, 10] is the most powerful tool to find the general solution of partial differential equations. We consider a one-parameter Lie group of infinitesimal transformations in (x, t, u). The associated Lie algebra of infinitesimal symmetries is the set of vector fields of the form

$$V = \xi(x,t,u)\frac{\partial}{\partial x} + \tau(x,t,u)\frac{\partial}{\partial t} + \phi(x,t,u)\frac{\partial}{\partial u}, \tag{3}$$

where $\xi(x,t,u)$, $\tau(x,t,u)$ and $\phi(x,t,u)$ are the infinitesimals. By requiring that this transformation leaves invariant the set of solutions of (1) we obtain an overdetermined, linear system of equations for the infinitesimals. This transformation leaves invariant the set of solutions of (1). After solving the determining equations, we give the following classification:

- For $f(u)$ arbitrary, the only symmetries admitted by equation (1) are the groups of space and time translations, which are defined by the infinitesimal generators

  $$V_1 = \frac{\partial}{\partial x}, \qquad V_2 = \frac{\partial}{\partial t}.$$

  In this case, we obtain travelling wave reductions,

  $$z = x - \lambda t, \qquad u = h(z),$$

  where $h(z)$, after integrating once with respect to z, satisfies

  $$\alpha\, h'''' + \lambda\beta\, h''' - \varphi_h\, h'' - \varphi_{hh}\, \left(h'\right)^2 - \lambda\nu\, h' = 0. \tag{4}$$

  Equation (4) is invariant under translations, what allows us to reduce the order by one.

- For $\beta = 0$, the only functional forms of $f(u)$, with $f(u) \neq$ const., for which equation (1) has extra symmetries are $f(u) = (au + b)^n$ and $f(u) = de^{au}$, where a, b, d, n are constants, $a \neq 0$, $d \neq 0$. These symmetries were obtained in [6]. For equation (4) it is possible to determine the form of the function f for which we can apply the simplest equation method. This method enables us to obtain exact travelling wave solutions for this class of nonlinear partial differential equations containing polynomial nonlinearities.

 ISBN: 978-84-617-8694-7

2 Conservation laws

We apply the direct method of the multipliers to set up the standard determining equations for finding the conservation laws admitted by equation (1) [1, 2, 3, 4, 5]. We find all multipliers by solving the determining equation

$$\frac{\delta}{\delta u}\Big((\nu u_t - \Delta(\varphi(u) + \alpha\Delta u - \beta u_t))Q\Big) = 0. \tag{5}$$

The nontrivial conservation laws are characterized by the multiplier Q satisfying

$$\hat{E}[u]\left(Q(\nu u_t - \Delta(\varphi(u) + \alpha\Delta u - \beta u_t))\right) = 0,$$

where

$$\hat{E}[u] := \frac{\partial}{\partial u} - D_t\frac{\partial}{\partial u_t} - D_x\frac{\partial}{\partial u_x} + D_x^2\frac{\partial}{\partial u_{xx}} + \ldots.$$

The conservation laws can be written

$$D_t T + D_x X = 0. \tag{6}$$

Given a multiplier Q, we can obtain the conserved density using a standard method

$$T = \int_0^1 dQ\ uQ(x, t, Qu, Qu_x, Qu_{xx}, ...)$$

and the flux X by

$$\begin{aligned} X &= -D_x^{-1}(Q(\nu u_t - \Delta(\varphi(u) + \alpha\Delta u - \beta u_t)) - \frac{\partial T}{\partial u_x}(\nu u_t - \Delta(\varphi(u) + \alpha\Delta u - \beta u_t) \\ &\quad +(\nu u_t - \Delta(\varphi(u) + \alpha\Delta u - \beta u_t)D_x\left(\frac{\partial T}{\partial u_{xx}}\right) + \ldots. \end{aligned}$$

From conservation laws, by using the corresponding conserved (potential) systems we can obtain potential symmetries of equation (1).

Acknowledgements

MS Bruzón, T.M. Garrido and E. Recio express their sincere thanks to the Plan Propio de Investigación de la Universidad de Cádiz and A.P. Márquez expresses its sincere gratitude to the Vicerrectorado de Alumnos de la Universidad de Cádiz.

References

[1] Anco S.C., Bluman G., *Direct construction of conservation laws from field equations.* Phys. Rev. Lett. **78** (1997) 2869–2873.

[2] S.C. Anco and G. Bluman, *Direct constrution method for conservation laws of partial differential equations Part I: Examples of conservation law classifications.* Euro. Jnl of Appl. Math. **13** (2002) 545–566.

[3] S.C. Anco and G. Bluman, *Direct constrution method for conservation laws of partial differential equations Part II: General treatment.* Euro. Jnl of Appl. Math. **13** (2002) 567–585.

[4] S.C. Anco, *Generalization of Noether's theorem in modern form to non-variational partial differential equations.* To appear in Fields Institute Communications: Recent progress and Modern Challenges in Applied Mathematics, Modeling and Computational Science, arXiv: mathph/1605.08734 (2016).

[5] Bluman G.W., Cheviakov A., Anco S.C., *Applications of symmetry methods to partial differential equations.* New York: Springer, 2009.

[6] M. S. Bruzón and M. Gandarias, *Symmetry reductions for a disipation-modified KdV equation*, Appl. Math. Lett. **16** (2003) 155–159.

[7] J. W. Cahn and J. E. Hilliard, *Free Energy of a Nonuniform System. I. Interfacial Free Energy*, J. Chem. Phys. **28** (1958) 258–267.

[8] B. Le Trong Thanh, F. Smarrazzo and A. Tesei, *Passage to the limit over small parameters in the viscous CahnHilliard equation*, J. Math. Anal. Appl. **420** (2014) 1265–1300.

[9] A. Novick–Cohn and L. A. Segel, *Nonlinear aspects of the Cahn-Hilliard equation*, J. Physica D **10** (1984) 277–298.

[10] Olver P.J., *Applications of Lie groups to differential equations.* 2nd ed., Graduate Texts in Mathematics, Springer–Verlag, Berlin (1993).

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

A mathematical model for a diseased orange tree

Iulia Martina Bulai[1], Ana Cristina Esteves[2] and Ezio Venturino[3]

[1] *Department of Information Engineering, University of Padova*

[2] *Department of Biology, CESAM, University of Aveiro*

[3] *Department of Mathematics "Giuseppe Peano, University of Torino*

emails: iuliam@live.it, acesteves@ua.pt, ezio.venturino@unito.it

Abstract

In this paper a three dimensional non linear ODE system is analysed. The model describes the interaction between the orange tree and two different microorganisms present on it: a pathogen fungus *Guignardia citricarpa* and a beneficial one *Trichoderma harzianum T1A*. Five different equilibrium points are found plus the trivial one. The feasibility conditions are needed only for the coexistence equilibrium. Their stability is analysed, only the pathogen fungus-free point and the coexistence equilibrium point are conditionally asymptotically stable while the remaining equilibrium points are unstable.

Key words: orange tree, Guignardia citricarpa, Trichoderma harzianum T1A, mathematical model

1 Introduction

We consider a mathematical model that describes the behaviour of a biological system composed by an orange tree, a pathogen fungus, *Guignardia citricarpa*, and a beneficial one, *Trichoderma harzianum T1A*. We assume that the orange tree is attacked by the pathogen fungus, this attack can be less disastrous if there is a beneficial fungus that secretes proteins related to the control of *G. citricarpa*. Furthermore we assume that the beneficial fungus has a good influence on the reduction of damage of the tree due to the pathogen fungus, [1].

 ISBN: 978-84-617-8694-7

2 The model

Let us denote the various variables considered in the model by

- O: Orange tree fruits
- P: Pathogen fungus,*G. citricarpa*
- F: Beneficial fungus, *T. harzianum T1A*

the three populations of the mathematical model.
The model reads:

$$\begin{aligned} \frac{dO}{dt} &= rO\left(1-\frac{O}{K}\right) - h(F)OP, \quad h(F) = \frac{1}{q+F} \\ \frac{dP}{dt} &= eh(F)OP - aPF \\ \frac{dF}{dt} &= baPF + sF\left(1-\frac{F}{H}\right) \end{aligned} \tag{1}$$

First equation: describes the evolution of the orange tree. It grows logistically with net reproduction rate r and carrying capacity K, and is negatively affected by the pathogen fungus, with a variable rate $h(F)$. We take $h(F)$ so that it is a decreasing function depending on F, this meaning that the beneficial fungus helps to reduce the damage of the tree, O, due to P. The more F there are on the tree the less will be the damage due to P. In fact *T. harzianum T1A* secrets proteins related to the control of *G. citricarpa*.

Second equation: we describe the evolution of the pathogen fungus. P feed on the tree parts, ($e < 1$), and are degraded by the good fungi, at rate a, in the sense that *Trichoderma*, F, is able to produce extracellular enzymes able to degrade the cell wall of the pathogenic fungi, P.

Third equation: the beneficial fungus gets food from degrading the bad fungi ($b < 1$) and we assume that it has other food resources for which in their presence it can grow logistically with net reproduction rate s and carrying capacity H also in the absence of the bad fungi.

3 The qualitative analysis of the model

Solving the system

 ISBN: 978-84-617-8694-7

$$\begin{cases} rO\left(1-\dfrac{O}{K}\right)-\dfrac{OP}{q+F}=0 \\ e\dfrac{OP}{q+F}-aPF=0 \\ baPF+sF\left(1-\dfrac{F}{H}\right)=0 \end{cases} \tag{2}$$

we get five different equilibrium points plus the trivial one:

- the trivial equilibrium point, $E_0=(0,0,0)$, is always feasible
- the pathogen fungus-beneficial fungus-free point, i.e. the healthy fruit-only equilibrium $E_1=(K,0,0)$ is always feasible
- the orange-beneficial fungus-free point $E_2=(0,P,0)$ is always feasible
- the orange-pathogen fungus-free point $E_3=(0,0,H)$ is always feasible
- the pathogen fungus-free point $E_4=(K,0,H)$ is always feasible
- the coexistence equilibrium point $E_*=(O^*,P^*,F^*)$ is feasible if $O^*>0$ and $P^*>0$

To study the stability of the equilibrium points we compute the characteristic polynomial evaluating each equilibrium point at the Jacobian of (1):

$$J=\begin{bmatrix} r-\dfrac{2rO}{K}-\dfrac{P}{q+F} & -\dfrac{O}{q+F} & \dfrac{OP}{(q+F)^2} \\ \dfrac{eP}{q+F} & \dfrac{eO}{q+F}-aF & -\dfrac{eOP}{(q+F)^2}-aP \\ 0 & baF & baP+s-\dfrac{2sF}{H} \end{bmatrix} \tag{3}$$

Proposition 1 The trivial equilibrium point, $E_0(0,0,0)$ as well as $E_1=(K,0,0)$, $E_2=(0,P,0)$ and $E_3=(0,0,H)$ are unconditionally unstable.

Proof. We evaluate E_0, E_1, E_2 and E_3 at the Jacobian matrix (3), and we get

- E_0 is unstable, in fact the characteristic polynomial is

$$\lambda(\lambda-s)(\lambda-r)=0$$

and has two positive eigenvalues $\lambda_1=s$ and $\lambda_2=r$ plus an eigenvalue equal to zero.

 ISBN: 978-84-617-8694-7

- E_1 is unstable, in fact the characteristic polynomial is

$$(\lambda - s)(\lambda + r)\left(\lambda - \frac{Ke}{q}\right) = 0$$

and has two positive eigenvalues $\lambda_1 = s$ and $\lambda_2 = Keq^{-1}$ plus a negative eigenvalue $\lambda_1 = -r$.

- E_2 is unstable, in fact the characteristic polynomial is

$$\lambda(\lambda - Pab - s)\left(\lambda + \frac{qr - P}{q}\right) = 0$$

and has one positive eigenvalue $\lambda_1 = Pab + s$, an eigenvalue equal to zero and one with indefinite sign $\lambda_3 = (qr - P)q^{-1}$.

- E_3 is unstable, in fact the characteristic polynomial is

$$(\lambda + s)(\lambda - r)(\lambda + Ha) = 0$$

and has one positive eigenvalue $\lambda_1 = Pab + s$, and two negative ones $\lambda_2 = -s$ and $\lambda_3 = -Ha$, respectively.

□

Proposition 2 The pathogen fungus-free point $E_4 = (K, 0, H)$ is stable if the condition

$$K < \frac{aH(q + H)}{e} \tag{4}$$

holds.

Proof. The characteristic polynomial associated to E_4 is

$$(\lambda + r)(\lambda + s)\left(\lambda + \frac{aH(q + H) - eK}{q + H}\right) = 0,$$

thus we have two negative eigenvalues $\lambda_1 = -r$ and $\lambda_2 = -s$. Requiring the negativity of

$$\lambda_3 = \frac{eK - aH(q + H)}{q + H},$$

we get condition (4). □

 ISBN: 978-84-617-8694-7

Proposition 3 There exists at least one feasible coexistence equilibrium $E_* = (O^*, P^*, F^*)$ if the following conditions hold:

$$\frac{s - Hbar}{s + barq} < \frac{H}{F} < 1 \tag{5}$$

and furthermore whenever it exists, it is stable if the Routh-Hurwitz conditions hold $R_3 > 0$ and $R_1 R_2 > R_3$ where these quantities are defined below in (13).

Proof. To get O_*, P_* and F_* we solve the system

$$\begin{cases} Krq + KrF - rqO - rFO - KP = 0 \\ eO - aqF - aF^2 = 0 \\ HbaP + Hs - sF = 0 \end{cases} \tag{6}$$

From the third equation of (6) we get

$$P_* = \frac{sF_* - Hs}{Hba} \tag{7}$$

and we substitute it in the first equation of (6) and we get

$$O_* = \frac{HKbarq + HKs + F_*(HKbar - Ks)}{Hba(rq + eF_*)} \tag{8}$$

where F_* is the real root of the third degree polynomial

$$\Psi(F) := AF^3 + BF^2 + CF + D = 0, \tag{9}$$

with

$$\begin{aligned} A &= Ha^2br > 0 \\ B &= 2Ha^2bqr > 0 \\ C &= Ha^2bq^2r - HKaber + Kes \\ D &= -(HKabeqr + sHKe) < 0. \end{aligned} \tag{10}$$

One can see that (9) has at least one positive root, in fact the limit for $F \to +\infty$ is $+\infty$ and $\Psi(0) = D < 0$ (i.e. (9) evaluated for $F = 0$). The continuity of this polynomial function $\Psi(F)$ ensures the existence of a positive root, then.

Furthermore to study the stability of the coexistence equilibrium we compute the characteristic polynomial associated at it evaluating E_* at (3) we get

$$J = \begin{bmatrix} -\frac{rO}{K} & -\frac{O}{q+F} & \frac{OP}{(q+F)^2} \\ \frac{eP}{q+F} & 0 & -\frac{eOP}{(q+F)^2} - aP \\ 0 & baF & -\frac{sF}{H} \end{bmatrix} \tag{11}$$

 ISBN: 978-84-617-8694-7

thus

$$det(J-\lambda) = \lambda^3 + R_1\lambda^2 + R_2\lambda + R_3 = 0 \tag{12}$$

with

$$\begin{aligned} R_1 &= \frac{HrO + KsF}{KH} > 0 \\ R_2 &= \frac{(q+F)^2 rsOF + KHbaeOPF + KHba^2FP + KHeOP}{KH(q+F)^2} > 0 \\ R_3 &= \frac{HbaerO^2PF + H(q+F)^2ba^2rOPF + KesOPF - HKbaeOP^2F}{HK(q+F)^2} \end{aligned} \tag{13}$$

As one can see the condition $R_1 > 0$ is satisfied, while the positivity of R_3 and $R_1R_2 > R_3$ must be requested for the stability of E_*.

□

4 Conclusions

A three dimensional mathematical model was introduced, that describes the evolution in time of an orange tree, O, and two different fungi *Guignardia citricarpa*, P, and *Trichoderma harzianum T1A*, F. Five equilibrium points are always feasible, while for the coexistence equilibrium feasibility conditions needs to be required. Between six equilibrium points only the pathogen fungus-free point and the coexistence equilibrium are conditionally asymptotically stable and the remaining four are unstable.

The mathematical analysis is instrumental in defining strategies for the biological control of this infestant. In the future we plan to devise policies to possibly add the beneficial fungus to orange cultures in order to fight the *Guignardia citricarpa* damages to these fruits and reduce economic losses for the farmers.

References

[1] de Lima, FB; Feix, C; Osoio, N; Alves, A; Vitorino, R; Domingues, P; Ribeiro, RTD; Esteves, AC, *Trichoderma harzianum T1A constitutively secretes proteins involved in the biological control of Guignardia citricarpa*, Biol. Control **17** (2017) 99-109.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

Spectral preconditioners for the efficient numerical solution of sequences of linear systems

Ángeles Martínez[1], Luca Bergamaschi[2], Enrico Facca[1] and Mario Putti[1]

[1] *Department of Mathematics – "Tullio Levi-Civita", University of Padova*

[2] *Department of Civil, Environmental and Architectural Engineering, University of Padova*

emails: acalomar@math.unipd.it, luca.bergamaschi@unipd.it, facca@math.unipd.it, putti@math.unipd.it

Abstract

We consider the problem of finding an appropriate preconditioner for the PCG method in the solution of sequences of SPD linear systems. We investigate several preconditioning strategies that incorporate partial approximated spectral information. In our approach, a number of approximated eigenvectors are computed for a given coefficient matrix in the sequence of linear systems to be solved and used to obtain an efficient preconditioner for the subsequents sytems in the sequence. We apply these techniques to the efficient numerical solution of a branched transport model whose long time solution for specific parameter settings is equivalent to the solution of the Monge-Kantorovich equations of optimal transport [7]. Galerkin FEM discretization combined with explicit Euler time stepping yield a linear system to be solved at each time step, characterized by a large sparse very ill conditioned symmetric matrix A. Extreme cases even prevent the convergence of PCG with standard preconditioners such as an IC (with partial fill-in) factorization of A, which can not always be computed. We present numerical evidence that the proposed techniques are efficient in reducing the condition number of the preconditioned systems, thus decreasing the number of PCG iterations and the CPU time.

Key words: linear systems solving, PCG method, preconditioning, optimal transport.

1 Introduction

Sequences of linear systems in which the coefficient matrix slightly change are very common in many applications such as those requiring the discretization of a time-dependent PDE or

the solution by Newton's method of a nonlinear set of equations. We consider the iterative solution of a sequence of large and sparse (and possibly ill-conditioned) SPD matrices by the Preconditioned Conjugate Gradient (PCG) method. Finding a good preconditioner is mandatory to assure convergence of iterative methods. In the solution of sequences of linear systems like

$$A\boldsymbol{x}_i = \boldsymbol{b}_i$$

a number of paper have investigated the benefits of using spectral information from matrix A to accelerate the iterative solution of such systems by deflation. Among the others we mention [13, 9] and also [14]. Another approach consists in using the partial spectral information to update a given preconditioner (such as an incomplete Cholesky factorization) computed for matrix A. This approach has been described in several papers such as [5, 6, 10]. In all these papers the authors start with an initial preconditioner P_0 and use an approximation of a few eigenvectors of the preconditioned matrix to update P_0 with a low-rank matrix. Another characteristic shared by all these previous papers is that the coefficient matrix of the linear systems to be solved remains unchanged throughout the whole sequence. This allow the incremental refinement of the set of eigenvectors used to update the low-rank correction matrix.

In a slightly different framework, i.e. acceleration of the linear systems within iterative inner-outer eigensolvers, a number of updating strategies that incorporate partial spectral information of A, have been proposed and discussed in [2, 12].

In this paper we consider sequences of linear systems in which the coefficient matrices slightly vary from one system to another. We propose a preconditioning technique based on the knowledge of a set of (roughly) approximated eigenvectors of the coefficient matrix (or the preconditioned matrix) at a given linear system in the sequence. This information is used to update a given initial preconditioner to produce an efficient preconditioner for the subsequent systems in the sequence. We will test our approach to the efficient numerical solution of a branched transport model whose long time solution for specific parameter settings is equivalent to the solution of the Monge-Kantorovich equations of optimal transport [7]. Galerkin FEM discretization combined with explicit Euler time stepping yield a linear system to be solved at each time step, characterized by a large sparse very ill conditioned symmetric matrix A. Extreme cases even prevent the convergence of PCG with standard preconditioners such as an IC (with partial fill-in) factorization of A, which can not always be computed.

We present numerical evidence that the proposed techniques are efficient in reducing the condition number of the preconditioned systems, thus decreasing the number of PCG iterations and the CPU time.

The remaining of the paper is organized as follows: Section 2 introduces the spectral preconditioner and two different strategies to obtain an approximated set of eigenvectors needed to form the low rank updating matrix. Section 3 describes the implementation details

 ISBN: 978-84-617-8694-7

regarding the construction of the preconditioner. We also include in this section detailed algorithms of the iterative solution phase by the PCG method for the two different spectral information recovering techniques. Numerical results are shown in Section 4. Section 5 summarizes the main conclusions.

2 The spectral preconditioner

In this paper we consider sequences of linear systems of the form

$$A_i \boldsymbol{x}_i = \boldsymbol{b}_i. \tag{1}$$

For a given linear system in the sequence $Ax = b$, we study the acceleration of the PCG solver provided by the following spectral preconditioner:

$$P = P_0 + V_p \Lambda_p^{-1} V_p^T, \tag{2}$$

where $V_p = [v_1, \ldots, v_p]$ and $v_j, j = 1, \ldots, p$ are approximate eigenvectors either of $P_0 A$ or of A; $\Lambda_p = \mathrm{diag}(\lambda_1, \ldots, \lambda_p)$, and $\lambda_j,\ j = 1, \ldots, p$ are the corresponding smallest eigenvalues, and P_0 is a standard preconditioner computed for A. When V_p contains eigenvectors of $P_0 A$, the effect of the low-rank correction is easily shown to be:

$$PAv_j = (\lambda_j + 1)v_j, \qquad j = 1, \ldots, m.$$

so that some of the eigenvalues of the new preconditioned matrix are incremented by 1 with an obvious reduction of the condition number.

We propose two different ways to obtain the approximated eigenvectors needed to construct the spectral preconditioner.

2.1 Approximating some of the smallest eigenpairs by DAGC

Following [2], we propose to approximate some of the leftmost eigenvectors of a given coefficient matrix A by performing some preliminary iterations of an eigenvalue solver. We chose the Deflation-Accelerated Conjugate Gradient (DACG) eigensolver [1], which is based on the preconditioned conjugate gradient (nonlinear) minimization of the Rayleigh Quotient (RQ) $q(\boldsymbol{x}) = \dfrac{\boldsymbol{x}^T A \boldsymbol{x}}{\boldsymbol{x}^T \boldsymbol{x}}$. The leftmost eigenpairs are computed sequentially, by minimizing the RQ over a subspace orthogonal to the previously computed eigenvectors. This method, which applies only to symmetric positive definite matrices, has been proven very efficient in the solution of eigenproblems arising from discretization of PDEs in [4]. DACG also proved very suited to parallel implementation as documented in [3] where an efficient parallel matrix vector product has been employed.

 ISBN: 978-84-617-8694-7

Convergence of DACG is strictly related to the relative separation between consecutive eigenvalues, namely $\xi_j = \dfrac{\lambda_j}{\lambda_{j+1} - \lambda_j}$. Also DACG takes advantage of preconditioning which can be chosen to be the IC factorization.

Once a small number of leftmost eigenvectors has been computed and stored as columns of V_p, different low-rank corrections of a given preconditioner P_0 can be defined as e.g. described in [12]. For example a BFGS-style preconditioner can be written as

$$\begin{aligned} P &= V_p(V_p^T A V_p)^{-1} V_p^T + \left(I - V_p(V_p^T A V_p)^{-1} V_p^T A\right) P_0 \left(I - A V_p(V_p^T A V_p)^{-1} V_p^T\right) \\ &\approx V_p \Lambda_p^{-1} V_p^T + \left(I - V_p V_p^T\right) P_0 \left(I - V_p V_p^T\right) \end{aligned} \tag{3}$$

In the previous paper, however, a simplified version of the BFGS preconditioner (3) has been defined, which neglects the left and right projector on P_0 and therefore takes the same form as (2):

$$P = V_p \Lambda_p^{-1} V_p^T + P_0.$$

In the above mentioned paper it is also shown that the preconditioned matrix PA has a better spectral distribution than $P_0 A$.

2.2 Recovering spectral information by the Lanczos process

Another strategy to obtain a set of approximated eigenvectors is to recover them from the Krylov subspace built by the linear solver. Denoting again by P_0 a preconditioner for matrix A, during the PCG method we save the first m preconditioned residuals as columns of a matrix W_m:

$$W_m = \left[\frac{P_0 r_0}{\sqrt{r_0^T P_0 r_0}}, \frac{P_0 r_1}{\sqrt{r_1^T P_0 r_1}}, \ldots, \frac{P_0 r_{m-1}}{\sqrt{r_{m-1}^T P_0 r_{m-1}}.} \right]$$

Matrix W_m thus satisfies: $W_m^T P_0^{-1} W_m = I_m$, in view of the P_0-orthogonality of the residuals generated by the PCG method. Moreover we form the Lanczos tridiagonal matrix using the PCG coefficients α_k, β_k as

$$T_m = \begin{bmatrix} \frac{1}{\alpha_0} & -\frac{\sqrt{\beta_1}}{\alpha_0} & & & \\ -\frac{\sqrt{\beta_1}}{\alpha_0} & \frac{1}{\alpha_1} + \frac{\beta_1}{\alpha_0} & -\frac{\sqrt{\beta_2}}{\alpha_1} & & \\ & & \ddots & & \\ & & & -\frac{\sqrt{\beta_{m-1}}}{\alpha_{m-2}} & \\ & & -\frac{\sqrt{\beta_{m-1}}}{\alpha_{m-2}} & \frac{1}{\alpha_{m-1}} + \frac{\beta_{m-1}}{\alpha_{m-2}} & \end{bmatrix}$$

 ISBN: 978-84-617-8694-7

Matrices W_m and T_m obey to the classical Lanczos relation i.e.: $W_m^T A W_m = T_m$. After eigensolving T_m we obtain $T_m = Q\Lambda_m Q^T$, where the diagonal coefficients of the diagonal matrix Λ_m approximate the eigenvalues of $P_0 A$ while the columns of $V_p = W_m Q_q$ (where Q_p contains the first p columns of Q) are approximations of the p leftmost eigenvectors of $P_0 A$.

In fact we have, first, that $V_p^T A V = Q_p^T W_m^T A W_m Q_p = Q_p^T T_m Q_p = \Lambda_p \equiv \mathrm{diag}(\lambda_1, \ldots, \lambda_p)$. Then, by setting $U = P_0^{-1/2} V_p$ we obtain U and Λ_p satisfying

$$U^T U = V_p^T P_0^{-1} V_p = I_m \tag{4}$$

$$\Lambda_p = V_m^T A V_m = U^T P_0^{1/2} A P_0^{1/2} U \tag{5}$$

which correspond to the Lanczos process applied to matrix $P_0^{1/2} A P_0^{1/2}$. Hence the columns of U approximate the eigenvectors of $P_0^{1/2} A P_0^{1/2}$ that is

$$P_0^{1/2} A P_0^{1/2} U \approx U\Lambda_p \iff P_0 A P_0^{1/2} U \approx P_0^{1/2} U \Lambda_p \Longleftrightarrow P_0 A V_p \approx V_p \Lambda_p,$$

so that the columns of V_p approximate the eigenvectors of $P_0 A$.

3 Implementation

The approximation of a number of the leftmost eigenpairs is a costly task which can not be performed at each linear system solution. To reduce the cost of this operation we devise different strategies depending on how we obtain the spectral information.

3.1 Eigenpairs of A obtained by DACG

The computation of a number of the leftmost eigenpairs by DACG is a preprocessing stage that in principle should be executed before every system solving. However, in view of the fact that the system matrices A_i do not change much during the simulation we propose to evaluate selectively the eigenpairs, whenever the PCG solution of a generic linear system $A_j \boldsymbol{x}_j = \boldsymbol{b}_j$ takes more than a fixed number of iterations ($k_i \geq \mathtt{it}_{switch}$). In this case, except for the very first linear system, we use as the initial DACG eigenvector guess, the previously computed eigenvectors.

Spectral PCG Algorithm: DACG variant

- INPUT: $it_{prec}, it_{chol}, p, \tau_{DACG}$.
- Set chol_switch = TRUE; switch = TRUE;
- FOR $i = 1$ TO n_sys

 ISBN: 978-84-617-8694-7

- IF chol_switch THEN
  compute $P_0 = IC(A_i)$; set chol_switch = FALSE;
- IF switch THEN
  1. Compute the p leftmost eigenpairs by the DACG procedure with preconditioner P_0 and accuracy τ_{DACG}.
  2. Form matrices V_p, Λ_p.
  3. Solve the $i-$th linear system by PCG preconditioned by $P_0 + V_p \Lambda_p^{-1} V_p^T$.
  4. switch= FALSE.
- IF $k_i > it_{prec}$ switch = TRUE
- IF $k_i > it_{chol}$ chol_switch = TRUE

END FOR

3.2 Eigenpairs of $P_0 A$ obtained by Lanczos-PCG

Computation of matrices T_m and W_m is carried out during the PCG process and hence it needs no additional cost. The main computational burden due to this strategy is given by the matrix-matrix product $V_m = W_m Q_p$ which is implemented via BLAS-3 subroutines, with a consequent optimal use of memory accesses. Due to the slow convergence of the Lanczos process to the smallest eigenvalues, and also for memory reasons, it is convenient to recover a relatively small number of eigenpairs (independently of the size m of V_m which should be taken sufficiently large). In the Lanczos process we use only the $p = \{10, 20\}$ smallest eigenvalues and corresponding eigenvectors thus obtaining a $n \times p$ matrix V_p and a $p \times p$ diagonal matrix Λ_p.

Spectral PCG Algorithm: Lanczos variant

- INPUT: $it_{prec}, it_{chol}, m_{\max}, p$.
- Set chol_switch = TRUE; switch = TRUE;
- FOR $i = 1$ TO n_sys
  - IF chol_switch THEN
    compute $P_0 = IC(A_i)$; set chol_switch = FALSE;
  - IF switch THEN
    1. Solve the $i-$th linear system by the PCG method preconditioned by P_0.
    2. Construct the tridiagonal Lanczos matrix T_m, with $m = \min\{m_{\max}, k_i\}$.
    3. Extract from T_m and W_m the p smallest eigenpairs and form matrices V_p, Λ_p.
    4. switch= FALSE.

 ISBN: 978-84-617-8694-7

- ELSE
  1. Solve the $i-$th linear system by PCG preconditioned by $P_0 + V_p \Lambda_p^{-1} V_p^T$.
- IF $k_i > it_{prec}$ switch = TRUE
- IF $k_i > it_{chol}$ chol_switch = TRUE

END FOR

4 Numerical results

In this section we illustrate the behaviour of the spectral preconditioner on a sequence of linear systems arising in the discretization of the PDE modeling the optimal transport (OT) problem. Starting from the PDE-version of the L^1 Monge-Kantorovich OT problem formulated by [7], Facca et al. in [8] have proposed a general formulation of a dynamical system whose infinite-time solution is conjecture to be related to the OT and BT problems. The model reads as: find (μ, u) such that:

$$-\nabla \cdot (\mu(t,x)\nabla u(t,x)) = f(x) = f^+(x) - f^-(x) \tag{6}$$

$$\mu'(t,x) = |\mu(t,x)\nabla u(t,x)|^\beta - \mu(t,x) \tag{7}$$

$$\mu(0,x) = \mu_0(x) > 0 \tag{8}$$

where μ is the transport density, u is the transport potential, f is the zero-average forcing function (f^+ and f^- being the initial and final mass configurations), and $\beta > 0$ is related to the branching exponent. The model is completed by imposing homogeneous Neumann boundary conditions to 6). Here, the gradient and divergence operators are computed with respect to the spatial coordinate x, while μ' indicates time differentiation.

Experimental evidence shows that for $\beta > 1$ the proposed formulation reaches and equilibrium state that resembles solutions of the branched transport problem. In this case, in fact, the transport density concentrates on sets that have a fractal nature, thus providing a naturally network-forming model of optimal transportation.

Numerically, the problem is solved by a combination of Galerkin FEM and Euler time stepping. Spatial discretization is obtained by employing two different triangulations $\mathcal{T}_h$ and $\mathcal{T}_{h/2}$, the second one obtained from the first by uniform refinement, we adopt a piecewise constant ($\mathcal{P}_0(\mathcal{T}_h)$) representation of μ_h on $\mathcal{T}_h$ and a linear ($\mathcal{P}_1(\mathcal{T}_{h/2})$) representation of u_h on $\mathcal{T}_{h/2}$. Forward Euler time-stepping completes the numerical approach, which reads:

$$A[\mu^k]u^k = b \tag{9}$$

$$\mu^{k+1} = (I + \Delta t^k B[u^k])\mu^k \tag{10}$$

$$\mu^0 = \Pi_h \mu_0 \tag{11}$$

 ISBN: 978-84-617-8694-7

where A is the classical Galerkin $\mathcal{P}_1$ stiffness matrix, vector μ collects the elemental values of μ_h, u is the vector of nodal values of u_h, B is the matrix defining the norm of the gradient of u_h, and Π_h denotes the L^2 projector on $\mathcal{T}_h$. Equilibrium is achieved by repeating the above algorithm until $(\|\mu_h^{k+1} - \mu_h^k\|_{L^2(\Omega)})/\Delta t^k \|\mu_h^k\|_{L^2(\Omega)} < \tau$. We generally impose $\tau = 10^{-8}$.

At each time step of the method, a linear system involving the sparse symmetric and positive definite system matrix $A[\mu^k]$ needs to be solved. PCG convergence becomes increasingly difficult as time progresses as the condition number of the system matrix grows indefinitely with β. In fact, the dynamics of the model is such that μ_h tends to zero in large portions of Ω.

The test case addressed in this work is shown in Figure 1. The simulation considers a mesh $\mathcal{T}_h$ of 412417 nodes and $n_{nz} = 1647617$ nonzero elements. The timestep Δt^k is increased by a factor 1.05 at each time step starting from $\Delta t^0 = 10^{-3}$ up to a maximum value of $\Delta t^k = 10^{-1}$. This leads to a sequence of 800 linear systems of type (1) to reach equilibrium at the chosen tolerance τ.

As the boundary conditions imposed to the PDE (1)-(3) are only Neumann, the resulting discretized system is singular, with $\ker A_i = \text{span}\{c\} = \text{span}\{\big(1, \ldots, 1\big)^T\}$. However following [11] we guarantee the consistency (and the PCG solution) of our linear systems by also modifying the right hand side: $\tilde{b} = b - \dfrac{c^T b}{\|c\|^2} c$. Moreover, all the linear systems has been symmetrically scaled with the diagonal of A_i in order to reduce the condition number. . .

The code is written in Fortran 90. All the experiments were run on a 2 x Intel Xeon CPU E5645 at 2.40GHz (six core) and with 4GB RAM for each core. Times are expressed in seconds. The stopping criterion for the linear solver is independent of the preconditioner used and it is based on the relative residual. The iterative procedure stops whenever

$$\frac{\left\| A_i \boldsymbol{x}_i^{(k)} - \boldsymbol{b}_i \right\|}{\|\boldsymbol{b}_i\|} < \varepsilon = 10^{-11}. \tag{12}$$

The initial preconditioner P_0 is computed by setting the maximum number of nonzero per row LFIL= 30 and the drop tolerance $\tau_{IC} = 10^{-4}$. Using lower LFIL and/or larger τ_{IC} values prevented the existence of the IC factorization for all systems in the simulations. This choice of parameters produced a rather dense L factor with a number of nonzero roughly 8 times that of the lower triangular part of A. Therefore computation of this preconditioner for each linear system of the sequence was not effective. We decided to compute the IC preconditioner for a given matrix A_i if $i = 1$ or the number of PCG iterations in the previous linear system was above a fixed value, it_{chol}. We used the previously computed IC preconditioner, otherwise.

In the sequel we will report the number of eigenvectors p used to build the low rank correction to the initial preconditioner, the way these eigenvectors have been approximated,

 ISBN: 978-84-617-8694-7

by using the Lanczos method to recover them from the Krylov space built by the linear solver (LAN) or explicitly approximated by the DACG eigensolver, the overall iteration count of the PCG solver for the total number of linear systems to be solved, and timings accounting for computation of the preconditioner (T_{prec}), of the approximated eigenvectors (T_{eig}), of the PCG solver (T_{PCG}) and the total CPU time (T_{tot}).

We first perform a preliminary study on the sensitivity of the preconditioner efficiency versus the accuracy of the eigencomputation. To this end we considered the first 200 linear systems and use three different tolerances for the relative eigenresidual test: $\tau_{DACG} \in \{0.1, 0.3, 0.5\}$. Other parameters were: $it_{prec} = 60, it_{chol} = 60, p = 20$. As a benchmark, we also solved the first 200 systems by the PCG method preconditioned by an IC factorization computed selectively (with $it_{chol} = 100$).

τ_{DACG}	ITER	T_{eig}	T_{prec}	T_{PCG}	T_{tot}
–	20646	0.00	187.9	1687.8	1875.7
0.1	9907	326.9	117.9	1002.1	1446.2
0.3	10006	198.5	117.2	1011.5	1327.8
0.5	10055	150.0	117.2	1017.7	1284.9

These results show that there is no need for high accuracy in the computation of the eigenvectors. With a very low accuracy ($\tau_{DACG} = 0.5$) the number of PCG iterations is halved and the CPU time reduced of a factor 1.5 with respect to the fixed IC preconditioner (first row in the previous Table).

We report in Table 1 the results of a complete simulation i.e. the cumulative number of PCG iterations and CPU times in solving the 800 linear systems. In addition to the previously described parameters we used as the maximum size of the Lanczos subspace $m_{\max} = 80$, which experimentally revealed the optimal value. We set $\tau_{DACG} = 0.5$.

We notice from Table 1 that the proposed low-rank updated spectral preconditioner is effective in both variants, providing an important reduction of the number of iterations as well of the CPU time. On the average, our spectral preconditioners provide a halving of the total CPU time and a $30\% - 40\%$ reduction of the number of iterations. Using $p = 10$ or $p = 20$ eigenvectors produces only slight variations in the number of iterations/CPU time. Hence, the choice $p = 10$ seems to be preferred in terms of memory storage.

Surprisingly, the DACG variant, though the eigenvector approximation is done outside the PCG and hence is expected to be more costly, reveals as effective as the Lanczos variant. This is mainly due to the fact that after an initial and costly assessing of the leftmost eigenvectors, the subsequent computations are very cheap since the previously computed eigenvectors are very good initial guesses for the next systems. However, we may expect a different behavior of the two techniques in cases of higher variations of the matrices involved. In this case, the DACG preprocessing time will increase as opposite to the Lanczos

 ISBN: 978-84-617-8694-7

Prec. (p)	s_1	s_2	ITER	T_{eig}	T_{prec}	T_{PCG}	T_{tot}
IC	–	–	64248	0.0	2841.2	5432.5	8273.7
IC	100	–	74511	0.0	447.4	6062.9	6510.3
LAN(10)	–	60	41148	29.6	2814.2	3372.1	6215.9
LAN(10)	50	70	44765	30.9	1767.2	3746.4	5544.5
LAN(10)	60	60	44041	196.0	572.7	3606.9	4375.6
LAN(20)	60	60	41738	190.4	459.2	3775.8	4425.4
DACG(10)	60	60	45502	185.4	516.0	3811.6	4512.9
DACG(20)	60	60	42050	263.5	272.2	3922.0	4457.7

Table 1: *Timings and total PCG iterations of the complete simulation for various combinations of parameters.* $s_1 = it_{\text{chol}}, s_2 = it_{\text{switch}}$.

technique. Moreover, the Lanczos approach can be accelerated by employing a method similar to that described in [14]. This aspect will be investigated in a future work.

5 Conclusions

We have proposed a class of spectral preconditioners to accelerate the PCG solution of a sequence of linear systems arising from the discretization of a continuous branched transport model, where Galerkin FEM discretization combined with explicit Euler time stepping yield a linear system to be solved at each time step, characterized by a large sparse very ill conditioned symmetric matrix A. Using the fact that the matrices involved display mild variations at close simulation times, we have used eigeninformation obtained at a given timestep, to accelerate the subsequent linear systems solution. Numerical results reveal that the proposed spectral preconditioner are able to consistently reduce the number of iterations and halve the CPU time with respect to keeping fixed the IC preconditioner. Moreover, the CPU time required by the eigenanalysis is almost negligible as shown in our experiments where it represents less than 5% of the overall CPU time when $p = 10$ approximated eigenvectors are computed.

References

[1] L. Bergamaschi, G. Gambolati, and G. Pini, *Asymptotic convergence of conjugate gradient methods for the partial symmetric eigenproblem*, Numer. Lin. Alg. Appl., 4 (1997), pp. 69–84.

 ISBN: 978-84-617-8694-7

[2] L. Bergamaschi and A. Martínez, *Two-stage spectral preconditioners for iterative eigensolvers*, Numer. Lin. Alg. Appl., 24 (2017), pp. 1–14.

[3] L. Bergamaschi, A. Martínez, and G. Pini, *Parallel Rayleigh Quotient optimization with FSAI-based preconditioning*, J. Applied Mathematics, 2012, Article ID 872901, 14 pages (2012).

[4] L. Bergamaschi and M. Putti, *Numerical comparison of iterative eigensolvers for large sparse symmetric matrices*, Comp. Methods App. Mech. Engrg., 191 (2002), pp. 5233–5247.

[5] B. Carpentieri, I. S. Duff, and L. Giraud, *A class of spectral two-level preconditioners*, SIAM J. Sci. Comput., 25 (2003), pp. 749–765 (electronic).

[6] I. S. Duff, L. Giraud, J. Langou, and E. Martin, *Using spectral low rank preconditioners for large electromagnetic calculations*, Int. J. Numer. Methods Engrg., 62 (2005), pp. 416–434.

[7] L. C. Evans and W. Gangbo, *Differential equations methods for the Monge-Kantorovich mass transfer problem*, Memoirs AMS, 137 (1999).

[8] E. Facca, F. Cardin, and M. Putti, *A continuous model of slime mold dynamics*, SIAM J. Applied. Math., submitted (2017).

[9] J. Frank and C. Vuik, *On the construction of deflation-based preconditioners*, SIAM J. Sci. Comput., 23 (2001), pp. 442–462. Copper Mountain Conference (2000).

[10] L. Giraud, S. Gratton, and E. Martin, *Incremental spectral preconditioners for sequences of linear systems*, Applied Numerical Mathematics, 57 (2007), pp. 1164 – 1180. Numerical Algorithms, Parallelism and Applications (2).

[11] E. F. Kaasschieter, *Preconditioned conjugate gradients for solving singular systems*, J. Comput. Appl. Math., 24 (1988), pp. 265–275.

[12] A. Martínez, *Tuned preconditioners for the eigensolution of large spd matrices arising in engineering problems*, Numer. Lin. Alg. Appl., 23 (2016), pp. 427–443.

[13] Y. Saad, M. Yeung, J. Erhel, and F. Guyomarc'h, *A deflated version of the conjugate gradient algorithm*, SIAM J. Sci. Comput., 21 (2000), pp. 1909–1926. Iterative methods for solving systems of algebraic equations (Copper Mountain, CO, 1998).

[14] A. Stathopoulos and K. Orginos, *Computing and deflating eigenvalues while solving multiple right-hand side linear systems with an application to quantum chromodynamics*, SIAM J. Sci. Comput., 32 (2010), pp. 439–462.

 ISBN: 978-84-617-8694-7

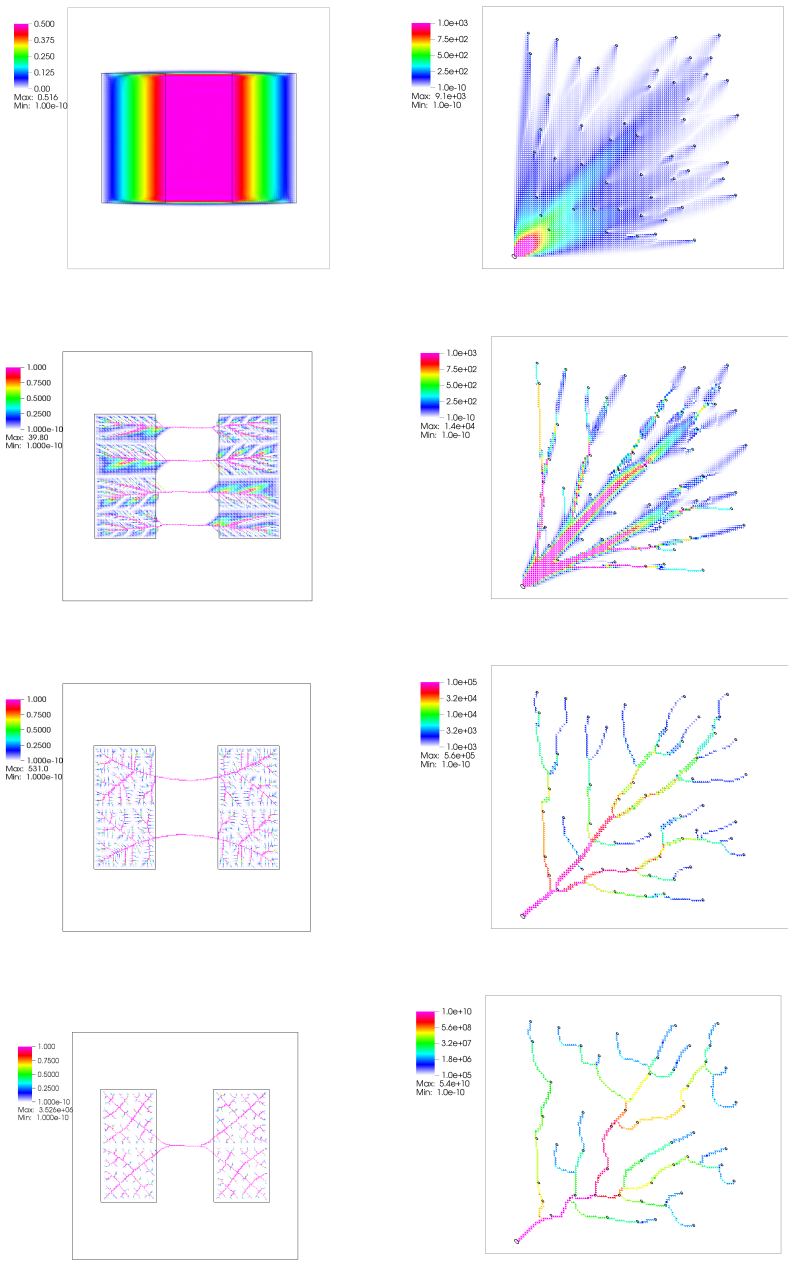


Figure 1: Spatial distribution of μ^* obtained with two different forcing terms. The left column shows a piecewise unitary forcing function, where the black rectangles indicates the supports of f^+ (left) and f^- (right). In right column f^+ is the sum of 30 Dirac sources randomly distributed in the square $[0.1, 0.9] \times [0.1, 0.9]$, while f^- is concentrated in $P = (0.05, 0.05)$ and is the sum of the 30 Dirac sources. Different β are represented by rows, with values ranging from top to bottom from $\beta = 1.0$, corresponds to the L^1 case, $1.05, 1.4$ and 3.0.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

Cyclic codes as function field codes

Cristina Martínez Ramírez[1]

[1] *Hamilton Institute, Maynooth University, Maynooth, Co. Kildare, Ireland*

emails: cmartinez@maths.nuim.ie

Abstract

We study cyclic codes which include generalized Reed-Solomon codes as function field codes. This geometric approach allows to construct longer codes and to get more information on the parameters defining the codes.

Key words: Algebraic code, function field, cyclic code

1 Introduction

Function fields are used ubiquitously in algebraic coding theory for their flexibility in constructions and have produced excellent linear codes. Suitable families of function fields, for example good towers of function fields, have been used to construct families of codes with parameters bound better than the asymptotic bound.

Let q a power of a prime number p. It is well known, that there exists exactly one finite field with q elements which is isomorphic to the splitting field of the polynomial $x^q - x$ over the prime field $\mathbb{F}_p$. Any other field F of characteristic p contains a copy of $\mathbb{F}_p$. We denote respectively by $\mathbb{A}^n(\mathbb{F}_q)$ and $\mathbb{P}^n(\mathbb{F}_q)$ the affine space and the projective space over $\mathbb{F}_q$. Let $\mathbb{F}_q[x_1, x_2, \ldots, x_n]$ be the algebra of polynomials in n variables over $\mathbb{F}_q$.

The encoding of an information word into a $k-$dimensional subspace is usually known as coding for errors and erasures in random network coding, [KK]. Namely, let V be an $N-$dimensional vector space over $\mathbb{F}_q$, a code for an operator channel with ambient space V is simply a nonempty collection of subspaces of V. The collection of subspaces is a code for error correcting errors that happen to data sent through an operator channel. The matrix coding the information is parametrized by random variables $a_1, a_2, \ldots, a_n$ which constitute the letters of an alphabet. Here the operator channel is an abstraction of the operator encountered in random linear network coding, when neither transmitter nor receiver has

 ISBN: 978-84-617-8694-7

knowledge of the channel transfer characteristics. The input and output alphabet for an operator channel is the projective geometry. A good code is capable of correcting error and erasures at the output of the operator channel. Thus in order to construct good codes one need to choose a metric consistent with channel errors and search of a set of vectors with given metric properties as a correcting code. The codes considered here are codes for channels whose errors are consistent with the weighted Hamming metric (WHM).

Let $\mathcal{C}$ be a non-singular, projective, irreducible curve defined over $\mathbb{F}_q$, as the vanishing locus of a polynomial $F \in \mathbb{F}_q[x_0, x_1, x_2]$. We define the number $N(q)$ of $\mathbb{F}_q-$rational points on the curve to be

$$N(q) = |\{(x_0, x_1, x_2) \in \mathbb{P}^2(\mathbb{F}_q) | F(x_0, x_1, x_2) = 0\}|.$$

It is a polynomial in q with integer coefficients, whenever q is a prime power.

The number of points $\overline{\mathcal{C}}(\mathbb{F}_{q^r})$ on $\mathcal{C}$ over the extensions $\mathbb{F}_{q^r}$ of $\mathbb{F}_q$ is encoded in an exponential generating series, called the zeta function of $\overline{\mathcal{C}}$:

$$Z(\mathcal{C}, t) = exp\left(\sum_{r=1}^{\infty} \sharp\overline{\mathcal{C}}(\mathbb{F}_{q^r}) \frac{t^r}{r}\right).$$

Garcia and Stichtenoth analyzed the asymptotic behavior of the number of rational places and the genus in towers of function fields, [GS]. From Garcia-Stichtenoth's second tower one obtains codes over any field $\mathbb{F}_q$ where q is an even power of a prime, [GMMR].

One of the main problems in coding theory is to obtain non-trivial lower bounds of the number $N(F_i)$ of rational places of towers of function fields $\{F_i/\mathbb{F}_q\}_{i=1}^{\infty}$ such that $F_i \subsetneq F_{i+1}$. Suitable families of function fields, for example good towers of function fields, have been used to construct families of codes that beat the Gilbert-Varshamov bound. This paper aims to explore this link for the study and construction of cyclic codes. For example good codes are obtained for curves of genus 0, they are in fact extended generalised Reed-Solomon codes.

2 Algebraic geometric codes

Let $\mathbb{F}_q$ be a finite field of q elements, where q is a power of a prime. We consider as an alphabet a set $\mathcal{P} = \{P_1, \ldots, P_N\}$ of $N - \mathbb{F}_q$ rational points lying on a smooth projective curve $\mathcal{C}$ of genus g and degree d defined over the field $\mathbb{F}_q$. Goppa in [Go], constructed algebraic geometric linear codes from algebraic curves over finite fields with many rational places.

Definition 2.1 *AGC codes are constructed by evaluation of the global sections of a line bundle or a vector bundle on the curve $\mathcal{C}$. Namely, let $F|\mathbb{F}_q$ be the function field of the*

 ISBN: 978-84-617-8694-7

curve and $\mathcal{D}$ a divisor of $F|\mathbb{F}_q$ such that $\operatorname{Supp} \mathrm{G} \cap \operatorname{Supp} \mathrm{D} = \emptyset$*, then the geometric Goppa code associated with the divisors D and G is defined by*

$$\mathbf{C}(D,G) = \{(x(P_1), \ldots, x(P_n)) |\, x \in \mathcal{L}(G)\} \subseteq \mathbb{F}_{q^n}.$$

Recall that $\mathbb{F}_{q^n}|\mathbb{F}_q$ is a cyclic Galois extension and by the Theorem of the primitive element is finitely generated by unique element $\alpha \in \mathbb{F}_q$. In particular, each element in $\mathbb{F}_{q^n}$ can be written as a quotient of two polynomials in α with coefficients in $\mathbb{F}_q$. In the sequel, an $[n,k]_q-$code C is a $k-$dimensional subspace of $(\mathbb{F}_q)^n$.

2.1 Generalized Reed-Solomon codes as cyclic codes

Another important family of Goppa codes is obtained considering the rational normal curve $\mathcal{C}^n$ defined over $\mathbf{F}_q$:

$$\mathcal{C}^n := \{\mathbb{F}_q(1, \alpha, \ldots, \alpha^n) :\ \alpha \in \mathbb{F}_q \cup \{\infty\}\}.$$

The Goppa codes of dimension n defined over $\mathcal{C}^n$ are constructed by evaluating non-zero polynomials of degree less than n over a sequence $\alpha_1, \ldots, \alpha_n$ of n distinct elements in $\mathbb{F}_q$, if $k \leq n$, then the map

$$\epsilon : \mathbb{F}_q[x] \to \mathbb{F}_q^n, \quad f \mapsto (f(\alpha_1, \ldots, \alpha_n)) \tag{1}$$

is injective, since the existence of a non-zero polynomial of degree less than k vanishing on all α_i implies $n < k$ by the fundamental theorem of algebra (a non-zero polynomial of degree r with coefficients in a field can have at most r roots). These are just Reed-Solomon codes of parameters $[n,k,d]$ over a finite field $\mathbb{F}_q$, with parity check polynomial $h(x) = \prod_{i=1}^{q}(x-\alpha^i)$, where α is a primitive root of $\mathbb{F}_q$ such that $\alpha^{k+1} = \alpha + 1$. Any codeword $(c_0, c_1, \ldots, c_{n-1})$ can be expanded into a q-ary k vector with respect to the basis $\{1, \alpha, \ldots, \alpha^{k-1}\}$.

Construction of generalized Reed-Solomon codes over $\mathbb{F}_q$ only employ elements of $\mathbb{F}_q$, hence their lengths are at most $q+1$. In order to get longer codes, one can make use of elements of an extension of $\mathbb{F}_q$, for instance considering subfield subcodes of Reed-Solomon codes. In this way, one gets cyclic codes. Recall that a linear cyclic code is an ideal in the ring $\mathbb{F}_q[x]/(x^n-1)$ generated by a polynomial $g(x)$ with roots in the splitting field $\mathbb{F}_q^l$ of $x^n - 1$, where $n|\, q^l - 1$. We shall identify the code with the set of its codewords.

Theorem 2.2 *Given a set of integers $\{0, 1, \ldots, n-1\}$ module n, there is a set J of k integers which is a set of roots, that is, there is a polynomial $h(x) = \prod_{j \in J}(x - \alpha^j)$, where α is a generator of $(\mathbb{F}_p)^m$ for some prime number p and m is the least integer such that $n|p^m - 1$. The ideal $h(x)$ generates in $\mathbb{F}_{p^m}[x]/(x^n-1)$ is a cyclic linear code of parameters $(n, k, n-k+1)$.*

Another important family of cyclic codes is obtained considering the roots of the polynomial $x^n - 1$ over its splitting field. These codes are of great importance in ADN-computing and as they are linear codes, they can be described as function fields.

 ISBN: 978-84-617-8694-7

Theorem 2.3 *Cyclic codes are function field codes constructed over the curve $\mathcal{C}_{n,m}$ with affine equation $y^m + x^n = 1$ defined over a finite field $\mathbb{F}_q$ of q elements, where q is a power of a prime p and n, m are integer numbers greater or equal than 2.*

Proof. Let us assume n is an integer even number, thus $n = 2^k \cdot s$, with s an integer odd number. We recall that a linear cyclic code is an ideal in the ring $\mathbb{F}_q[x]/(x^n-1)$ generated by a polynomial $g(x)$ with roots in the splitting field $\mathbb{F}_q^l$ of x^n-1, where $n|\, q^l-1$. If we consider the factorisation of the polynomial x^n-1 over $\mathbb{F}_p[x]$, we get $(x^{n/2}-1)\,(x^{n/2}+1) = (x^{n/4}-1)\,(x^{n/4}+1)\,(x^{n/2}+1) = (x^{n/2^k}-1)\,(x^{n/2^k}+1)\,(x^{n/2^{(k-1)}}-1)\,(x^{n/2^{(k-1)}}+1)\ldots\,(x^{n/2}+1)$. We see that the point $P_0 = (\alpha, 0) \in \mathbb{P}(\mathbb{F}_q^2)$ with $\alpha^{n/2} = p-1$ is an $\mathbb{F}_{q^2}-$rational place of the affine curve $y^m = (x^{n/2}+1)$. The other rational places are $P_k = (\beta, 0)$ with $\beta^{n/2^k} = p-1$,..., $P_2 = (\beta^2, 0)$, $P_1 = (1, 0)$, $P_0 = (-1, 0)$ and the place $P_\infty = (0, \alpha)$ at ∞.The cyclic code is realized as the algebraic geometric code associated to the divisors $D = P_0 + P_1 + \ldots + P_k$, $G = \mu\, P_\infty$ and the parameter μ satisfies the bound $\mu > 2g-2$, where g is the genus of the curve $\mathcal{C}_{n,m}$. Note that m is the least integer such that $n|\, p^m - 1$. In particular α is a generator of $(\mathbb{F}_p)^m$.

If n is an integer odd number, by Theorem 2.2, we know there is a set of roots $\{\alpha^j\}_{j\in J}$, such that α is a generator of $(\mathbb{F}_p)^m$. Now we consider the points $P_j = (\alpha^j, 0)$ with $j \in J$ and the point P at $P_\infty = (0, \alpha) = \infty$, and the cyclic code is realised as the function field code associated to the divisors $D = \sum_{j\in J} P_j$ and μP_∞.

Acknowledgments

This research has been partially supported by the COST Action IC1104 and the project ARES (Team for Advanced Research on Information Security and Privacy. Funded by Ministry of Economy and Competitivity).

References

[BM] A. Besana, C. Martínez, *Topological network coding, t−designs, and set partitions*, preprint 2017.

[CMP] A. Couvreur, I. Márquez, R. Pellikaan, *A polynomial Time Attack against Algebraic Geometry Code Based Public Key Cryptosystems*, arXiv:1401.6025

[GS] Arnaldo Garcia and Henning Stichtenoth. *On the asymptotic behaviour of some towers of function fields over finite fields.* Journal of Number Theory, **61** (2):248?273, 1996.

[GMMR] O. Geil, S. Martin, U. Martínez-Peas, D. Ruano, *Refined analysis of RGHWs of code pairs coming from Garcia-Stichtenoth's second tower*, Proceedings of 21^{st} conference on Applications of Computer Algebra.

 ISBN: 978-84-617-8694-7

[Go] Goppa, V. D., *Codes on algebraic curves (Russian)*. Dokl. Akad. Nauk. SSSR **259**, 1289-1290 (1981).

[KK] R. Kötter, F. R. Kschischang, *Coding for Errors and Erasures in Random Network Coding*, IEEE Transactions on information Theory, Vol. **54**, no. 8, 2008.

[NXL] H. Niederreiter, C. Xing, K. Y. Lam, *A new construction of Algebraic-Geometry codes*, AAECC **9**, 373-381 (1999).

[HNX] D. Hachenberger, H. Niederreiter, C. Xing, *Function field codes* AAECC (2008) **19**:201-211.

[MWS] F.J. Mac Williams, N.J.A. Sloane, *The Theory of Error-Correcting Codes*, North Holland.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

A specialized lazy learner for time series forecasting

Francisco Martínez[1], María P. Frías[2], Francisco Charte[1] and Antonio J. Rivera[1]

[1] *Department of Computer Science, University of Jaén (Spain)*

[2] *Department of Statistics and Operations Research, University of Jaén (Spain)*

emails: fmartin@ujaen.es, mpfrias@ujaen.es, fcharte@ujaen.es, arivera@ujaen.es

Abstract

In a time series context the nearest neighbour algorithm looks for the historical observations most similar to the latest observations of the time series. However, some nearest neighbours can be misleading. In this paper we propose that, if prior information about the structure of the time series is known, the search space of possible neighbours can be narrowed so that some possibly misleading neighbours are avoided. This way a more effective forecasting method can be obtained.

Key words: time series forecasting, lazy learners

1 Introduction

Time series forecasting has been traditionally done using statistical models, such as ARIMA [1] or exponential smoothing [2]. However, the last decades have seen the widespread use of computational intelligence techniques, such as artificial neural networks [3], to forecast univariate time series.

Lazy learners [4], such as nearest neighbours, are one of the computational intelligence techniques applied in time series forecasting. As the next section will explain in more detail, a nearest neighbours algorithm tries to forecast the next future values of a time series looking for past realizations that are similar to the last observations in the time series. The subsequent observations of the past similar realizations are used to forecast the time series.

 ISBN: 978-84-617-8694-7

However, we think that in certain time series, some similar past realizations can be misleading and, therefore, produce bad forecasts. To solve this problem, we propose to narrow the search space of possible neighbours, excluding possibly misleading instances.

The remainder of this paper is structured as follows. Section 2 describes how the nearest neighbour algorithm can be applied to predict the future values of a time series. Section 3 explains why to narrow the search space of neighbours could be effective. Section 4 describes the experimental setup and how the meta parameters of the nearest neighbour algorithm have been chosen. Section 5 shows the results of our experimentation and Section 6 draws some conclusions.

2 Using nearest neighbours in time series forecasting

The nearest neighbour is a classical supervised algorithm in machine learning. Originally, it was applied in classification tasks. Given an unlabeled example, we look for the most similar labeled example, according to some features of the examples and using a distance function to express similarity. The label of the most similar example is used to classify the unlabeled example. The labeled examples can contain outliers and errors that can distort the prediction, so instead of looking for the nearest neighbour normally the k nearest neighbours are found and their majority class is used as prediction.

The nearest neighbours algorithm can be easily extended to a regression task. Instead of having labeled examples, the examples contain a numeric target value and the target value of the k nearest neighbours is combined—for example, it is averaged—to produce the predicted value. In a time series forecasting scenario, the target value of an example is a time series observation and the features describing the example are lagged values of the observation. This way, the next value of a time series is predicted looking for examples in the time series similar to the last observations—see Figure 1. Figure 1 shows an example of one-step ahead forecasting using lags 1 to 4 as features and two nearest neighbours. The last four values of the time series are the features of the example to be regressed on and the two sets of consecutive white squares the 2 nearest neighbours, whose targets are the black triangles. The mean of the two targets is the forecast—the asterisk.

The underlying reason to using nearest neighbours in a time series scenario is that a time series contains repetitive patterns. Hence, given the last behaviour of a time series we try to look for similar past behaviours in the hope that their subsequent values can be similar to the future values of the time series.

3 Why to narrow the search space of neighbours

When we do not know too much about the structure of a time series, we simply select the features—lagged values—of the future observation to be forecast and look for the most

 ISBN: 978-84-617-8694-7

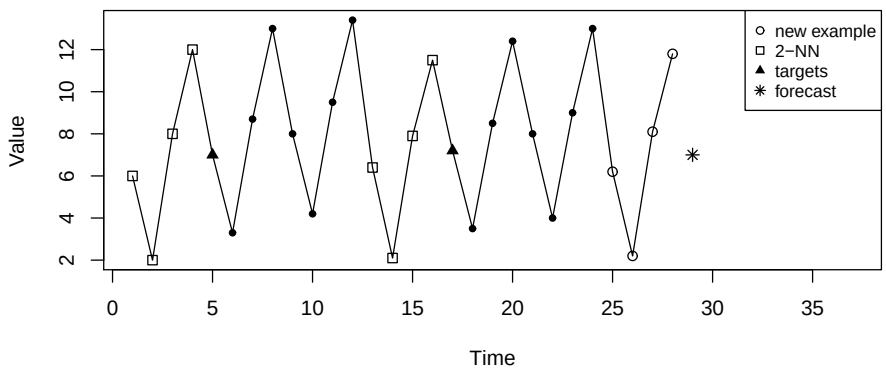


Figure 1: Example of 2 nearest neighbours for one-step-ahead forecasting.

similar examples in the time series. However, frequently we *do* know about the structure. For example, in the data explained in more detail in the next section, we are going to work with daily time series of cash money withdrawals at cash-machines. Apart from other patterns, these series have a strong weekly seasonality. Concretely, the withdrawals on weekends are very different from the withdrawals on workdays. Suppose now that we have information until a certain Friday and we want to predict the withdrawals on the following day, i.e., a Saturday. Let us also suppose that we are using the nearest neighbour algorithm and we use as feature the first lagged value. That is, to predict a Saturday we will find the day more similar to its previous day—a Friday. The most similar day found could be any day, let us suppose that it is a Tuesday. So we are going to predict the next Saturday withdrawals using the withdrawals on a Wednesday. This is not a desirable situation, because the withdrawals on Saturdays have nothing to do with the withdrawals on Wednesdays.

In our proposal, for this case we propose to narrow the search space looking only on previous Fridays to produce a more sensible forecast. That is, the prediction on a Saturday will be the average of several Saturdays whose previous Fridays are similar to the last Friday in the time series.

 ISBN: 978-84-617-8694-7

4 Experimental setup

In order to assess the effectiveness of our proposal we have used data from the NN5 time series competition[1]. In this competition 111 time series of 2 years of daily cash money withdrawals at various automatic teller machines (ATMs, or cash machines) at different locations in England were used. The goal was to forecast the cash money demand for the next 56 days ahead of every one of the 111 time series.

To assess forecast accuracy the symmetric mean absolute percentage error—SMAPE—was used. Given a vector forecast f for the next 56 days of a time series and the vector y of actual future values, the SMAPE is computed as follows:

$$SMAPE = \frac{1}{56}\sum_{t=1}^{56}\frac{|y_t - f_t|}{(|y_t| + |f_t|)/2}100$$

The SMAPE obtained for every one of the 111 time series by a given method is averaged to obtain a final SMAPE, which is used as the main forecast accuracy measure to compare the different methods.

Figure 2 includes a time series from the NN5 competition. Although these series contain multiple overlying seasonalities, in this work we are interested in experimenting with their strong weekly seasonality pattern.

4.1 The lazy learner meta parameters

In order to apply the nearest neighbours algorithm several meta parameters have to be chosen. Next, we describe our choices:

- As similarity function, i.e. to select how similar two examples are given their features, the Euclidian distance has been used.
- As combination function, i.e. to combine the target values of the k nearest neighbours, the arithmetic mean has been used.
- For k—the number of nearest neighbours—several values have been used to assess how robust the proposed method is.
- The final parameter is the lagged observations that are used as feature vector to compare the similarity among examples. We have used consecutive lags starting from lag 1. Concretely, we have experimented with lags 1, 1 to 7 and 1 to 14. We have chosen multiples of seven, because we are interested in the weekly seasonality.

[1] http://www.neural-forecasting-competition.com/NN5/

 ISBN: 978-84-617-8694-7

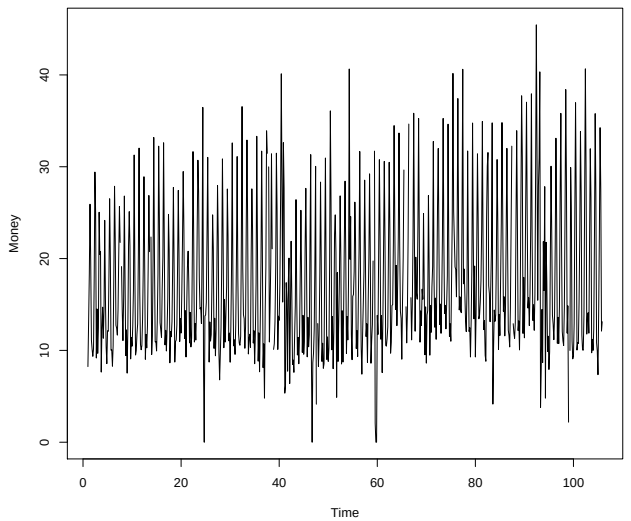


Figure 2: Example of time series from the NN5 competition.

Finally, we have to explain how the multiple—in this case 56—step-ahead forecasts are generated. When the forecast horizon is higher than 1 you have to chose among different strategies to generate the multiple forecasts, being the most used ones the direct, recursive and MIMO [5]. To date, there is no clear evidence that any strategy is superior than the others. Given this situation, we have chosen the recursive or iterative strategy because is quite straightforward. This strategy is used in the ARIMA and exponential smoothing methodologies. The recursive strategy uses a one-step-ahead model iteratively to get all the forecasts. That is, for forecasting horizon 1, the lagged observations of the time series are used as autoregressive variables for the model. For forecasting horizon 2, lacking a historical lagged one value to be used as autoregressive variable, the forecast for horizon 1 is used instead.

5 Experimentation

The goal of our experimentation is to assess whether, knowing information about the structure of a time series, the use of a selected space of possible neighbours can improve the forecast accuracy of the nearest neighbours algorithm. To that end, we have experimented with forecasting the 111 time series of the NN5 competition. We try to take advantage of our prior knowledge of the weekly seasonality pattern in the time series of this competition.

 ISBN: 978-84-617-8694-7

The way in which we are going to use this knowledge is as follows: in order to forecast the money withdrawals on one day, we take into account the day of the week to be forecast and only use as possible nearest neighbours examples, those examples whose targets have the same day of the week as the day being forecast.

In order to assess whether this strategy is effective we are going to execute the following nearest neighbours instances to predict 56 days ahead of the 111 time series of the NN5 competition:

- Classical nearest neighbours algorithm with k ranging from 1 to 10 and autoregressive lagged values at consecutive lags: 1, 1 to 7 and 1 to 14.
- Nearest neighbours algorithm with the same meta parameters, but using a selected space of possible neighbours, so that only examples whose targets have the same day of the week as the day of the value being forecasted are included.

Table 1: Global SMAPE of different algorithms over the 111 NN5 time series.

	1	2	3	4	5	6	7	8	9	10
C - 1	49.04	42.52	40.27	39.36	39.04	38.30	38.98	39.20	38.77	39.00
SS - 1	32.87	27.66	26.32	25.27	24.60	24.51	24.18	24.09	24.14	24.13
C - 1:7	31.55	28.21	25.33	23.95	23.61	23.43	23.36	23.51	23.60	23.48
SS - 1:7	30.16	25.42	23.56	23.08	22.56	22.37	22.32	22.44	22.51	22.58
C - 1:14	32.19	26.63	24.77	23.97	23.39	23.06	22.99	22.81	22.71	22.76
SS - 1:14	31.28	25.33	24.02	23.04	22.93	22.46	22.35	22.15	22.12	22.15

The results of our experiment are shown in Table 1. The first row of this table indicates the numbers of neighbours—i.e., the k meta parameter. The first column indicates whether the algorithm uses the classical approach (C) or our proposal of a selected search space (SS). After the hyphen the number of lags used as autoregressive variables is shown—for example, C - 1:7 means the classical algorithm with lags from 1 to 7. For every combination of type of algorithm, lags and k the table shows its forecast accuracy using the SMAPE measure over the 111 time series from the NN5 competition. The computation of SMAPE was described in the previous section. The most outstanding result is that for every combination of k and lags, the new approach outperform the results of the classical algorithm. The improvement in forecast accuracy is especially evident when only one lag is considered. This is a expected result, as with only a lagged value is easier to find misleading neighbours.

6 Conclusions

In this paper we have proposed to reduce the search space of possible neighbours in a nearest neighbours algorithm applied to time series forecasting when prior information about the

 ISBN: 978-84-617-8694-7

structure of the time series is known. The goal is to look for only significant neighbours. Our preliminary experimentation on the time series of the NN5 competition seems to indicate that the proposal is quite effective.

Acknowledgements

This paper has been partially supported by the project TIN2015-68854-R (FEDER Founds) of the Spanish Ministry of Economy and Competitiveness.

References

[1] G. E. Box, G. M. Jenkins and G. C. Reinsel, *Time Series Analysis: Forecasting and Control, 4th edition*, John Wiley & Sons, Hoboken, 2008.

[2] R. J. Hyndman, A. B. Koehler, J. K. Ord and R. D. Snyder, *Forecasting with Exponential Smoothing: The State Space Approach*, Springer–Verlag, Berlin, 2008.

[3] S. F. Crone, M. Hibon, K. Nikolopoulos, *Advances in forecasting with neural networks? Empirical evidence from the NN3 competition on time series prediction*, International Journal of Forecasting **27(3)** (2011) 635–660.

[4] I. H. Witten, E. Frank and M A. Mark, *Data Mining: Practical Machine Learning Tools and Techniques, 3rd edition*, Morgan Kaufmann Publishers Inc., San Francisco, 2011.

[5] B. T. Souhaib, G. Bontempi, A. F. Atiya and A. Sorjamaa, *A Review and Comparison of Strategies for Multi-step Ahead Time Series Forecasting Based on the NN5 Forecasting Competition*, Expert Syst. Appl. **39(8)** (2012) 7067–7080.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

Calibration estimator for Head Count Index

Sergio Martínez[1], María Illescas[2], Helena Martínez[1] and Antonio Arcos[3]

[1] *Department of Mathematics, University of Almería, Spain*

[2] *Department of Economics and Business, University of Almería, Spain*

[3] *Department of Statistics and Operational Research, University of Granada, Spain*

emails: spuertas@ual.es, millescasdesarrolla@gmail.com, hmp603@ual.es, arcos@ugr.es

Abstract

This paper considers the problem of estimating a poverty measure, the Head Count Index, using the auxiliary information available, which is incorporated into the estimation procedure by calibration techniques. The proposed method does not directly use the auxiliary information provided by auxiliary variables related to the variable of interest in the calibration process, but the auxiliary information, after a transformation, is incorporated by calibration techniques applied to the distribution function of the study variable. Monte Carlo experiments were carried out for simulated data and for real data taken from the Spanish living conditions survey to explore the performance of the new estimation methods of the Head Count Index.

Key words: Auxiliary information, calibration estimator, poverty index, survey sampling

1 Introduction

The estimation of a proportion in finite populations is a interesting topic in many areas such as medical and pharmaceutical statistics, marketing research, sociological studies and has important applications in the field of economics. Indeed, the analysis of poverty and social exclusion measures is a topic of increased interest to society. For governments is of high interest the estimation of poverty, inequality and life condition indicators and many social indicators related to the measurement of poverty are based upon binary variables or require the use of proportions to obtain such indicators. Among these poverty measures, we

 ISBN: 978-84-617-8694-7

can find the Head Count Index that is widely used by institutions to elaborate their reports on poverty. The Head Count Index (HCI) can be calculated as the proportion of persons (or households) with an equivalised disposable income below the 60% of the national median equivalised income. In the literature, numerous references discuss about the HCI and related poverty indicators. For instance, some references are [11], [1], [14], [13] and [12]. The real HCI is unknown in practice, but it is estimated by using survey data, therefore estimation methods for proportions are required, since the HCI can be expressed as a proportion. Usually the method for estimating the HCI is by using direct estimators without using auxiliary information, but official surveys on income and living conditions generally contain additional variables related to the variable of interest and the efficient insertion of the auxiliary information available would improve the precision of the estimations for the proportion of a categorical variable of interest. These additional variables includes numeric and binary attributes and the HCI can have stronger relationship with auxiliary quantitative variables. In the presence of auxiliary information, there exist several procedures to obtain more efficient estimators for the proportion of a categorical variable of interest, maybe some of them ([17] and [8]) assume that the auxiliary information is given by binary variables and consequently the auxiliary quantitative variables can not include at the estimation stage. In the case that the auxiliary information available includes both categorical and numerical attributes, we can use the logistic generalised regression estimator, proposed by [5] but has the problem of estimating the parameter associated to the logistic model.
In this paper, we consider the problem of estimating the population proportion of a categorical variable using the calibration framework. Calibration techniques were first employed by [2] to estimate the total population, but this approach is also applicable to the estimation of parameters more complex than the total population. [4], [15] and [16] use different ways to implement the calibration approach in the estimation of the distribution function and the quantiles. The use of calibration techniques in the estimation of population proportion of a categorical variable is not new. In [8] the authors proposed estimation procedures for a proportion and based on calibration framework but as we discussed previously, the estimator obtained cannot be applied for the estimation of the HCI, since they assume that the auxiliary information is exclusively given by binary variables. Another calibration alternative when the auxiliary information includes both categorical and numerical attributes is given in [7] where it was proposed a calibration estimator based on probit regression. In this paper, we consider the incorporation of the auxiliary information with calibration techniques applied to the distribution function of the study variable under simple random sampling. The article is arranged as follows. In Section 2, the HCI and indirect estimation methods are introduced. Section 3 gives a alternative calibration estimator for HCI based on the estimation of the distribution function. In Section 4, we derive optimum estimators in the sense of minimum variance when the sample is selected under simple random sampling without replacement (SRSWR). Finally, in Section 5, simulation studies are carried

 ISBN: 978-84-617-8694-7

out to analyze the performance of estimator proposed in this paper. Simulation studies are based upon real survey data and simulated finite populations. The real data is obtained from the Spanish living conditions survey. Section 5 gives some concluding remarks.

2 The Headcount Index and Indirect Estimation of population proportion

Let $U = \{1, 2, \ldots, N\}$ be a finite population consisting of N different elements. Let $s = \{1, 2, \ldots, n\}$ be the set of the units included in a sample, selected according to a specified sampling design with inclusion probabilities π_k and π_{kl} assumed to be strictly positive. We assume that y is the quantitative variable used to obtain the HCI and L is the poverty line used to classify the population into poor and nonpoor, that is, an individual (or households) is considered as poor if its income or expenditure y is less than the poverty line L. Thus, the real Head Count Index can be defined as the population proportion of the attribute A in the population U, $HCI = P_A = \sum_{k \in U} A_k/N$, where $A_k = 1$ if the unit k is classified as poor ($y_k \leq L$) and $A_k = 0$ otherwise. The value A_k is only available for the sample units. We assume that the poverty line L is established by the corresponding authority, i.e., L is fixed at some official quantity. For instance, Eurostat fixes the relative poverty line in the 60% of the median of the equivalised net income. To estimate P_A, the usual design-weighted Horvitz-Thompson estimator is $\widehat{P}_{AHT} = \sum_{k \in s} d_k A_k/N$, where $d_k = 1/\pi_k$. Now, we assume the existence of a vector $\mathbf{x} = (x_1, x_2, \ldots x_P)'$ of auxiliary information, such that for every population unit k the value $\mathbf{x}_k = (x_{1k}, x_{2k}, \ldots, x_{Pk})$ is known. We also assume that the variables included in the vector $\mathbf{x}$ can be either numeric or binary attributes of the same type as the study attribute A. Most official surveys on income and living conditions contain auxiliary variables related to the variable of interest, these auxiliary variables can be quantitative variables or qualitative attributes.

The Horvitz-Thompson estimator $\widehat{P}_{AH}$ is an unbiased estimator for P_A but does not use the auxiliary information provided by the vector $\mathbf{x}$. The incorporation of auxiliary information in estimating the population proportion P_A is not new and has been treated in many works. If the auxiliary vector $\mathbf{x}$ only includes binary attributes, we can use the estimation methods proposed by [17]. In the case that the vector $\mathbf{x}$ includes both categorical and numerical attributes, we can use the logistic generalised regression estimator, proposed by [5]. This estimator is given by:

$$\widehat{P}_{LGREG} = \frac{1}{N}\left(\sum_{k \in U} pl_k + \sum_{k \in s} \frac{A_k - pl_k}{\pi_k}\right) \tag{1}$$

 ISBN: 978-84-617-8694-7

where $pl_k = exp(x_k\widehat{\beta})/(1 + exp(x_k\widehat{\beta}))$ and $\widehat{\beta}$ is the BLUP estimator of the β parameter of the logistic regression. [3] provided some codes to compute the LGREG estimator and a Monte Carlo study to empirically investigate the accuracy of the confidence intervals when HT and LGREG estimators are used.

One way of incorporating auxiliary information provided by $\mathbf{x}$ in the estimation of P_A is via replacing the weights d_k of the estimators $\widehat{P}_{AH}$ by new weights ω_k, using calibration techniques. Following [2], to obtain a calibration estimator for the attribute A, we calculate the weights ω_k minimizing the chi-square distance

$$\Phi_s = \sum_{k\in s} \frac{(\omega_k - d_k)^2}{d_k q_k} \tag{2}$$

subject to a set of calibration constraints, where q_k are known positive constants unrelated to d_k. Thus, if $\mathbf{x}$ only includes binary attributes, we can estimate P_A through calibration techniques proposed by [8]. Calibration techniques have also recently been used in the estimation of P_A when the vector of auxiliary variables $\mathbf{x}$ contains both binary and numerical attributes. Thus, in [7], it was proposed a calibration estimator $\widehat{P}_{CP}$ based on probit regression, where the calibrated weights ω_k are obtained by minimizing (2) subject to the following conditions:

$$\frac{1}{N}\sum_{k\in s}\omega_k p_k = \bar{P} = \frac{1}{N}\sum_{k\in U} p_k; \text{ and } \frac{1}{N}\sum_{k\in s}\omega_k = 1 \tag{3}$$

with $p_k = \widehat{P}[A_k = 1] = F(\widehat{\beta}' \cdot \mathbf{x}_k)$, where F is the normal-standard distribution function and $\widehat{\beta}$ is the π-weighted likelihood estimator of the β parameter of the probit regression ([8]).

In the next section we consider the estimation of the population proportion by estimating the distribution function $F_A(t)$ of the attribute of study A.

3 Calibration Estimation of population proportion by estimating distribution function

In this section, we describe alternative calibration estimation methods for the problem of estimating P_A, based on auxiliary vector $\mathbf{x}$ that includes numeric and binary attributes. This calibration methods define a new indirect estimator for P_A through the estimation of the distribution function $F_A(t) = \sum_{k\in U} \Delta(t - A_k)/N$ of the attribute of study A, where $\Delta(t - A_k) = 1$ if $t \geq A_k$ and $\Delta(t - A_k) = 0$ otherwise. Since the aim is to estimate the population proportion P_A by estimating the distribution function, we will consider the

 ISBN: 978-84-617-8694-7

complementary attribute $\bar{A}$ of the attribute A, this is $\bar{A}_k = 1 - A_k$. Thus, if $F_{\bar{A}}(t)$ is the distribution function associated with the complementary attribute, it is clear that $P_A = F_{\bar{A}}(0)$ and the estimate of the population proportion P_A can be obtained by the methods of estimating the distribution function. The usual estimator of distribution function is the Horvitz-Thompson estimator given by

$$\widehat{F}_{\bar{A}HT}(t) = \frac{1}{N}\sum_{k\in s} d_k \Delta(t - \bar{A}_k) \tag{4}$$

From (4) it is easy to see that $\widehat{P}_{AH} = \widehat{F}_{\bar{A}HT}(0)$. Thus, we will obtain new indirect estimators of P_A through calibration techniques applied to $F_{\bar{A}}(t)$ at the point $t = 0$.
Recently, the calibration approach have been employed for the estimation of the distribution function and quantiles in different ways ([4], [15] and [16]). Following [15], we consider the definition of a pseudo-variable $g_k = \widehat{\beta}'\mathbf{x}_k$ for $k = 1, 2, \ldots, N$ with

$$\widehat{\beta}' = \left(\sum_{k\in s} d_k q_k \mathbf{x}_k \mathbf{x}_k'\right)^{-1} \sum_{k\in s} d_k q_k \mathbf{x}_k \bar{A}_k,$$

where q_k are known positive constants unrelated to d_k. With the pseudo-variable g, we consider the estimation of $P_A = F_{\bar{A}}(0)$ with the calibration estimator obtained with the minimization of (2) subject to the following conditions:

$$\frac{1}{N} = \sum_{k\in s} \omega_k \Delta(\mathbf{t_g} - g_k) = F_g(\mathbf{t_g}) \tag{5}$$

with $\mathbf{t_g} = (t_1, \ldots, t_P)'$ is a vector chosen arbitrarily, where $t_1 < t_2 < \ldots < t_P$. Assuming that the inverse of symmetric matrix $T = \sum_{k\in s} d_k q_k \Delta(\mathbf{t_g} - g_k)\Delta(\mathbf{t_g} - g_k)'$ exists, the resulting estimator ([15]) is given by

$$\widehat{P}_{AC} = \widehat{F}_{\bar{A}C}(0) = \widehat{P}_{AHT} + \left(F_g(\mathbf{t_g}) - \widehat{F}_{GHT}(\mathbf{t_g})\right)' \cdot \widehat{D} \tag{6}$$

where $\widehat{D} = T^{-1} \cdot \sum_{k\in s} d_k q_k \Delta(\mathbf{t_g} - g_k)\Delta(0 - \bar{A}_k)$ and $\widehat{F}_{GHT}(\mathbf{t_g})$ is the Horvitz-Thompson estimator of $F_g(\mathbf{t_g})$.

The asymptotic variance of $\widehat{F}_{\bar{A}C}(0)$ ([15]) is given by:

$$AV(\widehat{F}_{\bar{A}C}(0)) = \frac{1}{N^2}\sum_{k\in U}\sum_{l\in U} \Delta_{kl}(d_k E_k)(d_l E_l) \tag{7}$$

where $D = \left(\sum_{k\in U} q_k \Delta(\mathbf{t_g} - g_k)\Delta(\mathbf{t_g} - g_k)'\right)^{-1} \cdot \left(\sum_{k\in U} \Delta(\mathbf{t_g} - g_k)\Delta(0 - \bar{A}_k)\right)$ and $E_k = \Delta(0 - \bar{A}_k) - \Delta(\mathbf{t_g} - g_k) \cdot D$.

 ISBN: 978-84-617-8694-7

4 Determining optimal calibration estimators

The precision of $\widehat{F}_{\bar{A}C}(0)$ changes with the selection of $\mathbf{t_g}$. In [6] and [9], the authors studied, for a fixed P, the problem of selection the optimal vector $\mathbf{t_g}$ under simple random sampling and $q_k = 1$ for all $k \in U$, that gives the best estimation of $F_y(t)$ with the calibration estimator $\widehat{F}_{yc}(t)$ developed in [15], that is, the problem of determining an auxiliary vector $\mathbf{t_g} = (t_1, \ldots, t_P)'$, with $t_1 < t_2 < \ldots < t_P$ that minimizes the variance of the estimator $\widehat{F}_{yc}(t)$ given a point t for which we want to estimate $F_y(t)$. Moreover, in [10], the problem of the optimal dimension P of the auxiliary vector $\mathbf{t}_P$ and the optimal vector of this dimension is studied for the calibrated estimator of [15]. Following [9], the minimization of the asymptotic variance (7) under simple random sampling, is equivalent to minimizing the following function:

$$G(t_1, t_2, \ldots, t_P) = 2NP_A \cdot k_P - \sum_{j=1}^{P} \frac{(k_j - k_{j-1})^2}{(F_g(t_j) - F_g(t_{j-1}))} - k_P^2 \tag{8}$$

where $k_i = \sum_{k \in U} \Delta(0 - \bar{A}_k)\Delta(t_i - g_k)$ with $i = 1, 2 \ldots, P$; $k_0 = 0$ and t_0 is a value such that $F_g(t_0) = 0$.

If we consider the auxiliary vector $\mathbf{t}_P = t_1$ (dimension $P = 1$), the value of t_1 at which the calibration estimator $\widehat{P}_{AC}$ is optimum ([6]) is given by $t_{opt} = \arg\min_{a_k \in A_0} G(a_k)$, where $A_0 = \{g_k : k \in U, \bar{A}_k = 0\} = \{g_k : A_k = 1\} = \{a_1, a_2, \ldots, a_M\}$ with $a_1 < a_2 < \ldots < a_M$. The optimal value of t_1, (t_{opt}) depends on some unknown values, so we go to replace the optimal vector t_{opt} by sample-based estimates. For it, we consider the following set $A_{0_s} = \{g_k : k \in s, \bar{A}_k = 0\} = \{a_{1_s}, a_{2_s}, \ldots, a_{m_s}\}$, and the global minimum of the function $\widehat{G}(t_1)$ (the usual estimation of $G(t_1)$), is at one point of A_{0_s} ([6]). Thus, we can define a new calibration estimator $\widehat{P}_{AC1}$ based on the auxiliary point $\widehat{t}_{opt}$ that minimizes the function $\widehat{G}(t_1)$. The asymptotic behaviour of the estimator $\widehat{P}_{AC1}$ is the same as the estimator based on optimum point t_{opt} ([6]). Thus the asymptotic variance of $\widehat{P}_{AC1}$ is given by (7) with $\mathbf{t_g} = t_{opt}$.

On the other hand, if the dimension of the auxiliary vector is $P > 1$, the global minimum of the function $G(\mathbf{t}_g)$ ([9]) is a vector $\mathbf{t_{GP}} = (t_1, t_2, \ldots, t_P)$, with $t_1 < t_2 < \ldots < t_P$ and $t_i \in A_0$ or $t_i \in B_0$ for $i = 1, 2, \ldots, P$, where $B_0 = \{b_1, b_2, \ldots, b_M\}$ with $b_h = \max_{l \in U_h}\{g_l\}$ for $h = 1, \ldots, M$; $U_1 = \{l \in U : g_l < a_1\}$ and $U_h = \{l \in U : a_{h-1} \leq g_l < a_h\}$, $h = 2, 3, \ldots, M$. It is clear that $b_1 < b_2 < \ldots < b_M$. Since the sets A_0 and B_0 are finite, finding the global minimum is computationally simple. For some h in $1, 2, \ldots, M$ the corresponding point b_h may not exist, but in this case, the minimization problem is simpler than the current case

 ISBN: 978-84-617-8694-7

([9]). Again, the optimal auxiliary vector $\mathbf{t_{GP}}$ depends on some unknown values, therefore we will replace the optimal vector with sample-based estimates. For it, we consider the usual estimation of the function G denoted by $\widehat{G}(\mathbf{t}_g)$, the sample-based set A_{0_s} and the set $B_{0_s} = \{b_{1_s}, b_{2_s}, \ldots, b_{m_s}\}$ with $b_{h_s} = \max_{l \in U_{h_s}}\{g_l\}$ for $h = 1, \ldots, m$; $U_{1_s} = \{l \in s : g_l < a_{1_s}\}$ and $U_{h_s} = \{l \in s : a_{(h-1)_s} \leq g_l < a_{h_s}\}$, $h = 2, 3, \ldots, m$.
The potential points for the global minimum of $\widehat{G}(\mathbf{t}_g)$ are $\widehat{\mathbf{t}}_{\mathbf{GP}} = (\widehat{t}_1, \widehat{t}_2, \ldots, \widehat{t}_P)$ with $\widehat{t}_i \in A_{0_s}$ or $\widehat{t}_i \in B_{0_s}$ ([9]). The calibration estimator $\widehat{P}_{ACP}$ based on $\widehat{\mathbf{t}}_{\mathbf{GP}}$ has the same asymptotic behaviour that the estimator based on $\mathbf{t_{GP}}$ and the asymptotic variance is given by (7) with $\mathbf{t_g} = \mathbf{t_{GP}}$.

Following [10], the optimal dimension P of the auxiliary vector $\mathbf{t}_g$ is $2M$ if b_1 exists and for all $i = 1, \ldots, M-1$, $b_{i+1} \neq a_i$. The optimal vector in this case is $\mathbf{t_{OPT}} = (b_1, a_1, \ldots, b_M, a_M)$. If for some $i_1, i_2, \ldots i_R \in \{0, 1, \ldots, M-1\}$; $a_{i_1} = b_{i_1+1}$ with $R \leq M$ and $i_h \neq i_j$ if $h \neq j$ the optimal dimension is $P = 2M - R$ and the optimal auxiliary vector $\mathbf{t_{OP}} = (t_{O1}, \ldots, t_{O(2M-R)})$ is given by:

$$\mathbf{t_{OP}} = (b_1, a_1, b_2, a_2, \ldots, b_{i_1}, a_{i_1}, a_{i_1+1}, b_{i_1+2}, \ldots, b_{i_h}, a_{i_h}, a_{i_h+1}, b_{i_h+2}, \ldots b_M, a_M) \quad (9)$$

The optimal auxiliary vector $\mathbf{t_{OP}}$ depends on some unknown values, furthermore, although the vector $\mathbf{t_{OP}}$ be known, we could have incompatible restrictions in (5) when a sample s is selected. Thus, similarly to the previous cases, we can define a new calibration estimator $\widehat{P}_{ACOPT}$ based on $\widehat{\mathbf{t}}_{\mathbf{OP}}$, a sample-based estimation.

The Horvitz-Thompson estimator $\widehat{P}_{AHT}$, under SRSWOR, has the following shift invariance property $\widehat{P}_{AHT} = 1 - \widehat{Q}_{AHT}$, where $\widehat{Q}_{AHT}$ is the Horvitz-Thompson estimator for $Q_A = 1 - P_A$. Thus, $\widehat{P}_{AHT}$ has the same performance in the estimation of P_A as the performance of $\widehat{Q}_{AHT}$ in the estimation of Q_A. In general, this property is not satisfied by the calibration estimators considered $\widehat{P}_{AC1}$, $\widehat{P}_{ACP}$ and $\widehat{P}_{ACOPT}$. It is easy to see that this property is fulfilled by a calibration estimator if $N = \sum_{k \in U} \omega_k$. A way to obtain this condition consists in the incorporation of the value $g_{max} = \max_{k \in U} g_k$ in the auxiliary optimum vectors. Thus, we can define a calibration estimator $\widehat{P}_{AQ1}$ based on the auxiliary vector $(\widehat{t}_{opt}, g_{max})$, a calibration estimator $\widehat{P}_{AQP}$ based on $(\widehat{\mathbf{t}}_{GP}, g_{max})$ and a calibration estimator $\widehat{P}_{AQOPT}$ based on $(\widehat{\mathbf{t}}_{OP}, g_{max})$. Nothing guarantees that we can use this vector in the calibration constraints given by (5), when selecting a sample s, since we could have incompatible restrictions. In this case, we consider the calibration constraints given by (5) without $g_{max} = \max_{k \in U} g_k$.
It is easy to see that the incorporation of the value g_{max} in the calibration conditions (5) does not produces a negative effect in the asymptotic variance. For it, from equation (8) we have

$$G(\widehat{\mathbf{t}}_{\mathbf{GP}}) - G((\widehat{\mathbf{t}}_{GP}, g_{max}) = (NP_A - k_P)^2 - \frac{(NP_A - k_P)^2}{(1 - F_g(t_P))} \leq 0$$

 ISBN: 978-84-617-8694-7

Another way to incorporate shift in-variance property, consists in the minimization of the function (8) when the auxiliary vector considered is $(\mathbf{t_1}, g_{max})$. Following [6] the optimal auxiliary vector is $\mathbf{t_{GMAX}} = (t_1, g_{max})$ where $t_1 \in A_0$ or $t_1 \in B_0$. Similarly to the previous cases, we can define a new calibration estimator $\widehat{P}_{ACMAX}$ based on $\widehat{\mathbf{t}}_{\mathbf{GMAX}}$, a sample-based estimation.

5 Numerical Comparison

In this sections, we present the results of a Monte Carlo simulation study where we compare the precision of the proposed calibration estimators: $\widehat{P}_{AC1}$, $\widehat{P}_{ACP}$, $\widehat{P}_{ACOPT}$, $\widehat{P}_{AQ1}$, $\widehat{P}_{AQP}$, $\widehat{P}_{AQOPT}$ and $\widehat{P}_{ACMAX}$ with the Horvitz-Thompson estimator $\widehat{P}_{AHT}$; the multivariated ratio estimator $\widehat{P}_{AMratio}$ (see [17]); calibration estimators P_{AR} and the multivariate calibration estimator $\widehat{P}_{AWM}$ (see [8]); the logistic generalised regression estimator $\widehat{P}_{LGREG}$ ([5] and calibration estimator $\widehat{P}_{CP}$ based on probit regression (see [7]). The estimators $\widehat{P}_{ACP}$ and $\widehat{P}_{AQP}$ are based on auxiliary vector with dimension $P = 2$. Our simulations are programmed in `R`, with some new code developed to compute the estimators to be compared. The performance of each proportion estimator was measured and compared in terms of relative bias (RB) and relative efficiency (RE). The simulated values of RB and RE for a particular proportion estimator T were computed as

$$\text{RB} = B^{-1}\sum_{b=1}^{B}(T^b - P)/P, \quad \text{RE} = MSE(\widehat{P}_{AHT})/MSE(T^b)$$

where $MSE(T^b) = B^{-1}\sum_{b=1}^{B}(T^b - P)^2$, $MSE(\widehat{P}_{AHT}) = B^{-1}\sum_{b=1}^{B}(\widehat{P}^b_{AHT} - P)^2$, and T^b and $\widehat{P}^b_{AHT}$ are the values of T and $\widehat{P}_{AHT}$ from the bth simulation, respectively.

To investigate the efficiency of the proposed estimators under a variety of situations, we conside different stages. First, we will consider the estimation of a population proportion in simulated populations and secondly we will use the proposed estimators in the estimation of the Head Cont Index. For the estimation of population proportion, we consider 5 populations generated as a random sample of 10000 units from a Bernoulli distribution with parameter $P = 0.9$, and the attributes of interest were thus achieved with the aforementioned population proportion. Auxiliary attributes were also generated using the same distribution, but a given proportion of values was randomly changed so that Cramers V coefficient between the attribute of interest and the auxiliary attribute would range from 0.5 to 0.9. For each of the 5 populations, $B = 10000$ samples of sizes n =150, 250, 350 and 450 were selected, under simple random sampling, to compare the considered estimators in terms of relative bias (RB) and relative efficiency (RE). Tables 1 give the values of RB and RE in percentages for the binomial populations.

 ISBN: 978-84-617-8694-7

Table 1: RB % and RE % for several sample sizes of the estimators compared. SRSWOR from the BINOMIAL populations.

	RB%	RE%	RB%	RE%	RB%	RE%	RB%	RE%	RB%	RE%	RB%	RE%	RB%	RE%	RB%	RE%
	$\rho = 0.5$								$\rho = 0.6$							
Estimator	$n = 150$		$n = 250$		$n = 350$		$n = 450$		$n = 150$		$n = 250$		$n = 350$		$n = 450$	
$\widehat{P}_{AHT}$	0.014	100.00	-0.017	100.00	-0.003	100.00	-0.043	100.00	0.055	100	0.010	100	-0.016	100	-0.020	100
$\widehat{P}_{AMratio}$	0.014	100.00	-0.017	100.00	-0.003	100.00	-0.043	100.00	0.055	100	0.010	100	-0.016	100	-0.020	100
P_{AR}	0.031	109.21	-0.017	108.44	0.004	108.16	-0.037	109.90	0.040	109.78	-0.007	116.88	-0.009	117.98	-0.009	116.28
$\widehat{P}_{AWM}$	-0.002	112.48	-0.038	111.70	-0.008	112.08	-0.036	113.86	0.010	112.12	-0.035	127.92	-0.023	130.18	-0.020	126.98
$\widehat{P}_{LGREG}$	0.006	113.63	-0.034	112.99	-0.006	113.47	-0.035	115.42	0.018	113.39	-0.030	131.35	-0.021	134.15	-0.018	130.46
$\widehat{P}_{CP}$	0.002	113.29	-0.037	112.60	-0.007	113.13	-0.036	115.09	0.014	112.99	-0.037	130.66	-0.025	133.37	-0.021	129.62
$\widehat{P}_{AC1}$	0.014	100.00	-0.017	100.00	-0.003	100.00	-0.043	100.00	0.055	100	0.010	100	-0.016	100	-0.020	100
$\widehat{P}_{AQ1}$	0.014	100.00	-0.017	100.00	-0.003	100.00	-0.043	100.00	0.055	100	0.010	100	-0.016	100	-0.020	100
$\widehat{P}_{ACP}$	0.025	114.14	-0.029	114.51	0.001	115.12	-0.031	117.52	0.038	115.27	-0.010	133.62	-0.010	136.87	-0.012	131.63
$\widehat{P}_{AQP}$	0.025	114.14	-0.029	114.51	0.001	115.12	-0.031	117.52	0.038	115.27	-0.010	133.62	-0.010	136.87	-0.012	131.63
$\widehat{P}_{ACOPT}$	0.040	115.21	-0.027	115.07	0.003	115.79	-0.029	118.14	0.046	115.85	-0.002	135.30	-0.007	138.81	-0.010	133.41
$\widehat{P}_{AQOPT}$	0.040	115.21	-0.027	115.07	0.003	115.79	-0.029	118.14	0.046	115.85	-0.002	135.30	-0.007	138.81	-0.010	133.41
$\widehat{P}_{ACMAX}$	0.025	114.14	-0.029	114.51	0.001	115.12	-0.031	117.52	0.038	115.27	-0.010	133.62	-0.010	136.87	-0.012	131.63
	$\rho = 0.7$								$\rho = 0.8$							
$\widehat{P}_{AHT}$	0.015	100.00	0.013	100.00	0.008	100.00	0.006	100.00	0.004	100.00	-0.043	100.00	-0.007	100.00	-0.001	100.00
$\widehat{P}_{AMratio}$	0.015	100.00	0.013	100.00	0.008	100.00	0.006	100.00	0.004	100.00	-0.043	100.00	-0.007	100.00	-0.001	100.00
P_{AR}	0.011	127.74	-0.003	130.28	0.012	133.02	0.002	131.23	0.0157	165.15	-0.032	167.61	0.003	162.67	0.005	166.53
$\widehat{P}_{AWM}$	-0.13	140.46	-0.077	144.91	-0.029	148.35	-0.028	146.13	-0.145	187.40	-0.095	193.44	-0.048	189.55	-0.034	190.52
$\widehat{P}_{LGREG}$	-0.12	150.79	-0.070	156.15	-0.025	160.36	-0.025	158.11	-0.130	200.75	-0.081	209.88	-0.038	206.88	-0.028	208.68
$\widehat{P}_{CP}$	-0.15	148.42	-0.092	153.53	-0.038	157.63	-0.034	155.26	-0.160	197.77	-0.098	205.53	-0.050	202.48	-0.036	203.90
$\widehat{P}_{AC1}$	0.015	100.00	0.013	100.00	0.008	100.00	0.006	100.00	0.004	100.00	-0.043	100.00	-0.007	100.00	-0.001	100.00
$\widehat{P}_{AQ1}$	0.015	100.00	0.013	100.00	0.008	100.00	0.006	100.00	0.004	100.00	-0.043	100.00	-0.007	100.00	-0.001	100.00
$\widehat{P}_{ACP}$	-0.002	176.23	-0.004	181.61	0.014	184.02	0.001	181.07	-0.057	213.73	-0.020	225.64	0.005	223.49	-0.005	224.91
$\widehat{P}_{AQP}$	-0.002	176.23	-0.004	181.61	0.014	184.02	0.001	181.07	-0.057	213.73	-0.020	225.64	0.005	223.49	-0.005	224.91
$\widehat{P}_{ACOPT}$	0.004	173.86	-0.006	179.09	0.014	181.82	0.001	179.94	0.002	218.32	-0.011	225.62	0.005	221.95	-0.004	223.73
$\widehat{P}_{AQOPT}$	0.004	173.86	-0.006	179.09	0.014	181.82	0.001	179.94	0.002	218.32	-0.011	225.62	0.005	221.95	-0.004	223.73
$\widehat{P}_{ACMAX}$	-0.002	176.23	-0.004	181.61	0.014	184.02	0.001	181.07	-0.057	213.73	-0.020	225.64	0.005	223.49	-0.005	224.91
	$\rho = 0.9$															
$\widehat{P}_{AHT}$	-0.011	100.00	0.013	100.00	0.014	100.00	-0.002	100.00								
$\widehat{P}_{AMratio}$	-0.011	100.00	0.013	100.00	0.014	100.00	-0.002	100.00								
P_{AR}	-0.015	289.45	0.006	287.74	-0.001	286.77	-0.003	294.96								
$\widehat{P}_{AWM}$	-0.181	331.63	-0.123	344.48	-0.080	345.35	-0.057	356.19								
$\widehat{P}_{LGREG}$	-0.191	360.15	-0.127	377.50	-0.079	389.81	-0.053	400.60								
$\widehat{P}_{CP}$	-0.209	349.52	-0.143	364.93	-0.091	372.18	-0.062	382.58								
$\widehat{P}_{AC1}$	-0.011	100.00	0.013	100.00	0.014	100.00	-0.002	100.00								
$\widehat{P}_{AQ1}$	-0.011	100.00	0.013	100.00	0.014	100.00	-0.002	100.00								
$\widehat{P}_{ACP}$	-0.115	411.46	-0.050	437.04	-0.014	470.50	-0.006	469.58								
$\widehat{P}_{AQP}$	-0.115	411.46	-0.050	437.04	-0.014	470.50	-0.006	469.58								
$\widehat{P}_{ACOPT}$	-0.05	428.22	-0.020	445.22	-0.009	461.61	-0.006	463.43								
$\widehat{P}_{AQOPT}$	-0.05	428.59	-0.020	445.22	-0.009	461.61	-0.006	463.43								
$\widehat{P}_{ACMAX}$	-0.115	411.80	-0.050	437.04	-0.014	470.50	-0.006	469.58								

The results derived from this simulation study gave values for RB within a reasonable range. All the estimators considered produced absolute relative bias values of less than 0.5%. The estimators $\widehat{P}_{AHT}$, $\widehat{P}_{AMratio}$, $\widehat{P}_{AC1}$ and $\widehat{P}_{AQ1}$ has the same variance and have a larger variance than the other estimators considered. With large Cramer's V coefficient (ρ) values, the proposed estimators $\widehat{P}_{ACP}$, $\widehat{P}_{AQP}$, $\widehat{P}_{ACMAX}$, $\widehat{P}_{ACOPT}$ and $\widehat{P}_{AQOPT}$ produce good results. It can also be seen that as ρ increases, all the estimators achieve greater precision, which is particularly marked for very high correlations.
Of all the estimates that use auxiliary information, the calibration estimators $\widehat{P}_{ACOPT}$ and $\widehat{P}_{AQOPT}$ has the highest degree of efficiency for small values of ρ while for the large values, the estimators $\widehat{P}_{ACP}$, $\widehat{P}_{AQP}$ and $\widehat{P}_{ACMAX}$ present a greater efficiency. In all cases, the calibration estimators $\widehat{P}_{ACP}$, $\widehat{P}_{AQP}$, $\widehat{P}_{ACMAX}$, $\widehat{P}_{ACOPT}$ and $\widehat{P}_{AQOPT}$ perform better than

the estimators P_{AR}, $\widehat{P}_{AWM}$, $\widehat{P}_{LGREG}$ and $\widehat{P}_{CP}$.
The sample size produces a clear effect on the behaviour of the estimators: as the sample size increases, so does the efficiency of the estimators.

For the estimation of Head Count Index, we consider real data taken from the 2008 Spanish living conditions survey carried out by the Instituto Nacional de Estadística (INE) of Spain. For our simulation study, we considered the survey data collected as a population, from which samples are selected.The poverty threshold is calculated each year, using the distribution of the equivalised net income for the previous year. Following the criteria recommended by Eurostat, this threshold is set at 60% of the median of the equivalised net income. We considered the variable "Returns and additional revenue from adjustments in taxes" and the attribute "Home with own car" (1 for home with own car, 0 otherwise) as the auxiliary variables. Again, $B = 10000$ samples of sizes n =500, 600, 700 and 800 were selected, under simple random sampling, to compare the relative bias (RB) and relative efficiency (RE) of the considered estimators. Table 2 give the values of RB and RE in percentages for the real population.

Table 2: RB % and RE % for several sample sizes of the estimators compared under SRSWOR from the 2008 Spanish living conditions survey population.

	RB%	RE%	RB%	RE%	RB%	RE%	RB%	RE%
Estimator	$n = 500$		$n = 600$		$n = 700$		$n = 800$	
$\widehat{P}_{AHT}$	-0.066	100.00	-0.333	100.00	-0.256	100.00	0.044	100.00
$\widehat{P}_{AMratio}$	-0.066	100.00	-0.333	100.00	-0.258	100.00	0.044	100.00
P_{AR}	-0.064	94.80	-0.363	90.14	-0.286	96.98	0.027	96.65
$\widehat{P}_{AWM}$	-0.184	104.88	-0.227	102.49	-0.306	102.10	0.017	102.03
$\widehat{P}_{LGREG}$	-0.151	105.04	-0.190	102.50	-0.276	103.01	0.037	102.73
$\widehat{P}_{CP}$	-0.155	104.96	-0.221	102.47	-0.274	102.71	0.042	102.57
$\widehat{P}_{AC1}$	-0.199	99.08	-0.288	100.23	-0.342	99.53	-0.041	99.66
$\widehat{P}_{AQ1}$	-0.107	99.199	-0.287	100.85	-0.273	99.68	0.027	99.68
$\widehat{P}_{ACP}$	-0.648	104.31	-0.337	112.15	-0.532	103.16	-0.053	102.95
$\widehat{P}_{AQP}$	-0.638	104.39	-0.380	112.90	-0.522	103.22	-0.053	103.04
$\widehat{P}_{ACOPT}$	-0.167	105.05	0.388	102.45	-0.157	103.06	0.057	102.55
$\widehat{P}_{AQOPT}$	0.169	105.04	0.788	102.23	0.156	103.23	0.058	102.77
$\widehat{P}_{ACMAX}$	-0.238	105.70	-0.443	103.63	-0.238	104.32	0.050	103.60

In this population, the results are slightly different. The relative biases usually remain negligible (less than 0.5 %) but the efficiency is different:

- The calibration estimator P_{AR}, $\widehat{P}_{AC1}$ and $\widehat{P}_{AQ1}$ performs poorly and has worse efficiency than the estimator $\widehat{P}_{AHT}$.
- The remaining estimates are more efficient than the estimator $\widehat{P}_{AHT}$ but the gain in

 ISBN: 978-84-617-8694-7

efficiency is not as great as in the previous example.

- The proposed calibration estimator $\widehat{P}_{ACP}$, $\widehat{P}_{AQP}$, $\widehat{P}_{ACMAX}$, $\widehat{P}_{ACOPT}$ and $\widehat{P}_{AQOPT}$ often work better than the other estimators. The estimator $\widehat{P}_{ACMAX}$ is the most efficient for all sample sizes, except for the sample size $n = 600$ where the estimators $\widehat{P}_{ACP}$ and $\widehat{P}_{AQP}$ present the best results.

Acknowledgements

This study was partially supported by Consejería de Economía, Innovación, Ciencia y Empleo (grant SEJ2954, Junta de Andalucía, Spain) and by Ministerio de Economía y Competitividad (grant MTM2015-63609-R, MINECO/FEDER).

References

[1] E. Crettaz and C. Suter, *The impact of adaptive preferences on subjective indicators: An analysis of poverty indicators*, Social Indicators Research. **114** (2013) 139–152.

[2] J. C. Deville and C. E. Särndal, *Calibration Estimators in Survey Sampling*, Journal of the American Statistical Association. **87** (1992) 376–382.

[3] P. Duchesne, *Estimation of a proportion with survey data*, Journal of Statistics Education. [Online] **11(3)** (2003) (www.amstat.org/publications/jse/v11n3/duchesne.pdf).

[4] T. Harms and P. Duchesne, *On calibration estimation for quantiles*, Survey Methodology. **32** (2006) 37–52.

[5] R. Lehtonen and A. Veijanen, 1998. *Logistic generalized regression estimators*, Survey Methodology. **24** (1998) 51–55.

[6] S. Martínez, M. Rueda, A. Arcos and H. Martínez, 2010 *Optimum calibration points estimating distribution functions*, Journal of Computational and Applied Mathematics. **233(9)** (2010) 2265–2277.

[7] S. Martínez, M. Rueda, A. Arcos and H. Martínez, 2014. *Estimating the proportion of a categorical variable with probit regression*, Sociological Methods and Research. In Press.

[8] S. Martínez, A. Arcos, H. Martínez and S. Singh, 2015a. *Estimating Population Proportions by Means of Calibration Estimators*, Revista Colombiana de Estadistica. **38(1)** (2015a) 267–293.

 ISBN: 978-84-617-8694-7

[9] S. Martínez, M. Rueda, H. Martínez and A. Arcos, *Determining P optimum calibration points to construct calibration estimators of the distribution function*, Journal of Computational and Applied Mathematics. **275** (2015b) 281–293.

[10] S. Martínez, M. Rueda, H. Martínez and A. Arcos, *Optimal dimension and optimal auxiliary vector to construct calibration estimators of the distribution function*, Journal of Computational and Applied Mathematics. **318** (2017) 444–459.

[11] M. Medeiros, *The rich and the poor: The construction of an affluence line from the poverty line*, Social Indicators Research. **78** (2006) 1–18.

[12] D. Morales, M. Rueda and D. Esteban, *Model-assisted estimation of small area poverty measures: an application within the Valencia Region in Spain*, Social Indicators Research. In Press.

[13] J. F. Múñoz, E. Álvarez-Verdejo, R. M. García-Fernández and L. J. Barroso *Efficient Estimation of the Headcount Index*, Social Indicators Research. **123** (2015) 713–732.

[14] J. Navicke, O. Rastrigina and H. Sutherland, *Nowcasting Indicators of Poverty Risk in the European Union: A Microsimulation Approach*, Social Indicators Research. **119(1)** (2014) 101–119.

[15] M. Rueda, S. Martínez, H. Martínez and A. Arcos, *Estimation of the distribution function with calibration methods*, Journal of Statistical Planning and Inference. **137** (2007a) 435–448.

[16] M. Rueda, S. Martínez, H. Martínez and A. Arcos, *Calibration methods for estimating quantiles*, Metrika. **66** (2007b) 355–371.

[17] M. Rueda, J. F. Muñoz, A. Arcos, E. Álvarez and S. Martínez, 2011. *Estimators and confidence intervals for the proportion using binary auxiliary information with applications to pharmaceutical studies*, Journal of biopharmaceutical statistics. **21(3)** (2011) 526–554.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

New Lower Bounds for the Geometric Arithmetic index

Alvaro Martínez-Pérez[1] and José M. Rodríguez[2]

[1] *Departamento de Análisis Económico y Finanzas, Universidad de Castilla-La Mancha*

[2] *Departamento de Matemáticas, Universidad Carlos III de Madrid*

emails: alvaro.martinezperez@uclm.es, jomaro@math.uc3m.es

Abstract

The concept of geometric-arithmetic index was introduced in the chemical graph theory recently, but it has shown to be useful. The aim of this paper is to obtain new inequalities involving the geometric-arithmetic index GA_1 and characterize graphs extremal with respect to them. Our main results provide lower bounds on $GA_1(G)$ involving just the minimum and the maximum degree of the graph G.

Key words: Geometric-arithmetic index, Graph invariant, Vertex-degree-based graph invariant, Topological index.
MSC 2000: 05C07, 92E10

1 Introduction

A single number, representing a chemical structure in graph-theoretical terms via the molecular graph, is called a topological descriptor and if it in addition correlates with a molecular property it is called topological index, which is used to understand physicochemical properties of chemical compounds. Topological indices are interesting since they capture some of the properties of a molecule in a single number.

The first geometric-arithmetic index GA_1 was defined in [3] as

$$GA_1 = GA_1(G) = \sum_{uv \in E(G)} \frac{\sqrt{d_u d_v}}{\frac{1}{2}(d_u + d_v)}$$

where uv denotes the edge of the graph G connecting the vertices u and v, and d_u is the degree of the vertex u.

 ISBN: 978-84-617-8694-7

Herein, $G = (V(G), E(G))$ denotes a (nonoriented) finite simple (without multiple edges and loops) nontrivial ($E(G) \neq \emptyset$) graph. The aim of this paper is to obtain lower bounds on $GA_1(G)$ involving just the minimum and the maximum degree of the graph G and characterize graphs extremal with respect to them.

2 Main results

Let us recall Lemma 2.2 and Corollary 2.3 in [2].

Lemma 2.1. *Let f be the function $f(t) = \frac{2t}{1+t^2}$ on the interval $[0, \infty)$. Then f strictly increases in $[0, 1]$, strictly decreases in $[1, \infty)$, $f(t) = 1$ if and only if $t = 1$ and $f(t) = f(t_0)$ if and only if either $t = t_0$ or $t = t_0^{-1}$.*

Corollary 2.2. *Let g be the function $g(x, y) = \frac{2\sqrt{xy}}{x+y}$ with $0 < a \leq x, y \leq b$. Then $\frac{2\sqrt{ab}}{a+b} \leq g(x, y) \leq 1$. The equality in the lower bound is attained if and only if either $x = a$ and $y = b$, or $x = b$ and $y = a$, and the equality in the upper bound is attained if and only if $x = y$.*

Let us denote by $K_{\delta,\Delta}$ the complete bipartite graph with a partition K_1, K_2 with δ and Δ vertices respectively.

Given integers $0 < \delta \leq \Delta$, let us define $\mathcal{G}_{\delta,\Delta}$ as the set of graphs G with minimum degree δ, maximum degree Δ and such that:

(1) G is isomorphic to the complete graph with $\Delta + 1$ vertices $K_{\Delta+1}$, if $\delta = \Delta$,

(2) $|V(G)| = \Delta + 1$, there are Δ vertices with degree δ, if $\delta < \Delta$ and $\Delta(\delta + 1)$ is even,

(3) $|V(G)| = \Delta + 1$, there are $\Delta - 1$ vertices with degree δ and a vertex with degree $\delta + 1$, if $\delta < \Delta - 1$ and $\Delta(\delta + 1)$ is odd,

(4) $|V(G)| = \Delta + 1$, there are $\Delta - 1$ vertices with degree δ and two vertices with degree Δ, if $\delta = \Delta - 1$ and Δ is odd (and thus $\Delta(\delta + 1)$ is odd).

Proposition 2.3. *For any integers $1 < \delta \leq \Delta$, we have*

(1) *if $\frac{\Delta}{\delta} > \left(2 + \sqrt{3}\right)^2$, then $GA_1(H_{\delta,\Delta}) > GA_1(K_{\delta,\Delta})$,*

(2) *if $\frac{\Delta}{\delta} < \left(2 + \sqrt{3}\right)^2$ and $\Delta(\delta + 1)$ is even, then $GA_1(H_{\delta,\Delta}) < GA_1(K_{\delta,\Delta})$.*

It may be wondered if for any graph G with minimum degree δ and maximum degree Δ

$$GA_1(G) \geq \min\left\{GA_1(H_{\delta,\Delta}), GA_1(K_{\delta,\Delta})\right\}. \tag{1}$$

The following example shows that the answer is negative.

Example 2.4. *Let us suppose $\delta = 4$ and $\Delta = 56$. Consider a graph G with 57 vertices, two of them, a_1, a_2 with degree 56 and the rest, $b_1, \ldots, b_{55}$ with degree 4. Let us assume*

 ISBN: 978-84-617-8694-7

the edges are as follows. There is an edge a_ib_j for every i, j, an edge a_1a_2 and the vertices $b_1, \ldots, b_{55}$ induce a cycle of length 55.

Then, $GA_1(G) = \frac{2\cdot 110\sqrt{4\cdot 56}}{4+56} + 56 \approx 110.8776$.

However, $GA_1(H_{4,56}) = \frac{112\sqrt{224}}{60}+84 \approx 111.9377$, and $GA_1(K_{4,56}) = \frac{448\sqrt{224}}{60} \approx 111.7508$.

Nevertheless, we are able to find the critical lower bound in some cases:

Theorem 2.5. *Let G be a graph with minimum degree $\delta > 0$ and maximum degree $\Delta \geq 2$. If*

$$\frac{2\sqrt{\delta\Delta}}{\delta+\Delta} \geq \frac{\Delta(\delta-1)}{\Delta(\delta-1)+2}, \tag{2}$$

then

$$GA_1(G) \geq \frac{2\Delta\sqrt{\delta\Delta}}{\delta+\Delta} + \frac{\Delta(\delta-1)}{2}. \tag{3}$$

Furthermore, if $\Delta(\delta+1)$ is odd,

$$\frac{2\sqrt{\delta\Delta}}{\delta+\Delta} \geq \frac{\Delta(\delta-1)}{\Delta(\delta-1)+2} \quad \text{and} \quad \frac{3\sqrt{\delta\Delta}}{\delta+\Delta} + \delta - \frac{1}{2} \geq \frac{2\sqrt{(\delta+1)\Delta}}{\delta+1+\Delta} + \frac{2\delta\sqrt{\delta(\delta+1)}}{2\delta+1}, \tag{4}$$

then

$$GA_1(G) \geq \frac{2(\Delta-1)\sqrt{\delta\Delta}}{\delta+\Delta} + \frac{2\sqrt{(\delta+1)\Delta}}{\delta+1+\Delta} + \frac{2\delta\sqrt{\delta(\delta+1)}}{2\delta+1} + \frac{(\Delta-2)(\delta-1)-1}{2}. \tag{5}$$

If Δ and δ verify (2), then the equality in (3) is attained if and only if $\Delta(\delta+1)$ is even and $G \in \mathcal{G}_{\delta,\Delta}$. If Δ and δ verify (4) and $\Delta(\delta+1)$ is odd, then the equality in (5) is attained if and only if $G \in \mathcal{G}_{\delta,\Delta}$.

Corollary 2.6. *Let G be a graph with minimum degree $\delta > 0$ and maximum degree $\Delta = \delta + h \geq 2$. If we have*

(1) *$h = 0$ or $h = 1$, for every $\Delta \geq 2$,*
(2) *$h = 2$, for every $\Delta \geq 3$,*
(3) *$h = 3$, for every $\Delta \geq 4$,*
(4) *$h = 4$, for every $\Delta \geq 5$,*
(5) *$h = 5$, for every $\Delta \in \{6, 7, 8\}$,*
(6) *$h = 6$, for every $\Delta \in \{7, 8\}$,*
(7) *$h \geq 7$ and $\Delta = h + 1$,*

then

$$GA_1(G) \geq \frac{2\Delta\sqrt{\Delta(\Delta-h)}}{2\Delta-h} + \frac{\Delta(\Delta-h-1)}{2}.$$

Theorem 2.7. *Let G be a graph with minimum degree 2 and maximum degree $\Delta \geq 28$. Then,*

$$GA_1(G) \geq 2\Delta\frac{2\sqrt{2\Delta}}{\Delta+2},$$

and the equality is attained if and only if $G = K_{2,\Delta}$.

 ISBN: 978-84-617-8694-7

Acknowledgements

The firs author was partially supported by a grant from Ministerio de Economía y Competititvidad (MTM 2015-63612P), Spain, the second author by two grants from Ministerio de Economía y Competititvidad (MTM 2016-78227-C2-1-P and MTM 2015-69323-REDT), Spain, and a grant from CONACYT (FOMIX-CONACyT-UAGro 249818), México.

References

[1] A. Martínez-Pérez and J. M. Rodríguez *New lower bounds for the Geometric-Arithmetic Index*, MATCH Commun. Math. Comput. Chem. To appear.

[2] J. M. Rodríguez and J. M. Sigarreta, *On the Geometric-Arithmetic Index*, MATCH Commun. Math. Comput. Chem. **74** (2015) 103–120.

[3] D. Vukičević and B. Furtula, *Topological index based on the ratios of geometrical and arithmetical means of end-vertex degrees of edges*, J. Math. Chem. **46** (2009) 1369–1376.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

Low memory computation algorithm of recurrence plot of recurrence plots for long time series

Tomáš Martinovič[1]

[1] *IT4Innovations, VŠB - Technical University of Ostrava, Ostrava, Czech Republic*

emails: tomas.martinovic@vsb.cz

Abstract

Recurrence plot analysis is well established method to analyse time series in numerous areas of research. As the main result, the technique for computation of recurrence plot of recurrence plots(RPoRPs) is outlined. RPoRPs is extension of RP. This method significantly reduces spatial complexity of computation by computing RPoRPs directly from the time series. Additionally, a way to store RPoRPs in optimal format for diagonal recurrence quantitative analysis is proposed.

Key words: RPoRPs, recurrence plot, algorithms, time series
MSC 2000: 37M10,

1 Introduction

Recurrence quantitaive analysis (RQA) is a well established method for nonlinear time series analysis [5] [1]. RQA is the extension of a recurrence plot (RP), which is a visualisation tool for the recurrences in the state-space. Recurrence plots are used in number of disciplines such as medicine, engineering, economics and more, see e.g. [4], [8], [7] respectively.

Main advantages of RQA is its robustness to noise, possibility to study nonstationary time series and numerous extensions of RP, which allow study of stochastic and linear systems [2]. This paper is focused on computation of recurrence plot of recurrence plots (RPoRPs) [3], which is approximation of RP for very long time series. Main advantage of this technique is, that it computes RPoRPs directly from the time series and saves a lot of space.

 ISBN: 978-84-617-8694-7

2 Recurrence plot

Generally, recurrence plot is computed from the embedded time series. Embedding time series to the phase space is common practice in the nonlinear dynamical systems analysis. In case of observing one feature of multi dimensional dynamical system, embedding of this observed time series should re-create phase space with same dynamics as the original system see e.g. [9].

Let $\{x(t) \in \mathbb{R} | \, t = 1, 2, \ldots, n\}$ be observed time series of length n. Then embedded vector $X(t)$ at time t, is defined as $X(t) = [x(t), x(t+l), x(t+2l), \ldots, x(t+(m-1)l)]$, where t is observed time, l is delay time and m is embedding dimension.

First is computed the distance matrix of all the vectors $X(t)$. The euclidean distance matrix is computed as

$$D(t_1, t_2) = d(X(t_1), X(t_2)) = \sqrt{\sum_{k=0}^{m-1} (x(t_1 + kl) + x(t_2 + kl))^2}, \tag{1}$$

for all the pairs $X(t_1), X(t_2)$, where $t_1, t_2 \in \{1, 2, \ldots, n-(m-1)l\}$. Recurrence plot is then computed as

$$RP^{\varepsilon}(i,j) = \begin{cases} 1, & D(i,j) < \varepsilon \\ 0, & \text{otherwise,} \end{cases} \tag{2}$$

for $i, j \in \{1, 2, \ldots, n-(m-1)l\}$. This may be alternatively written as $RP^{\varepsilon}(i,j) = \Theta(D(i,j))$, where $\Theta(\cdot)$ is the Heaviside function.

3 Recurrence plot of recurrence plots

Recurrence plot of recurrence plots is a method to analyse long time series, for which it is normally impossible to compute recurrence plots since the number of elements in recurrence plot is exponential to the time series length. It was firstly proposed in Fukino[3]. RPoRPs is based on the windowed RP analysis as can be found in [6].

First the windowed RPs are computed. The elements of windowed RP are given by

$$RP_t^{\varepsilon}(i,j) = \Theta(d(S(ts+i), (S(ts+j)))), \tag{3}$$

where $t \in \{1, 2, \ldots, (t-(m-1)l-w)/s\}$ and $i, j \in \{1, 2, \ldots, w\}$. Analysis of windowed RPs is used to study local characteristics of the time series and to discover transitions in a dynamical system.

RPoRPs is defined as the thresholded distance matrix of all the windowed RPs.

$$RPoRP(i,j) = \Theta\left(\sqrt{\sum_{k,l=1}^{w} \left(RP_i(k,l) - RP_j(k,l)\right)^2}\right), \tag{4}$$

 ISBN: 978-84-617-8694-7

where i and j are RPoRPs element's indices. While windowed RPs are reflection of local characteristics. RPoRPs describes global characteristics of the time series.

RP and RPoRPs were tested on the Tent map T defined by

$$T(x) = 1 - |2x - 1|, \quad x \in [0, 1]. \tag{5}$$

The elements of time series are given by $x_t = T(x_{t-1})$, with the initial value $x_0 = \sqrt{(}2)/2$. Example of RP and RPoRPs of width 798 is shown in Figure 3 and Figure 3. 800 iterations of tent map were used to compute RP and 80050 iterations were used to compute the RPoRPs. The windowed RP used for the computation of RPoRPs had width of 400 and shift width was 100.

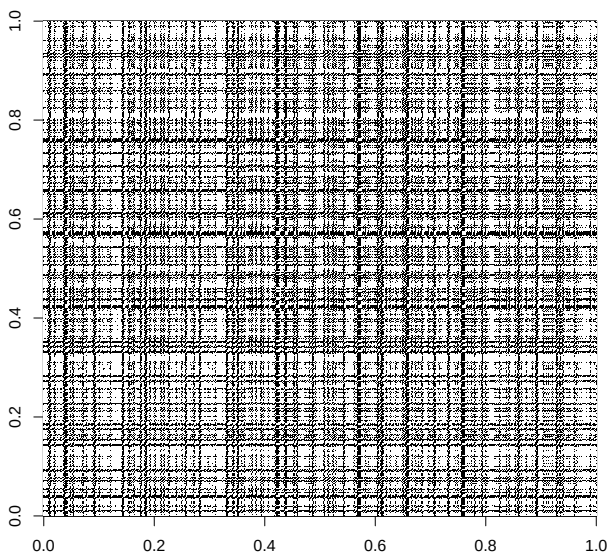


Figure 1: RP of tent map iterations

 ISBN: 978-84-617-8694-7

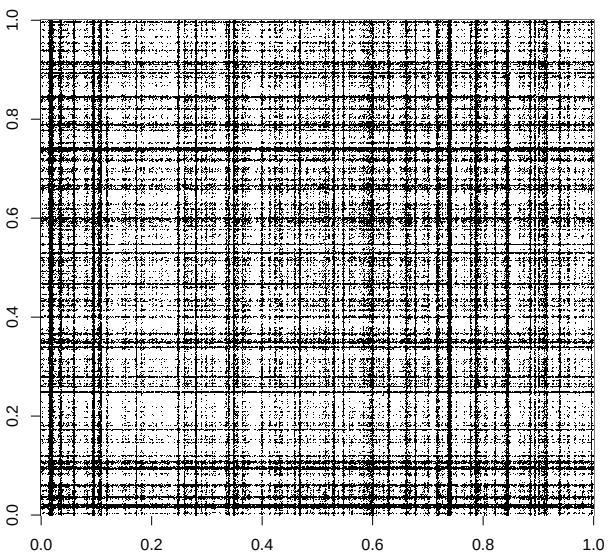


Figure 2: RPoRP of tent map iterations

4 Recurrence quantitative analysis

Although recurrence plots are powerful tool for time series analysis. Depending solely on the visual examination is often difficult. For long time series the visualisation of RPs is not exact, because the resolution of RP is higher than the screen resolution. Another thing is that, visual examination of RPs is not possible in automatic systems and therefore it is not suitable for practical use in automated tools.

Because of this, the RQA was created. It is quantification of main informations in the RPs. These can be divided into three categories: recurrence rate based, diagonal line based and vertical line based. Diagonal line based measures are generally connected with predictability of the system, while vertical line based measures are describing how long the system is trapped in the same state.

For RQA, measures of main interest are determinism, diagonal line length and entropy. Determinism is computed from the histogram of diagonal lines and is defined as

$$DET = \frac{\sum_{l=l_{min}}^{N} lP(l)}{\sum_{i,j=1}^{N} RP(i,j)}, \tag{6}$$

where l is the length of diagonal lines and $P(\cdot)$ is the histogram of diagonal lines. Deter-

 ISBN: 978-84-617-8694-7

minism is related to the predictability of the system, because deterministic process should contain many long diagonal lines in the RP and few single recurrence points. On the other hand, RP of noise, or random process should contain many single recurrence points with a few short diagonals.

5 Computation

Simple calculation of distance matrix requires computing n^2 distances for all pairs of n vectors. Naturally this may be reduced to computing only the lower/upper triangle, $n(n-1)/2$ pairs, since the distance matrix is symmetric and the main diagonal is always zero. Despite this reduction, the computational complexity grows exponentially. Equally grows memory requirements for storage of such distance matrix. For example to store the lower triangle of distance matrix for the time series of 100,000 observations would take approximately 372 GB. This means, it would be impossible to store it in RAM of most modern computers. This would make computation excessively difficult and computationally and spatially complex.

This issue is partially solved by windowed RPs, since their length is much smaller and therefore their memory requirements are acceptable. However to compute RPoRPs it is necessary to compute a large number of distance matrices, which are later used to compute RPoRPs.

However this is unnecessary, since the RPoRPs can be effectively computed from the original time series. The formula for computing RPoRPs directly from the time series is simple expansion of the equation (4)

$$d_{RPoRP}(i,j) = \sqrt{2\sum_{a=0}^{w-1}\sum_{b=a+1}^{w}\left(\sqrt{\sum_{n=0}^{m-1}\alpha_n^2}-\sqrt{\sum_{n=0}^{m-1}\beta_n^2}\right)}, \tag{7}$$

$$\alpha_n = x(t_{i+a}+nl) - x(t_{i+b}+nl), \tag{8}$$

$$\beta_n = x(t_{j+a}+nl) - x(t_{j+b}+nl), \tag{9}$$

where $x(t)$ is observation of time series x at time t, m is embedding dimension, l is time delay and w is distance matrix window width.

To reduce cache misses it is important to compute elements of RPoRPs in diagonal order. This allows to compute diagonal RQA measures by cycling through consecutive elements of array containing RPoRPs, so the prefetching mechanism may easily predict which elements will be used.

6 Conclusion

It was outlined how to compute RPoRPs directly from the time series by exploiting the definition of RPoRPs. This method introduces much lower spatial complexity of computa-

 ISBN: 978-84-617-8694-7

tion, significantly increasing performance and reducing cache misses when computing RQA. This should open up new possibilities in RP analysis, allowing study of much longer time series with multiple resolutions. Moreover, simulations of RP and RPoRPs were done on experimental example based on the Tent map.

Acknowledgements

This work was supported by The Ministry of Education, Youth and Sports from the National Programme of Sustainability (NPU II) project "IT4Innovations excellence in science - LQ1602" and by the IT4Innovations infrastructure which is supported from the Large Infrastructures for Research, Experimental Development and Innovations project "IT4Innovations National Supercomputing Center LM2015070". This work was partially supported by grant of SGS No. SP2017/182 "Solving graph problems on spatio-temporal graphs with uncertainty using HPC", VŠB - Technical University of Ostrava, Czech Republic.

References

[1] U. R. Acharya et al., *Application of recurrence quantification analysis for the automated identification of epileptic EEG signals*, Int. J. of Neural Systems **21**(3) (2011) 199–211.

[2] M. C. Casdagli, *Recurrence plots revisited*, Physica D **108** (1997) 12–44.

[3] M. Fukino, Y. Hirata, K. Aihara, *Coarse-graining time series data: Recurrence plot of recurrence plots and its application for music*, Chaos **26** 023116 (2016).

[4] L. Lancia, D. Volgt, G. Krasovitskiy, *Characterization of laryngealization as irregular vocal fold vibration and interaction with prosodic prominence*, J. of Speech, Language, and Hearing Research **54** (2016) 80–97.

[5] N. Marwan et al., *Recurrence plots for the analysis of complex systems* Physics Report **438** (2007) 237–329.

[6] N. Marwan, *How to avoid potential pitfalls in recurrence plot based data analysis*, Int. J. of Bifurcation and Chaos **21**(4) (2011) 1003–1017.

[7] M. D. McKenzie, *Chaotic behavior in national stock market indices*, Global Finance J. **12**(1) (2001) 35–53.

[8] P. Sharma, N. Murali, T. Jayakumar, *A time-frequency analysis of temperature fluctuations in a fast reactor*, Proceedings of the 5th International Congress on Image and Signal processing (CISP2012)(6469737), (2012) 1546–1550.

 ISBN: 978-84-617-8694-7

[9] F. Takens, *Detecting strange attractors in turbulence*, Dynamical Systems and Turbulence, Lecture Notes in Mathematics vol.898, Springer-Verlag (1981) 366–381.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

On Variance Equality for Gaussian Mixtures

Miguel Martins Felgueiras[1], Rui Filipe Santos[2] and João Paulo Martins[2]

[1] *School of Technology and Management and CIGS, Polytechnic Institute of Leiria, CEAUL — Center of Statistics and Applications of University of Lisbon*

[2] *School of Technology and Management, Polytechnic Institute of Leiria, CEAUL — Center of Statistics and Applications of University of Lisbon*

emails: mfelg@ipleiria.pt, rui.santos@ipleiria.pt, jpmartins@ipleiria.pt

Abstract

In this work we investigate convex Gaussian mixtures, namely an usually considered statistical hypothesis in data analysis, the variance equality. Under variance equality hypothesis it is possible to avoid a large number of parameters and to rewrite the mixture. Moreover, approximations to simpler distributions can sometimes be achieved. These approximations, namely to the Beta distribution, can be used to perform a variance equality test.

Key words: Gaussian mixtures, variance equality.

1 Introduction

Let us define a convex Gaussian mixture of N subpopulations as a random variable with density function

$$f_X(x) = \sum_{j=1}^{N} w_j \frac{1}{\sqrt{2\pi}\sigma_j} \exp\left\{-\frac{1}{2}\left(\frac{x-\mu_j}{\sigma_j}\right)^2\right\}, \quad \sigma_j > 0, w_j > 0, \sum_{j=1}^{N} w_j = 1, \tag{1}$$

where μ_i, σ_i and ω_i denote the mean, the variance and the weight of subpopulation i.

This type of mixtures is often used in statistical data analysis, because convex gaussian mixtures are very flexible. They can deal with multimodality, different shapes and always have finite moments, which is a plus. In a remarkable work of 1894 [6], Pearson used a two component Gaussian to fit biological data. This was, as far as we now, the first use of

 ISBN: 978-84-617-8694-7

Gaussian mixtures in a real data situation, together with the presentation of several theoretical aspects. [1] and its followers used these mixtures when dealing with financial data, usually with a small number of components. A recent application to image processing can be found in [4]. Image deconvolution is also a prolific field of Gaussian mixture's applications. A reference book for finite Gaussian (and non Gaussian) mixtures is [5], containing many applications to several fields of knowledge together with theoretical background.

2 Cumulants, Moments and Simplifications

For a random variable (r.v.) X with density as stated in (1), moments can be retrieved from the cumulant generation function,

$$\ln[\varphi_X(-it)] = \kappa_{1,X}t + \kappa_{2,X}\frac{t^2}{2!} + \kappa_{3,X}\frac{t^3}{3!} + \kappa_{4,X}\frac{t^4}{4!} + O\left(t^5\right) \tag{2}$$

with

$$\kappa_{1,X} = \mu'_{1,X} \quad \kappa_{2,X} = \mu_{(2,X)} \quad \kappa_{3,X} = \mu_{(3,X)} \quad \kappa_{4,X} = \mu_{(4,X)} - 3\mu^2_{(2,X)}, \tag{3}$$

where $\mu_{(k,X)}$ is the centred k-moment, $\mu'_{k,X}$ represents the raw k-moment and $\kappa_{k,X}$ denotes the k-cumulant of r.v. X.

One common assumption is to consider the same variance for all the components, that is to consider $\sigma_j = \sigma$ for $j = 1, ..., N$. When all the subpopulations have the same variance, [2] showed that

$$X \overset{d}{=} V + Y \tag{4}$$

where V and Y are independent random variables, $V \sim N(0, \sigma)$ and Y is a discrete random variable where $P(Y = \mu_j) = w_j$, for $j = 1, ..., N$.

Thus, using equations (3) and (4) when $\sigma_j = \sigma$ for $j = 1, \ldots, N$, the mixture cumulants can be rewritten as a sum. Using cumulants properties,

$$\kappa_{1,X} = \kappa_{1,Y} \quad \kappa_{2,X} = \kappa_{2,Y} + \sigma^2 \quad \kappa_{3,X} = \kappa_{3,Y} \quad \kappa_{4,X} = \kappa_{4,Y}. \tag{5}$$

The equality $\kappa_{2,X} = \kappa_{2,Y} + \sigma^2$ is important, because the knowledge of $\kappa_{2,Y} = \mu_{(2,Y)}$ or σ^2 is enough to estimate all the other parameters by the moments method, and even to fit (in some situations) an Y distribution.

3 Pearson System Approximation

Let $\beta_{1,X}$ (skewness) and $\beta_{2,X}$ (kurtosis) coefficients be defined as

$$\beta_{1,X} = \frac{\mu_{(3,X)}}{\mu^{3/2}_{(2,X)}} \qquad \beta_{2,X} = \frac{\mu_{(4,X)}}{\mu^2_{(2,X)}}. \tag{6}$$

 ISBN: 978-84-617-8694-7

Any unimodal mixture can be approximated by a Pearson system distribution [3]. The selected distribution depends on $\beta_{1,X}$ and $\beta_{2,X}$.

For unimodal gaussian mixtures where $\sigma_j = \sigma$ for $j = 1, ..., N$, the Pearson type I (four parameters beta) approach arises in multiple situations, namely when

$$\mu_{(2,Y)}\sqrt{\frac{1.5\beta_{1,Y}^2 - \beta_{2,Y} + 3}{3}} - \mu_{(2,Y)} < \sigma^2 < \frac{1.5\beta_{1,Y}^2\mu_{(2,Y)}}{\beta_{2,Y} - 3} - \mu_{(2,Y)}. \quad (7)$$

However, in most situations it is impossible to decompose the mixture into a random variable sum, since both $\mu_{(2,Y)}$ and σ^2 are unknown. Therefore, the use of the above equation in estimation issues or when testing variance equality is limited.

4 Gaussian Mixtures with Two Components

Since the general result stated in the previous section has limited usefulness, we will present an approximation that works well for unimodal Gaussian mixtures with two components. This approximation allows us to develop a variance equality test based on Beta distribution fit. The test quality is ascertain using both simulation studies and real data applications.

Acknowledgements

Funded by FCT - Fundação para a Ciência e a Tecnologia, Portugal, through the project UID/MAT/00006/2013.

References

[1] Fama E., *The behavior of stock-market prices*, J. Bus. **38** (1965) 34–105.

[2] Felgueiras, M., Santos, R., Martins, J., *Some Results on Gaussian Mixtures*, AIP Conf. Proc. **1618** (2014) 523–526.

[3] Johnson, N., Kotz, S., Balakrishnan, N., *Continuous Univariate Distributions, Volume I*, Wiley, New York, (1994).

[4] Kalti K., Mahjoub M., *Image Segmentation by Gaussian Mixture Models and Modified FCM Algorithm*, Int. Arab J. Inf. Techn. **11 1** (2014) 11–18.

[5] McLachlan G., Peel D., *Finite Mixture Models*, Wiley, New York, (2000).

[6] Pearson K., *Contributions to the mathematical theory of evolution*, Phil. Trans. R. Soc. **A 185** (1894) 71–110.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

The reinfection threshold in the SIRI model

José Martins[1], Alberto Pinto[2] and Nico Stollenwerk[3]

[1] *LIAAD-INESC TEC and Department of Mathematics, School of Technology and Management, Polytechnic Institute of Leiria, Portugal*

[2] *LIAAD-INESC TEC and Department of Mathematics, Faculty of Sciences, University of Porto, Portugal*

[3] *CMAF-CIO, Faculty of Science,University of Lisbon, Portugal*

emails: jmmartins@ipleiria.pt, aapinto@fc.up.pt, nico@ptmat.fc.ul.pt

Abstract

The reinfection SIRI model describes the spreading of an epidemics in a population of susceptible (S), infected (I), and recovered (R) individuals, where after an initial infection the recovered individuals only have partial immunity against a possible re-infection. Grassberger, Chaté and Rousseau considered similar models with partial immunization, and observed transitions between phases of no-growth, annular growth and compact growth (see [5]). The transition between no-growth and annular growth corresponds to the transition between the disease-free equilibrium and the endemic equilibrium and so it is well characterized. The characterization of the transition between annular growth and compact growth is much more dubious because it not corresponds to a sharp threshold. This is transition is known as the reinfection threshold. In this work, we propose a new approach to characterize the reinfection threshold based on the curvature of the infected individuals quantity.

Key words: Reinfection threshold, SIRI model

1 Introduction

In 2004, Gomes, White and Medley pointed out the attention to the epidemic models with partial immunity with the discovery of the reinfection threshold [3]. Besides the well known first infection threshold, between a disease free state and a non-trivial state with strictly positive endemic equilibrium, they found a second threshold characterized by the

 ISBN: 978-84-617-8694-7

ratio between the infection and reinfection rates. They called to this second threshold the "reinfection threshold". The existence of the reinfection threshold is an extremely important discovery because partial immunity is a frequent feature of many pathogens, like Neisseria meningitidis, Streptococcus pneumoniae, Mycobacterium tuberculosis (see [3, 8]). But the concept of the reinfection threshold was not consensual and was under debate in the following years (see [1, 4]) The main concern with this new concept was that the so called reinfection threshold does not exhibit a sharp threshold behaviour and therefore it is dubious that corresponds to a phase transition.

Some years before, Grassberger, Chaté and Rousseau considered a model for the spreading of an agent in a medium whose susceptibility changes irreversibly at the first infection, that shows transitions between phases of no growth below the direct percolation or dynamical percolation critical points which are joined by a phase transition line, and another transition between annular growth and compact growth in two and higher spatial dimensions (see [5, 6]). The model considered in their work has the possibility to model epidemics with partial immunization, where after a first infection the recovered host only have partial immunity against the pathogen or a genetically close mutant pathogen. Hence, the three phases of growth are expected to be observed in epidemic models with partial immunity and more than one phase transition should exist.

Here, we consider the basic reinfection SIRI epidemic model in a stochastic spatial setup, and we analyze dynamics of the mean total numbers of susceptibles, infected and recovered, in the mean field approximation. The deterministic models considered before by Gomes, White and Medley [3] where the reinfection threshold was found are automatically related with the mean field dynamics considered in this work. In the stochastic version of the SIRI model considered here, we introduce a transition from recovered to susceptibles as effect of temporary immunity with rate α. In the limit of vanishing α, we observe a sharp threshold behaviour characterized by the parameter values of the reinfection threshold of Gomes, White and Medley [3]. Indeed, for positive values of α the observed threshold is not a sharp threshold.

In this work, we propose a new approach to characterize analytically the reinfection threshold even for positive values of α. Our approach is based on the curvature of the infected individuals quantity at stationarity.

 ISBN: 978-84-617-8694-7

2 The reinfection SIRI model

For the basic reinfection SIRI epidemic model, we have the transitions identified by the following reaction schemes (see [7, 8]))

$$S + I \xrightarrow{\beta} I + I$$

$$I \xrightarrow{\gamma} R$$

$$R + I \xrightarrow{\tilde{\beta}} I + I$$

$$R \xrightarrow{\alpha} S$$

where the transition rates are: the infection rate β as a transition from susceptible S to infected I, the recovery rate γ from I into the recovered R class, the reinfection rate $\tilde{\beta}$ as effect of only partial immunization of the recovered, and a transition rate α from recovered R to susceptibles S as effect of temporary immunization. Since we consider the SIRI model in a spatial setup , we have for the variables S_i, I_i and $R_i \in \{0, 1\}$ the constraint that an individual i belongs to one of the three classes

$$S_i + I_i + R_i = 1 \quad .$$

Let J be the $N \times N$ adjacency matrix that describes the neighbouring structure of the N individuals of the population. The matrix J is symmetric, with zero diagonal elements, and with the elements $J_{i,j} \in \{0, 1\}$ defined such that: $J_{i,j} = 1$ if the individual i is a neighbour of j, and $J_{i,j} = 0$ if the individual i is not a neighbour of j. The time evolution of the probability of the state $S_1, I_1, R_1, S_2, I_2, R_2, ..., R_N$ occur at time t is stated by the master equation, and can be used to obtain the dynamic equation for the expected values of the state variables (see [8]).

Let $\langle I \rangle(t)$ denotes the mean value of the total number of infected hosts at a given time t

$$\langle I \rangle(t) \;=\; \sum_{S_1=0}^{1} \sum_{I_1=0}^{1} \sum_{R_1=0}^{1} ... \sum_{R_N=0}^{1} \left(\sum_{i=1}^{N} I_i \right) p(S_1, I_1, R_1, ..., R_N, t) \, .$$

Taking the time derivative of $\langle I \rangle(t)$ and applying the master equation, that defines $\frac{d}{dt} p(S_1, I_1, R_1, S_2, ..., R_N, t)$, we obtain after some calculations, the time evolution of $\langle I \rangle(t)$.

 ISBN: 978-84-617-8694-7

In terms of all variables S, I and R, we obtain the ODE system

$$\begin{aligned}
\frac{d}{dt}\langle S\rangle &= \alpha\langle R\rangle - \beta\,\langle SI\rangle_1 \\
\frac{d}{dt}\langle I\rangle &= \beta\,\langle SI\rangle_1 - \gamma\langle I\rangle + \tilde{\beta}\,\langle RI\rangle_1 \\
\frac{d}{dt}\langle R\rangle &= \gamma\langle I\rangle - \alpha\langle R\rangle - \tilde{\beta}\,\langle RI\rangle_1
\end{aligned}$$

where e.g.

$$\langle RI\rangle_1(t) = \sum_{S_1=0}^{1} ... \sum_{R_N=0}^{1} \left(\sum_{i=1}^{N}\sum_{j=1}^{N} (J^1)_{ij} R_i I_j \right) \cdot p(S_1, ..., R_N, t)$$

is the mean number of pairs of recovered next to infected. In the equation for the dynamics of $\langle RI\rangle_1$ also an expression $\langle RI\rangle_2$ could show up. These are longer range correlations, formally given by a power of two of the adjacency matrix J^2 and then its elements $(J^2)_{ij} = \sum_{k=1}^{N} J_{ik} \cdot J_{kj}$.

2.1 The mean field approximation

From now on we will only consider regular lattices where all sites has the same number of neighbours, hence, for all i,

$$\sum_{j=1}^{N} J_{ij} = Q_i = Q\,.$$

In mean field approximation we assume that the exact number of inhabited infected neighbors is replaced by the average number of infected individuals in the full system, acting like a mean field on the actually considered site

$$\sum_{j=1}^{N} J_{ij} I_j \approx \sum_{j=1}^{N} J_{ij} \frac{\langle I\rangle}{N} = \frac{Q}{N}\langle I\rangle$$

Hence, we obtain

$$\langle RI\rangle_1(t) \approx \frac{Q}{N}\langle R\rangle\langle I\rangle\,.$$

With similar approximations for the other pairs, we obtain closed ODE system for the

 ISBN: 978-84-617-8694-7

expectation values of S, I and R:

$$\begin{aligned}
\frac{d}{dt}\langle S\rangle &= \alpha\langle R\rangle - \beta\frac{Q}{N}\,\langle S\rangle\langle I\rangle \\
\frac{d}{dt}\langle I\rangle &= \beta\frac{Q}{N}\,\langle S\rangle\langle I\rangle - \gamma\langle I\rangle + \tilde{\beta}\frac{Q}{N}\,\langle R\rangle\langle I\rangle \\
\frac{d}{dt}\langle R\rangle &= \gamma\langle I\rangle - \alpha\langle R\rangle - \tilde{\beta}\frac{Q}{N}\,\langle R\rangle\langle I\rangle \quad .
\end{aligned}$$

2.2 Stationarity analysis

The previous ODE system can be studied in the scaled quantities, time changed to $\tau = t/\gamma$ and consequently

$$\rho = \beta Q/\gamma\,, \qquad \varepsilon = \alpha/\gamma$$

and the ratio of infectivities given by

$$\sigma = \tilde{\beta}/\beta \quad .$$

Further, we consider densities of susceptibles, infected, and recovered $s = \langle S\rangle/N$, $\quad i = \langle I\rangle/N$, $\quad r = \langle R\rangle/N = 1 - s - i$, and we obtain the two-dimensional ODE system:

$$\begin{aligned}
\frac{d}{d\tau}s &= \varepsilon(1 - s - i) - \rho s i \\
\frac{d}{d\tau}i &= \rho i(s + \sigma(1 - s - i)) - i
\end{aligned}$$

The stationary solution is either the disease-free state

$$i_1^* = 0$$

or the endemic state

$$i_2^* = -\frac{r}{2} + \sqrt{\frac{r^2}{4} - q}$$

with

$$r = \frac{1}{\rho\sigma}\,(1 - \rho\sigma + \varepsilon) \qquad \textit{and} \qquad q = \frac{\varepsilon}{\rho^2\sigma}\,(1 - \rho)\,.$$

The first infection threshold occurs when the stationary solutions i_1^* and i_2^* meet each other, i.e. $i_1^* = i_2^* = 0$, at

$$\rho = 1\,.$$

 ISBN: 978-84-617-8694-7

When $\varepsilon = 0$, we obtain another change of regime at

$$\rho = \frac{1}{\sigma}\,.$$

This second change of regime was called the reinfection threshold by Gomes *et al.* [3], and appears as a sharp threshold in the limit of ε decreasing to zero.

We observe that for large values of ε the behaviour of i^* is dominated by the first threshold behaviour around $\rho = 1$. In contrast, for small values of ε there is only the qualitative behaviour left from the behaviour around the second threshold $\rho = 1/\sigma$. The change between these two extremes is quite continuous, however, close to the reinfection threshold $\rho = 1/\sigma$, the solutions for i^* for small ε are a smoothened out version of that threshold. It is in this smoothened transition that we propose a new definition for the reinfection threshold. For positive values of $\epsilon = \alpha/\gamma$ we propose a new approach to the reinfection threshold based on the parameter values that maximize the curvature of the stationary value of infecteds

$$arg\max_{\rho} K\left(i(\rho)\right),$$

where K defines the curvature of the endemic state. The curvature of a smooth curve measures how fast the curve is changing direction at a given point. For the endemic stationary value of infecteds this maximum change is the closest behavior to a sharp threshold and therefore is the best possible approach to the reinfection threshold.

3 Conclusions and future work

We considered the spatial stochastic reinfection SIRI model and we made a stationary analysis in the mean field approximation. Taking the limit of vanishing the endemic equilibrium, we found the first infection threshold. In analogy with other spreading models, this first threshold corresponds to the transition between no-growth and annular growth (see [5, 2]). The second transition between annular growth and compact growth, called in an epidemic context "the reinfection threshold", was also found for the mean field dynamics. Since this second transition was under discussion for a while, we propose in this work a new approach to characterize the reinfection threshold. This approach will be used to make an analytical characterization of the reinfection threshold for some parameters values where the reinfection threshold was questionable. The characterization of the phase transition lines can also be made considering the pair approximation dynamics instead of the mean field dynamics [7].

 ISBN: 978-84-617-8694-7

Acknowledgements

The authors thank the financial support of LIAAD-INESC TEC and FCT Fundação para a Ciência e Tecnologia (Portuguese Foundation for Science and Technology) within project UID/EEA/50014/2013 and ERDF (European Regional Development Fund) through the COMPETE Program (operational program for competitiveness) and by National Funds through the FCT within Project "Dynamics, optimization and modelling", with reference PTDC/MAT-NAN/6890/2014 and Project "NanoSTIMA: Macro-to-Nano Human Sensing: Towards Integrated Multimodal Health Monitoring and Analytics/NORTE-01-0145-FEDER-000016" financed by the North Portugal Regional Operational Programme (NORTE 2020), under the PORTUGAL 2020 Partnership Agreement, and through the European Regional Development Fund (ERDF).

References

[1] R. Breban and S. Blower, *The reinfection threshold does not exist*, J. Theor. Biology, **235** (2005) 151-152.

[2] S.M. Dammer and H. Hinrichsen, *Spreading with immunization in high dimensions*, J. Stat. Mech: Theor Exp., **7** (2004) P07011.

[3] M.G.M. Gomes, L.J. White and G.F. Medley, *Infection, reinfection, and vaccination under suboptimal immune protection: epidemiological perspectives* J. Theor. Biol., **228** (2004) 539-549.

[4] M.G.M. Gomes, L.J. White and G.F. Medley, *The reinfection threshold*, J. Theor. Biology, **236** (2005) 111-113.

[5] P. Grassberger, H. Chat and G. Rousseau, *Spreading in media with long-time memeory*, Phys. Rev. E, **55** (1997) 2488-2495.

[6] P. Grassberger, *On the critical behavior of the general epidemic process and dynamical percolation*, Math. Biosci., **63** (1983) 157-172.

[7] N. Stollenwerk, J. Martins and A. Pinto, *The phase transition lines in pair approximation for the basic reinfection model SIRI*, Phys. Lett. A, **371** (2007) 379–388.

[8] N. Stollenwerk, S. van Noort, J. Martins, M. Aguiar, F. Hilker, A. Pinto and M.G. Gomes, *A spatially stochastic epidemic model with partial immunization shows in mean field approximation the reinfection threshold*, J. Biol. Dyn., **4** (2010) 634649.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

Glutamate dehydrogenase enzyme immunoassays: a meta-analysis with a Bayesian approach

João Paulo Martins[1], Miguel Felgueiras[2] and Rui Santos[1]

[1] *School of Technology and Management, Polytechnic Institute of Leiria, CEAUL – Center of Statistics and Applications of University of Lisbon*

[2] *School of Technology and Management and CIGS, Polytechnic Institute of Leiria, CEAUL – Center of Statistics and Applications of University of Lisbon*

emails: jpmartins@ipleiria.pt, mfelg@ipleiria.pt, rui.santos@ipleiria.pt

Abstract

Several commercial enzyme immunoassays are available for screening glutumate dehydrogenase. In this work, a meta-analysis is performed using the results provided in 22 papers published since 2009. The sensitivity and specificity of the index tests is assessed jointly using hierarchical summary receiver operatoring characteristic (HSROC) models estimated under a Bayesian setting. Results show a specificity of the index tests lower than what has been published in previous meta-analysis.

Key words: C. difficile, meta-analysis, HSROC
MSC 2000: 62F15, 62P10

1 Introduction

To the best of our knowledge, there are three meta-analysis published concerning the accuracy of the commercial EIA tests for GDH screening. The first one was published in 2011. It was performed by Shetty *et al.* [22] based on 13 full papers. In 2016, two meta-analysis were published almost simultaneously. Arimoto *et al.* [1] undertook a meta-analysis using a large number of studies (42), some of them obtained from conference abstracts or posters. However, despite being a recent review it uses tests no longer available in the market and that perform worse than those that are actually available (according to their results). This may have led to under estimate the sensitivity of the EIA tests. Crobach *et al.* work [5] aimed to update the guidelines for *C. difficile* diagnostic published in 2009. They updated their previous sample (and excluded tests that are not commercially available) resulting in the use of 16 full papers. Our meta-analysis is based on 22 papers [2-4;6-21;23-25].

 ISBN: 978-84-617-8694-7

Acknowledgements

Funded by FCT - Fundação para a Ciência e a Tecnologia, Portugal, through the project UID/MAT/00006/2013.

References

[1] J. Arimoto, N. Horita, S. Kato, A. Fuyuki, T. Higurashi, H. Ohkubo, H. Endo, N. Takashi, T. Kaneko and A. Nakajima, *Diagnostic test accuracy of glutamate dehydrogenase for Clostridium difficile: Systematic review and meta-analysis*, Sci. Rep. **6** (2016) 29754.

[2] J. A. Barkin, E. Miller, A. Grace, J. S. Barkin and D. A. Sussman, *Superiority of the DNA amplification assay for the diagnosis of C. difficile infection: a clinical comparison of fecal tests*, Dig. Dis. Sci. **57** (2012) 2592–2599.

[3] N. A. Brown, W. D. Lebar, C. L. Young, R. E. Hankerd and D. W. Newton, *Diagnosis of Clostridium difficile infection: comparison of four methods on specimens collected in Cary-Blair transport medium and tcdB PCR on fresh versus frozen samples*, Infect. Dis. Rep. **3** (2011) 15–19.

[4] M. J. Bruins, E. Verbeek, J. A. Wallinga, L. E. S. Bruijnesteijn van Coppenraet, E. J. Kuijper and P. Bloembergen, *Evaluation of three enzyme immunoassays and a loop-mediated isothermal amplification test for the laboratory diagnosis of Clostridium difficile infection*, Clin. Microbiol. Infect. **31** (2012) 3035–3039.

[5] M. J. T. Crobach, T. Planche, C. Eckert, F. Barbut, E. M. Terveer, O. M. Dekkers, M. H. Wilcox and E. J. Kuijper, *European Society of Clinical Microbiology and Infectious Diseases: update of the diagnostic guidance document for Clostridium difficile infection*, Clin. Microbiol. Infect. **22** (2016) 563–581.

[6] K. Eastwood, P. Else, A. Charlett and M. Wilcox, *Comparison of nine commercially available Clostridium difficile toxin detection assays, a real-time PCR assay for C. difficile tcdB, and a glutamate dehydrogenase detection assay to cytotoxin testing and cytotoxigenic culture methods*, J. Clin. Microbiol. **47** (2009) 3211–3217.

[7] C. Eckert, E. Holscher, A. Petit, V. Lalande and F. Barbut, *Molecular test based on isothermal helicase-dependent amplification for detection of the Clostridium difficile toxin A gene*, J. Clin. Microbiol. **52** (2014) 2386–2389.

[8] S. D. Goldenberg, N. M. Price, D. Tucker, P. Wade P and G. L. French, *Mandatory reporting and improvements in diagnosing Clostridium difficile infection: an incompatible dichotomy?*, J. Infect. **62** (2011) 363–370.

 ISBN: 978-84-617-8694-7

[9] W. Jamal, E. M. Pauline, V. O. Rotimi, *Comparative performance of the GeneXpert C. difficile PCR assay and C. diff Quik Chek Complete kit assay for detection of Clostridium difficile antigen and toxins in symptomatic community-onset infections*, Int. J. Infect. Dis. **29** (2014) 244–248.

[10] M. Kawada, M. Annaka, H. Kato, S. Shibasaki, K. Hikosaka, H. Mizuno, Y. Masuda and T. Inamatsu, *Evaluation of a simultaneous detection kit for the glutamate dehydrogenase antigen and toxin A/B in feces for diagnosis of Clostridium difficile infection*, J. Infect. Chemother. **17** (2011) 807–811.

[11] H. Kim, W. H. Kim, M. Kim, S. H. Jeong and K. Lee, *Evaluation of a rapid membrane enzyme immunoassay for the simultaneous detection of glutamate dehydrogenase and toxin for the diagnosis of Clostridium difficile infection*, Ann. Lab. Med. **34** (2014) 235–239.

[12] E. J. Kvach, D. Ferguson, P. F. Riska and M. L. Landry, *Comparison of BD GeneOhm Cdiff Real-Time PCR Assay with a two-step algorithm and a toxin A/B enzyme-Linked immunosorbent assay for diagnosis of toxigenic Clostridium difficile infection*, J. Clin. Microbiol. **48** (2010) 109–114.

[13] A. M. Larson, A. M. Fung and F. C. Fang, *Evaluation of tcdB real-time PCR in a three-step diagnostic algorithm for detection of toxigenic Clostridium difficile*, J. Clin. Microbiol. **48** (2010) 124–130.

[14] S. Miller, A. Wiita, C. Wright, H. Reyes and C. Liu, *Evaluation of glutamate dehydrogenase immunoassay screening with toxin confirmation for the diagnosis of Clostridium difficile infection* , Lab. Med. **44** (2013) 65–71.

[15] K. V. Ota and K. L. McGowan, *Clostridium difficile testing algorithms using glutamate dehydrogenase antigen and C. difficile toxin enzyme immunoassays with C. difficile nucleic acid amplification testing increase diagnostic yield in a tertiary pediatric population*, J. Clin. Microbiol. **50** (2012) 1185–1188.

[16] L. R. Peterson, M. S. Mehta, P. A. Patel, D. M. Hacek, M. Harazin, P. P. Nagwekar, R. B.Thomson and A. Robicsek, *Laboratory testing for Clostridium difficile infection: light at the end of the tunnel*, Am. J. Clin. Pathol. **136** (2011) 372–380.

[17] T. D. Planche, K. A. Davies, P. G. Coen, J. M. Finney, I. M. Monahan, K. A. Morris, L. O'Connor, S. J. Oakley, C. F. Pope, M. W. Wren, N. P. Shetty, D. W. Crook and M. H. Wilcox, *Differences in outcome according to Clostridium difficile testing method: a prospective multicentre diagnostic validation study of C difficile infection*, Lancet. Infect. Dis. **13** (2013) 936–945.

 ISBN: 978-84-617-8694-7

[18] M. O. Qutub, N. AlBaz, P. Hawken and A. Anoos, *Comparison between the two-step and the three-step algorithms for the detection of toxigenic Clostridium difficile*, Indian J. Med. Microbiol. **29** (2011) 293–296.

[19] M. E. Reller, R. C. Alcabasa, C. A. Lema and K. C. Carroll, *Comparison of two rapid assays for Clostridium difficile common antigen and a C. difficile toxin A/B assay with the cell culture neutralization assay* , Am. J. Clin. Pathol. **133** (2010) 107–110.

[20] S. B. Selvaraju, M. Gripka, K. Estes, A. Nguyen, M. A. Jackson and R. Selvarangan, *Detection of toxigenic Clostridium difficile in pediatric stool samples: an evaluation of Quik Check Complete Antigen assay, BD GeneOhm Cdiff PCR, and ProGastro Cd PCR assays*, Diagn. Microbiol. Infect. Dis. **71** (2011) 224–229.

[21] M. Senoh, H. Kato, T. Murase, H. Hagiya, Y. Tagashira, T. Fukuda, M. Iwaki, A. Yamamoto and K. Shibayama, *Reverse transcription polymerase chain reaction-based method for selectively detecting vegetative cells of toxigenic Clostridium difficile*, Microbiol. Immunol. **58** (2014) 615–620.

[22] N. Shetty, M. W. D. Wren, P. G. Coen, *The role of glutamate dehydrogenase for the detection of Clostridium difficile in faecal samples: a meta-analysis*, J. Hosp. Infect. **77** (2011) 1–6.

[23] B. M. Shin, E. J. Lee, J. W. Moon and S. Y. Lee, *Evaluation of the VIDAS glutamate dehydrogenase assay for the detection of Clostridium difficile*, Anaerobe **40** (2016) 68–72.

[24] J. Swindells, N. Brenwald, N. Reading and B. Oppenheim, *Evaluation of diagnostic tests for Clostridium difficile infection*, J. Clin. Microbiol. **48** (2010) 606–608.

[25] A. Walkty, P. R. Lagac-Wiens, K. Manickam, H. Adam, P. Pieroni, D. Hoban, J. A. Karlowsky and M. Alfa, *Evaluation of an algorithmic approach in comparison with the Illumigene assay for laboratory diagnosis of Clostridium difficile infection*, J. Clin. Microbiol. **51** (2013) 1152–1157.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

Effective fluid flow trough corrugated pipe and the Darcy-Weisbach law

Eduard Marušić-Paloka[1] **and Maja Starčević**[1]

[1] *Department of mathematics, University of Zagreb*

emails: emarusic@math.hr, mstarcev@math.hr

Abstract

We study an incompressible viscous fluid flow through a pipe with rough wall. Starting from the Stokes system, prescribing the pressure drop on pipe's ends and assuming the periodicity of the asperities, we find an effective boundary condition of the Navier type on the rough wall and the corrector for the Darcy-Weisbach friction coefficient. The results are justified by an error estimate. Asymptotic expansions, homogenization and boundary layer techniques are used in the analysis.

Key words: pipe flow, rough boundary, Navier law, Darcy-Weisbach law
MSC 2000: 35B27, 35B25, 76D05, 76M50, 76M45

1 Introduction

Fluid flows in pipes are of interest in applications. They can be found in natural circumstances, for instance in circulation systems of humans, animals and plants, but also in water supplies, irrigation and sewer water systems, etc. In engineering applications the pressure drop caused by the friction of the viscous fluid against the pipe's wall is computed from the Darcy-Weisbach law, saying that it is proportional with the square of the velocity. In case of smooth pipe and laminar flow the friction coefficient can be computed as $f_D = 32/Re$. Frequently for pipe flows we need to take into account the effects of rough boundaries. Pipe's walls are never perfectly smooth and some boundary roughness appears either due to its fabrication or because of the demage or the sedimentation of material during it's use. Flows in rough domains in general represent a challenge for numerical simulations. The subject of this article is the flow in a pipe that is corrugated (i.e. periodically wrinkled in the longitudinal direction) . We use the Stokes model of the incompressible fluid flow ([1]).

 ISBN: 978-84-617-8694-7

Important feature is the geometry of the pipe, in particular it's roughness. In applications, the pipe roughness amplitude could vary from 0.0015 mm for glass, drawn brass and copper pipe material, 0.26 mm for cast iron, 0.18-0.6 mm for concrete, 0.12 mm for PVC, and 45 mm for corrugated metal.
We derive here the corrector to the usual friction coefficient, due to the rugosities of the pipe. For derivation of the approximation and of the effective boundary condition we use asymptotic analysis, homogenization and boundary layer techniques.

2 Geometry and equations

Let $F : \mathbf{R} \to \mathbf{R}$ be a smooth periodic function, with period 1. We define the fluid flow domain and it's wall as

$$\Omega_\varepsilon = \{(r,\varphi,z) \in \mathbf{R}^3 \; ; \; 0 \le \varphi < 2\pi \; , \; 0 < z < 1 \; , \; 0 \le r < 1 + \varepsilon \; F(z/\varepsilon) \; \}$$
$$\Gamma_\varepsilon = \{(r,\varphi,z) \in \mathbf{R}^3 \; ; \; 0 \le \varphi < 2\pi \; , \; 0 < z < 1 \; , \; r = 1 + \varepsilon \; F(z/\varepsilon) \; \} \; .$$

A stationary flow of an incompressible, viscous, Newtonian fluid with, small Reynolds number, can be fairly described by the Stokes system. Our system can thus be written as

$$-\mu \Delta u^\varepsilon + \nabla p^\varepsilon = 0 \;\; , \;\; \operatorname{div} u^\varepsilon = 0 \;\text{ in }\; \Omega_\varepsilon \;\; , u^\varepsilon = 0 \;\text{ on } \Gamma_\varepsilon \; , \;\; u^\varepsilon \times \mathbf{n} = 0 \;\; , \;\; p^\varepsilon = p_z \;\text{ for }\; z = 0, 1 \; .$$

Here u^ε denotes velocity, p^ε pressure and $\mathbf{n}$ stands for the outward normal.

3 Formal computation

After introducing the change of the variables $\rho = r - \varepsilon \; F\left(\frac{z}{\varepsilon}\right)$ we define $\mathcal{U}^\varepsilon(\rho,\varphi,z) = u^\varepsilon(r,\varphi,z)$. Assuming that the pressure is linear$p^\varepsilon = p_0 + z \; \delta P$, $\delta P = p_1 - p_0$ and that the velocity has only the component along the pipe, i.e. $u^\varepsilon = \mathcal{U}^\varepsilon_z \mathbf{e}_z$, we arrive at

$$\mu \left\{ \frac{1}{\rho + \varepsilon \, F} \frac{\partial}{\partial \rho} \left((\rho + \varepsilon F) \frac{\partial \mathcal{U}^\varepsilon_z}{\partial \rho} \right) + \frac{\partial^2 \mathcal{U}^\varepsilon_z}{\partial z^2} - 2F' \frac{\partial^2 \mathcal{U}^\varepsilon_z}{\partial \rho \partial z} + (F')^2 \frac{\partial^2 \mathcal{U}^\varepsilon_z}{\partial \rho^2} - \frac{F''}{\varepsilon} \frac{\partial \mathcal{U}^\varepsilon_z}{\partial \rho} \right\} = \delta P$$
$$\frac{\partial \mathcal{U}^\varepsilon_z}{\partial z} - F' \; \frac{\partial \mathcal{U}^\varepsilon_z}{\partial \rho} = 0 \; . \tag{1}$$

We look for an asymptotic expansion of the form $\mathcal{U}^\varepsilon_z = U_0(\rho, z, z/\varepsilon) + \varepsilon \; U_1(\rho, z, z/\varepsilon) + \cdots$. We get by direct computation that $U_0(\rho) = \frac{1}{4\mu}(\rho^2 - 1) \; \delta P$ and $U_1 = F(\xi) \frac{\partial U_0}{\partial \rho} + A$. The constant A is to be determined from the appropriate boundary condition, which will be discussed in the next section.

 ISBN: 978-84-617-8694-7

4 Boundary layer

In this section we construct an additional corrector, using the boundary layers techniques, to fix the boundary condition on the wall of the pipe. The approximation

$$\overline{\mathcal{U}}^{\varepsilon}(\rho,\xi) = (U_0(\rho) + \varepsilon\, F(\xi)\frac{\partial U_0}{\partial \rho}(\rho) + \varepsilon\, A)\mathbf{e}_z$$

cannot satisfy the boundary condition $\overline{\mathcal{U}}^{\varepsilon}(1,\xi) = 0$ due to the term $\varepsilon\, F(\xi)\frac{\partial U_0}{\partial \rho}$ and we need to add a boundary layer corrector. Thus we postulate an approximation of the form

$$\mathcal{V}^{\varepsilon} = U_0(\rho)\mathbf{e_z} + \varepsilon\, F(\xi)\frac{\partial U_0}{\partial \rho}(\rho)\mathbf{e_z} + \varepsilon\, B\, \frac{\partial U_0}{\partial \rho}(1) + \varepsilon\, A\mathbf{e_z}\ , \tag{2}$$

where $B = B\,(\tau,\xi) = B_\tau(\tau,\xi)\mathbf{e}_r + B_\xi(\tau,\xi)\mathbf{e}_z$, with $\xi = \frac{z}{\varepsilon}$, $\tau = \frac{1-r}{\varepsilon} + F(\xi)$. We also need to add an additional term to the pressure approximation

$$p^{\varepsilon} = p_0 + z\ \delta P + b(\tau,\xi)\ \frac{\partial U_0}{\partial \rho}(1)\ \ .$$

The couple (B,b) satisfies the following boundary layer problem

$$-\mu\left((1+(F')^2)\frac{\partial^2 B_\tau}{\partial \tau^2} + \frac{\partial^2 B_\tau}{\partial \xi^2} + F'\frac{\partial^2 B_\tau}{\partial \xi \partial \tau} + \frac{\partial}{\partial \xi}\left(F'\frac{\partial B_\tau}{\partial \tau}\right)\right) - \frac{\partial b}{\partial \tau} = 0\ , \tag{3}$$

$$-\mu\left((1+(F')^2)\frac{\partial^2 B_\xi}{\partial \tau^2} + \frac{\partial^2 B_\xi}{\partial \xi^2} + F'\frac{\partial^2 B_\xi}{\partial \xi \partial \tau} + \frac{\partial}{\partial \xi}\left(F'\frac{\partial B_\xi}{\partial \tau}\right)\right) + \frac{\partial b}{\partial \xi} + F'\frac{\partial b}{\partial \tau} = 0\ ,$$

$$-\frac{\partial B_\tau}{\partial \tau} + \frac{\partial B_\xi}{\partial \xi} + F'\frac{\partial B_\xi}{\partial \tau} = 0 \quad \text{for} \quad \tau > 0\ \ ,\ \ 0 < \xi < 1\ , \tag{4}$$

$$(B_\tau(0,\xi), B_\xi(0,\xi)) + (0, F(\xi)) = K = (K_\tau, K_\xi) \tag{5}$$

$$\nabla B(\tau,\xi) \to 0\ \ ,\ b(\tau,\xi) \to 0 \text{ for } \tau \to +\infty\ ,\ \ (B,b) \quad \text{is} \quad 1-\text{periodic} \quad \text{in } \xi\ \ . \tag{6}$$

The solution of the above boundary layer problem decays exponentially as $\tau \to +\infty$ to some constant. As b is determined up to a constant, it can be chosen in a way that it decays to 0. As for the velocity B, due to the linearity of the equation, we can choose K in a way that it decays to zero too. Indeed, denoting $\omega = \{(\tau,\xi) \in \mathbf{R}^2\ \ ,\ \tau > 0\ \ ,\ \ 0 < \xi < 1\ \}$, we have:

Theorem 1 *The problem (3)-(6) has a unique solution $B \in V = \{w \in L^2_{loc}(\omega)^2\ \ ,\ \nabla w \in L^2(\omega)^4\}$, $b \in L^2(\omega)/\mathbf{R}$. Furthermore, the constant vector K can be chosen such that*

$$|B(\tau,\xi)| \le Ce^{-\beta\tau} \quad , \quad |b(\tau,\xi)| \le Ce^{-\beta\tau}\ , \tag{7}$$

for some $C > 0$ and $\beta > 0$. Finally $K_\tau = 0$.

The approximation with the boundary layer corrector can be justified by an appropriate error estimate (see [2])

 ISBN: 978-84-617-8694-7

5 Darcy-Weisbach friction coefficient

In engineering applications the Darcy-Weisbach formula is used to compute the pressure drop in the pipe, due to the friction against the wall. It s usually written in the form

$$\delta P = f_D \ \rho \frac{\langle \mathcal{V}_z^\varepsilon \rangle^2}{4} \ ,$$

where δP is the pressure drop, ρ is the fluid density, $\langle \mathcal{V}_z^\varepsilon \rangle$ is the mean velocity, and f_D is the friction coefficient. For smooth pipe, the Poiseuille formula gives $f_D = \frac{32}{Re}$, which is the well-known expression for the friction coefficient in case of laminar flow. Here the Reynolds number is defined as $Re = \rho V/\mu$. In case of rough pipe, a corrector due to the roughness is needed. Since, up to the order ε, we have

$$\langle \mathcal{V}_z^\varepsilon \rangle = \frac{1}{|\Omega_\varepsilon|} \int_{\Omega_\varepsilon} \mathcal{V}_z^\varepsilon = \frac{\delta P}{\mu} \left[-\frac{1}{8} + \varepsilon \left(\frac{7}{12} \langle F \rangle - \frac{1}{2} K_\xi \right) \right] \ ,$$

where $\langle F \rangle = \int_0^1 F(t) \ dt$. Taking the Poiseuille velocity $V = \frac{\delta P}{8\mu}$ as the characteristic velocity we get the Reynolds number $Re = \rho \ \frac{V}{\mu}$ and then

$$\langle \mathcal{V}_z^\varepsilon \rangle = -V \ \left[1 + \varepsilon \ \left(\frac{14}{3} \ \langle F \rangle - 4 \, K_\xi \right) \right] \ .$$

Since $f_D = \frac{4 \, \delta P}{\rho \, \langle \mathcal{V}_z^\varepsilon \rangle^2}$ neglecting the terms of order ε^2 and smaller leads to

$$f_D = \frac{32}{Re} \ \left[1 \ - 4 \, \varepsilon \ \left(\frac{7}{3} \, \langle F \rangle - 2 \, K_\xi \right) \right] \ .$$

Acknowledgements

This work was supported by the grant No 3955 *Mathematical modeling and numerical simulations of processes in thin or porous domains* by Croatian science foundation.

References

[1] Lauga E., Ston H. A.,*Effective slip in pressure-driven Stokes flow*, J. Fluid Mech., 489 (2003), 55-77.

[2] Marušić-Paloka E., Starčević M., *Effects of rough wall on incompressible flow through a pipe*, submitted.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

Solving large scale quasiseparable Lyapunov equations

Stefano Massei[1], Davide Palitta[2] and Leonardo Robol[3]

[1] *Applied Mathematics, EPF Lausanne, Switzerland*

[2] *Dipartimento di Matematica, Università di Bologna, Italy*

[3] *ISTI, Consiglio Nazionale delle Ricerche, Pisa, Italy*

emails: stefano.massei@epfl.ch, davide.palitta3@unibo.it,
leonardo.robol@isti.cnr.it

Abstract

We consider the problem of efficiently solving Lyapunov and Sylvester equations of medium and large scale, in the case where all the coefficients are quasiseparable, i.e., they have off-diagonal blocks of low-rank. This comprises the case with banded coefficients and right-hand side, recently studied in [6, 9].

We show that, under suitable assumptions, this structure is guaranteed to be numerically present in the solution, and we provide explicit estimates of the numerical rank of the off-diagonal blocks. Moreover, we describe an efficient method for approximating the solution, which relies on the technology of hierarchical matrices.

A theoretical characterization of the quasiseparable structure in the solution is presented, and numerically experiments confirm the applicability and efficiency of our approach. We provide a MATLAB toolbox that allows easy replication of the experiments and a ready-to-use interface for our solver.

Key words: Lyapunov equation, Sylvester equation, Quasiseparable structure

1 Introduction

Lyapunov and Sylvester equations, i.e., linear matrix equations of the form

$$AX + XA^T = C, \qquad AX - XB = C,$$

respectively, appear in many different areas of mathematics, and are a recurrent problem in numerical analysis. In many practical cases, such as the problems arising from control theory

 ISBN: 978-84-617-8694-7

[1, 3], the right-handside (RHS) of the above equations is often low-rank, or numerically low-rank, i.e., $C \approx uv^T$, with u, v tall and thin matrices.

In this case it is well-known that, if the coefficient matrices A and B satisfy certain spectral properties, the solution X can be well-approximated by a low rank matrix [2]. This has enabled the study of projection methods, based on several variants of Krylov subspaces (such as classical, extended, and rational Krylov spaces).

However, recently there has been a growing interest in considering Sylvester and Lyapunov equations arising from discretization of $2D$ PDEs [9]. On of the easiest and most classical examples can be obtained by the discretization of the Laplacian operator, but this approach can handle heat and convection-diffusion PDEs as well. In general, one can consider problems arising from equations of the form $Lu = f$, where L is a linear differential operator and u is defined on a rectangular domain.

In this framework the matrix C is nothing else that the discretization of the RHS of the PDE on the chosen grid. When the function is well-approximated by a separable function, i.e., $f(x, y) \approx f_1(x)f_2(y)$, or a linear combination of functions of this form, then the matrix C has numerically low rank, and Krylov methods can be applied. However, this is not always the case, and often the matrix C can have a more complicated structure.

We consider the situation where the RHS C is quasiseparable. Many interesting cases fall under this framework, such as a diagonal or diagonal plus low rank matrix, as well as banded matrices. These matrices can be full-rank in general, but they have a low rank structure in their off-diagonal blocks. We investigate the preservation of this structure in the solution X, and we propose a new method for the approximation of this structured solution. For this approach to work, also the matrices A and B need to have the same structure. This is not restrictive, since they are often banded, which is a particular case of quasiseparable matrices.

Some previous attempts to exploit a similar structure (relying on the band of these matrices), can be found in [6]. We show that our approach yields faster computation times and better accuracy.

2 An approximation result

A key result in our analysis is that, under certain spectral conditions on the matrices A and B similar to the ones needed for the results in the classical low-rank case, one can prove that the solution X has a numerical quasiseparable structure, i.e., its off-diagonal blocks have numerically low rank.

We provide numerical evidence of this fact as well as propose an argument to prove that the structure must be present in the solution, relying on arguments from approximation theory inspired by the theory of Krylov and rational Krylov subspaces. These result provide an extension of the approximation results in [4, 5] that prove the preservation of

 ISBN: 978-84-617-8694-7

the quasiseparable structure in the solution of quadratic matrix equations, which can also be described by means of integral formulas.

This opens the door to the treatment of large scale problems, where the size of X would make its storage unfeasible otherwise. In fact, to store a quasiseparable matrix of order n with a quasiseparable rank of k only $\mathcal{O}(nk)$ or $\mathcal{O}(nk \log n)$ storage is needed, depending on the chosen representation. It is clear that if $k \ll n$ this enables the study of large scale problems.

As a motivating example, consider the partial differential equation

$$\begin{cases} \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = \log\left(\epsilon + |x - y|\right), & (x, y) \in \Omega, \\ u(x, y) \equiv 0 & (x, y) \in \partial\Omega, \end{cases}$$

where Ω is some rectangular domain of the form $[a, b] \times [a, b]$ and ϵ is small. In this case the RHS $f(x, y)$ is large when $x \approx y$, and can be well approximated by a separable function elsewhere. In fact, the matrix C obtained sampling this function has full-rank, but a low quasiseparable order.

The coefficient matrix A obtained discretizing the Laplacian is tridiagonal, and every banded matrix is quasiseparable. Therefore, this problem is covered by our framework.

3 Approximating the solution

In order to approximate the solution X, which we have proved to be quasiseparable, we need to develop a method that exploits the quasiseparable structure of A, B, and C.

We propose to consider the integral representation of the solution X, that is

$$X = \int_0^\infty e^{-tA} C e^{-tA^T} dt,$$

and we discuss numerical strategies to approximate the above integral formula effectively using the available structure. This problem is linked with many different topics, such as the efficient approximation of matrix functions whose argument is quasiseparable [8], and the devise of efficient integration schemes for quasiseparable matrix valued functions.

We draw a connection with previous work of Hackbush et al. [7], and we discuss several improvements to the available tools in the literature, and the state of the art procedures.

Numerical experiments confirm that our approach is faster than all the other available (such as the one in [6]), and has a much higher accuracy. Moreover, our scheme does not suffer when the conditioning of the problem increases.

The algorithm presented has been implemented in a MATLAB toolbox, called `hm-toolobx`, that can be used to easily manipulate Hierarchical matrices. We provide some examples of its use, and on the solution of Lyapunov equations.

 ISBN: 978-84-617-8694-7

Acknowledgements

This work has been partially supported byby the Region of Tuscany (Project "MOSCARDO - ICT technologies for structural monitoring of age-old constructions based on wireless sensor networks and drones", 2016- 2018, FAR FAS), and by the GNCS/INdAM project "Metodi numerici avanzati per equazioni e funzioni di matrici con struttura".

References

[1] A. C. Antoulas. *Approximation of large-scale dynamical systems*, volume 6 of *Advances in Design and Control*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2005.

[2] B. Beckermann. An error analysis for rational galerkin projection applied to the sylvester equation. *SIAM Journal on Numerical Analysis*, 49(6):2430–2450, 2011.

[3] P. Benner, J.-R. Li, and T. Penzl. Numerical solution of large-scale lyapunov equations, riccati equations, and linear-quadratic optimal control problems. *Numerical Linear Algebra with Applications*, 15(9):755–777, 2008.

[4] D. A. Bini, S. Massei, and L. Robol. Efficient cyclic reduction for Quasi-BirthDeath problems with rank structured blocks. *Appl. Numer. Math.*, 2016.

[5] D. A. Bini, S. Massei, and L. Robol. On the decay of the off-diagonal singular values in cyclic reduction. *Linear Algebra Appl.*, 519:27–53, 2017.

[6] A. Haber and M. Verhaegen. Sparse solution of the Lyapunov equation for large-scale interconnected systems. *Automatica J. IFAC*, 73:256–268, 2016.

[7] W. Hackbusch. *Hierarchical matrices: algorithms and analysis*, volume 49 of *Springer Series in Computational Mathematics*. Springer, Heidelberg, 2015.

[8] S. Massei and L. Robol. Decay bounds for the numerical quasiseparable preservation in matrix functions. *Linear Algebra Appl.*, 516:212–242, 2017.

[9] D. Palitta and V. Simoncini. Matrix-equation-based strategies for convection-diffusion equations. *BIT*, 56(2):751–776, 2016.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

Numerical quasiseparable preservation in matrix functions

Stefano Massei[1] **and Leonardo Robol**[2]

[1] *Department of applied mathematics, Ecole polytechnique federale de Lausanne*

[2] *Istituto di Scienza e Tecnologie dellInformazione, CNR, Pisa*

emails: stefano.massei@epfl.ch, leonardo.robol@isti.cnr.it

Abstract

Given matrices A and B such that $B = f(A)$, where $f(z)$ is a holomorphic function, we analyze the relation between the singular values of the off-diagonal submatrices of A and B. We provide a family of bounds which depend on the interplay between the spectrum of the argument A and the singularities of the function. In particular, these bounds guarantee the numerical preservation of quasiseparable structures under mild hypotheses. We extend the Dunford-Cauchy integral formula to the case in which some poles are contained inside the contour of integration. We use this tool together with the technology of hierarchical matrices ($\mathcal{H}$-matrices) for the effective computation of matrix functions with quasiseparable arguments.

Key words: Matrix functions, Quasiseparable matrices, Hierarchical matrices

1 Introduction

Matrix functions are an evergreen topic in matrix algebra due to their wide use in applications. It is not hard to imagine why the interaction of structures with matrix functions is an intriguing subject [11]. In fact, in many cases structured matrices arise and can be exploited for speeding up algorithms, reducing storage costs or allowing to execute otherwise not feasible computations [4, 5]. The property we are interested in is the *quasi-separability*. That is, we want to understand whether the submatrices of $f(A)$ contained in the strict upper triangular part or in the strict lower triangular part, called *off-diagonal submatrices*, have a "small" numerical rank.

Studies concerning the numerical preservation of data-sparsity patterns were carried out recently [1, 2, 3, 7]. Regarding the quasiseparable structure, in [8, 9, 10] Gavrilyuk,

 ISBN: 978-84-617-8694-7

Hackbusch and Khoromskij addressed the issue of approximating some matrix functions using the hierarchical format [6]. In these works the authors prove that, given a low rank quasiseparable matrix A and a holomorphic function $f(z)$, computing $f(A)$ via a quadrature formula applied to the contour integral definition, yields an approximation of the result with a low quasiseparable rank. Representing A with a $\mathcal{H}$-matrix and exploiting the structure in the arithmetic operations provides an algorithm with almost linear complexity. The feasibility of this approach is equivalent to the existence of a rational function $r(z) = \frac{p(z)}{q(z)}$ which well-approximates the holomorphic function $f(z)$ on the spectrum of the argument A. More precisely, since the quasiseparable rank is invariant under inversion and sub-additive with respect to matrix addition and multiplication, if $r(z)$ is a good approximation of $f(z)$ of low degree then the matrix $r(A)$ is an accurate approximation of $f(A)$ with low quasiseparable rank. This argument explains the preservation of the quasiseparable structure, but still needs a deeper analysis which involves the specific properties of the function $f(z)$ in order to provide effective bounds to the quasiseparable rank of the matrix $f(A)$.

In this work we deal with the analysis of the quasiseparable structure of matrix functions by studying the interplay between the off-diagonal singular values of the matrices A and B such that $B = f(A)$. Our intent is to understand which parameters of the model come into play in the numerical preservation of the structure and to extend the analysis to functions with singularities.

2 Theoretical analysis

Because of the sub-additivity with respect to the matrix sum and to the matrix product of the quasiseparable rank, it is straightforward that a polynomial of degree d evaluated in an argument A of quasiseparable rank k provides a matrix of quasiseparable rank at most $d \cdot k$. Using this observation it is possible to provide estimates on the numerical quasiseparable rank of $f(A)$ based on the polynomial approximation of the function f. More precisely, if f is well approximated by a polynomial of low degree on the spectrum of A then $f(A)$ has off-diagonal blocks with fast decay in their singular values. So it is possible to obtain bounds on the off-diagonal singular values exploiting classical results of polynomial approximation in the complex plane like Bernstein theorem. In the case in which f has singularities close to the spectrum of A these bounds do not provide useful information. We introduce new technical tools that enable us to overcome the problem, getting meaningful bounds.

3 Exploiting the structure

In order to take advantage of the quasiseparable structure we need a representation that enable us to perform the storage and the matrix operations cheaply. We rely on the frame-

 ISBN: 978-84-617-8694-7

Operation	Computational complexity
Matrix-vector multiplication	$O(km\log(m))$
Matrix-matrix addition	$O(k^2m\log(m))$
Matrix-matrix multiplication	$O(k^2m\log(m)^2)$
Matrix-inversion	$O(k^2m\log(m)^2)$
Solve linear system	$O(k^2m\log(m)^2)$

work of Hierarchical representations originally introduced by Hackbusch [10] in the context of integral and partial differential equations. It consists in a class of recursive block representations with structured sub-matrices that allows the treatment of a number of data-sparse patterns. Here, we consider a particular member of this family — sometimes called Hierarchical off-diagonal low-rank representation (HODLR) — which has a simple formulation and an effective impact in handling quasiseparable matrices.

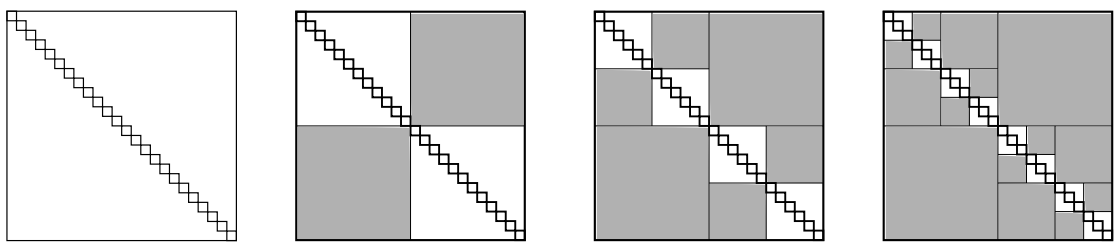


Figure 1: The behavior of the block partitioning in the HODLR-matrix representation. The blocks filled with grey are low rank matrices represented in a compressed form, and the diagonal blocks in the last step are stored as dense matrices.

The HODLR representation acts on a matrix by compressing many of its sub-blocks. Therefore, it is natural to perform the arithmetic operations in a block-recursive fashion. The basic steps of these procedures require arithmetic operations between low-rank matrices or small size matrices. If the rank of the off-diagonal blocks is small compared to m, then the algorithms performing the arithmetic operations have linear polylogarithmic complexities [6][Chapter 6]. The latter are summarized in the table where it is assumed that the constant k bounds the quasiseparable rank of all the matrices involved. Moreover, the operations are performed adaptively with respect to the rank of the blocks. This means that the result of an arithmetic operation will be an HODLR matrix with the same partitioning, where each low rank block is a truncated reduced SVD of the corresponding block of the exact result. This operation can be carried out with linear cost, assuming that the quasiseparable rank stays negligible with respect to m.

 ISBN: 978-84-617-8694-7

References

[1] M. Benzi and P. Boito. Decay properties for functions of matrices over C^*-algebras. *Linear Algebra Appl.*, 456:174–198, 2014.

[2] M. Benzi, P. Boito, and N. Razouk. Decay properties of spectral projectors with applications to electronic structure. *SIAM Rev.*, 55(1):3–64, 2013.

[3] M. Benzi and V. Simoncini. Decay bounds for functions of Hermitian matrices with banded or Kronecker structure. *SIAM J. Matrix Anal. Appl.*, 36(3):1263–1282, 2015.

[4] D. A. Bini, S. Massei, and L. Robol. Efficient cyclic reduction for Quasi-Birth-Death problems with rank structured blocks. *Appl. Numer. Math.*, 2016.

[5] D. A. Bini, S. Massei, and L. Robol. On the decay of the off-diagonal singular values in cyclic reduction. arXiv preprint arXiv:1608.01567, 2016.

[6] S. Börm, L. Grasedyck, and W. Hackbusch. Hierarchical matrices. *Lecture notes*, 21:2003, 2003.

[7] C. Canuto, V. Simoncini, and M. Verani. On the decay of the inverse of matrices that are sum of Kronecker products. *Linear Algebra Appl.*, 452:21–39, 2014.

[8] I. P. Gavrilyuk, W. Hackbusch, and B. N. Khoromskij. $\mathcal{H}$-matrix approximation for the operator exponential with applications. *Numer. Math.*, 92(1):83–111, 2002.

[9] I. P. Gavrilyuk, W. Hackbusch, and B. N. Khoromskij. Data-sparse approximation to a class of operator-valued functions. *Math. Comp.*, 74(250):681–708, 2005.

[10] W. Hackbusch. *Hierarchical matrices: algorithms and analysis*, volume 49 of *Springer Series in Computational Mathematics*. Springer, Heidelberg, 2015.

[11] S. Massei and L. Robol. Decay bounds for the numerical quasiseparable preservation in matrix functions. *Linear Algebra and its Applications*, 2016.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

Kantorovich method to solve an integral equation arising from a problem in mathematical biology

Abdelaziz Mennouni[1]

[1] *Department of Mathematics, LTM, University of Batna 2, Algeria*

emails: aziz.mennouni@yahoo.fr

Abstract

Kantorovich method is formulated and justified for solving numerically an integral equation arising from a problem in mathematical biology, using a sequence of orthogonal finite rank projections. The convergence analysis and associated theorems are considered in this work. Numerical examples illustrate the theoretical results and show the effectiveness of the method.

Key words: Kantorovich method, integral equations, orthogonal finite rank projections.
MSC 2000: 45A05, 45B05

1 Introduction

Let us consider the following integral equation of the second kind

$$x(s)\int_0^1 k(s-t)dt = \int_0^1 x(\tau)k(\tau-s)d\tau + f(s), \quad 0 \le s \le 1, \tag{1}$$

where $k(\cdot,\cdot)$ is a Fredholm kernel, and where f is a known function.

Equation (1) reads as

$$x(s) - \int_0^1 \frac{x(\tau)k(\tau-s)d\tau}{\int_0^1 k(s-t)dt} = f(s), \quad 0 \le s \le 1. \tag{2}$$

Define the integral operator T:

$$Tx(s) := \int_0^1 \frac{x(\tau)k(\tau-s)d\tau}{\int_0^1 k(s-t)dt}, \quad 0 \le s \le 1.$$

 ISBN: 978-84-617-8694-7

Set $\mathcal{H} := L^2[0,1]$. Suppose $k \in \mathcal{L}^1[0,1]$, $k > 0$ almost everywhere. We recall that for each $f \in \mathcal{H}$, T is compact from $\mathcal{H}$ into itself, (see [6]). Hence, the integral equation (1) has a unique solution $x \in \mathcal{H}$.

Let I denote the the identity operator on $\mathcal{H}$. Eq. (2) can be rewritten in operator form as follows:

$$(I - T)\, x = f.$$

The purpose of this work is to approximate x through the exact solution x_n of the Kantorovich equation

$$(I - \pi_n T)\, x_n = f. \tag{3}$$

2 Projection approximations using general grids

Let $(s_{n,j})_{j=0}^{n}$ be a grid on $[0,1]$ such that

$$0 < s_{n,0} < s_{n,1} < \ldots < s_{n,n} < 1.$$

Set

$$h_{n,i} := s_{n,i} - s_{n,i-1}, \quad i \in [\![1, n]\!], \quad h_n := (h_{n,1}, h_{n,2}, \ldots, h_{n,n}).$$

Let us consider $(\pi_n)_{n\geq 1}$, a sequence of bounded projections each one of finite rank, such that

$$\pi_n x := \sum_{j=1}^{n} \langle x, e_{n,j} \rangle\, e_{n,j},$$

where

$$e_{n,j} := \frac{\phi_{n,j}}{\sqrt{h_{n,j}}}, \quad \phi_{n,j}(s) := \begin{cases} 1 \text{ for } s \in]s_{n,j-1}, s_{n,j}[, \\ 0 \text{ otherwise.} \end{cases}$$

Let

$$J_n := \{s_{n,j}, \quad j \in [\![0, n]\!]\}.$$

Define the modulus of continuity of the function $\psi \in H$ relative to h_n as follows:

$$\omega_2(\psi, J_n) := \sup_{0\leq\delta\leq h_n} \left(\int_0^1 |\psi(\tau + \delta) - \psi(\tau)|^2\, d\tau \right)^{\frac{1}{2}}.$$

All functions are extended by 0 outside $[0,1]$. We recall that

$$\omega_2(\psi, J_n) \to 0 \text{ as } n \to \infty \text{ for all } \psi \in H,$$

and that, for all $\psi \in H$ (cf. [4]),

$$\|(I - \pi_n)\psi\| \leq \omega_2(\psi, J_n). \tag{4}$$

 ISBN: 978-84-617-8694-7

3 Kantorovich method

We have

$$\pi_n Tx := \sum_{j=1}^{n} \langle Tx, e_{n,j}\rangle \, e_{n,j}.$$

Applying T to both sides of equation (3), and performing the inner product with $e_{n,i}$ to both sides of this equation, we get

$$\langle Tx_n, e_{n,i}\rangle - \sum_{j=0}^{n-1} \langle Tx_n, e_{n,j}\rangle \langle Te_{n,j}, e_{n,i}\rangle = \langle Tf, e_{n,i}\rangle ,$$

or, equivalently,

$$(I_n - A_n)X_n = b_n, \tag{5}$$

where

$$X_n(j) := \langle Tx_n, e_{n,i}\rangle ,$$

and

$$\begin{aligned} A_n(i,j) &:= \langle Te_{n,j}, e_{n,i}\rangle , \\ b_n(i) &:= \langle Tf, e_{n,i}\rangle . \end{aligned}$$

Hence

$$\begin{aligned} A_n(i,j) &:= \frac{1}{\sqrt{h_{n,j}h_{n,i}}} \int_{s_{j-1}}^{s_j} \int_{s_{i-1}}^{s_i} \frac{k(\tau - s)}{\int_0^1 k(s-t)dt} d\tau ds, \\ b_n(i) &:= \frac{1}{\sqrt{h_{n,i}}} \int_{s_{i-1}}^{s_i} \int_0^1 \frac{f(\tau)k(\tau - s)}{\int_0^1 k(s-t)dt} d\tau ds. \end{aligned}$$

4 Convergence analysis

For all $x \in \mathcal{H}$,

$$\lim_{n\to\infty} \|\pi_n Tx - Tx\| = 0,$$

and since T is compact,

$$\lim_{n\to\infty} \| (\pi_n T - T) T\| = 0, \quad \lim_{n\to\infty} \| (\pi_n T - T) \pi_n T\| = 0.$$

 ISBN: 978-84-617-8694-7

Theorem 1 *There exists a positive constant M, such that*

$$\|x_n - x\| \leq M \left[\omega_2(x, J_n) + \omega_2(f, J_n)\right].$$

Proof. In fact

$$\pi_n x = \pi_n T x + \pi_n f.$$

Since

$$\begin{aligned} x - \pi_n x &= x - x_n + x_n - \pi_n x \\ &= x - x_n + (\pi_n T x_n + f) - (\pi_n T x + \pi_n f) \\ &= x - x_n + \pi_n T (x_n - x) + (I - \pi_n) f \\ &= (I - \pi_n T)(x - x_n) + (I - \pi_n) f. \end{aligned}$$

Hence

$$x - x_n = (I - \pi_n T)^{-1} \left[(I - \pi_n) x - (I - \pi_n) f\right],$$

and since T is compact, the

$$M := \sup_{n \geq N} \|I - \pi_n T\|,$$

is finite. Using (4), we get the desired result. ■

References

[1] M. Ahues, A. Largillier, B. V. Limaye, *Spectral Computations for Bounded Operators*, CRC, Boca Raton, 2001.

[2] K.E. Atkinson, Han, *Theoretical Numerical Analysis: A Functional Analysis Framework*, 3rd edition Springer-Verlag, 2009.

[3] M. Ahues, F.D. d'Almeida, R.R. Fernandes, *Piecewise Constant Galerkin Approximations of Weakly Singular Integral Equations*, International Journal of Pure and Applied Mathematics, **55** (2009), 569–580.

[4] A. Amosov, M. Ahues, *Superconvergence of Some Projection Approximations for Weakly Singular Integral Equations Using General Grids*, SIAM Journal on Numerical Analysis, **47** (2008), 646–674.

[5] A. Mennouni, *Two Projection Methods for Skew-Hermitian Operator Equations*, Mathematical and Computer Modelling, 55 (2012) pp. 1649-1654.

[6] S. Eveson, *An integral equation arising from a problem in mathematical biology*, Bull Lon Math Soc. **23** (1991) 293–299.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

Numerical Solution of a Cancer Invasion Model Using DRBEM and FDM

Gülnihal Meral[1]

[1] , *Department of Mathematics and Computer Science, Ankara Yıldırım Beyazıt University, 06010, Ankara, Turkey*

emails: gmeral@ybu.edu.tr

Abstract

In this study, the haptotaxis-only model for cancer cell invasion of tissue [1] is solved by using the combination of Dual Reciprocity Boundary Element Method (DRBEM) and the Finite Difference Method(FDM). The model consists of a system of reaction-diffusion-taxis equations, describing the interactions between the cancer cells, the normal cells and the matrix degrading enzyme (MDE) which is produced by the cancer cells.

The space derivatives coming from the Laplacian terms in the equations for MDE concentration and cancer cell density are discretized by DRBEM using the fundamental solution of Laplace equation and considering the rest of the terms as nonhomogenity. The nonhomogenity for the equation describing the change in the cancer cell density involves the first and second order space derivatives of the unknowns and their products and in order to handle this diffculty the finite difference method is made use of. The resulting time dependent ordinary differential equations(ODEs) obtained after the spatial discretization of the equations for the cancer cell density and for the concentration of the MDE and the ODE for the normal cell density are then solved by using a combination of the forward and backward Euler methods.

Because of the nonlinear terms in the model, the numerical solution in two-space dimensions has some difficulties and with the proposed method these diffculties are overcome mainly by using DRBEM in space discretization which only needs to discretize the boundary only and thus the method gives the solution with a lower computational cost. The haptotactic behaviour and the effect of the reestablishment of the cells are analysed by the numerical simulations and the results agree well with the expected behaviour of the invasion.

Key words: DRBEM, FDM, Mathematical Modelling, Cancer Invasion

 ISBN: 978-84-617-8694-7

1 Introduction

The crucial point of cancer is the metastasis, which is defined as the formation of a secondary tumour at a distant side of the body. Metastasis forms following the invasion of the tissue where the tumour arises. This is followed by the penetration of the cancer cells into adjacent tissues. Then the tumour cells migrate via blood or lymph system and form a secondary tumour, i.e. the metastasis. Thus the turning point is the invasion step. Chemotaxis and haptotaxis are two mechanisms effecting the movement of cancer cells through tissue network. Chemotaxis is the movement of the cancer cells in response to a chemoattractant in the solution [1]. On the other hand the cells have to adhere to the extracellular matrix (ECM) fibres in order to move and thus they start to migrate from a region of low concentration of the present adhesive molecule on ECM to an area with higher concentration, which is called haptotaxis. The contact with the tissue stimulates the production of the proteolytic enzyme called matrix degrading enzyme (MDE) which degrades the tissue fibres and this let the cancer cells to migrate towards neighbouring blood vessels.

In the last decades, several models are proposed on invasion considering different factors, such as the pH level of the environment [2, 3], role of heat shock proteins [4, 5] or matrix-degrading enzymes [1].

Generally, the analytical solutions for the cancer invasion models are not available and thus it is important to find accurate and efficient numerical solutions to these models. However, because of the nonlinearities arise in the most of the models it is not easy to solve these models numerically, either. The models are usually solved by using finite difference method [3, 6, 7, 8]. However, when one considers the two-dimensional regions more discretization points are needed and the numerical solution becomes computationally expensive and stability problems may occur. To handle these problems a combined application of DRBEM and FDM is applied recently for the acid mediated invasion model [9].

In this study, the combined application of DRBEM and FDM is applied to the haptotaxis-only model [1]. The method is applied by using the fundamental solution of Laplace equation considering the time derivative, the nonlinearities coming from the haptotaxis and proliferation terms; and the production, decay terms for the equation describing the temporal and spatial evolution of cancer cell density; and MDE concentration, respectively. The haptotaxis term includes the first and second order space derivatives of the unknowns and their products. For the discretization of these derivatives in the nonhomogenity the forward and central differences are made use of. For the time discretization of the ODEs obtained after the space discretization a combination of forward and backward Euler methods are used. The proposed method is tested to see the effect of haptotaxis and proliferation of the cells and the expected behaviour is obtained in each case with a small number of discretization points.

 ISBN: 978-84-617-8694-7

2 Model Problem

The nondimensionalized model problem [1] describing the interactions among the cancer cell density, normal cell density and the concentration of the MDE is given by a system of reaction diffusion equations

$$\frac{\partial c}{\partial t} = \nabla \cdot (D_c \nabla c) - \nabla \cdot (\xi_c c \nabla n) + \mu_1 c(1-c-n) \quad (1)$$

$$\frac{\partial n}{\partial t} = -\delta_n m n + \mu_2 n(1-c-n) \quad (2)$$

$$\frac{\partial m}{\partial t} = D_m \nabla^2 m + \mu_3 c - \delta_m m \quad (3)$$

where D_c, D_m denote the diffusion coefficients for the cancer cells and MDE, respectively; ξ_c is the haptotaxis coefficient, μ_1, μ_2 are the production rates for the cancer and normal cells, respectively. Moreover, δ_n denotes the normal cell degradation rate, μ_3 and δ_m denote the production and decay rates of the MDE, respectively. The first and the second terms seen on the right hand side of Equation (1) describe the random motion and the haptotaxis; respectively. The last terms in Equations (1) and (2) models proliferation of the cancer and normal cells, respectively; using the logistic growth. The MDE degradation of normal cells is modelled by the first term in Equation (2) where the last term in Equation (3) is for the decay for the proteolytic enzymes MDE. The first and second terms in Equation (3) are modelling the diffusion and the production of MDEs.

For the boundary conditions, it is assumed there is no flux of cancer cells and MDE across the boundary of the problem domain and thus the Neumann type boundary conditions are considered, i.e.,

$$\frac{\partial c}{\partial \nu} = \frac{\partial m}{\partial \nu} = 0 \quad (4)$$

on the boundary $\partial\Omega$ of the domain $\Omega \subset \mathbb{R}^2$, with ν denoting the outward unit normal to $\partial\Omega$.

Finally, the problem is completed with the initial condiitons,

$$c(\mathbf{x},0) = c_0, \quad n(\mathbf{x},0) = n_0, \quad m(\mathbf{x},0) = m_0 \quad (5)$$

where c_0, n_0, m_0 are strictly positive functions which are consistent with the boundary conditions.

3 Discretization of the Model

For the spatial discretization of the model, combined application of DRBEM and FDM is used. For the spatial discretization of Equation (3) the DRBEM is used [10]. To this end,

 ISBN: 978-84-617-8694-7

Equation (3) is weighted by the fundamental solution $u^* = \frac{1}{2\pi} \ln \frac{1}{r}$ of Laplace equation

$$D_m \int_\Omega \nabla^2 m u^* d\Omega = \int_\Omega b_1 u^* d\Omega \tag{6}$$

where the nonhomogenity b_1 can be approximated by using radial basis functions $f_j(x, y)$ as

$$b_1 \left(m, c, \frac{\partial m}{\partial t} \right) = \frac{\partial m}{\partial t} - \mu_3 c + \delta_m m \approx \sum_{j=1}^{N+L} \alpha_j(t) f_j(x, y) \tag{7}$$

resulting with a linear systems of equations

$$[F] \{\alpha\} = \{b_1\} \tag{8}$$

with N and L being the number of boundary and selected interior nodes, F is the $(N + L) \times (N + L)$ matrix of distance functions f_j related to other distance functions $\hat{u}_j(x, y)$ through $\nabla^2 \hat{u}_j = f_j$.

Application of Green's second identity to both sides of equation (6) yields to

$$[H] \{m\} - [G] \left\{ \frac{\partial m}{\partial \nu} \right\} = \left([H] \left[\hat{U} \right] - [G] \left[\hat{Q} \right] \right) \{\alpha\} \tag{9}$$

with the $(N + L) \times (N + L)$ matrices $[G]$ and $[H]$ containing the boundary integrals of the fundamental solution and its normal derivative, respectively. Because of the non-flux boundary conditions the second term on the left hand side of equation (9) vanishes and back substitution of $\{\alpha\}$ gives

$$D_m [H] \{m\} = [C] \left(\left\{ \frac{\partial m}{\partial t} \right\} - \mu_3 \{c\} + \delta_m \{m\} \right) \tag{10}$$

which is an ODE, that is going to be discretized by using a combination of backward and forward Euler methods,

$$\left[(1 + \Delta t \delta_m) I - \Delta t \bar{H}_m \right] m^{k+1} = m^k + \Delta t \mu_3 c^k \tag{11}$$

which calculates the decay term implicitly and the production term explicitly. In Equation (11) k is the time level, Δt is the length of the time interval, $i, j = 1, 2, \cdots N + L$.

The ordinary differential equation (2) describing the time evolution of the normal cell density is discretized by using a combination of backward and forward Euler Methods, which discretizes the degradation term implicitly and the proliferation term explicitly; i.e.,

 ISBN: 978-84-617-8694-7

$$n_{ij}^{k+1} = \frac{1}{1+\Delta t \delta_n m_{ij}^{k+1}} \left(n_{ij}^k + \Delta t \mu_2 n_{ij}^k \left(1 - c_{ij}^k - n_{ij}^k \right) \right) \tag{12}$$

For the discretization of equation (1) , DRBEM is once again made use of with the fundamental solution of Laplace equation:

$$D_c \int_\Omega \nabla^2 c u^* d\Omega = \int_\Omega b_2 u^* d\Omega \tag{13}$$

where the nonhomogenity b_2 can be approximated by using radial basis functions $f_j(x,y)$ as

$$\begin{aligned} b_2\left(m, c, \frac{\partial c}{\partial t}\right) &= \frac{\partial c}{\partial t} - \xi_c \left[\frac{\partial c}{\partial x}\frac{\partial n}{\partial x} + \frac{\partial c}{\partial y}\frac{\partial n}{\partial y} + c\nabla^2 n \right] - \mu_1 c(1-c-n) \\ &\approx \sum_{j=1}^{N+L} \beta_j(t) f_j(x,y) \end{aligned} \tag{14}$$

After the application of the Green's second identity and using the boundary condition (4), one has

$$[\bar{H}_c]c = [C]\left[\left\{ \tfrac{\partial c}{\partial t} \right\} - \mu_1 c(1-c-n) + \xi_c \left(\left[\tfrac{\partial F}{\partial x} \right] [F]^{-1} \{d_1\} + \left[\tfrac{\partial F}{\partial x} \right] [F]^{-1} \{d_2\} + c\{d_3\} \right) \right] \tag{15}$$

where $[\bar{H}_c] = D_c[H]$ and $\{d_1\}$, $\{d_2\}$ and $\{d_3\}$ contain the FDM discretizations of the vectors $\left\{ \frac{\partial n}{\partial x} c \right\}$, $\left\{ \frac{\partial n}{\partial y} c \right\}$ and $\left\{ \nabla^2 n \right\}$ vectors by using forward difference for the first order derivatives and central difference for the second order derivatives included in the Laplacian term. For the time discretization, as for the other equations in the model, a combination of the forward and backward Euler methods is used by making use of the updated values coming from Equations (11) and (12) are used. Finally the resulting discretized ODE can be written as

$$\left[(1 + \Delta t d_4) I - \Delta t \bar{H}_c \right] c^{k+1} = c^k + \Delta t \mu_1 c^k (1 - c^k - n^{k+1}) \tag{16}$$

The solution is thus obtained iteratively using discretized equations ((11), (12) and 16) starting from the initial condition (5).

4 Numerical Results

For the numerical simulations the domain is considered to be the unit square. The number of nodes are taken as 25 and 20 inside the region and on the boundary of the domain; respectively. The time step is taken to be $\Delta t = 0.1$. Initially it is assumed that the cancer

 ISBN: 978-84-617-8694-7

cells penetrate a short distance while the space is occupied mainly by the normal cells and since the MDEs are produced by the cancer cells its initial concentration is also assumed to be proportional to the initial cancer cell density (Figure 1) . The boundary conditions are the non-flux boundary conditions which are given by (4). The parameter ranges in [1] are used and the following parameters are fixed in calculations:

$$D_c = 10^{-4}, \quad D_m = 10^{-2}, \quad \delta = 10, \quad \mu_3 = 0.075 \quad \delta_m = 0.15$$

In the sets of Figures 2-4 the time evolution of cancer and normal cell densities and MDE concentration are shown. For the figures, except Figure 3, the haptotaxis coefficient is taken as $\xi_c = 0.002$ and $\xi_c = 0.01$ is taken in Figure 3. The production rates for the cancer and normal cells are taken as $\mu_1 = 0.3$ and $\mu_2 = 0.5$ in Figures 2, 3. In Figure 4 in order to see the effect of bigger proliferation rate for cancer cells, the proliferation rate is doubled, i.e. $\mu_1 = 0.6$ is taken.

In Figure 2 the general behaviour of the haptotaxis-only model (1)-(3) can be observed that as time evolves the cancer cells use the MDEs they produce to migrate through the region and invade more than 80 % of the region by $t = 50$. Moreover, the corresponding decay is visible in the normal cell density in Figures 2(d)-2(f).

In the next set of figures (Figure 3), the effect of haptotaxis coefficient is tested. The coefficient is increased by a factor of 5 in order to see the effect of haptotaxis on the invasion.In these figures the haptotactic hills are observed showing that the cells construct larger clusters at the leading edge of the tumour by the effect of increased haptotaxis.

In the last set of figures the effect of proliferation is analysed. One can observe that with a higher proliferation rate, the cancer cells invade more, with a larger cluster.

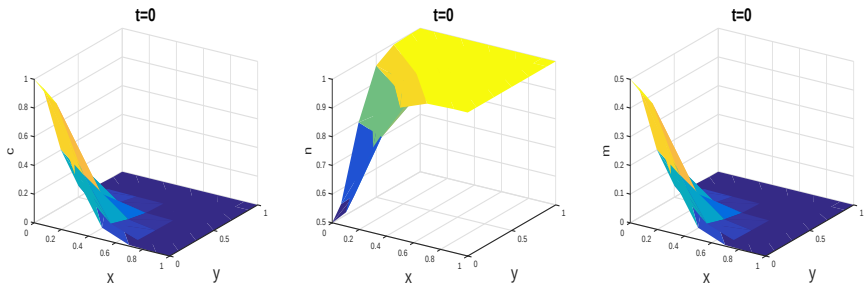
5 Conclusion

In this study, a numerical method is proposed for the haptotaxis-only model (1)-(3).In order to discretize the space derivatives in the model system, DRBEM is used with the fundamental solution of Laplace equation. However, the nonhomogenity for the cancer cell density contains the first and second order space derivatives and their products and to overcome this difficulty FDM is used in combination with DRBEM. The method is tested in a two dimensional region, which is not easy to handle numerically due to the nonlinearities. By using the boundary-only nature of DRBEM the proposed method gives consistent results with the expected behaviour of the model with a small number of discretization points.

References

[1] M.A.J. Chaplin, G. Lolas, *Mathematical Modeling of Cancer Invasion of Tissue: Dynamic Heterogeneity*, Networks and Heterogeneous Media, **1(3)** (2006) 399-439.

 ISBN: 978-84-617-8694-7



(a) (b) (c)

Figure 1: Initial Conditions

[2] R.A. Gatenby, E.T. Gawlinski, *A reaction-diffusion model of cancer invasion*, Cancer Research, **56** (1996) 5745-5753.

[3] C. Märkl, G. Meral, C. Surulescu, *Mathematical Analysis and numerical simulations for a system modeling acid-mediated tumor cell invasion*, International Journal of Analysis, **2013** (2013), 1-15.

[4] G. Meral, C. Surulescu, *Mathematical modelling, analysis and numerical simulations for the influence of heat shock proteins on tumour invasion*, Journal of Mathematical Analysis and Applications, **408**, (2013), 597-614.

[5] Z. Szymanska, J. Urbanski, A. Marciniak-Czochra, *Mathematical modelling of the influence of heat shock proteins on cancer invasion of tissue*, Journal of Mathematical Biology **58**, (2009), 819844.

[6] R. A. Gatenby, E.T. Gawlinski, A.F. Gmitro, B. Kaylor, R.J. Gillies, *Acid-Mediated tumor invasion: a multidisciplinary study*, Cancer Research, **66**, (2006), 5216-5223.

[7] N.K. Martin, E. A. Goffney, R.A. Gatenby, P.K. Maini, Tumour-stromal interactions in acid-mediated invasion: A mathematical model, Journal of Theoretical Biology, **267**, (2010), 461-470.

[8] G. Meral,C. Stinner, C. Surulescu, On a multiscale model involving cell contractivity and its effects on tumor invasion, Discrete and Continuous Dynamical Systems Series B, **20**, (2015), 189-213.

[9] G. Meral, DRBEM solution of the acid-mediated tumour invasion model with time-dependent carrying capacities,, European Journal of Computational, (2017) Mechanics,http://dx.doi.org/10.1080/17797179.2017.1306833

 ISBN: 978-84-617-8694-7

[10] P.W. Partridge, C. A. Brebbia and L.C. Wrobel, *The dual reciprocity boundary element method*, Computational Mechanics Publications, Southampton, Boston, 1992.

 ISBN: 978-84-617-8694-7

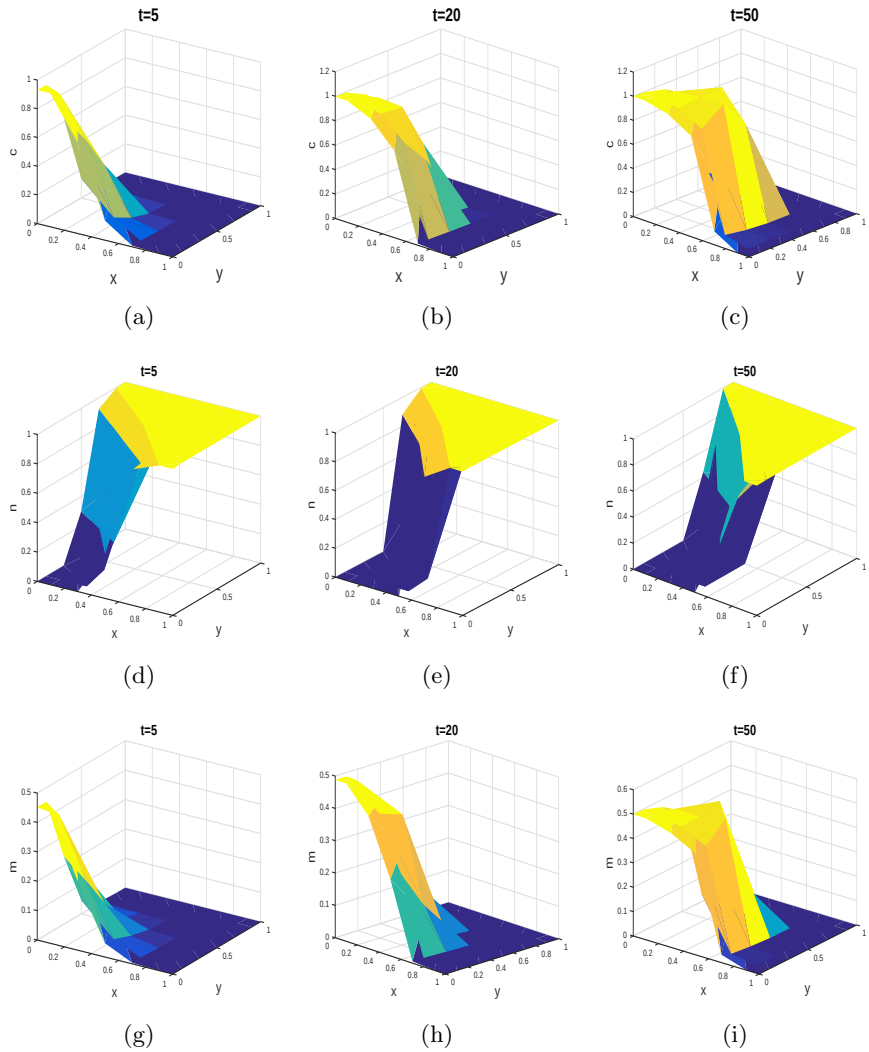
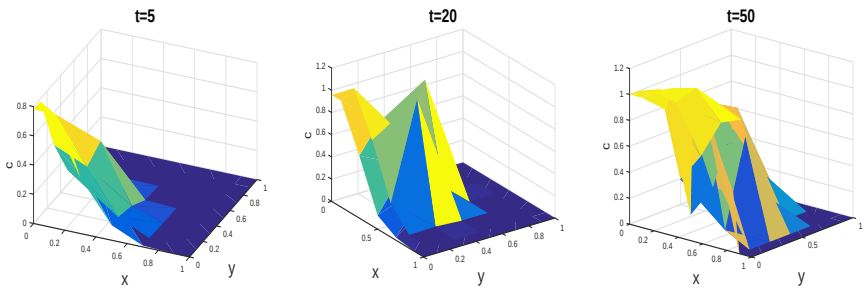


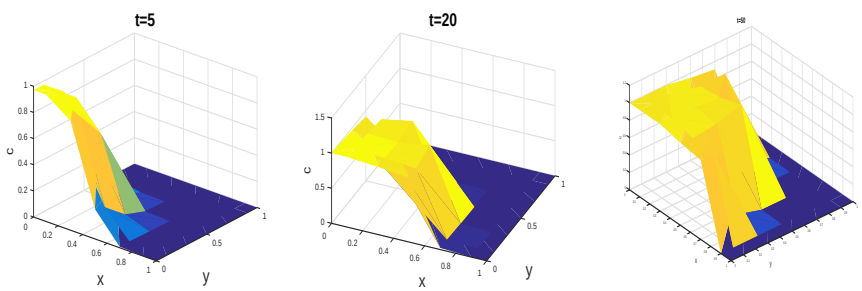
Figure 2: Behaviour of the system at different times

 ISBN: 978-84-617-8694-7



(a) (b) (c)

Figure 3: Effect of haptotaxis



(a) (b) (c)

Figure 4: Effect of higher proliferation rate for cancer cells

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

A parallel multi-step Power method for computing PageRank

H. Migallón[1], V. Migallón[2], J. A. Palomino[2] and J. Penadés[2]

[1] *Departamento de Física y Arquitectura de Computadores, Universidad Miguel Hernández, Spain*

[2] *Departamento de Ciencia de la Computación e Inteligencia Artificial, Universidad de Alicante, Spain*

emails: hmigallon@umh.es, violeta@ua.es, japb83@gmail.com, jpenades@ua.es

Abstract

In this work a multi-step technique based on the Power method for accelerating the parallel computation of PageRank is proposed. This new iterative scheme reduces the computation of the PageRank vector by eliminating synchronization points at which a process must wait for information from other processes. A hybrid MPI/OpenMP parallel implementation of this technique has been developed. The performance of the parallel model has been analyzed with realistic test data, showing that the multi-step algorithms can significantly speed up the convergence time with respect to the parallel Power algorithm.

Key words: PageRank, multi-step method, parallel algorithms, shared memory, distributed memory

1 Introduction

The PageRank algorithm, used by the Google search engine, is one of the most known and influential algorithms for determining the relevance of Web pages [12]. The core of this algorithm involves using the Power method to compute successive iterates that converge to the principal eigenvector of the Markov chain representing the Web link graph. The algorithm makes use of two key ideas: first, that links between Web pages provide information about their importance, and second, that the relationship between importance and linking is recursive. Given an ordered set of n Web pages, we can summarize the links between

 ISBN: 978-84-617-8694-7

them by means of the adjacency matrix of the Web graph $G = [g_{ij}]_{i,j=1}^{n}$, with elements $g_{ij} = 1$ when there is a link from page j to page i, with $i \neq j$, and zero otherwise. From this matrix we can construct a transition matrix $P = [p_{ij}]_{i,j=1}^{n}$ as follows: $p_{ij} = \frac{g_{ij}}{c_j}$ if $c_j \neq 0$ and 0 otherwise, where $c_j = \sum_{i=1}^{n} g_{ij}$, $1 \leq j \leq n$, represents the number of out-links from a page j. For pages with a nonzero number of out-links, i.e., $c_j \neq 0$ for all j, $1 \leq j \leq n$, the matrix P is column stochastic. Thus each element of this matrix has values between 0 and 1, and the sum of the components of each column is 1. In this case, the PageRank vector can be obtained by solving $Px = x$. Since we are interested in a probability distribution, the sum of the components of x is assumed to be one. Algorithm 1 shows the original Power method [14] for the PageRank computation where $e = (1, 1, \ldots, 1)^T$. Note that we use the L_1 norm $\|x\|_1 = \sum_{i=1}^{n} |x_i|$.

Algorithm 1: Power method.

Initialization $x^{(0)} = \frac{e}{n}$, $l = 0$;
repeat
  $x^{(l+1)} = Px^{(l)}$;
  $\delta = \|x^{(l+1)} - x^{(l)}\|_1$;
  $l = l + 1$;
until $\delta < \epsilon$;

When the matrix $P \geq 0$ is irreducible (i.e., its graph is strongly connected) and stochastic, its largest eigenvalue in magnitude is $\lambda_{max} = 1$. Thus, Algorithm 1 converges to the eigenvector corresponding to $\lambda_{max} = 1$, and when normalized, it is the stationary probability distribution over pages under a random walk on the Web.

However, the Web contains many pages without out-links, called dangling nodes. Dangling pages present a problem for the mathematical PageRank formulation because in this case the matrix P is non-stochastic and therefore Algorithm 1 can not be used. Moreover, the matrix irreducibility is not satisfied for a Web graph. In order to overcome these difficulties, Page and Brin [12] change the transition matrix P to a column stochastic matrix $\bar{P} = \alpha(P + vd^T) + (1-\alpha)ve^T$, where $d \in \Re^n$ is the dangling page indicator defined by $d_i = 1$ if and only if $c_i = 0$ and the vector $v \in \Re^n$ is some probability distribution over pages. This model means that the random surfer jumps from a dangling page according to a teleportation distribution v. Originally uniform teleportation $v = \frac{e}{n}$ was used. Then, setting α such that $0 < \alpha < 1$, Algorithm 1 can be reformulated using the matrix $\bar{P}$, obtaining Algorithm 2.

A key parameter in this model is the damping factor α that determines the weight given to the Web link graph in the model. In the original formulation of PageRank [12], the Power method was applied using $\alpha = 0.85$. Notice that, a higher value of α (close to 1) yields to a model that is mathematically closer to the actual link structure of the

 ISBN: 978-84-617-8694-7

Algorithm 2: Power method for solving $\bar{P}x = x$.

Initialization $x^{(0)} = \frac{e}{n},\ l = 0$;
repeat
  $x^{(l+1)} = \alpha P x^{(l)}$;
  $\gamma = \|x^{(l)}\|_1 - \|x^{(l+1)}\|_1$;
  $x^{(l+1)} = x^{(l+1)} + \gamma v$;
  $\delta = \|x^{(l+1)} - x^{(l)}\|_1$;
  $l = l + 1$;
until $\delta < \epsilon$;

Web but makes the computation more difficult [8] being that this parameter α controls the asymptotic rate of convergence and as $\alpha \to 1$, the expected number of iterations required for convergence increases dramatically [10]. Therefore, new approaches for accelerating the PageRank computation are required.

In the last years, several techniques to accelerate the Power method have been developed such as extrapolation methods [7], adaptive methods [6] or Arnoldi-type algorithms [5, 15, 16] and approaches based on linear system formulations [3, 4, 13]. Due to the large size of the Web link graph, to deal with realistic problems, a promising way of accelerating PageRank is the parallel processing, see e.g., [1, 3, 4] and the references cited therein.

In this paper a multi-step technique based on the Power method for accelerating the parallel computation of PageRank is proposed. This technique aims to reduce the number of power iterations by eliminating synchronization points at which a process must wait for information from other processes. Section 2 describes this new iterative technique and its formulation in order to assure the convergence. Section 3 reports the experimental results showing the behavior of the designed algorithms for realistic test data on a current Symmetric Multi-Processing (SMP) supercomputer. Finally, in Section 4 we give some conclusions.

2 Multi-step Power Algorithm

Consider that P is partitioned into p row blocks of the form $P = \left[P_1^t\ P_2^t\ \cdots\ P_p^t\right]^t$, where each P_i, $1 \le i \le p$, is a matrix of order $n_i \times n$, with $\sum_{i=1}^{p} n_i = n$. Analogously, we consider the iterate vectors $x^{(l)}$ and v partitioned according to the block structure of P. Obviously, the Power method for solving $\bar{P}x = x$ can be executed in parallel. In this case each process actualizes a block of the vector $x^{(l+1)}$ and a synchronization of all processes is performed at each iteration in order to construct the global iterate vector $x^{(l+1)}$. Due to this synchronization, we can use the formulation of Algorithm 2 because the property of

preserving the L_1 norm remains valid.

In order to reduce the work involved in this iterative algorithm we propose a parallel multi-step Power (MSTEP) algorithm (Algorithm 3). In this algorithm each part of $x^{(l+1)}$ is updated more than once ($q > 1$ times) without waiting for the other parts of $x^{(l+1)}$ to be updated and therefore eliminating some synchronization points. In this case the condition of preserving the L_1 norm is not assured and therefore, Algorithm 2, in its current formulation, can not be used for our purpose. For this reason, Algorithm 3 has been formulated such that the condition on the L_1 norm is not needed.

Algorithm 3: Parallel multi-step method for solving PageRank (MSTEP).

Initialization $x^{(0)} = \frac{e}{n}, v = \frac{e}{n},\ q,\ l = 0;$
for $i = 1, 2, \ldots, p$, **do in parallel**
  repeat
    $y^{(0)} = x^{(l)};$
    for $k = 1, 2, \ldots, q$, **do**
      $y_i^{(k)} = \alpha P_i y^{(k-1)};$
      $\gamma = \alpha d^T y^{(k-1)} + (1-\alpha)\|y^{(k-1)}\|_1;$
      $y_i^{(k)} = y_i^{(k)} + \gamma v_i;$ **(1)**
      $y_j^{(k)} = y_j^{(k-1)},\ j \neq i\ ;$
    end
    $x_i^{(l+1)} = y_i^{(q)};$
    Perform an all-gather operation to obtain $x^{(l+1)} = [x_1^{(l+1)}, \ldots, x_p^{(l+1)}];$
    Compute $\|x_i^{(l+1)} - x_i^{(l)}\|_1;$
    Perform a sum all-to-all reduction over $\|x_i^{(l+1)} - x_i^{(l)}\|_1$ to obtain
    $\delta = \|x^{(l+1)} - x^{(l)}\|_1;$
    $l = l + 1;$
  until $\delta < \epsilon;$
  Compute $\|x_i^{(l)}\|_1;$
  Perform a sum all-to-all reduction over $\|x_i^{(l)}\|_1$ to obtain $\|x^{(l)}\|_1;$
  Compute $\pi_i = \frac{x_i^{(l)}}{\|x^{(l)}\|_1};$
  Perform a gather operation over π_i to obtain $\pi = \frac{x^{(l)}}{\|x^{(l)}\|_1}$ in a root process;
end

Note that a relaxation parameter $\beta > 0$ can be introduced in the MSTEP Algorithm (that is, in Algorithm 3) and replace the computation of $y_i^{(k)}$ in (1) with the equation

 ISBN: 978-84-617-8694-7

$y_i^{(k)} = \beta(y_i^{(k)} + \gamma v_i) + (1-\beta)y_i^{(k-1)}$. In this case we obtain a parallel relaxed multi-step algorithm, that we have called RMST Algorithm. Clearly, with $\beta = 1$ equation (1) is recovered.

3 Numerical experiments

In order to investigate and analyze the MSTEP and RMST algorithms described here, we have used several datasets of different sizes, available from the Laboratory for Web Algorithmics [9]; see Table 1. These transition matrices have been generated from a web-crawl [2].

Graph	n	nnz	Dgn (%)	Dens	M (GB)
uk-2002	18,520,486	298,113,762	14.91	16.01	1.32
it-2004	41,291,594	1,150,725,436	12.76	27.87	4.75
webbase-2001	118,142,155	1,019,903,190	23.41	8.63	5.12
uk-2006-10	93,436,772	3,130,910,405	13.52	33.50	12.71
uk-2007-05	105,896,555	3,738,733,648	12.23	35.31	15.11

Table 1: Graphs collection. n =number of nodes, nnz =number of arcs, Dgn=percentage of dangling nodes, Dens=density (arcs/nodes), M=memory requirements using CSR' format.

To compute PageRank for large domains there is no possible way to work with the matrix in its full format because the memory requirements would be too high. Therefore, to take advantage of the large number of zero elements, special schemes are required to store sparse matrices. The main goal is to represent only the nonzero elements, and to be able to perform the common matrix operations. The modified Compressed Sparse Row (CSR') format used for storing the sparse matrices [11] has involved a reduction of memory requirements of about $63-73\%$ with respect to the original CSR format.

The algorithms described here have been implemented on an HPC cluster of 26 nodes HP Proliant SL390s G7 connected through a network of low-latency QDR Infiniband-based. Each node consists of two Intel XEON X5660 hexacore at up to 2.8 GHz and 12MB cache per processor, with 48 GB of RAM. The operating system is CentOS Linux 5.6 for x86 64 bit. The parallel environment has been managed using a hybrid MPI/OpenMP model: each process is assigned to a core as follows: let p be the number of physical cores used, $p = s \times c$ indicates that s nodes of the parallel platform have been used and for each one of these nodes, c cores have been considered. Therefore, we use a philosophy of distributed shared memory (DSM) using $p = s \times c$ processes or threads. Particularly, if $s = 1$, the algorithms are executed in shared memory (SM) using $p = c$ cores on a single node. Conversely, if $c = 1$, we are working on distributed memory (DM) using $p = s$ nodes.

 ISBN: 978-84-617-8694-7

Matrix	Number of processes				
uk-2002	2	4	8	16	32
MSTEP(2)	4.36	9.91	20.35	24.91	31.29
MSTEP(4)	7.82	10.87	26.75	34.28	46.10
RMST(2)	9.61	12.11	19.11	25.82	33.58
RMST(4)	8.11	11.59	26.51	34.89	46.39
it-2004	2	4	8	16	32
MSTEP(2)	2.81	8.62	11.88	20.32	28.54
MSTEP(4)	0.94	7.23	9.34	27.80	30.64
RMST(2)	5.55	11.91	15.11	23.29	26.20
RMST(4)	4.69	8.63	14.69	28.03	34.49
webbase-2001	2	4	8	16	32
MSTEP(2)	10.75	12.55	21.30	27.16	37.06
MSTEP(4)	8.84	17.22	25.27	38.78	50.89
RMST(2)	13.35	17.62	25.73	29.43	39.78
RMST(4)	8.84	16.94	29.48	39.19	53.08
uk-2006-10	2	4	8	16	32
MSTEP(2)	3.98	9.36	16.13	21.43	29.44
MSTEP(4)	0.02	9.32	15.35	23.35	35.67
RMST(2)	3.65	8.74	14.85	28.16	26.62
RMST(4)	0.61	7.60	16.07	32.47	35.89
uk-2007-05	2	4	8	16	32
MSTEP(2)	2.41	6.60	12.03	19.66	25.02
MSTEP(4)	2.61	4.51	13.59	21.04	27.39
RMST(2)	2.17	6.25	13.93	19.70	20.80
RMST(4)	1.93	4.31	12.96	20.79	26.41

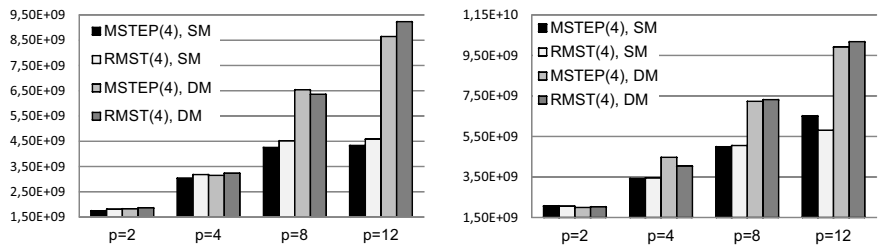
Table 2: Percentage of time reduction of the parallel MSTEP and RMST algorithms with respect to the parallel Power algorithm, $\beta = 0.98$, DM ($p = 2(2 \times 1)$, $4(4 \times 1)$, $8(8 \times 1)$), DSM ($p = 16(4 \times 4)$, $32(8 \times 4)$), $\epsilon = 10^{-6}$, $\alpha = 0.85$.

 ISBN: 978-84-617-8694-7

Matrix	Number of processes				
uk-2002	2	4	8	16	32
MSTEP(2)	4.97	12.85	24.09	25.99	34.43
MSTEP(4)	9.49	18.33	33.77	37.90	51.22
RMST(2)	27.98	32.07	36.64	42.66	48.83
RMST(4)	29.18	35.59	46.91	50.90	60.95
it-2004	2	4	8	16	32
MSTEP(2)	2.81	9.65	9.93	21.11	25.46
MSTEP(4)	5.56	12.92	16.70	35.63	45.59
RMST(2)	28.01	33.13	36.35	42.41	47.43
RMST(4)	27.74	32.66	37.53	48.33	53.05
webbase-2001	2	4	8	16	32
MSTEP(2)	3.36	13.13	20.92	27.82	38.10
MSTEP(4)	10.74	19.35	31.60	41.49	56.40
RMST(2)	36.01	40.16	44.54	50.38	57.47
RMST(4)	38.55	43.14	51.96	59.16	68.71
uk-2006-10	2	4	8	16	32
MSTEP(2)	4.05	6.07	13.11	16.06	26.53
MSTEP(4)	5.78	10.13	15.59	23.75	38.25
RMST(2)	17.30	19.28	23.19	27.50	35.86
RMST(4)	17.89	20.93	29.15	37.25	46.22
uk-2007-05	2	4	8	16	32
MSTEP(2)	3.65	7.39	12.68	21.06	24.81
MSTEP(4)	4.73	8.42	20.07	28.33	36.15
RMST(2)	14.74	18.17	22.63	30.38	37.57
RMST(4)	15.89	20.84	28.51	38.37	43.91

Table 3: Percentage of time reduction of the parallel MSTEP and RMST algorithms with respect to the parallel Power algorithm, $\beta = 0.98$, DM ($p = 2(2 \times 1)$, $4(4 \times 1)$, $8(8 \times 1)$), DSM ($p = 16(4 \times 4)$, $32(8 \times 4)$), $\epsilon = 10^{-6}$, $\alpha = 0.99$.

 ISBN: 978-84-617-8694-7



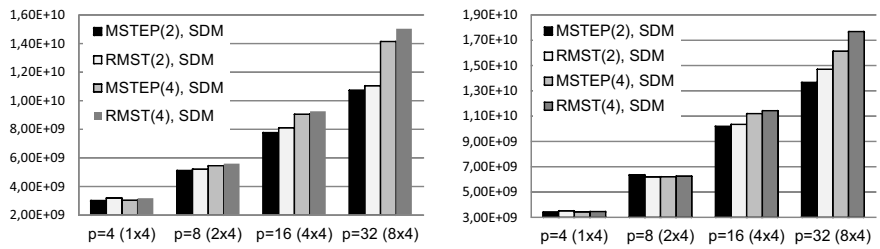
(a) uk-2002 (Flops). (b) uk-2007-05 (Flops).

Figure 1: Performance of parallel MSTEP and RMST algorithms, shared versus distributed memory, $\epsilon = 10^{-6},\ \beta = 0.99,\ \alpha = 0.99$.

The most expensive operation of these algorithms is the matrix-vector product. This is a perfectly parallel operation with several possible methods for partitioning both the matrix and the vector. In the link matrices used to calculate PageRank, the number of nonzero elements per row can differ immensely. In order to balance the calculations, a nonzero elements partitioning is used, where each node has to handle approximately the same amount of nonzero elements. Generally, for the PageRank calculations, this distribution strategy behaves better than row-wise distributions; see e.g. [1].

In the experiments reported here, we have used for the stopping criterion $\epsilon = 10^{-6}$, and we have run our algorithms for several values of α and β. The notation MSTEP(q) indicates that q steps are used in Algorithm 3. Similar notation is used for the RMST method. The synchronization points in the MSTEP methods are reduced in relation to the Power method as the number of steps q, in Algorithm 3, increases. However, applying a certain number of q steps in this algorithm without waiting for the update of other parts of the global iterate vector, is more computationally intense than applying one iteration of the Power method. As long as this overhead is minimal, the proposed acceleration is beneficial. Therefore, we must keep q small in Algorithm 3. Good choices of q, for this algorithm and its relaxed counterpart, are between 2 and 4. Tables 2 and 3 illustrate the gain obtained by these methods in relation to the parallel Power method varying the number of processes on both distributed and distributed shared memory, and using several values of α. Good choices of the relaxation parameter β, in the proposed RMST algorithms, are between 0.98 and 0.99. In these tables we have chosen $\beta = 0.98$. As it can be seen, the proposed algorithms can significantly speed up the convergence time relative to the parallel Power method, more specially for values of α close to 1. In fact, for $\alpha = 0.99$, the RMST algorithms achieve a gain in relation to the parallel Power method between 15.89% and 38.55% using 2 processes, while using 32 processes, the achieved gain by means of these algorithms was between 43.91% and 68.71%.

 ISBN: 978-84-617-8694-7



(a) uk-2002 (Flops). (b) uk-2007-05 (Flops).

Figure 2: Performance of parallel MSTEP and RMST algorithms, distributed shared memory, $\epsilon = 10^{-6},\ \beta = 0.99,\ \alpha = 0.99$.

	$\epsilon = 10^{-3}$	$\epsilon = 10^{-5}$	$\epsilon = 10^{-7}$
POWER It.	145	531	964
In/Out It. gain	13.8%	12.8%	11.1%
In/Out time gain	15.8%	15.3%	13.7%
MSTEP It. gain	74.5%	74.7%	74.9%
MSTEP time gain	14.8%	17.03%	14.5%
RMST It. gain	77.8%	78.3%	77.8%
RMST time gain	23.72%	29.7%	24.0%

Table 4: Multistep versus In/Out methods, uk-2007-05, SM $p = 8$ (1×8), $\alpha = 0.99$.

Figures 1 and 2 compare the parallel algorithms treated herein in floating-point operations per second (Flops) for several configurations. We obtain an acceptable performance using the 12 available cores of one node but not comparable with the performance obtained using distributed memory.

However, to deal with larger problems, the strategies of parallelization require to use at the same time the benefits of shared and distributed memory. Generally, the best parallel results have been obtained for the RMST algorithms using from 1 to 4 cores in each node. Table 4 compares, using several values of ϵ for the stopping criterion, the multi-step algorithms treated here with the inner-outer methods proposed in [3]. As it can be seen, the parallel RMST algorithms accelerate the PageRank computation more significantly than these parallel inner-outer algorithms.

4 Conclusions

In this paper a multi-step technique based on the Power method for accelerating the parallel computation of PageRank is proposed. This technique aims to reduce the number of power iterations by eliminating synchronization points at which a process must wait for information from other processes. The parallel implementation has been developed using a mixed MPI/OpenMP model to exploit parallelism beyond a single level. In order to investigate and analyze the parallel MSTEP and RMST algorithms described here, we have used several realistic large datasets. Taking into account the characteristics of the transition matrices, a modified Compressed Sparse Row format has been used to store these sparse matrices that has involved a reduction of memory requirements of about $63 - 73\%$ with respect to the popular CSR format. In order to balance the calculations among nodes, a nonzero elements partitioning is used, where each node has to handle the same amount of nonzero elements. The numerical experiments show that these new parallel algorithms can considerably improve the convergence rate relative to the Power method.

Acknowledgements

This research was supported by the Spanish Ministry of Economy and Competitiveness (MINECO) and the European Commission (FEDER funds) under Grant Number TIN2015-66972-C5-4-R.

References

[1] J. Arnal, H. Migallón, V. Migallón, J. A. Palomino and J. Penadés, *Parallel relaxed and extrapolated algorithms for computing PageRank*, J. Supercomput. **70** (2014) 637–648.

[2] P. Boldi, B. Codenotti, M. Santini and S. Vigna, *Ubicrawler: A scalable fully distributed Web crawler*, Softw. Pract. Expe. **34** (2004) 711–726.

[3] D. Gleich, A. Gray, C. Greif and T. Lau, *An inner-outer iteration for computing PageRank*, SIAM J. Sci. Comput. **32(1)** (2010) 349–371.

[4] D. Gleich, L. Zhukov and P. Berkhin, *Fast parallel PageRank: A linear system approach*, Technical Report YRL-2004-038, Yahoo! Research Labs, 2004.

[5] G. H. Golub and C. Greif, *An Arnoldi-type algorithm for computing PageRank*, BIT **46(4)** (2006) 759–771.

 ISBN: 978-84-617-8694-7

[6] S. D. Kamvar, T. H. Haveliwala and G. H. Golub, *Adaptive methods for the computation of PageRank*, Linear Algebra Appl. **386** (2004) 51–65.

[7] S. D. Kamvar, T. H. Haveliwala, C. D. Manning and G. H. Golub, *Extrapolation methods for accelerating PageRank computations*, In: Proceedings of the Twelfth International World Wide Web Conference, pp. 261–270. ACM Press, 2003.

[8] S. D. Kamvar, *Numerical Algorithms for Personalized Search in Self-organizing Information Networks*, Princeton University Press, Princeton, New Jersey, 2010.

[9] Laboratory for Web Algorithmics, http://law.di.unimi.it, 2002.

[10] A. Langville and C. D. Meyer, *Google's Pagerank and Beyond: The Science of Search Engine Rankings*, Princeton University Press, Princeton, New Jersey, 2006.

[11] H. Migallón, V. Migallón, J. A. Palomino and J. Penadés, *Parallelization strategies for computing PageRank*, In: Topping, B.H.V., Adam, J.M., Pallarés, F.J., Bru, R., Romero, M.L. (eds.) Proceedings of the Seventh International Conference on Engineering Computational Technology, Paper 29. Civil-Comp Press, Stirlingshire, United Kingdom, doi:10.4203/ccp.94.29, 2010.

[12] L. Page, S. Brin, R. Motwani and T. Winograd, *The PageRank citation ranking: Bringing order to the Web*, Technical Report, Stanford Digital Library Technologies Project, 1999.

[13] B. Y. Pua, T. Z. Huanga and C. Wena, *A preconditioned and extrapolation-accelerated GMRES method for PageRank*, Appl. Math. Lett. **37** (2014) 95–100.

[14] J. H. Wilkinson, *The algebraic eigenvalue problem*, Oxford University Press, Oxford, 1998.

[15] G. Wu and Y. Wei, *An Arnoldi-Extrapolation algorithm for computing PageRank*, J. Comput. Appl. Math. **234** (2010) 3196–3212.

[16] G. Wu and Y. Wei, *A Power Arnoldi algorithm for computing PageRank*, Numer. Linear Algebra Appl. **14** (2007) 521–546.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

Nash equilibria and negotiation with quadratic functions

P. Millán[1], L. Orihuela[1] and J.F. Carbonell-Márquez[2]

[1] *Departamento de Ingeniería, Universidad Loyola Andalucía*

[2] *Departamento de Mecánica, Universidad de Córdoba*

emails: pmillan@uloyola.es, dorihuela@uloyola.es, jcarbonell@uco.es

Abstract

This paper extends some results on the existence, uniqueness, and stability of a Nash equilibrium in a distributed, negotiated decision process with a class of quadratic convex functions. Based on these results, it is derived the optimal decision of any given player when all the others act as its opponents. Furthermore, the paper conjectures the properties of the Nash equilibrium for the case with constrained decisions, provided the existence of a unique, stable Nash equilibrium for the equivalent unconstrained problem.

Key words: Noncooperative Games, Optimal Control and Differential Games, Quadratic Programming

MSC 2000: 91A10, 49N90, 90C20, 91A80, 91A25, 91B06.

1 Introduction

Nowadays, it is becoming more and more frequent the shift from centralized to distributed or agent-based decision making, which finds important applications such as energy management in electrical grids, robotic systems, or cloud computing. Typically, each of these agents (also called players) has its local cost function and, therefore, the very definition of *optimal decisions* gets blurred and is replaced by concepts like Pareto efficiency and Nash equilibrium. On the one hand, a set of decisions is Pareto optimal when it is not possible to make any player better off without making at least another one worse off. On the other hand, a Nash equilibrium is a set of decisions that every player maintains if the decision of the others do not change. It is known that in general Nash equilibriums are not-Pareto efficient, but in negotiated competitive situations, Pareto optimals are commonly unstable, see [1].

 ISBN: 978-84-617-8694-7

This paper studies some interesting properties of a the distributed negotiation process in the work [2], which introduced conditions for the existence, uniqueness, and stability of a Nash equilibrium in a distributed, negotiated decision process with a class of quadratic convex functions. First, it is derived the optimal decision of any player when the goal of the rest of them consist of trying to harm it by maximizing its cost function. After that, the paper conjectures, based on simulation results, the existence and location of a unique Nash equilibrium for the case with constrained decisions, provided the existence of a unique, stable Nash equilibrium for the equivalent unconstrained problem.

2 Problem description

Consider a game-theoretical scenario defined over a set of p players, $i = 1, 2, \ldots, p$, connected through a network described by an undirected graph $G = (V, E)$, with vertexes $V = \{1, 2, \ldots, p\}$ and edges $E \subseteq V \times V$.

At each negotiation step k, each player makes its decision by minimizing an associated quadratic cost function $J_i(k) \triangleq \mathbf{x_i}(k)^\top A_i \mathbf{x_i}(k) - 2\mathbf{x_i}(k)^\top \Big(\mathbf{b_i} + \sum_{j \in N_i} C_{ij}\mathbf{x_j}(k-1)\Big)$, where $\forall i \in V$, $A_i \in \mathbb{R}^{q_i \times q_i}$ is a diagonal positive definite matrix, $\mathbf{b_i} \in \mathbb{R}^{q_i}$ is a vector and $C_{ij} \in \mathbb{R}^{q_i \times q_j}$ are matrices of appropriate dimensions, N_i is the set comprised by those players connected with player i, and $\mathbf{x_j}(k-1)$ are the decisions made and communicated by other players at the previous negotiation step.

Definition 1. A Nash equilibrium is a situation in which no player changes its decisions as long as the rest of the players keep the same decisions as well. It is defined by a set of decisions $x^* \triangleq \left[\mathbf{x_1^*}^\top \; \mathbf{x_2^*}^\top \; \cdots \; \mathbf{x_p^*}^\top\right]^\top$ such that $\mathbf{x_i^*} = arg\min_{\mathbf{x_i}} \; \mathbf{x_i}^\top A_i \mathbf{x_i} - 2\mathbf{x_i}^\top \left(\mathbf{b_i} + \sum_{j \in N_i} C_{ij}\mathbf{x_j^*}\right), \forall i.$

Throughout this paper, if A is a matrix and $\mathbf{b}$ is a vector, then $[A]_{ij}$ is the element at the i-th row and j-th column of matrix A, and $[\mathbf{b}]_i$ is the i-th element of vector $\mathbf{b}$.

3 The unconstrained case

This section studies the case with unconstrained decision variables. In this case, at every negotiation step k, the decisions that minimize the cost $J_i(k)$ can be found by determining $\mathbf{x_i}(k)$ such that $\frac{\partial J_i(k)}{\partial \mathbf{x_i}(k)} = 0$, since functions $J_i(k)$ are convex with respect to $\mathbf{x_i}(k)$.

Let us define $\mathcal{A} \triangleq diag\{A_1, A_2, \ldots, A_p\}$, $\mathcal{C} \triangleq [c_{ij}]$, where $C_{ij} \equiv 0$ if edge $(i, j) \notin E$. It was stated in [2] that the game admits a unique Nash equilibrium x^*, given by $x^* = (I - \mathcal{A}^{-1}\mathcal{C})^{-1}\mathcal{A}^{-1}\mathcal{b}$ iff $\det(I - \mathcal{A}^{-1}\mathcal{C}) \neq 0$ holds. The equilibrium x^* is globally asymptotically stable iff the eigenvalues of matrix $\mathcal{A}^{-1}\mathcal{C}$ belong to the unit circle.

 ISBN: 978-84-617-8694-7

Let us move now to the particular situation in which agent i takes a decision assuming that the rest of the players' objective is $J_j(k) = -J_i(k), \forall j$. This represents a competitive situation in which agent i decides its options taking into account that the other players are opponents pursuing to damage agent i. We study the max-min optimization problem, since agent i takes a decision before the rest of the players.

Theorem 1 *Consider the max-min problem in which agent i minimizes J_i and, later on, the rest of agents maximize the same cost function. Then, if $rank\,[C_{i1}\dots C_{ip}] \leq q_i$ holds, the optimal decision for player i taking into account every possible decision of the other players can be obtained as $\mathbf{x_i^*} = arg\min \mathbf{x_i}^\top A_i \mathbf{x_i} - 2\mathbf{x_i}^\top \mathbf{b_i}$ subject to $[C_{i1}\dots C_{ip}]^\top \mathbf{x_i}$. Otherwise, the optimal decision is the trivial one, that is, $\mathbf{x_i^*} = \mathbf{0}$.*

When $rank\,[C_{i1}\dots C_{ip}] \leq q_i$ holds, agent i can choose its decision in such a way that J_i is minimized for any decision of the other players.

Example 1 *Consider a two-player game with* $A_1 = \begin{pmatrix} 3 & 0 \\ 0 & 5 \end{pmatrix}, b_1 = \begin{pmatrix} 1 \\ 3 \end{pmatrix}, C_{12} = \begin{pmatrix} -5 \\ 2 \end{pmatrix}$. *Figure 1 illustrates the paraboloids for agent* 1 *at instant k for three decisions made by player 2 at $k-1$. Since this example satisfies the condition in Theorem 1 ($rank\,[C_{12}] = 1$,$q_1 = 2$), we see that there exists a region of the state-space (the intersection between the paraloboloids) different to the origin that is a solution for the max-min problem. The minimum of the cost is marked with a red dot. The reader can check that this minimum is lower that that obtained with the trivial decision, in a green dot.*

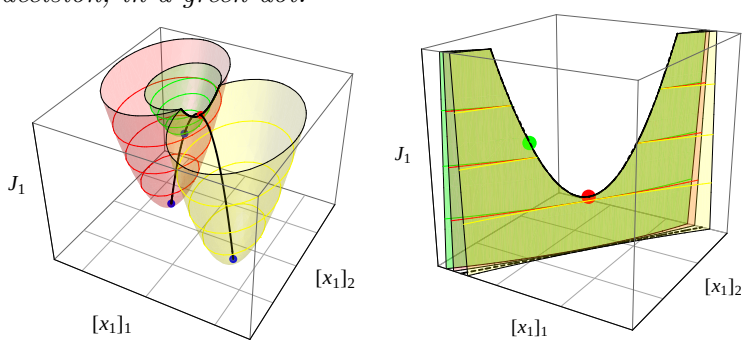


Figure 1: Paraboloids of J_1 for different $x_2(k-1)$ and its intersection.

4 The constrained case

Let us consider now the situation in which the decision variables are subjected to some restrictions. In this case, each player must find its decision vector $\mathbf{x_i}(k)$ by solving the optimization problem $\min_{\mathbf{x_i}(k)} J_i$, subject to $[\underline{\mathbf{x}}_\mathbf{i}]_l \leq [\mathbf{x_i}(k)]_l \leq [\overline{\mathbf{x}}_\mathbf{i}]_l$, $\forall l \in \{1, 2\dots, q_i\}$, where

 ISBN: 978-84-617-8694-7

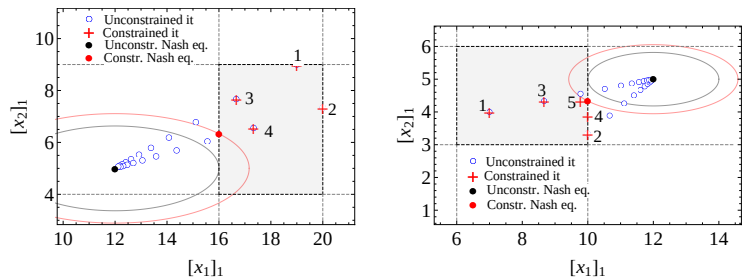


Figure 2: Decisions of both players for the constrained and unconstrained cases.

$\underline{\mathbf{x}}_\mathbf{i}$,$\overline{\mathbf{x}}_\mathbf{i}$ are two vectors containing the minimum and maximum bounds, respectively, for each decision variable.

It was proved in [2] that the constrained negotiation will converge to the Nash equilibrium for the unconstrained game provided that i) this Nash equilibrium belongs to the feasible set, and ii) the system $\tilde{y}(k+1) = \mathcal{A}^{-1}\mathcal{C}\tilde{y}(k)$ admits a quadratic Lyapunov fuction (QLF) with associated matrix P. The next example studies, nonetheless, the situation in which condition i) is not satisfied.

Example 2 *Consider a two-player game with one variable to optimize. The cost functions are described by* $A_1 = 1,\ b_1 = 2,\ C_{12} = 2;\ A_2 = 3,\ b_2 = 3,\ C_{21} = 1$. *Figure 2 shows two negotiation for different feasible sets. Note that the negotiation converges to an equilibrium although the Nash equilibrium for the unconstrained game is outside the feasible set.*

Conjecture 1 *There exists a unique Nash equilibrium for the constrained negotiation when the equivalent unconstrained problem has a unique Nash equilibrium outside of the feasible region. The Nash equilibrium belongs to the border of the feasible set, but it is not the point with minimum* P*-norm, being* P *the associated Lyapunov matrix for the unconstrained game.*

Acknowledgements

This work has been partially supported by MCyT (Grant DPI2013-44135-R) and AEI/FEDER (Grant TEC2016-80242-P).

References

[1] T. Basar and G. J. Olsder. *Dynamic noncooperative game theory*. SIAM, 1999.

[2] L. Orihuela, P. Millán, and J. F. Carbonell-Márquez. Distributed negotiation with a class of quadratic cost functions. In *IFAC World Conference*, page to appear, Toulouse, France, July 2017.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

On third order generalized periodic impulsive problems

Feliz Minhós[1] and Rui Carapinha[2]

[1] *Departamento de Matemática, Escola de Ciências e Tecnologia, Universidade de Évora. Rua Romão Ramalho, 59, 7000-671 Évora, Portugal*

[2] *Centro de Investigação em Matemática e Aplicações (CIMA), Instituto de Investigação e Formação Avançada, Universidade de Évora. Rua Romão Ramalho, 59, 7000-671 Évora, Portugal*

emails: fminhos@uevora.pt, gene.destro@gmail.com

Abstract

This work concerns to third order impulsive problems with the fully nonlinear equation

$$u'''(x) = f(x, u(x), u'(x), u''(x))$$

for a.e. $x \in [0,1] \setminus \{x_1, ..., x_m\}$ where $f : [0,1] \times \mathbb{R}^4 \to \mathbb{R}$ is a L^1-Carathéodory function, the periodic boundary conditions

$$u^{(i)}(0) = u^{(i)}(1), \ i = 0, 1, 2,$$

and the impulsive effects given by the generalized functions

$$u^{(i)}\left(x_j^+\right) = I_{ij}\left(u(x_j), u'(x_j), u''(x_j)\right),$$

with $i = 0, 1, 2$, $x_j \in (0,1)$, for $j = 1, ..., m$, such that $0 = x_0 < x_1 < ... < x_m < x_{m+1} = 1$, and $I_{ij} : \mathbb{R}^3 \to \mathbb{R}$ are continuous and nondecreasing functions in all variables.

2010 Mathematics Subject Classification: 34B37 ; 34B15 ; 92E20.

Keywords: Higher order boundary value problems, generalized impulsive conditions, upper and lower solutions, fixed point theory.

 ISBN: 978-84-617-8694-7

1 Introduction

This paper presents a nonlinear periodic third order impulsive problem composed by the fully differential equation

$$u'''(x) = f\left(x, u(x), u'(x), u''(x)\right) \tag{1}$$

for a.e. $x \in J \backslash \{x_1, ..., x_m\}$ with $J := [0,1]$, where $f : J \times \mathbb{R}^3 \to \mathbb{R}$ is a L^1-Carathéodory function, the periodic boundary conditions

$$u^{(i)}(0) = u^{(i)}(1), \ i = 0, 1, 2, \tag{2}$$

and the impulsive effects given by some generalized function in the form

$$u^{(i)}\left(x_j^+\right) = I_{ij}\left(u(x_j), u'(x_j), u''(x_j)\right), \tag{3}$$

where $i = 0, 1, 2$, $x_j \in (0,1)$, for $j = 1, ..., m$, such that $0 = x_0 < x_1 < ... < x_m < x_{m+1} = 1$ and $I_{ij} : \mathbb{R}^3 \to \mathbb{R}$ are continuous and nondecreasing functions in all variables.

Different types of third order boundary value problems (separated, periodic, multipoint, with delays, integro-differential, functional,...) have been studied by many authors and several methods, such as fixed point theory, topological and coincidence degree, lower and upper solutions, cone theory,..., and can describe real phenomena in medicine, physics, agriculture, biology, economics,...(see, for example, [2, 3, 4] and the references therein).

Impulsive problems are particularly well adapted to models where there are sudden changes at some moments, and they have been the subject of growing interest (see, for instance, [5, 6]). These jump situations may happen in many fields, such as, population dynamics, control theory, chemistry, ...

To our best knowledge, it is the first time where third order periodic impulsive problems are considered with the instantaneous changes, depending on the unknown function and its first and second derivatives, given by generalized functions. In this way, problem (1)-(3) covers cases where the jumps in each moment depend not only on the value of the function on this instant, but also on the velocity and the convexity of the solution in the referred moment.

The main tools rely on a perturbed and truncated auxiliary problem, on an iterative technique, not necessarily monotone, as in [1], and lower and upper solutions method. We

 ISBN: 978-84-617-8694-7

point out that, the nonlinear part must verify only a local monotone condition (see (11)) and no assumption on its periodicity or asymptotic growth is needed.

In Section 2 we describe the class of functions to be considered and an explicit expression for the solution of the associated linear problem. Section 3 contains the main result: an existence and localization theorem, that is, with some qualitative information on the solution. In Section 4 we present an example to illustrate the potentialities of the main theorem.

2 Definitions and auxiliary results

In this section we present some notations, definitions and auxiliary results, to be used forward.

For $m \in \mathbb{N}$, let $0 = x_0 < x_1 < ... < x_m < x_{m+1} = 1$, $D = \{x_1, ..., x_m\}$ and

$$u(x_j^{\pm}) := \lim_{x \to x_j^{\pm}} u(x).$$

Definition 1 *Denote by $PC(J)$ the set of functions $u : J \to \mathbb{R}$ continuous on $J \setminus D$ where $u\left(x_j^+\right)$ and $u\left(x_j^-\right)$ exist with $u\left(x_j^-\right) = u(x_j)$, for $k = 1, 2, ..., m$.*

For $u \in PC(J)$, we define the norm by

$$\|u\| = \sup_{x \in J} |u(x)|.$$

Consider $PC^l(J)$, $l = 1, 2$, as the space of the real-valued functions u, such that $u^{(l)} \in PC(J)$, $u^{(l)}\left(x_j^+\right)$ and $u^{(l)}\left(x_j^-\right)$ exist with $u^{(l)}\left(x_j^-\right) = u^{(l)}(x_j)$, for $l = 0, 1, 2$ and $j = 1, 2, ..., m$.

Therefore $u \in PC^2(J)$ can be written as

$$u(x) = \begin{cases} u_0(x) & \text{if } x \in [0, x_1], \\ u_1(x) & \text{if } x \in (x_1, x_2], \\ \vdots & \\ u_m(x) & \text{if } x \in (x_m, 1], \end{cases} \tag{4}$$

 ISBN: 978-84-617-8694-7

where $u_i := u_i|_{(x_i,x_{i+1}]}$ with $u_i \in C^2(x_i, x_{i+1}]$ for $i = 0, 1, ..., m$.

Denote, for $n \in \mathbb{N}$,

$$PC_D^n(J) = \left\{ u \in PC^n(J) : u^{(n)} \in AC(x_i, x_{i+1}],\ i = 0, 1, ..., m \right\}$$

and for each $u \in PC_D^n(J)$ we set the norm

$$\|u\|_D = \|u\| + \left\|u'\right\| + \cdots + \left\|u^{(n)}\right\|$$

Moreover for $p \in L^1(J)$ we consider the usual norm

$$\|p\|_1 := \int_J |p(t)|\, dt.$$

Throughout this paper the following hypothesis will be assumed :

(H1) $f : [0,1] \times \mathbb{R}^3 \to \mathbb{R}$ is a L^1-Carathéodory function, that is, $f(x, \cdot, \cdot, \cdot)$ is a continuous function for a. e. $x \in J$;

$f(\cdot, y_0, y_1, y_2)$ is measurable for $(y_0, y_1, y_2) \in \mathbb{R}^3$; and for every $M > 0$ there is a real-valued function $\psi_M \in L^1([0,1])$ such that

$$|f(x, y_0, y_1, y_2)| \leq \psi_M(x), \text{ for a. e. } x \in [0,1]$$

and for every $(y_0, y_1, y_2) \in \mathbb{R}^3$ with $|y_i| \leq M$, for $i = 0, 1, 2$;

(H2) the real valued functions I_{ij}, for $i = 0, 1, 2$ and $j = 1, ..., m$ are nondecreasing in all variables.

Definition 2 *A function $u \in PC_D^2(J)$ is a solution of (1)-(3) if it satisfies (1) almost everywhere in $J \setminus D$, the periodic conditions (2) and the impulse conditions (3).*

Next Lemma will give the unique solution for a linear Cauchy problem..

Lemma 3 *et $p : [0,1] \times \mathbb{R} \to \mathbb{R}$ be a L^1-Carathéodory function. Then for each interval $(x_j, x_{j+1}]$, $j = 0, 1, ..., m$, and a_j, b_j, $c_j \in \mathbb{R}$, the initial value problem composed by the equation*

$$u'''(x) = p(x), \text{ for } x \in (x_j, x_{j+1}] \tag{5}$$

 ISBN: 978-84-617-8694-7

and the conditions

$$u\left(x_j^+\right) = a_j, \quad u'\left(x_j^+\right) = b_j, \quad u''\left(x_j^+\right) = c_j, \tag{6}$$

has a unique solution $u_j \in C^2(x_j, x_{j+1}]$, *given by*

$$u_j(x) = a_j + b_j\,(x - x_j) + c_j \frac{(x - x_j)^2}{2} + \int_{x_j}^{x} \frac{(x - r)^2}{2} p\,(r)\,dr. \tag{7}$$

Therefore, $u \in PC_D^2\,(J)$, *given by (4), is the unique solution of*

$$u'''\,(x) = p\,(x)\,, \text{ for a.e. } x \in [0, 1], \tag{8}$$

verifying (6), for each $j = 0, 1, ..., m$.

The solution $u(x)$ given by (7) can be obtained by iterate integrations of (5).
Strict lower and upper solutions are defined by the following inequalities:

Definition 4 *A function* $\alpha \in PC_D^3\,(J)$ *is said to be a strict lower solution of the problem (1)-(3) if:*

(i) $\alpha'''\,(x) < f\,(x, \alpha\,(x)\,, \alpha'\,(x)\,, \alpha''\,(x))$, *for a.e.* $x \in (0, 1)\,.$

(ii) $\alpha\,(0) \leq \alpha\,(1)\,, \quad \alpha'\,(0) \leq \alpha'\,(1)\,, \quad \alpha''\,(0) \leq \alpha''\,(1)\,,$

(iii) $\alpha\left(x_j^+\right) \leq I_{0j}\,(\alpha(x_j), \alpha'(x_j), \alpha''(x_j))\,,$
$\alpha'\left(x_j^+\right) \leq I_{1j}\,(\alpha\,(x_j)\,, \alpha'\,(x_j)\,, \alpha''\,(x_j))\,,$
$\alpha''\left(x_j^+\right) \leq I_{2j}\,(\alpha(x_j), \alpha'(x_j), \alpha''(x_j))$

A function $\beta \in PC_D^3\,(J)$ *is a strict upper solution of problem (1)-(3) if the reversed inequalities hold.*

 ISBN: 978-84-617-8694-7

3 Existence and localization result

The main theorem provides not only the existence of a solution, but also gives some qualitative data about its behavior:

Theorem 5 *Let $\alpha, \beta \in PC_D^2(J)$ be, respectively, lower and upper solutions of (1)-(3) such that*

$$\alpha''(x) \leq \beta''(x) \text{ on } J \setminus D \tag{9}$$

and

$$\alpha^{(i)}(0) \leq \beta^{(i)}(0), \ i = 0, 1. \tag{10}$$

Assume that

$$f\left(x, \alpha, \alpha', y_2\right) \leq f(x, y_0, y_1, y_2) \leq f\left(x, \beta, \beta', y_2\right), \tag{11}$$

for fixed $(x, y_2) \in J \times \mathbb{R}$, $\alpha^{(i)} \leq y_i \leq \beta^{(i)}$, for $i = 0, 1$.

If conditions (H1) and (H2) hold, then the problem (1)-(3) has a solution $u(x) \in PC_D^2(J)$, such that

$$\alpha^{(i)}(x) \leq u^{(i)}(x) \leq \beta^{(i)}(x), \text{ on } J, \text{ for } i = 0, 1, 2.$$

Remark 6 *As one can notice by (10), the inequalities $\alpha^{(i)}(x) \leq \beta^{(i)}(x)$ hold for $i = 0, 1, 2$, and every $x \in J$.*

Consider the following modified problem composed by the equation

$$u'''(x) = f\left(x, \delta_0(x, u(x)), \delta_1\left(x, u'(x)\right), \delta_2\left(x, u''(x)\right)\right), \tag{12}$$

for $x \in (0, 1)$ and $x \neq x_j$ where the continuous functions $\delta_i : \mathbb{R}^2 \to \mathbb{R}$, for $i = 0, 1, 2$ are given by

$$\delta_i(x, y_i) = \begin{cases} \beta^{(i)}(x) & , \quad y_i > \beta^{(i)}(x) \\ y_i & , \quad \alpha^{(i)}(x) \leq y_i \leq \beta^{(i)}(x) \\ \alpha^{(i)}(x) & , \quad y_i < \alpha^{(i)}(x) \end{cases} \tag{13}$$

with the boundary conditions (2) and the impulse conditions (3).

 ISBN: 978-84-617-8694-7

To prove the existence of solution for the problem (12),(2),(3) we apply an iterative technique, not necessarily monotone. Let $(u_n)_{n\in\mathbb{N}}$ be the sequence in $PC^2_D(J)$ defined as follows

$$u_0(x) = \alpha(x), \tag{14}$$

and for $n = 1, 2, ...$ the problem composed by the equation

$$u'''_n(x) = f\left(x, \delta_0\left(x, u_{n-1}(x)\right), \delta_1\left(x, u'_{n-1}(x)\right), \delta_2\left(x, u''_n(x)\right)\right), \tag{15}$$

for a.e. $x \in [0,1]$, with the boundary conditions

$$u_n(0) = u_{n-1}(1), \quad u'_n(0) = u'_{n-1}(1), \quad u''_n(0) = u''_{n-1}(1), \tag{16}$$

and the impulsive conditions, for $j = 1, ..., m$,

$$\begin{aligned} u_n\left(x_j^+\right) &= I_{0j}\left(u_{n-1}(x_j), u'_{n-1}(x_j), u''_{n-1}(x_j)\right), \\ u'_n\left(x_j^+\right) &= I_{1j}\left(\delta_0(x_j, u_{n-1}(x_j)), u'_{n-1}(x_j), u''_{n-1}(x_j)\right), \\ u''_n\left(x_j^+\right) &= I_{2j}\left(\delta_0(x_j, u_{n-1}(x_j)), \delta_1\left(x_j, u'_{n-1}(x_j)\right), u''_{n-1}(x_j)\right), \end{aligned} \tag{17}$$

Remark that the initial conditions (16) are a particular case of the initial conditions (6). In fact

$$\begin{aligned} u_n(0) &= u_n(x_0) = u_{n-1}(1) := a_0, \\ u'_n(0) &= u'_n(x_0) = u'_{n-1}(1) := b_0, \\ u''_n(0) &= u''_n(x_0) = u''_{n-1}(1) := c_0. \end{aligned}$$

Therefore, by Lemma 3, the sequence $(u_n)_{n\in\mathbb{N}}$, is well defined.

Remark that the initial value problem (15)-(17) will become the periodic impulsive problem (1)-(3), if the two following claims hold:

- *Every solution $u_n(x)$ of the problem (15)-(17) verifies*

$$\alpha^{(i)}(x) \le u_n^{(i)}(x) \le \beta^{(i)}(x), \text{ for } i = 0, 1, 2,$$

 ISBN: 978-84-617-8694-7

for all $n \in \mathbb{N}$ *and every* $x \in J$, *which implies that*

$$\delta_i\left(x, u_n^{(i)}(x)\right) = u_n^{(i)}(x), \text{ for } i = 0, 1, 2, \ n \in \mathbb{N} \text{ and every } x \in J,$$

and, consequently, (15) become

$$u_n'''(x) = f\left(x, u_{n-1}(x), u_{n-1}'(x), u_n''(x)\right), \text{ for a.e. } x \in [0,1].$$

- *There is a subsequence of* $(u_n)_{n\in\mathbb{N}}$, *denoted by simplicity as* (u_n), *uniformly convergent to* $u \in PC_D^2$, *solution of problem (1)-(3).*

4 Example

Consider the third order fully differential equation

$$u'''(x) = e^{u(x)} + \left(u'(x)\right)^3 - 280\sqrt[5]{u''(x)}, \tag{18}$$

with the periodic boundary conditions (2) and the impulsive functions

$$\begin{aligned} u\left(\frac{1}{2}^+\right) &= \frac{1}{10}u\left(\frac{1}{2}\right) + \frac{1}{100}\left[u'\left(\frac{1}{2}\right) + u''\left(\frac{1}{2}\right)\right], \\ u'\left(\frac{1}{2}^+\right) &= \frac{1}{10}\left[u\left(\frac{1}{2}\right) + u'\left(\frac{1}{2}\right) + u''\left(\frac{1}{2}\right)\right], \\ u''\left(\frac{1}{2}^+\right) &= \frac{1}{5}\left[u\left(\frac{1}{2}\right) + u'\left(\frac{1}{2}\right)\right] + \frac{1}{10}u''\left(\frac{1}{2}\right). \end{aligned} \tag{19}$$

This problem is a particular case of (1)-(3) with

$$f(x, y_0, y_1, y_2) = e^{y_0} + (y_1)^3 - 280\sqrt[5]{y_2},$$

for all $x \in [0,1] \setminus \left\{\frac{1}{2}\right\}$, $m = 1$, $x_1 = \frac{1}{2}$ and the nondecreasing functions I_{i1}, $i = 0, 1, 2$, are given by

$$\begin{aligned} I_{01}(a,b,c) &= \frac{1}{10}a + \frac{1}{100}(b+c), \\ I_{11}(a,b,c) &= \frac{1}{10}(a+b+c), \\ I_{21}(a,b,c) &= \frac{1}{5}(a+b) + \frac{1}{10}c. \end{aligned} \tag{20}$$

 ISBN: 978-84-617-8694-7

The piecewise continuous $\alpha, \beta \in PC_D^2(J)$, with $D = \left\{\frac{1}{2}\right\}$, defined as

$$\alpha(x) = \begin{cases} -x^2 - 2x - 2 & , \ x \in \left[0, \frac{1}{2}\right] \\ -x^2 - \frac{1}{4} & , \ x \in \left(\frac{1}{2}, 1\right] \end{cases}$$

and

$$\beta(x) = \begin{cases} x^2 + 4x + 3 & , \ x \in \left[0, \frac{1}{2}\right] \\ x^2 + \frac{x}{2} + \frac{1}{4} & , \ x \in \left(\frac{1}{2}, 1\right] \end{cases},$$

are lower and upper solutions, respectively, for problem (18), (2), (19), assuming

$$\alpha'(x) = \begin{cases} -2x - 2 & , \ x \in \left[0, \frac{1}{2}\right] \\ -2x & , \ x \in \left(\frac{1}{2}, 1\right] \end{cases}, \quad \beta'(x) = \begin{cases} 2x + 4 & , \ x \in \left[0, \frac{1}{2}\right] \\ 2x & , \ x \in \left(\frac{1}{2}, 1\right] \end{cases},$$

and $\alpha''(x) = -2$, $\beta''(x) = 2$, for $x \in [0, 1]$.

As f satisfies assumption (H_1) and (11), the jump functions (20) verify (H_2), then, by Theorem 5 there is a periodic solution $u(x) \in PC_D^2(J)$ of problem (18), (2), (19), such that

$$\alpha^{(i)}(x) \leq u^{(i)}(x) \leq \beta^{(i)}(x), \text{ for } i = 0, 1, 2. \tag{21}$$

Remark that this solution can not be a trivial periodic one, as constants do not verify (18).

References

[1] Z. Benbouziane, A. Boucherif, S. Bouguima, *Existence result for impulsive third order periodic boundary value problems*, Appl. Math. Comput. 206 (2008) 728–737.

[2] A. Cabada, *The method of lower and upper solutions for third-order periodic boundary value problems*, J. Math. Anal. Appl., 195 (1995) 568–589.

[3] A. Cabada, F. Minhós, A. Santos, *Solvability for a third order discontinuous fully equation with nonlinear functional boundary conditions,* J. Math. Anal. Appl. 322 (2006) 735–748

[4] J. Fialho, F. Minhós, *On higher order fully periodic boundary value problems*, J. Math. Anal. Appl., 395 (2012) 616-625.

 ISBN: 978-84-617-8694-7

[5] J. Fialho, F. Minhós, *High order Boundary Value Problems: Existence, localization and multiplicity results*, Mathematics Research Developments Series, Nova Science Publishers, Inc., New York, 2014.

[6] I. Rachůnková, M. Tvrdý, *Existence results for impulsive second-order periodic problems.* Nonlinear Anal. 59 (2004) 133–146

[7] J. Ren, S. Siegmund, Y. Chen, *Positive periodic solutions for third-order nonlinear differential equations*, Electronic Journal of Differential Equations, Vol. 2011 (2011), No. 66, pp. 1–19.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

Numerical Investigations of Synthetic Jet Actuators

A. Miró[1], M. Soria[1], I. Rodríguez[2] and J. C. Cajas[3]

[1] *Department of Physics, Aerospace Section, Universitat Politècnica de Catalunya*

[2] *Turbulence and Aerodynamics in Mechanical and Aerospace Engineering Research Group, Universitat Politècnica de Catalunya*

[3] *Barcelona Supercomputing Center (BSC), Spain*

emails: arnau.miro@upc.edu, manel.soria@upc.edu, ivette.rodriguez@upc.edu, juan.cajas@bsc.es

Abstract

Synthetic jet actuators (SJA) consist of a cavity with a mechanically moving diaphragm, whose actuation causes external fluid to enter and leave through a small orifice resulting in a net jet able to transfer of kinetic energy and momentum to a fluid medium without the addition of external flow. They are are expected to play a key role in active flow control (the application that motivates the present study), cooling and mixing.

This paper is focused on the interaction between the flow inside the actuator cavity and the external flow. Solving the coupling between the internal and external flows adds considerable complication and cost. It would be of interest to characterize the SJA outlet velocity and implement it as a boundary condition. To investigate the accuracy of this approach, an impinging jet configuration has been implemented with three different models: cavity and moving surface, cavity and imposed velocity and imposed velocity at the outlet. Time and phase averages of the external flows obtained with the different implementations are compared, and the effect of three different models is discussed.

Key words: AFC, synthetic jets, computational fluid dynamics, impingement

1 Introduction

Synthetic Jet (or Zero Net Mass Flow) actuators [1–6] consist of a cavity with a mechanically moving diaphragm (or piston). Its actuation changes the cavity volume periodically, causing external fluid to enter and leave through a small slot.

The expelled fluid creates a shear layer that rolls up forming vortices. Under certain conditions [7,8], when the cavity volume increases again they are too far to be ingested back, and therefore a train of vortices is created without the addition of mass flux. This property allows the transfer of kinetic energy and momentum to a fluid medium without the need of piping systems. This finds important applications in different areas such as active flow control (AFC) [9, 10], heat transfer enhancement [11, 12] and mixing enhancement [13, 14].

The flow patterns resulting from the interaction of the currents entering and leaving the cavity is very complex and have been the subject of many numerical and experimental studies. To mention just a few: [2] studied an incompressible laminar and turbulent Synthetic Jet using a URANS model (Spalart-Allmaras), assuming it to be two-dimensional. [17] performed a DNS three-dimensional periodic simulations of a synthetic jet previously analyzed with PIV [18]. [11] performed an experimental investigation on two-dimensional impinging synthetic jets and managed to successfully separate the effects of the Reynolds and Stokes numbers.

The present work is focused on the interaction between the external flow and the SJA internal flow. In the ingestion phase, vorticity, turbulent kinetic energy and passive scalars (such as temperature or concentration of chemical species) enter again the cavity where they interact with the internal flow before being expelled. These effects, together with the moving membrane and the large size difference between the actuator and the external flow, add considerable complication and cost to the simulation of the devices.

Different models have been used to model flows with a SJA. In decreasing order of computational cost, they are:

- Model M1: the internal SJA flow is simulated and the movement of the membrane is described in a realistic way [15, 21, 22].

- Model M2: the internal flow is simulated but the time derivative of the membrane position respect to the bottom of the cavity is used as a velocity boundary condition [17]. With this approach it is difficult to assign correct values for turbulent quantities or passive scalars at the membrane during the expulsion cycle.

- Model M3: The cavity flow is not simulated and a SJA outlet velocity is imposed [23, 24]. Probably this is most used approach active flow control applications.

The goal of this paper is to compare the different SJA models, to assess up to which point it is necessary to simulate the jet actuator cavity, and to analyze the effect of considering or not a realistic membrane movement in the simulation. To do so, the discharge and impingement of a SJA into a hot wall is studied.

 ISBN: 978-84-617-8694-7

1.1 Governing numbers and phenomenology

For laminar flows, the incompressible Navier-Stokes equations are used to model the SJA flow. In dimensionless form they are:

$$\frac{\partial u_j}{\partial x_j} = 0, \tag{1}$$

$$\mathrm{Sr}\frac{\partial u_i}{\partial t} + \frac{\partial (u_i u_j)}{\partial x_j} = -\frac{\partial p}{\partial x_i} + \frac{\partial}{\partial x_j}\left[\frac{1}{Re}\left(\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i}\right)\right], \tag{2}$$

$$\mathrm{Sr}\frac{\partial T}{\partial t} + \frac{\partial (u_j T)}{\partial x_j} = \frac{\partial}{\partial x_j}\left[\frac{1}{Re\,Pr}\frac{\partial T}{\partial x_j}\right]. \tag{3}$$

For the case of turbulent flows, even at the moderate Reynolds numbers, the computing cost associated with the direct integration of Navier-Stokes equations (DNS) [33] is too high, specially considering that (as will be seen) a large number of actuator cycles has to be solved to reach a stationary state. Instead, a URANS approach is used, where the turbulent time scales of the flow are filtered to obtain:

$$\frac{\partial \overline{u}_j}{\partial x_j} = 0, \tag{4}$$

$$\mathrm{Sr}\frac{\partial \overline{u}_i}{\partial t} + \frac{\partial (\overline{u}_i \overline{u}_j)}{\partial x_j} = -\frac{\partial \overline{p}}{\partial x_i} + \frac{\partial}{\partial x_j}\left[\frac{1}{Re}\left(\frac{\partial \overline{u}_i}{\partial x_j} + \frac{\partial \overline{u}_j}{\partial x_i}\right) - \overline{u_i' u_j'}\right], \tag{5}$$

$$\mathrm{Sr}\frac{\partial \overline{T}}{\partial t} + \frac{\partial (\overline{u}_j \overline{T})}{\partial x_j} = \frac{\partial}{\partial x_j}\left[\frac{1}{Re\,Pr}\frac{\partial \overline{T}}{\partial x_j} - \overline{u_j' T'}\right] \tag{6}$$

Where $\overline{\phi}(t) = \frac{1}{\Delta t}\int_t^{t+\Delta t} \phi dt$. The correlations $\overline{u_i' u_j'}$ and $\overline{u_i' T'}$ can not be obtained from the averaged flow and have to be modelled. In our case, the $k-\omega$ SST model has been used to do so. For brevity, the reader is referred to [27] for a description of the model and to [28] for the implementation details. In [20], numerical simulations of SJA with $k-\omega$ SST at $Re = 305$ and $Re = 508$ are show to be in good agreement with experimental results.

Incompressible regime is assumed as the ratio of the Helmholtz frequency of the actuator and the drive frequency is less than 0.5 [25, 26]. Richardson number has been assumed to be low enough as to disregard the buoyancy effect.

Prandtl number is $Pr = (\rho \nu c_p)/k$, a value of 0.71 has been assumed. Reynolds number is defined along the standard set by Smith and Glezer [6], $\mathrm{Re} = \frac{\rho U_0 d}{\mu}$, where d is the orifice diameter (or width, for rectangular actuators as in our case) and U_0 a characteristic velocity defined in terms of the stroke length L_0 as $U_0 = L_0 f$, with f being the drive frequency. The stroke length is defined as $L_0 = \frac{1}{d}\int_0^{\tau/2}\int_0^d u(x,0,t)dx\ dt$, where $u(x,0,t)$ is the instantaneous velocity at the orifice and τ is the period. The stroke length times the orifice surface is the

volume of fluid displaced during the expulsion part of the cycle. The Strouhal number Sr is $\mathrm{Sr} = \frac{2\pi f d}{U_0} = \frac{\mathrm{Sk}^2}{\mathrm{Re}}$. Alternatively [11, 17], the Stokes number $\mathrm{Sk} = \sqrt{\frac{\rho 2\pi f d^2}{\mu}}$ can be used to characterize the flow instead of the Strouhal number.

An important parameter that defines the jet behaviour is its formation criteria [7, 8],

$$\mathrm{JFC} = \frac{1}{\mathrm{Sr}_{\overline{U}}} = \frac{\mathrm{Re}_{\overline{U}}}{\mathrm{Sk}^2_{\overline{U}}} > K, \tag{7}$$

which is based on a time and space averaged velocity $\overline{U}$ during on the expulsion stroke ($\overline{U} = 2U_0$). If JFC is below K, the vortices expelled by the SJA are ingested back and the jet is not formed. For the two-dimensional rectangular jets considered in the present work, K can be approximated to 2 [7].

If a detailed description of the SJA membrane is included in the simulations, its position is modeled as $y(x,t) = -\delta(x)\cos(2\pi f t)$, where $\delta(x)$ is a diaphragm shape function. A mean amplitude can be defined as

$$\overline{A} = \frac{1}{W}\int_{-W/2}^{W/2} \delta(x)dx, \tag{8}$$

where W is the actuator cavity width. Under these circumstances, the characteristic velocity U_0 can be related to the mean amplitude and the drive frequency $U_0 = 2\overline{A}fW/d$, thus forbidding the decoupling between f and U_0. Also, under this definition, the jet formation criteria becomes

$$\mathrm{JFC} = \frac{2}{\pi}\left(\frac{\overline{A}}{d}\right)\left(\frac{W}{d}\right), \tag{9}$$

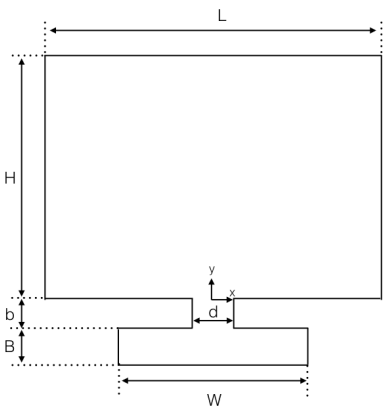
a pure geometrical parameter. High amplitudes and large jet cavity aspect ratios favour the formation of jets, independently of the frequency. Changing the value of the drive frequency enables increasing both Re and Sk in a way that the jet formation criteria remains the same. Therefore, changing the excitation frequency not only changes the jet period but also the velocity in which the jet is expelled.

2 Configuration studied

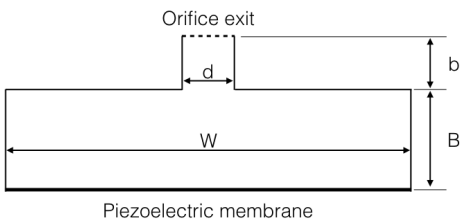
A rectangular actuator (Fig. 1) in a domain homogeneous in the z direction has been considered. The medium is assumed to be wide enough ($L = 60$) as to not interfere with the jet, but small H values of 5 and 10 have been imposed, in the range of the optimal distances for cooling applications [29]. The actuator cavity width W has been chosen to impose a jet formation criteria of $JFC = 3$ and a mean amplitude $\overline{A}/d = 0.2$ (equations 8 and 9). The other dimensions of the SJA are based on Liu [31], with $b/d = 0.3$ and $B/d = 1.67$. Two Reynolds numbers have been studied, $Re = 50$ and $Re = 500$.

 ISBN: 978-84-617-8694-7

Non-slip boundary conditions are imposed at the top ($y = H$) and bottom ($y = 0$) of the discharge cavity as well as at the actuator walls ($y < 0$ and $y = \pm d/2$), except for the active bottom wall later described. Free-flow boundary conditions have been prescribed at all the vertical boundaries with $y > 0$. Respect to the energy equation, the cavity top wall is hot ($T = 1$), the bottom wall cold ($T = 0$) and the lateral boundaries adiabatic ($dT/dt = 0$) where the flow leaves the domain and cold where it enters. The SJA walls ($y < 0$) are assumed to be at $T = 0$.



(a) Schematic view of the studied configuration (not to scale).



(b) Detail of the actuator (not to scale).

Figure 1: Computational domain and actuator.

The SJA has been implemented in the aforementioned three different ways:

- M1, with a moving membrane of position $y(x,t) = -\delta(x)\cos(2\pi ft)$.
- M2, imposing velocity boundary conditions at the SJA bottom as $v = 2\pi f\delta(x)\sin(2\pi ft)$. A cosine shape function has been used, $\delta(x) = \delta_C \cos\left(\pi \frac{x}{W}\right)$, where as δ_C is a scaling parameter that is set so $\overline{A}$ can have the desired value. Other membrane position functions can be found in the literature [4, 16, 21].
- M3, disregarding the SJA and imposing the velocity at the orifice exit as $u(x,0,t) = 2\pi f\overline{A}\sin(2\pi ft)W/d$ and $T = 0$.

2.1 Solution procedure and grid convergence

The configuration is solved using the finite volume method by means of the software *Code_Saturne*. The PISO algorithm is selected to solve the pressure-velocity coupling. No turbulence model is used to solve the low Reynolds numbers ($Re = 50$), while $k-\omega$ SST has

been used for $Re = 500$. Each simulation is allowed to run from quiescent initial conditions until a stationary state is reached. A large number of actuator cycles (≈ 200) is needed. The solution obtained is then transferred to a finer mesh and used as initial condition. This reduces the number of actuator cycles needed to reach again a stationary state to about ≈ 30. As an example of the grid convergence studies, Nusselt number distributions obtained with three different meshes for $Re = 50$ and $H = 10$ are shown in Fig. 2.

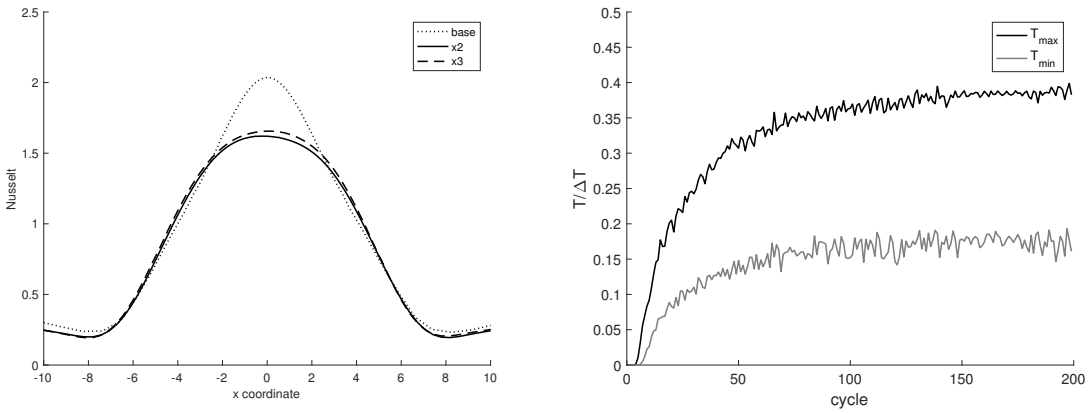


Figure 2: Left: Nusselt numbers obtained with the different meshes for Model 1, $Re = 50$ and $H = 10$, with 4348, 18424 and 41528 CV. Right: Maximum and minimum temperature at the actuator neck for $Re = 500$, as a function of the SJA cycle.

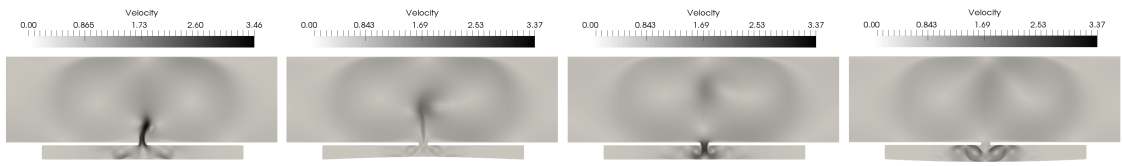
3 Results and Discussion

3.1 Phase and time averaged flow

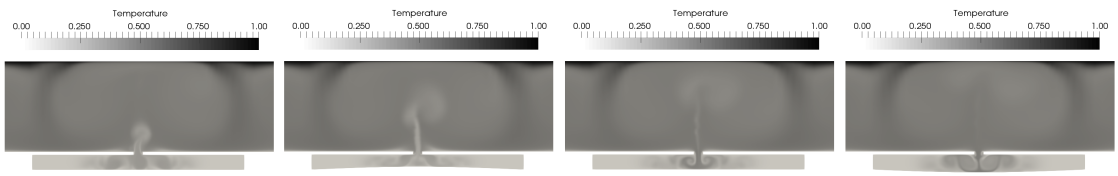
Fig. 3 shows a sequence of phase averaged velocity, temperature and turbulent kinetic energy maps that are representative of the jet sequence once the stationary state has been reached. Due to the motion of the membrane, during the forward stroke (second column), a pair of vortices are generated in the actuator lips and expelled to the cavity. If the jet formation criteria is high enough, when the suction stroke begins (third column) the vortices are too far to be ingested back and therefore generate a nonzero momentum flux in the streamwise direction. Also, another pair of vortices are created inside the cavity. During the ingestion (fourth column), k and T from the external flow are admitted in the SJA cavity, where they interact with the internal flow to be expelled back (first column). To illustrate the interaction between the internal and the external flows, the evolution of the maximum and minimum temperatures in the actuator neck is shown in Fig. 2 (right). The maximum

 ISBN: 978-84-617-8694-7

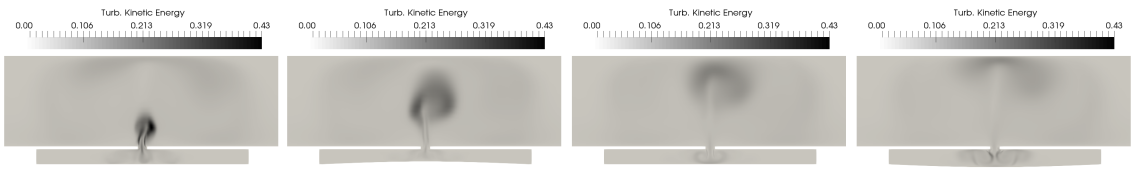
temperature corresponds to the ingestion phase. As can be seen, after a few cycles the fluid entering the SJA is quite hot and despite the cooling in the SJA cavity (where the walls are assumed to be at $T = 0$), when it exits at a temperature $T \approx 0.18$. This value, that has an effect on the cooling efficiency, depends on the interaction between both flows can not be imposed as a boundary condition without simulating the SJA chamber.



(a) Velocity (non-dimensionalized by U_0).



(b) Temperature (non-dimensionalized by ΔT).



(c) Turbulent kinetic energy (non-dimensionalized by U_0^2).

Figure 3: Sequences of velocity, temperature and turbulent kinetic energy fields during a representative cycle for $Re = 500$. From left to right $\phi = 0$, $\phi = 90$, $\phi = 180$ and $\phi = 270$.

If a time average is performed, instead of phase averages, the overall flow field can be obtained, as can be seen in Fig. 4. For $Re = 500$, $H = 10$, a series of large vortices emerge from the SJA strokes, that transport energy from the hot to the cold wall. At first glance, the large central jet would suggest that a fluid is entering the domain; however, close inspection of the streamlines reveals that the net mass flow is zero.

3.2 Effect of simulating the actuator cavity

The different models of the synthetic jet actuator have been compared by assessing the difference in the vertical velocities at different distances from the actuator outlet and the time averaged Nusselt number, for $Re = 50$ and $H/d = 10$. As can be seen in Fig. 5 (left), as the distance from the actuator increases, the velocity profile becomes more uniform and

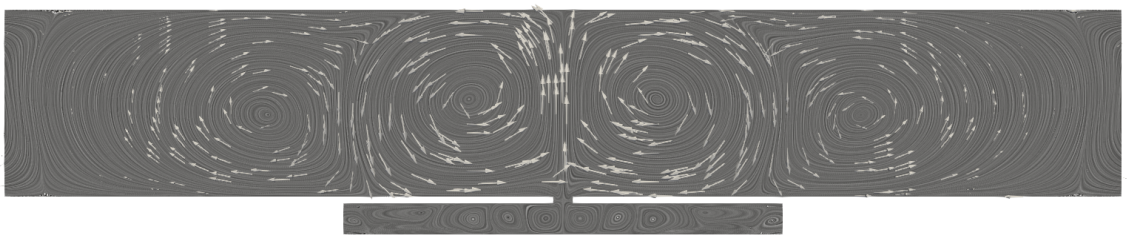


Figure 4: Streamlines for $Re = 500$, $H = 10$ and M1.

the difference between the models is smaller. However, at the hot wall ($H = 10$), there is a significant difference in the Nusselt number distribution (Fig. 5, right), specially between M3 (that doesn't solve for the cavity) and M1 and M2. To find out the reason of this difference in the Nusselt values, despite the similar velocity profiles, vertical temperature profiles at the maximum ingestion and expulsion phases have been represented in Fig. 6. Model M3, that only solves the domain for $y > 0$ (where it imposes an arbitrary value of $T = 0$), predicts a temperature too low near the hot wall, that overestimates the Nusselt number. This difference would be smaller if at least the correct mean value of the outlet temperature could be imposed in M3, but this value can not be obtained without solving first the interaction between the SJA cavity and the external flow.

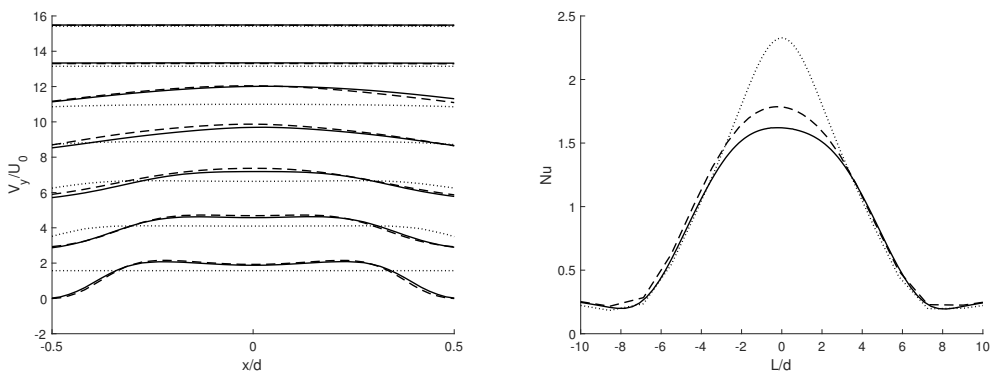


Figure 5: Comparison of the three implementations of the SJA for $Re = 50$, $H = 10$. Implementation 1: Solid line; 2: Dashed line; 3: Dotted line. Left: Vertical velocity profiles at $y = 0, 0.5, 1, 1.5, 2, 5, 8$ (each shifted 2.5 up units to avoid superposition with the previous). Right. Time averaged Nusselt numbers at the hot wall.

 ISBN: 978-84-617-8694-7

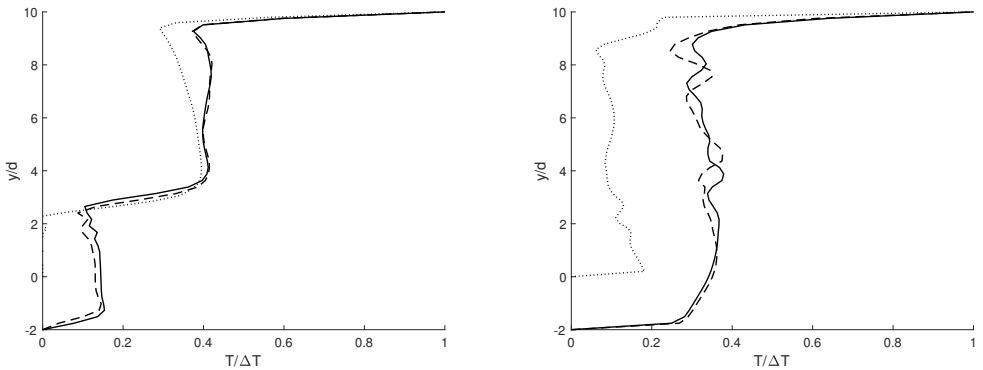


Figure 6: Comparison of the three implementations of the SJA for $Re = 50$, $H = 10$. Implementation 1: Solid line; 2: Dashed line; 3: Dotted line. Left: Phase averaged vertical temperature profiles at the maximum expulsion. Right: Phase averaged vertical temperature profiles at the maximum ingestion.

4 Conclusions

A SJA impinging into a hot wall has been studied, considering three different models: Moving membrane (M1), velocity boundary conditions and cavity (M2) and boundary conditions at the exit orifice (M3). Domains of two different vertical sizes ($H = 5$ and $H = 10$), and two different Reynolds numbers ($Re = 50$ and $Re = 500$) have been studied.

The resulting flows are quite complex, and a large number of actuator cycles has to be solved in order to reach a stationary state. All the implementations can solve the basic features of the external flow created by the SJA. Moreover, as the distance from the actuator outlet is increased, the velocity profiles predicted by the three implementations become very similar.

However, when the Nusselt number and the temperature profiles predicted by the different models are analyzed, there are significant differences. The reason seems to be related with the value of the temperature that has to be imposed at the SJA outlet in M3. Moreover, it is quite possible that similar effects are found in other passive scalars such as concentration in mixing applications or turbulent kinetic energy. While it remains to be studied if such differences are actually important for active flow control applications, M2 seems a good compromise between accuracy and cost. Although many of the features of the internal SJA flow can not be well predicted due to the simplified boundary conditions imposed at the membrane in M2, the errors introduced by these assumptions will decrease in the external flow.

Acknowledgements

This work has been partially supported by Red Española de Supercomputación (RES) under the project number FI-2016-3-0014 and the Spanish Ministry project (MEC) FIS2016-77849-R.

References

[1] A. Glezer and M. Amitay, *Synthetic Jets*, Ann. Rev. Fluid Mechanics **503** (2002).

[2] L. D. Kral, J. F. Donovan, A. B. Cain and A. W. Cary, *Numerical simulation of synthetic Jet Actuators*, AIAA Journal (1997).

[3] S. G. Mallinson, G. Hong and J. A. Reizes, *Some characteristics of synthetic jets*, AIAA Paper (1999).

[4] S. G. Mallinson, C. Y. Kwok and J. A. Reizes, *Numerical simulation of micro-fabricated zero mass-flux jet actuators*, Sensors and Actuators, A: Physical **3** (2003) 229–236.

[5] C. L. Rumsey, T. B. Gatski, W. L. Sellers III, V. N. Vasta and S. A. Viken, *Summary of the 2004 Computational Fluid Dynamics Validation Workshop on Synthetic Jets*, AIAA Journal **2** (2006) 194–207.

[6] B. L. Smith and A. Glezer, *The formation and evolution of synthetic jets*, Phys. Fluids **9** (1998).

[7] R. Holman, Y. Utturkar, R. Mittal, B. L. Smith, and L. N. Cattafesta, *Formation Criterion for Synthetic Jets*, AIAA Journal **10** (2005) 2110–2116.

[8] Y. Utturkar, R. Holman, R. Mittal, B. Carroll, M. Sheplak and L. N. Cattafesta, *A Jet Formation Criterion for Synthetic Jet Actuators*, 41st Aerospace Sciences Meeting & Exhibit (2003).

[9] Sheplak, M. and Cattafesta, L. N., *Actuators for Active Flow Control*, Ann. Rev. of Fluid Mechanics **1** (2011) 247–272.

[10] A. Glezer, *Some aspects of aerodynamic flow control using synthetic-jet actuation*, P. Trans. of the Royal Society A: Mathematical, Physical and Engineering Sciences **1940** (2011) 1476–1494.

[11] L. Silva-Llanca, A. Ortega, and I. Rose, *Experimental convective heat transfer in a geometrically large two-dimensional impinging synthetic jet*, Intl. J. of Thermal Sciences **90** (2015) 339–350.

 ISBN: 978-84-617-8694-7

[12] O. Ghaffari, S. A. Solovitz and M. Arik, *An investigation into flow and heat transfer for a slot impinging synthetic jet*, Intl. J. of Heat and Mass Transfer **100** (2016) 634–645.

[13] H. Wang and S. Menon, *Fuel-Air Mixing Enhancement by Synthetic Microjets*, AIAA Journal **12** (2001) 2308–2319.

[14] Y. Chen, S. Liang, K. Anug, A. Glezer and J. Jagoda, *Enhanced Mixing in a Simulated Combustor Using Synthetic Jet Actuators*, 37th AIAA Aerospace Sciences Meeting and Exhibit **17** (1999).

[15] C. Y. Lee and D. B. Goldstein, *Two Dimensional Synthetic Jet Simulation*, AIAA Journal **3** (2002) 510–516.

[16] P. Mane, K. Mossi, A. Rostami, R. Bryant and N. Castro, *Piezoelectric Actuators as Synthetic Jets: Cavity Dimension Effects*, Journal of Intelligent Material Systems and Structures **11** (2007) 1175–1190.

[17] R. B. Kotapati, R. Mittal and L. N. Cattafesta, *Numerical study of a transitional synthetic jet in quiescent external flow*, J. of Fluid Mechanics **581** (2007) 287–321.

[18] C. Yao and F. J. Chent, *Synthetic Jet Flow Field Database for CFD Validation*, 2nd AIAA Flow Control Conference (2004) 1–11.

[19] Q. Xia and S. Zhong, *An experimental study on the behaviours of circular synthetic jets at low Reynolds numbers*, J. of Mechanical Engineering Science **11** (2012) 2686–2700.

[20] L. Silva-Llanca and A. Ortega, *Vortex dynamics and mechanisms of heat transfer enhancement in synthetic jet impingement*, Intl. J. of Thermal Sciences **112** (2017) 153–164.

[21] H. Xia and N. Qin, *Dynamic Grid and Unsteady Boundary Conditions for Synthetic Jet Flow*, 43rd AIAA Aerospace Sciences Meeting and Exhibit (2005) 1–9.

[22] D. P. Rizzetta, M. R. Visbal and M. J. Stanek, *Numerical Investigation of Synthetic-Jet Flowfields*, AIAA Journal **8** (1999).

[23] W. Zhang and R. Samtaney, *A direct numerical simulation investigation of the synthetic jet frequency effects on separation control of low-Re flow past an airfoil*, Phys. Fluids **27** (2015) 1–22.

 ISBN: 978-84-617-8694-7

[24] J. P. D'Alencon and L. Silva-Llanca, *Two-dimensional numerical analysis of a low-re turbulent impinging synthetic jet*, Proceedings of the 15th InterSociety Conference on Thermal and Thermomechanical Phenomena in Electronic Systems (2016) 921–929.

[25] Q. Gallas, *On the Modeling and Design of Zero-Net Mass Flux Actuators*, University of Florida, 2005.

[26] M. A. Feero, P. Lavoie and P. E. Sullivan, *Influence of cavity shape on synthetic jet performance*, Sensors and Actuators, A: Physical **223** (2015) 1–10.

[27] F. R. Menter, *Two-equation eddy-viscosity turbulence models for engineering applications*, AIAA journal **8** (1994) 1598–1605.

[28] A. Keshmiri, J. Uribe and N. Shokri, *Benchmarking of Three Different CFD Codes in Simulating Natural, Forced, and Mixed Convection Flows*, Numerical Heat Transfer, Part A: Applications, **67** (2015) 1324–1351.

[29] A. Pavlova and M. Amitay, *Electronic Cooling Using Synthetic Jet Impingement*, J. Heat Transfer **9** (2006).

[30] Valiorgue, P. and Persoons, T. and McGuinn, A. and Murray, D. B., *Heat transfer mechanisms in an impinging synthetic jet for a small jet-to-surface spacing*, Exp. Thermal and Fluid Science **4** (2009) 597–603.

[31] Y. H. Liu, S. Y. Tsai and C. C. Wang, *Effect of driven frequency on flow and heat transfer of an impinging synthetic air jet*, Applied Thermal Engineering **75** (2015) 289–297.

[32] Hunt, J. C. R. and Wray, A. A. and Moin, P., *Eddies, Streams, and Convergence Zones in Turbulent Flows*, Proceedings of the Summer Program, 1988.

[33] Soria M., et al., *Direct numerical simulation of a three-dimensional natural-convection flow in a differentially heated cavity of aspect ratio 4*, Numerical Heat Transfer, Part A Applications **45** (2004) 649–673.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

Optimal approximate solution for optimization problems via best proximity point theorem and variational principle in generalized distance functions

Chirasak Mongkolkeha[1]

[1] *Department of Mathematics, Statistics, and Computer Science, Faculty of Liberal Arts and Science, Kasetsart University, Kamphaeng-Saen Campus, Nakhonpathom 73140, Thailand*

emails: faascsm@ku.ac.th

Abstract

The aim of this talk is to approximate solution for solving minimization problems via best proximity point theorems and variational principle in generalized distance functions by giving an algorithm for determining such an is furnished. Also, we give a necessary and sufficient conditions for finding the existence of best proximity point and minimal elements with illustrative example of our main results.

Key words: Best proximity point, Generalized distance functions, Minimization problems, Minimal elements.
MSC 2000: 54E40, 7H10, 58E30.

1 Introduction

In 1989, Bakhtin [2] (see also Czerwik [3]) introduced the concept of a b-metric space and proved some fixed point theorems for some contractive mappings in b-metric spaces which are generalizations of Banach's contraction principle in metric spaces.

Definition 1.1 *Let X be a nonempty set and $s \geq 1$ be a given real number. A functional $d : X \times X \to [0, \infty)$ is called a b-metric if, for all $x, y, z \in X$, the following conditions are satisfied:*

1. *$d(x, y) = 0$ if and only if $x = y$;*

 ISBN: 978-84-617-8694-7

2. $d(x,y) = d(y,x)$;

3. $d(x,z) \leq s[d(x,y) + d(y,z)]$.

A pair (X,d) is called a *b-metric space* with coefficient s. Since, every metric space is a b-metric space with $s = 1$ and hence the class of b-metric spaces is larger than the class of metric spaces. In 1996, Kada et al. [5] introduced some generalized metric which is difference from b-metric space and called w-distance as follow:

Definition 1.2 *Let (X,d) be a metric space. A function $p : X \times X \longrightarrow [0,\infty)$ is said to be the w-distance on X if the following are satisfied:*

1. *$p(x,z) \leq p(x,y) + p(y,z)$ for all $x,y,z \in X$;*

2. *for any $x \in X$, $p(x,\cdot) : X \longrightarrow [0,\infty)$ is lower semi-continuous (i.e., if $x \in X$ and $y_n \longrightarrow y \in X$, then $p(x,y) \leq \liminf_{n\longrightarrow\infty} p(x,y_n)$;*

3. *for any $\varepsilon > 0$, there exists $\delta > 0$ such that $p(z,x) \leq \delta$ and $p(z,y) \leq \delta$ imply $d(x,y) \leq \varepsilon$.*

They also improved Caristi's fixed point theorem, Ekeland's variational principle and the nonconvex minimization theorem of Takahashi [6]. Later, Shioji et al. [7] studied the relationship between weakly contractive mappings and weakly Kannan mappings under the conditions, the w-distance and the symmetric w-distance. In 2012, Imdad and Rouzkard [8] proved some fixed point theorems in a complete metric space equipped with a partial ordering via the w-distance.

Later, In 2014, Hussain et al. [4] introduced the concept of the wt-distance in generalized b-metric spaces, which is a generalization of the w-distance.

Definition 1.3 *Let (X,d) be a b-metric space with constant $s \geq 1$. A function $P : X \times X \longrightarrow [0,\infty)$ is called the wt-distance on X if the following are satisfied:*

1. *$P(x,z) \leq s(P(x,y) + P(y,z))$ for all $x,y,z \in X$;*

2. *for any $x \in X$, $P(x,\cdot) : X \longrightarrow [0,\infty)$ is s-lower semi-continuous (i.e., if $x \in X$ and $y_n \longrightarrow y \in X$, then $P(x,y) \leq \liminf_{n\longrightarrow\infty} sP(x,y_n)$;*

3. *for any $\varepsilon > 0$, there exists $\delta > 0$ such that $P(z,x) \leq \delta$ and $P(z,y) \leq \delta$ imply $d(x,y) \leq \varepsilon$.*

Recently, Abdou et al. [1] proved some common fixed point theorems in Menger probabilistic metric type spaces by using the wt-distance.

In this talk, we prove the existence of best proximity point of some nonlinear mappings in complete metric spaces via the w-distance and also give some examples to for support our

 ISBN: 978-84-617-8694-7

results. Furthermore, we prove some Ekeland's variational principle in *wt*-distance and also, we give a necessary and sufficient conditions with an approximate algorithms for finding the existence of minimal elements to establish enhanced Ekeland's variational principle. Our result improve, extend and generalize several results given by some authors in literatures.

Acknowledgements

This work has been supported by the Kasetsart University Research and Development Institute (KURDI) and Department of Mathematics Statistic and Computer, Faculty of Liberal Arts and Science, Kasetsart University, Thailand.

References

[1] Afrah A.N. Abdou, Y.J. Cho and R. Saadati, *Distance type and common fixed point theorems in Menger probabilistic metric type spaces*, Appl. Math. Comput. **265** (2015), 1145–1154.

[2] A. Bakhtin, *The contraction mapping principle in quasimetric spaces*, Funct. Anal. Unianowsk Gos. Ped. Inst. **30** (1989), 26–37.

[3] S. Czerwik, *Contraction mappings in b-metric spaces*, Acta Math. Inform. Univ. Ostrav. **1** (1993), 5–11.

[4] N. Hussain, R. Saadati and R. P Agrawal, *On the topology and wt-distance onmetric type spaces*, Fixed Point Theory Appl. (2014), 2014:88.

[5] O. Kada, T. Suzuki and W. Takahashi, *Nonconvex minimization theorems and fixed point theorems in complete metric spaces*, Math. Japon. **44** (1996), 381–391.

[6] W. Takahashi, *Existence theorems generalizing fixed point theorems for multivalued mappings, in Fixed Point Theory and Applications*, Marseille, 1989, Pitman Res. Notes Math. Ser. 252: Longman Sci. Tech., Harlow, (1991), 39–406.

[7] N. Shioji, T. Suzuki and W. Takahashi , *Contractive mappings, Kanan mapping and metric completeness*, Proc. Amer. Math. Soc. **126** (1998), 3117–3124.

[8] M. Imdad and F. Rouzkard, *Fixed point theorems in ordered metric spaces via w-distances*, Fixed Point Theory Appl. (2012), 2012:222.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

Recent Convex Tools for Nonlinear Programming

P. Montiel López[1] and M. Ruiz Galán[2]

[1] *Centro de Estudios Superiores La Inmaculada, Department of Sciences, University of Granada, Spain*

[2] *Department of Applied Mathematics, University of Granada, Spain*

emails: pablomontiel@eulainmaculada.com, mruizg@ugr.es

Abstract

In this work, we present some results on Nonlinear Programming that generalize the classical theorems of the Lagrange multipliers, Kuhn–Tucker and Fritz John, and, moreover, they are sharp, in the sense that the validity of each of them is equivalent to the fact that a certain family of functions associated to the nonlinear problem under consideration satisfies a weak convexity condition. These results are derived from adequate versions of the Hahn–Banach theorem.

Key words: Hahn–Banach theorem, Nonlinear Optimization.
MSC 2000: 90C30, 46A22.

1 Hahn–Banach, Mazur–Orlicz and Nonlinear Programming

The development of Convex Analysis has always been together with that of Optimization, benefiting reciprocally. A remarkable example is the Hahn–Banach theorem and its geometric reformulations –the seminal versions of the theorems of the alternative such as Gordan's theorem or Farkas' lemma, the convex separation theorem– which have definitely become some of the most important tools in the study of problems on Linear and Nonlinear Programming.

First, we make use of a generalization of the Hahn–Banach theorem, known as the Mazur–Orlicz theorem, [8, 9, 14, 15, 17], which states that if C is a convex subset of a real vector space in which a sublinear (subadditive and positively homogeneous) functional is defined, then it is possible to find a linear functional less than it and in such a way that their infimum on C coincide. As a consequence, we derive some existence results along the

 ISBN: 978-84-617-8694-7

lines of the theorems of König of the maximum and the supremum [6, 7], which establish the concept of optimal convexity for its validity: the *infsup-convexity*, a weak concept of convexity that arose in minimax theory ([5, 13, 16]). Specifically, the more general result states that given two nonempty sets X and Λ, $f : X \longrightarrow \mathbb{R}$ and $(f_\lambda)_{\lambda\in\Lambda}$ a family of real valued functions defined on X such that

$$x \in X \;\Rightarrow\; (f_\lambda(x))_{\lambda\in\Lambda} \in \ell^\infty(\Lambda),$$

then, the family $(f_\lambda - f)_{\lambda\in\Lambda}$ is infsup-convex on X if, and only if, for each $\alpha \in \mathbb{R}$ with

$$x \in X \;\Rightarrow\; f(x) + \alpha \leq \sup_{\lambda\in\Lambda} f_\lambda(x),$$

there exists a bounded and linear functional $\Phi : \ell^\infty(\Lambda) \longrightarrow \mathbb{R}$ such that

$$\Phi \leq \sup_{\Lambda}$$

and

$$x \in X \;\Rightarrow\; f(x) + \alpha \leq \Phi((f_\lambda(x))_{\lambda\in\Lambda}).$$

When applying the König-type results we establish general versions of the theorems of the Lagrange multipliers, Fritz John and Kuhn–Tucker for an optimization problem with an arbitrary number of constraints,

$$\inf_{x\in F} f(x),$$

with

$$F := \left\{ x \in X : \; \sup_{\lambda\in\Lambda} f_\lambda(x) \leq 0, \; \sup_{\omega\in\Omega} |g_\omega(x)| = 0 \right\},$$

X being a nonempty set, Λ and Ω sets, $f : X \longrightarrow \mathbb{R}$ a given funtion and $(f_\lambda)_{\lambda\in\Lambda}$ and $(g_\omega)_{\omega\in\Omega}$ families of real valued functions on X. Such results not only extend some known statements [1, 2, 3, 4, 18], but also are sharp in terms of the infsup-convexity of certain families of functions. Some particular cases have been established in [10, 11, 12].

Acknowledgement

Research partially supported by project MTM2016-80676-P (AEI/FEDER, UE) and by Junta de Andalucía Grant FQM359.

References

[1] O. Brezhnevaa and A.A. Tret'yakov, *An elementary proof of the Karush–Kuhn–Tucker theorem in normed linear spaces for problems with a finite number of inequality constraints*, Optimization **60** (2011), 613–618.

[2] D. Fang, X. Luo and X. Wang, *Strong and total Lagrange dualities for quasiconvex programming*, J. Appl. Math. **2014** (2014), Article ID 453912.

[3] F. Flores-Bazán, *Fritz John necessary optimality conditions of the alternative-type*, J. Optim. Theory Appl. **161** (2014), 807–818.

[4] K. Ito and K. Kunisch, *Karush–Kuhn–Tucker conditions for nonsmooth mathematical programming problems in function spaces*, SIAM J. Control Optim. **49** (2011), 2133–2154.

[5] G. Kassay and J. Kolumbán, *On a generalized sup-inf problem*, J. Optim. Theory Appl., **91** (1996), 651–670.

[6] H. König, *Sublinear functionals and conical measures*, Arch. Math. **77** (2001), 56–64.

[7] H. König, *Sublineare funktionale*, Arch. Math. **23** (1972), 500–508.

[8] F.-C. Liu, *Mazur–Orlicz equality*, Studia Math. **189** (2008), 53–63.

[9] S. Mazur and W. Orlicz, *Sur les espaces métriques linéaires II* Studia Math. **13** (1953), 137–179.

[10] P. Montiel López and M. Ruiz Galán, *Nonlinear programming via König's maximum theorem*, J. Optim. Theory Appl. **170** (2016), 838–852.

[11] P. Montiel López and M. Ruiz Galán, *Revisiting the Hahn–Banach theorem and nonlinear infinite programming*, preprint.

[12] M. Ruiz Galán, *A sharp Lagrange multiplier theorem for nonlinear programs*, J. Global Optim. **65** (2016), 513–530.

[13] M. Ruiz Galán, *An intrinsic notion of convexity for minimax*, J. Convex Anal. **21** (2014), 1105–1139.

[14] S. Simons, *Bootstrapping the Mazur–Orlicz–König theorem and the Hahn-Banach Lagrange theorem*, J. Convex Anal. **25** (2018), to appear.

[15] S. Simons, *The Hahn–Banach–Lagrange theorem*, Optimization **56** (2007), 149–169.

[16] A. Stefanescu, *A theorem of the alternative and a two-function minimax theorem*, J. Appl. Math. **2004** (2004), 167–177.

[17] C. Sun, *The Mazur–Orlicz theorem for convex functionals*, J. Convex Anal. **24** (2017), to appear.

[18] S. Suzuki and D. Kuroiwa, *Optimality conditions and the basic constraint qualification for quasiconvex programming*, Nonlinear Anal. **74** (2011), 1279–1285.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

Quaternionic Mathieu functions for the heat-conduction equation in elliptical confocal coordinates

J. Morais[1] and K.I. Kou[2]

[1] *Department of Mathematics, ITAM, Rio Hondo #1, Col. Progreso Tizapán, México, DF 01080, México.*

[2] *Department of Mathematics, Faculty of Science and Technology, University of Macau, Macau.*

emails: joao.morais@itam.mx, kikou@umac.mo

Abstract

In this work, we show that there exists a theory of functions with quaternionic values and of two real variables, which is determined by a Cauchy-Riemann-type operator with quaternionic variable coefficients and that is intimately related to the well-known Mathieu functions. As a result, we introduce the *Quaternionic Mathieu Functions* and explain their connections to the solutions of the Heat-Conduction equation in the elliptical confocal coordinate system.

Key words: Quaternionic analysis, Heat-Conduction equation, elliptical confocal coordinates, Mathieu functions, hyperholomorphic functions.

1 Introduction

The majority of functions used in technical and applied mathematics were originated as the result of investigating practical problems. A relevant example is the *Mathieu functions*, which are solutions to the ordinary differential equations arising in the separation of the Helmholtz equation in two dimensions using *elliptical confocal coordinates*. These functions were introduced by Émile Léonard Mathieu in 1868 in his "*Mémoire sur le mouvement vibratoire d'une membrane de forme elliptique*" in connection with the study of the vibrational modes of a stretched membrane with an elliptic boundary [3]. Since that time Mathieu functions have been widely used in many areas of classical and modern physics, engineering and applied mathematics.

 ISBN: 978-84-617-8694-7

Mathieu functions occur in two main categories of physical problems: on one hand, in applications involving elliptical geometries, for example in the analysis of the vibrating modes in elliptical membranes, the propagating modes in elliptical pipes and the oscillations of water in a lake of elliptical shape. On the other hand, Mathieu functions may arise in problems involving periodic motion, such as the trajectory of an electron in a periodic array of atoms, the mechanics of the quantum pendulum, and the oscillations of floating vessels. Recently, in the works of Sato [8, 9] it is mentioned the importance of computing the thermal stresses resulting from the considerable temperature differences in solids bounded by ellipses. In particular, in [8] it is shown that the resulting solution of the heat conduction problem of a cylinder can be obtained in the form of an infinite Mathieu function series. Those papers deal with the heat conduction problem of an infinite and a confocal elliptical cylinder in which the effect of its surface resistance is included. In the above analysis the condition is that the cylinder surface is keeping heated or cooled at a constant temperature. The method of separation of variables is then applied to solving the *Heat-Conduction equation* using the elliptical cylinder coordinate system and the corresponding boundary conditions are accordingly satisfied by using the orthogonality of the Mathieu functions. Despite the existence of many applications and classical properties such as identities, recursions and asymptotics, Mathieu functions are barely mentioned in modern textbooks, and even older texts that gave some account of Mathieu functions are now out-of-print. One reason for the absence of theory of Mathieu functions in the existing literature compared to that for other special functions is because the behavior of Mathieu functions is relatively rich and, consequently, more difficult to understand. Another reason is probably the complicated and various notations that exist in the literature.

A higher-dimensional extension of the Mathieu functions to the framework of *Quaternionic Analysis* was firstly considered in [2]. Analogues of the basic integral formulae of complex analysis for this version of quaternionic function theory were established, which turned out to be in the same relation with the Schrödinger operator with a special potential as the usual holomorphic functions in one complex variable, or *quaternionic hyperholomorphic functions*, are with the corresponding Laplace operator. A major part of the present work is concerned with exploring what analytical properties of the Quaternionic Mathieu Functions (QMFs) arise from this extension. According to our current knowledge, central questions regarding a proper definition of the QMFs and general properties such as orthogonality relations and identities, and their connections to the solutions of the Heat-Conduction equation in the elliptical coordinate system remain untouched so far. We honestly think that the QMFs introduced in this paper give, through the factorization of the Heat-Conduction operator, a finer structure of the classical Mathieu functions. This understanding may be fruitful both to enrich the understanding of the properties of these functions in higher dimensions and as tools in such applications of quaternionic analysis in mathematical physics and related fields.

 ISBN: 978-84-617-8694-7

2 Heat-Conduction Equation in Elliptical Confocal Coordinates

As the present problem considers no temperature change along the lenght of an infinite hollow confocal elliptical cylinder, the heat-conduction equation within can be written in rectangular coordinates (x, y) as

$$\left(\Delta - \frac{1}{\kappa}\frac{\partial}{\partial t}\right)[\theta] := \frac{\partial^2\theta}{\partial x^2} + \frac{\partial^2\theta}{\partial y^2} - \frac{1}{\kappa}\frac{\partial\theta}{\partial t} = 0, \tag{1}$$

where $\theta(x, y, t)$ is the *temperature*, t the *time* and κ the *thermal diffusivity* given by $\kappa = K/(\rho\, C)$ with *thermal conductivity* K, *density* ρ and *specific heat* C, which are assumed to be constant. The problem of heat conduction under consideration can be rigorously analyzed by introducing the *elliptical confocal coordinates* (ξ, η), which are related to the rectangular coordinates by the relation

$$x + iy = c\cosh(\xi + i\eta) = c\cosh\xi\cos\eta + ic\sinh\xi\sin\eta,$$

where c denotes the *semi-focal length*, ξ is the *radial coordinate* which varies from zero along the line of foci to ξ_0 at the rings, and η is the *angular coordinate* which varies from 0 to 2π in passing once round an ellipse. We note that $c^2 = a^2 - b^2$ in which $a = 2c\cosh\xi_0$ and $b = 2c\sinh\xi_0$ represent the *semi-major* and *semi-minor axis length* of a family of confocal ellipses within the hollow cylinder bounded by the outer surface $\xi = \xi_0$:

$$\frac{x^2}{c^2\cosh^2\xi_0} + \frac{y^2}{c^2\sinh^2\xi_0} = 1. \tag{2}$$

We note first that the *heat-conduction operator*

$$\Delta - \frac{1}{\kappa}\frac{\partial}{\partial t}$$

acts on the space of functions $\mathcal{C}^2(\mathbb{R}^2) \times \mathcal{C}^1(\mathbb{R}_0^+)$. In the present study, we consider its restriction onto $\mathcal{C}^2(\Omega_{x,y}) \times \mathcal{C}^1(\mathbb{R}_0^+)$, where

$$\Omega_{x,y} := \mathbb{R}^2 \setminus \{(x, y) \in \mathbb{R}^2 \,|\, x = 0 \vee y = 0\}.$$

Henceforth we consider a domain Ξ in a copy of $\mathbb{R}^2$ with the coordinates (ξ, η). Now, we define a change of variables in the domain Ξ, i.e., there exists a mapping

$$\varphi : (\xi, \eta) \in \Xi \mapsto \varphi(\xi, \eta) = (x = \varphi_1(\xi, \eta), y = \varphi_2(\xi, \eta)) \in \Omega_{x,y},$$

such that $\varphi \in \mathcal{C}^2(\Xi)$ makes a one-to-one correspondence between both domains. Now, assume that $\psi = (\psi_1(x, y), \psi_2(x, y))$ is the inverse mapping, $\psi : \Omega_{x,y} \mapsto \Xi$, i.e., so that $\varphi(\psi(x, y)) = (x, y)$ for any $(x, y) \in \Omega_{x,y}$ and $\psi(\varphi(\xi, \eta)) = (\xi, \eta)$ for any $(\xi, \eta) \in \Xi$.

 ISBN: 978-84-617-8694-7

We introduce the operators of the change of variables:

$$W_\varphi : u \in \mathcal{C}^2(\Omega_{x,y}) \mapsto u \circ \varphi =: \tilde{u} \in \mathcal{C}^2(\Xi),$$
$$W_\psi = W_\varphi^{-1} : \tilde{u} \in \mathcal{C}^2(\Xi) \mapsto \tilde{u} \circ \psi =: u \in \mathcal{C}^2(\Omega_{x,y}).$$

It is of interest to point that W_φ is an isomorphism of $\mathcal{C}^2(\Omega_{x,y})$ onto $\mathcal{C}^2(\Xi)$, whereas W_ψ is an isomorphism of $\mathcal{C}^2(\Xi)$ onto $\mathcal{C}^2(\Omega_{x,y})$. Next, let A be an arbitrary linear operator acting on $\mathcal{C}^2(\Omega_{x,y})$ and B be an arbitrary operator acting on $\mathcal{C}^2(\Xi)$. We define the operators $\widetilde{A}$ and $\widetilde{B}$ as follows:

$$W_\varphi A W_\psi =: \widetilde{A} \quad \text{and} \quad W_\psi B W_\varphi =: \widetilde{B}.$$

Obviously, $\widetilde{A}$ acts on $\mathcal{C}^2(\Xi)$ while $\widetilde{B}$ acts on $\mathcal{C}^2(\Omega_{x,y})$.

Let $\mathcal{L}(\mathcal{C}^2(\Omega_{x,y}))$ and $\mathcal{L}(\mathcal{C}^2(\Xi))$ denote the algebras of all linear operators acting on the respective function spaces. Hence the mapping

$$A \in \mathcal{L}(\mathcal{C}^2(\Omega_{x,y})) \mapsto W_\varphi A W_\psi = \widetilde{A} \in \mathcal{L}(\mathcal{C}^2(\Xi))$$

is an isomorphism of algebras.

Taking $A = \Delta - \frac{1}{\kappa}\frac{\partial}{\partial t}$, a direct observation shows that

$$W_\varphi A W_\psi = W_\varphi \Delta W_\psi - \frac{1}{\kappa} W_\varphi \frac{\partial}{\partial t} W_\psi.$$

Furthermore, for any $\tilde{u} \in \mathcal{C}^2(\Xi)$ we find that

$$W_\varphi \frac{\partial}{\partial t} W_\psi[\tilde{u}] = W_\varphi \frac{\partial}{\partial t}[u] = W_\varphi \left[\frac{\partial}{\partial t} u\right] = \widetilde{\frac{\partial}{\partial t}} u = \frac{\partial}{\partial t}\tilde{u},$$

which leads to

$$W_\varphi \frac{\partial}{\partial t} W_\psi = \frac{\partial}{\partial t} I,$$

where I is the identity operator.

We call attention to the fact that

$$\begin{aligned} W_\varphi \Delta W_\psi &= W_\varphi \left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}\right) W_\psi \\ &= W_\varphi \frac{\partial^2}{\partial x^2} W_\psi + W_\varphi \frac{\partial^2}{\partial y^2} W_\psi. \end{aligned}$$

We now finally apply all the above to the aforementioned elliptical confocal change of variables:

$$\begin{cases} \varphi_1(\xi,\eta) &= c\cosh\xi\cos\eta, \\ \varphi_2(\xi,\eta) &= c\sinh\xi\sin\eta. \end{cases} \tag{3}$$

 ISBN: 978-84-617-8694-7

Straightforward computations show that

$$W_{\varphi}\left(\Delta-\frac{1}{\kappa}\frac{\partial}{\partial t}\right)W_{\psi}=\frac{1}{c^2(\cosh^2\xi-\cos^2\eta)}\left[\mathcal{W}_{\xi,\eta}-\frac{c^2(\cosh^2\xi-\cos^2\eta)}{\kappa}\frac{\partial}{\partial t}\right], \tag{4}$$

where

$$\mathcal{W}_{\xi,\eta}:=\frac{\partial^2}{\partial\xi^2}+\frac{\partial^2}{\partial\eta^2}. \tag{5}$$

Applying (4) to (1) leads to the equation

$$\frac{1}{c^2(\cosh^2\xi-\cos^2\eta)}\left[\mathcal{W}_{\xi,\eta}[\theta]-\frac{c^2(\cosh^2\xi-\cos^2\eta)}{\kappa}\frac{\partial\theta}{\partial t}\right]=0, \tag{6}$$

with $\theta=\theta(\xi,\eta,t)$. Then (6) is the two-dimensional Heat-Conduction equation (1) expressed in elliptical confocal coordinates.

The preceding conclusions are true if (3) is a one-to-one correspondence. Hence we assume henceforth that the operator

$$\mathcal{W}_{\xi,\eta}-\frac{c^2(\cosh^2\xi-\cos^2\eta)}{\kappa}\frac{\partial}{\partial t}$$

acts on $C^2(\Omega_{\xi,\eta})\times\mathcal{C}^1(\mathbb{R}_0^+)$, where

$$\Omega_{\xi,\eta}:=(0,\xi_0)\times(0,2\pi)\bigcup\{(0,\eta)\,|\,\eta\in(0,\pi)\}. \tag{7}$$

Upon introducing the symbol θ_o as the *constant temperature of the surrounding medium*, the *boundary condition* along the cylinder surface throughout heating ($\theta>\theta_o$) or coolling ($\theta<\theta_o$) is

$$\frac{\partial\theta}{\varrho\partial\xi}=-\frac{H}{K}(\theta-\theta_o),\ \text{ at }\ \xi=\xi_0, \tag{8}$$

in which H denotes the *convective heat-transfer coefficient*, and $\varrho^2:=c^2(\cosh^2\xi-\cos^2\eta)$. The *initial condition* is

$$\theta=\theta_0,\ \text{ at }\ t=0. \tag{9}$$

In the present study we are interested in the analysis of Eq. (6) and of its solutions under the conditions (8) and (9). Setting $\vartheta:=\theta-\theta_o$, Eqs. (6), (8) and (9) can be written as

$$\frac{1}{c^2(\cosh^2\xi-\cos^2\eta)}\left[\mathcal{W}_{\xi,\eta}[\vartheta]-\frac{c^2(\cosh^2\xi-\cos^2\eta)}{\kappa}\frac{\partial\vartheta}{\partial t}\right]=0, \tag{10}$$

$$\frac{\partial\vartheta}{\varrho\partial\xi}=-\frac{H}{K}\vartheta,\ \text{ at }\ \xi=\xi_0, \tag{11}$$

$$\vartheta=\vartheta_0,\ \text{ at }\ t=0. \tag{12}$$

 ISBN: 978-84-617-8694-7

As usual, we use the method of separation of variables to find a solution of the Heat-Conduction equation (10). Let the desired form of the solution be

$$\vartheta(\xi,\eta,t)=\zeta(\xi,\eta)T(t),$$

where ζ is a function of ξ and η, and T is a function of t alone. Substituting this expression into (10), leads to the following two separate equations for $\zeta(\xi,\eta)$ and $T(t)$:

$$\frac{1}{c^2(\cosh^2\xi-\cos^2\eta)}\Big[\mathcal{W}_{\xi,\eta}[\zeta]+4q\left(\cosh^2\xi-\cos^2\eta\right)\zeta\Big]=0, \tag{13}$$

$$\frac{\partial T}{\partial t}+\kappa\alpha^2T=0, \tag{14}$$

where α denotes the separation constant and $q:=\frac{\alpha^2c^2}{4}$ is a real (positive) parameter. A direct observation shows that Eq. (14) has solution

$$T(t)=e^{-\kappa\alpha^2t}.$$

Suppose now that

$$\zeta(\xi,\eta)=\psi(\xi;q)\phi(\eta;q)$$

is a solution of Eq. (13), where ψ is a function of ξ alone, and ϕ a function of η alone. We obtain the two *second-order ordinary differential equations with variable coefficients*:

$$\frac{\partial^2\phi}{\partial\eta^2}+(a-2q\cos 2\eta)\,\phi=0, \tag{15}$$

$$\frac{\partial^2\psi}{\partial\xi^2}-(a-2q\cosh 2\xi)\,\psi=0, \tag{16}$$

where a is the *separation constant*. A solution of (13) comprises the product of any two functions which are solutions of (15) and (16), respectively, *for the same values of a and q*. Since a may have any value, the number of solutions is unlimited.

Eqs. (15) and (16) are known as the *ordinary* and the *modified Mathieu equations*. However, in applications involving elliptical confocal coordinates, Eqs. (15) and (16) are better identified as the *angular* and *radial Mathieu equations*. Their solutions are, respectively, the *angular Mathieu functions* and the *radial Mathieu functions*. As a matter of fact, since a and q are arbitrary constants, one has two families of ordinary differential equations parametrized by the same parameters a and q.

Remark 2.1 *We shall remark that if in (15) we write $\pm i\xi$ for η, then (15) is transformed into (16), while the latter is transformed into (15) if $\pm i\eta$ is written for ξ.*

 ISBN: 978-84-617-8694-7

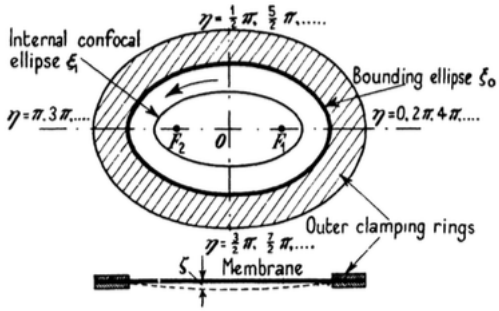


Fig. 2: Diagram for membranal problem, from which the Mathieu functions originated [4].

Referring to Fig. 2, if we start at $\eta = 0$ and move counter-clockwise round a confocal ellipse $\xi = \xi_1 < \xi_0$ the displacement $\zeta(\xi, \eta)$ at any instant alters continuously. It follows that ζ is single-valued and periodic in the coordinate η. The period is π or 2π in η, so that

$$\zeta(\xi, \eta) = \zeta(\xi, \eta + \pi) \quad \text{or} \quad \zeta(\xi, \eta) = \zeta(\xi, \eta + 2\pi)$$

as the case may be. In [7] it is shown that the solutions of Eq. (15), having period π or 2π consist entirely of cosine or sine terms, and not a combination of the two. Moreover, in [6] it is shown that if one solution is even, the other must be odd. Thus two independent even or two independent odd solutions cannot occur. We denote the *even* and the *odd angular Mathieu functions* of order m by

$$\phi_m(\eta; q) := \begin{cases} ce_m(\eta; q), & m = 0, 1, 2, \ldots \\ se_m(\eta; q), & m = 1, 2, 3, \ldots \end{cases} \tag{17}$$

or a constant multiple thereof.

Since m may be any positive integer, there is an infinite number of solutions of type $ce_m(\eta, q)$ and $se_m(\eta, q)$; the parity and periodicity of ce_m and se_m are exactly the same as their trigonometric counterparts, namely, ce_m is an even function and se_m an odd function of η, and they have period π when m is even, or period 2π when m is odd.

3 Preliminaries

Let $\mathbb{H}$ be the set of real quaternions [5]. Each *quaternion* $\mathbf{w}$ is represented in the form

$$\mathbf{w} := w_0 + w_1\mathbf{i} + w_2\mathbf{j} + w_3\mathbf{k}.$$

The set $\{w_i\}$ is in $\mathbb{R}$ and $\{\mathbf{i}, \mathbf{j}, \mathbf{k}\}$ are the *quaternionic imaginary units*, which obey the usual laws of multiplication

$$\mathbf{i}^2 = \mathbf{j}^2 = \mathbf{k}^2 = \mathbf{ijk} = -1$$

and the usual component-wise defined addition.

 ISBN: 978-84-617-8694-7

Similarly to the complex case the *scalar* and *vector parts* of $\mathbf{w}$, $\mathrm{Sc}(\mathbf{w})$ and $\mathrm{Vec}(\mathbf{w})$, are defined as the w_0 and $w_1\mathbf{i}+w_2\mathbf{j}+w_3\mathbf{k}$ terms, respectively. For a quaternionic number $\mathbf{w}$ we consider its *conjugate* $\overline{\mathbf{w}}$ defined by $\overline{\mathbf{w}} := w_0 - \mathrm{Vec}(\mathbf{w})$, and the *norm* $|\mathbf{w}|$ of $\mathbf{w}$ is defined by

$$|\mathbf{w}| = (\mathbf{w}\overline{\mathbf{w}})^{1/2} = (\overline{\mathbf{w}}\mathbf{w})^{1/2} = \left(\sum_{i=0}^{3} w_i^2\right)^{1/2}.$$

In the sequel, let Ω be a domain in $\mathbb{R}^2$ (open and connected) bounded by $\Omega_{\xi,\eta}$ given as (7), with a piecewise smooth boundary. We say that

$$\mathbf{f}:\Omega\to\mathbb{H}, \qquad \mathbf{f}(x,y) := [\mathbf{f}(x,y)]_0 + [\mathbf{f}(x,y)]_1\mathbf{i} + [\mathbf{f}(x,y)]_2\mathbf{j} + [\mathbf{f}(x,y)]_3\mathbf{k}$$

is a *quaternionic-valued function*, where $[\mathbf{f}]_i$ $(i=0,1,2,3)$ are real-valued functions defined in Ω. We will focus on a special class of quaternionic-valued functions analogous to complex holomorphic functions and connected with them via the following concept.

Definition 3.1 *For a given $q>0$, a function $f\in C^1(\Omega,\mathbb{H})$ is called $\mathcal{D}_q$-hyperholomorphic if it is a solution of the differential equation $\mathcal{D}_q[f]=0$ with variable and quaternionic coefficients, where*

$$\mathcal{D}_q := \frac{2\sqrt{q}}{c} + \frac{1}{c(\cosh^2\xi-\cos^2\eta)}\left[(\mathbf{i}\sinh\xi\cos\eta+\mathbf{j}\cosh\xi\sin\eta)\frac{\partial}{\partial\xi}\right.$$
$$\left. + (\mathbf{j}\sinh\xi\cos\eta-\mathbf{i}\cosh\xi\sin\eta)\frac{\partial}{\partial\eta}\right]. \tag{18}$$

Analogously, for a solution of the equation $\overline{\mathcal{D}}_q[f]=0$ where

$$\overline{\mathcal{D}}_q := \frac{2\sqrt{q}}{c} - \frac{1}{c(\cosh^2\xi-\cos^2\eta)}\left[(\mathbf{i}\sinh(\xi)\cos\eta+\mathbf{j}\cosh\xi\sin\eta)\frac{\partial}{\partial\xi}\right.$$
$$\left. + (\mathbf{j}\sinh\xi\cos\eta-\mathbf{i}\cosh\xi\sin\eta)\frac{\partial}{\partial\eta}\right] \tag{19}$$

a reasonably natural name is $\mathcal{D}_q$-anti-hyperholomorphic function.

For a given $q>0$, it can be easily checked that the operators (18) and (19) factorize the operator induced by Eq. (13):

$$\mathcal{W}_{\xi,\eta} + 4q(\cosh^2\xi-\cos^2\eta) =: \mathcal{W}$$

in the sense that

$$\mathcal{D}_q\overline{\mathcal{D}}_q = \overline{\mathcal{D}}_q\mathcal{D}_q = \frac{1}{c^2(\cosh^2\xi-\cos^2\eta)}\mathcal{W},$$

where $\mathcal{W}_{\xi,\eta}$ is given by (5).

 ISBN: 978-84-617-8694-7

The above means that $\mathcal{D}_q$-hyperholomorphic functions play indeed the same role for the $\mathcal{W}$ operator as the usual holomorphic functions in one complex variable, or quaternionic hyperholomorphic functions, play for the corresponding Laplace operator [1]. At the same time, there exists a deep difference since the operators (18) and (19) have *variable* and *non-constant coefficients*, and it is well-known that function theories using such operators are much more sophisticated.

Consider now the space of all square integrable quaternionic-valued functions on Ω such that each component is in the usual $L_2(\Omega)$ and define

$$L_2(\Omega,\mathbb{H}) := \left\{ \mathbf{f} \,|\, \mathbf{f}:\Omega\to\mathbb{H},\ \|\mathbf{f}\|_{L_2(\Omega,\mathbb{H})} := \left(\int_\Omega |\mathbf{f}(x,y)|^2 dxdy\right)^{1/2} < \infty \right\}.$$

It follows that the space $L_2(\Omega,\mathbb{H})$ endowed with the *right quaternionic inner product*

$$< \mathbf{f}, \mathbf{g} >_{L_2(\Omega,\mathbb{H})} := \int_\Omega \overline{\mathbf{f}(x,y)}\, \mathbf{g}(x,y)\, dxdy \tag{20}$$

is a $\mathbb{H}$-*linear Hilbert space* under right multiplication by quaternion numbers.

4 The Quaternionic Mathieu Functions (QMFs)

4.1 Definition and properties of QMFs

We now show how to directly relate the solutions of Eqs. (15) and (16) with the above established hyperholomorphic function theory. The strategy adopted is the following: we start by considering a null-solution to the operator

$$\frac{1}{c^2(\cosh^2\xi-\cos^2\eta)}\Big[\mathcal{W}_{\xi,\eta} + 4q(\cosh^2\xi-\cos^2\eta)\Big]$$

where $\mathcal{W}_{\xi,\eta}$ is given by (5).

We identify such a solution $\zeta(\xi,\eta)=\psi(\xi;q)\phi(\eta;q)$ as of the form

$$\begin{cases} \zeta_m^e(\xi,\eta;q) := Ce_m(\xi;q)ce_m(\eta;q), \quad m=0,1,2,\ldots \\ \zeta_m^o(\xi,\eta;q) := Se_m(\xi;q)se_m(\eta;q), \quad m=1,2,\ldots \end{cases}$$

where the corresponding values of q and m must be the same as those entering the angular equation (15), as q and m together determine the separation constant a. For the sake of simplicity, we denote by ζ_m any Mathieu function of the set

$$\{\zeta_m^e,\ \zeta_m^o : m=0,1,2,\ldots\}. \tag{21}$$

 ISBN: 978-84-617-8694-7

Definition 4.1 (QMFs) *For a given $q > 0$, the family of pairs*

$$\{\, \boldsymbol{\zeta}_m,\, \overline{\boldsymbol{\zeta}_m} : \, m = 1, 2, \dots \},$$

where we set

$$\boldsymbol{\zeta}_m := \frac{c}{4\sqrt{q}}\, \overline{\mathcal{D}}_q[\zeta_m] \quad \text{and} \quad \overline{\boldsymbol{\zeta}_m} := \frac{c}{4\sqrt{q}}\, \mathcal{D}_q[\zeta_m]$$

is called Quaternionic Mathieu Functions (QMFs).

Remark 4.1 *A direct observation shows that*

$$\mathcal{D}_q[\boldsymbol{\zeta}_m] = \frac{1}{4\sqrt{q}}\, \frac{1}{c(\cosh^2\xi - \cos^2\eta)} \Big[\mathcal{W}_{\xi,\eta} + 4q(\cosh^2\xi - \cos^2\eta)\Big][\zeta_m] = 0,$$

and

$$\overline{\mathcal{D}}_q[\overline{\boldsymbol{\zeta}_m}] = \frac{1}{4\sqrt{q}}\, \frac{1}{c(\cosh^2\xi - \cos^2\eta)} \Big[\mathcal{W}_{\xi,\eta} + 4q(\cosh^2\xi - \cos^2\eta)\Big][\zeta_m] = 0.$$

Hence each element of (21) can be decomposed into

$$\zeta_m^e = \boldsymbol{\zeta}_m^e + \overline{\boldsymbol{\zeta}_m^e} \qquad \Big(\text{resp.},\ \zeta_m^o = \boldsymbol{\zeta}_m^o + \overline{\boldsymbol{\zeta}_m^o}\Big)$$

where $\boldsymbol{\zeta}_m^e$ (resp. $\boldsymbol{\zeta}_m^o$) are $\mathcal{D}_q$-hyperholomorphic functions, and $\overline{\boldsymbol{\zeta}_m^e}$ (resp. $\overline{\boldsymbol{\zeta}_m^o}$) are $\mathcal{D}_q$-anti-hyperholomorphic functions.

Thus, we have the following result.

Lemma 4.1 *For a given $q > 0$, the QMFs are of the following form:*

$$\begin{aligned}
\boldsymbol{\zeta}_m(\xi,\eta) \;=\; & \frac{1}{2}\psi(\xi)\phi(\eta) \\
& + \mathbf{i}\, \frac{c}{4\sqrt{q}}\, \frac{1}{(\cos^2\eta - \cosh^2\xi)} \left[\sinh\xi\cos\eta\,\psi'(\xi)\phi(\eta) - \cosh\xi\sin\eta\,\psi(\xi)\phi'(\eta)\right] \\
& + \mathbf{j}\, \frac{c}{4\sqrt{q}}\, \frac{1}{(\cos^2\eta - \cosh^2\xi)} \left[\cosh\xi\sin\eta\,\psi'(\xi)\phi(\eta) + \sinh\xi\cos\eta\,\psi(\xi)\phi'(\eta)\right]
\end{aligned}$$

and

$$\begin{aligned}
\overline{\boldsymbol{\zeta}_m(\xi,\eta)} \;=\; & \frac{1}{2}\psi(\xi)\phi(\eta) \\
& - \mathbf{i}\, \frac{c}{4\sqrt{q}}\, \frac{1}{(\cos^2\eta - \cosh^2\xi)} \left[\sinh\xi\cos\eta\,\psi'(\xi)\phi(\eta) - \cosh\xi\sin\eta\,\psi(\xi)\phi'(\eta)\right] \\
& - \mathbf{j}\, \frac{c}{4\sqrt{q}}\, \frac{1}{(\cos^2\eta - \cosh^2\xi)} \left[\cosh\xi\sin\eta\,\psi'(\xi)\phi(\eta) + \sinh\xi\cos\eta\,\psi(\xi)\phi'(\eta)\right]
\end{aligned}$$

where $\psi(\xi)$ is a solution of Eq. (16) and $\phi(\eta)$ is a solution of Eq. (15).

 ISBN: 978-84-617-8694-7

Referring to Fig. 2, at any point $(0,\eta)$ and the corresponding point $(0,-\eta)$ across the line segment $\xi = 0$, we have continuity of displacement and gradient in crossing the interfocal line orthogonally.

Theorem 4.1 (a) *continuity of displacement, i.e.* $\boldsymbol{\zeta}(0,\eta) = \boldsymbol{\zeta}(0,-\eta)$;

(b) *continuity of gradient, i.e.* $\dfrac{\partial}{\partial\xi}\left[\boldsymbol{\zeta}(\xi,\eta)\right]_{\xi\to 0} = -\dfrac{\partial}{\partial\xi}\left[\boldsymbol{\zeta}(\xi,-\eta)\right]_{\xi\to 0}$.

The *orthogonality* of the QWFs for the same q but different a over confocal ellipses with respect to the quaternionic inner product (20) is given in the following.

Theorem 4.2 *For a given $q>0$, the set $\{\boldsymbol{\zeta}_m^e, \boldsymbol{\zeta}_m^o : m = 1,2,\dots\}$ is orthogonal over the confocal ellipses (2) in the sense of the quaternionic inner product (20).*

4.2 Traveling wave function in the confocal elliptical cylinder

The QMFs will now be used to define the total wave function produced in a medium of arbitrary q under conditions (11) and (12). Since each QMF is linked to a specific characteristic constant a, the complete in-medium function $\boldsymbol{\vartheta}(\xi,\eta,t)$ is obtained by multiplying each QMF by the appropriate t-dependent exponential for which the associated a values cause them to satisfy the boundary conditions (11) and (12). Thus the *total wave function in the medium* is

Definition 4.2 (Total Wave Function) *For a given $q>0$, the total wave function associated with the QMFs is defined by*

$$\boldsymbol{\vartheta}(\xi,\eta,t) := \sum_{m=1}^{\infty} \boldsymbol{\zeta}_m(\xi,\eta) e^{-\kappa\alpha_m^2 t}\boldsymbol{\beta}_m \tag{22}$$

for integral quaternionic constants $\boldsymbol{\beta}_m$.

The coefficients in (22) *may* be evaluated numerically using the orthogonality of the underlying QMFs. Using the unknown constants $\boldsymbol{\beta}_m$, the initial condition (12) reads now as follows:

$$\boldsymbol{\vartheta}_0 = \sum_{m=1}^{\infty} \boldsymbol{\zeta}_m(\xi,\eta)\boldsymbol{\beta}_m. \tag{23}$$

Integrating, with respect to ξ from 0 to ξ_0 and to η from 0 to 2π, both sides of the equation obtained by multiplying from the left Eq. (23) by $c^2(\cosh^2\xi - \cos^2\eta)\overline{\boldsymbol{\zeta}_p(\xi,\eta)}$, it follows that

$$\begin{aligned}&\left[\int_0^{2\pi}\int_0^{\xi_0} \overline{\boldsymbol{\zeta}_p(\xi,\eta)} c^2(\cosh^2\xi - \cos^2\eta) d\xi d\eta\right]\boldsymbol{\vartheta}_0\\ &= \sum_{m=0}^{\infty}\left[\int_0^{2\pi}\int_0^{\xi_0} \overline{\boldsymbol{\zeta}_p(\xi,\eta)}\boldsymbol{\zeta}_m(\xi,\eta) c^2(\cosh^2\xi - \cos^2\eta) d\xi d\eta\right]\boldsymbol{\beta}_m,\quad p = 1,2,\dots.\end{aligned}$$

 ISBN: 978-84-617-8694-7

From Theorem 4.2, we readily have

$$\boldsymbol{\beta}_m = \boldsymbol{f}_m(\xi_0)\boldsymbol{\vartheta}_0, \quad m = 1, 2, \ldots$$

where

$$\boldsymbol{f}_m(\xi_0) := \frac{\displaystyle\int_0^{2\pi}\int_0^{\xi_0} \overline{\boldsymbol{\zeta}_m(\xi,\eta)}c^2(\cosh^2\xi - \cos^2\eta)d\xi d\eta}{\displaystyle\int_0^{2\pi}\int_0^{\xi_0} |\boldsymbol{\zeta}_m(\xi,\eta)|^2 c^2(\cosh^2\xi - \cos^2\eta)d\xi d\eta}.$$

The authors are currently attempting to explore the properties of the function (22) and its transition to circular membranes in more detail.

References

[1] V. Kravchenko and M. Shapiro. *Integral representations for spatial models of mathematical physics*. Addison-Wesley-Longman. Pitman Research Notes in Mathematics, 1996.

[2] M.E. Luna-Elizarrarás, M.A. Pérez-de la Rosa, R.M. Rodríguez-Dagnino and M. Shapiro. *On quaternionic analysis for the Schrödinger operator with a particular potential and its relation with the Mathieu functions*. Math. Meth. Appl. Sci. **36** (2013) 1080–1094.

[3] E. Mathieu. *Mémoire sur le mouvement vibratoire d'une membrane de forme elliptique*. J. Math. Pures Appl. **13** (1868) 137–203.

[4] N. McLachlan. *Theory and applications of Mathieu functions*. Oxford Press, London, 1951.

[5] J. Morais, S. Georgiev and W. Sprössig. *Real quaternionic calculus handbook*. Birkhäuser, Basel, 2014.

[6] E. Poole. *Introduction to the theory of linear differential equations*. Oxford: Clarendon press, 1936.

[7] E. Sarchinger. *Beiträge zur theorie der funktionen des elliptischen zylinders*. D., Leipzig, 1894.

[8] K. Sato. *Heat conduction in infinite elliptical cylinder during heating or cooling*. Theoretical and Applied Mechanics Japan, **55** (2006) 157–168.

[9] K. Sato. *Transient heat conduction in infinite hollow confocal elliptical cylinder*. Theoretical and Applied Mechanics Japan, **58** (2010) 167–175.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

Discovering the composition of audio files by Audio-to-MIDI alignment

A. J. Muñoz-Montoro[1], P. Cabañas-Molero[1], F. J. Bris-Peñalver[1], E. F. Combarro[2], R. Cortina[2] and P. Alonso[3]

[1] *Department of Telecommunication Engineering, University of Jaén*

[2] *Department of Computer Science, University of Oviedo*

[3] *Department of Information Systems and Computation, Universitat Politècnica de València*

emails: jmontoro@ujaen.es, pcabanas@ujaen.es, fbris@ujaen.es, efernandezca@uniovi.es, raquel@uniovi.es, palonso@upv.es

Abstract

This paper presents a framework for the automatic identification of the composition of small portions of music files. The task is performed by aligning the portion of audio with the stored MIDI scores. The system provides a ranked list of the compositions that best match the audio query, as well as their similarity measures and their aligned MIDI files. Preliminary results show the precision and robustness of the proposal.

Key words: Audio-to-MIDI Alignment, Dynamic Time Warping, Audio Identification

1 Introduction

Over the last years, a large amount of multimedia content has been placed on the Internet and made accessible through streaming sites. Many of the available files include performances of well-known classical music pieces, but the identity of the compositions is not always annotated. In this scenario, it is of great interest to identify automatically the composition corresponding to a given audio excerpt. This task involves computing the similarity between the audio input and every entry in a database of musical pieces, finally choosing the entry with the highest similarity. Since two performances of the same composition may

 ISBN: 978-84-617-8694-7

differ in tempo or interpretation, and usually the input audio is only a small portion of the whole piece, the common way to compute this similarity is through audio alignment techniques.

Audio alignment can be defined as the synchronization between two musical sequences which may have differences in interpretation. Alignment tools have been widely used in other related applications, such as cover song identification [1], query-by-tapping [2] or query-by-humming [3]. Most of the recent works are focused on the alignment between a musical score (such as MIDI or any other symbolic representation) and an audio excerpt [4, 5, 6]. Based on MIDI-to-audio alignment, a large number of approaches have been proposed to identify a query of a music excerpt using a large database of music score representations [7, 8, 9, 3]. In these cases, the alignment result is used as a criteria to measure how well the content in a MIDI score matches an input audio segment, such that the audio can be correctly identified if the correct composition is contained in the database. In the context of classical music, it is possible to gather large collections of accurately transcribed MIDI files, enabling to construct a reliable database for identification purposes.

To measure the similarity between the audio and the score, a set of features that characterize the musical content is first extracted from the audio signal. Features proposed by authors are often related to the target application, and a large variety of them can be found, such as chroma vectors [8, 1, 10, 11], beat-tracking [12], decaying locally adaptive normalized chroma onset (DLNCO) [5], peak structure distance (PSD) [13, 4] or measures derived from analysis with Non-negative Matrix Factorization (NMF) [14].

The alignment is performed by finding the best match between the extracted feature sequence and the score. Commonly, the process consists in filling a similarity matrix (or cost matrix) of the form $\mathbf{D}(\tau, t)$, containing distance values between discrete times τ in the MIDI and discrete times t in the audio. The most common approach to find the correspondence between both temporal axes is dynamic time warping (DTW) [11, 15, 16]. DTW uses dynamic programming to find an alignment such that the sum of the distance between aligned time instants is minimized. This output global distance is a natural measure of similarity between two sequences, and can be used to decide the best matching in a corpora of MIDI files. Consequently, for audio-to-MIDI matching applications, the performance of the system heavily depends on the robustness of the alignment core, particularly on its ability to compensate changes in tempo and interpretation.

In this paper we propose an audio-to-MIDI matching approach that identifies a music excerpt query among a database of classical music MIDI scores. The alignment kernel is based on our robust system proposed in [17, 18], employing a fast spectral factorization algorithm and DTW. As a result, the system provides a ranked list of the compositions that best match the audio query, as well as their similarity measures and their aligned MIDI files.

 ISBN: 978-84-617-8694-7

2 Proposed audio-to-MIDI matching system

The proposed framework for audio-to-MIDI matching is composed of two main modules. The MIDI preprocessing module must be carried out beforehand for each MIDI file in the database, with the aim of learning a set of parameters that adequately represent each score. The matching module performs the alignment between the input audio excerpt and each entry of the MIDI database, returning a confidence alignment measure.

2.1 MIDI Preprocessing Module

The aim of this module is to adequately represent the information given by a MIDI file to be used for alignment purposes. This task is performed only once, when a new MIDI entry is incorporated to the database, and not during the matching stage. The input MIDI score can be represented by a binary matrix $\mathbf{GT}(n,\tau)$, where τ is the time-frame index referenced to the score (MIDI time) and n are the notes in MIDI scale. This matrix has dimensionality $N \times L_m$, where N is the total number of notes in the composition and L_m is the number of frames in MIDI time. If the score contains different instruments, N is obtained as the sum of the different notes per instrument.

Each unique occurrence of individual or concurrent notes will be denoted here as a *score unit*. K is the number of units in the score and will be most likely smaller than the number of notes ($K << N$). In terms of score units, the score matrix $\mathbf{GT}(n,\tau)$ can be decomposed as follows

$$\mathbf{GT}(n,\tau) = \mathbf{Q}(n,k)\mathbf{R}(k,\tau), \tag{1}$$

where $\mathbf{Q}(n,k)$ is the binary notes-to-units matrix, k the index of each unique unit and $\mathbf{R}(k,\tau)$ represents the binary activation of each unit. Observe that $\mathbf{Q}(n,k)$ informs about the notes belonging to each unit, whereas $\mathbf{R}(k,\tau)$ retains the MIDI time activation per unit.

Since our alignment algorithm is based on spectral decomposition, the preprocessing module learns in advance a single spectral pattern for each score unit. To this end, a synthetic signal is first generated from the score using a MIDI synthesizer. Let us denote the magnitude spectrogram of the synthetic signal as $\mathbf{Y}(f,\tau)$, with f being the frequency bin index. This signal is then decomposed according to the following model:

$$\mathbf{Y}(f,\tau) \approx \hat{\mathbf{Y}}(f,\tau) = \mathbf{B}(f,k)\mathbf{G}(k,\tau), \tag{2}$$

where $\hat{\mathbf{Y}}(f,\tau)$ is the estimated spectrogram, $\mathbf{G}(k,\tau)$ matrix represents the gain of the spectral pattern for unit k at frame τ, and $\mathbf{B}(f,k)$ matrix represents the spectral patterns for all the units defined in the score. The parameters are estimated using NMF with β-divergence and multiplicative update rules, where $\mathbf{G}(k,\tau)$ is initialized to $\mathbf{R}(k,\tau)$, and $\mathbf{B}(f,k)$ to random positive numbers.

As a result of the preprocessing step, the system stores the matrices $\mathbf{R}(k,\tau)$ and $\mathbf{B}(f,k)$ for the MIDI file.

 ISBN: 978-84-617-8694-7

2.2 MIDI-to-audio Alignment Module

To perform the alignment between the input signal and a certain MIDI, the first step is the computation of a similarity measure between the audio and the different units defined by the score. In our approach, this measure is given by the distortion between the frequency transform of the input and the spectral patterns learned per unit.

Lets denote the frequency-domain input signal vector at time t as $\mathbf{x}_t(f)$, and the k-th unit spectral pattern as $\mathbf{b}_k(f)$. Assuming a signal model in which only a single pattern can be active at t, the gain $g_{k,t}$ that minimizes the β-divergence between the input frame and pattern k is computed by [19]:

$$g_{k,t} = \frac{\sum_f \mathbf{x}_t(f)\mathbf{b}_k(f)^{(\beta-1)}}{\sum_f \mathbf{b}_k(f)^{\beta}}. \tag{3}$$

Finally, the distortion matrix for each unit at each frame is defined by

$$\mathbf{\Phi}(k,t) = D_\beta\left(\mathbf{x}_t(f)|g_{k,t}\mathbf{b}_k(f)\right), \tag{4}$$

where $D_\beta(\cdot)$ is the β-divergence function and β can take values in the range $\in [0,2]$.

As can be inferred, the distortion matrix $\mathbf{\Phi}(k,t)$ provides us information about the similitude of each k-th unit spectral pattern with the real signal spectrum at each frame t. Using this information, we can directly compute the cost matrix between the MIDI time τ and the time of the input signal t as

$$\mathbf{D}(\tau,t) = \mathbf{R}^T(\tau,k)\mathbf{\Phi}(k,t), \tag{5}$$

where superscript "T" stands for matrix transposition. This matrix has dimensions $T_m \times T_r$, where T_r is the number of frames of the input signal.

To perform the alignment, the optimum path across matrix $\mathbf{D}(\tau,t)$ is obtained with DTW. Essentially, DTW consists in filling recursively a warping matrix $\mathbf{C}$ as follows:

$$\mathbf{C}(\tau,t) = \min \left\{ \begin{array}{c} \mathbf{C}(\tau-1,t-1) + \mathbf{D}(\tau,t) \\ \mathbf{C}(\tau-2,t-1) + \sigma_{2,1}\mathbf{D}(\tau,t) \\ \vdots \\ \mathbf{C}(\tau-\alpha_\tau,t-1) + \sigma_{\alpha_\tau,1}\mathbf{D}(\tau,t) \\ \mathbf{C}(\tau-1,t-2) + \sigma_{1,2}\mathbf{D}(\tau,t) \\ \vdots \\ \mathbf{C}(\tau-1,t-\alpha_t) + \sigma_{1,\alpha_t}\mathbf{D}(\tau,t) \end{array} \right\} \tag{6}$$

where the step size at each dimension has a range from 1 to α_τ and 1 to α_t, respectively. α_τ and α_t are the maximum step size at each dimension. Parameter σ controls the bias toward

 ISBN: 978-84-617-8694-7

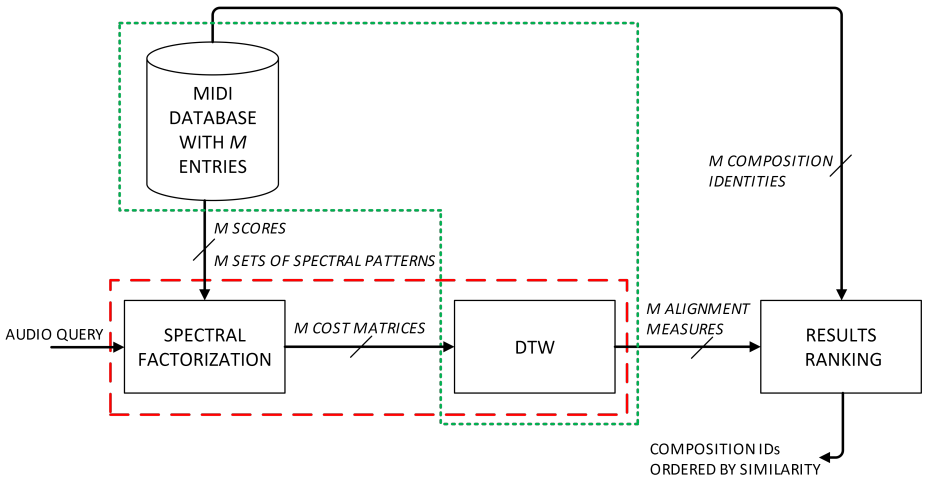


Figure 1: Block diagram of the proposed audio-to-MIDI matching engine.

diagonal steps as $\sigma_{x,y} = \sqrt{x^2 + y^2}$. $\mathbf{C}(\tau, t)$ is the accumulated cost of the minimum-cost path up to (τ, t), and $\mathbf{C}(\tau, 1) = \mathbf{D}(\tau, 1)\ \forall \tau$.

Once the warping matrix has been filled, it is possible to find the minimum-cost path $\mathbf{w} = w_1, ..., w_l, ..., w_L$, where each w_l is an ordered pair (τ_l, t_l) meaning that instant τ_l must be aligned with t_l. The path is obtained by tracing the recursion backwards from $\mathbf{C}(\tau_L, T_r)$, where $\tau_L = \arg\min_\tau \mathbf{C}(\tau, T_r)$. Since the audio query is usually a small fragment of the whole composition, the first and last elements of the path can be at any point along the τ axis. Globally, the path has to satisfy the following three conditions:

i Boundary condition: $w_1 = (\tau, 1)$ and $w_L = (\tau, T_r)$.

ii Monotonicity condition: $\tau_{l+1} \geq \tau_l$ and $t_{l+1} \geq t_l$.

iii Step size condition: $\tau_{l+1} \leq \tau_l + \alpha_\tau$ and $t_{l+1} \leq t_l + \alpha_t$.

This approach has the advantage of being computationally simple, and can be applied to perform audio-to-score matching over large datasets. For a certain score, the value $\mathbf{C}(\tau_L, T_r)$ can be considered as a confidence measure expressing the quality of the matching.

Figure 1 shows a block diagram of the system. In response to an audio query (typically, a excerpt of a few seconds), the system computes the alignment confidence measure $\mathbf{C}(\tau_L, T_r)$ for each MIDI in the database, returning a list of results ordered according to this value.

In [20] the software ReMAS (Real-time Musical Accompaniment System) designed to track the reproduction of a musical piece with the aim to match the score position into its symbolic representation on a digital sheet was presented. ReMAS shows that it is possible

to exploit efficiently several cores of an ARM® processor, or a GPU accelerator, reducing the processing time per frame in a few milliseconds in most of the cases. On the other hand, [21] proposes a parallel online DTW solution based on a client–server architecture implemented for multi-core architectures (86, 64 and ARM®). Looking at Figure 1, the blocks within the rectangle of dashed red lines have a high degree of similarity to ReMAS, whilst the shape of dashed green lines does so with [21]. Thereby the framework proposed in this work aims to adapt, extend and test the parallel heterogeneous algorithms of [21] and [20] to the new problem.

3 EXPERIMENTS AND RESULTS

The MIDI database for our audio-to-MIDI matching system has been obtained from the web *www.piano-midi.de*. It is composed of about 1123 minutes of score with 679986 played notes in 336 files. It contains several classical music pieces by 23 classical composers.

The system has been evaluated with 63 real audio performances downloaded from the Internet. For each audio file, six segments of 5, 10, 15, 20, 25 and 30 seconds with random start times along the file were selected and entered as a query into the system. Table 1 shows the percentage of searches in which the correct composition is in the 1st position of the list of results. As shown, for very short queries of only 5s, the correct composition is detected as the best result for the $85,7\%$ of the audio files. This accuracy rises to 98,4% for queries with a duration of 25s or longer.

Table 1: Accuracy measures (%)

5s	**10s**	**15s**	**20s**	**25s**	**30s**
85,7	93,6	95,2	96,8	98,4	98,4

Acknowledgements

This work has been supported by the "Ministerio de Economía y Competitividad" of Spain / FEDER under projects TEC2015-67387-C4-{1, 2, 3}-R.

References

[1] J. Serrà, E. Gómez, P. Herrera, and X. Serra, "Chroma binary similarity and local alignment applied to cover song identification," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 16, pp. 1138–1151, 2008.

 ISBN: 978-84-617-8694-7

[2] P. Hanna and M. Robine, "Query by tapping system based on alignment algorithm," in *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*, april 2009, pp. 1881–1884.

[3] Matti Ryynanen and Anssi Klapuri, "Query by humming of midi and audio using locality sensitive hashing," in *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*. IEEE, 2008, pp. 2249–2252.

[4] J. Devaney and D.P.W. Ellis, "Handling asynchrony in audio-score alignment," in *Proceedings of the International Computer Music Conference Computer Music Association*, 2009, pp. 29–32.

[5] S. Ewert, M. Muller, and P. Grosche, "High resolution audio synchronization using chroma onset features," in *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*, april 2009, pp. 1869–1872.

[6] A. Cont, "A coupled duration-focused architecture for real-time music-to-score alignment," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 32, no. 6, pp. 974–987, june 2010.

[7] H Lin, H Chen, and T Ma, "Web-based music lecture database framework with aligned midi score and real performance audio," *Multimedia, IEEE*, vol. PP, no. 99, pp. 1, 2009.

[8] Riccardo Miotto and Nicola Orio, "A music identification system based on chroma indexing and statistical modeling.," in *ISMIR*, Juan Pablo Bello, Elaine Chew, and Douglas Turnbull, Eds., 2008, pp. 301–306.

[9] Hung-Chen Chen, Yi-Hung Wu, Yu-Chi Soo, and Arbee Chen, "Continuous query processing over music streams based on approximate matching mechanisms," *Multimedia Systems*, vol. 14, pp. 51–70, 2008.

[10] C. Joder, S. Essid, and G. Richard, "A conditional random field framework for robust and scalable audio-to-score matching," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. PP, no. 99, 2011.

[11] Ning Hu, R.B. Dannenberg, and G. Tzanetakis, "Polyphonic audio matching and alignment for music retrieval," in *Applications of Signal Processing to Audio and Acoustics, 2003 IEEE Workshop on.*, oct. 2003, pp. 185–188.

[12] T. Otsuka, T. Takahashi, H. G. Okuno, K. Komatani, T. Ogata, K. Murata, and K. Nakadai, "Incremental polyphonic audio to score alignment using beat tracking for singer robots," in *2009 IEEE/RSJ International Conference on Intelligent Robots and Systems*, Oct 2009, pp. 2289–2296.

 ISBN: 978-84-617-8694-7

[13] Robert J. Turetsky and Daniel P. W. Ellis, "Ground-truth transcriptions of real music from force-aligned MIDI syntheses," in *Proceedings of the 4th International Society for Music Information Retrieval Conference*, 2003, pp. 135–141.

[14] A. Cont, "Realtime audio to score alignment for polyphonic music instruments, using sparse non-negative constraints and hierarchical hmms," in *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, May 2006, vol. 5, pp. V–V.

[15] Nicola Orio and Diemo Schwarz, "Alignment of monophonic and polyphonic music to a score," in *in Proceedings of the ICMC*, 2001, pp. 155–158.

[16] S. Dixon, "Live tracking of musical performances using on-line time warping," in *Proc. International Conference on Digital Audio Effects (DAFx)*, 2005, pp. 92–97.

[17] J.J. Carabias-Orti, F.J. Rodríguez-Serrano, P. Vera-Candeas, N. Ruiz-Reyes, and F.J. Cañadas-Quesada, "An audio to score alignment framework using spectral factorization and dynamic time warping," in *Proc. ISMIR*, 2015, pp. 742–748.

[18] F.J. Rodríguez-Serrano, J.J. Carabias-Orti, P. Vera-Candeas, and D. Martinez-Muñoz, "Tempo driven audio-to-score alignment using spectral decomposition and online dynamic time warping," *ACM Trans. Intell. Syst. Technol.*, vol. 8, no. 2, pp. 22:1–22:20, 2016.

[19] J.J. Carabias-Orti, F.J. Rodríguez-Serrano, P. Vera-Candeas, F.J. Cañadas-Quesada, and N. Ruiz-Reyes, "Constrained non-negative sparse coding using learnt instrument templates for realtime music transcription," *Engineering Applications of Artificial Intelligence*, vol. 26, no. 7, pp. 1671 – 1680, 2013.

[20] P. Alonso, P. Vera-Candeas, R. Cortina, and J. Ranilla, "An efficient musical accompaniment parallel system for mobile devices," *The Journal of Supercomputing*, vol. 73, no. 1, pp. 343–353, 2017.

[21] P. Alonso, R. Cortina, F.J. Rodríguez-Serrano, P. Vera-Candeas, M. Alonso-González, and J. Ranilla, "Parallel online time warping for real-time audio-to-score alignment in multi-core systems," *The Journal of Supercomputing*, vol. 73, no. 1, pp. 126–138, 2017.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

Computation of periodic orbits in a three level trophic chain model

Juan F. Navarro[1] **and Rubén Poveda**[1]

[1] *Department of Applied Mathematics, University of Alicante*

emails: jf.navarro@ua.es, rpm52@alu.ua.es

Abstract

This contribution deals with an adaptation of the Poincaré–Lindstedt method for the determination of periodic orbits in three–dimensional nonlinear differential systems. We describe here a general symbolic algorithm to implement the method in a specific symbolic system which works with modified Poisson series. The sufficient conditions to make secular terms dissapear from the approximate series solution are given here.

Key words: Lotka–Volterra, periodic orbits, Poincaré–Lindstedt method, symbolic computation.

1 Introduction

The Poincaré–Lindstedt technique is a classical perturbation method used to continue a periodic orbit with respect to a small perturbation parameter, when fixing the amplitude (or the energy) of the system. This method has been used extensively to the analysis of a wide variety of systems in many branches of science: from galactic ([11], [12]) to atomic models ([9]), and covering also applications in population biology, ecology and mathematical biology ([1]). Nowadays, many researchers ([2], [4], [5], [6], [10] and [14] to cite some examples) make use of this method to study dynamical systems.

However, the efforts devoted to the application of the Poincaré–Lindstedt technique to systems of differential equations presenting a periodic orbit have not been so abundant. In [13], an approximation to the periodic solutions of the general Lotka–Volterra prey–predator system is obtained using the Poincaré–Lindstedt method. In [8], the method of Poincaré–Lindstedt is adapted to compute periodic solutions in perturbed two–dimensional systems. In particular, an approximate solution to a Lotka–Volterra model for two species

 ISBN: 978-84-617-8694-7

is computed. The computation of periodic solutions in Lotka–Volterra systems is an open problem where the Poincaré–Lindstedt method could play a key role in the computation of periodic orbits and the understanding of the way the phase space is structured not only in two species systems.

The aim of this paper is to present a general algorithm for implementing the standard Poincaré–Lindstedt method to three–dimensional perturbed systems of differential equations of first order. This adaptation is successfully applied to compute periodic solutions in Lotka–Volterra systems modeling a three–species food chain interaction. In the following section, we describe how to adapt the standard method to three–dimensional systems of the Lotka–Volterra type and give sufficient conditions to make secular terms dissapear from the solution. This result is key to set the stage to adapt the perturbation method to three–dimensional systems of differential equations.

2 Adaptation of the Poincaré–Lindstedt Method for Three–Dimensional Systems

Let us consider the problem defined by the following nonlinear differential system of first order,

$$\begin{aligned} \dot{x} + \alpha_{12} y &= \epsilon\, f_1(x,y,z)\,, \\ \dot{y} - \alpha_{21} x + \alpha_{23} z &= \epsilon\, f_2(x,y,z)\,, \\ \dot{z} - \alpha_{32} y &= \epsilon\, f_3(x,y,z)\,, \end{aligned} \tag{1}$$

where $0 < \epsilon \ll 1$ is a small parameter and functions $f_1(x,y,z)$, $f_2(x,y,z)$ and $f_3(x,y,z)$ can be arranged as follows,

$$\begin{aligned} f_1(x,y,z) &= \sum_{0\le q\le M} \sum_{0\le \nu_1\le q} \sum_{0\le \nu_2\le \nu_1} f_{1,\nu_2,q-\nu_1,\nu_1-\nu_2}\, x^{\nu_2} y^{q-\nu_1} z^{\nu_1-\nu_2}\,, \\ f_2(x,y,z) &= \sum_{0\le q\le M} \sum_{0\le \nu_1\le q} \sum_{0\le \nu_2\le \nu_1} f_{2,\nu_2,q-\nu_1,\nu_1-\nu_2}\, x^{\nu_2} y^{q-\nu_1} z^{\nu_1-\nu_2}\,, \\ f_3(x,y,z) &= \sum_{0\le q\le M} \sum_{0\le \nu_1\le q} \sum_{0\le \nu_2\le \nu_1} f_{3,\nu_2,q-\nu_1,\nu_1-\nu_2}\, x^{\nu_2} y^{q-\nu_1} z^{\nu_1-\nu_2}\,. \end{aligned} \tag{2}$$

where $f_{1,\nu_2,q-\nu_1,\nu_1-\nu_2}$, $f_{2,\nu_2,q-\nu_1,\nu_1-\nu_2}$ and $f_{3,\nu_2,q-\nu_1,\nu_1-\nu_2} \in \mathbb{R}$ for $0 \le q \le M$, $0 \le \nu_1 \le q$, $0 \le \nu_2 \le \nu_1$ and $M \in \mathbb{N}$.

If the unperturbed system ($\epsilon = 0$) has periodic solutions and ϵ is a measure of the size of the perturbing terms, then the trajectories for the full system will remain pretty close to those of the non–perturbed system, for any finite period of time $t_0 < t < t_0 + \alpha$ ($\alpha > 0$) with an error not larger than $O(\alpha)$. In general, even a small perturbation is enough to

 ISBN: 978-84-617-8694-7

destroy periodicity, that is, nonlinearity will finish with most of the periodic orbits of the unperturbed system, but some of them may persist. The Poincaré–Lindstedt technique is used to find those periodic solutions by expanding the solution of the system in the form

$$\begin{aligned} x(t) &= x_0(T) + \epsilon\, x_1(T) + \epsilon^2\, x_2(T) + \cdots , \\ y(t) &= y_0(T) + \epsilon\, y_1(T) + \epsilon^2\, y_2(T) + \cdots , \\ z(t) &= z_0(T) + \epsilon\, z_1(T) + \epsilon^2\, z_2(T) + \cdots , \end{aligned} \tag{3}$$

where $x_\nu = x_\nu(T)$, $y_\nu = y_\nu(T)$ and $z_\nu = z_\nu(T)$ are 2π–periodic in T, and $T = \omega t$ is the stretched time variable, with

$$\omega = 1 + \epsilon\omega_1 + \epsilon^2\omega_2 + \cdots , \tag{4}$$

being ω_ν real constants. Thus, the nonlinear period is $2\pi/\omega$.

To apply this technique, one has to start by rewriting (1) in terms of the new independent variable T, to obtain

$$\begin{aligned} \omega x' + \alpha_{12} y &= \epsilon \sum_{0\le q\le M} \sum_{0\le \nu_1\le q} \sum_{0\le \nu_2\le \nu_1} f_{1,\nu_2,q-\nu_1,\nu_1-\nu_2}\, x^{\nu_2} y^{q-\nu_1} z^{\nu_1-\nu_2} , \\ \omega y' - \alpha_{21} x + \alpha_{23} z &= \epsilon \sum_{0\le q\le M} \sum_{0\le \nu_1\le q} \sum_{0\le \nu_2\le \nu_1} f_{2,\nu_2,q-\nu_1,\nu_1-\nu_2}\, x^{\nu_2} y^{q-\nu_1} z^{\nu_1-\nu_2} , \\ \omega z' - \alpha_{32} y &= \epsilon \sum_{0\le q\le M} \sum_{0\le \nu_1\le q} \sum_{0\le \nu_2\le \nu_1} f_{3,\nu_2,q-\nu_1,\nu_1-\nu_2}\, x^{\nu_2} y^{q-\nu_1} z^{\nu_1-\nu_2} . \end{aligned} \tag{5}$$

Here, $\dot{}$ stands for d/dt and $'$ for d/dT. If expansions (3) and (4) are substituted into (5), and terms in equal powers of ϵ are collected, we get an equation for each order of the approximation in the expansions (3). In order to simplify the expression of these equations, let us introduce here the following notation: S_ν denotes the ν–th order coefficient of the expansion of S, so that

$$S = S_0 + \epsilon S_1 + \epsilon^2 S_2 + \cdots .$$

For instance, if $S = x^2$, then $(x^2)_0 = x_0 x_0$, $(x^2)_1 = 2x_0 x_1$, and in general, $(x^2)_q = \sum_{0\le\nu\le q} x_\nu x_{q-\nu}$. This notation eases the way to express the formulae for the computation of the coefficients of the expansion of the solution at any order.

The solution to (1) is constructed from the order zero, which corresponds with the unperturbed problem, and can be written as

$$\begin{aligned} \omega_0 x_0' + \alpha_{12} y_0 &= 0 , \\ \omega_0 y_0' - \alpha_{21} x_0 + \alpha_{23} z_0 &= 0 , \\ \omega_0 z_0' - \alpha_{32} y_0 &= 0 . \end{aligned} \tag{6}$$

 ISBN: 978-84-617-8694-7

The first order system is given by

$$\begin{aligned}
\omega_0 x_1' + \alpha_{12} y_1 &= \sum_{0\le q\le M}\ \sum_{0\le\nu_1\le q}\ \sum_{0\le\nu_2\le\nu_1} f_{1,\nu_2,q-\nu_1,\nu_1-\nu_2}\, x_0^{\nu_2} y_0^{q-\nu_1} z_0^{\nu_1-\nu_2} - \omega_1 x_0' \,, \\
\omega_0 y_1' - \alpha_{21} x_1 + \alpha_{23} z_1 &= \sum_{0\le q\le M}\ \sum_{0\le\nu_1\le q}\ \sum_{0\le\nu_2\le\nu_1} f_{2,\nu_2,q-\nu_1,\nu_1-\nu_2}\, x_0^{\nu_2} y_0^{q-\nu_1} z_0^{\nu_1-\nu_2} - \omega_1 y_0' \,, \\
\omega_0 z_1' - \alpha_{32} y_1 &= \sum_{0\le q\le M}\ \sum_{0\le\nu_1\le q}\ \sum_{0\le\nu_2\le\nu_1} f_{3,\nu_2,q-\nu_1,\nu_1-\nu_2}\, x_0^{\nu_2} y_0^{q-\nu_1} z_0^{\nu_1-\nu_2} - \omega_1 z_0' \,.
\end{aligned} \tag{7}$$

The order Q of the expansion is obtained by solving the system

$$\begin{aligned}
\omega_0 x_Q' + \alpha_{12} y_Q &= \sum_{0\le q\le M}\ \sum_{0\le\nu_1\le q}\ \sum_{0\le\nu_2\le\nu_1} f_{1,\nu_2,q-\nu_1,\nu_1-\nu_2}\, \left[x^{\nu_2} y^{q-\nu_1} z^{\nu_1-\nu_2}\right]_{Q-1} - \\
&\quad - \sum_{1\le\nu\le Q-1} x_\nu' \omega_{Q-\nu} - \omega_Q x_0' \,, \\
\omega_0 y_Q' - \alpha_{21} x_Q + \alpha_{23} z_Q &= \sum_{0\le q\le M}\ \sum_{0\le\nu_1\le q}\ \sum_{0\le\nu_2\le\nu_1} f_{2,\nu_2,q-\nu_1,\nu_1-\nu_2}\, \left[x^{\nu_2} y^{q-\nu_1} z^{\nu_1-\nu_2}\right]_{Q-1} - \\
&\quad - \sum_{1\le\nu\le Q-1} y_\nu' \omega_{Q-\nu} - \omega_Q y_0' \,, \\
\omega_0 z_Q' - \alpha_{32} y_Q &= \sum_{0\le q\le M}\ \sum_{0\le\nu_1\le q}\ \sum_{0\le\nu_2\le\nu_1} f_{3,\nu_2,q-\nu_1,\nu_1-\nu_2}\, \left[x^{\nu_2} y^{q-\nu_1} z^{\nu_1-\nu_2}\right]_{Q-1} - \\
&\quad - \sum_{1\le\nu\le Q-1} z_\nu' \omega_{Q-\nu} - \omega_Q z_0' \,.
\end{aligned} \tag{8}$$

At each order p of the perturbation method, one has to calculate x_p, y_p, z_p and ω_p from the equation above, but also x_p', y_p', z_p' and the collection of products $(x^{\nu_1} y^{\nu_2} z^{\nu_3})_p$ for each $\nu_1, \nu_2, \nu_3 \in \mathbb{Z}$ such that $0 \le \nu_1, \nu_2, \nu_3 \le M$, in order to compute the right–hand side of equation (8) for the $(p+1)$–th order of the perturbation method. At the p–th order of the Poincaré–Lindstedt approximation, one first fits the value of ω_p to assure that no secular terms exist, expressing it as a function of some constants which depend on the initial conditions of the problem. This is done by applying the proposition 1, given below. Once ω_p has been obtained, x_p, y_p and z_p can be computed by solving the system (8).

 ISBN: 978-84-617-8694-7

Proposition 1 *The differential system*

$$\begin{aligned}
\omega_0 x' + \alpha_{12} y &= A\cos T + B\sin T + \sum_{n>1}(A_n \cos nT + B_n \sin nT)\,, \\
\omega_0 y' - \alpha_{21} x + \alpha_{23} z &= C\cos T + D\sin T + \sum_{n>1}(C_n \cos nT + D_n \sin nT)\,, \\
\omega_0 z' - \alpha_{32} y &= E\cos T + F\sin T + \sum_{n>1}(E_n \cos nT + F_n \sin nT)\,,
\end{aligned} \tag{9}$$

where

$$\omega_0 = \sqrt{\alpha_{12}\alpha_{21} + \alpha_{23}\alpha_{32}}\,,$$

has no secular terms if

$$\begin{aligned}
\omega_0 C + \alpha_{23} F - \alpha_{21} B &= 0\,, \\
\omega_0 D - \alpha_{23} E + \alpha_{21} A &= 0\,.
\end{aligned}$$

Proof. Let us take, without loss of generality,

$$A_n = B_n = C_n = D_n = D_n = E_n = F_n = 0\,,$$

for any $n > 1$. The solution $x(t)$ to (9) can be written as

$$\begin{aligned}
x(t) &= \frac{1}{\omega_0^2}\left(\alpha_{12}\omega_0 k_3 - \frac{\alpha_{12}\alpha_{21} + 2\alpha_{23}\alpha_{32}}{2\omega_0} B - \frac{\alpha_{12}}{2} C - \frac{\alpha_{12}\alpha_{23}}{2\omega_0} F\right)\cos T + \\
&+ \frac{1}{\omega_0^2}\left(\frac{\alpha_{12}\alpha_{21}}{2\omega_0} A + \frac{\alpha_{12}}{2} D - \frac{\alpha_{12}\alpha_{23}}{2\omega_0} E\right) T\cos T - \\
&- \frac{1}{\omega_0^2}\left(\alpha_{12}\omega_0 k_2 + \alpha_{12} D - \frac{\alpha_{23}\alpha_{32}}{\omega_0} A - \frac{\alpha_{12}\alpha_{23}}{\omega_0} E\right)\sin T - \\
&- \frac{1}{\omega_0^2}\left(-\frac{\alpha_{12}\alpha_{21}}{2\omega_0} B + \frac{\alpha_{12}}{2} C + \frac{\alpha_{12}\alpha_{23}}{2\omega_0} F\right) T\sin T + k_1\,,
\end{aligned}$$

where k_1, k_2 and k_3 are integration contants depending on the initial conditions of the problem. The condition to make the secular term $T\cos T$ dissapear is given by

$$\frac{\alpha_{12}}{2} D - \frac{\alpha_{12}\alpha_{23}}{2\omega_0} E + \frac{\alpha_{12}\alpha_{21}}{2\omega_0} A = 0\,,$$

that is,

$$\omega_0 D - \alpha_{23} E + \alpha_{21} A = 0\,.$$

In the same way, the condition to make the secular term $T\sin T$ dissapear is given by

$$\frac{\alpha_{12}}{2} C - \frac{\alpha_{12}\alpha_{21}}{2\omega_0} B + \frac{\alpha_{12}\alpha_{23}}{2\omega_0} F = 0\,,$$

 ISBN: 978-84-617-8694-7

that is,

$$\omega_0 C + \alpha_{23} F - \alpha_{21} B = 0\,.$$

This procedure can be carried out also for $y(t)$ and $z(t)$, obtaining the same set of conditions. ■

3 Application to Lotka–Volterra Systems

In 1925, Lotka ([7]) proposed a differential equation model to describe the population dynamics of two interacting species, a predator and its prey. A species $x(t)$ serves as food to another species $y(t)$, so that, in this sense, $x(t)$ becomes transformed into $y(t)$. The formulation of Volterra was as follows: A species $y(t)$ feeds on a species $x(t)$, which, in turn feeds on some source presented in such large excess that the mass of this source may be considered constant during the period of time under consideration. Then, under this assumption, we have that the rate of increase of $x(t)$ per unit of time is equal to the difference between the mass of newly formed of $x(t)$ per unit of time and the mass of $x(t)$ destroyed by $y(t)$ per unit of time. The Lotka–Volterra model for two species consists of the following differential equations:

$$\begin{aligned} \dot{x} &= x(r_1 - a_{12}x_2)\,, \\ \dot{y} &= y(-r_2 + a_{21}x_2)\,. \end{aligned} \tag{10}$$

Here, $y(t)$ and $x(t)$ represent, respectively, the predator population and the prey population as functions of time. The parameters $r_1, r_2, a_{12}, a_{21} > 0$ are interpreted as follows: r_1 represents the natural growth rate of the prey in the absence of predators, a_{12} represents the effect of predation on the prey, r_2 is the natural death rate of the predator in the absence of prey, and a_{21} the efficiency and propagation rate of the predator in the presence of prey.

In this section, in order to illustrate the effectiveness of the perturbation procedure, we will focus our interest in a linear three species food chain where the lowest level prey, x is preyed upon by a mid–level species, y who, in turn, is preyed upon by a top level predator z. The equations of the system are

$$\begin{aligned} \dot{x} &= x(r_1 - a_{12}y)\,, \\ \dot{y} &= y(-r_2 + a_{21}x - a_{23}z)\,, \\ \dot{z} &= z(-r_3 + a_{32}y)\,. \end{aligned} \tag{11}$$

Here, $r_1, r_2, r_3 > 0$ and $a_{12}, a_{21}, a_{23}, a_{32} > 0$. The parameters r_1, r_2, a_{12} and a_{21} have the same meaning as in the Lotka–Volterra equations for two species, a_{23} represents the effect of predation on species y by species z, r_3 the natural death rate of the predator z in the absence of prey, and a_{32} stands for the efficiency and propagation rate of the predator z in

 ISBN: 978-84-617-8694-7

the presence of prey. Since populations are non–negative, we will restrict our attention to the non–negative octant $\Omega = \{(x, y, z) : \ x \geq 0\,,\ y \geq 0\,,\ z \geq 0\} \subset \mathbb{R}^3$, and the positive octant $\Omega^+ = \{(x, y, z) : \ x > 0\,,\ y > 0\,,\ z > 0\} \subset \mathbb{R}^3$. Each coordinate plane is invariant with respect to (11) (see for instance [3]).

Let us study now the equilibrium points of (11). This system has equilibrium points in the interior of Ω^+ if

$$\begin{aligned} r_1 - a_{12}y &= 0\,, \\ -r_2 + a_{21}x - a_{23}z &= 0\,, \\ -r_3 + a_{32}y &= 0\,. \end{aligned}$$

By solving these equations, we get that

$$y = \frac{1}{a_{12}} r_1\,, \quad r_3 a_{12} = r_1 a_{32}\,,$$

and

$$x = \frac{1}{a_{21}} r_2 + \frac{a_{23}}{a_{21}} z\,.$$

Thus, it results straightforward to find that there is ray of equilibrium points of the form

$$P_\lambda \left(\frac{r_2}{a_{21}} + \frac{a_{23}}{a_{21}}\lambda, \frac{r_1}{a_{12}}, \lambda \right)\,,$$

with $\lambda \in \mathbb{R}$, in the case $r_3 a_{12} = r_1 a_{32}$. The Jacobian matrix of the system in any of these equilibrium points is given by

$$A(P_\lambda) = \begin{pmatrix} 0 & -a_{12}(r_2 + a_{23}\lambda)/a_{21} & 0 \\ a_{21}r_1/a_{12} & 0 & -a_{23}r_1/a_{12} \\ 0 & a_{32}\lambda & 0 \end{pmatrix}\,,$$

with eigenvalues

$$\begin{aligned} \lambda_1 &= 0\,, \\ \lambda_2 &= +\frac{1}{a_{12}} \sqrt{-a_{12}r_1 \left(a_{23}a_{32}\lambda + r_2 a_{12} + a_{23}a_{12}\lambda\right)}\,, \\ \lambda_3 &= -\frac{1}{a_{12}} \sqrt{-a_{12}r_1 \left(a_{23}a_{32}\lambda + r_2 a_{12} + a_{23}a_{12}\lambda\right)}\,. \end{aligned}$$

As the three eigenvalues have zero real part, each such equilibrium point has a three–dimensional center manifold, which does not help us determine the dynamics near these fixed points. The numerical exploration of the solutions suggests that the system contains invariant surfaces. In particular, it can be proved that the surfaces $z = Kx^{-r_3/r_1}$ are

 ISBN: 978-84-617-8694-7

invariant in Ω^+. These surfaces are filled with periodic orbits enclosing the ray of quilibrium points P_λ (see [3]).

In order to determine these periodic orbits, we perturb the system around the equilibrium point $(r_2/a_{21} + \lambda a_{23}/a_{21}, r_1/a_{12}, \lambda)$,

$$\begin{aligned} x(t) &= \frac{r_2}{a_{21}} + \frac{a_{23}}{a_{21}}\lambda + \epsilon X(t)\,, \\ y(t) &= \frac{r_1}{a_{12}} + \epsilon Y(t)\,, \\ z(t) &= \lambda + \epsilon Z(t)\,, \end{aligned}$$

to obtain the perturbed system

$$\begin{aligned} \dot{X} &= -\frac{a_{12}}{a_{21}}\left(r_2 + a_{23}\lambda\right) Y - \epsilon\, a_{12} XY\,, \\ \dot{Y} &= r_1\frac{a_{21}}{a_{12}} X - r_1\frac{a_{23}}{a_{12}} Z + \epsilon\left(a_{21}XY - a_{23}YZ\right), \\ \dot{Z} &= a_{32}\lambda Z + \epsilon\, a_{32} YZ\,. \end{aligned}$$

The introduction of stretched time variable $T = \omega t$ gives

$$\begin{aligned} \omega x' &= -\frac{a_{12}}{a_{21}}\left(r_2 + a_{23}\lambda\right) y - \epsilon\, a_{12} xy\,, \\ \omega y' &= r_1\frac{a_{21}}{a_{12}} x - r_1\frac{a_{23}}{a_{12}} z + \epsilon\left(a_{21}xy - a_{23}yz\right), \\ \omega z' &= a_{32}\lambda y + \epsilon\, a_{32} yz\,, \end{aligned} \tag{12}$$

after recalling $X = x$, $Y = y$ and $Z = z$ for the sake of simplicity. In order to apply the Poincaré–Lindstedt technique as described in section 2, we expand the solution of the system and ω as expressed in equations (3) and (4),

$$\begin{aligned} x(T) &= x_0(T) + \epsilon x_1(T) + \epsilon^2 x_2(T) + \cdots\,, \\ y(T) &= y_0(T) + \epsilon y_1(T) + \epsilon^2 y_2(T) + \cdots \\ z(T) &= z_0(T) + \epsilon z_1(T) + \epsilon^2 z_2(T) + \cdots \end{aligned}$$

and

$$\omega = \omega_0 + \epsilon\omega_1 + \epsilon^2\omega_2 + \cdots$$

By substituting these expasions into equation (12), and collecting terms in equal powers of ϵ, we get an equation for each order of the approximation. The order zero is given by

$$\begin{aligned} \omega_0 x_0' &= -\frac{a_{12}}{a_{21}}\left(r_2 + a_{23}\lambda\right) y_0\,, \\ \omega_0 y_0' &= r_1\frac{a_{21}}{a_{12}} x_0 - r_1\frac{a_{23}}{a_{12}} z_0\,, \\ \omega_0 z_0' &= a_{32}\lambda y_0\,. \end{aligned} \tag{13}$$

 ISBN: 978-84-617-8694-7

The solution to (13) is

$$\begin{aligned} x_0(T) &= \frac{a_{23}}{a_{21}}A + \frac{(r_2+a_{23}\lambda)a_{12}}{a_{21}\omega_0}B\cos T - \frac{(r_2+a_{23}\lambda)a_{12}}{a_{21}\omega_0}C\sin T\,, \\ y_0(T) &= C\cos T + B\sin T\,, \\ z_0(T) &= A - \frac{a_{32}\lambda}{\omega_0}B\cos T + \frac{a_{32}\lambda}{\omega_0}C\sin T\,, \end{aligned} \tag{14}$$

where A, B and C depend on the initial conditions of the solution. The derivatives of x_0, y_0 and z_0 are

$$\begin{aligned} x_0'(T) &= -\frac{(r_2+a_{23}\lambda)a_{12}}{a_{21}\omega_0}C\cos T - \frac{(r_2+a_{23}\lambda)a_{12}}{a_{21}\omega_0}B\sin T\,, \\ y_0'(T) &= B\cos T - C\sin T\,, \\ z_0'(T) &= \frac{a_{32}\lambda}{\omega_0}C\cos T + \frac{a_{32}\lambda}{\omega_0}B\sin T\,. \end{aligned} \tag{15}$$

The first order system is given by

$$\begin{aligned} \omega_0 x_1' + \frac{a_{12}}{a_{21}}\left(r_2+a_{23}\lambda\right)y_1 &= -a_{12}x_0y_0 - \omega_1 x_0'\,, \\ \omega_0 y_1' - r_1\frac{a_{21}}{a_{12}}x_1 + r_1\frac{a_{23}}{a_{12}}z_1 &= a_{21}x_0y_0 - a_{23}y_0z_0 - \omega_1 y_0'\,, \\ \omega_0 z_1' - a_{32}\lambda y_1 &= a_{32}y_0z_0 - \omega_1 z_0'\,. \end{aligned}$$

The substitution of equations (14) and (15) in the first order system yields

$$\begin{aligned} &\omega_0 x_1' + \frac{a_{12}}{a_{21}}\left(r_2+a_{23}\lambda\right)y_1 = \\ &= \frac{a_{12}}{a_{21}\omega_0}C\left(\omega_1(r_2+a_{23}\lambda) - \omega_0 a_{23}A\right)\cos T + \\ &\quad + \frac{a_{12}}{a_{21}\omega_0}B\left(\omega_1(r_2+a_{23}\lambda) - \omega_0 a_{23}A\right)\sin T - \\ &\quad - \frac{a_{12}^2BC}{a_{21}\omega_0}(r_2+a_{23}\lambda)\cos 2T + \frac{a_{12}^2}{2a_{21}\omega_0}\left((C^2-B^2)(r_2+a_{23}\lambda)\right)\sin 2T\,, \\ &\omega_0 y_1' - \frac{a_{21}r_1}{a_{12}}x_1 + \frac{a_{23}r_1}{a_{12}}z_1 = \\ &= -\omega_1 B\cos T + \omega_1 C\sin T + \frac{BC}{\omega_0}\left(a_{12}(r_2+a_{23}\lambda) + a_{23}a_{32}\lambda\right)\cos 2T + \\ &\quad + \frac{B^2-C^2}{2\omega_0}\left(a_{12}r_2 + (a_{12}+a_{32})a_{23}\lambda\right)\sin 2T\,, \\ &\omega_0 z_1' - a_{32}\lambda y_1 = \\ &= \frac{a_{32}}{\omega_0}C(A\omega_0 - \omega_1\lambda)\cos T + \frac{a_{32}}{\omega_0}B(A\omega_0 - \omega_1\lambda)\sin T - \\ &\quad - \frac{a_{32}^2\lambda}{\omega_0}BC\cos 2T + \frac{a_{32}^2\lambda}{2\omega_0}(C^2-B^2)\sin 2T\,. \end{aligned} \tag{16}$$

 ISBN: 978-84-617-8694-7

The application of proposition 1 to the system (16) gives

$$\omega_1 = -\frac{r_1 a_{23} A\,(a_{12} - a_{32})\,\omega_0}{a_{12}(\omega_0^2 + r_1 r_2) + \lambda r_1 a_{23}(a_{12} + a_{32})}\,.$$

From the analysis of this example, it follows that the application of the Poincaré–Lindstedt method involves working with expressions and procedures that should be automated to avoid mistakes in the algebraic manipulation of those developments.

4 Computation of a periodic orbit in a three–dimensional Lotka–Volterra model

Let us consider the system

$$\begin{aligned} \dot{x} &= x(1-y)\,, \\ \dot{y} &= y(-2+x-z)\,, \\ \dot{z} &= z(-1+y)\,. \end{aligned} \tag{17}$$

The equilibrium points of the system are $(2+\lambda, 1, \lambda)$, $\lambda \in \mathbb{R}^+$. Let us take, for instance, $\lambda = 1/2$, and the equilibrium point $(5/2, 1, 1/2)$. Following the procedure employed in section 3, equation (17) is transformed into

$$\begin{aligned} \omega x' &= -\frac{5}{2}y - \epsilon yz\,, \\ \omega y' &= x - z + \epsilon(xy - yz)\,, \\ \omega z' &= \frac{1}{2}y + \epsilon yz\,. \end{aligned} \tag{18}$$

We compute ω_0 from the unperturbed system ($\epsilon = 0$), $\omega_0 = \sqrt{3}$. By substituting (3) and (4) into equation (18), and collecting terms in equal powers of ϵ, we get an equation for each order of the series solution. In Figure 1, we compare between the numerical solution to equation (17) with initial conditions $x(0) = 2.864670867$, $y(0) = 1.187254264$, $z(0) = 0.669094392$, obtained by using a fourth order Runge–Kutta method, and a fifth order approximation to the solution computed through the Poincaré–Lindstedt method.

5 Conclusion

In this paper, a symbolic algorithm for a general application of the Poincaré–Lindstedt method for the computation of periodic solutions in three–dimensional differential systems of first order has been presented. We have adapted the standard method to three–dimensional nonlinear systems, giving the sufficient conditions to avoid the occurrence of secular terms in the perturbation series solution. The algorithm has been used to compute periodic solutions in a three–dimensional Lotka–Volterra system modeling a chain food interaction.

 ISBN: 978-84-617-8694-7

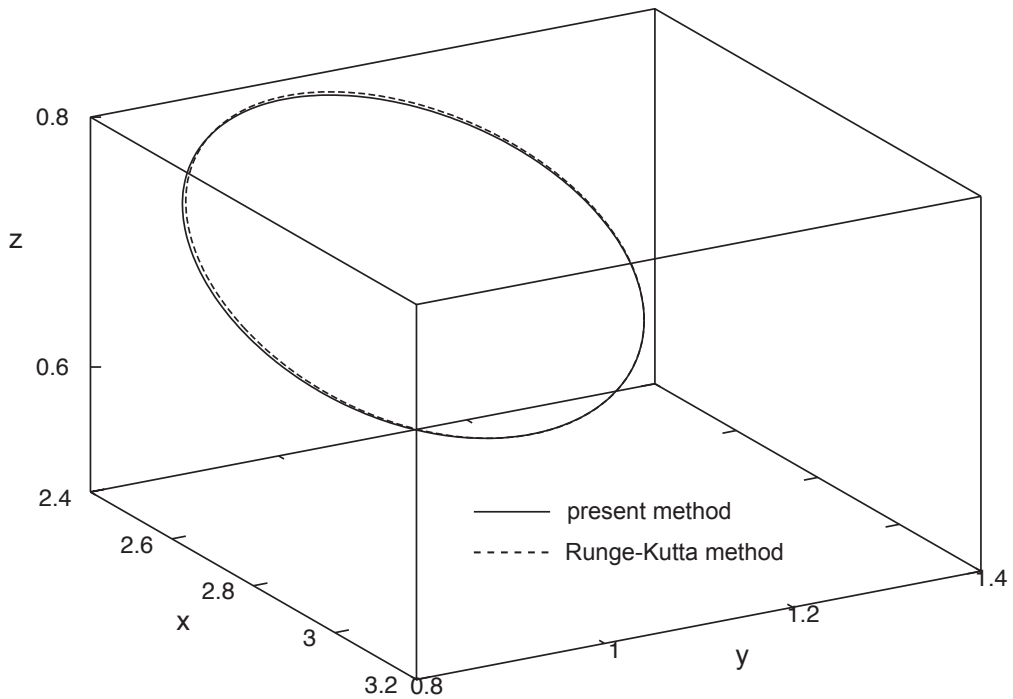


Figure 1: Comparison between the approximate solution to equation (17) with initial conditions $x(0) = 2.864670867$, $y(0) = 1.187254264$, $z(0) = 0.669094392$, obtained by using a fourth order Runge–Kutta method and a fifth order approximation to the solution computed through the Poincaré–Lindstedt method.

 ISBN: 978-84-617-8694-7

References

[1] S. Bhattacharjee and J. K. Bhattacharjee, *Lindstedt Poincaré technique applied to molecular potentials*, J. Math. Chem. **50** (2012) 1398–1410.

[2] A. Buonomo, *The periodic solution of van der Pol equation*, SIAM J. Appl. Math. **59 (1)** (1998) 156–171.

[3] E. Chauvet, J. E. Paullet, J. P. Previte and Z. Walls, *A Lotka–Volterra Three-Species Food Chain*, Mathematics Magazine **75 (4)** (2002) 243–255.

[4] S. H. Chen and Y. K. Cheung, *An elliptic Lindstedt–Poincaré method for certain strongly non–linear oscillators*, Nonlinear Dyn. **12** (1997) 199–213.

[5] A. Dena, M. Rodríguez, S. Serrano and R. Barrio, *High–Precision Continuation of Periodic Orbits*, Abstract and Applied Analysis (2012) 12 pages.

[6] H. Hu, Z. G. Xiong ZG, *Comparison of two Lindstedt–Poincaré–type perturbation methods*, J. Sound Vibr. **278** (2004) 437–444.

[7] A. J. Lotka, *Elements of Physical Biology*, Baltimore, Williams and Wilkins Company, 1925.

[8] Juan F. Navarro, *A Symbolic Algorithm for the Computation of Periodic Orbits in Non–Linear Differential Systems*, Journal of Advances in Applied Mathematics **1 (3)** (2016) 160–174.

[9] T. Penati and S. Flach, *Tail resonances of Fermi–Pasta–Ulam q–breathers ans their impact on the pathway to equipartition*, Chaos **17 (2)** (2007) 023102–023116.

[10] J. I. Ramos, *On Lindstedt–Poincaré techniques for the quintic Duffing equation*, Appl. Math. Comput. **193** (2007) 303–310.

[11] R. Scuflaire, *Periodic orbits in analytical planar galactic potentials*, Celest. Mech. **71** (1991) 203–228.

[12] R. Scuflaire, *Stability of axial orbits in analytic galactic potentials*, Celest. Mech. **61** (1995) 261–285.

[13] D. Venu Gopala Rao, P. Sri Hari Krishna, *An Extensive Study of Approximating the Periodic Solutions of the Prey–Predator System*, Applied Mathematical Sciences **4 (58)** (2010) 2851–2864.

[14] D. Viswanath, *The Lindstedt–Poincaré technique as an algorithm for computing periodic orbits*, *SIAM Rev.* **43 D** (2001) 478–495.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

Two dimensional approximation of Jackson type

M.A. Navascués[1] and M.V. Sebastián[2]

[1] *Departamento de Matemática Aplicada, Universidad de Zaragoza*

[2] *Centro Universitario de la Defensa, Academia General Militar de Zaragoza*

emails: manavas@unizar.es, msebasti@unizar.es

Abstract

Some procedures of approximation of three-dimensional data on a grid are summarized. The first part proposes a generalization of a discrete periodic approximation defined by Dunham Jackson. The functions used have the advantage of owning an analytical explicit expression in terms of the samples (specific values) of the original function or the data. In the second part we deal with a continuous approximation function for the same problem, defined through an integral. Some results of the rate of convergence and bounds of the approximation error are presented, with the single hypothesis of continuity of the original function.

Key words: Trigonometric approximation, trigonometric interpolation, smoothing, surface fitting

MSC 2000: 42A10 42A15 42A05 65D10 65D05

1 Introduction

In 1885 Weierstrass proved that every continuous function defined in a compact interval is approximated by a polynomial with arbitrary precision. This fact motivated the scientists of later generations and, specially, Dunham Jackson, American mathematician who published his work in the first decades of the twentieth century. This author wrote several books ([7], [9]) and numerous articles (see for instance [4], [5], [6], [8]) on polynomial and trigonometric approximation of continuous and discontinuous functions. His writings led to transcendental mathematical outcomes as for instance the inequalities named after him, which describe the degree of approximation of a continuous function by means of polynomials (algebraic and trigonometric). For instance, if

$$d_n^*(f) = d(f, \mathcal{P}_n)$$

 ISBN: 978-84-617-8694-7

represents the uniform distance from f to the space of polynomials of degree (order) at most n, the following inequalities hold ([2]).

Theorem 1.1. *For all f continuous with period 2π ($f \in \mathcal{C}[-\pi,\pi]$) which satisfy a Lipschitz condition $|f(x)-f(y)| \leq \lambda|x-y|$,*

$$d_n^*(f) \leq \frac{\pi\lambda}{2(n+1)}.$$

Theorem 1.2. *For all f continuous with period 2π ($f \in \mathcal{C}[-\pi,\pi]$)*

$$d_n^*(f) \leq \omega\left(\frac{\pi}{(n+1)}\right),$$

where $\omega(\delta)$ is the modulus of continuity of f.

Hereafter we evoke the procedures used by Jackson for periodic functions, generalizing the approximants in the two dimensional case. One of the formulae proposed is an explicit model in terms of the data on a two-dimensional grid. Some error bounds are deduced for more general exponents of the basic functions, and the convergence is proved with the single hypothesis of continuity on a compact interval, unlike the standard trigonometric interpolation.

2 Two-dimensional discrete approximant of Jackson type

In this Section we consider a trigonometric fitting mapping for functions or data defined on a grid on the two-dimensional interval $[-\pi,\pi]\times[-\pi,\pi]$, assuming periodicity in both variables. They are inspired (generalizing them) in approximations of Jackson, and they are defined explicitly in terms of the data (function values). Due to this fact, we name them discrete approximants, in order to distinguish them from other functions defined in later Section.

Let us consider a set of three-dimensional data on a grid:

$$\{(x_i, y_j, f(x_i,y_j)) : i = 1,2,\ldots,m; j = 1,2,\ldots,n\},$$

where $x_{i+1}-x_i = \pi/m$, $i = 1,2,\ldots,2m-1$; $y_{j+1}-y_j = \pi/n$, $j = 1,2,\ldots,2n-1$.

Let T^1 represent the unit circle and let us consider $f \in \mathcal{C}(T^1\times T^1)$ (continuous and periodic with period 2π in both variables), and an exponent $\gamma > 0$ ([6]). The approximation of f on a grid is defined by the expression:

$$\mathcal{J}_{mn\gamma}(f)(x,y) = K_{mn\gamma}(x,y)\sum_{i=1}^{2m}\sum_{j=1}^{2n} f(x_i,y_j)\left|\frac{\sin\left(\frac{1}{2}m(x_i-x)\right)}{m\sin\left(\frac{1}{2}(x_i-x)\right)}\right|^{\gamma}\left|\frac{\sin\left(\frac{1}{2}n(y_j-y)\right)}{n\sin\left(\frac{1}{2}(y_j-y)\right)}\right|^{\gamma},$$

 ISBN: 978-84-617-8694-7

where

$$K_{mn\gamma}^{-1}(x,y)=\sum_{i=1}^{2m}\sum_{j=1}^{2n}\left|\frac{\sin\left(\frac{1}{2}m(x_i-x)\right)}{m\sin\left(\frac{1}{2}(x_i-x)\right)}\right|^{\gamma}\left|\frac{\sin\left(\frac{1}{2}n(y_j-y)\right)}{n\sin\left(\frac{1}{2}(y_j-y)\right)}\right|^{\gamma}.$$

In the case $\gamma=4$, one has

$$K_{mn4}^{-1}(x,y)=\sum_{i=1}^{2m}\left(\frac{\sin\left(\frac{1}{2}m(x_i-x)\right)}{m\sin\left(\frac{1}{2}(x_i-x)\right)}\right)^4\sum_{j=1}^{2n}\left(\frac{\sin\left(\frac{1}{2}n(y_j-y)\right)}{n\sin\left(\frac{1}{2}(y_j-y)\right)}\right)^4=H_m^{-1}H_n^{-1}.$$

H_m (or H_n) is a constant such that ([6]) for all m:

$$1/2\leq H_m<3/4.$$

Consequently, K_{mn4} is a constant such that for all m, n :

$$1/4\leq K_{mn4}<9/16.$$

$\mathcal{J}_{mn4}(f)$ is a trigonometric polynomial of order at most $2(m-1)$ in x and $2(n-1)$ in y. If γ is a multiple of 2, the function is a trigonometric rational.

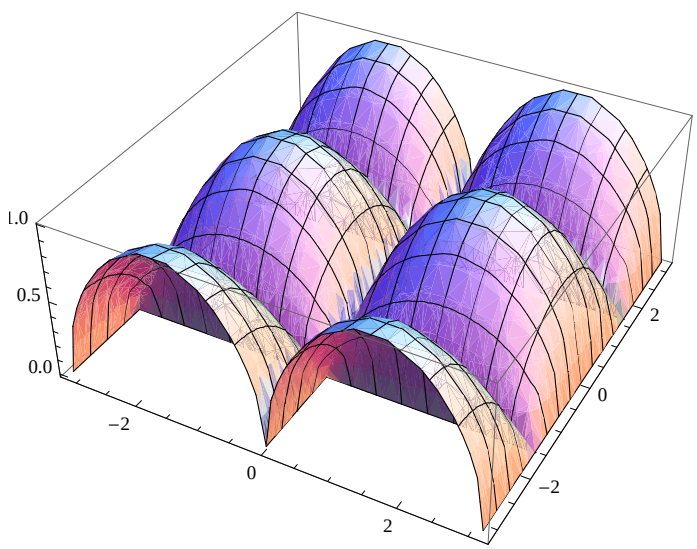


Figure 1: Graph of the function $f(x,y)=\sqrt{|\sin(x)\cos(y)|}$ in the square $[-\pi,\pi]\times[-\pi,\pi]$.

Lemma 2.1. *For all* $m=1,2,\ldots;\ \gamma>0,$ *and* $v\in\mathbb{R}$:

$$\left|\frac{\sin(mv)}{m\sin(v)}\right|^{\gamma}\leq 1.$$

 ISBN: 978-84-617-8694-7

Theorem 2.2. *For any continuous function* $f \in \mathcal{C}(T^1 \times T^1)$, $\gamma > 2$ *and* $(x, y) \in T^1 \times T^1$,

$$|\mathcal{J}_{mn\gamma}(f)(x,y) - f(x,y)| \leq \omega\left(\frac{\pi}{4m} + \frac{\pi}{4n}\right) C'(\gamma) \leq \omega\left(\frac{1}{m} + \frac{1}{n}\right) C(\gamma),$$

where $C'(\gamma)$, $C(\gamma)$ *do not depend on* m, n.

Remark 2.3. If $\gamma > 2$, the rate of convergence of the error when the partition is indefinitely refined is that of the modulus $\omega(\frac{1}{m} + \frac{1}{n})$ (it does not depend on the exponent γ).

Remark 2.4. For any continuous function $f \in \mathcal{C}(T^1 \times T^1)$ and $\gamma > 2$, the approximant $\mathcal{J}_{mn\gamma}(f)$ converges uniformly to f as m and n tend to infinity.

Remark 2.5. The uniform convergence on the compact interval implies the convergence in the p-norm for any $1 \leq p < \infty$.

Remark 2.6. If f satisfies a Lipschitz condition of order q $(0 < q \leq 1)$, then there is a wider range of convergence values: $\gamma > 1 + q$. The case $\gamma = 2$ is within this interval.

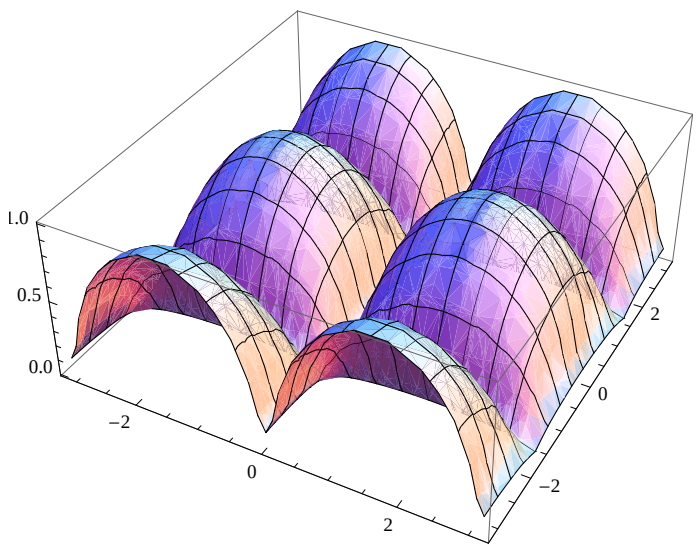


Figure 2: Graph of the discrete approximant of $f(x, y) = \sqrt{|\sin(x)\cos(y)|}$ for $m = 10$, $n = 10$ and $\gamma = 5$ in the square $[-\pi, \pi] \times [-\pi, \pi]$.

Let us denote by

$$\|\mathcal{J}_{mn\gamma}\|$$

the norm of the operator $\mathcal{J}_{mn\gamma}$ with respect to the uniform (supremum) norm $\|\cdot\|_\infty$ in $\mathcal{C}(T^1 \times T^1)$.

 ISBN: 978-84-617-8694-7

Theorem 2.7. *For $m, n \in \mathbb{N}$ and $\gamma > 2$, $\|\mathcal{J}_{mn\gamma}\| = 1$.*

Remark 2.8. The operator $\mathcal{J}_{mn\gamma}$ is linear and bounded. It does not amplify the errors in the z-values since let $f, \widetilde{f}$ be the functions corresponding to data (x_i, y_j, z_{ij}) and $(x_i, y_j, \widetilde{z}_{ij})$. Considering the linearity and the norm of the approximation operator:

$$\|\mathcal{J}_{mn\gamma}(f) - \mathcal{J}_{mn\gamma}(\widetilde{f})\|_\infty \leq \|f - \widetilde{f}\|_\infty.$$

3 Two dimensional continuous approximants of Jackson type

We consider here the approximant of a continuous and periodic function $f \in \mathcal{C}(2\pi)$. This model was proposed by Jackson in the article "On approximation by trigonometric sums and polynomials" ([5]). We consider a more general case, where the exponent 4 appearing in the paper is replaced by any positive exponent $\gamma > 0$:

$$\mathcal{F}_{m\gamma}(f)(x) = h_{m\gamma} \int_{-\pi/2}^{\pi/2} f(x+2u) \left| \frac{\sin(mu)}{m \sin(u)} \right|^\gamma du,$$

where

$$h_{m\gamma}^{-1} = \int_{-\pi/2}^{\pi/2} \left| \frac{\sin(mu)}{m \sin(u)} \right|^\gamma du. \tag{3.1}$$

If $\gamma = 4$, $\mathcal{F}_{m\gamma}(f)$ is a trigonometric polynomial of order at most $2(m-1)$. In the general case the kernels

$$\left| \frac{\sin(mu)}{m \sin(u)} \right|^\gamma$$

have an "order" of $\gamma(m-1)/2$ (although we admit here non-integer values of the exponent). If $\gamma = 2q$, where $q \in \mathbb{N}$, $\mathcal{F}_{m\gamma}(f)$ is a trigonometric polynomial or order at most $q(m-1)$.

Lemma 3.1. *If $p < -1$ and $\gamma > -(p+1)$ then*

$$K_{p\gamma} = \int_0^{+\infty} u^p |\sin(u)|^\gamma du \leq \frac{1}{(p+\gamma+1)} - \frac{1}{(p+1)}.$$

If $\gamma > 2$,

$$i_{m\gamma}^1 = \int_0^{\pi/2} u \left| \frac{\sin(mu)}{m \sin(u)} \right|^\gamma du \leq \frac{1}{m^2} \left(\frac{\pi}{2} \right)^\gamma \left(\frac{1}{2} - \frac{1}{2-\gamma} \right),$$

and if $\gamma > 1$,

$$i_{m\gamma}^0 = \int_0^{\pi/2} \left| \frac{\sin(mu)}{m \sin(u)} \right|^\gamma du \leq \frac{1}{m} \left(\frac{\pi}{2} \right)^\gamma \left(1 - \frac{1}{1-\gamma} \right).$$

 ISBN: 978-84-617-8694-7

If $\gamma > 0$,

$$h_{m\gamma} \leq \frac{m}{2} C_\gamma,$$

where

$$C_\gamma^{-1} = \int_0^{\pi/2} \left(\frac{\sin(u)}{u} \right)^\gamma du. \tag{3.2}$$

Table 1: Approximation errors of function values for different choices of m, n and γ, using the discrete model.

		$\gamma = 4$	$\gamma = 4.5$	$\gamma = 5$
$m = n = 5$	$\lvert \sin(\pi/3)\cos(\pi/6) \rvert$	0.0781323	0.0700892	0.0637317
$m = n = 10$	$\lvert \sin(\pi/3)\cos(\pi/6) \rvert$	0.0214228	0.0213215	0.0226161
$m = n = 5$	$\sqrt{\lvert \sin(\pi/3)\cos(\pi/6) \rvert}$	0.0572732	0.0507959	0.0458076
$m = n = 10$	$\sqrt{\lvert \sin(\pi/3)\cos(\pi/6) \rvert}$	0.0150111	0.0143546	0.0147956
$m = n = 5$	$\lvert \sin(\pi/3)\cos(4\pi/9) \rvert$	-0.124229	-0.120218	-0.117719
$m = n = 10$	$\lvert \sin(\pi/3)\cos(4\pi/9) \rvert$	0.000659161	0.0014973	0.00138685
$m = n = 5$	$\sqrt{\lvert \sin(\pi/3)\cos(4\pi/9) \rvert}$	-0.127787	-0.125574	-0.124311
$m = n = 10$	$\sqrt{\lvert \sin(\pi/3)\cos(4\pi/9) \rvert}$	0.100651	0.100422	0.099083

Let us consider now a biperiodic function $f \in \mathcal{C}(I_x \times I_y)$ or $f \in \mathcal{C}(T^1 \times T^1)$. Let p, q, m, n such that

$$2(p-1) \leq m \leq 2p, \quad 2(q-1) \leq n \leq 2q. \tag{3.3}$$

Let us define the operator

$$\mathcal{G}_{pq\gamma}(f)(x,y) = H_{pq\gamma} \int_{-\pi/2}^{\pi/2} \int_{-\pi/2}^{\pi/2} f(x+2u, y+2v) G_{pq\gamma}(u,v) du dv,$$

where

$$G_{pq\gamma}(u,v) = \left| \frac{\sin(pu)\sin(qv)}{p\sin(u) q \sin(v)} \right|^\gamma,$$

and

$$H_{pq\gamma}^{-1} = \int_{-\pi/2}^{\pi/2} \int_{-\pi/2}^{\pi/2} G_{pq\gamma}(u,v) du dv = h_{p\gamma}^{-1} h_{q\gamma}^{-1},$$

where $h_{p\gamma}^{-1}, h_{q\gamma}^{-1}$ are defined by the expression (3.1).

Hereafter we present bounds for the uniform error committed in this two-dimensional approximation. For it, we present the following Lemmas.

 ISBN: 978-84-617-8694-7

Lemma 3.2. *For p, q, m, n as in (3.3),*

$$u \geq \frac{1}{p} \Rightarrow \omega(u) \leq 4pu\omega\left(\frac{1}{m}\right),$$

$$v \geq \frac{1}{q} \Rightarrow \omega(v) \leq 4qv\omega\left(\frac{1}{n}\right).$$

Lemma 3.3. *Let us consider the following integrals:*

$$C^0_{pq\gamma} = \int_0^{\pi/2}\int_0^{\pi/2} G_{pq\gamma}(u,v)dudv,$$

$$C^{1u}_{pq\gamma} = \int_0^{\pi/2}\int_0^{\pi/2} uG_{pq\gamma}(u,v)dudv,$$

$$C^{1v}_{pq\gamma} = \int_0^{\pi/2}\int_0^{\pi/2} vG_{pq\gamma}(u,v)dudv.$$

Then, for $\gamma > 2$,

$$C^0_{pq\gamma} = H^{-1}_{pq\gamma}/4,$$

$$pC^{1u}_{pq\gamma}H_{pq\gamma} < \frac{1}{2}\left(\frac{\pi}{2}\right)^{\gamma}\left(\frac{1}{2} - \frac{1}{2-\gamma}\right)C_\gamma,$$

$$qC^{1v}_{pq\gamma}H_{pq\gamma} < \frac{1}{2}\left(\frac{\pi}{2}\right)^{\gamma}\left(\frac{1}{2} - \frac{1}{2-\gamma}\right)C_\gamma,$$

where C^{-1}_γ is defined in (3.2).

Theorem 3.4. *If $\gamma > 2$, $f \in \mathcal{C}(T^1 \times T^1)$ and $(x, y) \in T^1 \times T^1$, the two-dimensional approximation satisfies the following inequality*

$$|\mathcal{G}_{pq\gamma}(f)(x,y) - f(x,y)| \leq K_\gamma \omega\left(\frac{1}{m} + \frac{1}{n}\right),$$

where K_γ is a constant depending only on γ.

Remark 3.5. If $\gamma > 2$, the two-dimensional approximant $\mathcal{G}_{pq\gamma}(f)$ is convergent to $f \in \mathcal{C}(T^1 \times T^1)$ as m, n tend to infinity. The rate of convergence is that of $\omega(\frac{1}{m} + \frac{1}{n})$.

Remark 3.6. For functions satisfying a Lipschitz condition of order β, the range of convergence values of γ is extended to $\gamma > \beta + 1$.

The statements about the norm of the operator (Theorem 2.7 and Remark 2.8) are valid for $\mathcal{G}_{pq\gamma}$ as well.

 ISBN: 978-84-617-8694-7

Acknowledgements

This work has been partially supported by the Proyects: CUD-ID: 2013-05 and CUD-ID: 2015-05 of the Centro Universitario de la Defensa de Zaragoza.

References

[1] N. I. Achieser, *Theory of Approximation*, Dover Publ., New York, 1992.

[2] E. W.Cheney, *Approximation Theory*, AMS Chelsea Pub., Providence, 1982.

[3] P. J. Davis, *Interpolation and Approximation*, Dover Publ., New York (2nd. ed.) 1976.

[4] D. Jackson, *On the degree of convergence of the development of a continuous function according to Legendre polynomials*, Trans. Am. Math. Soc. **13** (1912) 305–318.

[5] D. Jackson, *On approximation by trigonometric sums and polynomials*, Trans. Am. Math. Soc. **13** (1912) 491–515.

[6] D. Jackson, *On the accuracy of trigonometric interpolation*, Trans. Am. Math. Soc. **14** (1913) 453–461.

[7] D. Jackson, *Theory of Approximation*, Amer. Math. Soc. Colloquium Publ. **11** (1930).

[8] D. Jackson, *Problems of closest approximation on a two-dimensional region*, Amer. J. Math. **60** (1938) 436–446.

[9] D. Jackson, *Fourier series and orthogonal polynomials*, Carus Math. Mongraph **6** (1941).

[10] G. G. Lorentz, *Approximation of Functions*, AMS Chelsea Publ., Providence (Rhode Island), 1986.

[11] J. Szabados, P. Vértesi, *Interpolation of Functions*, World Sci. Publ., Singapore, 1990.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

Computational procedures for parameter estimation in extremes: a review

M. Manuela Neves[1] and D. Prata Gomes[2]

[1] *Instituto Superior de Agronomia and CEAUL, Universidade de Lisboa, Portugal*

[2] *Mathematics Department/FCT and CMA, Universidade Nova de Lisboa, Portugal*

emails: manela@isa.ulisboa.pt, dsrp@fct.unl.pt

Abstract

When modelling extreme events there are a few primordial parameters among which we refer to the *extreme value index*, denoted by ξ, and the *extremal index*, denoted by θ. The *extreme value index* measures the right tail-weight of the underlying distribution and the *extremal index* characterizes the degree of local dependence in the extremes of a stationary sequence. Most of the semi-parametric estimators of these parameters present the well known type of behaviour: nice asymptotic properties but a high variance for small k, the number of upper order statistics used in the estimation, and an increasing bias with k. Recently, computer intensive procedures have revealed to be highly fruitful in extreme value parameter estimation. The role of computer intensive methodologies and adaptive algorithms for an adequate estimation of the aforementioned parameters are here revisited. Real data illustrations will also be provided.

Key words: extremal index, extreme value index, extreme value theory, resampling procedures.

MSC 2000: AMS codes (optional)

1 Introduction and preliminaries

Extreme Value Theory (EVT) aims to study and to predict the occurrence of extreme or even rare events, outside of the range of available data. These events are part of the real world but environmental extreme or rare events may have a massive impact on everyday life and may have catastrophic consequences for human activities.

 ISBN: 978-84-617-8694-7

The classical assumption in EVT, is that we have a set of independent and identically distributed (i.i.d.) random variables (r.v.'s), $X_1, \ldots, X_n$, from an unknown cumulative distribution function (c.d.f.) F and we are concerned with the limit behaviour of $M_n \equiv X_{n:n} = \max(X_1, \ldots, X_n)$ as $n \to \infty$. Whenever it is possible to linearly normalize M_n so that we get a non-degenerate limit, as $n \to \infty$, such a limit is of the type of the extreme value (EV) d.f.,

$$EV_\xi(x) := \begin{cases} \exp[-(1+\xi x)^{-1/\xi}], & 1+\xi x > 0 \quad \text{if} \quad \xi \neq 0 \\ \exp[-\exp(-x)], & x \in \mathbb{R} \qquad\quad \text{if} \quad \xi = 0. \end{cases} \tag{1}$$

We then say that F is in the domain of attraction for maxima of EV_ξ, denoting this by $F \in D_{\mathcal{M}}(EV_\xi)$. The parameter ξ is the *extreme value index* (EVI) and it measures essentially the weight of the right tail function, $\overline{F} = 1 - F$. The estimation of ξ, in (1), is then of primordial importance not only by itself but also because it is the basis for the estimation of all other parameters of extreme events.

In most fields of applications, the independence assumption is not valid. Stationary sequences are realistic for many real problems and dependence in stationary sequences can assume several forms. Provided that a stationary sequence $\{X_n\}_{n\geq 1}$ has limited long-range dependence at extreme levels, the maxima of this sequence follow the same distributional limit law as the associated independent sequence, $\{Y_n\}_{n\geq 1}$, but with other values for the parameters of EV d.f., Leadbetter *et al.*, (1983). Let us assume to be working with a strictly stationary sequence of r.v.'s, $\{X_n\}_{n\geq 1}$, with marginal d.f. F, under the long range dependence condition **D** (Leadbetter *et al.*, 1983) and the local dependence condition **D"** (Leadbetter and Nandagopalan, 1989). The stationary sequence $\{X_n\}_{n\geq 1}$ is said to have an extremal index (EI) θ, $0 < \theta \leq 1$, if for each $\tau > 0$, we can find a sequence of levels $u_n = u_n(\tau)$ such that, with $\{Y_n\}_{n\geq 1}$ the associated i.i.d. sequence (i.e. from the same F),

$$\mathbb{P}\left(Y_{n:n} \leq u_n\right) = F^n(u_n) \underset{n\to\infty}{\longrightarrow} e^{-\tau} \quad \text{and} \quad \mathbb{P}\left(X_{n:n} \leq u_n\right) \underset{n\to\infty}{\longrightarrow} e^{-\theta\tau}. \tag{2}$$

Under the validity of those conditions the *extremal index* can also be defined as

$$\theta = \frac{1}{\text{limiting mean size of clusters}} = \lim_{n\to\infty} \mathbb{P}(X_2 \leq u_n | X_1 > u_n) = \lim_{n\to\infty} \mathbb{P}(X_1 \leq u_n | X_2 > u_n), \tag{3}$$

where $u_n :\ F(u_n) = 1 - \tau/n + o(1/n),$ as $n \to \infty$, with $\tau > 0$, fixed.

D and **D**" are straightforwardly valid for iid data, and $\theta = 1$.

2 EVI and EI estimation and computational procedures

Under a semi-parametric framework it is only necessary to assume that $F \in D_{\mathcal{M}}(EV_\xi)$ and the estimators of EVI are based on the k largest observations. Considering heavy-tailed parents, the sample $(X_1, X_2, \cdots X_n)$ and the associated sample of ascending order statistics

 ISBN: 978-84-617-8694-7

(o.s.'s), $(X_{1:n} \leq \ldots \leq X_{n:n})$, the most famous classical EVI-estimator is the Hill estimator (Hill, 1975), here denoted $H(k)$ and given by

$$H(k) := \frac{1}{k} \sum_{i=1}^{k} \{\ln X_{n-i+1:n} - \ln X_{n-k:n}\}, \qquad k = 1, 2, \ldots, n-1. \tag{4}$$

Consistency of (4) is achieved if $X_{n-k:n}$ is an *intermediate* o.s., i.e., if $k \equiv k_n \to \infty$ and $k/n \to 0$, as $n \to \infty$, but that estimator presents a very high variance for small values of k and a very high bias for large values of k. A simple class of second-order *minimum-variance reduced-bias* (MVRB) EVI-estimators is the one in Caeiro *et al.* (2005). This class, here denoted $\overline{H}(k)$, depends upon the estimation of second-order parameters (β, ρ) and has the functional form:

$$\overline{H}(k) := H(k)(1 - \widehat{\beta}(n/k)^{\widehat{\rho}}/(1 - \widehat{\rho})), \tag{5}$$

with $H(k)$ the Hill estimator in (4), and where $(\widehat{\beta}, \widehat{\rho})$ needs to be an adequate consistent estimator of (β, ρ), as given, for example, in Gomes and Pestana (2007). Gomes *et al.* (2013) considered to remove the non-null asymptotic bias through the use of the Generalized Jackknife (GJ) methodology, Gray and Schucany (1972), by using an adequate pair of EVI-estimators to build a reduced-bias affine combination of them. As an example, we can refer to a class of GJ–EVI estimators, parameterised in a tuning parameter $\alpha \in (0, 1)$, defined as

$$\overline{H}^{GJ}(k) := \frac{\overline{H}(k) - \alpha^{2\hat{\rho}} \overline{H}(\lfloor \alpha k \rfloor)}{1 - \alpha^{2\hat{\rho}}}, \tag{6}$$

where $\lfloor x \rfloor$ denotes, as usual, the integer part of x.

Concerning the θ parameter, the classical up-crossing, UC-estimator, $\widehat{\Theta}^{UC}$ (Nandagopalan, 1990), is a naive estimator that comes directly as an empirical counterpart of (3),

$$\widehat{\Theta}^{UC}(u_n) := \frac{\sum_{i=1}^{n-1} I\left(X_i \leq u_n < X_{i+1}\right)}{\sum_{i=1}^{n} I(X_i > u_n)}, \tag{7}$$

for a suitable threshold u_n, where $I(A)$ denotes, as usual, the indicator function of A. Consistency of this estimator is obtained provided that the high level u_n is a normalized level, i.e. if with $\tau \equiv \tau_n$ fixed, the underlying d.f. F verifies $F(u_n) = 1 - \tau/n + o(1/n)$, $n \to \infty$ and $\tau/n \to 0$.

Gomes *et al.*(2008) considered the level u_n as a deterministic level $u \in [X_{n-k:n}, X_{n-k+1:n}[$ and derived a reduced-bias GJ estimator of order 2, based on the estimator $\widehat{\Theta}^{UC}$ computed at the three levels, k, $\lfloor k/2 \rfloor + 1$ and $\lfloor k/4 \rfloor + 1$, given by

$$\widehat{\Theta}_n^{GJ} := 5\widehat{\Theta}^{UC}(\lfloor k/2 \rfloor + 1) - 2(\widehat{\Theta}^{UC}(\lfloor k/4 \rfloor + 1) + \widehat{\Theta}^{UC}(k)). \tag{8}$$

Adaptive choice of thresholds k or u_n, through the bootstrap methodology and also through some heuristic computational procedures have been studied in several works, such as Gomes *et al.* (2012) and Neves *et al.* (2015), to cite only a few.

Those procedures will be reviewed here.

 ISBN: 978-84-617-8694-7

Acknowledgements

This work has been supported by FCT–Fundação para a Ciência e a Tecnologia through the projects UID/MAT/00006/2013 (CEAUL) and UID/MAT/00297/2013 (CMA).

References

[1] F. Caeiro, M.I. Gomes and D.D. Pestana, *Direct reduction of bias of the classical Hill estimator*, Revstat **3**:2 (2005) 111-136.

[2] M. I. Gomes, F. Figueiredo, and M. M. Neves, *Adaptive estimation of heavy right tails: resampling-based methods in action*, Extremes **15** (2012) 463–489.

[3] M. I. Gomes, M.I., A. Hall and C. Miranda, *Subsampling techniques and the Jackknife methodology in the estimation of the extremal index*, J. Comput. Statist. and Data Analysis **52**:4 (2008) 2022–2041.

[4] M. I. Gomes, M. J. Martins and M. M. Neves, *Generalised Jackknife-based estimators for univariate extreme-value modeling*, Comm. Statist. Theory Methods **42**:7 (2013) 1227–1245.

[5] M.I. Gomes and D. Pestana, *A sturdy reduced-bias extreme quantile (VaR) estimator*, J. Amer. Statist. Assoc. **102** (2007) 280–292.

[6] H.L. Gray and W.R. Schucany, *The Generalized Jackknife Statistic*, Marcel Dekker. New York, 1972.

[7] B. Hill, *A simple general approach to inference about the tail of a distribution*, Ann. Statist. **3** (1975) 1163–1174.

[8] M. R. Leadbetter, G. Lindgren and H. Rootzén, *Extremes and related properties of random sequences and series*, Springer-Verlag, New York, 1983.

[9] M.R. Leadbetter and S. Nandagopalan, *On exceedance point process for stationary sequences under mild oscillation restrictions*, *In Extreme Value Theory.* J. Hüsler and R. D. Reiss (edts), 69–80, Springer-Verlag, Berlim, 1989.

[10] S. Nandagopalan, *Multivariate Extremes and Estimation of the Extremal Index*, PhD Thesis, University of North Carolina, Chapel Hill, 1990.

[11] M.M. Neves, M.I. Gomes, F. Figueiredo and D. Prata Gomes, *Modeling Extreme Events: Sample Fraction Adaptive Choice in Parameter Estimation*, J. Stat. Theory Pract. **9**:1 (2015) 184–199.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

Some notes on the convergence of GMRES for compact operator equations

Paolo Novati[1]

[1] *Department of Mathematics and Geosciences, University of Trieste, Italy*

emails: novati@units.it

Abstract

We study some properties of GMRES for solving infinite dimensional linear equations involving compact operators. These problems are intrinsically ill-posed since a compact operator does not admit a bounded inverse. We study the convergence and the regularization properties with respect to the decay rate of the singular values of the operator.

Key words: Linear ill-posed problem, Compact operator, Hilbert-Schmidt operator, Arnoldi algorithm, GMRES
MSC 2000: 47A52, 65F10

1 Introduction

We consider linear equations of the type

$$Af = g, \tag{1}$$

where f and g belong to a separable Hilbert space $\mathcal{H}$, and $A : \mathcal{H} \to \mathcal{H}$ is a compact linear operator. Even assuming that A is not of finite rank, the problem (1) is ill-posed since a compact operator does not possess a bounded inverse. For this reason, after a suitable discretization the arising algebraic linear system is generally solved in the least square sense through some kind of regularization such as the popular Tikhonov method (see e.g. [4, Chapter 5] for an overview) or by means of iterative methods that are known to be 'self-regularizing' such as the CG in the Hermitian case and the LSQR ([8]) or the CGLS in the general nonhermitian one. The main drawback of the CGLS and the LSQR is that they need to work with the operator adjoint A^* that in some important applications is not

 ISBN: 978-84-617-8694-7

known since A is only defined through its action. For this reason the Arnoldi based methods such as the well-known GMRES have been recently employed in this field and they have been shown to be a valid alternative to the transpose based method. In this sense, the first attempt was presented in [2]. We also quote here [3] for a recent survey.

In this work, working in the continuous framework defined by (1), we give a theoretical justification of some important features of the Arnoldi based methods, that are commonly considered true from experimentation in the finite dimensional setting. Denoting by $\{\sigma_n\}_{n\in\mathbb{N}}$ the sequence of the singular values of A (not of finite rank) arranged in decreasing order of magnitude we demonstrate that if $\{\sigma_n\}_{n\in\mathbb{N}} \in \ell_2$ (as it occurs in many practical situations) then the GMRES converges to the exact solution of (1). The rate of convergence is also put in relation with the extendibility of the Krylov subspaces that can be measured in terms of the decay rate of the singular values.

2 Background

Let $\mathcal{H}$ be a separable Hilbert space (it admits a countable orthonormal basis $\{\varphi_n\}_{n\in\mathbb{N}}$), with scalar product $<,>$ and norm $\|\cdot\|$ defined as

$$\|x\| = \langle x, x\rangle^{1/2}.$$

For any given $p > 0$, we denote by ℓ_p the set of the positive sequences $\{a_j\}_{j\geq 1}$ such that

$$\sum\nolimits_{j\geq 1} a_j^p < \infty.$$

Let $A : \mathcal{H} \to \mathcal{H}$ be a linear operator. Given $g \in \mathcal{H}$, we denote by $\mathcal{K}_m = \text{span}\{g, Ag, \ldots, A^{m-1}g\}$ the Krylov subspaces generated by A and g. Setting $N = \sup_m(\dim \mathcal{K}_m)$, the Arnoldi algorithm computes an orthonormal basis $\{w_1, ..., w_m\}$ of $\mathcal{K}_m$ for each $m \leq N$. In particular, we have

$$w_1 = \frac{g}{\|g\|}, \quad w_{m+1} = \frac{(I - P_m)Aw_m}{\|(I - P_m)Aw_m\|},$$

where P_m is the orthogonal projection onto $\mathcal{K}_m$. If, for some m, $(I - P_m)Aw_m = 0$, then $N = m$ and $w_{N+1} = 0$.

Now consider the sequence $\{f_m\}_{m\geq 1}$, $f_m \in \mathcal{K}_m$, such that the corresponding residual norm $\|Af_m - g\|$ is minimized over all the elements of $\mathcal{K}_m$, for $1 \leq m \leq N$. As before, if $N < \infty$, we have that $Af_N - g = 0$, so that, f_N is the solution of $Af = g$. Such a sequence can be constructed with the well known GMRES algorithm ($f_0 = 0$). It is also well known that $Af_m - g \perp A\mathcal{K}_m$, that is, $Q_mAf_m - g = 0$ where Q_m is the projection onto $\mathcal{K}_m$ orthogonal to $A\mathcal{K}_m$. The GMRES approximation is uniquely defined if the operator $Q_mA|_{\mathcal{K}_m} : \mathcal{K}_m \to \mathcal{K}_m$ is invertible for each $m \leq N$, and this condition is ensured if A is not of finite rank.

 ISBN: 978-84-617-8694-7

Theorem 1 *([9, Th.1.9.3]) Let $A : \mathcal{H} \to \mathcal{H}$ be a compact linear operator. Then there exists a decreasing sequence of positive real number $\{\sigma_n\}_{n\in\mathbf{S}}$ (finite or countably infinite and converging to 0) and two orthonormal sequences $\{\varphi_n\}_{n\in\mathbf{S}}$, $\{\psi_n\}_{n\in\mathbf{S}}$, such that*

$$Ax = \sum\nolimits_{n\in\mathbf{S}} \sigma_n \langle x, \varphi_n\rangle \psi_n, \quad x \in \mathcal{H}. \tag{2}$$

The sequence $\{\sigma_n\}_{n\in\mathbf{S}}$ is uniquely determined and consists of the eigenvalues of the positive self-adjoint operator $(A^*A)^{1/2}$ (the singular values of A) counted according to their multiplicities; $\{\varphi_n\}_{n\in\mathbf{S}}$ is the corresponding sequence of eigenvectors.

Assuming that a compact linear operator is not of finite rank, for each $g \in \mathcal{H}$, the equation $Af = g$ has a candidate solution f given by

$$f = \sum\nolimits_{n\geq 1} \frac{\langle g, \psi_n\rangle}{\sigma_n}\varphi_n.$$

Since $\|f\|^2 = \sum_{n\geq 1} \left|\frac{\langle g,\psi_n\rangle}{\sigma_n}\right|^2$ by Parseval identity, $f \in \mathcal{H}$ if and only if

$$\left\{\frac{|\langle g, \psi_n\rangle|}{\sigma_n}\right\}_{n\geq 1} \in \ell_2. \tag{3}$$

Assuming that $\{\sigma_n\}_{n\geq 1} \in \ell_p$, $p > 0$, by the generalized Hölder inequality (see e.g. [6, §2.7]) we have that

$$\{|\langle g, \psi_n\rangle|\}_{n\geq 1} \in \ell_{\frac{2p}{2+p}}. \tag{4}$$

Since $\frac{2p}{2+p} < p$, the condition (4) expresses what is commonly called Picard Condition, that is, the coefficients $|\langle g, \psi_n\rangle|$ must decay faster than the singular values, [4, §1.2.3].

3 Convergence analysis

Definition 2 *Let $A : \mathcal{H} \to \mathcal{H}$ be a bounded linear operator, and let $\{\varphi_n\}_{n\in\mathbb{N}}$ be any orthonormal basis of $\mathcal{H}$. If*

$$\sum\nolimits_{n\in\mathbb{N}} \|A\varphi_n\|^2 < \infty \tag{5}$$

then A is a Hilbert-Schmidt operator. We denote by $\mathcal{C}_2(\mathcal{H})$ this class.

Remark 3 *([9, Chapter 2]) Relation (5) ensures that a Hilbert-Schmidt operator is also compact. Moreover, for each orthonormal basis $\{\varphi_n\}_{n\in\mathbb{N}}$*

$$\sum\nolimits_{n\in\mathbb{N}} \|A\varphi_n\|^2 = \sum\nolimits_{j\in\mathbb{N}} \sigma_j^2$$

 ISBN: 978-84-617-8694-7

Theorem 4 *[7] Let $A : \mathcal{H} \to \mathcal{H}$ be a compact linear operator with a singular value expansion (2) and let $\{f_m\}_{m\geq 1}$ be the sequence of GMRES approximations. If g satisfies the condition (3) then $\|f_m - f\| \to 0$. Moreover, if $A \in \mathcal{C}_2(\mathcal{H})$ then there exists a non negative sequence $\{a_i\}_{i\geq 1} \in \ell_2$, such that*

$$\|Af_m - g\| \leq \left(\sum\nolimits_{i>m} a_i^2\right)^{1/2}. \tag{6}$$

The connection between the residuals of FOM and GMRES, expressed by the famous peak-plateau phenomenon (see e.g. [1]), ensures that the GMRES convergence implies the FOM convergence. This means that the sequence of FOM approximations is bounded, that is, we have $\|f_m\| \leq M$. As consequence, we have that the FOM residual (and hence the GMRES one) is bounded by $Mh_{m+1,m}$, where $h_{m+1,m} := \langle w_{m+1}, Aw_m\rangle = \|(I - P_m)Aw_m\|$. Indeed, since the FOM approximation satisfies $P_m\,(Af_m - g) = 0$, and $P_m g = g$, we have

$$\|Af_m - g\| = h_{m+1,m}\,|\langle f_m, w_m\rangle| \leq h_{m+1,m}\,\|f_m\|\,. \tag{7}$$

Definition 5 *Let $A : \mathcal{H} \to \mathcal{H}$ be a compact operator and let $p > 0$. Then A is p-nuclear and we write $A \in \mathcal{C}_p(\mathcal{H})$ if $\{\sigma_j\}_{j\in\mathbb{N}} \in \ell_p$.*

The above definition implies that Hilbert-Schmidt operators are 2-nuclear operators. We can state the following.

Theorem 6 *[7] Let $A \in \mathcal{C}_p(\mathcal{H})$ with $p > 0$. The following statements hold.*

1. *If $p \geq 1$ then $\{h_{j+1,j}\}_{j\in\mathbb{N}} \in \ell_p$.*
2. *If $\{h_{j+1,j}\}_{j\in\mathbb{N}}$ is non increasing then $\{h_{j+1,j}\}_{j\in\mathbb{N}} \in \ell_p$.*
3. *If $h_{j+2,j+1} \leq \left(\prod_{i=1}^{j} h_{i+1,i}\right)^{1/j}$, $j \geq 1$, then $\{h_{j+1,j}\}_{j\in\mathbb{N}} \in \ell_{p+\varepsilon}$ for each $\varepsilon > 0$.*

Theorem 7 *[7] Let $A \in \mathcal{C}_p(\mathcal{H})$, $p > 0$. If the condition (3) is satisfied then for the GMRES residual it holds*

$$\{\|Af_m - g\|\}_{m\geq 1} \in \ell_p. \tag{8}$$

Example 8 *In Figure 1 we consider the plot of the sequences $\{\|Af_m - g\|\}_{m\geq 1}$, $\{h_{m+1,m}\}_{m\geq 1}$, $\{\sigma_m\}_{m\geq 1}$, for two test problems taken from [5], that is, BAART and I-LAPLACE, that are linear systems arising from the discretization of Fredholm integral equations of the first kind. In both cases the singular values decay exponentially so that $\{\sigma_m\}_{m\geq 1} \in \ell_p$ for each $p > 0$.*

Acknowledgements

This work was partially supported by GNCS-INdAM and by FRA - University of Trieste.

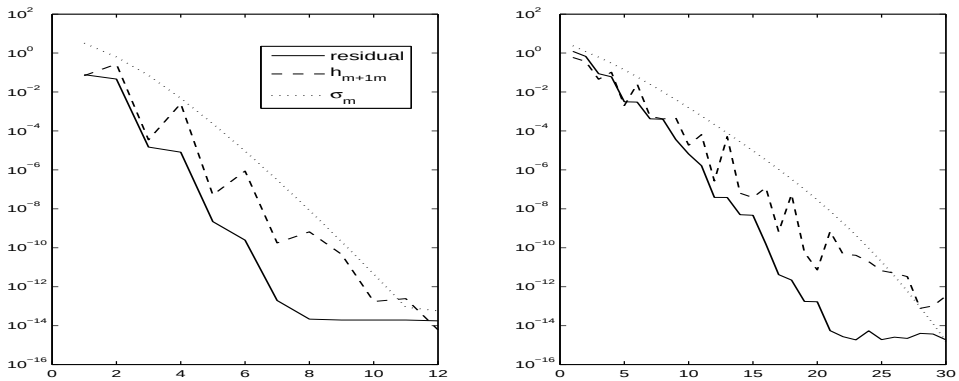


Figure 1: GMRES residual history for the problem BAART on the left and I-LAPLACE on the right.

References

[1] P. N. Brown, *A theoretical comparison of the Arnoldi and GMRES algorithms*, SIAM J. Sci. Statist. Comput. **12** (1991) 58-78.

[2] D. Calvetti, S. Morigi, L. Reichel and F. Sgallari, *Tikhonov regularization and the L-curve for large discrete ill-posed problems*, J. Comput. Appl. Math. **123** (2000) 423–446.

[3] S. Gazzola, P. Novati and M. R. Russo, *On Krylov projection methods and Tikhonov regularization*, Electron. Trans. Numer. Anal. **44** (2015) 83–123.

[4] P. C. Hansen, *Rank-Deficient and Discrete Ill-Posed Problems: Numerical Aspects of Linear Inversion*, SIAM, Philadelphia, 1998.

[5] P. C. Hansen, *Regularization Tools Version 4.0 for Matlab 7.3*, Numer. Algorithms **46** (2007) 189–194.

[6] G. H. Hardy, J. E. Littlewood and G. Pólya, *Inequalities. Second Edition*, Cambridge University Press, 1952.

[7] P. Novati, *Some properties of the Arnoldi based methods for linear ill-posed problems*, SIAM J. Numer. Anal. (2017) in press.

[8] C. C. Paige, M. A. Saunders, *LSQR: an algorithm for sparse linear equations and and sparse least squares*, ACM Trans. Math. Software **8** (1982) 43–71.

 ISBN: 978-84-617-8694-7

[9] J. R. Ringrose, *Compact non-self-adjoint operator*, Van Nostrand Reinhold Company, London, 1971.

 ISBN: 978-84-617-8694-7

Volume V

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

Towards co-execution of massive data-parallel OpenCL kernels on CPU and Intel Xeon Phi

Raúl Nozal[1], Borja Pérez[1] and Jose Luis Bosque[1]

[1] *Computer and Electronics Engineering Department, University of Cantabria*

emails: raul.nozal@unican.es, borja.perezpavon@unican.es,
joseluis.bosque@unican.es

Abstract

Heterogeneous systems composed by a CPU and a set of different hardware accelerators are very compelling thanks to their excellent performance and energy consumption features. One of the most important problems of those systems is the workload distribution among their devices. This paper describes an extension of the Maat library to allow the co-execution of a data-parallel OpenCL kernel on a heterogeneous system composed by a CPU and an Intel Xeon Phi. Maat provides an abstract view of the heterogeneous system as well as a set of load balancing algorithms to squeeze the performance out of the node. Experimental results show that this approach always outperforms the baseline with only a Xeon Phi. Furthermore, the load balancing algorithm has a huge impact in the system performance, therefore, the right selection is essential.

Key words: Heterogeneous computing, co-execution CPU-Xeon Phi, load balancing, OpenCL

1 Introduction

One of the most important challenges of high performance computing today is to reach Exascale computers. Nowadays one of the most promising ways is the use of cluster-based architectures with nodes that have great computing capacity. These nodes are based on multi-core processors and include hardware accelerators, such as GPUs or Intel Xeon Phi.

Nonetheless these nodes introduce new challenges, because the use of accelerators turns them heterogeneous. Hence, the software development to efficiently exploit all the available resources has to take into account this heterogeneity. Therefore the use of a portable programming language on heterogeneous platforms is mandatory. Open Computing Language

 ISBN: 978-84-617-8694-7

(OpenCL) [1] is a framework for developing programs to be executed across heterogeneous platforms such as many-core and multi-core architectures.

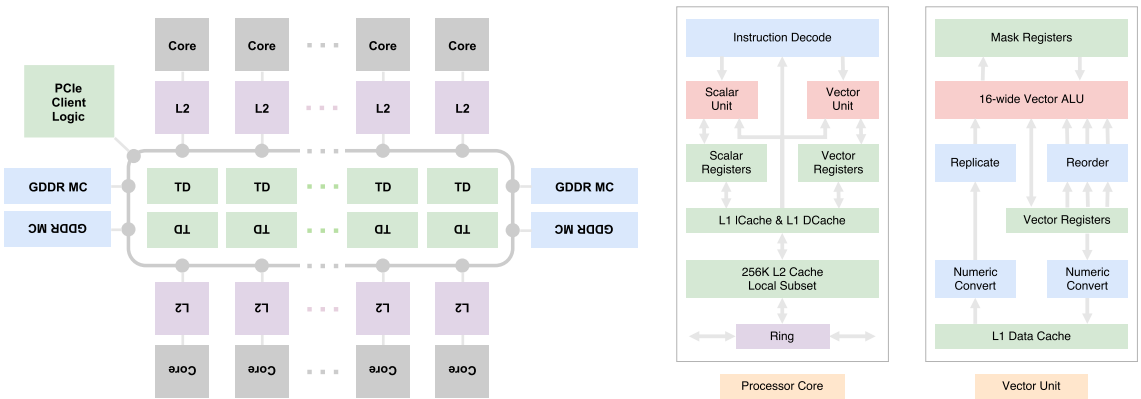
However, when using OpenCL, the programmer is responsible for explicitly selecting and managing the devices as well as for partitioning the data among them. Therefore load balancing becomes one of the most challenging problems, having a tremendous impact on performance and programmability. The objective of load balancing algorithms is distributing the workload proportionally to the devices' computational power. This problem is more acute in heterogeneous systems with irregular applications.

This paper extends Maat [2] to allow the co-execution of massively data-parallel kernels on heterogeneous systems composed by a multi-core CPU and an Intel Xeon Phi accelerator. Maat is an OpenCL library that provides the programmer with a unified and abstract view of the heterogeneous system that guarantees code portability. Furthermore, Maat provides a set of load balancing algorithms that allow the programmer to choose the most appropriate depending on the behaviour of the application at hand. In this paper a diverse set of applications it is considered. They are grouped as regular, when every work unit represents the same running time, and irregular, if different work units have different running times. This decision is critical in order to achieve the best performance, with different applications, as will be shown in section 5. Experimental results show speedups close to the maximum achievable, reaching an efficiency of more than 97% and 86% for both regular and irregular applications, respectively.

This topic has given rise to an interest in both understanding how to efficiently use this kind of devices and making their programming easier. The work presented in [3] studies the performance of the Xeon Phi through the use offload directive based approaches in a single heterogeneous node. [4] also addresses the use of directives with Xeon Phi, but in this case with a focus on dynamic load balancing. The authors of [5] also focus on improving the performance of an application running on a Xeon Phi, buy they deem load balancing unsuitable for the kind of computations of a Xeon Phi. As a result, they propose a load imbalancing based optimisation technique. To the problem of cooperatively executing a single task among the available devices, [6] proposes an analytical model to identify when co-execution is worth it in terms of performance. Finally, [7] presents a dynamic load balancing for the co-execution of a single data-parallel kernel. However, these two works do not consider the Xeon Phi as their target device.

The remainder of the paper is structured as follows. Section 2 introduces some basic concepts that are important to the article. Section 3 describes the main characteristics of the three load balancing techniques which are studied. Section 4 explains the extensions developed on Maat library. Next, in section 5 a performance evaluation of the three load balancing algorithms with different benchmarks is developed. Finally, section 6 summarises the main conclusions and future work.

 ISBN: 978-84-617-8694-7



(a) Block diagram architecture.

(b) Microarchitecture of a Xeon Phi's Core.

Figure 1: Intel Xeon Phi Architecture.

2 Background

The Intel Xeon Phi Knights Corner is a coprocessor to speed-up parallel applications; therefore it requires at least one processor in the system [8]. It is a cache-coherent shared memory many-core processor (SMP) based on the x86 architecture. The coprocessor is connected via the PCI express bus to other devices and the host, and they are not hardware cache coherent with other devices in the node. It consists of up to 61 cores connected by a high performance on-die bidirectional ring interconnect and 8 memory controllers supporting 16 GB of GDDR5 channels at most, with up to 352 GB/s bandwidth. Figure 1(a) shows the basic architecture of the MIC coprocessor.

The cores are in-order dual issue x86 processor 1.2 GHz cores, with 32 KB of data cache and 32 KB of instruction cache, with 64-bit support and as it is shown in the figure 1(b), it provides more interesting extensions:

- Four hardware threads on each core, resulting in up to 244 hardware threads available on a single device. They are primarily used to hide memory latencies implicit in an in-order microarchitecture. As such, it is much more important that applications use these multiple hardware threads on Intel Xeon Phi than on Xeon processors.
- Each core has a 512-KB L2 cache locally with high speed access to all other L2 caches, making a collective L2 cache size over 30 MB.
- Wide vectorization capability via 512-bit wide vector registers and functional units, which can execute SIMD instructions, with 16 single precision or 8 double precision operations simultaneously.

The Xeon Phi runs Linux operating system. Every card has its own IP address. The Xeon Phi supports many common tools for developing, debugging and executing parallel applications. It supports the standard parallel programming models, such as MPI and OpenMP.

3 Load balancing Algorithms

Since there is no load balancing strategy for data parallelism that can improve the execution of all applications, in this paper we propose a set of load balancing algorithms. The basis of these is that the load experienced by a computing device is only dependent on the amount of data it must process.

3.1 Static algorithm

This algorithm works before the task is executed by dividing the data-set in as many *packages* as devices are in the system. The division relies on knowing the computing power of the devices in advance. Then the execution time of each device can be equalised by proportionally dividing the data-set among the devices.

Considering a heterogeneous system with n devices. Each device i has *computational power* P_i, which is defined as the amount of work that a device can complete per time unit, including the communication overhead. These powers are parameters that must be given to the algorithm and can be extracted by a simple profiled execution. Then, the total computational power of the heterogeneous system is the sum of the individual powers of the devices $P_H = \sum_{i=1}^{n} P_i$.

The application will execute a kernel over W work-items, grouped in G work-groups of fixed size $L_s = \frac{W}{G}$. Since the work-groups do not communicate among themselves, it makes sense to distribute the workload taking the work-group as the atomic unit. Each device will have an execution time of T_i. Then the execution time of the heterogeneous system will be that of the last device to finish its work, or $T_H = max_{i=1}^{n} T_i$.

The goal of the Static algorithm is to determine the number of work-groups to assign each device, so that all the devices finish their work at the same time. This means finding a tuple $\{\alpha_1, ...\alpha_n\}$, where α_i is the number of work-groups assigned to the device i. For this reason, the expression used by the algorithm is:

$$\alpha_i = \left\lfloor \frac{P_i G}{\sum_{i=1}^{n} P_i} \right\rfloor$$

If there is not an exact solution with integers then $\sum_{i=1}^{n} \alpha_i < G$. In this case, the remaining work-groups are assigned to the most powerful device.

The beauty of the Static algorithm is that it minimises the number of synchronisation points. This makes it perform well when facing regular loads with known computing powers

 ISBN: 978-84-617-8694-7

that are stable throughout the data-set. However, it is not adaptable, so its performance might not be as good with irregular loads.

3.2 Dynamic algorithm

Some applications do not present a constant load during their executions. To adapt to their irregularities, the Dynamic algorithm divides the data-set in small packages of equal size. The number of packages is well above the number of devices in the heterogeneous system. During the execution of the kernel, a master thread in the host is in charge of assigning packages to the different devices, including the CPU, following the next strategy:

1. The master splits the G work-groups in packages, each with the package size specified by the programmer. This number must be a multiple of the work-group size. If the number of work-items is not divisible by the package size, the last package will be smaller.

2. The master launches one package on each device, including the host itself if it is desired.

3. The master waits for the completion of any package.

4. When device i completes the execution of a package:

   (a) The device returns the partial results corresponding to the processed package.

   (b) The master stores the partial results.

   (c) If there are outstanding packages, a new one is launched on device i.

   (d) If all the devices are idle and there are no more packages, the master jumps to step 5.

   (e) The master returns to step 3.

5. The master ends when all the packages have been processed and the results have been received.

This algorithm adapts to the irregular behaviour of some applications. However, each completed package represents a synchronisation point between the device and the host, where data is exchanged and a new package is launched. This overhead has a noticeable impact on performance. The Dynamic algorithm takes the size of the packages as a parameter.

 ISBN: 978-84-617-8694-7

3.3 HGuided algorithm

The previous strategies have their strong points and their weak spots. Although neither is the best for every application, both give hints toward an optimal data-division algorithm. The Heterogeneous Guided algorithm (*HGuided*) is an attempt to reduce the synchronisation points of the Dynamic while retaining most of its adaptiveness.

The same algorithm used in the Dynamic approach is applicable to the HGuided, except for how the data-set is divided. The HGuided algorithm makes larger packages at the beginning and reduces the size of the subsequent ones as the execution progresses. This reduces the number of synchronisation points and the corresponding overhead, while retaining a small package granularity towards the end of the execution to allow all devices to finish simultaneously.

Since it is an algorithm for heterogeneous systems the size of the packets is also dependent on the computing power of the devices. For instance, the size of the package for device i is calculated as follows:

$$packet\_size_H = \left\lfloor \frac{G_r P_i}{n \sum_{j=1}^{n} P_j} \right\rfloor$$

Where G_r is the number of pending work-groups and is updated with every package launch. The HGuided takes the computing powers of the devices as parameters. It also requires a minimum package size, which is a lower bound for the $packet\_size_H$.

4 Design and Implementation

Maat is a library that acts as an OpenCL wrapper to simplify the programming of heterogeneous devices and squeeze the performance out of them. Maat is specially designed to be used in large data-parallel applications and it provides the three load balancing algorithms presented previously. While OpenCL forces the programmer to consider the devices as individual entities, Maat defines an abstraction layer over all the accelerator devices in a machine. It presents a single virtual device to operate with; hiding the underlying hardware details. Thus, the library effectively divides a single task among all the real devices based on the load balancing algorithm selected by the programmer.

This single virtual device is accessed through a *super context*. It is created by the programmer specifying the target devices. In contrast to OpenCL contexts, a *super context* can hold devices from several different manufacturers. Maat offers a set of functions that resemble typical OpenCL calls, to manage the *super context*. The *super context* transparently manages the data structures of all the target devices, like the command queues.

While in OpenCL the programmer will need to allocate many buffers to communicate with the different target devices, Maat simplifies this task with the *super buffer*. When one

 ISBN: 978-84-617-8694-7

of these is created through the *super context*, the latter transparently allocates the required buffers on each target device. If the data will only flow from the host to the device, it is considered an *in super buffer*. Otherwise it is considered an *out super buffer*.

The common reading and writing procedure in *super buffers* assumes that each work-item will use the position indicated by its index. The copy is not performed until the kernel is launched in the device. Such behaviour is necessary in Dynamic and HGuided algorithms where there is no way of knowing in advance which device will compute what package of data. The *out super buffer* creation function requires two extra parameters to be able to automatically copy back the results from the device to the host: a pointer to where the results should be stored in the host memory and the size of the result obtained by each work-group. Therefore, the requirement of writing buffers based on the work-item index denotes the type of applications supported. This may seem a strong requirement, but not only common benchmarks but also many kernels widely used in the industry meet this condition.

OpenCL encapsulates the code to be executed by a device, in a data structure called kernel, which is bound to the device. Maat simplifies the kernel configuration and execution process by using the idea of *super kernel*. When such an entity is created, the *super context* transparently sets up a different kernel for each device. Similarly, parameter assignment to the *super kernel* is forwarded to all the kernels in the *super context*. The *super kernel* launch resembles that of a standard OpenCL kernel in that it receives exactly the same global work size, local work size and global work offset it would receive if working with a single device, achieving an easy portability of common OpenCL applications to Maat.

When a *super kernel* is enqueued, Maat transparently performs as many executions of individual kernels as required by the selected load balancing algorithm. Each execution will use the adequate OpenCL parameters to represent the correct package of work. The on-demand launch of additional packages in the Dynamic and HGuided algorithms has been implemented using OpenCL callbacks.

The callbacks have been adapted to perform non-blocking readings. Most of the logic has been transferred to functions managed by an independent thread in the host because the new Intel Xeon Phi's OpenCL driver behaves different from the CPU's regarding callbacks. This increases the performance because it simplifies the callbacks and the synchronisations are dead-lock free (host functions are triggered using *pthread_cond_wait*).

Finally, a final function that waits for the completion of every callback guarantees the correctness of the non-blocking calls and improves the performance due to the maximum interleaving of kernel execution with buffer management operations. This waiting call serves as a common interface between the three load balancing algorithms to monitor every performed callback, follow its event state and wait for its completion, simplifying the kernel execution of the three algorithms.

 ISBN: 978-84-617-8694-7

Application	Type	Problem size	Local work size
Binomial	Regular	8192000 samples, 255 steps	256
Gaussian	Regular	13000×13000 px, 101×101 px filter	128
NBody	Regular	1536000 bodies	128
Ray1	Irregular	12000×12000 px, 3 lights in 10 objects	64
Ray2	Irregular	5000×5000 px, 11 lights in 30 objects	64

Table 1: Parameters for each application.

5 Evaluation

5.1 Experimental Setup

The machine on which the experimentation was carried out has two processor chips and one Intel Xeon Phi. The CPUs are Intel Sandy Bridge Xeon E5-2620, with six cores that can run two threads each at 2.0 GHz. It has hardware support for 24 threads, 15 MB of L3 cache memory and 16 GB of DDR3 main memory. The CPUs are connected via QPI, which allows OpenCL to detect them as a single device. Therefore, throughout the remainder of this document, any reference to the CPU includes both Xeon E5-2620 processors.

The Intel Xeon Phi coprocessor is a Knights Corner 7120P which consists of up to 61 cores connected by a high performance on-die bidirectional ring interconnect and 8 memory controllers supporting 16 GB of GDDR5 channels at most, with up to 352 GB/s bandwidth. It is connected to the system using an independent PCI 2.0 slot.

5.2 Benchmarks and Metrics

Four benchmarks have been chosen for the experiments. While Binomial, Gaussian and NBody are regular applications, Ray Tracing is irregular. The last one renders realistic images by modelling different light rays with independent threads each one involving an unpredictable amount of work depending on the number of times it bounces on the objects of the scene. Two different scenes with different complexity (resolution, lights and objects), referred as Ray1 and Ray2, have been considered for experimentation. As it will be seen later, by changing the input data, the behaviour of the application varies wildly, due to the irregular behaviour of the benchmark.

The parameters for each of the applications are shown in table 1. Each benchmark has been run using a problem size big enough to justify its distribution among all the available devices. Local work size has been set so the performance of the fastest device, namely the Xeon Phi, is maximised. The reason for this is that almost no performance difference was detected when varying local work size for the CPU.

The metric used to evaluate the performance of the algorithms is the total response time,

 ISBN: 978-84-617-8694-7

including the input data and results communications. From that, the speedup is calculated as the ratio between the execution time on the Phi and on the heterogeneous system. Due to the heterogeneity of the system and the different behaviour of the benchmarks, the maximum achievable speedups depend on each benchmark. These values were derived from the response time T_i of each device:

$$S_{max} = \frac{1}{max_{i=1}^{n}\{T_i\}} \sum_{i=1}^{n} T_i \tag{1}$$

Additionally, the efficiency of the heterogeneous system has been computed as the ratio between the maximum achievable speedup and the real observed speedup for each benchmark. $Eff = \frac{S_{real}}{S_{max}}$

The computational power needed for the Static and HGuided algorithms has been computed for each benchmark as follows. First, the response times of the benchmark have been measured in both the CPU and the Phi, including the time required to complete the kernel execution, performing the data distribution, kernel launch overhead and result collection. Then, considering the computational power of the CPU equal to 1, the computational power of the Phi is calculated, as the ratio among the CPU time and the Phi time: $P_{Phi} = \frac{T_{CPU}}{T_{Phi}}$.

5.3 Experimental Results

This section presents the performance results achieved in the heterogeneous system with different load balancing algorithms. The baseline to compare the results obtained with Maat on the heterogeneous system is the total response time of the benchmarks in a scenario with only one Xeon Phi. Therefore, the benefits presented in this section are due to the co-execution of the benchmarks on the CPU and the Phi simultaneously.

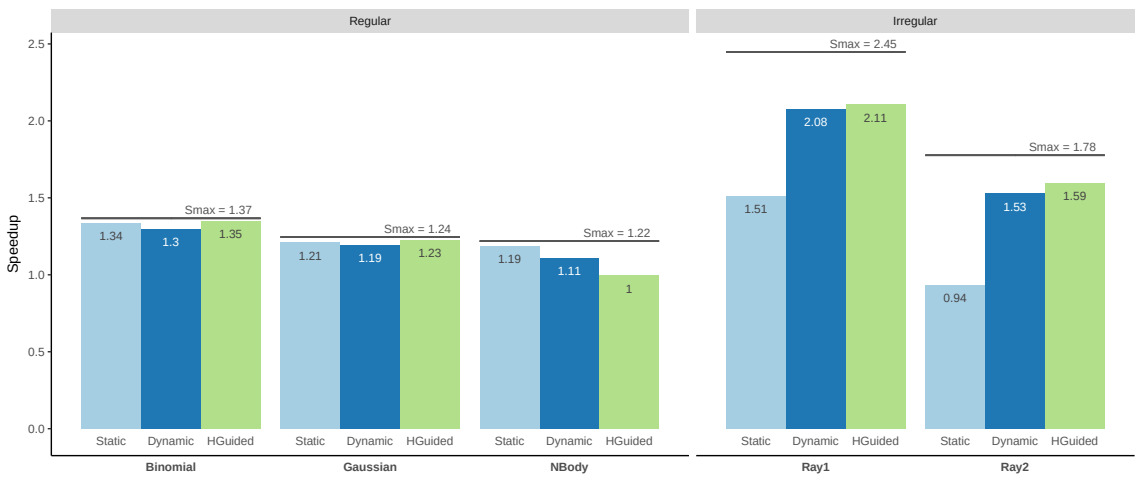


Figure 2: Speedups of the benchmarks with the three load balancing algorithms.

 ISBN: 978-84-617-8694-7

To give an idea of performance of the load balancing algorithms, the figure 2 shows the speedups reached by the Maat implementation of the benchmarks, compared with the baseline. The theoretical maximum speedup that can be obtained with each benchmark is shown as a horizontal line above the bars of each benchmark. The speedups reveal that, for all benchmarks, there is at least one algorithm that gives excellent results. This indicates that Maat can adapt to different kinds of loads obtaining outstanding performance. The gap between the measured and the theoretical maximum values is a consequence of the extra communication overhead.

Analysing the speedup measurements in detail, it can be seen that, Static and HGuided deliver excellent results in Binomial and Gaussian. However, while Static excels, HGuided decreases its performance with NBody due to the small ratio of computing time compared with communication time. The Static algorithm yields speedups very close to the maximum and it is the most consistent between the regular applications.

On the other hand, the analysis of the Ray Tracing, the irregular benchmark, shows that the HGuided method obtains the best results, followed nearly by the Dynamic. This method achieves a better load balance and reduces the communication overhead, because it divides the workload between a smaller number of packets than in the Dynamic algorithm.

Finally, the load balancing efficiency gives an idea of how well a load is balanced. A value of 100% represents that all the devices have been working all the time, thus achieving the maximum obtainable speedup. As shown in figure 3 the regular applications reach at least 97% of efficiency, while the irregular ones achieve an efficiency between 86% and 89%.

Coming up with a balanced work distribution is significantly harder for Ray due to the resolution of the rendered image and the complexity of interaction between the input

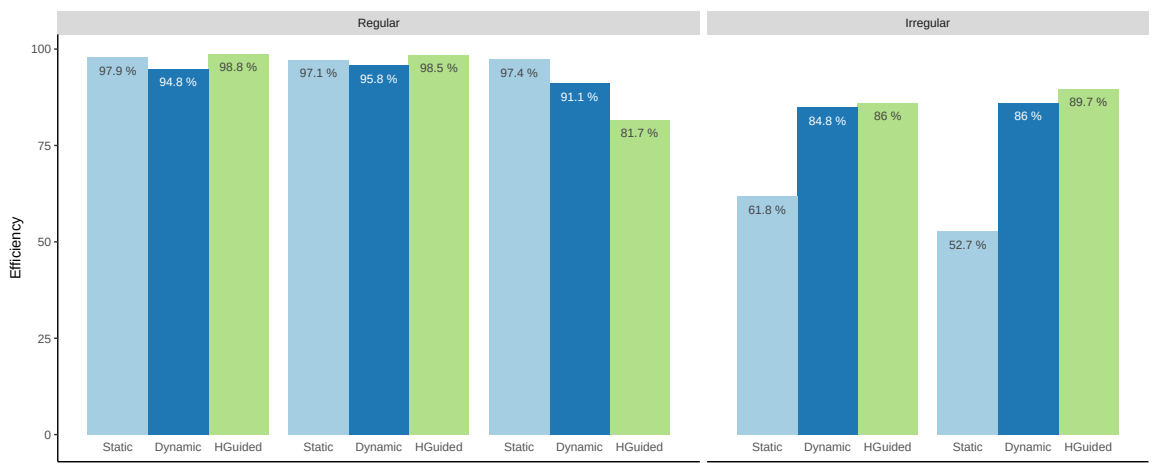


Figure 3: Balancing efficiency of the heterogeneous system.

 ISBN: 978-84-617-8694-7

scene and the ray tracing algorithm. However, it should be noted that if the correct load balancing algorithm is not used for each benchmark, the results can be quite bad. Thus we can see efficiencies around 50% in some cases, which imply a very poor use of resources.

6 Conclusions and Future Work

This paper extends the Maat library to allow the co-execution of massively data-parallel OpenCL kernels on heterogeneous systems composed by an Intel Xeon Phi coprocessor. A set of load balancing algorithms has been implemented, in order to give the best performance to both regular and irregular applications.

From the experimental results presented in this paper a set of conclusions can be remarked. The use of the whole heterogeneous system is always beneficial, from the performance point of view, at least with one of the load balancing methods. A second conclusion is that applications with different behaviour, regular or irregular ones, need different load balancing algorithms to get the best performance on the heterogeneous system. With respect to the algorithms, the Static approach is the most adequate for regular applications as it minimises overheads, although the HGuided is even better in specific cases. In the case of irregular applications the HGuided method excels and the use of Static is strongly discouraged as it may result in poor performance due to imbalance. Lastly, the Dynamic algorithm is a good all-around option when a priori information of the computing powers of the system is not available, but at the cost of more modest speedups and efficiencies.

Future work includes the optimisation of some implementation issues to reduce the synchronisation and communication overhead that are one of the most important challenges to improve the performance of Maat. Moreover, a study about the energy efficiency of the heterogeneous system will be developed.

Acknowledgements

This work has been supported by the University of Cantabria, grant CVE-2014-18166, the Spanish Science and Technology Commission under contracts TIN2016-76635-C2-2-R and TIN2016-81840-REDT (CAPAP-H6 network), the European Research Council (G.A. No 321253) and the European HiPEAC Network of Excellence. The Mont-Blanc project has received funding from the European Unions Horizon 2020 research and innovation programme under grant agreement No 671697.

References

[1] John E. Stone, David Gohara, and Guochun Shi. OpenCL: A parallel programming standard for heterogeneous computing systems. *IEEE Des. Test*, 12(3):66–73, May

2010.

[2] Borja Pérez, José Luis Bosque, and Ramón Beivide. Simplifying programming and load balancing of data parallel applications on heterogeneous systems. In *Proceedings of the 9th Annual Workshop on General Purpose Processing Using Graphics Processing Unit*, GPGPU '16, pages 42–51, New York, NY, USA, 2016. ACM.

[3] M. G. Lopez, V. V. Larrea, W. Joubert, O. Hernandez, A. Haidar, S. Tomov, and J. Dongarra. Towards achieving performance portability using directives for accelerators. In *2016 Third Workshop on Accelerator Programming Using Directives (WACCPD)*, pages 13–24, Nov 2016.

[4] X. Xiao, S. Hirasawa, H. Takizawa, and H. Kobayashi. The importance of dynamic load balancing among openmp thread teams for irregular workloads. In *2016 Fourth International Symposium on Computing and Networking (CANDAR)*, pages 529–535, Nov 2016.

[5] A. Lastovetsky, L. Szustak, and R. Wyrzykowski. Model-based optimization of eulag kernel on intel xeon phi through load imbalancing. *IEEE Transactions on Parallel and Distributed Systems*, 28(3):787–797, March 2017.

[6] F. Zhang, J. Zhai, B. He, S. Zhang, and W. Chen. Understanding co-running behaviors on integrated cpu/gpu architectures. *IEEE Transactions on Parallel and Distributed Systems*, 28(3):905–918, March 2017.

[7] Prasanna Pandit and R. Govindarajan. Fluidic kernels: Cooperative execution of opencl programs on multiple heterogeneous devices. In *Proceedings of Annual IEEE/ACM International Symposium on Code Generation and Optimization*, CGO '14, pages 273:273–273:283, New York, NY, USA, 2014. ACM.

[8] James Jeffers and James Reinders. *Intel Xeon Phi Coprocessor High Performance Programming*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1st edition, 2013.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

Random sample sizes in one-way fixed effects models

Célia Nunes[1], Gilberto Capistrano[2], Dário Ferreira[1], Sandra S. Ferreira[1] and João T. Mexia[3]

[1] *Department of Mathematics and Center of Mathematics and Applications, University of Beira Interior, Covilhã, Portugal*

[2] *School of Business and Development of Excellence - ENDEX, Pouso Alegre, Brazil*

[3] *Center of Mathematics and its Applications, Faculty of Science and Technology, New University of Lisbon, Portugal*

emails: celian@ubi.pt, gilbertocapistrano@gmail.com, dario@ubi.pt, sandraf@ubi.pt, jtm@fct.unl.pt

Abstract

Analysis of variance (ANOVA) is one of the most frequently used statistical analysis in several research areas, namely in medical research. Despite its wide use, it has been applied assuming that sample dimensions are known. In this work we aim to carry out ANOVA like analysis of one-way fixed effects models, to situations where the samples sizes may not be previously known. Assuming that the samples were generated by Poisson counting processes we obtain the unconditional distribution of the test statistic, under the assumption that we have random sample sizes. The applicability of the proposed approach is illustrated considering a real data example on cancer registries. The results obtained suggested that false rejections may be avoid by applying our approach.

Key words: ANOVA, random sample sizes, fixed effects models, counting processes, cancer registries.
MSC 2000: 62J12, 62J10, 62J99

1 Introduction

Analysis of variance (ANOVA) is one of the most frequently used statistical analysis in practical applications. It is routinely used in several research areas, namely in medical research. Despite its wide use, it has been applied assuming that sample dimensions are

 ISBN: 978-84-617-8694-7

known. In this work we aim to carry out ANOVA like analysis of one-way fixed effects models, to situations where the samples sizes may not be previously known. This often occurs when there is a fixed time span for collecting the observations. An illustrative example of this is the collection of observations during a fixed time period in a study comparing, for example, several pathologies of patients arriving at a hospital, see e.g. [6, 7].

In these situations it is more appropriate, assuming there are m different levels, to consider the sample sizes as realizations, $n_1, ..., n_m$, of independent random variables, $N_1, ..., N_m$, [4, 5, 6, 7, 8].

This new approach must be based on an adequate choice of the distribution of $N_1, ..., N_m$. We assume that the numbers collected in non overlapping intervals are independent and simultaneous arrivals are not to be expected. We are thus led to consider, possible non homogeneous, Poisson counting processes. So for fixed collection periods our sample sizes $N_1, ..., N_m$ will have Poisson distribution with parameters $\lambda_1, ..., \lambda_m$. We put $N_i \sim P(\lambda_i)$, $i = 1, ..., m$. Moreover $n = \sum_{i=1}^{m} n_i$ will be a realization of the random variable

$$N = \sum_{i=1}^{m} N_i$$

which, through independence of N_i, $i = 1, ..., m$,

$$N \sim P(\lambda),$$

with $\lambda = \sum_{i=1}^{m} \lambda_i$. Furthermore the vector $\boldsymbol{n} = (n_1, ..., n_m)'$ will be a realization of $\boldsymbol{N} = (N_1, ..., N_m)'$.

We are interested in testing the hypothesis

$$H_0 : \mu_1 = ... = \mu_m,$$

which may be rewritten as

$$H_0 : \boldsymbol{A}\boldsymbol{\mu} = \boldsymbol{0}, \tag{1}$$

where $\boldsymbol{\mu}$ is the mean vector of the treatment means with components $\mu_1, ..., \mu_m$, and $\boldsymbol{A} = [\boldsymbol{I}_{m-1} | -\boldsymbol{1}_{m-1}]$, with $\boldsymbol{I}_c$ the $c \times c$ identity matrix and $\boldsymbol{1}_c$ the vector with c components equal to 1.

In the next section we present the test statistics and their conditional and unconditional distributions, under the assumption that we have random sample sizes. In section 3 the presented approach is illustrated through an application on real medical data, using cancer registries. We conclude this work with some closing remarks.

2 Statistic and their distributions

When $N_i = n_i$, $i = 1, ..., m$, we have the samples $Y_{i,1}, ..., Y_{i,n_i}$, $i = 1, ..., m$, with averages $\overline{Y}_{i,\bullet}$, $i = 1, ..., m$. Assuming that the observations are normal and independent with

 ISBN: 978-84-617-8694-7

variance σ^2, when $N_i = n_i$, $i = 1, ..., m$, the vector of treatment means, $\boldsymbol{Y}_\bullet$, which has components $\overline{Y}_{1,\bullet}, ..., \overline{Y}_{m,\bullet}$, will be normal with mean vector $\boldsymbol{\mu}$ and variance-covariance matrix $\sigma^2 D(\frac{1}{n_1}, ..., \frac{1}{n_m})$, where $D(\frac{1}{n_1}, ..., \frac{1}{n_m})$ is the diagonal matrix with principal elements $\frac{1}{n_1}, ..., \frac{1}{n_m}$.

So, when $N_i = n_i$, $i = 1, ..., m$, see for instance [2, 3], the sum of squares for testing the hypothesis $H_0 : \boldsymbol{A\mu} = 0$ will be

$$S_{num} = (\boldsymbol{AY}_\bullet)' \left(\boldsymbol{A} D \left(\frac{1}{n_1}, ..., \frac{1}{n_m} \right) \boldsymbol{A}' \right)^{-1} (\boldsymbol{AY}_\bullet), \tag{2}$$

which corresponds to the product by σ^2 of a noncentral chi-square with $g = m - 1$ degrees of freedom and non-centrality parameter

$$\delta(n) = \frac{1}{\sigma^2} (\boldsymbol{A\mu})' \left(\boldsymbol{A} D \left(\frac{1}{n_1}, ..., \frac{1}{n_m} \right) \boldsymbol{A}' \right)^{-1} (\boldsymbol{A\mu}). \tag{3}$$

We put $S_{num} \sim \sigma^2 \chi^2_{g,\delta(n)}$.

The sum of the sums for the errors will be given by, see e.g [1] and [9],

$$S = \sum_{i=1}^{m} \sum_{k=1}^{n_i} (Y_{i,k} - \overline{Y}_{i,\bullet})^2.$$

Moreover S will be the product by σ^2 of a central chi-square with

$$g(n) = n - m$$

degrees of freedom, $S \sim \sigma^2 \chi^2_{g(n)}$, and will be conditionally independent from S_{num}.

Therefore, when $N = n$, the conditional distribution of

$$\Im = \frac{S_{num}}{S}$$

will be a noncentral $\overline{F}$ distribution, which correspond to the distribution of the quotient of independent chi-squares with g and $g(n)$ degrees of freedom and non-centrality parameters $\delta(n)$ and 0, denoted by $\overline{F}(.|g, g(n), \delta(n))$.

Given $N = n$, when H_0 holds, $\delta(n) = 0$ and the conditional distribution of $\Im$ will be a central $\overline{F}$ distribution with g and $g(n)$ degrees of freedom, $\overline{F}(z|g, g(n))$.

2.1 The unconditional distribution

For carrying out the inference we will assume that $N_i \geq n_i^\bullet$, $i = 1, ..., m$, which means that we have a minimum dimension for each sample. In this case the global minimum dimension

 ISBN: 978-84-617-8694-7

will be $n^\bullet = \sum_{i=1}^m n_i^\bullet$. So, since we have m different treatments, denoting $\boldsymbol{n}^\bullet = (n_1^\bullet, ..., n_m^\bullet)'$, we consider

$$p_{\boldsymbol{n}^\bullet}(n) = pr(N = n|\boldsymbol{N} \geq \boldsymbol{n}^\bullet) = \sum_{n_1=n_1^\bullet}^{n-\sum_{i=2}^m n_i^\bullet} ... \sum_{n_\ell=n_\ell^\bullet}^{n-(\sum_{i=1}^{\ell-1} n_i+\sum_{i=\ell+1}^m n_i^\bullet)} ... \sum_{n_m=n-\sum_{i=1}^{m-1} n_i}^{n-\sum_{i=1}^{m-1} n_i} pr(\boldsymbol{N} = \boldsymbol{n}|\boldsymbol{N} \geq \boldsymbol{n}^\bullet),\ n_i = n_i^\bullet, ...,\ i = 1, ..., m, \tag{4}$$

where, through the independence of N_i, $i = 1, ..., m$,

$$pr(\boldsymbol{N} = \boldsymbol{n}|\boldsymbol{N} \geq \boldsymbol{n}^\bullet) = \prod_{i=1}^m pr(N_i = n_i|N_i \geq n_i^\bullet) = \prod_{i=1}^m \frac{pr(N_i = n_i)}{pr(N_i \geq n_i^\bullet)}$$
$$= \prod_{i=1}^m \frac{e^{-\lambda_i}(\lambda_i^{n_i}/n_i!)}{1 - \sum_{u_i=0}^{n_i^\bullet-1} e^{-\lambda_i}(\lambda_i^{u_i}/u_i!)} = \prod_{i=1}^m \frac{\lambda_i^{n_i}}{n_i!(e^{\lambda_i} - \sum_{u_i=0}^{n_i^\bullet-1} \frac{\lambda_i^{u_i}}{u_i!})},\ n_i = n_i^\bullet, \ldots,\ i = 1, \ldots, m. \tag{5}$$

Thus the unconditional distribution of $\Im$, when H_0 holds, will be given by, see e.g. [4] and [7],

$$\overline{\overline{F}}(z) = \sum_{n=n^\bullet}^\infty pr(N = n|\boldsymbol{N} \geq \boldsymbol{n}^\bullet)\overline{F}(z|g, g(n)) = \sum_{n=n^\bullet}^\infty p_{\boldsymbol{n}^\bullet}(n)\overline{F}(z|g, g(n)), \tag{6}$$

considering $p_{\boldsymbol{n}^\bullet}(n)$ as defined in (4).

When we known that $N \leq \overline{n}$, we may not consider in (6) the terms for $n > \overline{n}$, and we have $\overline{\overline{F}}(z)$ bounded by

$$\overline{\overline{F}}_{\overline{n}}(z) \leq \overline{\overline{F}}(z) \leq \overline{\overline{F}}^*_{\overline{n}}(z), \tag{7}$$

where

$$\overline{\overline{F}}_{\overline{n}}(z) = \sum_{n=n^\bullet}^{\overline{n}} pr(N = n|\boldsymbol{N} \geq \boldsymbol{n}^\bullet)\overline{F}(z|g, g(n)) \tag{8}$$

and

$$\overline{\overline{F}}^*_{\overline{n}}(z) = \overline{\overline{F}}_{\overline{n}}(z) + \sum_{n=\overline{n}+1}^\infty pr(N = n|\boldsymbol{N} \geq \boldsymbol{n}^\bullet). \tag{9}$$

So $\overline{n}$ denote the upper bound needed to control the truncation error of the unconditional distribution $\overline{\overline{F}}(z)$.

 ISBN: 978-84-617-8694-7

It is important to note that

$$\sum_{n=\overline{n}+1}^{\infty} pr(N = n|\boldsymbol{N} \geq \boldsymbol{n}^{\bullet}) = 1 - \sum_{n=n^{\bullet}}^{\overline{n}} pr(N = n|\boldsymbol{N} \geq \boldsymbol{n}^{\bullet}). \tag{10}$$

3 Application

In this section we evaluate our approach under a real data example. To construct this experiment we resort to a dataset which was provided by the Brazilian National Cancer Institute (INCA)[1]. The dataset gathers information regarding the age of patients with cancer disease. The data considered is from 2010 and refers to the City of São Paulo, Brazil.

In our model the factor considered is the *Type of Cancer*, with three levels: *Soft tissues of the thorax*, *Intestinal tract* and *Nasal cavity*. Table 1 illustrates the number of patients and the sample mean age for each type of cancer.

Table 1: Number of patients and sample means.

Type of Cancer	Number of patients	Sample means
Soft tissues of the thorax	18	49.50
Intestinal tract	22	61.7727
Nasal cavity	25	62.40

We will test the hypothesis

$$H_0 : \boldsymbol{A\mu} = 0,$$

with

$$\boldsymbol{A} = \begin{bmatrix} 1 & 0 & -1 \\ 0 & 1 & -1 \end{bmatrix}.$$

The numerator of the $\Im$ statistic is now given by

$$S_{num} = (\boldsymbol{AY}_{\bullet})' \left(\boldsymbol{A}D\left(\tfrac{1}{18}, \tfrac{1}{22}, \tfrac{1}{25}\right)\boldsymbol{A}'\right)^{-1} (\boldsymbol{AY}_{\bullet}) \ ,$$

which is, when H_0 holds, the product by σ^2 of a central chi-square with $g = m - 1 = 2$ degrees of freedom, $S_{num} \sim \sigma^2\chi^2_2$. So, we obtain

$$\left(\boldsymbol{A}D\left(\frac{1}{18}, \frac{1}{22}, \frac{1}{25}\right)\boldsymbol{A}'\right)^{-1} = \begin{bmatrix} 13.0154 & -6.0923 \\ -6.0923 & 14.5538 \end{bmatrix} \text{ e } \boldsymbol{Ay}_{\bullet} = \begin{bmatrix} -12.9000 \\ -0.6273 \end{bmatrix},$$

[1] http://www2.inca.gov.br/wps/wcm/connect/inca/portal/home

 ISBN: 978-84-617-8694-7

where $\boldsymbol{y}_\bullet$ has components $\overline{y}_{1,\bullet} = 49.50$; $\overline{y}_{2,\bullet} = 61.7727$; $\overline{y}_{3,\bullet} = 62.40$. Therefore for the numerator of the statistic we obtain

$$S_{num} = 2073.021.$$

The denominator of the statistic is, when $N = n$, the product by σ^2 of a central chi-square with $g(n) = n - 3$ degrees of freedom, $S \sim \sigma^2 \chi^2_{n-3}$. In this case we obtain

$$S = \sum_{j=1}^{18}(y_{1,j} - \overline{y}_{1,\bullet})^2 + \sum_{j=1}^{22}(y_{2,j} - \overline{y}_{2,\bullet})^2 + \sum_{j=1}^{25}(y_{3,j} - \overline{y}_{3,\bullet})^2 = 26632.364.$$

So, the statistic's value, $\Im_{Obs}$, is given by

$$\Im_{Obs} = \frac{2073.021}{26632.364} = 0.07784.$$

Given $N = n$, when H_0 holds, the common conditional distribution of $\Im$ is a central $\overline{F}$ distribution with $g = 2$ and $g(n) = 65 - 3 = 62$ degrees of freedom, since $n = 65$, $\overline{F}(z|2, 62)$.

The quantiles of the conditional and unconditional distribution were performed using R software. The quantiles, $z_{1-\alpha}$, of the conditional distribution are given in Table 2, which are obtained considering $z_\alpha = \frac{2}{62} f_{1-\alpha,2,62}$, where $f_{1-\alpha,2,62}$ corresponds to the $1-\alpha$ quantile of a central F distribution with 2 and 62 degrees of freedom. So we can conclude that using the common approach we reject H_0 for $\alpha = 0.1$, since $\Im_{Obs} > z_{1-\alpha}$, and we do not reject for $\alpha = 0.05$ and 0.01.

Table 2: The quantiles of the conditional distribution.

Values of α	0.1	0.05	0.01
$z_{1-\alpha}$	0.07711	0.10146	0.16016

To carry out the computation we are led to use our previous information assuming that $\lambda_i,\ i = 1, 2, 3$ correspond to the average numbers of occurrences per year. So we take $\lambda_1 = 18$; $\lambda_2 = 22$ and $\lambda_3 = 25$, which means that $N_1 \sim P(18)$, $N_2 \sim P(22)$ and $N_3 \sim P(25)$. Let us also assume that we have at least 5 observations per level, which means that $n_i^\bullet = 5,\ i = 1, 2, 3$, $n^\bullet = \sum_{i=1}^3 n_i = 15$ and consequently $\boldsymbol{n}^\bullet = (5, 5, 5)'$.

To compute the quantiles for the unconditional distribution we obtain the minimum value $\overline{n} = 97$ (considering in expression (8)) such that

$$\sum_{n=\overline{n}+1}^{\infty} pr(N = n|\boldsymbol{N} \geq \boldsymbol{n}^\bullet) = 1 - \sum_{n=n^\bullet}^{\overline{n}} pr(N = n|\boldsymbol{N} \geq \boldsymbol{n}^\bullet) < 10^{-4}. \tag{11}$$

 ISBN: 978-84-617-8694-7

Therefore, the infinite serie in (6) is truncated not considering the terms for $n > 97$. So, when H_0 holds, we have the distribution

$$\begin{aligned}\overline{\overline{F}}_{\overline{n}}(z) &= \sum_{n=15}^{97} p_{\boldsymbol{n}^\bullet}(n)\overline{F}(z|2,n-3) \\ &= \sum_{n=15}^{97}\sum_{n_1=5}^{n-10}\sum_{n_2=5}^{n-(n_1+5)}\sum_{n_3=n-(n_1+n_2)}^{n-(n_1+n_2)} pr(\boldsymbol{N}=\boldsymbol{n}|\boldsymbol{N}\geq\boldsymbol{n}^\bullet)\overline{F}(z|2,n-3),\end{aligned}$$

with

$$pr(\boldsymbol{N}=\boldsymbol{n}|\boldsymbol{N}\geq\boldsymbol{n}^\bullet) = \prod_{i=1}^{3}\frac{\lambda_i^{n_i}/n_i!}{e^{\lambda_i}-\sum_{u_i=0}^{4}\frac{\lambda_i^{u_i}}{u_i!}}.$$

The quantiles, $z^t_{1-\alpha}$, for the probability $1-\alpha$, of this distribution are presented in Table 3. Since $\Im_{Obs} < z^t_{1-\alpha}$, we can conclude that we do not reject H_0 for the usual level of significance. So these results lead us to take a contrary decision that we had taken using the conditional approach, for $\alpha = 0.1$.

Table 3: The quantiles of the unconditional distribution.

Values of α	0.1	0.05	0.01
$z^t_{1-\alpha}$	0.07856	0.10341	0.16341

4 Closing remarks

In this paper we resorted to the Poisson distributions as the adequate choice for the distributions of the sample sizes when such are unknown. We open room to a new field based on the assumption that we have a minimum dimension for each sample considering one-way ANOVA. Through the application presented we can confirm that the quantiles may exceed those of the common ANOVA when random sample sizes are considered, giving relevance to the unconditional approach in avoiding false rejections.

Acknowledgements

This work was partially supported by national founds of FCT-Foundation for Science and Technology under UID/MAT/00212/2013 and UID/MAT/00297/2013.

References

[1] A.I. Khuri, T. Mathew and B.K. Sinha, *Statistical Tests for Mixed Linear Models*, Wiley series in Probability and Statistics, John Wiley & Sons, New York, 1998.

[2] E.L. Lehmann, *Testing statistical hypotheses*, John Wiley & Sons, Inc., New York, 1959.

[3] J.T. Mexia, *Best linear unbiased estimates, duality of F tests and the Scheffé multiple comparison method in presence of controled heterocedasticity*, Computational Statistics & Data Analysis **10**(3) (1990) 271–281.

[4] J.T. Mexia, C. Nunes, D. Ferreira, S.S. Ferreira and E. Moreira, *Orthogonal fixed effects ANOVA with random sample sizes*, Proceedings of the 5th International Conference on Applied Mathematics, Simulation, Modelling (ASM'11) (2011) 84-90.

[5] E.E. Moreira, J.T. Mexia and C.E. Minder, *F tests with random sample size. Theory and applications*, Statistics & Probability Letters **83**(6) (2013) 1520-1526.

[6] C. Nunes, D. Ferreira, S.S. Ferreira and J.T. Mexia, *F-tests with a rare pathology*, Journal of Applied Statistics **39**(3) (2012) 551-561.

[7] C. Nunes, D. Ferreira, S.S. Ferreira and J.T. Mexia, *Fixed effects ANOVA: an extension to samples with random size*, Journal of Statistical Computation and Simulation **84**(11) (2014) 2316-2328.

[8] C. Nunes, G. Capistrano, D. Ferreira, S.S. Ferreira and J.T. Mexia, *One-Way Fixed Effects ANOVA with Missing Observations*, Proceedings of the 12th International Conference on Numerical Analysis and Applied Mathematics, AIP Conf. Proc. **1648** (2015) 110008-110011

[9] S.R. Searle, G. Casella and C.E. McCulloch, *Variance Components*, Wiley series in Probability and statistics, John Wiley & Sons, New York, 1992.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

Preconditioning of Linear Systems Using LU Factors

Takeshi Ogita[1]

[1] *Division of Mathematical Sciences, Tokyo Woman's Christian University*

emails: ogita@lab.twcu.ac.jp

Abstract

This study aims to compute accurate numerical solutions of ill-conditioned linear systems. For this purpose, we have developed several preconditioning methods [2, 3]. We have shown that, using such preconditioning methods, the condition number of the coefficient matrix can be reduced efficiently, and an accurate solution of the linear system can be obtained. However, the computational cost for such preconditioning methods is considerably larger than the standard numerical algorithm such as LU factorization, since some matrix multiplication in higher-precision arithmetic is required. In this study, we modify this point by exploiting the structure of the coefficient matrix, and develop an accurate and efficient algorithm for solving ill-conditioned linear systems. As a result, the computational cost for the preconditioning can significantly be reduced with similar quality to the previous methods.

Key words: linear systems, preconditioning, accurate numerical algorithms
MSC 2000: 65F05, 65G50

1 Introduction

Let us consider solving a linear system

$$Ax = b, \quad A \in \mathbb{R}^{n\times n},\ b \in \mathbb{R}^n \tag{1}$$

using floating-point arithmetic. To solve (1) is one of the significant tasks in scientific computing.

Let $\mathbf{u}$ denote the rounding error unit, e.g., $\mathbf{u} = 2^{-53} \approx 10^{-16}$ for IEEE 754 binary64. The purpose of this study is to develop efficient methods for solving (1) in ill-conditioned cases such that

$$\kappa(A) \gtrsim \mathbf{u}^{-1},$$

 ISBN: 978-84-617-8694-7

where $\kappa(A)$ denotes the condition number of A defined by $\kappa(A) := \|A\| \cdot \|A^{-1}\|$.

When using ordinary floating-point arithmetic, such as IEEE 754 binary64 arithmetic, the difficulty of the problem is as follows. Let x^* be the exact solution of (1). Let $\widehat{x}$ be a computed solution of (1) obtained by some backward stable algorithm in floating-point arithmetic with the rounding error unit $\mathbf{u}$. Then, as a rule of thumb (cf. e.g., [1]), $\widehat{x}$ satisfies

$$\frac{\|\widehat{x} - x^*\|}{\|x^*\|} \approx p(n) \cdot \kappa(A) \cdot \mathbf{u},$$

where $p(n)$ is a modestly growing function of n. From this, we can see that approximate solutions may become inaccurate if $\kappa(A) \gtrsim \mathbf{u}^{-1}$.

To overcome the problem, we have the following two possibilities:

1. Preconditioning with higher-precision arithmetic
2. Multiple precision arithmetic (MP)

However, MP is significantly slower than ordinary floating-point arithmetic, and we prefer to utilize ordinary floating-point arithmetic in terms of computational speed. Moreover, for modestly well-conditioned problems, standard numerical algorithms often provide sufficiently good results in ordinary floating-point arithmetic. In such cases, the use of MP is not necessary. However, we do not know in advance how many digits suffice to achieve desired result accuracy. Thus, we prefer some preconditioning method utilizing the results obtained by a standard numerical algorithm in ordinary floating-point arithmetic. As a standard numerical algorithm, we assume that LU factorization with partial pivoting is used.

2 Preconditioning

We briefly review an approach of preconditioning methods. We choose some nonsingular matrix $M \in \mathbb{R}^{n \times n}$ such that

$$\kappa(MA) \approx \mathbf{u} \cdot \kappa(A).$$

As a choice of M, we use an approximate inverse of an LU factor of A as in [2, 3].

For simplicity, we omit a permutation matrix according to the partial pivoting. Suppose the Crout's version of LU factorization of A is done in ordinary floating-point arithmetic such that $A \approx LU$. Then, heuristics suggest that $\kappa(A) \approx \kappa(L)$. By setting $M := L^{-1}$, we obtain

$$\kappa(L^{-1}A) \approx \mathbf{u} \cdot \kappa(A)$$

as shown in [4]. In practice, an approximate inverse X_L of L is used instead of L^{-1}. Here, higher-precision arithmetic is necessary for computing $X_L \cdot A$. For example, we can use

 ISBN: 978-84-617-8694-7

high-precision dot product algorithm `Dot2` proposed in [5], which computes dot product as if computed in twice the working precision. Major computational cost in $\mathcal{O}(n^3)$ flops required for the preconditioning is as follows:

- LU factorization $A \approx LU$: $\frac{2}{3}n^3$ flops
- Triangular matrix inversion $X_L \approx L^{-1}$: $\frac{1}{3}n^3$ flops
- Matrix multiplication $X_L \cdot A$ in twice the working precision: $c \cdot n^3$ flops with relatively large constant c such that $10 \leq c \leq 22$.

Therefore, total computational cost for the preconditioning is considerably larger than LU factorization. Thus, we aim to reduce the cost.

Detailed discussions and numerical results will be presented in the talk.

References

[1] N. J. Higham, *Accuracy and Stability of Numerical Algorithms, 2nd ed.*, SIAM, Philadelphia PA, 2002.

[2] Y. Kobayashi, T. Ogita, *A fast and efficient algorithm for solving ill-conditioned linear systems*, JSIAM Letters **7** (2015) 1–4.

[3] Y. Kobayashi, T. Ogita, *Accurate and efficient algorithm for solving ill-conditioned linear systems by preconditioning methods*, Nonlinear Theory and Its Applications, IEICE **7** (2016) 374–385.

[4] T. Ogita, *Accurate matrix factorization: Inverse LU and inverse QR factorizations*, SIAM J. Matrix Anal. Appl. **31** (2010) 2477–2497.

[5] T. Ogita, S. M. Rump, S. Oishi, *Accurate sum and dot product*, SIAM J. Sci. Comput., **26** (2005) 1955–1988.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

Tridiagonal Kernel Enhanced Multivariance Products Representation (TKEMPR) for Univariate Linear Operators: Continuous Singular Value Decomposition

Ayla OKAN[1] and Metin DEMİRALP[2]

[1] *Informatics Institute, Computational Science and Engineering Department, İstanbul Technical University*

emails: aylaokan@itu.edu.tr, metin.demiralp@gmail.com

Abstract

This paper is concerned with the analysis of the linear integral operators whose kernels are a square integrable bivariate function over the square $[a,b]^2$. The basic aim is to use high dimensional modeling. Tridiagonal Kernel Enhanced Multivariance Products Representation (TKEMPR) has been a novel extension high dimensional modeling representation (HDMR) which has been derived form of the Enhanced Multivariance Products Representation (EMPR), proposed by Demiralp and his group.

Key words: Linear Integral Operators, Enhanced Multivariance Products Representation, (EMPR), Tridiagonal Kernel Enhanced Multivariance Products Representation, (TKEMPR), Singular Value Decomposition, Decomposition Method
MSC 2000: 41A35, 49R50, 65R20

1 Introduction

In this section, to make a strong background for TKEMPR, we will give the following main lines and basic technicalities of EMPR method.

The Enhanced Multivariance Products Representation (EMPR) for a given N-variate function $f(x_1,...,x_N)$ is given by the finite expansion

$$\begin{aligned} f(x_1,...,x_N) \;=\; & f_0 \prod_{i=1}^{N} s_i(x_i) + \sum_{j=1}^{N} f_j(x_j) \prod_{\substack{i=1 \\ i\neq j}}^{N} s_i(x_i) + \sum_{\substack{j_1,j_2=1 \\ j_1<j_2}}^{N} f_{j_1,j_2}(x_{j_1},x_{j_2}) \prod_{\substack{i=1 \\ i\neq j_1,j_2}}^{N} s_i(x_i) \\ & + \sum_{\substack{j_1,j_2,j_3=1 \\ j_1<j_2<j_3}}^{N} f_{j_1,j_2,j_3}(x_{j_1},x_{j_2},x_{j_3}) \prod_{\substack{i=1 \\ i\neq j_1,j_2,j_3}}^{N} s_i(x_i) + \cdots + f_{1,2,...,N}(x_1,x_2,...,x_N) \end{aligned} \tag{1}$$

 ISBN: 978-84-617-8694-7

where subindexed fs are respectively called constant, univariate, bivariate EMPR components. The s_is appearing in the formulation are univariate support functions[1, 2]. x_j stands for the j-th independent variable remaining in a closed interval on the real line $[\,a_i, b_i\,]$, $i = 1, ..., N$.

The main purpose of this representation is to determine the EMPR components. To this end, one needs to impose certain conditions like Sobol's vanishing integral conditions involving support functions[3]. To determine these terms, the normalization conditions

$$\int_{a_i}^{b_i} dx_i W_i(x_i) = 1, \quad \int_{a_i}^{b_i} dx_i W_i(x_i) \left(s_i(x_i)\right)^2 = 1; \quad 1 \leq i \leq N \tag{2}$$

and the vanishing–under–integration conditions

$$\int_{a_i}^{b_i} dx_i W_i\left(x_i\right) s_i\left(x_i\right) f_{i_1 \ldots i_k}\left(x_{i_1}, ..., x_{i_k}\right) = 0, \quad x_i \in \left(x_{i_1}, ..., x_{i_k}\right), \quad 1 \leq i \leq k \leq N \tag{3}$$

are assumed to hold. These conditions allow us to obtain the EMPR terms as the following forms

$$f_0 = \int_{a_1}^{b_1} dx_1 W_1(x_1) ... \int_{a_N}^{b_N} dx_N W_N(x_N) \prod_{j=1}^{N} s_j(x_j) f\left(x_1, ..., x_N\right) \tag{4}$$

$$\begin{aligned} f_i(x_i) \;=\; & \int_{a_1}^{b_1} dx_1 W_1(x_1) ... \int_{a_{i-1}}^{b_{i-1}} dx_{i-1} W_{i-1}(x_{i-1}) \int_{a_{i+1}}^{b_{i+1}} dx_{i+1} W_{i+1}(x_{i+1}) ... \int_{a_N}^{b_N} dx_N W_N(x_N) \\ & \times \prod_{\substack{j=1 \\ j \neq i}}^{N} s_j(x_j) f(x_1, ..., x_N) - s_i f_0 \end{aligned} \tag{5}$$

$$\begin{aligned} f_{i_1 i_2}(x_{i_1}, x_{i_2}) \;=\; & \int_{a_1}^{b_1} dx_1 W_1(x_1) ... \int_{a_{i_1-1}}^{b_{i_1-1}} dx_{i_1-1} W_{i_1-1}(x_{i_1-1}) \int_{a_{i_1+1}}^{b_{i_1+1}} dx_{i_1+1} W_{i_1+1}(x_{i_1+1}) ... \\ & \times \int_{a_{i_2-1}}^{b_{i_2-1}} dx_{i_2-1} W_{i_2-1}(x_{i_2-1}) \int_{a_{i_2+1}}^{b_{i_2+1}} dx_{i_2+1} W_{i_2+1}(x_{i_2+1}) ... \int_{a_N}^{b_N} dx_N W_N(x_N) \\ & \times \prod_{\substack{j=1 \\ j \neq i_1, i_2}}^{N} s_j(x_j) f(x_1, ..., x_N) - s_{i_1}(x_{i_1}) f_{i_1}(x_{i_1}) - s_{i_2}(x_{i_2}) f_{i_2}(x_{i_2}) \\ & - s_{i_1}(x_{i_1}) s_{i_2}(x_{i_2}) f_0 \end{aligned} \tag{6}$$

EMPR is like most of its kind based on divide–and–conquer philosophy. For a given N–variate function, there are 2^N EMPR undetermined components at the right side of the expansion. However in practise, the main purpose of this matter is not to evaluate all of them to lower the computational cost. Then, EMPR turns into an approximation method.

 ISBN: 978-84-617-8694-7

2 Tridiagonal Kernel Enhanced Multivariance Products Represantation (TKEMPR)

In this work, we specifically focus on the case of bivariate functions which are assumed to be square integrable over the interval $[\,a,b\,]^2$. These bivariate functions can be taken into the consideration as the kernel the univariate integral operators also.

By using bivariate EMPR, the kernel bivariate target function $K(x,y)$ can be rewritten with the following expansion

$$K(x,y) = \mathcal{K}_0 u(x)v(y) + \mathcal{K}_1(x)v(y) + u(x)\mathcal{K}_2(y) + \mathcal{K}_{1,2}(x,y) \tag{7}$$

where subindexed $\mathcal{K}$s are called constant, univariate, and bivariate EMPR components. $u(x)$ and $v(y)$ are the given support functions.

Predefined normalization and vanishing conditions enable us to evaluate the EMPR terms of the kernel $K(x,y)$ uniquely as

$$\mathcal{K}_0 = \int_{a_1}^{b_1} dx \int_{a_2}^{b_2} dy u(x)v(y)W_1(x)W_2(y)K(x,y), \tag{8}$$

$$\mathcal{K}_1(x) = \int_{a_2}^{b_2} dy v(y)W_2(y)K(x,y) - \mathcal{K}_0 u(x), \tag{9}$$

$$\mathcal{K}_2(y) = \int_{a_1}^{b_1} dx u(x)W_1(x)K(x,y) - \mathcal{K}_0 v(y), \tag{10}$$

and finally

$$\mathcal{K}_{1,2}(x,y) = K(x,y) - \mathcal{K}_0 u(x)v(y) - \mathcal{K}_1(x)v(y) - \mathcal{K}_2(y)u(x). \tag{11}$$

These four equations can also be rewritten in much more concise form by using the following integral operators

$$\mathcal{K}_1(x) = \left(\widehat{I}_x - \widehat{P}_u\right) \int_{a_2}^{b_2} dy v(y)W_2(y)K(x,y), \tag{12}$$

$$\mathcal{K}_2(y) = \left(\widehat{I}_y - \widehat{P}_v\right) \int_{a_1}^{b_1} dx u(x)W_1(x)K(x,y), \tag{13}$$

$$\mathcal{K}_{1,2}(x,y) = \left(\widehat{I}_x - \widehat{P}_u\right)\left(\widehat{I}_y - \widehat{P}_v\right) K(x,y) \tag{14}$$

 ISBN: 978-84-617-8694-7

where $\widehat{I}_x$ and $\widehat{I}_y$ are unit operators in the spaces of univariate functions depending on x and y respectively. $\widehat{P}$ and $\widehat{Q}$ are idempotent integral operators which project their operands into the one dimensional space spanned by the related $u(x)$ and $v(y)$ functions respectively. When we take x and y as the variable of interest, we can define these integral operators as the following form

$$\widehat{P}_u f(x) \equiv u(x) \int_{a_1}^{b_1} d\xi u(\xi) W_1(\xi) f(\xi), \qquad \widehat{P}_v g(y) \equiv v(y) \int_{a_2}^{b_2} d\eta v(\eta) W_2(\eta) g(\eta) \tag{15}$$

Herein, it is also possible to define the inner product through the integrals over the x and y dependent functions separately under the related weight functions as

$$(f_1, f_2)_x \equiv \int_{a_1}^{b_1} dx f_1(x) W_1(x) f_2(x), \qquad (g_1, g_2)_y \equiv \int_{a_2}^{b_2} dy g_1(y) W_2(y) g_2(y) \tag{16}$$

These definitions allow us to get the following inner product equalities as follows

$$(u, \mathcal{K}_1)_x = 0, \qquad (v, \mathcal{K}_2)_y = 0 \tag{17}$$

$$(u, \mathcal{K}_{1,2})_x = 0, \qquad \forall y \in [\, a_2, b_2 \,], \qquad (v, \mathcal{K}_{1,2})_y = 0, \qquad \forall x \in [\, a_1, b_1 \,] \tag{18}$$

The operators $\left(\widehat{I}_x - \widehat{P}\right)$ and $\left(\widehat{I}_y - \widehat{Q}\right)$ project to the subspaces orthogonal to the axes spanned by $u(x)$ and $v(y)$ functions respectively. To construct the recursive structure, we can rewrite the initial support functions for the target bivariate function with u_1 and v_1 separately. Then, we can create new $u_2(x)$ and $v_2(y)$ generated support functions to use them on bivariate $\mathcal{K}_{1,2}$ term which becomes as the target function to apply the recursive EMPR method for the next step.

$$u_1(x) \equiv u(x), \qquad u_2(x) \equiv \frac{1}{\beta_1} \mathcal{K}_1(x), \qquad \beta_1 \equiv (K_1, K_1)^{\frac{1}{2}} \tag{19}$$

$$v_1(y) \equiv v(y), \qquad v_2(y) \equiv \frac{1}{\gamma_1} \mathcal{K}_2(y), \qquad \gamma_1 \equiv (K_2, K_2)^{\frac{1}{2}} \tag{20}$$

Then, by using the first recursion step, the target bivariate $K(x, y)$ function can be given as the following expansion

$$K(x, y) = \alpha_1 u_1(x) v_1(y) + \beta_1 u_2(x) v_1(y) + \gamma_1 u_1(x) v_2(y) + K_{1,2}(x, y) \tag{21}$$

where $\alpha_1 \equiv K_0$. The target and obtained bivariate functions can be redefined for recursion as

$$K^{(1)}(x, y) \equiv K(x, y), \qquad K^{(2)}(x, y) \equiv K_{1,2}(x, y) \tag{22}$$

 ISBN: 978-84-617-8694-7

Then, for the given bivariate $K^{(1)}(x,y)$ kernel function, recursive EMPR expansion for the first step is

$$K^{(1)}(x,y) = \alpha_1 u_1(x)v_1(y) + \beta_1 u_2(x)v_1(y) + \gamma_1 u_1(x)v_2(y) + K^{(2)}(x,y) \tag{23}$$

where

$$\alpha_1 = \int_{a_1}^{b_1} dx \int_{a_2}^{b_2} dy u_1(x)v_1(y)W_1(x)W_2(y)K^{(1)}(x,y), \tag{24}$$

$$\beta_1 \equiv \left(\int_{a_1}^{b_1} dx K_x^{(1)}(x) W_1(x) \left[\widehat{I}_x - \widehat{P}_{u_1} \right] K_x^{(1)}(x) \right)^{\frac{1}{2}},$$

$$K_x^{(1)}(x) \equiv \int_{a_2}^{b_2} dy v_1(y) W_2(y) K^{(1)}(x,y), \tag{25}$$

and

$$\gamma_1 \equiv \left(\int_{a_2}^{b_2} dy K_y^{(1)}(y) W_2(y) \left[\widehat{I}_y - \widehat{P}_{v_1} \right] K_y^{(1)}(y) \right)^{\frac{1}{2}}$$

$$K_y^{(1)}(y) \equiv \int_{a_1}^{b_1} dx u_1(x) W_1(x) K^{(1)}(x,y), \tag{26}$$

The generated support functions which will be used for the next step can be written by using integral operators as follows

$$u_2(x) \equiv \frac{1}{\beta_1} \left(\widehat{I}_x - \widehat{P}_{u_1} \right) K_x^{(1)}(x), \qquad v_2(y) \equiv \frac{1}{\gamma_1} \left(\widehat{I}_y - \widehat{P}_{v_1} \right) K_y^{(1)}(y) \tag{27}$$

$K^{(2)}(x,y)$, the reminder bivariate term, which will be used as the new target function for the next step can be given as the following concise form

$$K^{(2)}(x,y) = \left(\widehat{I}_x - \widehat{P}_{u_1} \right) \left(\widehat{I}_y - \widehat{P}_{v_1} \right) K^{(1)}(x,y) \tag{28}$$

More generally, we can write this expansion for the jth step recursion

$$K^{(j)}(x,y) = \alpha_j u_j(x)v_j(y) + \beta_j u_{j+1}(x)v_j(y) + \gamma_j u_j(x)v_{j+1}(y) + K^{(j+1)}(x,y). \tag{29}$$

where

$$\alpha_j = \int_{a_1}^{b_1} dx \int_{a_2}^{b_2} dy u_j(x)v_j(y)W_1(x)W_2(y)K^{(j)}(x,y), \tag{30}$$

$$\beta_j \equiv \left(\int_{a_1}^{b_1} dx K_x^{(j)}(x) W_1(x) \left[\widehat{I}_x - \widehat{P}_{u_j} \right] K_x^{(j)}(x) \right)^{\frac{1}{2}}, \quad K_x^{(j)}(x) \equiv \int_{a_2}^{b_2} dy v_j(y) W_2(y) K^{(j)}(x,y) \tag{31}$$

 ISBN: 978-84-617-8694-7

$$\gamma_j \equiv \left(\int_{a_2}^{b_2} dy K_y^{(j)}(y) W_2(y) \left[\widehat{I}_y - \widehat{P}_{v_j}\right] K_y^{(j)}(y)\right)^{\frac{1}{2}}, \quad K_y^{(j)}(y) \equiv \int_{a_1}^{b_1} dx u_j(x) W_1(x) K^{(j)}(x,y) \tag{32}$$

$$u_{j+1}(x) \equiv \frac{1}{\beta_j}\left(\widehat{I}_x - \widehat{P}_{u_j}\right) K_x^{(j)}(x), \qquad v_{j+1}(y) \equiv \frac{1}{\gamma_j}\left(\widehat{I}_y - \widehat{P}_{v_j}\right) K_y^{(j)}(y), \tag{33}$$

The new target bivariate function given in (29) satisfies the following concise form

$$K^{(j+1)}(x,y) = \left(\widehat{I}_x - \widehat{P}_{u_j}\right)\left(\widehat{I}_y - \widehat{P}_{v_j}\right) K^{(j)}(x,y) \tag{34}$$

The expansion given in (23) can be established by all these recursions and define

$$\begin{aligned} K^{(1)}(x,y) &= \sum_{j=1}^{n} \left(\alpha_j u_j(x) v_j(y) + \beta_j u_{j+1}(x) v_j(y) + \gamma_j u_j(x) v_{j+1}(y)\right) \\ &+ K^{(n+1)}(x,y) \end{aligned} \tag{35}$$

where

$$\begin{aligned} K^{(n+1)}(x,y) &= \left[\prod_{i=1}^{n}\left(\widehat{I}_x - \widehat{P}_{u_i}\right)\right]\left[\prod_{i=1}^{n}\left(\widehat{I}_y - \widehat{P}_{v_i}\right)\right] K^{(1)}(x,y) \\ &= \left(\widehat{I}_x - \sum_{i=1}^{n}\widehat{P}_{u_i}\right)\left(\widehat{I}_y - \sum_{i=1}^{n}\widehat{P}_{v_i}\right) K(x,y). \end{aligned} \tag{36}$$

When n grows unboundedly, it becomes quite noticeable that $K^{(n+1)}(x,y)$ tends to vanish because of the idempotency and mutual annihilating properties of the projection operators over the $u(x)$ and $v(y)$ respectively. By this way, we have

$$K(x,y) = \sum_{j=1}^{\infty} \left(\alpha_j u_j(x) v_j(y) + \beta_j u_{j+1}(x) v_j(y) + \gamma_j u_j(x) v_{j+1}(y)\right). \tag{37}$$

which we call TKEMPR for the kernel function[4, 5]. This method also can be considered as the concise matrix format like three matrix product representation as follows

$$K(x,y) = \mathbf{u}(x)^T \mathbf{\Sigma} \mathbf{v}(y) \tag{38}$$

whose kernel $\mathbf{\Sigma}$ is tridiagonal denumerable infinite square matrix and $\mathbf{u}(x)$ and $\mathbf{v}(y)$ are denumerable infinite vectors. These vectors are composed of the elements $u_i(x)$ and $v_i(y)$ respectively. Kernel matrix $\mathbf{\Sigma}$ has αs, βs and γs as the downward ordered elements of the main diagonal, and, lower nearest neighbour of the main diagonal, and, upper nearest neighbour of the main diagonal respectively [6, 7, 8].

 ISBN: 978-84-617-8694-7

3 Continuous Singular Value Decomposition

The problem considered in this work is described as follows:

- For a given bivariate function which is also the kernel of the considered linear operator, a concise three matrix factor product decomposition[9, 10, 11] has been constructed by using recently developed decomposition, TKEMPR.
- The focused linear operator can be decomposed by using this square integrable kernel function's decomposition through Continuous Singular Value Decomposition.
- And as the final stage, we exhibit the main lines and relationships of these two decompositions.

For these reasons, let us now focus on the following linear operator decomposition

$$\widehat{\mathcal{I}}f(x) \equiv \int_a^b dy K(x,y) f(y) \tag{39}$$

where $f(x)$ belongs to the linear vector space of univariate functions which are square integrable over $[a,b]$ and $K(x,y)$ is the kernel function, assuming to be square integrable over $[a,b]^2$[13, 14]. For this work these functions are assumed to be real–valued to facilitate the analysis.

In the symmetric kernel case, the integral operator $\widehat{\mathcal{I}}$ given in (39) is self–adjoint. Therefore, there is no need to define its adjoint. Also, the left and right eigenfunctions are the same. In the nonsymmetric kernel we can focus on "Continouos Singular Value Decomposition" to protect the real–valued case for eigenvalues[13, 14, 15]. These lead us to write the following equations

$$\widehat{\mathcal{I}}v(x) = \sigma u(x), \qquad \widehat{\mathcal{I}}^{\dagger}u(x) = \sigma v(x) \tag{40}$$

Herein, we assume that there is nonsymmetric kernel case. By using the general definiton of adjoint operator, which can be found in scientific literature, we can get

$$\widehat{\mathcal{I}}^{\dagger}\widehat{\mathcal{I}}v(x) = \sigma^2 v(x), \qquad \widehat{\mathcal{I}}\widehat{\mathcal{I}}^{\dagger}u(x) = \sigma^2 u(x) \tag{41}$$

Here we can use the following notation

$$\widehat{\mathcal{I}}_L v(x) = \int_a^b dy K_L(x,y)\sigma^2 v(x), \qquad \widehat{\mathcal{I}}_R u(x) = \int_a^b dy K_R(x,y)\sigma^2 u(x) \tag{42}$$

such that

$$K_L(x,y) \equiv \int_a^b d\eta K(\eta,x)K(\eta,y), \qquad K_R(x,y) \equiv \int_a^b d\eta K(x,\eta)K(y,\eta) \tag{43}$$

 ISBN: 978-84-617-8694-7

The subscripts L and R symbolize the words "left" and "right" respectively. It is obvious that $K_L(x,y)$ and $K_R(x,y)$ are now symmetric kernels even if $K(x,y)$ is assumed to be nonsymmetric. And also, all σ^2 values are non–negative. This means, σ values which can be chosen as positive valued are defined as "Singular Values" of the operator $\widehat{\mathcal{I}}$ and $u(x)$s and $v(y)$s are called "the Left and Right Singular Functions" respectively. Then, the kernel function can be decomposed as the following form

$$K(x,y) = \sum_{i=1}^{\infty} \sigma_i u_i(x) v_i(y) \tag{44}$$

This expansion can also be given in the following concise form

$$K(x,y) = \mathbf{u}^T(x)\ \mathbf{\Sigma}\ \mathbf{v}(y) \tag{45}$$

where $\mathbf{u}(x)$ and $\mathbf{v}(y)$ are denumerable infinite vectors whose elements are $u_i(x)$s and $v_i(y)$s respectively while the denumerable infinite square matrix $\mathbf{\Sigma}$ is diagonal such that its main diagonal elements are singular values.

It is possible to combine the above mentioned form with the linear operator decomposition given in (39)

$$\widehat{\mathcal{I}} f(x) = \mathbf{u}^T(x)\ \mathbf{\Sigma}\ \int_a^b dy \mathbf{v}(y) f(y), \qquad \widehat{\mathcal{I}}^\dagger f(x) = \mathbf{v}^T(x)\ \mathbf{\Sigma}\ \int_a^b dy \mathbf{u}(y) f(y) \tag{46}$$

Then we can write the following spectral counterparts as

$$\widehat{\mathcal{I}}^\dagger \widehat{\mathcal{I}} f(x) = \mathbf{v}^T(x)\ \mathbf{\Sigma}^2\ \int_a^b dy \mathbf{v}(y) f(y) \tag{47}$$

$$\widehat{\mathcal{I}} \widehat{\mathcal{I}}^\dagger f(x) = \mathbf{u}^T(x)\ \mathbf{\Sigma}^2\ \int_a^b dy \mathbf{u}(y) f(y) \tag{48}$$

where we have used the symmetry in $\mathbf{\Sigma}$ and it leads us to

$$\int_a^b dx \mathbf{u}(x)\mathbf{u}(x)^T = \mathbf{I}_\infty, \qquad \int_a^b dx \mathbf{v}(x)\mathbf{v}(x)^T = \mathbf{I}_\infty \tag{49}$$

where $\mathbf{I}_\infty$ is the denumerable infinite identity matrix.

The basic aim of this work has been the decomposition of linear operator $\widehat{\mathcal{I}}$ which resulted with the iteratively structured eigenvalue problems. These problems can be taken into consideration as matrix diagonalization problems and increase computational cost. It is also possible to use TKEMPR procedures to decompose the linear operator as

$$\widehat{\mathcal{I}} = \sum_{i=1}^{\infty} \left(\alpha_i \widehat{P}_i + \beta_i \widehat{T}_{i+1,i} + \gamma_i \widehat{T}_{i,i+1} \right) \tag{50}$$

 ISBN: 978-84-617-8694-7

such that

$$\widehat{P}_i f(x) \equiv \int_a^b dy u_i(x) v_i(y) f(y), \quad \widehat{T}_{i,j} f(x) \equiv \int_a^b dx u_i(x) v_j(y) f(y), \quad i,j = 1,2,3,... \quad (51)$$

Herein, the resulting operations do not only turn into tridiagonalization instead of diagonalization but also they are recursive instead of iterative procedures [16].

4 Numerical Implementations

Let us begin our numerical experiments by using the below mentioned kernel function

$$K(x,y) = e^{(x+y)} + e^{(-x-y)} \quad (52)$$

which is a symmetric function over the interval $[0,1]$. For convenience in calculations we prefer to use symmetric function. But in the previous section we give the details of integral operator decomposition for nonsymmetric kernel functions. Therefore, the function given in (52) can be considered as being symmetrized function over the interval $[0,1]$. The exact solution of the spectral problem

$$\int_0^1 dy K(x,y)\phi(y) = \lambda\phi(y) \qquad 0 \leq x,y \leq 1 \quad (53)$$

can be given as

$$\lambda = 3.518549337, \qquad \phi(x) = 0.507153053e^x + 0.164328385e^{-x} \quad (54)$$

such that this eigenfunction is normalized over $[0,1]$. These results are taken from[17]. By keeping the greatest eigenvalue λ and corresponding normalized eigenfunction $\phi(x)$ given in (54), we can focus on our decomposition method by using TKEMPR method. In the three factor matrix product type representation of TKEMPR, the kernel matrix $\mathbf{\Sigma}$ is defined in general form as

$$\mathbf{\Sigma} = \begin{bmatrix} \alpha_1 & \gamma_1 & 0 & ... & 0 & ... & ... \\ \beta_1 & \alpha_2 & \gamma_2 & 0 & \ddots & ... & ... \\ 0 & \beta_2 & \alpha_3 & \ddots & 0 & \ddots & ... \\ \vdots & \ddots & \ddots & \ddots & \gamma_{m-1} & 0 & ... \\ 0 & ... & 0 & \beta_{m-1} & \alpha_m & 0 & ... \end{bmatrix} \quad (55)$$

where $\mathbf{\Sigma}$ is tridiagonal denumerable infinite square matrix. The implementation results will be given by using table including α_i values computing by TKEMPR steps.

 ISBN: 978-84-617-8694-7

	$u_1[x]=1,$ $v_1[y]=1$	$u_1[x]=ae^x,$ $v_1[y]=be^{-y}$	$u_1[x]=ae^x$ $v_1[y]=ae^y$	$u_1[x]=a(e^x+e^{-x})$ $v_1[y]=a(e^y+e^{-y})$	$u_1[x]=ae^x+be^{-x}$ $v_1[y]=ae^y+be^{-y}$
α_1	3.35207	3.08616	3.50757	3.49140	3.51839
α_2	0.27166	5.56393×10^{-11}	0.119297	0.13546	0.10847
α_3	0.00313	-2.38675×10^{-12}	8.71743×10^{-11}	-1.18143×10^{-15}	-3.56266×10^{-16}

In this table, a and b values are used for constant real valued coefficients which come from normalization of the relevant support functions. When we analyzed the table, we can suggest that in the first step of TKEMPR we obtain very close values to the greatest eigenvalue. It is given in the last column of the table that, TKEMPR three factor product decomposition method can develop a relationship between the methods solving eigenvalue problem. It is also possible to show the efficiency of this method with the following figure which includes the approximated function, shown with blue colour, obtained from 3×3 three factor product TKEMPR decomposition and the exact kernel function, shown with yellow colour. The norm value for this functions is 1.52656×10^{-32} with the working precision 16.

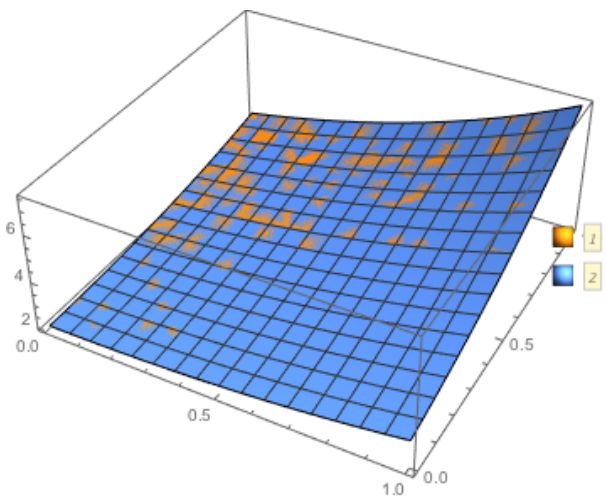


Figure 1: A plot of the exact and approximate functions

5 Concluding Remarks

This work has been designed for the combined utilization of TKEMPR decomposition together with Continouos Singular Value Decomposition. The basic formulation and a confir-

 ISBN: 978-84-617-8694-7

mative implementation. We have much more implementations, some of which will be given during our conference presentation. Now we can use this formulation to test the quality of TKEMPR results in our TKEMPR studies.

References

[1] B. TUNGA AND M. DEMİRALP, *The influence of the support functions on the quality of Enhanced Multivariance Product Representation*, Journal of Mathematical Chemistry, **48**, (2010), 827–840.

[2] A. OKAN, N.A. BAYKARA AND M. DEMİRALP, *Weight optimization in Enhanced Multivariance Product Representation (EMPR) Method*, Int. Conf. on Numer. Anal. and Appl. Math., AIP Conference Proceedings, **1281**, (2010), 1935–1938.

[3] I. M. SOBOL, *Sensitivity estimates for nonlinear mathematical models*, Mathematical Modeling and Computational Experiments 1, (1993), 407–414.

[4] A. OKAN AND M. DEMİRALP, *Tridiagonal Kernel Enhanced Multivariance Products Representation (TKEMPR) for univariate integral operator kernels*, The 2014 International Conference Mathematics and Computers in Sciences and Industry (MCSI 2014), (2014), 195–200, doi: 10.1109/MCSI.2014.26, print ISBN: 978-1-4799-4744-7

[5] A. OKAN AND M. DEMİRALP, *Tridiagonal Kernel Enhanced Multivariance Products Representation (TKEMPR) for outer product sums: Arrowheading EMPR for Kernel (AEMPRK)*, The 12th International Conference of Numerical Analysis and Applied Mathematics (ICNAAM 2014), (2014).

[6] E. DEMİRALP, AND M. DEMİRALP, *Tridiagonal Matrix Enhanced Multivariance Products Representation(TMEMPR) for matrix decomposition*, Proceedings of 14th International Conference Computational and Mathematical Methods in Science and Engineering, **2**, (2014), 446–455.

[7] A. OKAN AND M. DEMİRALP, *Arrowheading Enhanced Multivariance Products Representation for a Kernel (AEMPRK) in a Taylor series expansion*, 11th International Conference of Computational Methods in Sciences and Engineering, ICCMSE 2015, (2015), doi: 10.1063/1.4912452

[8] A. OKAN AND M. DEMİRALP, *Numerical implementations for Tridiagonal Kernel Enhanced Multivariance Products Representation (TKEMPR) method: bivariate case*, The Proceedings of International Journal of Signal Processing,**1**, (2016), 102–107, ISSN: 2367-8984.

[9] M. Demİralp and E. Demİralp, *An Orthonormal decomposition method for multidimensional matrices* AIP Proceedings for the International Conference of Numerical Analysis and Applied Mathematics (ICNAAM 2009), **1168**, (2009), 424–427, doi:http://dx.doi.org/10.1063/1.3241487.

[10] M. Demİralp, and E. Demİralp, *Dimensionality reduction and approximation via space extension and multilinear array decomposition* AIP Proceedings for the International Conference of Computational Methods in Science and Engineering (ICCMSE 2009), Mini Symposium on Recent Developments in Numerical Schemes for Hilbert Space Related Issues in Science and Engineering, (2009), 837–840, doi:http://dx.doi.org/10.1063/1.4771824, 2009.

[11] M. Demİralp, and E. Demİralp, *A new straightforward decomposition method without iteration to approximate matrices via dominant basis matrices*, The International Conference on Scientific Computing - WorldComp09 (CSC09), (2009), 79–83.

[12] E. Demİralp, and M. Demİralp, *Reductive multilinear array decomposition based support functions in Enhanced Multivariance Products Representation (EMPR)* Proceedings for the 1st IEEEAM Conference on Applied Computer Science (ACS), 448–454.

[13] S. Tuna, N. A. Baykara, and M. Demİralp, *Weighted singular value decomposition for folded matrices* Proceedings of the 2nd International Conference on Applied Informatics and Computing Theory (AICT11), (2011), 70–75.

[14] F.G. Tricomi, *Integral equations*, Courier Dover Publications, (1985).

[15] M. Masujima, *Applied mathematical methods in theoritical physics*, John Wiley and Sons, (2006).

[16] C.D. LaBudde, *The reduction of an arbitrary real square matrix to tridiagonal form using similarity transformations*, Mathematics of Computation (American Mathematical Society), **17** (84), (1963), 433–437, doi:10.2307/2004005. JSTOR 2004005. MR 0156455.

[17] S. Tuna, and M. Demİralp, *Zero interval limit perturbation expansion for the spectral entities of Hilbert-Schmidt operators combined with most dominant spectral component extraction: convergence and confirmative implementations*, Journal of Mathematical Chemistry, (2017), 1–23., doi:10.1007/s10910-017-0740-1.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

A Gaussian biparametric model for over- and underdispersed count data

María José Olmo-Jiménez[1] **and José Rodríguez-Avi**[1]

[1] *Department of Statistics and Operations Research, University of Jaén, Spain*

emails: mjolmo@ujaen.es, jravi@ujaen.es

Abstract

Although count data often exhibit overdispersion, there are also count datasets with underdispersed behaviour. However, there are not many distributions that can account for underdispersion. The most well-known distributions that handle both over- and underdispersion are the Conway-Maxwell Poisson or CMP and the hyper Poisson or HP, among others. These distributions are flexible but have serious computational problems, since there are not explicit expressions for the normalizing constant nor the moments, which have to be computed numerically or by solving link equations. In this work we present a biparametric distribution for count data that can also cope with over- and underdispersion, but does not have the computational limitations of the aforementioned models. In fact, it is a particular case of the complex triparametric Pearson (CTP) distribution. Firstly, we show the main probabilistic properties of this new distribution; secondly, we compare it with the CMP and the HP through their probability mass function and finally, we include an application example that shows the versatility of this new model.

Key words: count data, CTP distribution, overdispersion, underdispersion
MSC 2000: 60E05

1 Introduction

It is often found that data exhibit overdispersion. In this case the negative binomial (NB) distribution is the preferred alternative model, although there are many other distributions for this kind of data, such as the generalized Poisson (GP), the univariate generalized Waring (UGW), the extended generalized Waring (EGW), the zero-inflated (ZI) models and so on (see, e.g. [3]). Underdispersion is also common, for instance, in counts obtained for rare

 ISBN: 978-84-617-8694-7

events but there are not so many distributions that handle with it. The most well-known are the weighted Poisson or WP [2], the CMP [8] and the HP [1], among others. These distributions are flexible but have the disadvantage of the absence of explicit expressions for the normalizing constant and the moments, so it is necessary to use approximations or link equations for computing them, which may be a serious problem.

The CTP distribution developed by [5] is also a count data model that can cope with over- and underdispersion and does have explicit expressions for its moments in terms of the parameters. Nevertheless it has three parameters: a, b and γ. In order to compare it with other biparametric models, in [4] the authors considered the particular case $a = 0$, called the complex biparametric Pearson (CBP) distribution. However, this model is always overdispersed and so has been compared with the NB, ZIP and GP (see [6]). Now, we develop another biparametric model that is also a particular case of the CTP but when $b = 0$. It keeps the property that it can account for both over- and underdispersion and so it competes with the CMP and HP.

Then the work is structured as follows. In Section 2 we introduce the new model and study its main properties. In Section 3 a comparison among this model, the CMP and the HP is carried out through the graphical representation of the pmf in function of the mean and the variance. Finally, Section 4 includes an example that illustrates the versatility and utility of this distribution.

2 The new distribution

The CTP distribution was developed by [5]. It is a triparametric discrete distribution of infinite range generated by the Gaussian hypergeometric function with complex parameters, so it belongs to the GHD family [3]. Specifically, its pmf has the expression:

$$f(x) = f_0 \frac{(a+ib)_x (a-ib)_x}{(\gamma)_x} \frac{1}{x!} = f_0 \prod_{j=1}^{x} \frac{(a+j-1)^2 + b^2}{(\gamma+j-1)j}, \quad x = 0, 1, \ldots \tag{1}$$

where i is the imaginary unit, $a \in \mathbb{R}$, $b \geq 0$ and $\gamma > 0$. $(\alpha)_r = \Gamma(\alpha + r)/\Gamma(\alpha)$ is the Pochhammer symbol and f_0 the normalizing constant given by

$$\frac{\Gamma(\gamma - a - ib)\Gamma(\gamma - a + ib)}{\Gamma(\gamma)\Gamma(\gamma - 2a)}.$$

This distribution arises as a solution of the next difference equation:

$$G(x) f_{x+1} - L(x) f_x = 0$$

where G and L are quadratic polynomials, $G(x) = (\gamma + x)(x + 1)$ and $L(x) = x^2 + \theta_1 x + \theta_2, \theta_1, \theta_2 \in \mathbb{R}$, the latter with complex roots $a \pm ib$.

 ISBN: 978-84-617-8694-7

One of the main properties of this distribution is that it can cope with over- and underdispersion. In fact, it is underdispersed if $a < -(\mu+1)/2$, equidispersed if $a = -(\mu+1)/2$ or overdispersed if $a > -(\mu+1)/2$. In particular, if $a \geq 0$ the CTP is always overdispersed.

In [4] the authors study the particular case $a = 0$ and they call it the complex biparametric Pearson (CBP) distribution. This model is always overdispersed and has been compared with other biparametric models (see [6]).

On this occasion, we consider the case $b = 0$ in such a way that we obtain another biparametric distribution which is a particular case of the CTP. We call this new model CBP type II or CBP_{II} with parameters a and γ. Actually, this model is generated by the Gaussian hypergeometric function ${}_2F_1(a, a; \gamma; 1)$, so it has no complex parameters. Regarding the polynomial $L(r)$, we are considering the case of a double root a.

Taking (1) into account, its pmf has the expression

$$f(x) = f_0 \frac{(a)_x^2}{(\gamma)_x} \frac{1}{x!} = f_0 \prod_{j=1}^{x} \frac{(a+j-1)^2}{(\gamma+j-1)j}, \quad x = 0, 1, \ldots \tag{2}$$

where $a \in \mathbb{R}, \gamma > 0$ and $f_0 = \frac{\Gamma(\gamma-a)^2}{\Gamma(\gamma)\Gamma(\gamma-2a)}$. So (2) can be rewritten as

$$f(x) = \frac{\Gamma(\gamma-a)^2}{\Gamma(\gamma-2a)\Gamma(a)^2} \frac{\Gamma(a+x)^2}{\Gamma(\gamma+x)\Gamma(x+1)}, \quad x = 0, 1, \ldots$$

Let us observe that the CBP_{II} has finite range when $a \in \mathbb{Z}$ since $f(x) = 0, x > -a$.

2.1 Properties

As the CBP_{II} distribution is a particular case of the CTP, the former inherits most of its properties:

1. There are explicit expressions for the mean and the variance in terms of the parameters of the model, that is,

$$\mu = \frac{a^2}{\gamma - 2a - 1}, \quad \sigma^2 = \frac{a^2(\gamma - a - 1)^2}{(\gamma - 2a - 1)^2(\gamma - 2a - 2)} = \mu \frac{\mu + \gamma - 1}{\gamma - 2a - 2}.$$

   To guarantee the existence of the mean and the variance it is clear that $\gamma > 2a + 1$ and $\gamma > 2a + 2$, respectively.

2. If $\frac{(a-1)^2}{\gamma-2a+1} \in \mathbb{Z}$, the distribution has two consecutive modes in this value and the previous one. Otherwise, the distribution is unimodal with mode in the integer part of that value. As a consequence, the pmf is $J-$shaped or bell-shaped. Moreover, it is a right skewed distribution.

 ISBN: 978-84-617-8694-7

3. It is underdispersed if $a < -(\mu+1)/2$, equidispersed if $a = -(\mu+1)/2$ or overdispersed if $a > -(\mu+1)/2$. Specifically, it can be shown that

   (a) a necessary condition to be underdispersed is $a < -0.5$. If $a \leq -1$, it is always underdispersed, but if $-1 < a < -0.5$ it is underdispersed when $\gamma > \frac{3a^2+4a-1}{2a+1}$.

   (b) a necessary condition to be equidispersed is $a < -0.5$. If $-1 < a < -0.5$ it is equidispersed when $\gamma = \frac{3a^2+4a+1}{2a+1}$.

   (c) a necessary condition to be overdispersed is $a > -1$. If $a \geq -1$, it is always overdispersed, but if $-1 < a < -0.5$ it is overdispersed when $\gamma > \frac{3a^2+4a-1}{2a+1}$.

   This property makes the CBP_{II} distribution more versatile to model a dataset.

4. A sufficient condition to be infinitely divisible (i.d.) is that $a > -0.5$ and $\gamma > a^2/(1+2a)$. So, if $a < -0.5$ the CBP_{II} distribution is not i.d. As a consequence, if the distribution is underdispersed it cannot be i.d.

5. It converges to the Poisson distribution when γ and $a^2 \to \infty$ with the same order of convergence and to the normal distribution when γ and $|a|$ have the same order of convergence.

Let us observe that when $a > 0$ the $CBP_{II}(a, \gamma)$ is a $UGW(a, a, \gamma - 2a)$, so the proposed model could be considered an underdispersed extension of a biparametric UGW.

3 Comparison with other distributions for count data

In order to find differences between the CBP_{II} and the CMP and HP distributions, we have compared their shapes, i.e., we have plotted their pmf for several values of the mean and the variance. So, we have obtained the expressions of the parameters of each model in terms of their mean and variance:

- for the $CBP_{II}(a, \gamma)$:

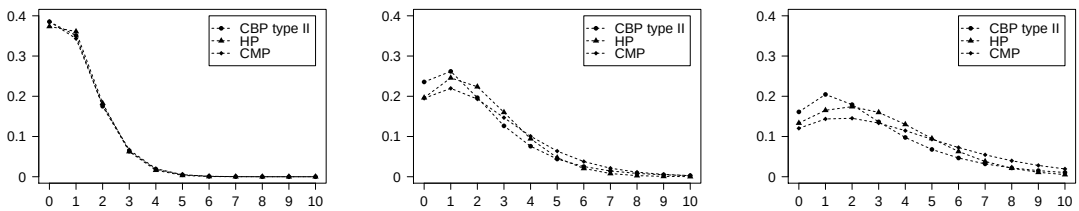
$$a = \frac{\mu^2 \pm \sqrt{\sigma^2 \mu \left[\sigma^2 + \mu\left(\mu - 1\right)\right]}}{\sigma^2 - \mu}, \quad \gamma = \frac{a^2}{\mu} + 2a + 1$$

  so, there are two solutions for each parameter. However, it can be shown that only one is possible for $\mu \geq 1$, which is the most usual case in count data.

- for the $HP(\gamma, \lambda)$ with $\gamma, \lambda > 0$ and pmf given by

$$P(X = x) = \frac{\Gamma(\gamma)}{{}_1F_1(1; \gamma; \lambda)} \frac{\lambda^x}{\Gamma(\gamma + x)}, \quad x = 0, 1, \ldots,$$

 ISBN: 978-84-617-8694-7



(a) $\mu = 1, \sigma^2 = 1.1$ (b) $\mu = 2, \sigma^2 = 4$ (c) $\mu = 3, \sigma^2 = 9$

Figure 1: Pmf of a CBP_{II}, HP and CMP for several values of the mean and the variance in an overdispersed scenario.

it is not possible to express the parameters in terms of μ and σ^2. [7] use the following link equations

$$\begin{aligned} \mu &= \lambda - (\gamma - 1)\left[1 - \frac{1}{{}_1F_1(1;\gamma;\lambda)}\right] \\ \sigma^2 &= \lambda + [\lambda - (\gamma - 1)]\,\mu - \mu^2. \end{aligned}$$

- for the $CMP(\lambda, v)$ with $\lambda > 0$, $v \geq 0$ and pmf

$$P(X = x) = \frac{\lambda^x}{(x!)^v}\frac{1}{Z(\lambda, v)}, \quad x = 0, 1, \ldots,$$

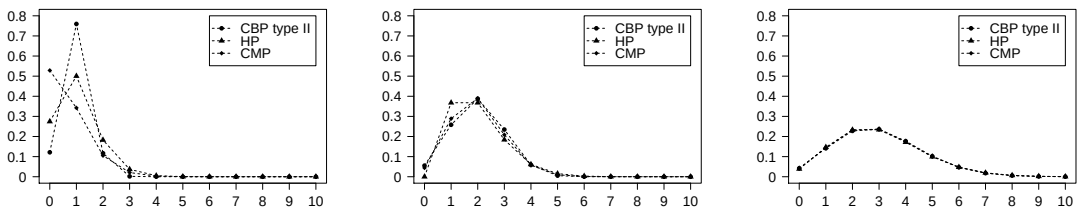
where $Z(\lambda, v) = \sum_{j=0}^{\infty} \frac{\lambda^j}{(j!)^v}$, we use the following approximate formulas (more details in [8])

$$\mu \approx \lambda^{1/v} - \frac{v-1}{2v}, \quad \sigma^2 \approx \frac{1}{v}\lambda^{\frac{1}{v}}.$$

These aproximations are specially good for $v \leq 1$ or $\lambda > 10^v$.

The comparison between the pmfs of the CBP_{II} and the CMP and HP is made in Figures 1 and 2 for overdispersed and underdispersed scenarios, respectively. Firstly, it is clear that the pmfs become $J-$shaped as the mean increases. It can be seen that the differences are more evident as the overdispersion or the underdispersion is more severe. Moreover, the CBP_{II} has higher probability in the modal value of the variable than the CMP and HP.

 ISBN: 978-84-617-8694-7



(a) $\mu = 1, \sigma^2 = 0.25$ (b) $\mu = 2, \sigma^2 = 1$ (c) $\mu = 3, \sigma^2 = 2.7$

Figure 2: Pmf of a CBP_{II}, HP and CMP for several values of the mean and the variance in an underdispersed scenario.

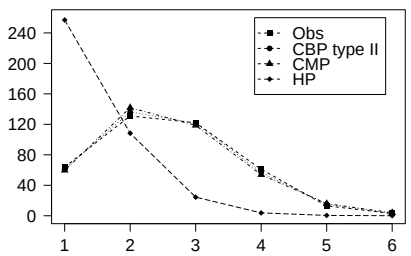


Figure 3: Observed and expected frequencies for data about the number of syllables of a Turkish poem.

4 Example

We consider data about the word length (in terms of number of syllables) in the turkish poem *Gidisat* by Ercüment Behzat Lâv available in [9]. Following these authors, the count for 1 is treated as a count for 0, and in general the count for y is treated as $y-1$, as though the data are generated by adding 1 to the distribution. These data exhibit underdispersion with a variance-mean ratio of 0.7372. Table 1 contains the parameter estimates, their standard errors (in parenthesis), observed and expected frequencies and the corresponding Pearson χ^2 test for the CMP, HP and CBP_{II} models. Observed and expected frequencies for each fit are represented in Figure 3. Although the CBP_{II} and CMP fits are very similar and really good, the former is slightly better and more accurate than the latter taking into account the AIC, the expected frequencies and the $p-$value.

 ISBN: 978-84-617-8694-7

		Expected		
Y	*Observed*	CBP_{II}	CMP	HP
1	64	61.2382	59.6867	257.0274
2	131	136.2322	141.8662	108.3037
3	122	121.6777	118.6984	24.4242
4	61	56.9254	53.9223	3.7603
5	13	15.2708	15.8818	0.4395
≥ 6	3	2.6557	3.9446	0.0449
		$\widehat{a} = -10.5302$ (2.1537)	$\widehat{\lambda} = 2.3768$ (0.2755)	$\widehat{\lambda} = 1.1514$ (0.1040)
		$\widehat{\gamma} = 49.8442$ (24.3691)	$\widehat{v} = 1.5062$ (0.1369)	$\widehat{\gamma} = 0.4852$ (0.0987)
	AIC	**1158.309**	1160.692	1164.391
	$\chi^2 - distance$	**1.0003**	2.9140	1964.384
	$p - value$	**0.8012**	0.4051	0

Table 1: Parameter estimates, standard errors (in parenthesis), observed and expected frequencies, AIC and χ^2 test for fits to data about the word length of a Turkish poem.

References

[1] G. E. Bardwell and E. L. Crow, *A two-parameter family of hyper-Poisson distributions*, Journal of the American Statistical Association **9**(305) (1964) 133–141.

[2] J. Del Castillo and M. Pérez-Casany, *Overdispersed and Underdispersed Poisson generalizations*, Journal of Statistical Planning and Inference **134** (2005) 486–500.

[3] N. L. Johnson, A. W. Kemp and S. Kotz, *Univariate discrete distributions*, Wiley: New York, 2005.

[4] J. Rodríguez-Avi, A. Conde-Sánchez, A. and A. J. Sáez-Castillo, *A new class of discrete distributions with complex parameters*, Statistical Papers **44** (2003) 67–88.

[5] J. Rodríguez-Avi, A. Conde-Sánchez, A., A. J. Sáez-Castillo and M. J. Olmo-Jiménez, *A triparametric discrete distribution with complex parameters*, Statistical Papers **45**(1) (2004) 81–95.

[6] J. Rodríguez-Avi and M. J. Olmo-Jiménez, *A regression model for overdispersed data without too many zeros*, Statistical Papers doi:10.1007/s00362-015-0724-9 (2015).

[7] A. J. Sáez-Castillo and A. Conde-Sánchez, *A hyper-Poisson regression model for overdispersed and underdispersed count data*, Computational Statistics and Data Analysis **61** (2013) 148–157.

 ISBN: 978-84-617-8694-7

[8] K. F. Sellers, S. Borle and G. Shmueli, *The COM-Poisson model for count data: a survey of methods and applications*, Applied Stochastic Models in Business and Industry **28** (2012) 104–116.

[9] G. Wimmer, R. Köhler, R. Grotjahn and G. Altmann, *Towards a Theory of Word Length Distribution*, Journal of Quantitative Linguistics **1** (1994) 98–106.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

Curvature study for PPH reconstruction operator and applications to smoothing splines

P. Ortiz[1] and J.C. Trillo[1]

[1] *Departamento de Matemática Aplicada y Estadística, Universidad Politcnica de Cartagena*

emails: portiz@navantia.es, jc.trillo@upct.es

Abstract

This paper is devoted to study the curvature of the Lagrange and PPH [1] reconstruction operators. We also compare the curvatures obtained in order to know which operator could be more suitable for some uses. Particularly we have considered the similarity with smoothing splines definition in order to develop an algorithm that propose nearly optimal weights in smoothing splines.
Key words: interpolation, curvature minimization, nonlinear, splines

MSC 2000: 41A05, 41A10, 65D17.

1 Introduction

Reconstruction operators are widely used in computer aided geometric design with interesting applications. They can be connected with subdivision schemes and in turn with fast algorithms for the generation of curves and surfaces.

Starting from a discrete set of data, the reconstruction operators aim at obtaining a piecewise function $p(x)$ which interpolates or approximates the data maintaining in many cases certain desirable properties. In particular, smoothing splines (see [6], [14]) in a given interval $[a, b]$ are based on polynomial reconstruction pieces which are connected with smooth joints at control knots and they minimize a functional of the type

$$J(p) := \int_a^b p''(x)^2 \, dx + \sum_j \mu_j (p(x_j) - f_j)^2, \tag{1}$$

 ISBN: 978-84-617-8694-7

which suppose a certain compromise, depending on the weights μ_j, between reducing the term approximating the curvature and keeping close to the initial set of data (x_j, f_j).

PPH reconstruction (see [1], [10]) is by construction a nonlinear interpolatory technique with several desirable properties. Among them, we would like to mention the following: each polynomial piece is builded with a fixed centered stencil, it is fourth order accurate on smooth convex regions, in the presence of singularities it reduces to second order but it does not lose all accuracy as it occurs in the linear case, moreover the reconstruction is free of Gibb's effects.

In this work we study the term of curvature of the above functional for Lagrange reconstruction and for PPH reconstruction given in [1] for the uniform case and in [13] for the non uniform case. We compare this term for both reconstruction operators and we deduce that PPH is an interesting candidate for minimization. On the other hand we calculate the weights that locally minimize the functional for the PPH reconstruction, and we use this result to relate this nonlinear reconstruction with splines theory, opening up doors to research about new nearly optimal weights in smoothing splines.

2 Curvature study on uniform grids

Let us consider the set of values $f_{j-1}, f_j, f_{j+1}, f_{j+2}$ corresponding to subsequent abcisas $x_{j-1}, x_j, x_{j+1}, x_{j+2}$ of a regular grid X of step h. The set of polynomials $p(x)$ which pass through the central points (x_j, f_j) and (x_{j+1}, f_{j+1}) can be written in terms of two free variables A and B as follows,

$$\begin{aligned} p(x) &:= \frac{x_{j+1}-x}{h} f_j + \frac{x-x_j}{h} f_{j+1} \\ &- \frac{1}{6}(x-x_j)(x_{j+1}-x)\left[A(1+\frac{x_{j+1}-x}{h}) + B(1+\frac{x-x_j}{h})\right]. \end{aligned} \tag{2}$$

At the boundary points x_{j-1}, x_{j+2} of the interval, the distance of each polynomial of the set to the initial data is given by

$$\begin{aligned} p(x_{j-1}) - f_{j-1} &= h^2(A - 2D_j), \\ p(x_{j+2}) - f_{j+2} &= h^2(B - 2D_{j+1}). \end{aligned} \tag{3}$$

where D_j and D_{j+1} are the following divided differences

$$\begin{aligned} D_j &= f[x_{j-1}, x_j, x_{j+1}] = \frac{f_{j-1} - 2f_j + f_{j+1}}{2h^2}, \\ D_{j+1} &= f[x_j, x_{j+1}, x_{j+2}] = \frac{f_j - 2f_{j+1} + f_{j+2}}{2h^2}, \end{aligned} \tag{4}$$

Introducing the second derivative of (2) in the curvature term in (1) we get

 ISBN: 978-84-617-8694-7

$$C(p) = \int_{x_j}^{x_{j+1}} p''(x)^2 \, dx = \frac{h}{3}(A^2 + AB + B^2). \tag{5}$$

For the sake of simplicity, from here in advance we particularize with the subscript L or P the parameters A, B or the curvature C associated to Lagrange or PPH polynomial.

Let be $PL(x)$ the Lagrange polynomial. It also interpolates the boundary points, i.e. it is the polynomial of set (2) that verifies

$$\begin{array}{l} p(x_{j-1}) = f_{j-1}, \\ p(x_{j+2}) = f_{j+2}. \end{array} \tag{6}$$

These conditions and equations (3) give us the parameters A and B associated to Lagrange polynomial,

$$\begin{array}{l} A_L = 2D_j, \\ B_L = 2D_{j+1}. \end{array} \tag{7}$$

The curvature term (5) in this case will be

$$C_L = \frac{h}{3}(D_j^2 + D_j D_{j+1} + D_{j+1}^2). \tag{8}$$

When $p(x)$ is the PPH polynomial (see [1] for more details) there are two possible cases depending on the absolute values of the divided differences D_j, D_{j+1}.

Case 1. $|D_j| \leq |D_{j+1}|$, i.e, the possible singularity is at $[x_{j+1}, x_{j+2}]$,

$$\begin{array}{l} p(x_{j-1}) = f_{j-1}, \\ p(x_{j+2}) = \widetilde{f}_{j+2}. \end{array} \tag{9}$$

where $\widetilde{f}_{j+2}$ and $\widetilde{D}$ are given by

$$\widetilde{f}_{j+2} = f_{j+2} - 4h^2 \left(\frac{D_j + D_{j+1}}{2} - \widetilde{D} \right). \tag{10}$$

$$\widetilde{D} = \begin{cases} \frac{2D_j D_{j+1}}{D_j + D_{j+1}} & \text{if } \; D_j D_{j+1} > 0, \\ 0 & otherwise. \end{cases} \tag{11}$$

From conditions (9) and equations (9) result

 ISBN: 978-84-617-8694-7

$$\begin{aligned} A_P &= 2D_j, \\ B_P &= 4\widetilde{D} - 2D_j. \end{aligned} \tag{12}$$

Therefore the coefficient B_P and the curvature term C_P will depend on the sign of the product D_jD_{j+1}.

Case 1.1. $D_jD_{j+1} > 0$.

$$B_P = \frac{2D_j(3D_{j+1} - D_j)}{D_j + D_{j+1}} \tag{13}$$

$$C_P = \frac{4hD_j^2(D_j^2 - 2D_jD_{j+1} + 13D_{j+1}^2)}{3(D_j + D_{j+1})^2}. \tag{14}$$

Comparing curvature terms associated to Lagrange interpolation (8) and to PPH reconstruction (14) we can see that

$$C_L - C_P = \frac{4hD_{j+1}(D_{j+1} - D_j)^2(5D_j + D_{j+1})}{3(D_j + D_{j+1})^2} \geq 0. \tag{15}$$

Case 1.2. $D_jD_{j+1} \leq 0$.

$$B_P = -2D_j, \tag{16}$$

$$C_P = \frac{4hD_j^2}{3}. \tag{17}$$

And the difference $C_L - C_P$ becomes

$$C_L - C_P = \frac{4hD_{j+1}(D_j + D_{j+1})}{3} \geq 0. \tag{18}$$

Case 2. $|D_j| > |D_{j+1}|$, i.e, the possible singularity lies at $[x_{j-1}, x_j]$

$$\begin{aligned} p(x_{j-1}) &= \widetilde{f}_{j-1}, \\ p(x_{j+2}) &= f_{j+2}. \end{aligned} \tag{19}$$

where [1]

$$\widetilde{f}_{j-1} = f_{j-1} - 4h^2\left(\frac{D_j + D_{j+1}}{2} - \widetilde{D}\right). \tag{20}$$

Working in a similar way as in case 1 we obtain

 ISBN: 978-84-617-8694-7

$$\begin{aligned} A_P &= 4\widetilde{D} - 2D_{j+1}, \\ B_P &= 2D_{j+1}, \end{aligned} \tag{21}$$

and depending on the sign of the product D_jD_{j+1} results

Case 2.1. $D_jD_{j+1} > 0$.

$$A_P = \tfrac{2D_{j+1}(3D_j - D_{j+1})}{D_j + D_{j+1}}, \tag{22}$$

$$C_P = \frac{4hD_{j+1}^2(13D_j^2 - 2D_jD_{j+1} + D_{j+1}^2)}{3(D_j + D_{j+1})^2}. \tag{23}$$

$$C_L - C_P = \frac{4hD_j(D_j - D_{j+1})^2(D_j + 5D_{j+1})}{3(D_j + D_{j+1})^2} \geq 0. \tag{24}$$

Case 2.2. $D_jD_{j+1} \leq 0$.

$$A_P = -2D_{j+1}, \tag{25}$$

$$C_P = \frac{4hD_{j+1}^2}{3}. \tag{26}$$

$$C_L - C_P = \frac{4hD_j(D_j + D_{j+1})}{3} \geq 0. \tag{27}$$

We have just seen that when data are in a regular grid the curvature term in equation (1) associated to PPH reconstruction operator is smaller than the one associated to the Lagrange operator. Let us remark that a similar study can be carried out with non uniform grids. This result is the motivation and the central idea to carry out a nonlinear definition of the weights in (1) which aims at improving the performance of smoothing splines according to the initial data.

 ISBN: 978-84-617-8694-7

Acknowledgements

This research was partially supported by the project: Métodos numéricos para algunos problemas no lineales, Ref: 19374/PI/14 under the program Ayudas para la realización de proyectos de investigación destinados a proyectos competitivos (Programa Seneca 2014), also by the national research project MTM2015-64382.

References

[1] S. Amat, R. Donat, J. Liandrat and J.C. Trillo, *Analysis of a new nonlinear subdivision scheme. Applications in image processing*, Found. Comput. Math. **6(2)** (2006) 193–225.

[2] S. Amat, K. Dadourian, J. Liandrat, J. C. Trillo, *High order nonlinear interpolatory reconstruction operators and associated multiresolution schemes*, J. Comput. Appl. Math. **253** (2013), 163–180.

[3] S. Amat, R. Donat, J. C. Trillo, *Proving convexity preserving properties of interpolatory subdivision schemes through reconstruction operators*, Appl. Math. Comput. **219(14)** (2013) 7413–7421.

[4] S.Amat and J.Liandrat, *On the stability of PPH nonlinear multiresolution*, Appl. Comp. Harm. Anal. **18(2)** (2005) 198–206.

[5] F. Aràndiga and R. Donat, *Nonlinear Multi-scale Decomposition: The Approach of A.Harten*, Numerical Algorithms. **23** (2000) 175–216.

[6] C. de Boor, *A practical guide to splines*, New York Springer-Verlag, 1978.

[7] N. Dyn, J. A. Gregori, and D. Levin, *A 4-point interpolatory subdivision scheme for curve design*, Comput. Aided. Geom. Design. **4** (1987) 257–268.

[8] N. Dyn and D. Levin, *Stationary and non-stationary binary subdivision schemes*, Math. Meth. Comput. Aided. Geom. Design II. (1991) 209–216.

[9] N. Dyn, F. Kuijt, D. Levin, and R. van Damme, *Convexity preservation of the four-point interpolatory subdivision scheme*, Comput. Aided. Geom. Design. **16(8)** (1999) 789–792.

[10] M.S. Floater and C.A. Michelli, *Nonlinear stationary subdivision*, Approximation theory: in memory of A.K. Varna, edt: Govil N.K, Mohapatra N., Nashed Z., Sharma A., Szabados J. (1998) 209–224.

 ISBN: 978-84-617-8694-7

[11] A. Harten, *Multiresolution representation of data II*, SIAM J. Numer. Anal. **33(3)** (1996) 1205–1256.

[12] F. Kuijt and R. van Damme, *Convexity preserving interpolatory subdivision schemes*, Const. Approx. **14** (1998) 609–630.

[13] P. Ortiz, and J.C. Trillo, *A Nonlinear interpolatory reconstruction operator on non uniform grids*, Submitted (2016).

[14] C.H. Reinsch, *Smoothing by spline functions*, Numerische Mathematik. **10** (1967) 177–183.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

HPC Tool for Multidimensional Scaling

Francisco Orts[1], Ernestas Filatovas[2], Gloria Ortega[1], Olga Kurasova[2] and Ester M. Garzón[1]

[1] *Group of Supercomputation-Algorithms, Dpt. of Informatics, ceiA3, Univ. of Almería, 04120, Almería, Spain*

[2] *Inst. of Mathematics and Informatics, Vilnius Univ., Akademijos str. 4, LT-08663, Vilnius, Lithuania*

emails: francisco.orts@ual.es, ernest.filatov@gmail.com, gloriaortega@ual.es, olga.kurasova@mii.vu.lt, gmartin@ual.es

Abstract

The reduction of the dimensionality is of great interest in the context of big data processing. Multidimensional scaling methods, or MDS, are techniques for dimensionality reduction, where data from a high-dimensional space are mapped into a lower-dimensional space. They consume relevant computational resources and an intensive research has been developed to accelerate them. In this work, several MDS methods are revised and accelerated on modern MultiGPU clusters.

Key words: Dimensionality Reduction; Multidimensional Scaling; High Performance Computing; GPU computing

1 Background

Real-world data, such as speech signals, images, biomedical, financial, telecommunication and other data usually have a high dimensionality. Each data instance (point) is characterized by some features. The dimensionality of such data, as well as the amount of processed data, is constantly increasing but the requirement of processing these data within a reasonable time frame still remains an open problem. Recent development in graphics hardware allows performing generic parallel computations on powerful hardware and provides an opportunity to solve many time-consuming problems.

 ISBN: 978-84-617-8694-7

Multidimensional Scaling (MDS) is one of the most popular dimensionality reduction method [1, 3]. MDS aims at finding points $Y_1, Y_2, \ldots, Y_m$ in the low-dimensional space $\mathbb{R}^s$, $s < n$, such that the distances between them are as close as possible to the distances between the original points $X_1, X_2, \ldots, X_m$ in the multidimensional space $\mathbb{R}^n$. This is achieved by minimizing the stress function

$$E_{MDS} = \sum_{i<j} \Big(d(X_i, X_j) - d(Y_i, Y_j) \Big)^2. \quad (1)$$

Here $d(\cdot, \cdot)$ is the distance between two points in the corresponding space. MDS is widely-used in different fields [6, 8, 9, 11].

Recently, power of GPU has been employed to speedup MDS algorithms. Fester et al. [5] proposed a CUDA implementation of MDS algorithm based on the high throughput multidimensional scaling (HiT-MDS). In [13], authors suggested a new efficient parallel algorithm for MDS based on virtual particle dynamics (VPD-MDS) [4] and explored this algorithm on GPU. In [7], the multi-level MDS Glimmer algorithm was developed for GPU, by dividing the input data into hierarchical levels and executing the algorithm recursively. The method is based on a stochastic approach [2]. Another CUDA-based technique to get MDS approximation is CFMDS [11]. It implements both single-level and multi-level approaches: if the data can fit the memory of the GPU entirely, a classical algorithm is executed; otherwise, the input data are divided into smaller portions that fit the global memory. In [10], authors proposed a correlation clustering framework which uses MDS for layout and GPU-acceleration to speedup Visual feedback. In [12], the fast sampling-based multidimensional scaling (SBMDS) on a multi-core GPU architecture was proposed to improve content-based image retrieval (CBIR) systems. Although the research on this field is being carried out actively, it remains relevant as the new GPU architecture and heterogeneous platforms constantly appear that should be effectively exploited for solving dimensionality reduction problems of different complexities.

2 MultiGPU implementation of MDS methods

In this work the main focus is the acceleration of MDS methods on modern GPU architectures. According to this idea our contributions are:

1. Update the CUDA MDS versions on the modern GPUs. For example, MDS is dominated by the reduction operations to compute the distances among data. On modern GPU the reduction among the local data of different threads can be optimized by the use of shuffled operations avoiding the use of the shared memory. These operations are only available from the Kepler GPU architecture.

2. Explore what is the best GPU-threads organization according to the particular number of data dimensions and items of every instance of the problem.

 ISBN: 978-84-617-8694-7

3. Develop a parallel MDS version on MultiGPU clusters.

4. Provide multicore and GPU versions and define an approach to identify the best option between both, according to the particular characteristics of every application example.

Our proposal is evaluated on a set of test applications which are well-known as examples where the MDS approaches are used.

Acknowledgements

This work has been partially supported by the Spanish Ministry of Science throughout projects TIN15-66680 and CAPAP-H6 network TIN2016-81840-REDT, by J. Andalucía through projects P12-TIC-301 and P11-TIC7176, and by the European Regional Development Fund (ERDF).

References

[1] Ingwer Borg and Patrick JF Groenen. *Modern multidimensional scaling: Theory and applications*. Springer Science & Business Media, 2005.

[2] Matthew Chalmers. A linear iteration time layout algorithm for visualising high-dimensional data. In Roni Yagel and Gregory M. Nielson, editors, *IEEE Visualization*, pages 127–132. IEEE Computer Society and ACM, 1996.

[3] Gintautas Dzemyda, Olga Kurasova, and Julius Žilinskas. *Multidimensional Data Visualization: Methods and Applications*, volume 75. Springer Science & Business Media, 2013.

[4] Witold Dzwinel and Jan Blasiak. Method of particles in visual clustering of multi-dimensional and large data sets. *Future Generation Computer Systems*, 15(3):365–379, 1999.

[5] Thilo Fester, Falk Schreiber, and Marc Strickert. CUDA-based Multi-core Implementation of MDS-based Bioinformatics Algorithms". In Ivo Grosse, Steffen Neumann, Stefan Posch, Falk Schreiber, and Peter F. Stadler, editors, *GCB*, volume 157 of *LNI*, pages 67–79. GI, 2009.

[6] Ernestas Filatovas, Dmitry Podkopaev, and Olga Kurasova. A visualization technique for accessing solution pool in interactive methods of multiobjective optimization. *International Journal of Computers Communications and Control*, 10:508–519, 2015.

 ISBN: 978-84-617-8694-7

[7] Stephen Ingram, Tamara Munzner, and Marc Olano. Glimmer: Multilevel MDS on the GPU. *IEEE Trans. Vis. Comput. Graph.*, 15(2):249–261, 2009.

[8] Olga Kurasova, Tomas Petkus, and Ernestas Filatovas. Visualization of pareto front points when solving multi-objective optimization problems. *Information Technology And Control*, 42(4):353–361, 2013.

[9] Viktor Medvedev, Olga Kurasova, Jolita Bernatavičienė, Povilas Treigys, Virginijus Marcinkevičius, and Gintautas Dzemyda. A new web-based solution for modelling data mining processes. *Simulation Modelling Practice and Theory*, 2017.

[10] Eric Papenhausen, Bing Wang, Sungsoo Ha, Alla Zelenyuk, Dan Imre, and Klaus Mueller. GPU-accelerated incremental correlation clustering of large data with visual feedback. In *Proceedings of the 2013 IEEE International Conference on Big Data, 6-9 October 2013, Santa Clara, CA, USA*, pages 63–70, 2013.

[11] Sungin Park, Soo-Yong Shin, and Kyu-Baek Hwang. CFMDS: CUDA-based fast multidimensional scaling for genome-scale data. *BMC Bioinformatics*, 13(17):S23, 2012.

[12] Piotr Pawliczek and Witold Dzwinel. Interactive data mining by using multidimensional scaling. *Procedia Computer Science*, 18:40 – 49, 2013.

[13] Piotr Pawliczek, Witold Dzwinel, and David A. Yuen. Visual exploration of data by using multidimensional scaling on multicore CPU, GPU, and MPI cluster. *Concurr Comput*, 26(3):662–682, 2014.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

Generation of Test Matrices with Exact Singular Values for Numerical Computations

Katsuhisa Ozaki[1] **and Takeshi Ogita**[2]

[1] *Department of Mathematical Sciences, Shibaura Institute of Technology*

[2] *Division of Mathematical Sciences, Tokyo Woman's Christian University*

emails: ozaki@sic.shibaura-it.ac.jp, ogita@lab.twcu.ac.jp

Abstract

This paper considers test matrices for numerical linear algebra. The proposed method generates real symmetric or unsymmetric matrices whose singular values are known exactly. This is useful for checking the accuracy of numerically computed results. The computational cost of the proposed method is significantly less than that of matrix multiplication, and the strategy can be straightforwardly extended to a real symmetric matrix with exact eigenvalues.

Key words: test matrix, floating-point arithmetic, numerical linear algebra,
MSC 2000: 15A18, 65F15,

1 Introduction

This paper considers test matrices for numerical linear algebra. Test matrices have been well summarised in the literature [1, Section 28]. Here we denote two orthogonal matrices as $U \in \mathbb{R}^{m\times m}$ and $V \in \mathbb{R}^{n\times n}$. A matrix $\Sigma \in \mathbb{R}^{m\times n}$ is a diagonal matrix of singular values. Then, test matrices are sometimes generated by the following singular value decomposition form:

$$A := U\Sigma V^T. \tag{1}$$

This can be used by 'gallery' functions with the 'randsvd' option in MATLAB [2]. Let $\mathbb{F}$ be a set of floating-point numbers as defined by IEEE 754 [3]. If we use floating-point numbers and floating-point arithmetic, there are problems when obtaining a matrix with exact singular values based on (1).

 ISBN: 978-84-617-8694-7

Problem 1 Although both U and V are orthogonal matrices, their elements may not be representable by floating-point numbers ($\mathbb{F} \not\ni u_{ij}, v_{ij} \in \mathbb{R}$). One could obtain U and V by QR decomposition from a non-singular matrix. However, the exact orthogonal matrix is difficult to obtain due to accumulation of rounding errors.

Problem 2 Even if U, Σ and V are represented by floating-point numbers ($U \in \mathbb{F}^{m\times m}, \Sigma \in \mathbb{F}^{m\times n}$ and $V \in \mathbb{F}^{n\times n}$), rounding errors may occur in the evaluation of matrix multiplications $U\Sigma V^T$. Then, the singular values of a computed result are not σ_{ii}.

We propose a method that produces a matrix with exact singular values using a Hadamard matrix. With the proposed method, the user inputs Σ to obtain diagonal matrix Σ' ($\approx \Sigma$), where σ'_{ii} is the exact singular value of $U\Sigma' V^T \in \mathbb{F}^{m\times n}$.

2 Floating-point Arithmetic and Hadamard Matrix

The function $\texttt{diag}(x, m, n)$, $x \in \mathbb{F}^{\min(m,n)}$ generates an m-by-n diagonal matrix D with $d_{ii} = x_i$. $\texttt{fl}(\cdot)$ and $\texttt{fl}_\triangle(\cdot)$ indicate computed results obtained by floating-point arithmetic with rounding to the nearest mode (roundTiesToEven) and a rounding upward mode (roundTowardPositive), respectively. Assume that neither overflow nor underflow occurs in $\texttt{fl}(\cdot)$. Here, let $\texttt{u}$ and $\texttt{u}_S$ be a roundoff unit and the smallest positive number in $\mathbb{F}$, e.g., $\texttt{u} = 2^{-53}$ and $\texttt{u}_S = 2^{-1074}$ for binary64 in IEEE 754, respectively. A constant $\texttt{realmax}$ is the largest floating-point number in $\mathbb{F}$. The function $\texttt{ufp}(\cdot)$, i.e. the unit in the first place, is defined as follows

$$\texttt{ufp}(a) = 2^{\lfloor \log_2 |a| \rfloor} \text{ (for } a \neq 0), \quad \texttt{ufp}(a) = 0 \text{ (for } a = 0), \quad a \in \mathbb{R}.$$

Here, we introduce several lemmas used to prove the proposed method.

Lemma 1 (well-known)**.** *For $a \in \mathbb{F}$, $a \in 2\texttt{u} \cdot \texttt{ufp}(a)\mathbb{Z}$ holds.*

Lemma 2 (well-known)**.** *For $a \in \mathbb{R}$ ($\texttt{u}_S\mathbb{Z} \ni |a| \leq \texttt{realmax}$), $k = 2^w, w \in \mathbb{Z}$, if $\texttt{u}k\mathbb{Z} \ni |a| \leq k$, then $a \in \mathbb{F}$.*

Lemma 3 ([4])**.** *For $a, b \in \mathbb{F}$, $a + b = \texttt{fl}(a + b) + \delta$, $|\delta| \leq \texttt{u} \cdot \texttt{ufp}(a + b)$ is satisfied.*

Lemma 4 ([4])**.** *For $a, b \in \mathbb{F}$ ($|a| \geq |b|$), $c = \texttt{fl}(a + b)$, $\texttt{fl}(c - a) = c - a$ is satisfied.*

A Hadamard matrix H is a matrix whose entries are either +1 or -1, such that $H^T H = HH^T = nI$. For $n = 2^k, k \in \mathbb{N}_0$, we can generate a $2n$-by-$2n$ Hadamard matrix as follows [5]:

$$\begin{pmatrix} H & H \\ H & -H \end{pmatrix}, \tag{2}$$

 ISBN: 978-84-617-8694-7

where $H = 1$ for $n = 1$. Let $H^{(n)}$ be an n-by-n Hadamard matrix generated by (2). Note that this matrix is symmetric. Assume that $\sqrt{mn} \in \mathbb{N}$. A matrix A can be obtained by $A := H^{(m)}\Sigma H^{(n)}/\sqrt{mn}$. All elements in $H^{(m)}$ and $H^{(n)}$ can be represented exactly using floating-point numbers. No rounding error occurs in the division, because $\sqrt{mn}$ is a power of two. Therefore, Problem 1 (Section 1) is solved. We explain how to avoid rounding errors in matrix multiplication in the following.

3 Proposed Method

Assume that a user gives a vector $\sigma \in \mathbb{F}^{\min(m,n)}$. We produce a vector σ' $(\approx \sigma) \in \mathbb{F}^{\min(m,n)}$ to satisfy $\mathtt{fl}(H^{(m)}\Sigma' H^{(n)}/\sqrt{mn}) = H^{(m)}\Sigma' H^{(n)}/\sqrt{mn}$, where $\Sigma' = \mathtt{diag}(\sigma', m, n)$. Here, let $t \in \mathbb{F}$ be the following:

$$t := 1.5\mathtt{ufp}\left(\mathtt{fl}_{\triangle}\left(2\sum_{i=1}^{n}\sigma_i\right)\right). \tag{3}$$

We compute

$$\sigma'_i = \mathtt{fl}((t + \sigma_i) - t) \tag{4}$$

for all i. Then, no rounding error occurs in $\mathtt{fl}(H^{(m)}\Sigma' H^{(n)})$. Therefore, the singular values of $\mathtt{fl}(H^{(m)}\Sigma' H^{(n)}/\sqrt{mn})$ are σ'_i.

Theorem 1. *$H^{(m)}$ and $H^{(n)}$ are generated by (2). Assume that $2\min(m,n)\mathtt{u} \le 1$ and vector $\sigma' \in \mathbb{F}^n$ is obtained by (4). Let $\Sigma' = \mathtt{diag}(\sigma', m, n)$. Then,*

$$B := \begin{cases} H^{(m)}\Sigma' H^{(n)} = \mathtt{fl}(H^{(m)}(\Sigma' H^{(n)})) & (m \le n) \\ H^{(m)}\Sigma' H^{(n)} = \mathtt{fl}((H^{(m)}\Sigma')H^{(n)}) & (otherwise) \end{cases}$$

holds.

Proof. Here, we provide a proof for $H^{(m)}\Sigma' H^{(n)} = \mathtt{fl}(H^{(m)}(\Sigma' H^{(n)}))$ for $m \le n$, because $H^{(m)}\Sigma' H^{(n)} = \mathtt{fl}((H^{(m)}\Sigma')H^{(n)})$ for $m > n$ can be proved similarly. First, $\Sigma' H^{(n)} = \mathtt{fl}(\Sigma' H^{(n)})$ trivially holds. Then, b_{ij} becomes the following:

$$b_{ij} = \sum_{k=1}^{m} h_{ik}^{(m)} \sigma'_k h_{kj}^{(n)}.$$

The definition of t in (3) yields $\mathtt{ufp}(t) \le \mathtt{ufp}(t + \sigma_i)$ for all i. From Lemma 1, we obtain the following:

$$\sigma'_i \in 2\mathtt{u} \cdot \mathtt{ufp}(t). \tag{5}$$

 ISBN: 978-84-617-8694-7

From Lemma 4 and 3, we obtain

$$\sigma_i' = \mathtt{fl}((t+\sigma_i)-t) = \mathtt{fl}(t+\sigma_i) - t = \sigma_i + \delta_i, \quad |\delta_i| \le \mathtt{u}\cdot\mathtt{ufp}(t+\sigma_i) \le 2\mathtt{u}\cdot\mathtt{ufp}(t),$$

which derives the following:

$$|\sigma_i'| \le \sigma_i + 2\mathtt{u}\cdot\mathtt{ufp}(t). \tag{6}$$

From the assumption for n, the definition of t in (3), (5) and (6), we obtain the following:

$$2\mathtt{u}\cdot\mathtt{ufp}(t) \ni \sum_{k=1}^{m} h_{ik}^{(m)}\sigma_k' h_{kj}^{(n)} \le \sum_{k=1}^{m}|\sigma_k'| \le \sum_{k=1}^{m}(\sigma_k + 2\mathtt{u}\cdot\mathtt{ufp}(t)) \le \mathtt{ufp}(t) + 2m\mathtt{u}\cdot\mathtt{ufp}(t) \le 2\mathtt{ufp}(t).$$

Therefore, from Lemma 2, $H^{(m)}\Sigma' H^{(n)} = \mathtt{fl}(H^{(m)}(\Sigma' H^{(n)}))$ holds. □

Note that we do not need to compute full matrix multiplication for $\mathtt{fl}(H^{(m)}(\Sigma' H^{(n)}))$ using recursive approaches. We assume that both m and n are powers of two. Otherwise, the following matrix

$$\begin{pmatrix} H & O \\ O & P \end{pmatrix} \tag{7}$$

can be used rather than a Hadamard matrix and similar discussion is possible, where $P \in \mathbb{F}^{(n-\mathtt{ufp}(n))\times(n-\mathtt{ufp}(n))}$ and O are a permutation matrix and the zero matrix, respectively.

Note that $\sigma_i = \sigma_j \Rightarrow \sigma_i' = \sigma_j'$ holds, i.e., multiple singular values can be set. However, if $\sigma_i \neq \sigma_j$ but σ_i is very close to σ_j, then σ_i' may be equal to σ_j'. Therefore, the proposed method fails to produce clustered singular values.

References

[1] N. J. Higham, *Accuracy and Stability of Numerical Algorithms*, SIAM, Philadelphia, 2nd ed., 2002.

[2] MATLAB gallery function: https://www.mathworks.com/help/matlab/ref/gallery.html

[3] IEEE Standard for Floating-Point Arithmetic, Std 754-2008, 2008.

[4] S.M. Rump, T. Ogita, and S. Oishi, *Accurate floating-point summation part I: Faithful rounding*, SIAM J. Sci. Comput., **31**(2008), 189–224.

[5] J.J. Sylvester, *Thoughts on inverse orthogonal matrices, simultaneous sign successions, and tessellated pavements in two or more colours, with applications to Newton's rule, ornamental tile-work, and the theory of numbers*, Philosophical Magazine, **34** (1867), 461–475.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

Algorithm based on splitting deblurring and denoising for image recovery

Anantachai Padcharoen[1], Poom Kumam[2], Parin Chaipunya[3] and Dinh The Luc[4]

[1] *Department of Mathematics, King Mongkut's University of Technology Thonburi (KMUTT), 126 Pracha Uthit Rd., Bang Mod, Thung Khru, Bangkok 10140, Thailand.*

[2] *KMUTT-Fixed Point Theory and ApplicationsResearch Group, Theoretical and Computational Science Center (TaCS), Science Laboratory Building, Faculty of Science, King Mongkuts University of Technology Thonburi (KMUTT), 126 Pracha-Uthit Road, Bang Mod, Thrung Khru, Bangkok 10140, Thailand*

[3] *KMUTTFixed PointResearch Laboratory, Department of Mathematics, Room SCL 802 Fixed Point Laboratory, Science Laboratory Building, Faculty of Science, King Mongkut's University of Technology Thonburi (KMUTT), 126 Pracha Uthit Rd., Bang Mod, Thung Khru, Bangkok 10140, Thailand.*

[4] *Department of Mathematics, University of Avignon, LANG, France*

emails: apadcharoen@yahoo.com, poom.kum@kmutt.ac.th, chaipunya.p@gmail.com, dtluc@univ-avignon.fr

Abstract

In this paper, we use the popular splitting strategy to design a fast iterative algorithm for image restoration. We divide the algorithm into three steps via fast method, called fast iterative shrinkage/thresholding algorithm with backtracking to reduce image noise. we also give the convergence analysis for the proposed method. Numerical results demonstrate the efficiency and viability of the proposed algorithm, applied to l_1 regularization model and total-variation (TV) regularization model.

Key words: splitting deblurring and denoising, image recovery
MSC 2000: 47H10, 54H25.

 ISBN: 978-84-617-8694-7

1 Introduction

In the field of engineering, many application problems including image processing, compressed sensing are aiming to recover underlying image or signal from a degraded version. A degraded image or signal g can be modeled as

$$g = Ax + b, \tag{1}$$

where $x \in \mathbb{R}^n$ is the image or signal to be reconstructed, A is a $m \times n$ matrix that models the measurement process, and $b \in \mathbb{R}^m$ is an additive noise. In linear inverse problem (1), the goal is to recover image x when g and A are given. For different choices of A, recovering x becomes different application problems. For instance, it becomes the deblurring problem if A represents a blurring matrix; it becomes the inpainting problem if A represents a projection of an image onto some known pixels domain. If A is the identity matrix, it reduces to the denoising problem. But linear inverse problem (1) is usually ill-posed in image processing. For instance, in image deblurring, linear inverse problem (1) is ill-posed in the sense that the blurring matrix A is ill-conditioned and solution could be sensitive to the additive noise. The linear inverse problem (1) has infinite number of solutions.

2 Preliminaries

Definition 2.1. *[8] An operator $P : \mathbb{R}^n \to \mathbb{R}^n$ is called nonexpansive if, for any $u, v \in \mathbb{R}^n$, we have*

$$\|P(u) - P(v)\|_2 \leq \|u - v\|_2. \tag{2}$$

If there exist a number $\beta \in (0, 1)$ and a nonexpansive operator $T : \mathbb{R}^n \to \mathbb{R}^n$ such that $P = (1 - \beta)I + \beta T$ is nonexpansive, then P is called β-averaged. In particular, when $\beta = 1/2$, P is called a firmly nonexpansive operator.

Remark 2.2. *[9] The firmly nonexpansive operator is defined as*

$$\|P(u) - P(v)\|_2^2 \leq (P(u) - P(v))^T(u - v). \tag{3}$$

Lemma 2.3. *[10] Let α be a positive number and $\mathcal{R}$ be a convex and semicontinuous function. Suppose*

$$x^* = argmin_x \|u - x\|_2^2 + \alpha\mathcal{R}(x) \tag{4}$$

and define $S : \mathbb{R}^n \to \mathbb{R}^n$ such that $x = S(u)$. Then S is firmly nonexpansive.

Lemma 2.4. *[11] Let P_1 and P_2 be β_1-averaged nonexpansive operators ($\beta_1, \beta_2 \in (0, 1)$), respectively. Then P_1P_2 is the $(\beta_1 + \beta_2 - \beta_1\beta_2)$-averaged nonexpansive.*

Theorem 2.5. *[12] Let $P : \mathbb{R}^n \to \mathbb{R}^n$ be a β-averaged nonexpansive operator. If the set of fixed points of P is nonempty, then for any u_0, the sequence $\{u_k\}$, where $u_k = P(u_{k-1})$ for $k = 1, 2, 3, \ldots$, converges to a fixed point in $\mathbb{R}^n$.*

 ISBN: 978-84-617-8694-7

3 An iteration algorithm and the convergence analysis

3.1 A modified split fast iterative shrinkage/thresholding algorithm

In this section, we propose a modified split fast iterative shrinkage/thresholding algorithm to solve (1). Our method is based on the decoupling model of deblurring and denoising which is proposed in [6]. Then (1) could be rewritten as (an initial guess x_0 is used)

$$\begin{cases} \widehat{x}_k = \text{argmin}_x \|Ax - g\|_2^2 + \alpha_1 \|x - x_{k-1}\|_2^2 \\ \widetilde{x}_k = \text{argmin}_x \|x - \widehat{x}_k\|_2^2 + \alpha_2 \mathcal{R}_1(x) \\ x_k = \text{argmin}_x \|x - \widetilde{x}_k\|_2^2 + \alpha_3 \mathcal{R}_2(x), \end{cases} \tag{5}$$

where α_1 is a positive parameters for deblurring, α_2, α_3 are the positive parameters for denoising and for both $\mathcal{R}(x) = \|x\|_1$ or $\mathcal{R}(x) = \|x\|_{TV}$.

Theorem 3.1. *Let $S_c S_b S_a : \mathbb{R}^n \to \mathbb{R}^n$. If the set of fixed points of $S_c S_b S_a$ is nonempty, then algorithm* (5) *converges to a fixed point.*

Proof. By applying Theorem 2.5, we have algorithm (5) converges to a fixed point. □

Remark 3.2. *When A is full rank, i.e., the smallest eigenvalue of $A^T A$ is larger than 0, algorithm* (5) *converges to a unique fixed point for any initial vector x_0.*

4 Acknowledgments

This project was supported by the Theoretical and Computational Science (TaCS) Center under Computational and Applied Science for Smart Innovation Cluster (CLASSIC), Faculty of Science, KMUTT. The first author thanks for the support of Petchra Pra Jom Klao Doctoral Scholarship for Ph.d. student of King Mongkut's University of Technology Thonburi (KMUTT). This work was completed while the first author visit Professor Dinh The Luc at UniversityofAvignon. He thanks Professor Dinh The Luc and the University for their hospitality and support.

References

[1] L.M. Bregman, *The relaxation method for fnding common fixed points of convex sets and its application to the solution of problems in convex programming*, USSR Comput. Math. Math. Phys., **7** (1976) 200-217.

[2] G. Wang, Y. Wei, S. Qiao, Generalized Inverses: Theory and Computations (Science Press, Beijing, 2004), pp. 1-8, 129-165.

[3] S. Wang, Z. Yang, Generalized Inverse Matrix and Its Applications. Beijing University of Technology Press, Beijing (1996).

[4] J.F. Cai, S. Oher, Z. Shen, Linearized Bregman iterations for compressed sensing, Math. Comput. 78 (267) (2009) 15151536.

[5] J.F Cai, S. Osher, Z. Shen, Linearized Bregman iterations for frame-based image deblurring. SIAM J.Imag. Sci. 2(1), 226252 (2009)

[6] Y.W. Wen, M.K. Ng, W.K. Ching, Iterative algorithms based decoupling of deblurring and denoising for image restoration, SIAM J. Sci. Comput. 30 (5) (2008) 2655-2674.

[7] L.J. Deng, H. Guo, T.Z. Huang, A fast image recovery algorithm based on splitting deblurring and denoising, J. Comput. Appl. Math. 287 (2015) 88-97.

[8] P. L. Combettes, Solving monotone inclusions via compositions of nonexpansive averaged operators, Optimization, 53 (2004) 475-504.

[9] P. L. Combettes and V. R. Wajs, Signal recovery by proximal forward-backward splitting, Multiscale Model. Simul., 4 (2005) 1168-1200.

[10] Y.M. Huang, M.K. Ng, Y.W. Wen, A fast total variation minimization method for image restoration, Multiscale Model. Simul. 7 (2) (2008) 774-795.

[11] L. Sendur, I.W. Selesnick, Bivariate shrinkage functions for wavelet-based denoising exploiting interscale dependency, IEEE Trans. Signal Process. 50 (2002) 2744-2756.

[12] X.X. Guo, F. Li, M.K. Ng, A fast l1-TV algorithm for image restoration, SIAM J. Sci. Comput. 31 (3) (2009) 2322-2341.

[13] I. Daubechies, M. Defrise, and C. De Mol, An iterative thresholding algorithm for linear inverse problems with a sparsity constraint, Comm. Pure Appl. Math., 57 (2004) 1413-1457.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference on Computational and Mathematical Methods in Science and Engineering, CMMSE 2017 4–8 July, 2017.

Lattice Sums (Lennard-Jones Ingham Coefficients) for Cubic and Hexagonal Lattices

Elke Pahl[1], Antony Burrows[1] and Peter Schwerdtfeger[1]

[1] *Centre for Theoretical Chemistry and Physics, Institute for Natural and Mathematical Sciences (INMS) and The New Zealand Institute for Advanced Study (NZIAS), Massey University Albany, Private Bag 102904, Auckland 0745, New Zealand*

emails: e.pahl@massey.ac.nz, , p.a.schwerdtfeger@massey.ac.nz

Abstract

The main contribution to the cohesive (binding) energy of a solid is the two-body part containing all dimer interactions of the extended system. Using an extended Lennard-Jones functional form, $\sum_{n>3} c_n x^{-n}$, for the description of the dimer interactions one can find the following analytical expression for the two-body cohesive energy per atom [1]:

$$E_{\mathrm{ELJ}}^{\mathrm{coh}} = \frac{1}{2} \sum_{n>3} c_n L_n r_s^{-n} \tag{1}$$

This formula has the beauty of being only dependent on the potential parameters, c_n, the next-nearest neighbor distance in the crystal, r_s, and the so-called Lennard-Jones Ingham coefficients, L_n. The availability of an analytical form also allow us to derive analytical expression for other solid-state properties like pressure, bulk modulus and even zero-point energy.

The Lennard-Jones coefficients are lattice sums which only depend on the underlying symmetry of the lattice and as thus, have to be calculated only once for every lattice symmetry. Nevertheless, an accurate computation is difficult because these three-dimensional lattice sums present very slowly converging series connected to the well-known number-theoretical problem of finding the number of representations of a sum of N squares. Fast-converging formulae can be found when the lattice sums can be written in form of homogeneous quadratic equations ([2] and references therein). While such formulae already exist for cubic lattices, the only know formula by Kane and Goeppert-Mayer [3] is plagued by inhomogeneities.

Here, we will illustrate the origin of and connection between different cubic lattice sums by visulazing the underlying 3D structures. For the hexagonal lattice we derive

a new lattice formula only consisting of sums over homogeneous quadratic forms which makes a fast and accurate evaluation accessible.

Key words: Lattice Sums, Cohesive Energy, Lennard-Jones Ingham Coefficients

Acknowledgements

This work has been supported by the Marsden fund administered by the Royal Society of New Zealand.

References

[1] P. Schwerdtfeger, N. Gaston, R. P. Krawczyk, R. Tonner, and G. E. Moyano, *Extension of the Lennard-Jones potential: Theoretical investigations into rare-gas clusters and crystal lattices of He, Ne, Ar and Kr using many-body interaction expansions*, Phys. Rev. B **73** (2006) 064112–1–19.

[2] J. M. Borwein, M. Jonathan, M. L. Glasser, R. C. McPhedran, J.G. Wan, and I. J. Zucker, IJ, *Lattice sums then and now*, **150** (2013), Cambridge University Press.

[3] G. Kane and M. Goeppert-Mayer, *Lattice Summations for Hexagonal Close-Packed Crystals*, JCP, **8** (1940) 642.

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

Yinyang K-means clustering for hyperspectral image analysis

Mercedes Eugenia Paoletti[1], Juan Mario Haut[1], Javier Plaza[1] and Antonio Plaza[1]

[1] *Department of Technology of Computers and Communications, University of Extremadura, Escuela Politecnica, Avda. de la Universidad s/n*

emails: mpaolett@alumnos.unex.es, juanmariohaut@unex.es, jplaza@unex.es, aplaza@unex.es

Abstract

Hyperspectral images are widely used in remote sensing applications due to their wealth of information in the spectral domain, that allows for very detailed scene classification. Clustering is one of the most used unsupervised techniques for the analysis of these scenes. Popular clustering techniques such as K-means are computationally expensive, particularly when applied to hyperspectral images characterized by their large dimensionality. An efficient implementation of K-means is the so-called *Yinyang K-means*, which outperforms K-means algorithms by clustering the centers in the initial stage, and leveraging efficiently maintained lower and upper bounds between each point and the cluster centers. In this work, we have adapted an efficient implementation of this algorithm using graphics processing units (GPUs) for hyperspectral image analysis. We have carried out a comparison of this technique with other existing implementations with the aim of demonstrating its usefulness in hyperspectral imaging. Our obtained results suggest that this technique is ideal for working with big hyperspectral data repositories.

Key words: Hyperspectral imaging, k-means clustering, YinYang K-means, GPUs.

1 INTRODUCTION

Current Earth observation (EO) sensors acquire and produce high-dimensional data cubes with hundreds of spectral channels and millions of pixels. For instance, NASA's Jet Propulsion Laboratory's Airbone Visible/Infrared Imaging Spectrometer (AVIRIS) [1] measures the solar reflected spectrum from 400nm to 2500nm at intervals of 10nm. The

 ISBN: 978-84-617-8694-7

EO-1 Hyperion imaging spectrometer collects bands in the range of 400nm to 2500nm too [2, 3]. The resulting hyperspectral datasets [4] provide information corresponding to large observation areas on the surface of the Earth, using hundreds of contiguous spectral bands. As a result, these instruments can produce three-dimensional data cubes with size significantly larger than traditional images. These images can be exploited in many practical applications, such as monitoring and management of the environment and agriculture, urban and regional planning, detection of relevant geological zones (e.g. mineral detection) or defense and intelligence issues (e.g. target detection or mine detection).

However, hyperspectral images present many challenges in terms of storage and processing due to their large dimensionality. In addition, modern sensors are producing an almost continuous stream of data [3]. For example, AVIRIS has a data collection rate of 2.5 MB/s and Hyperion collects almost 71.9 GB/hour. On the other hand, most of the satellite missions that will be soon in operation, such as the environmental mapping and analysis program (EnMAP http://www.enmap.org/) present similar data collection ratios. This creates the need for scalable and efficient processing techniques for hyperspectral data in the context of different applications [3].

Many techniques (supervised and unsupervised) have been developed to address the aforementioned challenges. One of the most widely used unsupervised methods is clustering, which aims to organize the data so that pixels with similar spectral content are clustered together in the same class [5]. In this case, there is no need for labeled samples which are common in supervised techniques [6]. Although clustering offers an unsupervised alternative that has been widely used in various fields, it is also a very challenging task due to the large spectral variability and complex spatial structures present in hyperspectral images. The most popular and widely used family of clustering algorithms is represented by centroid-based clustering methods such as K-means [7].

K-means assumes that similar pixels always form clusters in feature space. By applying this method to hyperspectral images we can obtain satisfactory results, but K-means is hampered by its computational complexity. There are a handful of studies which aim to address this issue, either adapting the algorithm to parallel processing structures such as field-programmable gate array (FPGAs) [5, 8]or cloud computing architectures [9]. Other works aim at developing improved implementations such as K-means++ [10, 11], the $AFKMC^2$ [12], the K-means projective clustering [13] or the filtered K-means [7]. The *Yinyang* K-means [14] is a recent improvement of K-means. This method features a space-conscious elastic design that adaptively uses the upper and lower bound based filters while maintaining various space constraints. The upper and lower bound based filters are continuously and carefully maintained, and provide an efficient evolvement and interplay mechanism.

Our main goal in this paper is to adapt the *Yinyang* K-means to hyperspectral image processing. The implementation that we have adopted is optimized for GPUs using

 ISBN: 978-84-617-8694-7

NVIDIA Compute Device Unified Architecture (CUDA). To test the effectiveness of the implementation, we have compared it with other existing k-means implementations and made an exhaustive analysis of the pros and cons of all the implementations used.

The remainder of the paper is organized as follows. Section 2 will delve into the K-means method and the improvements provided by the *Yinyang* K-means version, presenting its theoretical foundations. Section 3 validates the *Yinyang* K-means algorithm by comparing it with other implementations in terms of execution times. Finally, section 4 concludes with some remarks and hints at plausible future research lines.

2 K-means method: an overview

2.1 The K-means clustering algorithm

K-means is one of the easiest unsupervised learning algorithms and most widely used method to group data in a specified number of clusters. Suppose a set of n observations $X = (x_1, x_2, ..., x_n)$, where each observation is $x_i \in \mathbb{R}^d$, i.e. $x_i = [x_{i_1}, x_{i_2}, ..., x_{i_d}]$ (where d is the number of spectral channels). The goal is to group each observation into a number of *clusters* k fixed a priori ($k <= n$). Iteratively, K-means calculates the centers of the k groups, optimizing the error of each group as $\min \sum_{j=1}^{k} \sum_{i=1}^{n_k} \parallel x_i^j - c_j, \parallel^2$ where $\parallel x_i^j - C_j \parallel^2$ is the euclidean distance between a data point x_i^j of the cluster j (n_k is the observations within each cluster) and the cluster center c_j, which if it is the point that minimizes the equation also called *centroid of cluster j*.

K-means algorithm successfully performs the task of obtaining useful information from the dataset, such as the best distance metric for the data [15]. However, the results can vary greatly due to a small change in parameters and in the choice of the initial centers. So, a proper initialization will result in a final best solution. In order to obtain a set of good initial cluster centers, several methods have been proposed, as K-means++ [10, 11]. This algorithm obtains a set of k initial centers which are generally very close to the final solution.

2.2 *Yinyang* K-means method

The main problem with the traditional K-means implementation is that it performs a lot of redundant work when recalculating distances corresponding to samples which are not going to change the cluster. With this in mind, *Yinyang* K-means optimizes two important points in the K-means algorithm: the assignment steps and the update steps. In order to do that, it implements two filters and a new center update method.

In the standard K-means, the assignment step computes the distances between every point x_i and every cluster center c_j in order to find out the closest center to each point.

 ISBN: 978-84-617-8694-7

The *Yinyang* K-means instead uses two filters to detect which distance calculations are unnecessary and avoids computing them. These filters are based on the triangle inequality:

- Group filtering groups the k clusters into t groups $G = g_1, g_2, ..., g_t$, where each $g_i \in G$ is a set of the clusters and t must no be greater than $k/10$. t provides a design knob for controlling the space overhead and redundant distance elimination. For each group, it calculates:

  1. Upper bound: for each point x in cluster j ($j = j(x)$) it sets the upper bound to $uj(x) = d(x, j(x))$, i.e. the distance between x and cluster. The upper bound is updated as $uj'(x) = uj(x) + \delta(j)$, where $\delta(j)$ is the distance of the cluster j.
  2. Lower bound: for each point x in cluster j ($j = j(x)$) it sets the lower bound $lj(x, g_i)$ as the shortest distance between x and all centers in g_i excluding $j(x)$. The lower bound is updated as $lj'(x, g_i) = lj(x, g_i) - max_{c \in g_i}\delta(c)$, $\delta(c) = d(c, c')$ is the shift of cluster center due to the center update.

  If the updated lower bound is major than the updated upper bound, $lj'(x, g_i) > uj'(x)$, no reassignment is needed for point x and all the group-level comparisons can be avoided.

- Local filtering is used for get the new centroid of a cluster, avoiding unnecessary distance operations. A new center $c' \in g'_i$ cannot be the closest center to a point x if there is another center $p' \neq c'$ such that $d(x, p') < lj(x, g_i) - \delta(c)$

On the other hand, in the updating step, K-means computes the new center for each cluster. In *Yinyang* K-means new centers are computed as $c' = \frac{c \cdot |V| - \left(\sum_{y \in V - OV} y\right) + \sum_{y' \in V' - OV} y'}{|V'|}$ where V' and V denote a cluster in the current and previous iteration, OV is $V \cap V'$, c is the old center and c' is the new center.

2.3 Parallel GPU *Yinyang* K-means

The parallel implementation of *Yinyang* K-means that we have adopted is available in a library. It has been optimized for low memory consumption and use of a large number of clusters, bearing in mind the limitations of the K-means algorithm related to the calculation of the centroids that requires having all the data be available in the same place. In the context of the CUDA-based GPU implementation, if the data occupies a large amount of storage, it is impossible to store them in the memory of a single GPU, so it is necessary to cut the samples into as many intervals as GPUs are available. As a result, in our adopted implementation each GPU will work with its own interval, calculating the distances and the local centroids, writing the local assignments and broadcasting its results to other GPUs.

The parameters used by our adopted implementation are the number of clusters, k, the number of cluster groups (t), and the value for *tolerance*, which will be used by the algorithm

 ISBN: 978-84-617-8694-7

to stop its execution (if the number of reassignments drop below the tolerance value). The parallel *Yinyang* K-means execution is initialized with random centroids (or it can also be initialized by intelligently produced centroids such as those produced by K-means++) and calculates the Euclidean distance (L2) to perform the centroid selection calculations as $d_2(\overrightarrow{x}, \overrightarrow{y}) = \sqrt[2]{\sum_i (x_i - y_i)^2}$.

The output of the method is a vector of centroids and a vector with the cluster index for each sample (numerical labels for each pixel).

3 Experiments and results

3.1 Experimental Configuration

In order to evaluate the performance of the adopted *Yinyang* K-means implementation, we use a hardware environment composed by a 6th Generation Intel® Core™i7-6700K processor with 8M of Cache and up to 4.20GHz (4 cores/8 way multitask processing), 32GB of DDR4 RAM with a serial speed of 2400MHz, a GPU NVIDIA GeForce GTX 1080 with 8GB GDDR5X of video memory and 10Gbps of memory frequency, a Toshiba DT01ACA HDD with 7200RPM and 2TB of capacity, and an ASUS Z170 pro-gaming motherboard. On the other hand, the software environment is composed by Ubuntu 16.04.4 x64 as operating system, CUDA 8 and Python.

3.2 Hyperspectral data sets

In our experiments, we use four different hyperspectral images. The first one was collected by AVIRIS [1] in 1992 over a set of agricultural fields with regular geometry and with a multiple crops and irregular patches of forest in Northwestern Indiana. This scene, Indian Pines, has 145x145 pixels with 224 spectral bands in the range from 400 to 2500nm, with 10nm of spectral resolution, 20nm moderate spatial resolution and 16 bits radiometric resolution. 4 zero bands plus 20 bands with lower signal-to-noise ratio (SNR) have been removed, retaining 200 spectral channels. Dataset has 16 ground-truth classes(Fig. 1).

Also, we use a larger version of the Indian Pines scene. This one has a much larger size of 2678×614 pixels. It was collected over the same area that small Indian Pines, but spanning a much larger extent. It contains 220 spectral bands in the range from 400 to 2500 nm, with spectral resolution of 10 nm, moderate spatial resolution of 20 nm and 16 bits of radiometric resolution. The total number of classes is 58. (Fig. 4).

The third dataset was collected by AVIRIS over Salinas Valley, California (Fig. 3). The area covered has 512×217 samples and the spatial resolution is 3.7 m per pixel. 204 out of the 224 bands are kept after 20 water absorption bands are removed. Dataset has 16 land-cover classes.

 ISBN: 978-84-617-8694-7

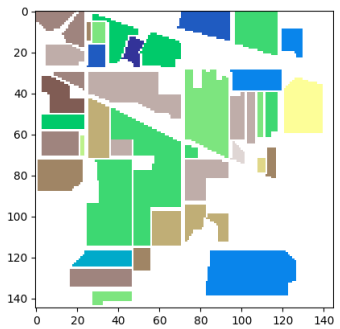


Figure 1: Ground-truth of small Indian Pines scene.

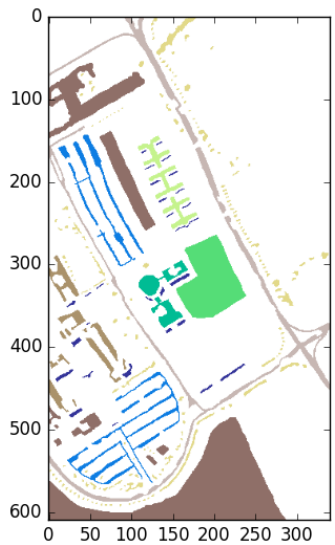


Figure 2: Ground-truth of University of Pavia.

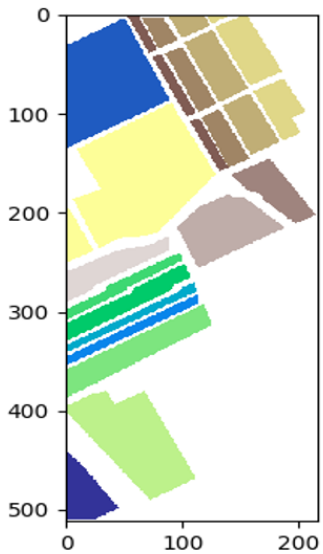


Figure 3: Ground-truth of Salinas scene.

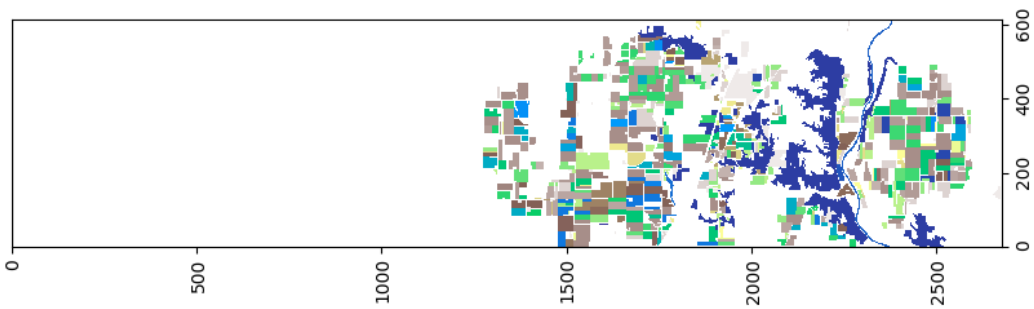


Figure 4: Ground-truth of big Indian Pines scene.

The fourth hyperspectral dataset was collected by the Reflective Optics System Imaging Spectrometer (ROSIS) sensor [16] during a flight campaign over Pavia, northern Italy. The dataset covers an urban environment, with various solid structures, natural objects and shadows (9 classes in total). The scene (see Fig. 2) contains 103 spectral bands of 610×340 pixels in the spectral range from 0.43 to 0.86μm, with spatial resolution of 1.3m/pixel.

3.3 Performance evaluation

In order to evaluate the performance of parallel *Yinyang* K-means, several experiments have been executed. The first one is a comparison between the GPU version of *Yinyang*

 ISBN: 978-84-617-8694-7

K-means and other iterative and parallel GPU implementations of the original (Lloyd) K-means algorithm setting the maximum number of centroids to the number of classes. The tolerance value was set to 0.001. To complete the experiment, we have tested two initiations of centroids: 1) completely random and 2) K-means++ method. Each version has been executed 10 times and the average times are reported for statistical significance. The obtained results are reported in Table 1.

Initialization	**Lloyd Iterative**		**Lloyd CUDA**		**Yinyan CUDA**	
	Small Indian Pines					
	Time	*Speed up*	*Time*	*Speed up*	*Time*	*Speed up*
k-means++	0.183 (0.015)	1	0.421 (0.062)	0.435	0.441 (0.066)	0.415
random	0.185 (0.007)	1	0.520 (0.050)	0.355	0.511 (0.056)	0.361
	Pavia University					
	Time	*Speed up*	*Time*	*Speed up*	*Time*	*Speed up*
k-means++	0.238 (0.017)	1	0.839 (0.189)	0.283	0.837 (0.242)	0.284
random	0.223 (0.07)	1	0.998 (0.124)	0.223	1.088 (0.159)	0.205
	Salinas					
	Time	*Speed up*	*Time*	*Speed up*	*Time*	*Speed up*
k-means++	0.726 (0.078)	1	2.436 (0.616)	0.298	2.083 (0.567)	0.348
random	0.715 (0.5)	1	1.978 (0.331)	0.361	2.231 (0.384)	0.320
	Big Indian Pines					
	Time	*Speed up*	*Time*	*Speed up*	*Time*	*Speed up*
k-means++	38.584 (2.694)	1	46.167 (10.158)	0.836	40.085 (10.753)	0.963
random	37.217 (2.960)	1	69.518 (6.057)	0.535	56.321 (11.650)	0.661

Table 1: Average time executions (standard deviation) and speed-up for each implementation of K-means initialized with random centroids and K-means++.

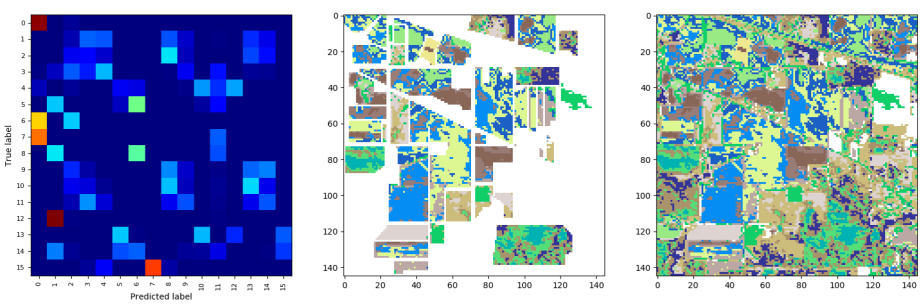


Figure 5: Small Indian Pines *Yinyang* K-means classification results: the confusion matrix and the classification maps without background and with background.

For small Indian Pines image, the fastest K-mean implementation is the original Lloyd

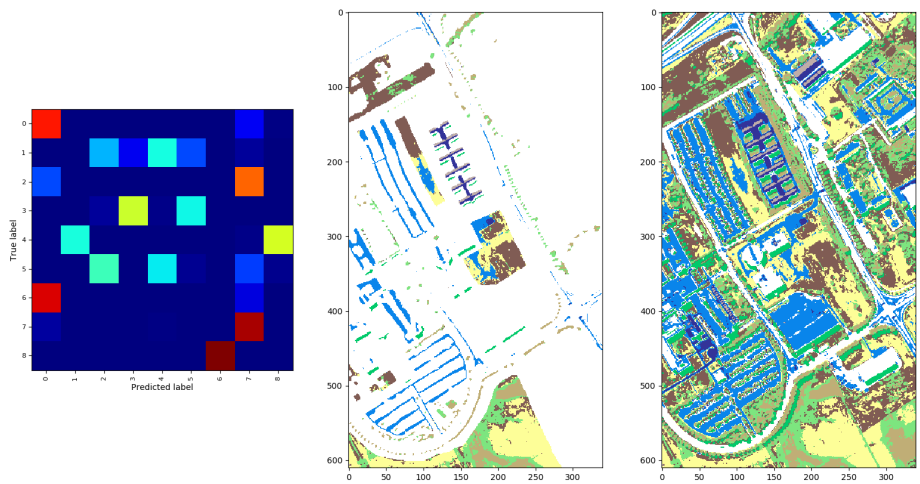


Figure 6: Pavia University *Yinyang* K-means classification results: the confusion matrix and the classification maps without background and with background.

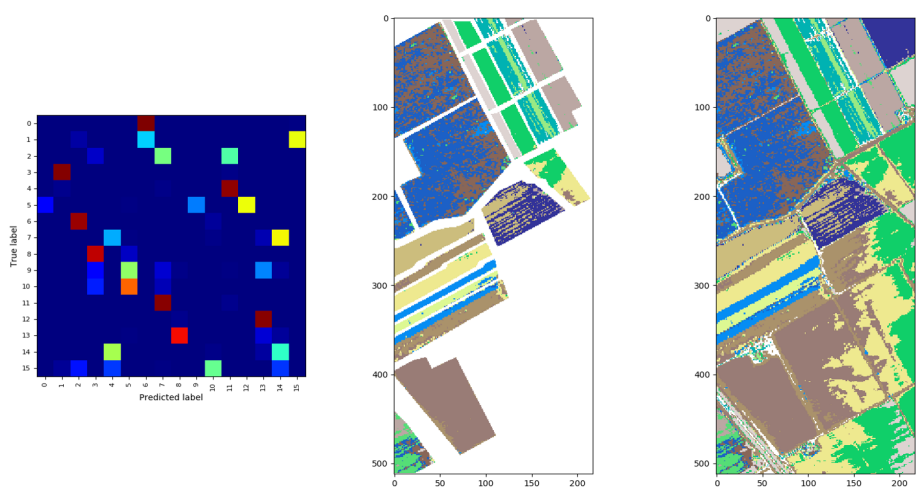


Figure 7: Salinas *Yinyang* K-means classification results: the confusion matrix and the classification maps without background and with background.

iterative version with K-means++ initialization. It is 2.41 times faster than the *Yinyang* cuda implementation with K-means++ initialization and 2.79 times faster than the same algorithm with random initialization. This is due to the small size of the image (only 145×145 pixels in only 16 groups), which is not enough to get the most out of GPU versions. In Fig. 5 we can see the classification results of *Yinyang* K-means. The confusion

matrix is a typical mechanism to evaluate unsupervised clustering methods, where in our representation warm colors indicate a high value and warm colors indicate a low value. The two classification maps show the clustering result without and with background pixels. As we can observe, the obtained result is reasonable for a K-means method, despite some noise at the borders of the classes.

For Pavia University dataset, the fastest implementation is also the iterative version and, among the two parallel GPU versions, the original LLoyd algorithm is faster than *Yinyang*. In this case we have more data than Indian Pines, but still enough complexity (only 9 centroids). In Fig. 6 we can observe the confusion matrix of the classification with *Yinyang*, whose results are better than in the Small Indian Pines. The classification maps reveal less noise at the borders.

For Salinas we have similar results: although we have a lot of data, the complexity is not enough (only 16 centroids). So, the iterative version is still the fastest. But we can see that *Yinyang* with K-means++ initialization is better than the original Lloyd algorithm parallelized in GPU: the differences between GPU versions are already starting to appear. In Fig. 7 we can see that the classification results with *Yinyang* are better than in the small Indian Pines and Pavia University images. Specifically, border pixels are better identified.

Finally, for big Indian Pines the fastest implementation is the iterative Lloyd algorithm (2678×614 with 58 centroids), but if we compare the two GPU versions, the *YinYang* is faster than the parallel GPU Lloyd. Since the classification results are similar to the ones already reported for the small Indian Pines image, we do not include these results for space considerations.

In summary, our results indicate that *YinYang* K-means works better than the CUDA version of Lloyd algorithm when more data needs to be processed, but higher complexity appears to be needed in order to improve the iterative version. So, we repeated the first experiment increasing the number of centroids to be calculated in a second experiment, which compares the parallel GPU implementation of *Yinyang* K-means with the same implementations of the original K-means algorithm, using a maximum number of centroids set to one hundred times the number of classes in each scene (i.e. 1600, 900, 1600 and 5800 centroids, respectively). Again, this is intended to increase the analysis complexity. The tolerance value is 0.001 in all cases. Again,we tested with random and K-means++ [10, 11] initializations. The obtained results are reported in Table 2.

For small Indian Pines dataset the fastest implementation is the *Yinyang* K-means with random initialization. It reaches a speed up of 7.071 over the iterative version. In the first experiment, with 16 classes the execution times were 0.441 and 0.511, at this time with 1600 centroids to calculate the execution times of *Yinyang* increase in just one second. However for iterative version, it needed 0.183-0.185 seconds and now it needs 6 seconds more. Also, for Pavia University scene *Yinyang* K-means is the fastest implementation, with K-means++ initialization. With the same number of pixels and 900 centroids, *Yinyang*

 ISBN: 978-84-617-8694-7

Initialization	Iterative		CUDA		Yinyan CUDA	
	Small Indian Pines					
	Time	*Speed up*	*Time*	*Speed up*	*Time*	*Speed up*
k-means++	6.034 (0.310)	1	1.912 (0.134)	3.191	1.542 (0.024)	3.956
random	6.687 (0.245)	1	1.131 (0.050)	5.410	0.865 (0.025)	7.071
	Pavia University					
	Time	*Speed up*	*Time*	*Speed up*	*Time*	*Speed up*
k-means++	14.124 (0.845)	1	4.166 (0.221)	3.350	2.884 (0.266)	4.839
random	12.941 (0.780)	1	3.526 (0.278)	4.119	2.573 (0.209)	5.644
	Salinas					
	Time	*Speed up*	*Time*	*Speed up*	*Time*	*Speed up*
k-means++	40.308 (1.938)	1	13.230 (0.767)	3.077	5.879 (0.196)	6.926
random	37.507 (2.460)	1	13.630 (1.229)	3.155	5.602 (0.146)	7.678
	Big Indian Pines					
	Time	*Speed up*	*Time*	*Speed up*	*Time*	*Speed up*
k-means++	2331.929 (233.417)	1	461.223 (11.586)	5.022	121.726 (1.909)	19.029
random	2022.694 (260.634)	1	463.165 (16.392)	5.132	113.459 (3.653)	20.952

Table 2: Average time executions (standard deviation) and speed-up for each implementation of K-means initialized with random centroids and K-means++.

K-means needs only two more seconds. The iterative version needs 12 or 14 seconds more, from the first experiment where it needed 0.23-0.22 seconds. For Salinas dataset we observe the same behavior: with 1600 centroids to calculate, the fastest one is *Yinyang* in GPU reaching a speedup of 7.678. For the big Indian Pines image, *Yinyang* K-means reaches a significant speedup: a 20.95. These results show that, the more the complexity of the analysis, the better the performance of the *Yinyang* GPU which is intended for big data problems involving not only massive data repositories but also complex analysis scenarios.

4 Conclusions and Future Lines

In this paper, we have proved a recent variant of K-means, the *Yinyang* K-means algorithm, to hyperspectral image analysis, in particular a parallel GPU implementation, which has been shown to obtain good processing results in hyperspectral image analysis when compared with other popular K-means implementations. Specifically, our experimental results show the effectiveness of the parallel GPU implementation of *Yinyang* K-means using four different hyperspectral scenes. The algorithm performs particularly effectively when we need to process big data and calculate a large set of centroids. The method not only improves as more data become available, but also with the increase of the complexity of the clusterization. On the other hand, the ranking results are in line with those obtained by any K-means algorithm. As future work, we are planning on using the *Yinyang* K-means in

 ISBN: 978-84-617-8694-7

conjunction with other techniques for hyperspectral image classification (e.g. supervised and semi-supervised techniques) with the aim of improving the obtained classification results.

Acknowledgements

This work has been supported by Ministerio de Educación (Resolución de 26 de diciembre de 2014 y de 19 de noviembre de 2015, de la Secretaría de Estado de Educación, Formación Profesional y Universidades, por la que se convocan ayudas para la formación de profesorado universitario, de los subprogramas de Formación y de Movilidad incluidos en el Programa Estatal de Promoción del Talento y su Empleabilidad, en el marco del Plan Estatal de Investigación Científica y Técnica y de Innovación 2013-2016). This work has also been supported by Junta de Extremadura (decreto 297/2014, ayudas para la realización de actividades de investigación y desarrollo tecnológico, de divulgación y de transferencia de conocimiento por los Grupos de Investigación de Extremadura, Ref. GR15005).

References

[1] Robert O. Green, Michael L. Eastwood, Charles M. Sarture, Thomas G. Chrien, Mikael Aronsson, Bruce J. Chippendale, Jessica A. Faust, Betina E. Pavri, Christopher J. Chovit, Manuel Solis, Martin R. Olah, and Orlesa Williams. Imaging spectroscopy and the Airborne Visible/Infrared Imaging Spectrometer (AVIRIS). *Remote Sensing of Environment*, 65(3):227–248, 1998.

[2] Amin Beiranvand Pour and Mazlan Hashim. ASTER, ALI and Hyperion sensors data for lithological mapping and ore minerals exploration. *SpringerPlus*, 3(1):130, 2014.

[3] A. Plaza, J. Plaza, A. Paz, and S. Sanchez. Parallel Hyperspectral Image and Signal Processing. *IEEE Signal Processing Magazine*, 28(3):119–126, 2011.

[4] Chein-I Chang. *Hyperspectral Imaging: Techniques for Spectral Detection and Classification*. Springer US, 2003.

[5] Miriam Leeser, Pavle Belanovic, Michael Estlick, Maya Gokhale, John J Szymanski, and James Theiler. Applying Reconngurable Hardware to the Analysis of Multispectral and Hyperspectral Imagery. In International Society for Optics and Photonics, editor, *International Symposium on Optical Science and Technology*, pages 100–107, 2002.

[6] Antonio Plaza, Javier Plaza, Gabriel Martín, and Sergio Sánchez. Hyperspectral Data Processing Algorithms. In Alfredo Huete Prasad S. Thenkabail, John G. Lyon, editor, *Hyperspectral Remote Sensing of Vegetation*, chapter 5, pages 121–137. Taylor & Francis, 2011.

[7] Tapas Kanungo, David M Mount, Nathan S Netanyahu, Christine D Piatko, Ruth Silverman, and Angela Y Wu. An Efficient k-Means Clustering Algorithm: Analysis and Implementation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(7):881–892, 2002.

[8] Abel Guilhermino, Da S Filho, Alejandro C Frery, Cristiano Coêlho De Araújo, Haglay Alice, Jorge Cerqueira, Juliana A Loureiro, Manoel Eusebio De Lima, Maria Das, Graas S Oliveira, and Michelle Matos Horta. Hyperspectral Images Clustering on Reconfigurable Hardware using the K-Means Algorithm. In *16th Annual Symposium on Integrated Circuits and Systems Design (SBCCI)*, pages 8–11, Sao Paulo (Brasil), 2003.

[9] J.M. Haut, M. Paoletti, J. Plaza, and A. Plaza. Cloud implementation of the K-means algorithm for hyperspectral image analysis. *Journal of Supercomputing*, 73(1), 2017.

[10] David Arthur and Sergei Vassilvitskii. k-means++: The Advantages of Careful Seeding. In ACM, editor, *Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1027–1035. Society for Industrial and Applied Mathematics, 2007.

[11] Olivier Bachem, Mario Lucic, S Hamed Hassani, and Andreas Krause. Approximate K-Means++ in Sublinear Time. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, pages 1459–1467, Phoenix, Arizona, 2016. AAAI Press.

[12] Olivier Bachem, Mario Lucic, S Hamed Hassani, and Andreas Krause. Fast and Provably Good Seedings for k-Means, 2016.

[13] K. Agarwal, Pankaj and Nabil H. Mustafa. K-Means Projective Clustering. In *Proceedings of the Twenty-third ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, pages 155–165, Paris- France, 2004. ACM.

[14] Yufei Ding, Yue Zhao, Ncsuedu Xipeng Shen, Madanlal Musuvathi, and Microsoftcom Todd Mytkowicz. Yinyang K-Means: A Drop-In Replacement of the Classic K-Means with Consistent Speedup. In *Proceedings of the 32nd International Conference on Machine Learning*, page 579587, Lille, France, 2015. JMLR: W&CP.

[15] Justin Sunu. *Applications of K-means and Spectral Clustering to Hyperspectral Video Sequences*. PhD thesis, California State University, Long Beach, 2014.

[16] B. Kunkel, F. Blechinger, R. Lutz, R. Doerffer, and H. van der Piepen. ROSIS (Reflective Optics System Imaging Spectrometer) - A candidate instrument for polar platform missions. In J. Seeley and S. Bowyer, editors, *Optoelectronic technologies for remote sensing from space*, pages 134–141, 1988.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

Hardware implemented ECC co-processor for High-Performance Cryptographic Servers

L. Parrilla[1], J.A. Alvarez-Bermejo[2], E. Castillo[1], J.A. Lopez-Ramos[3] and D.P. Morales[1]

[1] *Department of Electronics and Computer Technology, University of Granada*

[2] *Department of Informatics, University of Almeria*

[2] *Department of Mathematics, University of Almeria*

emails: lparrilla@ditec.ugr.es, jaberme@ual.es, encas@ditec.ugr.es, jlopez@ual.es, diegopm@ugr.es

Abstract

Security threats affecting electronics communications in the current world make necessary the encryption and authentication of every transaction. The increasing levels of security required are leading to an overload of transactions servers due to cryptographic tasks. In this paper a hardware-implemented co-processor for Elliptic Curve Cryptography operations for accelerating secure services is presented. The proposed co-processor can be implemented in last generations of FPGAs allowing to have in the same chip the secure web/database server and the cryptographic co-processor.

Key words: High-Performance cryptographic servers, security, FPGAs, hardware-software co-design.

1 Introduction

The number of electronic transactions is increasing every day in all aspects of daily life: financial, on-line shopping, domotics, automotive, an so on. Security of these transactions constitutes a challenge requiring more and more computing resources because of the number of devices connected, and the proliferation of cyber-criminals. In fact, the interchange of information trough computing networks must be secured by means of public-key cryptography combined with symmetric cryptosystems and secure hash functions. Some of the

ISBN: 978-84-617-8694-7

algorithms involved have high computing requirements, specially those related to public-key cryptography. Moreover, every year attackers have available more powerful computing systems, making some of these algorithms obsolete and/or insecure like the recent case of SHA1 [1]. In this context, servers with heavy loads of secure electronics transactions require high computing resources leading to high power consumption. Hardware accelerators [2, 3, 4] can be a solution for freeing the server from cryptographic computations, but they must be carefully designed for achieving an optimal trade-off between computation resources and area requirements. In fact, a powerful Elliptic Curve cryptographic co-processor can perform scalar-point operations in a microseconds, but probably the CPU of the server can not generate requests at those rates, thus maintaining the crypto-processor in idle state the majority of the time. Consequently, it would be more efficient to use more compact crypto-processors requiring less area, but implemented in smaller devices (and with less power consumption). In the other hand, new generations of Field Programmable Devices, such as Zynq from Xilinx [5] or Stratix 10 from Altera [6] include powerful microprocessors enabling the possibility of having a complete software server together with hardware-implemented co-processors in the same chip. Taking advantage of the co-design possibilities enabled by these devices, in this paper, three different implementations of an ECC (Elliptic Curve Cryptography) hardware co-processor co-existing with a CPU in the same Programmable Device are presented and compared in terms of performance an area requirements.

2 Background

Secure Sockets Layer (SSL) and Transport Layer Security (TLS) [7, 8] are the standard protocols widely used for exchanging securely information trough insecure channels such as the Internet. Within the algorithms included in the cipher-suites [8] of these protocols, those related to public-key exchange (RSA, ECDH) and authentication (RSA,DSA, ECDHA) are the most resources consuming. Hardware implementation of these algorithms would free the main server of being busy with heavy cryptographic operations and thus would allow better attending and managing the queues of clients asking for transactions. Among the algorithms dedicated to public-key exchange and authentication, those based on ECC are the most suitable for being implemented in hardware [9]. Moreover, ECC presents some other advantages with respect to RSA related to the key size [10, 11, 12, 13] (1024-bit RSA is equivalent to 163-bit ECC), thus minimizing communications between the main CPU and the cryptographic co-processor. Therefore, in this paper we propose using ECC defined over binary fields for implementing hardware co-processors for helping servers to attend massive secure transactions over computer networks.

 ISBN: 978-84-617-8694-7

2.1 Elliptic Curve Cryptography

An elliptic curve E defined over a finite field $GF(q)$ consists on a set of points $P = (x_p, y_p)$ where x_p and y_p are elements of $GF(q)$ satisfying the Weirstrass equation [15], together with the point at infinite, O. The curves are defined by means of two coefficients $a \in GF(q)$ and $b \in GF(q)$, named the coefficients of E. IF q is a power of 2, must be $b \neq 0$ in $GF(2^m)$, and the points $P = (x_p, y_p)$ over E (except O), satisfy the equation:

$$y_p{}^3 + x_p y_p = x_p{}^3 + a x_p{}^2 + b \tag{1}$$

If $q = 2^m$, the field elements can be represented using polynomial bases, by choosing an irreducible polynomial $f(t)$. In this case, an element a of the field is represented using the bit string:

$$a = (a_{m-1}...a_2 a_1 a_0) \tag{2}$$

corresponding to the polynomial

$$a(t) = a_{m-1}t^{m-1} + ...a_2 t^2 + a_1 t + a_0 \tag{3}$$

where a_i are elements of $GF(2)$. One of the characteristics enabling the use of elliptic curves in cryptography is the possibility of defining an internal operation into the curve, named elliptic addition. Geometrically, the sum of two points P, Q is given by a point $R = P + Q$ with the property that P, Q, and $-R$ lies on the same straight line. From the full elliptic addition, the scalar product of a point P of the curve E and a natural n is defined as:

$$nP = P + .^n. + P \tag{4}$$

This operation, that is the additive equivalent to the exponentiation in multiplicative abelian groups, constitutes the base of the cryptography using elliptic curves [10].

Given a curve E defined over a field $GF(q)$, r a positive prime integer dividing the number of points on E, $\#E$, and a curve point G of order r (generator of a subgroup of order r), an EC key pair can be defined. Specifically, taking $s \in [1, r-1]$ as an EC private key, and $W = sG$ as the associated EC public key, a public key cryptosystem for interchanging a secret value can be established. In fact, if two communicants A and B generate their corresponding EC key pairs $(s_A, W_A), (s_B, W_B)$, party A can compute the secret value $P = s_A \cdot W_B$ using the public key of B. Then, B can recover the secret value from the public key of A, making $P = s_B \cdot W_A$ because:

$$P = s_A \cdot W_B = s_A \cdot s_B \cdot G = s_B \cdot s_A \cdot G = s_B \cdot W_A \tag{5}$$

This is the basic principle for public key cryptography using EC, and the conditions to be met by the parameters involved and the details of the different primitives can be found in [10, 14, 15, 16].

 ISBN: 978-84-617-8694-7

2.2 Selection of the cipher-suite and ECC curves

As SSL is considered obsolete and insecure [17], we have selected the following cipher-suites defined in TLS 1.2 [8] and based in ECC[18]:

- TLS_ECDH_ECDSA_WITH_NULL_SHA
- TLS_ECDH_ECDSA_WITH_RC4_128_SHA
- TLS_ECDH_ECDSA_WITH_3DES_EDE_CBC_SHA
- TLS_ECDH_ECDSA_WITH_AES_128_CBC_SHA
- TLS_ECDH_ECDSA_WITH_AES_256_CBC_SHA
- TLS_ECDHE_ECDSA_WITH_NULL_SHA
- TLS_ECDHE_ECDSA_WITH_RC4_128_SHA
- TLS_ECDHE_ECDSA_WITH_3DES_EDE_CBC_SHA
- TLS_ECDHE_ECDSA_WITH_AES_128_CBC_SHA
- TLS_ECDHE_ECDSA_WITH_AES_256_CBC_SHA

When using these cipher-suites, public-key exchange is performed using ECDH (ECC Diffie-Hellman) or ECDHE (Ephemeral ECDH) [19], and digital signature is performed by means of ECDSA [20, 21]. In both cases, ECC scalar-point operations are required. Statistically, the two most used curves in TLS 1.2 [19] are:

- secp256r1 (NIST P-256)
- sect233r1 (NIST B-233)

The first one is defined over a $GF(p)$ field, while sect233r1 is defined over a binary field ($GF(2^{233})$), thus being more affordable in terms of resources and performance for a hardware implementation. Therefore, it will be the standard curve used in our design.

2.3 ECC domain parameters

As pointed out in previous subsection, the NIST B-233 [21] curve has been selected for the hardware implementation of the ECC co-processor. Thus, the EC domain parameters will be the following:

- $m = 233$

 ISBN: 978-84-617-8694-7

- $f = t^{233} + t^{74} + 1$
  (reduction polynomial for the field)
- $a = 1$
- b= 066 647EDE6C 332C7F8C 0923BB58 213B333B 20E9CE42 81FE115F 7D8F90AD
- r = 6901746346790563787434755862277025555839812737345013555379383634485463
  (number of elements of the subgroup)
- $h = 2$ (cofactor)
- G_x= 0FA C9DFCBAC 8313BB21 39F1BB75 5FEF65BC 391F8B36 F8F8EB73 71FD558B
- G_y= 100 6A08A419 03350678 E58528BE BF8A0BEF F867A7CA 36716F7E 01F81052
  (base point)

3 Co-processor design

The processor design depends on the algorithm used for computing the scalar-point operation. Montgomery-Ladder algorithm [14], presents some advantages for hardware (and software) implementations because it avoids inversion field operation in the main loop when projective or mixed coordinates are used. In this case, inversion only is required for coordinates conversion, and the recovering of the y coordinate (the algorithm operates only over the x coordinate) [14].

If binary representation of scalar k is $k = k_{m-1}2^{m-1} + ...k_2 2^2 + k_1 2 + k_0$, Algorithm 1 provides the scalar-point product kP in projective coordinates:

Algorithm 1 Montgomery ladder over projective coordinates

Require: k, P
Ensure: kP
1: $P_1 \leftarrow P$ *and* $P_2 \leftarrow 2P$
2: **for** $i = m - 2$ **downto** 0 **do**
3: **if** $k_i = 0$ **then**
4: $P_1 \leftarrow 2P_1$ *and* $P_2 \leftarrow P_1 + P_2$
5: **else**
6: $P_1 \leftarrow P_1 + P_2$ *and* $P_2 \leftarrow 2P_1$
7: **end if**
8: **end for**
9: **return** P_1

 ISBN: 978-84-617-8694-7

Operations required in *Algorithm 1* are point addition and doubling. When using projective coordinates [14], these operations results in field multiplications, additions and squarings. Field addition and squarings are combinational operations, and inversion is only required in coordinates conversion, resulting an execution time of [22]:

$$T \approx 6 \times m \times T_{mul} + 3 \times T_{inv} + 2 \times T_{mul} \tag{6}$$

Therefore, the execution time mainly depends of T_{mul}, the time required for completing the field-multiplication operation. In the next subsection, implementation of field-multiplier is analyzed.

3.1 Field multiplication

As pointed out in equation (6), the design of the multiplier determines the total execution time of the ECC scalar-point operation. There are several options for field-multiplier implementation, that can be grouped into two main sets:

- **Combinational multipliers**. These are multipliers based on Karatsuba-Ofman algorithm [23], and allow completing the field multiplication in only one clock cycle. As a drawback, these multipliers have high area requirements.
- **Sequential multipliers**. These are based on Digit Serial (DS) operations [22], with lower area needs, but requiring more number of clock cycles for completing the operation.

Performance of DS multipliers is based on operating at high clock frequencies, but this generates high power consumption. In the other hand, area requirements of Karatsuba-Ofman (KOA) multipliers are immoderate. In this work, we propose a mixed solution by means of introducing registers into an improvement of the KOA multiplier, named NOKOA (Non-Overlapping KOA) multiplier [24]. KOA and NOKOA multipliers are defined recursively: at step j the multiplier of 2^j inputs is built using three 2^{j-1} inputs multipliers. This structure allows saving area with respect to classical school multipliers, but generate higher delays. In order to improve delay, in the first stages (low values of j), classical school multipliers can be used, resulting the known as hybrid-KOA [25] (and hybrid-NOKOA) multipliers. It is possible to improve area figures of NOKOA multipliers reusing one multiplier at the last stage by means of registering the partial results. In this way, NOKOA3C multiplier is obtained allowing area savings around 50% with respect to NOKOA, at expenses of requiring 3 clock cycles for completing multiplication. If this process is repeated in the second stage of recursion, the NOKOA9C multiplier is obtained, requiring 9 clock cycles, with a 75% of area reduction.

Table 1 shows the implementation results for DS (Digit-Serial) multipliers with 1-bit, 8-bit, 24-bit and 39-bit digits from [22, 26] and NOKOA, NOKOA3C, and NOKOA9C

 ISBN: 978-84-617-8694-7

multipliers over the binary field $GF(2^{233})$ on a Virtex-5 device (xc5vlx110-3ff1760). This device has been selected for comparison purposes with DS multipliers proposed in [26], and all implementations were performed by using ISE 14.2.

Results shown in Table 1 have been obtained at four different operating frequencies: 25MHz, 50MHz, 100MHz, and the maximum operating frequency supported by the design. As can be observed, multipliers based on NOKOA algorithm are, in general, more efficient than DS multipliers. In fact, if we compare similar-size multipliers like DS_g24 and NOKOA9C (although NOKOA9C has around 30% less area than DS_g24), NOKOA9C achieves the final result in less time than DS_g24. Only when operating at the maximum operation frequency, DS_g24 presents better results (27ns) than NOKOA9C (42ns). However, if we take into account that dynamic power consumption is proportional to the clock frequency, DS_g24 requires more energy for completing the field multiplication than NOKOA9C.

In the case of NOKOA3C and DS_g39, (they have similar area), NOKOA3C always presents better results. Regarding NOKOA, it presents the best results at expenses on more area requirements. Therefore, the best trade-off taking into account area requirements, performance and total time for completing the field operation, is shown by NOKOA3C.

3.2 Field inversion/division

Field inversion is the most costly operation in binary finite fields, and the different proposals for their implementation are based on two mathematical results

1. Extended Euclides Algorithm (EEA). There exist efficient implementations like [27, 28], allowing the inversion in m clock cycles, or digit-serial implementations like [22], reducing the number of clock cycles at expenses of a higher area requirements.

2. Little Fermat Theorem (LTF). The Little Fermat Theorem establishes that the multiplicative inverse in a finite field can be obtained from:

$$p^{-1} = p^{2^m-2} = (p^{2^{m-1}-1})^2 \tag{7}$$

   and the standard [10] proposes an algorithm applying successive squarings, completing the inversion in m clock cycles. Another possibility is the use of the Itoh-Tsujii Algorithm (ITA) [29], optimizing the number of steps for the exponentiation calculus. In [30], ITA implementations minimizing the number of clock cycles are presented.

As Montogomery ladder algorithm requires inversion only when converting coordinates (see equation 6), we will use DS division with 1-bit digit size [22] for saving area resources.

 ISBN: 978-84-617-8694-7

Table 1: Implementation results for different polynomial multipliers over $GF(2^{233})$ on Virtex 5 devices

Design	# LUTS	#Max.freq (MHz)	#cycles	Freq. (MHz)	Total time
DS_g1 [26]	714	561	233	25	9.32 us
DS_g8 [26]	1413	448	30	25	1.20 us
DS_g24 [26]	3291	366	10	25	0.40 us
DS_g39 [26]	5167	340	6	25	0.24 us
NOKOA	11451	107	1	25	0.04 us
NOKOA3C	4891	214	3	25	0.12 us
NOKOA9C	2366	214	9	25	0.36 us
DS_g1 [26]	714	561	233	50	4.66 us
DS_g8 [26]	1413	448	30	50	0.60 us
DS_g24 [26]	3291	366	10	50	0.20 us
DS_g39 [26]	5167	340	6	50	0.12 us
NOKOA	11451	107	1	50	0.04 us
NOKOA3C	4891	214	3	50	0.06 us
NOKOA9C	2366	214	9	50	0.18 us
DS_g1 [26]	714	561	233	100	2.33 us
DS_g8 [26]	1413	448	30	100	0.30 us
DS_g24 [26]	3291	366	10	100	0.10 us
DS_g39 [26]	5167	340	6	100	0.06 us
NOKOA	11451	107	1	100	0.02 us
NOKOA3C	4891	214	3	100	0.03 us
NOKOA9C	2366	214	9	100	0.09 us
DS_g1 [26]	714	561	233	561	415 ns
DS_g8 [26]	1413	448	30	448	67 ns
DS_g24 [26]	3291	366	10	366	27 ns
DS_g39 [26]	5167	340	6	340	18 ns
NOKOA	11451	107	1	107	9 ns
NOKOA3C	4891	214	3	214	14 ns
NOKOA9C	2366	214	9	214	42 ns

 ISBN: 978-84-617-8694-7

3.3 Design of the ECC scalar-point Unit

Using the blocks described in previous subsections for field operations, Figure 1 shows the block diagram corresponding to the proposed ECC scalar-point unit.

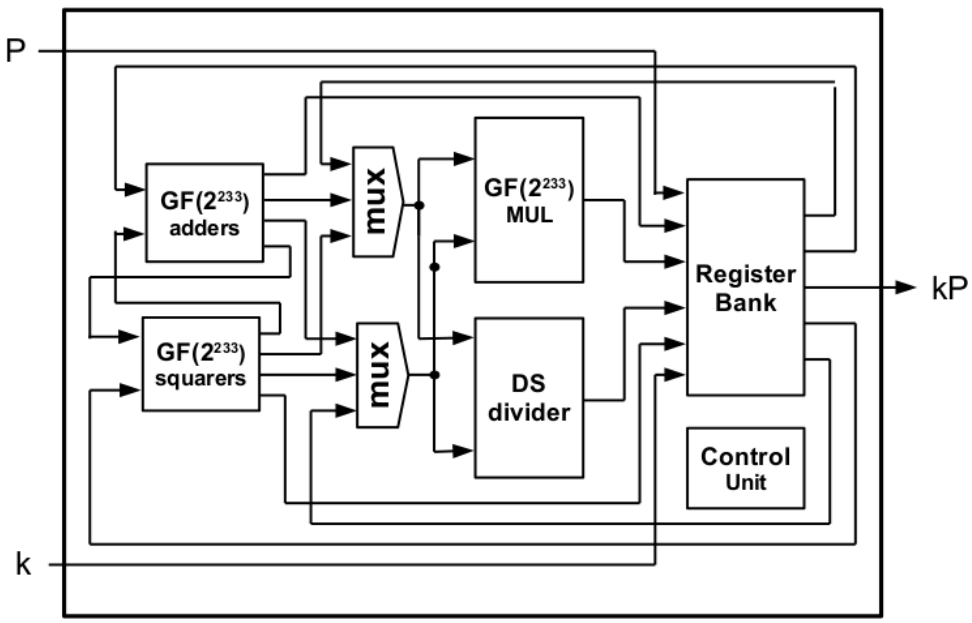


Figure 1: ECC Scalar-point Unit for B-233 curve

Table 2 shows implementation results corresponding to the three proposed co-processors, called ECC_B-233_NOKOA, ECC_B-233_NOKOA3C, and ECC_B-233_NOKOA9C. They have been implemented using Vivado 2016.4 over a Zynq [5] ZC102 device (Xc7z020clg484-1), with a target operating frequency of 50MHz for a contained power consumption.

The target device has around 53200 LUTs, thus it is possible to implement several

Table 2: ECC_B-233 implementation results using NOKOA, NOKOA3C and NOKOA9C multipliers

Design	# LUTS	#cycles	Freq. (MHz)	Total time
ECC_B-233_NOKOA	16320	2813	50	56.26 us
ECC_B-233_NOKOA3C	4891	5612	50	112.24 us
ECC_B-233_NOKOA9C	2366	14013	50	280.26 us

 ISBN: 978-84-617-8694-7

co-processors in the same device and take advantage or parallelism to achieve higher performance, if required. The selection of the co-processor to use will depend on the performance needed, the area available for implementing the co-processor (it could be necessary to include other hardware accelerators), and power consumption restrictions. Also, higher operating frequencies can be used for improving performance at expenses of increasing power consumption.

4 Conclusion

An ECC hardware-implemented co-processor has been presented. The co-processor has been implemented over a Zynq device, which includes an ARM Cortex A9 micro-processor, together with 53200 LUTs for hardware implementations. The designed co-processor, which shows contained area requirements, fits into the device, thus allowing to implement a complete high-performance cryptographic server together with web/database services into an unique chip.

References

[1] M. Stevens, E. Bursztein, P. Karpman, A., Albertini, and Y. Markov, *The first collision for full SHA-1*, URL: https://shattered.it/static/shattered.pdf, (2017).

[2] H. Marzouqi, M. Al-Qutayri, K. Salah, D. Schinianakis, and T. Stouraitis *A high-speed FPGA implementation of an RSD-based ECC processor*, IEEE Transactions on Very Large Scale Integration (VLSI) Systems, **24(1)**, (2016), 151–164.

[3] D.B. Roy, S. Agrawal, C. Rebeiro, and D. Mukhopadhyay, *Accelerating OpenSSL's ECC with low cost reconfigurable hardware*. In Integrated Circuits (ISIC), IEEE 2016 International Symposium on, (2016, December), 1–4.

[4] A.U. Ay, E. Ozturk, F.R. Henriquez, and E. Savas, *Design and implementation of a constant-time FPGA accelerator for fast elliptic curve cryptography*, In IEEE ReConFigurable Computing and FPGAs (ReConFig), 2016 International Conference on, (2016, November), 1–8.

[5] L.H. Crockett, R.A. Elliot, M.A. Enderwitz, and R.W. Stewart, *The Zynq Book: Embedded Processing with the Arm Cortex-A9 on the Xilinx Zynq-7000 All Programmable Soc* Strathclyde Academic Media. (2014).

[6] D. Lewis, G. Chiu, J. Chromczak, D. Galloway, B. Gamsa, V. Manohararajah, and J. Van Dyken, *The stratix 10 highly pipelined fpga architecture*, In Proceed-

ings of the 2016 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays, (2016, February), 159–168).

[7] D. Wagner, and B. Schneier *Analysis of the SSL 3.0 protocol*, In The Second USENIX Workshop on Electronic Commerce Proceedings **(Vol. 1, No. 1)**, (1996, November), 29–40.

[8] T. Dierks and E. Rescorla, *The Transport Layer Security (TLS) Protocol Version 1.2*, RFC 5246 (Proposed Standard), Internet Engineering Task Force, 2008

[9] T. Eisenbarth, and S. Kumar *A survey of lightweight-cryptography implementations*, IEEE Design & Test of Computers, **24(6)**, (2007).

[10] IEEE, *IEEE Standard Specifications for Public-Key Cryptography*, IEEE Std 1363-2000, (2000).

[11] IEEE, *IEEE Standard Specifications for Public-Key Cryptography - Amendment 1: Additional Techniques*, IEEE Std 1363a-2004, (2004).

[12] S.A Vanstone, *Next generation security for wireless: elliptic curve cryptography*, Computers & Security, **22(5)**, (2003), 412–415.

[13] K. Lauter, *The Advantages of Elliptic Curve Cryptography for Wireless Security*, IEEE Wireless Communications, **11(1)**, (2004), 62–67.

[14] H. Cohen, et al. (ed.), *Handbook of elliptic and hyperelliptic curve cryptography*, Chapman & Hall/CRC, (2005).

[15] N. Koblitz *Elliptic curve cryptosystems*, Mathematics of computation, **48(177)**, (1987), 203–209.

[16] V.S. Miller *Use of elliptic curves in cryptography*, In Advances in Cryptology – CRYPTO'85 Proceedings Springer Berlin Heidelberg, (1986), 417–426

[17] B. Moeller, T. Duong, and K. Kotowicz, *This POODLE bites: exploiting the SSL 3.0 fallback*, Security Advisory (Google), (2014).

[18] S. Blake-Wilson, B. Moeller, V. Gupta, C. Hawk, and N. Bolyard Wheeler, *Elliptic curve cryptography (ECC) cipher suites for transport layer security (TLS)*, RFC 4492, 2006.

[19] L. S. Huang, S. Adhikarla, D. Boneh, and C. Jackson, *An experimental study of TLS forward secrecy deployments*, IEEE Internet Computing **18(6)** (2014) 43–51.

 ISBN: 978-84-617-8694-7

[20] D. Johnson, A. Menezes, and S. Vanstone *The elliptic curve digital signature algorithm (ECDSA)*, International journal of information security, **1(1)**, (2001) 36–63.

[21] NIST *FIPS PUB 186-4, Digital signature standard (DSS)*, (2013).

[22] G. Sutter, J. Deschamps, and J. Imaña, *Efficient Elliptic Curve Point Multiplication using Digit Serial Binary Field Operations*, IEEE Transactions on Industrial Electronics, **60(1)**, (2013).

[23] A. Karatsuba, *The complexity of computations*, Proc. Steklov Inst. Math, **(211)**, (1995), 169–183.

[24] H. Fan, J. Sun, M. Gu, and K.Y. Lam, *Overlap-free Karatsuba-Ofman polynomial multiplication algorithms*, Information Security, IET, **4(1)**, (2010), pp.8–14.

[25] F. Rodríguez-Henríquez and C.K. Koc, *On Fully Parallel Karatsuba Multipliers for $GF(2^m)$*, Proc. Int. Conf. Computer Sci. and Tech. (CST 2003), (2003), 405–410.

[26] G. Sutter, J. Deschamps, and J. Imaña, *Efficient Elliptic Curve Point Multiplication using Digit Serial Binary Field Operations (Additional Material for published paper*, URL: http://www.arithmetic-circuits.org/ECC_PointMult/EllipticCurvePointMultiplication.html, (2013).

[27] H. Brunner, A. Curiger and M. Hofstetter, *On Computing Multiplicative Inverses in $GF(2^m)$*, IEEE Trans. on Comp., **42 (8)**, (1993), 1010–1015.

[28] Z. Yan, D.V. Sarwate, *New Systolic Architectures for Inversion and Division in $GF(2^m)$*, IEEE Trans. on Computers, **52(11)**, (2003), 1514–1519.

[29] T. Itoh, S. Tsujii, *A fast algorithm for computing multiplicative inverses in $GF(2^m)$ using normal bases* ", Inf. Comput., **78(3)**, (1988), 171–177.

[30] L. Parrilla, A. Lloris, E. Castillo and A. Garcia, *Minimum-clock-cycle Itoh-Tsujii algorithm hardware implementation for cryptography applications over $GF(2^m)$ fields*, Electronics Letters, **48(18)**, (2012), 1126–1128.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

Introductory elements for the development of a multiplicative statistic

C. Pavez-Rojas [1], F. Córdova-Lepe[2] and K. Vilches-Ponce[1]

[1] *Facultad de Ciencias Básicas, Universidad Católica del Maule, Talca, Chile*

[2] *Depto. de Matemática, Univ. Metropolitana de Cs. de la Educación, Santiago, Chile*

emails: carolpavez87@gmail.com, fcordova@ucm.cl, kvilches@ucm.cl

Abstract

Indicators such as the arithmetic mean and the variance have a construction based on the additive character of the real numbers field. Here we explore the analogues of some measures of central tendency and dispersion of data, but in the context of the real positive field, the base framework of the multiplicative (bigeometric) calculus.

Key words: statistics, geometric mean, bigeometric calculus
MSC 2000: 92B05

1 Introduction

In statistics the most known central tendency for a set of numerical data is the arithmetic mean, as we know, is constructed by adding the products of these data by certain weights. Therefore, such data is in a structure in which addition and multiplication operations exist and also have certain minimum properties (eg, an ordered and complete field), so a minimal context is necessary. In this work, we will use the fact that the real numbers set has, in its interior, an isomorphic image of itselves to derive a series of statistical indicators. Mainly, to asociate those with aritmethic origin with the geometric ones.

Let us begin by pointing out that: **the arithmetic mean and the geometric mean are isomorphic measures of central tendency**. That is to say, these can be thought like ways to measure the same, except that in equivalent contexts.

Indeed, the exponential function $\exp :]-\infty, +\infty[\to]0, +\infty[$ establishes an isomorphism between the complete ordered field of the real numbers $(\Re, +, \cdot)$ and its image the positive

 ISBN: 978-84-617-8694-7

reales numbers $(\Re^+, \cdot, *)$, where $\Re^+ =]0, \infty[$ and $*$ defines the operation $a * b = a^{\ln(b)}$, for any a, $b \in \Re^+$. Regarding the implications (theoretical and application) associated with the construction of the proportional calculation from the positive real field, see [1, 2, 3]. The main intention of the present work is to introduce some basic elements that allow to pass from an additive vision of the statistic to one of multiplicative type or, that we could also call, of proportional type.

This small work is organized as follows: In section 2 the arithmetic and geometric mean is linked. In section 3, the geometric version of a measure of data dispersion is constructed from its arithmetic analogue. Finally, section 4 formalizes and generalizes a series of types of averages (isomorphic to each other) within the real field.

2 Geometric mean

Notice that, given $a, b \in \Re^+$, obviously we have

$$a \cdot b = \exp\{\ln(a) + \ln(b)\},$$

and then

$$\sqrt{a \cdot b} = \exp\left\{\frac{\ln(a) + \ln(b)}{2}\right\}$$

In other words, the isomorphism $\exp : (\Re, +, \cdot) \to (\Re^+, \cdot, +)$ translates the mean (*arithmetic*) M_A with respect to the first operation (the addition) of the framework $(\Re, +, \cdot)$, in the mean (*geometric*) M_G with respect to the first operation (multiplication) of the framework $(\Re^+, \cdot, *)$, by the identity:

$$M_G(a, b) = \exp\{M_A(\ln(a), \ln(b))\}.$$

In a reverse reading, the function $\ln :]0, +\infty[\to]-\infty, +\infty[$, transposes the geometric mean of $(\Re^+, \cdot, *)$, in the arithmetic mean of $(\Re, +, \cdot)$, by the formula:

$$M_A(a, b) = \frac{a + b}{2} = \ln\{\sqrt{\exp(a) \cdot \exp(b)}\} = \ln\{M_G(e^a, e^b)\}.$$

In spite of the equivalence of these measures, as the isomorphism is inward, that is, towards the field of real positive, a part of the totality of the real numbers, there is also a gain. This allows us to see the arithmetic mean and the geometric mean as measures based on the first and second operation del campo respectively.

Moreover, since the weighted arithmetic average for a set of $\{x_1, \cdots, x_n\}$ of real numbers and $\{\alpha_1, \cdots, \alpha_n\}$, $0 \leq \alpha_i \leq 1$, $\alpha_1 + \cdots + \alpha_n = 1$, as weights is given by

$$M_A^*(\{x_i\}, \{\alpha_i\}) = \alpha_1 x_1 + \cdots + \alpha_n x_n,$$

ISBN: 978-84-617-8694-7

using isomorphism the corresponding weighted geometric mean for a set of positive real number $\{y_1, \cdots, y_n\}$ is given by:

$$M_G^*(\{y_i\}, \{\beta_i\}) = \exp(M_A^*(\{\ln(y_i)\}, \{\ln(\beta_i)\})),$$

Expression that is equal to $\exp\{\ln(\beta_1)\ln(y_1) + \cdots + \ln(\beta_n)\ln(y_n)\}$ and which can also be written as

$$y_1^{\ln(\beta_1)} \cdots y_n^{\ln(\beta_n)}.$$

Finally, we have

$$M_G^*(\{y_i\}, \{\beta_i\}) = (\beta_1 * y_1) \cdots (\beta_n * y_n) = y_1^{\alpha_1} \cdots y_n^{\alpha_n},$$

where the set of $\{\beta_i\}$, $\alpha_i = \beta_i$, are subject to the following conditions: $1 \leq \beta_i \leq e$ and $\beta_1 \cdots \beta_n = e$.

3 Geometric dispersion

Recall that if $\{x_1, \cdots, x_n\}$ is a certain set of data, then the average arithmetic deviation that the data moves away from a value y is given by the expression:

$$V_A(y, \{x_i\}) = \frac{1}{n}\sum_{i=1}^{n} |y - x_i|.$$

The absolute value function $|\cdot| : \Re \to \Re^+$ allows us to convert the field $(\Re, +, \cdot)$ in a metric space. Note that, the isomorphic field $(\Re^+, \cdot, *)$ is a *relative metric space* now performed by the so-called *relative value* defined by $[\,\cdot\,] : \Re^+ \to]1, \infty[$ with

$$[a] = \begin{cases} a & \text{if} \quad a \leq 1, \\ a^{-1} & \text{if} \quad a < 1. \end{cases}$$

The analogous *geometric deviation* $V_G(\cdot, \cdot)$ should be given by:

$$V_G(y, x_i) = \exp\{V_A(\ln(y), \{\ln(x_i)\})\}.$$

This is,

$$V_G(y, \{x_i\}) = \exp\left\{\frac{1}{n}\sum_{i=1}^{n} |\ln(y) - \ln(x_i)|\right\} = \exp\sum_{i=1}^{n}\left|\ln\{(y/x_i)^{1/n}\}\right|,$$

since $|\ln(u)| = \ln([u])$, we have

$$V_G(y, \{x_i\}) = \exp\left\{\sum_{i=1}^{n} \ln[(y/x_i)^{1/n}]\right\} = \left\{\prod_{i=1}^{n}\left[\frac{y}{x_i}\right]\right\}^{1/n}.$$

With no further effort we have reciprocally:

$$V_A(y, \{x_i\}) = \ln\{V_G(e^y, \{e^{x_i}\})\}.$$

 ISBN: 978-84-617-8694-7

4 Generalization and formalism

Inductively we introduce the sequence:

$$e_{k+1} = \exp\{e_k\}, \quad \text{with} \quad e_0 = 0.$$

Notice that, for each $k \geq 1$, e_k is the neutral element of the operation $\oplus_k$ defined on the interval $I_k =]e_{k-1}, \infty[$ recursively by

$$a \oplus_{k+1} b = \exp\{\ln(a) \oplus_k \ln(b)\},$$

where $\oplus_0$ is the usual sum.

With this operation the family of structures $\{(I_k, \oplus_k, \oplus_{k+1})\}_{k\geq 1}$ is a sequence of ordered and complete fields, therefore, all isomorphic to each other, *i.e.* isomorphic to $(\Re, +, \cdot)$, by means of the exponential function.

The interesting thing is that, as I_k is a subset of $\Re$, $k \geq 0$, with $e_{-1} = -\infty$, all the concepts derived by isomorphism in any of these fields have a reading in the basal field $(\Re, +, \cdot)$, which also, from now on we can denote by $(I_0, \oplus_0, \oplus_1)$.

Definition: The *k-th average*, $k \geq 0$, among the numbers $x_1, \cdots, x_n$ of I_k, is defined by:

$$M_k(x_i) = (x_1 \oplus_k \cdots \oplus_k x_n) \ominus_{k+1} (e_{k+1} \oplus_k \cdots \oplus_k e_{k+1}), \tag{1}$$

where $\ominus_j$ is the inverse operation of $\oplus_j$, which is also defined recursively by means of $a \ominus_{k+1} b$ equals to $\exp\{\ln(a) \ominus_k \ln(b)\}$, $\ominus_0$ the usual subtraction.

Remark: Some examples to illustrate, using only two elements, are :

- $M_0(a, b) = (a \oplus_0 b) \ominus_1 (e_1 \oplus_0 e_1) = (a + b)/(1 + 1) = M_A(a, b)$.
- $M_1(a, b) = (a \oplus_1 b) \ominus_2 (e_2 \oplus_1 e_2) = (a \cdot b) \ominus_2 (e \cdot e)$. Then, continuing with the calculations $M_1(a, b) = \exp\{\ln(ab) \ominus_1 \ln(e^2)\}$, which is equal to $\exp\{\ln(ab)/2\}$, this is, a $(ab)^{1/2}$, *i.e.*, $M_G(a, b)$.

Theorem: *Given a set* $\{x_i\} \subset I_{k+1}$, *some* $k \geq 0$, *then*

- $M_{k+1}(\{x_i\})) = \exp(M_k(\{\ln(x)\}))$, *and*
- $M_{k+1}(\{x_i\})) \leq M_k(\{x_i\})$.

ISBN: 978-84-617-8694-7

References

[1] Córdova-Lepe, F., & Pinto, M. *From quotient operation toward a proportional calculus*. International Journal of Mathematics, Game Theory and Algebra, **18(6)**, (2009) 527-536.

[2] Córdova-Lepe, F., & Pinto, M. *From quotient operation toward a proportional calculus*. In Mathematics Research Perspectives. Editors: Michael C. Leung (2012) 253-266.

[3] Córdova-Lepe, F. *The multiplicative derivative as a measure of elasticity in economics*. TMAT Revista Latinoamericana de Ciencias e Ingeniera,**2(3)**, (2006).

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

Stabilization of switched systems with state-dependent switching noise

C. Pérez[1], F. Benítez[1] and J. B. García-Gutiérrez[1]

[1] *Department of Mathematics, University of Cádiz*

emails: carmen.perez@uca.es, quico.benitez@uca.es,
juanbosco.garciagutierrez@alum.uca.es

Abstract

This paper studies the stabilization of second-order switched linear systems with state-dependent switching noise. Based on the method of stabilization known for this kind of switched systems without noise, we define some regions in the plane where the noise can be produced and the stabilization is also assured. Moreover, we give a procedure for determining completely these regions. Finally, in order to illustrate the results, several numerical examples are presented.

Key words: switched systems, switching noise, robust stabilization, time-delay, Lyapunov function

1 Introduction

A switched system is a kind of hybrid system that consists of several subsystems and a switching law determining at any time instant which subsystem is active. In recent years, the study of switched systems has received growing attention in Control Theory and its applications ([8], [9], [14]). They have been studied from several points of view such as the controllability ([17]), reachability ([14], [18], or stability ([15], [10]).

The problem of stability has been studied from several viewpoints ([8], [10]). One of these viewpoints is the problem of constructing a switching law that makes the switched system asymptotically stable. It is important to note that in this problem the only control action is the switching law. In the literature, there are several two approaches to stabilization for switched systems. In papers such as [7], [16] or [4] Lyapunov theory is employed to construct switching laws for a class of switched systems totally composed of unstable subsystems.

 ISBN: 978-84-617-8694-7

Another approach to this problem is based on detailed analysis of the vector field. For second-order systems, some necessary and sufficient conditions for stability/stabilizability have been obtained in [12] and [19] through analysis on the structure of the vector field. In [19] the problem of stabilizing two second-order linear systems is solved; i.e., a method to define switching laws that decide when a switched system is stabilizable is given. Also, recently, Cong [3] characterises the stability of these switched linear systems by classifying the switching laws into only two types.

A common assumption in the above results is that the detection of the switching signal is instant. However, in many real switched systems, the switchings cannot be instant, i.e., the changing of the switching law cannot be detected instantly, but only after a time period. For this reason, it is important to study the situation when switching noise is produced.

Based on the method of stabilization in [19] where the switchings are produced in rays, we define some regions in the plane where the switchings can be produced and the stabilization is also assured. With these results, we study the problem of stabilization with state-dependent switching noise. Related to this problem, in [5] and [6] the authors present definitions and results of these stochastic systems.

The remainder of this paper is organized as follows. In Section 2 some definitions and some results are included. In Section 3 we present the results that assure the regions of stability. In Section 4 we present the main result of this work. Several numerical examples are given in Section 4 in order to illustrate the results. Finally, the conclusions are provided in Section 5.

2 Preliminaries

Consider a switched linear system

$$\dot{x}(t) = A_{\sigma(t)}x(t), \tag{1}$$

where A_i is a 2×2 matrix, for $i = 1, 2$, $x \in \mathbb{R}^2$, and $\sigma : [0, \infty) \longrightarrow \{1, 2\}$ is a piecewise constant function called *switching law*. Such a function σ has a finite number of discontinuities, which we call *switching times*, on every bounded time interval and takes a constant value on every interval between two consecutive switching times. The role of σ is to indicate the active subsystem at each time instant t.

Before presenting the method of stabilization we need this definition:

Definition 1 *Let r_1 and r_2 be two rays starting from the origin. The set given by*

$$\{x \in \mathbb{R}^2 : x = \mu z_1 + (1 - \mu)z_2,\ z_1 \in r_1,\ z_2 \in r_2,\ 0 \leq \mu \leq 1\}$$

will be called the cone delimited by r_1 and r_2 and denoted by $\mathrm{C}(0, r_1, r_2)$.

 ISBN: 978-84-617-8694-7

In [19] the stabilizing switching law is given in the following form:

$$\sigma(t) = \delta(x(t)) \tag{2}$$

where $\delta : \mathbb{R}^2 \to \{1, 2\}$. Moreover, the switchings are produced in two rays r_1 and r_2, i.e., when the trajectory intersects these rays, we switch. These rays are always given by the equation $\det(A_1x, A_2x) = 0$.

By the results in [19] and [3], in [13] only two kinds of switching laws are used to study the stabilization: type I or type II. Under switching laws of type I, the solution rotates around the origin clockwise or counterclockwise direction (spinning switching in [3]). On the contrary, under switching laws of type II, the trajectories of each subsystem are of opposite directions, thus, the solution remains in a cone (chattering switching in [3]).

The switching laws of type I are defined as follows:

- $\sigma(t) = 1$ (resp. $\sigma(t) = 2$) when $x(t) \in \{x \in \mathbb{R}^2 : \det(A_1x, A_2x) > 0\}$ and each subsystem at $x(t)$ is of the clockwise (resp. counterclockwise) direction.

- $\sigma(t) = 2$ (resp. $\sigma(t) = 1$) when $x(t) \in \{x \in \mathbb{R}^2 : \det(A_1x, A_2x) < 0\}$ and each subsystem at $x(t)$ is of the clockwise (resp. counterclockwise) direction.

For switching laws of type II, we need to suppose that there exists a cone where the subsystems are of opposite directions and the set $\{x \in \mathbb{R}^2 : \det(A_1x, A_2x) > 0\}$ or $\{x \in \mathbb{R}^2 : \det(A_1x, A_2x) < 0\}$ is contained in this cone. Hence, if S_1 is a cone where A_1 is of clockwise (resp. counterclockwise) direction, A_2 is of counterclockwise (resp. clockwise) direction and $\{x \in \mathbb{R}^2 : \det(A_1x, A_2x) > 0\}$ (resp. $\{x \in \mathbb{R}^2 : \det(A_1x, A_2x) < 0\}$) is contained in this set, the switchings under a switching law of type II are produced when the solution intersects the rays $\det(A_1x, A_2x) = 0$.

In both cases (type I or II), the switchings are produced in rays. The problem of this is that the stabilization is not assured if the switchings are not produced in these rays. In [13], the problem of time-delay in detection of switching law is studied. From this study it is deduced the following result:

Theorem 1 *[13] Let A_1 and A_2 be two unstable 2×2 matrices. Suppose that for the switched system (1) the switching laws of type I (resp. II) are stabilizing. Then there exists $T_0 \in (0, \overline{T})$ such that the switched system under the switching law of type I (resp. II) and with time-delay equal to T is*

- *stable if $T \in [0, T_0)$.*

- *non stable if $T \in (T_0, \overline{T}]$.*

 ISBN: 978-84-617-8694-7

From the proof of this result, it is deduced that for time-delay T_0, a periodic solution is obtained. Moreover, we obtain two rays, denoted by s_1 and s_2, where the switchings for this time-delay are produced.

Under this notation, we prove in the following section that when the switchings from A_1 to A_2 are produced in $\mathrm{C}(0, r_1, s_1)$ and the switchings from A_2 to A_1 are produced in $\mathrm{C}(0, r_2, s_2)$, the switched system remains stable.

In order to prove this, we need to introduce a concept studied in [11]. This concept will be employed in the next section.

Definition 2 *[11] A function $H_A : \mathbb{R}^2 \to \mathbb{R}^+$ is a generalized first integral of $\dot{x} = Ax$ if $H_A(x)$ is not constant on any open subset of $\mathbb{R}^2$, $H_A(x(t))$ is piecewise constant along the trajectories of $\dot{x} = Ax$, and attains a finite set of values for all $t \in [0, T]$ (T finite).*

For a system $\dot{x} = Ax$ with A a 2×2 real matrix, a generalized first integral is constructed in [11] by considering two cases. In the first, the eigenvalue of A are complex and in the second, they are real.

The relation between the system $\dot{x} = Ax$ and its generalized first integral is the following $H(x(t)) = H(x(0))$ along the trajectories of $\dot{x} = Ax$, i.e., $x(t)$ coincides with the contour $H_A(x(t)) = H_A(x_0)$ if x is the solution of $\dot{x} = Ax$ with $x(0) = x_0$ (see [11] for more details). Therefore, when H is differentiable, $\frac{dH(x(t))}{dt} = \nabla H(x(t))Ax(t) = 0$.

3 Existence of the regions of stabilization

In this section we present the results under which the existence of the regions of stabilization is assured.

Theorem 2 *Let A_1 and A_2 be two unstable 2×2 matrices. Suppose that for the switched system (1) the switching laws of type I (resp. II) are stabilizing. Then there exists two cones C_1 and C_2 such that the switched system is stable under a switching law σ if we switch from A_1 to A_2 when $x(t) \in \mathrm{C}_1$ and we switch from A_2 to A_1 when $x(t) \in \mathrm{C}_2$.*

PROOF. Firstly, we suppose that the switching law of type I is stabilizing and the solution under this switching law rotates around the origin in clockwise direction. Under the nota-tion in the previous section, we define $\mathrm{C}_1 = \mathrm{C}(0, r_1, s_1)$ and $\mathrm{C}_2 = \mathrm{C}(0, r_2, s_2)$ (see Figure 1).

Now, if σ is a switching law under which the switchings from A_1 to A_2 are produced in C_1 and the switchings from A_2 to A_1 are produced in C_2, we have to prove that the switched system for σ is stable. In order to do that, we define a Lyapunov function for this system.

 ISBN: 978-84-617-8694-7

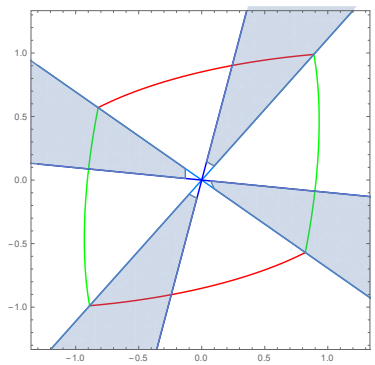


Figure 1: The cones $\mathrm{C}_1 = \mathrm{C}(0, r_1, s_1)$ and $\mathrm{C}_2 = \mathrm{C}(0, r_2, s_2)$ and the periodic solution for T_0.

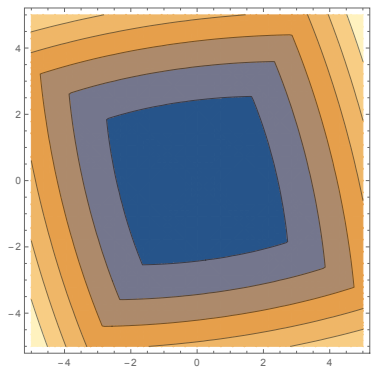


Figure 2: The level curves of V defined for a system with switching law of type I.

The Lyapunov function is defined as follows:

$$V(x) = \begin{cases} H_1(x) & \text{if} \quad x^T P x \leq 0 \\ kH_2(x) & \text{if} \quad x^T P x < 0 \end{cases}$$

where H_i is the first integral of A_i for $i = 1, 2$, P is the matrix that defines the quadratic form given by s_1 and s_2, and k is a positive constant defined in order to V be continuous (see Figure 2 V tipoI).

It is easy to prove that $V(x) \geq 0$ for any $x \in \mathbb{R}^2$ and, from definition, V is continuous. However, V is non differentiable at x with $x^T P x = 0$. For this reason, in order to prove the stability, we use the upper right Dini derivative for V, that is,

$$D^+ V(x(t)) = \limsup_{\tau \to 0^+} \frac{V(x(t+\tau)) - V(x(t))}{\tau}$$

 ISBN: 978-84-617-8694-7

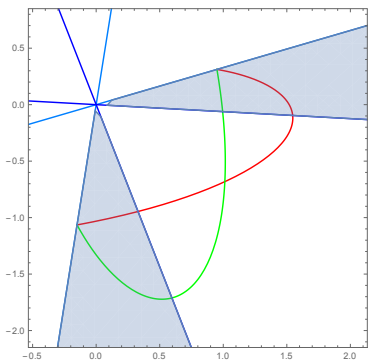


Figure 3: The cones $\mathrm{C}_2 = \mathrm{C}(0, s_1, r_1)$ and $\mathrm{C}_1 = \mathrm{C}(0, r_2, s_2)$ and the periodic solution for T_0.

It is proved (see [1] or [2] for more details) that if x is the solution of (1) under the switching law σ and $D^+V(x(t)) \leq 0$ for each $t \geq 0$, the system is stable.

Therefore, we only have to prove that $D^+V(x(t)) \leq 0$ for each $t \geq 0$. Let consider several cases for $x(t)$. If $x(t) \in \mathrm{C}(0, s_1, r_2) \cup r_2$, then, by the definition of σ, the solution of the system is given by A_1. Therefore, as $x(t)^T P x(t) > 0$, $D^+V(x(t)) = \nabla V(x(t))\dot{x}(t) = \nabla H_1(x(t))A_1 x(t) = 0$ by the definition of H_1. Analogously, it is proved that if $x(t) \in \mathrm{C}(0, s_2, r_1) \cup r_1$, then, $D^+V(x(t)) = 0$.

If $x(t) \in \mathrm{C}(0, r_2, s_2)$, then, by the definition of σ, the solution of the system can be given by A_1 or A_2. If the system is given by A_1, as $x(t)^T P x(t) > 0$, $D^+V(x(t)) = \nabla V(x(t))\dot{x}(t) = \nabla H_1(x(t))A_1 x(t) = 0$. If, on the contrary, the system is given by A_2, as $x(t)^T P x(t) > 0$, $D^+V(x(t)) = \nabla V(x(t))\dot{x}(t) = \nabla H_1(x(t))A_2 x(t)$. But, we know that, in this region, $\det(A_1 x(t), A_2 x(t)) > 0$, thus, $D^+V(x(t)) = \nabla H_1(x(t))A_2 x(t) < 0$. Now, if $x(t) \in \mathrm{C}(0, r_1, s_1)$, it can be proved that $D^+V(x(t)) < 0$.

If $x(t) \in s_2$, by the definition of σ, the solution of the system is given by A_2 and, $D^+V(x(t)) = \nabla H_2(x(t))A_2 x(t) = 0$. Similarly, if $x(t) \in s_1$, it is proved that $D^+V(x(t)) = \nabla H_1(x(t))A_1 x(t) = 0$. And the result is proved for switching laws of type I.

Now, we suppose that the switching law of type II is stabilizing. And we define $\mathrm{C}_2 = \mathrm{C}(0, s_1, r_1)$ and $\mathrm{C}_1 = \mathrm{C}(0, r_2, s_2)$ (see Figure 3). In this case, given a swiching law σ such that the switchings from A_1 to A_2 are produced in C_1 and the switchings from A_2 to A_1 are produced in C_2, we have to prove that (1) under this switching law is stable.

Again, we define a function V and we show that is a Lyapunov function for this system.

$$V(x) = \begin{cases} H_1(x) & \text{if} \quad x \in \mathrm{C}(0, r_1, r_2) \\ kH_2(x) & \text{if} \quad x \in \mathrm{C}(0, s_1, r_1) \cup r_1 \\ lH_2(x) & \text{if} \quad x \in \mathrm{C}(0, r_2, s_2) \cup r_2 \end{cases}$$

 ISBN: 978-84-617-8694-7

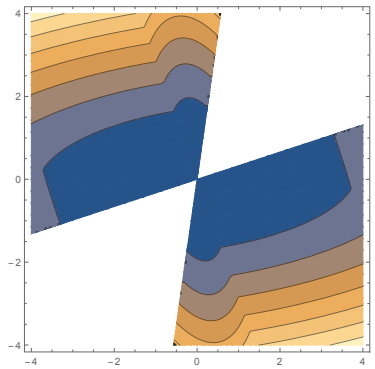


Figure 4: The level curves of V defined for a system with switching law of type II.

where H_i is the first integral of A_i for $i = 1, 2$ and k and l are positive constants defined in order to V be continuous.

It is important to note that V, in this case, is only defined in the region $\mathrm{C}(0, s_1, s_2)$ but, by definition of σ, the solution of the system is contained in this region. (see Figure 4).

Moreover, in this case, V is differentiable because at $x \in r_1$ or $x \in r_2$, A_1x and A_2x are parallel. Thus, we can consider the derivative of $V(x(t))$. If we prove that is non positive, we obtain that the system is stable.

If $x(t) \in \mathrm{C}(0, r_1, r_2) \cup r_1 \cup r_2$, then, by the definition of σ, the solution of the system can be given by A_1 or A_2. If the system is given by A_1, $\frac{dV(x(t))}{dt} = \nabla V(x(t))\dot{x}(t) = \nabla H_1(x(t))A_1x(t) = 0$. If, on the contrary, the system is given by A_2, $\frac{dV(x(t))}{dt} = \nabla V(x(t))\dot{x}(t) = \nabla H_1(x(t))A_2x(t)$. But, we know that, in this region, $\det(A_1x(t), A_2x(t)) > 0$, thus, $\frac{dV(x(t))}{dt} < 0$.

If $x(t) \in \mathrm{C}(0, s_1, r_1)$, then, by the definition of σ, the solution of the system can be given by A_1 or A_2. If the system is given by A_2, $\frac{dV(x(t))}{dt} = \nabla V(x(t))\dot{x}(t) = \nabla H_2(x(t))A_2x(t) = 0$. If, on the contrary, the system is given by A_1, $\frac{dV(x(t))}{dt} = \nabla V(x(t))\dot{x}(t) = \nabla H_2(x(t))A_1x(t)$. But, we know that, in this region, $\det(A_1x(t), A_2x(t)) > 0$, thus, $\frac{dV(x(t))}{dt} < 0$. Analogously, we can prove that if $x(t) \in \mathrm{C}(0, r_2, s_2)$, then $\frac{dV(x(t))}{dt} \leq 0$.

□

4 Stabilization with state-dependent switching noise

Now, we consider a stochastic switched linear system

$$\dot{x}(t) = A_{\sigma(x(t))}x(t), \tag{3}$$

 ISBN: 978-84-617-8694-7

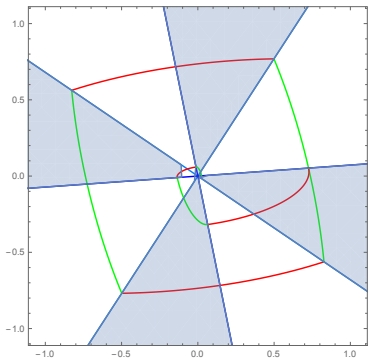


Figure 5: Stable solution and periodic solution for Example 1

where A_i is a 2×2 matrix, for $i = 1, 2$, $x \in \mathbb{R}^2$, and now $\{\sigma(x(t)); t \geq 0\}$ is a stochastic process taking values in $\{1, 2\}$. This stochastic process determines the switchings between the subsystems by the intensity function h.

By using the results in the previous section, we obtain the following result:

Theorem 3 *Let A_1 and A_2 be two unstable 2×2 matrices. Suppose that for the switched system (1) the switching laws of type I (resp. II) are stabilizing. Then there exists $\epsilon > 0$ and a intensity function $h : [-\epsilon, \epsilon] \to (0, \infty)$ such that the switched system (3) is stable with probability 1.*

5 Numerical examples

Now, in order to illustrate these results, three numerical examples are presented.

Example 1 *Consider a switched linear system given by (1) where the matrices are the following:*

$$A_1 = \begin{pmatrix} -2 & 52 \\ -8 & 6 \end{pmatrix} \qquad A_2 = \begin{pmatrix} 11 & 10 \\ -50 & -9 \end{pmatrix}$$

By applying the method in [19], we obtain that the switching law of type I is stabilizing and the solution rotates around the origin in clockwise direction (see Figure 5).

Moreover, applying Theorem 1, a periodic solution is obtained and two rays s_1 and s_2 (see Figure 5). Finally, by Theorem 3, we have that if we switch in the cones C_1 and C_2, the solution is stable .

 ISBN: 978-84-617-8694-7

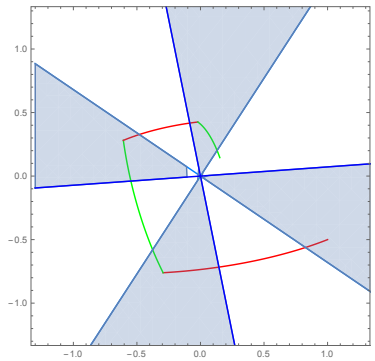


Figure 6: Stable solution for initial condition $x_0 = (1, -0.5)$ and cones where the switchings are produced for Example 1

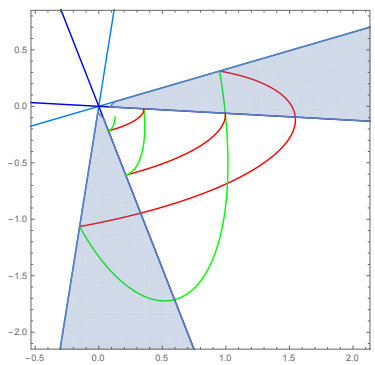


Figure 7: Stable solution and periodic solution for Example 2

Example 2 *Consider a switched linear system given by (1) where the matrices are the following:*

$$A_1 = \begin{pmatrix} 1 & 13 \\ -2 & 3 \end{pmatrix} \qquad A_2 = \begin{pmatrix} -1 & -2 \\ 10 & 3 \end{pmatrix}$$

By applying the method in [19], we obtain that the switching law of type II is stabilizing and the solution remains in a region where the trajectory of A_1 is of clockwise direction and the trajectory of A_2 is of counterclockwise direction (see Figure 7).

Moreover, applying Theorem 1, a periodic solution is obtained and two rays s_1 and s_2 (see Figure 7). Finally, by Theorem 3, we have that if we switch in the cones C_1 and C_2, the solution is stable (see Figure 8).

 ISBN: 978-84-617-8694-7

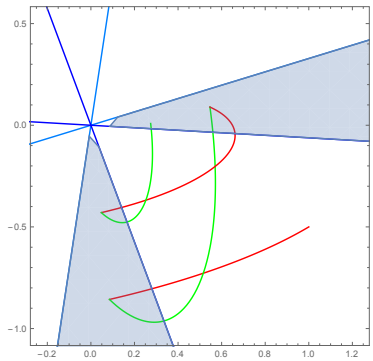


Figure 8: Stable solution for initial condition $x_0 = (1, -0.5)$ and cones where the switchings are produced.

6 Conclusions

In this work the stabilization of second-order switched linear system with state-dependent switching noise has been studied. Based on the method of stabilization of [19] and [3], we have presented a result that assures the existence of regions of stabilization where although the switchings are produced and the system is also stability. A method of finding these regions have been given.

By using these regions, we have presented a result that assures the stability when the switchings are given by a stochastic process.

As future work, it can be studied more regions where the switchings can be produced. The objective can be obtain necessary and sufficient conditions for stabilization.

References

[1] A. Bacciotti and L. Rosier, *Liapunov functions and stability in control theory*, Springer, 2005.

[2] F. Blanchini, *Nonquadratic Lyapunov functions for robust control*, Automatica **31** (1995) 451–461.

[3] S. Cong *Characterising the stabilisability for second-order linear switched systems*, Int. J. Control **86** (2013) 519-528.

[4] E. De Santis, M. D. Di Benedetto and G. Pola, *Stabilizability of linear switching systems*, Nonlinear Analysis: Hybrid Systems **2** (2008) 750–764.

 ISBN: 978-84-617-8694-7

[5] J. Leth, J. G. Rasmussen, H. Scholer, R. Wisniewski, *A class of stochastic hybrid systems with state-dependent switching noise*, 51st IEEE Conference on Decision and Control (2012)

[6] J. Leth, J. G. Rasmussen, H. Scholer, R. Wisniewski, *Parameter estimation for a class of stochastic hybrid systems with state-dependent switching noise*, European Control Conference (2015)

[7] Z. G. Li, C. Y. Wen and Y. C. Soh, *Stabilization of a class of switched systems via designing switching laws*, IEEE Transactions on Automatic Control **46** (2001) 665–670.

[8] D. Liberzon and A. S. Morse, *Basic problems in stability and design of switched systems*, IEEE Control Systems Magazine **19** (1999) 59–70.

[9] D. Liberzon, *Switching in Systems and Control*, Birkhauser, 2003.

[10] H. Lin and P. J. Antsaklis, *Stability and stabilizability of switched linear systems: a survey of recent results*, IEEE Transactions on Automatic Control **54** (2009) 308–322.

[11] M. Margaliot and G. Langholz *Necessary and sufficient conditions for absolute stability: the case of second-order systems*, IEEE Transactions on circuits and systems I **50** (2003) 227-234.

[12] C. Pérez and F. Benítez, *Switched Convergence of Second-Order Switched Nonlinear Systems*, Int. J. of Control, Autom., and Syst. **10** (2012).

[13] C. Pérez and F. Benítez, *Convergence of switched linear systems with time-delay in detection of switching law*, IET Control Theory and Applications **8** (2014) 647-654.

[14] Z. Sun and S. S. Ge, *Switched linear systems*, Springer, 2005.

[15] Z. Sun and S. S. Ge, *Stability theory of switched dynamical systems*, Springer, 2011.

[16] M. A. Wicks, P. Peleties and R. A. DeCarlo, *Construction of piecewise Lyapunov functions for stabilizing switched systems*, Proc. of the 33rd Conference on Decision and Control (1994) 3492–3497.

[17] G. Xie and L. Wang, *Controllability and stabilizability of switched linear systems*, Systems and Control Letters **48** (2003) 135–155.

[18] X. Xu and P. J. Antsaklis, *On the reachability of a class of second-order switched systems*, Proc. of the American Control Conference (1999).

[19] X. Xu and P. J. Antsaklis *Stabilization of second-order LTI switched systems*, Int. J. Control **73** (2000) 1261–1279.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

Enabling the Use of Fish Tank Virtual Reality Systems with Curved Monitors

Mariano Pérez[1], Silvia Rueda[1] and Juan M. Orduña[1]

[1] *Departamento de Informática, Universidad de Valencia*

emails: mariano.perez@uv.es, silvia.rueda@uv.es, juan.orduna@uv.es

Abstract

Curved monitors are getting more and more popular nowadays, due to the immersive sensation they provide to users in front of planar monitors. The display in curved monitors are cylindrical surfaces with significant curvatures, which must be taken into account for a correct projection of the rendered scene.

In this paper, we propose new equations for non-linear image projection over cylindrical surfaces (adapted to the geometry of the curved monitor displays) instead of the planar projection. These equations enable the implementation of a Fish Tank Virtual Reality system based on curved monitors. In each frame, and once the position of the observer eyes has been detected, this system computes the correct stereographic projection of the scene, in order to achieve a realistic representation with a high degree of immersion. This Fish Tank Virtual Reality system including curved monitors significantly improves the user immersive sensation.

Key words: FishTank VR, Curved Monitors, Non-Planar Projections, View-Dependent Stereoscopy

1 Introduction

Fish Tank Virtual Reality (FTVR) systems render the stereoscopic image of a 3D scene on a monitor by using a perspective projection coupled to the position of the observer eyes [1]. These systems were proposed for the first time to explore the effect that would have in the user experience the tracking of the user head in the exploring of 3D virtual environments [2]. In that work, the system was composed of three main elements: a pair of stereoscopic glasses, a high-frequency planar monitor, and a mechanical device for the head tracking.

 ISBN: 978-84-617-8694-7

Stereoscopic glasses provide different images to left and right eyes, significantly improving the depth perception and the immersive sensation. Typically, there are two kind of glasses, passive, or active ones. The former ones can be based on either color filters (anaglyph glasses) or polarization filters (polarized glasses). The latter ones use switching shutters (LC shutter glasses) [4]. The technology behind these devices has become stable, and it has hardly evolved in the last five years. However, the technology for the head tracking has significantly changed since its original proposal, allowing the replacement of the mechanical devices by a simple web camera [3]. This improvement has reduced costs and has made the process much more comfortable for the user. Finally, computer monitors technology has suffered an incredible evolution, changing even their geometry and appearing curved instead of planar monitors. However, in spite of this evolution the current implementation of existing FTVR systems still use a planar monitor as physical displays. As a consequence, the graphical software still produces a planar perspective projection.

In this paper, we propose a Fish Tank Virtual Reality system for curved monitors. This system uses a non-linear projection over cylindrical surfaces (adapted to the geometry of the curved monitor) instead of the typical planar projection. In each frame, and once the position of the observer eyes has been detected, this system computes the correct stereographic projection of the scene, in order to achieve a realistic representation. This system enables the use of curved monitors in FTVR systems, increasing the user immersive sensation.

The rest of the paper is organized as follows: Section 2 introduces some preliminary concepts about FTVR systems required for understanding the rest of the paper. Section 3 describes the non-planar projection carried out on curved monitors. Next, section 4 shows the results obtained with the proposed implementation. Finally, section 5 shows some concluding remarks.

2 Background

Curved monitors are becoming increasingly popular because they yield a better user experience than planar monitors, even though the experienced improvement depends on then monitor width, the shape of the monitor curve, and the distance to the user eyes. Unlike the case of curved TVs, this distance is usually small. One of the reasons for the extended use of curved monitors is the fact that they improve the user immersion. They slightly curve the image forward, yielding a wider field of view for the same width, as shown in figure 1.

Other reason for the increasing use of curved monitors is the superior image quality. On one hand, curved monitors focus the light coming from the monitor directly to the user eyes, increasing the light contrast in a factor ranging between 1.5x and 1.8x in regard to the one achieved with a planar monitor. On other hand, curved monitors yield a more uniform image quality between the central and peripheral pixels. This is due to the fact that current technologies like IPS or OLED require that the user view is as perpendicular

 ISBN: 978-84-617-8694-7

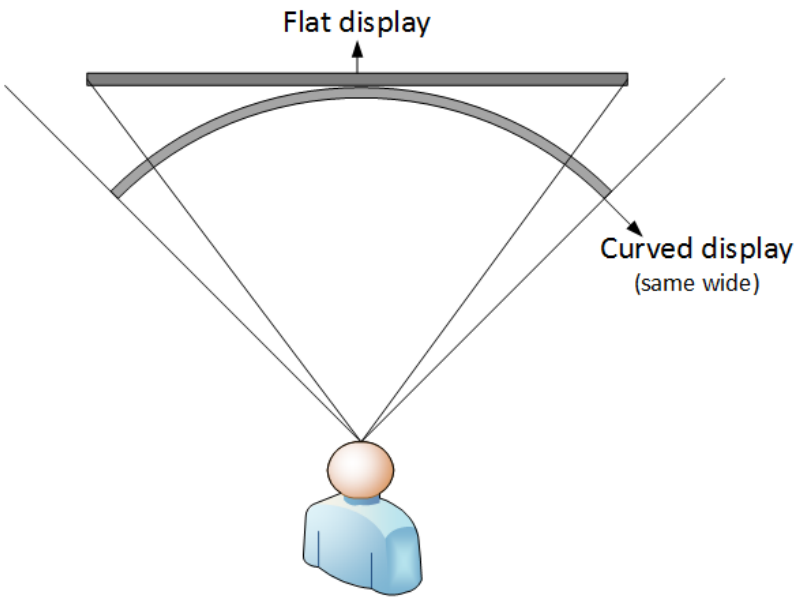


Figure 1: A curved monitor yields a wider field of view than a planar monitor of the same width.

as possible to the display surface. In the case of a large monitor with a user located close to the monitor surface (a common situation), the peripheral pixels are not display with a high quality because there is a small angle between the user view and the surface. However, in a curved monitor ideally all the pixels are located at the same distance and forming the same angle with the user view, as illustrated in figure 2.

A monitor in a FTVR system acts as a window through which the user can explore the virtual world. The graphics pipeline used for rendering the scene should project the scene on the display geometry of the monitor in order to produce credible effects. Traditionally, planar monitors have been used, so the graphics software and hardware (rasterization algorithms and the graphic pipeline in graphic cards) are optimized for planar projections. When a curved monitor is used, a non-planar projection adjusted to the display geometry must be used to map the 3D virtual scene on the final viewport. This task cannot be directly carried out with the current graphic hardware, and it may imply a high difficulty, specially if an asymmetric view frustum is used (the projection center is not located on the perpendicular line passing through the center of the projection plane, but on any other point, and changing in function of the relative position of the user in regard to the monitor).

Traditionally, there have been two different kinds of non-planar projections: image-based or geometry-based rendering methods [5]. Image-based implementations carry out

 ISBN: 978-84-617-8694-7

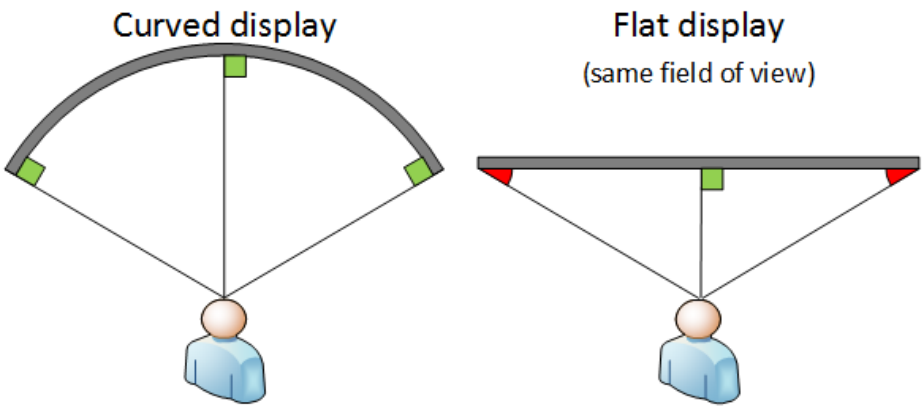


Figure 2: The apparent size and vision angle is the same for both central and peripheral pixels in the case of a curved monitor.

two steps: first, they render the 3D scene onto a texture using planar projections, and the second step renders the final view by applying the texture with the desired non-planar projection equations. The main potential weaknesses of this method are the appearance of sampling artifacts and the incompatibility with the antialiasing filtering hardware [8]. In its turn, the geometry-based rendering methods can be split into two groups: the first one splits the viewport into several narrow, rectangular slices, and then it uses planar perspective projections on each of them [6, 7]. These implementations usually yield a good performance with a reduced number of slices, but the performance decreases with the number of slices. Also, an additional drawback is the difficulty for implementing this method on current graphic architectures. The second group of geometry-based rendering methods directly compute the equation of the non-planar projection on mesh vertices, and they render the final image using the usual linear rasterization method implemented in the graphic pipeline. This method is conceptually simple and it can be easily implemented in the vertex shader (thus it could work even in early programmable GPUs), and even using instructions of the old non-programmable graphic libraries [9]. However, the main drawback of this approach is that the vertices are transformed using non-linear equations, but the inner fragments of the primitives are linearly interpolated using the rasterization process of the graphic pipeline. This problem becomes specially serious when coarse geometry is rendered, although it can be reduced by splitting the primitives through the use of geometry shaders [5] or tessellation shaders[10].

Currently, the implementations of non-planar projections described in the literature are restricted to the particular case where the view volume corresponds to a symmetric view frustum (that is, the line connecting the Center of Projection (COP) with the center of the display is perpendicular to the display surface, as shown on the left part if figure 3). Also,

 ISBN: 978-84-617-8694-7

in the case of spheric or cylindric projections the COP is located always at the center of the sphere or cylinder. This limitation represents an additional problem if the user is not located exactly at that position or when stereo is used (since each user eye must be the COP, the surface of the sphere/cylinder is different for each eye, the physical display cannot be the projection surface for both eyes simultaneously), because the projection rendered on the display provides the user with a wrong perspective from his/her point of view. Precisely, this is the typical situation in a FTVR system, since in these systems the display is located on a fixed position, the user can freely move, and the projection must be adapted on each moment to the position of the user eyes (once they have been detected through any tracking system). This is the problem motivating this work, which is aimed to develop a non-planar projection suitable for the use of curved monitors in FTVR systems.

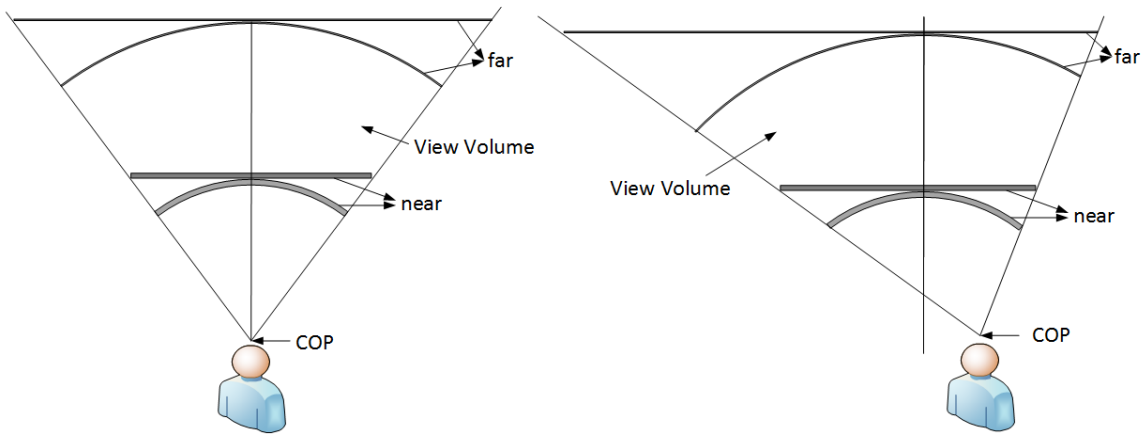


Figure 3: Planar and non-planar projection with symmetric view frustum (left), and asymmetric view frustum (right).

3 A Rendering Method for Asymmetric view frustum Rendered on Curved Displays

This method assumes a FTVR using a curved monitor as the physical display where the scene included in the view volume is projected. Since each eye COP is located at the position of the user eyes, and the user head can be located at any relative position in regard to the monitor (located at a fixed position), the view frustum corresponding to each eye must be necessarily asymmetric.

The geometry of curved monitor displays corresponds to a cylinder portion, and therefore the projection must be cylindrical. Figure 4 shows a scheme of the starting situation. In this figure, R y θ_{max} are physical parameters of the curved monitor, representing the display

curvature and the maximum angle obtained from the center of the cylinder, respectively. The center of that cylinder is taken as the origin of the world coordinates system.

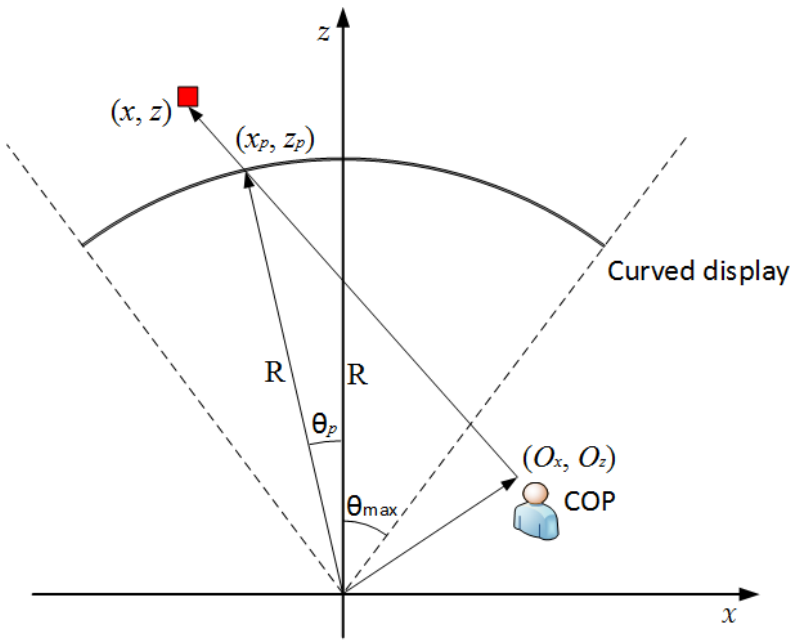


Figure 4: Cylindrical projection scheme

Figure 4 shows a case where the COP is not located at the coordinates origin, but on a different position (O_x, O_y). Following the straight line connecting the position (x, z) of any point of an scene object to the COP, this point is projected onto the cylinder surface point (x_p, z_p), which in polar coordinates corresponds to (R, θ_p). The transformation from Cartesian to polar coordinates is carried out using the following equations:

$$(x_p, z_p) = (R \sin \theta_p, R \cos \theta_p) \tag{1}$$

In order to obtain the projection of a given point, we compute the intersection point of the straight line passing through points (x, z) and (O_x, O_z), whose slope is $m = (x - O_x)/(z - O_z)$, and the circumference of the cylinder $x_p^2 + z_p^2 = R^2$. Combining these equations, we obtain the following equation:

$$z_p^2(1 + m^2) + z_p(2mO_x - 2m^2O_z) + (m^2O_z^2 + O_x^2 - 2mO_zO_x - R^2) = 0 \tag{2}$$

Equation 2 is a second order equation. While the COP is located inside the cylinder (the usual situation) this equation has two solutions: one of them corresponds to the first

 ISBN: 978-84-617-8694-7

intersection of the straight with the circumference (the point we are looking for) and the other one corresponds to the intersection point behind the COP. However, the latter point can be ignored because it will not be visible from this point of view. If we call $a = 1 + m^2$, $b = 2mO_x - 2m^2O_z$ y $c = m^2O_z^2 + O_x^2 - 2mO_zO_x - R^2$, since a is a positive value, the coordinates of the searched intersection point are obtained from the following equations:

$$z_p = \frac{-b + \sqrt{b^2 - 4ac}}{2a} \tag{3}$$

$$x_p = m(z_p - O_z) + O_x \tag{4}$$

However, these equations are incomplete, since we have not considered the height component y, which is the main cylinder axis. Taking as origin the center of the display, the component y_p is obtained in a similar way of that of x_p, from the straight line connecting points (y, z) and (O_y, O_z), and whose slope is $m' = (y - O_y)/(z - O_z)$:

$$y_p = m'(z_p - O_z) + O_y \tag{5}$$

3.1 Implementation

We propose a geometry-based solution which applies the cylindrical projection equations directly to the position of the mesh vertices. The transformations are applied in the vertex shader, and they transform all the vertices included within the view volume to vertices included within the typical canonical clipping volume: a cube whose edges length is 2 coordinate units, centered at the coordinates origin, whose points have coordinates $x_c, y_c, z_c \in [-1, 1]$. The primitives which are totally or partially included within the cube follow from that point the usual stages of the graphic pipeline.

The starting view volume, which corresponds to an asymmetric view frustum, is truncated on both the top and bottom part by two cylindrical surfaces of radio R_{far} y R_{near} centered on the coordinates origin. Since the *near* surface corresponds to the surface of the curve monitor, it should be taken $R_{near} = R$, being R the curvature radius of the curved monitor. R_{far} can be taken as any value greater than R_{near}. The other two parameters defining the display geometry are the angle θ_{max} shown in figure 4 and the monitor display height h. Starting from these parameters, we transform this non-linear volume into a regular volume (a cube) by using these equations:

$$x_c = \frac{\theta_p}{\theta_{max}} \tag{6}$$

$$y_c = \frac{2y_p}{h} \tag{7}$$

$$z_c = \frac{2\sqrt{x^2 + z^2}}{R_{far} - R_{near}} - \frac{R_{far} + R_{near}}{R_{far} - R_{near}} \tag{8}$$

 ISBN: 978-84-617-8694-7

Where θ_p is the angular component of the intersection point polar coordinates, which is obtained from coordinate x_p by applying the inverse transformation of equation 1. y_p is the value obtained from equation 5, and x and z are the coordinates of the vertex included in the view volume for which the transformation is done.

As commented above, a drawback of this approach is that it only transforms in a non-linear way the vertices, but not the inner fragments of the triangles corresponding to these vertices (which are linearly interpolated using the rasterization process of the graphic pipeline). This problem can be visually perceptible in the edges of a relatively large triangle (a triangle including a significant number of pixels once projected on the display) as the fact that the sides appear as straight lines instead of appearing as curve lines, as id could be expected in a cylindrical projection. In order to minimize this problem, we include tessellation shaders in the graphic pipeline. These shaders take as input primitives each of the triangles included in the view volume, using the vertices coming from the output of the vertex shader as input control points. By applying the the viewport transformation on the triangles we obtain their coordinates in the window space (in pixels), and we can determine the apparent size of each triangle in the display, as well as decide if it should be split it into more triangles. When triangles are subdivided, the same transformations made in the vertex shader (equations 6, 7 and 8) are applied to each of the new generated vertices in the tessellation shaders.

4 Results

In this section, we analyze the results obtained when the proposed transformations are applied, compared to the results obtained with the linear projection transformations. We have implemented a FTVR based on a personal computer with an Intel Core i7-4790 processor, 12 GB memory and an NVIDIA GeForce GTX 650 GPU; a Logitech C170 webcam for the tracking of user head and eyes; some pairs of anaglyph 3D glasses for obtaining an stereoscopic vision; one planar and one curved monitor, both with a 21 inch display, for comparing the obtained results. Both display dimensions are 67 cm. wide and 28 cm. high.

The rendered scene for performance evaluation purposes is a quite simple scene, composed of a matrix of 5x3x3 cubes, with a 2D wood texture of 512x512 pixels mapped on every cube face. The number of vertices in the cubes is variable, and we have changed it in order to study how it affects to system performance.

Figures 5 and 6 show the visual effects obtained with both kinds of projections: planar (left figures) and cylindrical (right figures). The first figure shows a scene using symmetric view frustum, while the second one shows the same scene using asymmetric view frustum, that is, changing the position of the user head (the COP has been moved 25 cm. right and 10 cm. up with respect to the monitor center).

Both figures show significant differences between the shape of the pictures on the left

 ISBN: 978-84-617-8694-7

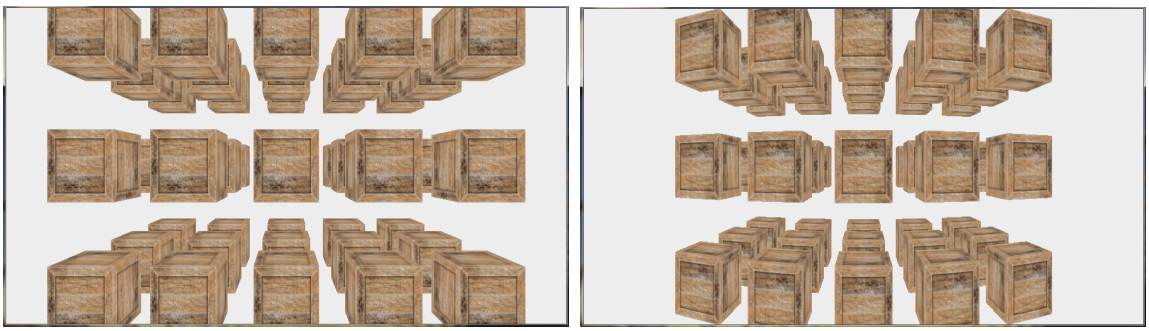


Figure 5: Scene rendered using symmetric view frustum with linear (left) and cylindrical (right) projections.

Figure 6: Scene rendered using asymmetric view frustum with linear (left) and cylindrical (right) projections.

and the right sides. Moreover, the right picture in figure 6 shows how the proposed transformation allow the correct rendering on a curved monitor even when the user is not located at the COP. These results show that the proposed transformations efficiently adapt the rendered image to the curved monitor geometry, achieving a more realistic effect regardless of the head user position.

Next, figure 7 corresponds to the same scene, but in this case using anaglyph images for achieving stereoscopic 3D effect. That is, this figure shows the same scene ready to be seen by users wearing anaglyph glasses. Nevertheless, since the image is ready for curved monitors, the reader can experience the result achieved on a curved monitor by printing the image on a paper and curving the paper. By moving the head right and up with respect to the line running through the center of the image, the reader will obtain the correct view of the image. Again, this figure shows that the proposed transformation provides the user

 ISBN: 978-84-617-8694-7

with realistic effects in the case of 3D images.

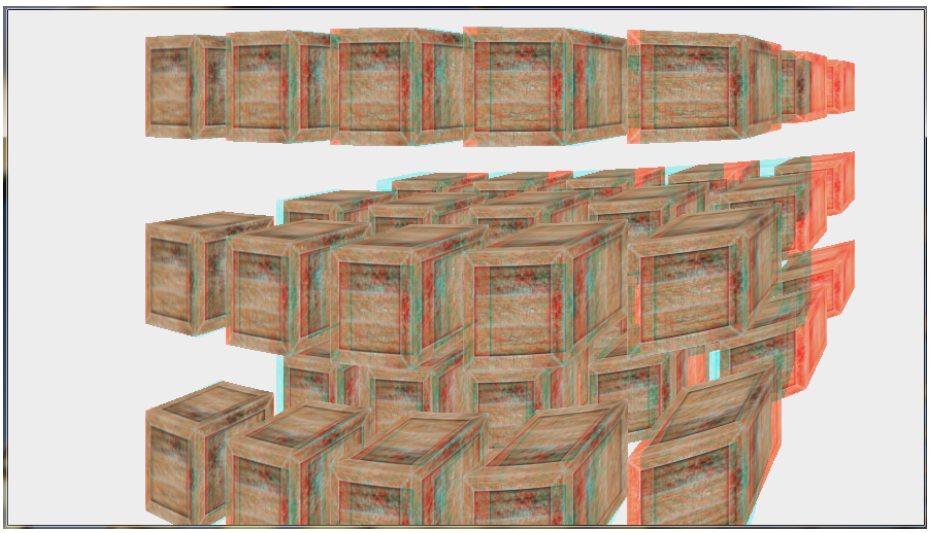


Figure 7: Stereoscopic rendering (red-cyan anaglyph) of the same scene using asymmetric view frustum with cylindrical projection.

On other hand, we have evaluated the quantitative effects of the proposed transformations. Table 1 shows comparative results, in terms of the frames per second rate achieved by the FTVR system, when both the proposed transformations and the typical planar projection transformations are applied. We have considered different number of vertices over which the transformations are applied, ranging from a few thousands to 400 million vertices.

	Num. of Vertices								
	6.5K	25K	100K	400K	1.5M	6.5M	25M	100M	400M
Linear Proj.	1430	1425	1420	1417	625	187	51	12.9	3.3
Cylind. Proj.	1427	1423	1420	1415	620	186	50	12.8	3.2

Table 1: Frames per second rendered as a function of the number of vertices in the scene.

Table 1 shows that the use of the proposed transformation does not significantly affect the system performance, since the frames per second achieved with both transformations are very similar even when the number of the triangles in the scene is huge (even bigger than the number of rendered pixels). These results prove that the proposed transformations do not add any overhead to the graphic system.

 ISBN: 978-84-617-8694-7

5 Conclusions

In this paper, we have proposed new transformations that enable the implementation of a FishTank VR system by using curved monitors instead of the planar ones, in order to increase the user immersive sensation. We have reformulated the projection equations in the literature not only to adapt them to the cylindrical geometry of curved monitor displays, but also to take into account the used head location at every moment, which may be not centered in regard to the position of the monitor. Also, we have implemented the new equations in the vertex and the tessellation shaders, using cylindrical projections with asymmetric view frustum. We have evaluated that implementation in a FTVR system with a curved monitor. The performance evaluation results show that the user perceives a correct perspective of the rendered scene at every moment, regardless of his/her head relative position to the monitor. Also, the quantitative results show that the proposed equations have no significant effects on the graphic system performance, achieving the same frame rates than when using the linear equations in the literature. These results prove that the proposed transformations enable the use of curved monitors in FTVR systems, improving the user immersive sensation.

Acknowledgments

This work has been supported by Spanish MINECO and EU FEDER funds under grants TIN2015-66972-C5-5-R, TIN2016-81850-REDC, and TIN2016-81840-REDT.

References

[1] J. D. Mulder, R. Van Liere, *Enhancing fish tank VR*, Proceedings IEEE Virtual Reality (2000) 91–98.

[2] C. Ware, K. Arthur, K.S. Booth, *Fish tank virtual reality*, Proceedings of the INTERACT'93 and CHI'93 conference on Human factors in computing systems (1993) 37–42.

[3] E. Murphy-Chutorian, M. Trivedi, *Head pose estimation in computer vision: A survey.*, IEEE transactions on pattern analysis and machine intelligence **31**.4 (2009) 607–626.

[4] P. May, *A survey of 3-D display technologies*, Information Display **32** (2005) 28–33.

[5] M. Trapp, H. Lorenz, J. Dllner, *Interactive stereo rendering for non-planar projections of 3d virtual environments*, Proceedings of GRAPP (2009) 199–204.

 ISBN: 978-84-617-8694-7

[6] A. Simon, R. C. Smith, R. R. Pawlicki, *Omnistereo for panoramic virtual environment display systems*, Proceedings of IEEE Virtual Reality (2004) 67–73.

[7] H. Lorenz, J. Dllner, *Real-time Piecewise Perspective Projections*, Proceedings of GRAPP (2009) 147–155.

[8] K. Petkov, C. Papadopoulos, M. Zhang, A. E. Kaufman, X. Gu, *Interactive visibility retargeting in vr using conformal visualization*, IEEE transactions on visualization and computer graphics **18**.7 (2012) 1027–1040.

[9] S. Bayarri, *Computing non-planar perspectives in real time*, Computers and Graphics **19**.3 (2009) 431–440.

[10] J. Ardouin, A. Lcuyer, M. Marchal, E. Marchand, *Stereoscopic rendering of virtual environments with wide Field-of-Views up to 360*, Proceedings of IEEE Virtual Reality (2014) 3–8.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

Fuzzy fixed point theorems for (β_M, ψ, ξ) fuzzy contractive mappings

Supak Phiangsungnoen[1] and Wiyada Kumam[2]

[1] *Department of Mathematics, Faculty of Liberal Arts, Rajamangala University of Technology Rattanakosin (RMUTR), 264 Chakkrawat Rd., Chakkrawat, Samphanthawong, Bangkok 10100, Thailand.*

[2] *Program in Applied Statistics, Department of Mathematics and Computer Science, Faculty of Science and Technology, Rajamangala University of Technology Thanyaburi, Thanyaburi, Pathumthani 12110, Thailand.*

emails: supak.pia@rmutr.ac.th, wiyada.kum@mail.rmutt.ac.th

Abstract

In this paper, we introduce the notion of (β_M, ψ, ξ) fuzzy contractive mappings and investigate the existence of fuzzy fixed points for such mappings. We also give illustrative example to show that our result is more general than the results in literatures.

Key words: Admissible mapping; β_M-admissible mapping; Fuzzy mapping; Fuzzy fixed point.
MSC 2000: 46T99, 47H10, 47H09, 54H25.

1 Introduction

Fuzzy concepts play the one importance role in real world problems which were used to modeled mathematically for the purpose of automation, like regulation, production control and many others. The study of fuzzy concepts for contractive mappings was instigated by Heilpern [2]. He introduced fuzzy contractive mappings and proved the existence of fuzzy fixed point theorem which is a generalization of Nadler's [3]. Afterward, there are mathematicians generalized and proved a fuzzy fixed point theorems for fuzzy contractive mappings in metric spaces and other spaces (see [4–12]).

The notion of β_*–admissible for multivalued mapping was introduced by Asl *et al.* [13]. Mohammadi *et al.* [14] introduced the concept of β-admissible for multivalued mappings

 ISBN: 978-84-617-8694-7

which is different from the notion of β_*–admissible. In 2014, Phiangsungnoen *et al.* [10] introduced the notion of β_M admissible for fuzzy mappings in the sense of Mohammadi *et al.* [14]. In the same year, Ali *et al.* [16]introduced and proved the notion of (β, ψ, ξ)-contractive multivalued mappings to generalized and extended the notion of contractive mappings to closed valued multifunctions via β_*–admissible.

In this paper, by using the notion of β_M admissible of Phiangsungnoen *et al.* [10]. We introduce the concept of (β_M, ψ, ξ) fuzzy contractive mappings and investigate the existence of fuzzy fixed points for such mappings. We also give illustrative example to show that our result is more general than the results in literatures.

2 Preliminaries

In this section, we recall some basic definitions and preliminaries that will be needed in this paper.

Let X be a space of points with generic elements of X denoted by x and $I = [0, 1]$. A fuzzy subset of X is characterized by a membership function such that each element in X is associated with a real number in the interval I.

Let (X, d) be a metric space and A be a fuzzy set in X. If X is endowed with topology, for $\alpha \in [0, 1]$, the α-level set of A, denoted by $[A]_\alpha$, is defined as $[A]_\alpha = \{x : A(x) \geq \alpha\}$, where $\alpha \in (0, 1]$ and for $\alpha = 0$ we have $[A]_0 = \overline{\{x : A(x) > 0\}}$, where $\overline{B}$ denotes the closure of the non-fuzzy set B.

Let X be a nonempty set. For $x \in X$, we write $\{x\}$ the characteristic function of the ordinary subset $\{x\}$ of X. For $\alpha \in (0, 1]$, the fuzzy point x_α of X is the fuzzy set of X given by

$$x_\alpha(y) = \begin{cases} \alpha & ; \quad x = y \\ 0 & ; \quad x \neq y \end{cases} .$$

In the sequel, $I = [0, 1]$, (X, d) and I^X denote the metric space and the collection of all fuzzy subsets of X. For $A, B \in I^X$, a fuzzy set A is said to be more accurate than a fuzzy set B (denoted by $A \subset B$) if and only if $Ax \leq Bx$ for each x in X, where $A(x)$ and $B(x)$ denote the membership function of A and B, respectively. For $x \in X$, $S \subset X$, $A, B \in I^X$ and $\alpha \in [0, 1]$, we define $d(x, S) = \inf\{d(x, a); a \in S\}$,

$$p_\alpha(x, A) = \inf\{d(x, a); a \in [A]_\alpha\},$$

$$p_\alpha(A, B) = \inf\{d(a, b); a \in [A]_\alpha, b \in [B]_\alpha\},$$

$$p(A, B) = \sup_\alpha p_\alpha(A, B),$$

 ISBN: 978-84-617-8694-7

$$D_\alpha(A,B) = H\big([A]_\alpha,[B]_\alpha\big) = \max\Big\{\sup_{a\in[A]_\alpha} d(a,[B]_\alpha), \sup_{b\in[B]_\alpha} d(b,[A]_\alpha)\Big\},$$

$$d_\infty(A,B) = \sup_\alpha D_\alpha(A,B).$$

and

$$\mathcal{W}_\alpha(X) = \{C \in I^X : [C]_\alpha \text{ is nonempty and compact }\}.$$

It is easy to see that H is the Hausdorff metric on $\mathcal{W}_\alpha(X)$ induced by the metric d.

Definition 2.1 ([4]) *Let (X,d) be a metric space. $Q : X \to I^X$ and $\alpha \in [0,1]$. A fuzzy point x_α in X is called fuzzy fixed point of Q if $x_\alpha \subset Qx$. If $\{x\} \subset Qx$, then x is called fixed point of Q.*

Lemma 2.2 ([2]) *Let (X,d) be a metric space, $x \in X$ and $A \in I^X$. For $\alpha \in [0,1]$, if $p_\alpha(x,A) = 0$ and $[A]_\alpha$ is closed subset of X, then $x_\alpha \subset A$.*

Definition 2.3 ([13]) *Let X be a nonempty set, $Q : X \to 2^X$, where 2^X is a collection of nonempty subsets of X and $\beta : X \times X \to [0,\infty)$. We say that Q is β_*-admissible if for $x,y \in X$ with $\beta(x,y) \geq 1$we have$\beta_*(Qx,Qy) \geq 1$, where $\beta_*(Qx,Qy) := \inf\{\beta(a,b) : a \in Qx$ and $b \in Qy\}$.*

Definition 2.4 ([14]) *Let X be a nonempty set, $Q : X \to 2^X$, where 2^X is a collection of nonempty subsets of X and $\beta : X \times X \to [0,\infty)$. We say that Q is β-admissible whenever for each $x \in X$ and $y \in Qx$ with $\beta(x,y) \geq 1$, we have $\beta(y,z) \geq 1$ for all $z \in Qy$.*

Remark 2.5 *If Q is β_*-admissible mapping, then Q is also β-admissible mapping.*

Definition 2.6 ([10]) *Let (X,d) be a metric space, $\beta : X \times X \to [0,\infty)$, $\alpha \in [0,1]$ and $Q : X \to W_\alpha(X)$. A mapping Q is said to be β_M-admissible if for each $x \in X$ and $y \in [Qx]_\alpha$, with $\beta(x,y) \geq 1$, we have $\beta(y,z) \geq 1$ for all $z \in [Qy]_\alpha$.*

In this results we let Ψ be the family of nondecreasing $\psi : [0,\infty) \to [0,\infty)$ such that $\sum_{n=1}^{\infty} \psi^n(t) < \infty$ for each $t > 0$ and ψ is continuous function. It easy to see that for $\psi \in \Psi$, $\psi(t) < t$ for all $t > 0$ and $\psi(0) = 0$. Moreover, we let Ξ be the family of functions $\xi : [0,\infty) \to [0,\infty)$ such that ξ is continuous, nondecreasing on $[0,\infty)$, $\xi(0) = 0$ and $\xi(t) > 0$ for all $t \in (0,\infty)$ and ξ is subadditive function. The example of ξ function showed in Example 2.1 [16].

Lemma 2.7 *Let (X,d) is a metric space and let $\xi \in \Xi$. Then $(X, \xi \circ d)$ is a metric space.*

 ISBN: 978-84-617-8694-7

3 Fuzzy fixed point theorems for (β_M, ψ, ξ) fuzzy contractive mappings

Theorem 3.1 *Let (X, d) be a complete metric space, $\alpha \in [0, 1]$ and Q be fuzzy mapping from X to $\mathcal{W}_\alpha(X)$. Suppose that there exist $\psi \in \Psi$, $\xi \in \Xi$ and $\beta : X \times X \to [0, \infty)$ such that*

$$\beta(x, y) \geq 1 \Rightarrow \xi(D_\alpha(Qx, Qy)) \leq \psi(\xi(J(x, y) + lK(x, y))), \tag{3.1}$$

for all $x, y \in X$, where $l \geq 0$ and

$$\begin{aligned} J(x, y) &= \max\Big\{d(x, y), p_\alpha(x, Qx), p_\alpha(y, Qy), \frac{p_\alpha(x, Qy) + p_\alpha(y, Qx)}{2}\Big\}, \\ K(x, y) &= \min\big\{p_\alpha(x, Qx), p_\alpha(y, Qy), p_\alpha(x, Qy), p_\alpha(y, Qx)\big\}. \end{aligned}$$

If the following condition holds :

(i) Q is β_M-admissible,

(ii) there exist $x_0 \in X$ and $x_1 \in [Qx_0]_\alpha$ such that $\beta(x_0, x_1) \geq 1$,

(iii) if $\{x_n\}$ is sequence in X such that $\beta(x_n, x_{n+1}) \geq 1$ and $x_n \to x$ as $n \to \infty$, then $\beta(x_n, x) \geq 1$,

then there exists $x \in X$ such that x_α is a fuzzy fixed point of Q.

Proof 1 *By hypothesis, there exist $x_0 \in X$ and $x_1 \in [Qx_0]_\alpha$ in condition (ii), we get $\beta(x_0, x_1) \geq 1$. If $x_0 = x_1$, then we have nothing to prove. Let $x_0 \neq x_1$. If $x_1 \in [Qx_1]_\alpha$, then x_1 is fuzzy fixed point of Q. Let $x_1 \notin [Qx_1]_\alpha$. Since $[Qx_1]_\alpha$ is nonempty compact subset of X, there exists $x_2 \in [Qx_1]_\alpha$, such that*

$$0 < \xi(d(x_1, x_2)) = \xi(p_\alpha(x_1, Qx_1)) \leq \xi(D_\alpha(Qx_0, Qx_1)). \tag{3.2}$$

From (3.2) and the fact that $\beta(x_0, x_1) \geq 1$, we have

$$\begin{aligned} 0 < \xi(d(x_1, x_2)) &\leq \xi(D_\alpha(Qx_0, Qx_1)) \\ &\leq \psi(\xi(J(x_0, x_1) + lK(x_0, x_1))) \end{aligned} \tag{3.3}$$

where

$$\begin{aligned} J(x_0, x_1) &= \max\Big\{d(x_0, x_1), p_\alpha(x_0, Qx_0), p_\alpha(x_1, Qx_1), \frac{p_\alpha(x_0, Qx_1) + p_\alpha(x_1, Qx_0)}{2}\Big\} \\ &= \max\Big\{d(x_0, x_1), p_\alpha(x_1, Qx_1)\Big\}, \end{aligned}$$

 ISBN: 978-84-617-8694-7

and

$$\begin{aligned} K(x_0,x_1) &= \min\{p_\alpha(x_0,Qx_0),p_\alpha(x_1,Qx_1),p_\alpha(x_0,Qx_1),p_\alpha(x_1,Qx_0)\} \\ &= \min\{p_\alpha(x_0,x_1),p_\alpha(x_1,x_2),p_\alpha(x_0,x_2),0\}. \end{aligned}$$

Assume that $\max\{d(x_0,x_1),p_\alpha(x_1,Qx_1)\} = p_\alpha(x_1,Qx_1)$. *Then from equation (3.3), we have*

$$\begin{aligned} 0<\xi(d(x_1,x_2)) &\leq \xi(D_\alpha(Qx_0,Qx_1)) \\ &\leq \psi(\xi(J(x_0,x_1)+lK(x_0,x_1))) \\ &\leq \psi(\xi(\max\{d(x_0,x_1),p_\alpha(x_1,Qx_1)\}+0)) \\ &= \psi(\xi(p_\alpha(x_1,Qx_1))) \\ &\leq \psi(\xi(d(x_1,x_2))) < \xi(d(x_1,x_2)) \end{aligned}$$

which is a contradiction. Hence, $\max\{d(x_0,x_1),p_\alpha(x_1,Qx_1)\} = d(x_0,x_1)$. *Then from equation (3.3), we have*

$$0<\xi(d(x_1,x_2)) \leq \xi(D_\alpha(Qx_0,Qx_1)) \leq \psi(\xi(d(x_0,x_1))).$$

By the same argument, for $x_2 \in X$, *we have* $[Qx_2]_\alpha$ *is a nonempty compact subset of* X *and then there exists* $x_3 \in [Qx_2]_\alpha$ *such that*

$$\xi(d(x_2,x_3)) = \xi(p_\alpha(x_2,Qx_2)) \leq \xi(D_\alpha(Qx_1,Qx_2)). \tag{3.4}$$

For $x_0 \in X$ *and* $x_1 \in [Qx_0]_\alpha$ *with* $\beta(x_0,x_1) \geq 1$, *by definition of* β_M*-admissible, we get*

$$\beta(x_1,x_2) \geq 1. \tag{3.5}$$

From (3.1), (3.4) and (3.5), we have

$$\begin{aligned} \xi(d(x_2,x_3)) &\leq \xi(D_\alpha(Qx_1,Qx_2)) \\ &\leq \psi(\xi(J(x_1,x_2)+lK(x_1,x_2))) \end{aligned} \tag{3.6}$$

where

$$\begin{aligned} J(x_1,x_2) &= \max\left\{d(x_1,x_2),p_\alpha(x_1,Qx_1),p_\alpha(x_2,Qx_2),\frac{p_\alpha(x_1,Qx_2)+p_\alpha(x_2,Qx_1)}{2}\right\} \\ &= \max\{d(x_1,x_2),p_\alpha(x_2,Qx_2)\}, \end{aligned}$$

and

$$\begin{aligned} K(x_1,x_2) &= \min\{p_\alpha(x_1,Qx_1),p_\alpha(x_2,Qx_2),p_\alpha(x_1,Qx_2),p_\alpha(x_2,Qx_1)\} \\ &= \min\{p_\alpha(x_1,x_2),p_\alpha(x_2,x_3),p_\alpha(x_1,x_3),0\}. \end{aligned}$$

 ISBN: 978-84-617-8694-7

Assume that $\max\{d(x_1,x_2), p_\alpha(x_2,Qx_2)\} = p_\alpha(x_2,Qx_2)$. *Then from equation (3.6), we have*

$$\begin{aligned} 0 < \xi(d(x_2,x_3)) &\leq \xi(D_\alpha(Qx_1,Qx_2)) \\ &\leq \psi(\xi(J(x_1,x_2)+lK(x_1,x_2))) \\ &\leq \psi(\xi(\max\{d(x_1,x_2),p_\alpha(x_2,Qx_2)\}+0)) \\ &= \psi(\xi(p_\alpha(x_2,Qx_2))) \\ &\leq \psi(\xi(d(x_2,x_3))) < \xi(d(x_2,x_3)) \end{aligned}$$

which is a contradiction. Hence, $\max\{d(x_1,x_2), p_\alpha(x_2,Qx_2)\} = d(x_1,x_2)$. *Then from equation (3.6), we have*

$$0 < \xi(d(x_2,x_3)) \leq \xi(D_\alpha(Qx_1,Qx_2)) \leq \psi(\xi(d(x_1,x_2))) = \psi^2(\xi(d(x_0,x_1))).$$

By induction, we can construct a sequence $\{x_n\}$ *in* X *such that, for each* $n \in \mathbb{N}$, $x_n \in [Qx_{n-1}]_\alpha$ *with* $\beta(x_{n-1},x_n) \geq 1$ *and*

$$0 < \xi(d(x_n,x_{n+1})) \leq \xi(D_\alpha(Qx_{n-1},Qx_n)) \leq \psi(\xi(J(x_{n-1},x_n)+lK(x_{n-1},x_n)))$$

where

$$\begin{aligned} J(x_{n-1},x_n) &= \max\left\{d(x_{n-1},x_n), p_\alpha(x_{n-1},Qx_{n-1}), p_\alpha(x_n,Qx_n), \frac{p_\alpha(x_{n-1},Qx_n)+p_\alpha(x_n,Qx_{n-1})}{2}\right\} \\ &= \max\{d(x_{n-1},x_n), p_\alpha(x_n,x_{n+1})\}, \end{aligned}$$

and

$$\begin{aligned} K(x_{n-1},x_n) &= \min\left\{p_\alpha(x_{n-1},Qx_{n-1}), p_\alpha(x_n,Qx_n), p_\alpha(x_{n-1},Qx_n), p_\alpha(x_n,Qx_{n-1})\right\} \\ &= \min\left\{p_\alpha(x_{n-1},x_n), p_\alpha(x_n,x_{n+1}), p_\alpha(x_{n-1},x_{n+1}), 0\right\}. \end{aligned}$$

Hence

$$\xi(d(x_n,x_{n+1})) \leq \psi(\xi(\max\{d(x_{n-1},x_n), p_\alpha(x_n,x_{n+1})\})) \tag{3.7}$$

for all $n \in \mathbb{N}$. *If there exists* $n^* \in \mathbb{N}$ *which* $p_\alpha(x_{n^*},Qx_{n^*}) = 0$, *then from Lemma 2.2, we have* $(x_{n^*})_\alpha \subset Qx_{n^*}$, *that is* $(x_{n^*})_\alpha$ *is a fuzzy fixed point of* Q. *Therefore, we suppose that for each* $n \in \mathbb{N}$, $p_\alpha(x_n,Qx_n) > 0$ *and thus* $d(x_{n-1},x_n) > 0$ *for all* $n \in \mathbb{N}$. *So, if* $d(x_n,x_{n+1}) > d(x_{n-1},x_n)$ *for some* $n \in \mathbb{N}$, *then from (3.7) and* $\psi(t) < t$ *for* $t \in (0,\infty)$, *we have*

$$\xi(d(x_n,x_{n+1})) \leq \psi(\xi(d(x_n,x_{n+1}))) < \xi(d(x_n,x_{n+1}))$$

 ISBN: 978-84-617-8694-7

which is a contradiction to our assumption. Therefore, we have

$$\begin{aligned}\xi(d(x_n,x_{n+1})) &\leq \psi\big(\xi(d(x_{n-1},x_n))\big)\\ &\leq \psi\big(\psi\big(\xi(d(x_{n-2},x_{n-1}))\big)\big)\\ &\vdots\\ &\leq \psi^n\big(\xi(d(x_0,x_1))\big). \end{aligned} \tag{3.8}$$

Next, we will show that $\{x_n\}$ is a Cauchy sequence in X. Since function ψ is belong to Ψ, there exist $\nabla > 0$ and positive integer $g = g(\nabla)$ such that

$$\sum_{n\geq h}\psi^n(\xi(d(x_0,x_1))) < \nabla.$$

Let $m > n > g$. Using the triangular inequality, previous relation and (3.8), we have

$$\xi(d(x_n,x_m)) \leq \sum_{i=n}^{m-1}\xi(d(x_i,x_{i+1})) \leq \sum_{i=n}^{m-1}\psi^i(\xi(d(x_0,x_1))) \leq \sum_{n\geq g}\psi^n(\xi(d(x_0,x_1))) < \nabla.$$

This implies that $\lim_{n,m\to\infty} d(x_m,x_n) = 0$. Hence $\{x_n\}$ is a Cauchy sequence in (X,d). By completeness of (X,d), there exists $z \in X$ such that $x_n \to z$ as $n \to \infty$. Next, we claim that $p_\alpha(z,Qz) = 0$ for each $\alpha \in [0,1]$. If not there exixts $\alpha^\dagger \in [0,1]$ such that $p_{\alpha^\dagger}(z,Qz) > 0$. By condition (iii), we have $\beta(x_n,x) \geq 1$ for all $n \in \mathbb{N}$. Now we have

$$\begin{aligned}\xi(p_{\alpha^\dagger}(z,Tz)) &\leq \xi(d(z,x_{n+1}) + p_\alpha(x_{n+1},Qz))\\ &\leq \xi(d(z,x_{n+1})) + \xi(D_\alpha(Qx_{n+1},Qz))\\ &\leq \xi(d(z,x_{n+1})) + \psi\Big(\xi\big(J(z,x_n) + lK(z,x_n)\big)\Big)\end{aligned}$$

where

$$\begin{aligned}J(z,x_n) &= \max\left\{d(z,x_n), p_{\alpha^\dagger}(z,Qz), p_{\alpha^\dagger}(x_n,Qx_n), \frac{p_{\alpha^\dagger}(z,Qx_n)+p_{\alpha^\dagger}(x_n,Qz)}{2}\right\}\\ &= \max\left\{d(z,x_n), p_{\alpha^\dagger}(z,Qz), p_{\alpha^\dagger}(x_n,x_{n+1}), \frac{p_{\alpha^\dagger}(z,x_{n+1})+p_{\alpha^\dagger}(x_n,Qz)}{2}\right\}\end{aligned}$$

and

$$\begin{aligned}K(z,x_n) &= \min\left\{p_{\alpha^\dagger}(z,Qz), p_{\alpha^\dagger}(x_n,Qx_n), p_{\alpha^\dagger}(z,Qx_n), p_{\alpha^\dagger}(x_n,Qz)\right\}\\ &= \min\left\{p_{\alpha^\dagger}(z,Qz), p_{\alpha^\dagger}(x_n,x_{n+1}), p_{\alpha^\dagger}(z,x_{n+1}), p_{\alpha^\dagger}(x_n,Qz)\right\}\end{aligned}$$

Taking limit as $n \to \infty$ gives that

$$\xi_{\alpha^\dagger}(z,Qz)) \leq \psi\big(\xi(p_{\alpha^\dagger}(z,Qz))\big) < \xi_{\alpha^\dagger}(z,Qz)),$$

which is a contradiction. Therefore, we have $\xi(p_{\alpha^\dagger}(z,Qz))$ that means $p_\alpha(z,Qz) = 0$ for each $\alpha \in [0,1]$. By Lemma 2.2, we get $x_\alpha \subset Qx$. This complete the proof.

 ISBN: 978-84-617-8694-7

□

Next, we give some examples to support the validity of our result.

Example 3.2 *Let $X = [0,1]$ and $d : X \times X \to [0,\infty)$ as $d(x,y) = |x-y|$ for all $x, y \in X$. Define a mapping $Q : X \to I^X$ as follows:*

$$(Qx)(t) = \begin{cases} \frac{2}{5} & , 0 \le t \le \frac{x}{6}, \\ 0 & , \frac{x}{6} < t \le \frac{x+1}{6}, \\ \frac{1}{4} & , \frac{x+1}{6} < t \le 1. \end{cases}$$

Let $\alpha = \frac{2}{5}$. We observe that

$$[Qx]_\alpha = [Qx]_{\frac{2}{5}} = \left[0, \frac{x}{6}\right]$$

for all $x \in X$. Therefore, Q is fuzzy mapping from X to $\mathcal{W}_\alpha(X)$.

Define $\beta : X \times X \to [0,\infty)$ by

$$\beta(x,y) = \begin{cases} 1 & , x = y, \\ \frac{1}{|x-y|} & , x \ne y. \end{cases}$$

Then it is easy to check that Q is an β_M-admissible. For each $x, y \in X$, we get

$$\begin{aligned} \xi(D_\alpha(Qx, Qy)) &= \xi\big(H([Qx]_\alpha, [Qy]_\alpha)\big) \\ &= \xi(\frac{1}{6}|x-y|) \\ &= \xi(\frac{1}{6}d(x,y)) \\ &< \frac{1}{2}\xi(d(x,y)) \\ &\le \psi(\xi(J(x,y) + lK(x,y))), \end{aligned}$$

where $\psi(t) = \frac{t}{2}$ and $\xi(t) = \sqrt{t}$ for all $t > 0$ and $l \ge 0$. It is easy to see that conditions (ii) and (iii) in Theorem 3.1 hold. Therefore all conditions of Theorem 3.1 hold. Thus Q has an α-fuzzy fixed point $x \in X$, that is, a point $x = 0$.

□

 ISBN: 978-84-617-8694-7

By using Remark 2.5, we get the following result.

Theorem 3.3 *Let (X,d) be a complete metric space, $\alpha \in [0,1]$ and Q be fuzzy mapping from X to $\mathcal{W}_\alpha(X)$. Suppose that there exist $\psi \in \Psi$, $\xi \in \Xi$ and $\beta : X \times X \to [0,\infty)$ such that*

$$\beta(x,y) \geq 1 \Rightarrow \xi(D_\alpha(Qx,Qy)) \leq \psi(\xi(J(x,y)+lK(x,y))), \tag{3.9}$$

for all $x,y \in X$, where $l \geq 0$ and

$$\begin{aligned} J(x,y) &= \max\Big\{d(x,y), p_\alpha(x,Qx), p_\alpha(y,Qy), \frac{p_\alpha(x,Qy)+p_\alpha(y,Qx)}{2}\Big\}, \\ K(x,y) &= \min\big\{p_\alpha(x,Qx), p_\alpha(y,Qy), p_\alpha(x,Qy), p_\alpha(y,Qx)\big\}. \end{aligned}$$

If the following condition holds :

(i) Q is β_M-admissible,

(ii) there exist $x_0 \in X$ and $x_1 \in [Qx_0]_\alpha$ such that $\beta(x_0,x_1) \geq 1$,

(iii) if $\{x_n\}$ is sequence in X such that $\beta(x_n,x_{n+1}) \geq 1$ and $x_n \to x$ as $n \to \infty$, then $\beta(x_n,x) \geq 1$,

then there exists $x \in X$ such that x_α is a fuzzy fixed point of Q.

In Theorems 3.1 and 3.3, we taking $\xi(t) = t$ then we have the following corollary.

Corollary 3.4 *Let (X,d) be a complete metric space, $\alpha \in [0,1]$ and Q be fuzzy mapping from X to $\mathcal{W}_\alpha(X)$. Suppose that there exist $\psi \in \Psi$ and $\beta : X \times X \to [0,\infty)$ such that*

$$\beta(x,y) \geq 1 \Rightarrow D_\alpha(Qx,Qy) \leq \psi(J(x,y)+lK(x,y)), \tag{3.10}$$

for all $x,y \in X$, where $l \geq 0$ and

$$\begin{aligned} J(x,y) &= \max\Big\{d(x,y), p_\alpha(x,Qx), p_\alpha(y,Qy), \frac{p_\alpha(x,Qy)+p_\alpha(y,Qx)}{2}\Big\}, \\ K(x,y) &= \min\big\{p_\alpha(x,Qx), p_\alpha(y,Qy), p_\alpha(x,Qy), p_\alpha(y,Qx)\big\}. \end{aligned}$$

If the following condition holds :

(i) Q is β_M-admissible,

(ii) there exist $x_0 \in X$ and $x_1 \in [Qx_0]_\alpha$ such that $\beta(x_0,x_1) \geq 1$,

(iii) if $\{x_n\}$ is sequence in X such that $\beta(x_n,x_{n+1}) \geq 1$ and $x_n \to x$ as $n \to \infty$, then $\beta(x_n,x) \geq 1$,

 ISBN: 978-84-617-8694-7

then there exists $x \in X$ such that x_α is a fuzzy fixed point of Q.

Corollary 3.5 *([10]) Let (X, d) be a complete metric space, $\alpha \in [0, 1]$ and Q be fuzzy mapping from X to $\mathcal{W}_\alpha(X)$. Suppose that there exist $\psi \in \Psi$ and $\beta : X \times X \to [0, \infty)$ such that*

$$\beta(x, y) D_\alpha(Qx, Qy) \leq \psi(J(x, y)) + lK(x, y), \tag{3.11}$$

for all $x, y \in X$, where $l \geq 0$ and

$$\begin{aligned} J(x, y) &= \max\Big\{d(x, y), p_\alpha(x, Qx), p_\alpha(y, Qy), \frac{p_\alpha(x, Qy) + p_\alpha(y, Qx)}{2}\Big\}, \\ K(x, y) &= \min\big\{p_\alpha(x, Qx), p_\alpha(y, Qy), p_\alpha(x, Qy), p_\alpha(y, Qx)\big\}. \end{aligned}$$

If the following condition holds :

(i) Q is β_M-admissible,

(ii) there exist $x_0 \in X$ and $x_1 \in [Qx_0]_\alpha$ such that $\beta(x_0, x_1) \geq 1$,

(iii) if $\{x_n\}$ is sequence in X such that $\beta(x_n, x_{n+1}) \geq 1$ and $x_n \to x$ as $n \to \infty$, then $\beta(x_n, x) \geq 1$,

then there exists $x \in X$ such that x_α is a fuzzy fixed point of Q.

In corollary 3.4, we taking $\psi(t) = kt$ where $k \in (0, 1)$ then we have the following corollary.

Corollary 3.6 *Let (X, d) be a complete metric space, $\alpha \in [0, 1]$ and Q be fuzzy mapping from X to $\mathcal{W}_\alpha(X)$. Suppose that there exist $k \in (0, 1)$ and $\beta : X \times X \to [0, \infty)$ such that*

$$\beta(x, y) \geq 1 \Rightarrow D_\alpha(Qx, Qy) \leq kJ(x, y) + lK(x, y)), \tag{3.12}$$

for all $x, y \in X$, where $l \geq 0$ and

$$\begin{aligned} J(x, y) &= \max\Big\{d(x, y), p_\alpha(x, Qx), p_\alpha(y, Qy), \frac{p_\alpha(x, Qy) + p_\alpha(y, Qx)}{2}\Big\}, \\ K(x, y) &= \min\big\{p_\alpha(x, Qx), p_\alpha(y, Qy), p_\alpha(x, Qy), p_\alpha(y, Qx)\big\}. \end{aligned}$$

If the following condition holds :

(i) Q is β_M-admissible,

(ii) there exist $x_0 \in X$ and $x_1 \in [Qx_0]_\alpha$ such that $\beta(x_0, x_1) \geq 1$,

 ISBN: 978-84-617-8694-7

(iii) if $\{x_n\}$ is sequence in X such that $\beta(x_n, x_{n+1}) \geq 1$ and $x_n \to x$ as $n \to \infty$, then $\beta(x_n, x) \geq 1$,

then there exists $x \in X$ such that x_α is a fuzzy fixed point of Q.

If we set $\beta(x, y) = 1$ for all $x, y \in X$ in Corollary 3.4, we get the following result.

Corollary 3.7 *Let (X, d) be a complete metric space, $\alpha \in [0, 1]$ and T be fuzzy mapping from X to $W_\alpha(X)$. Suppose that there exist $\psi \in \Psi$ such that*

$$D_\alpha(Qx, Qy) \leq \psi(J(x, y) + lK(x, y)), \tag{3.13}$$

for all $x, y \in X$, where $l \geq 0$ and

$$\begin{aligned} J(x, y) &= \max\left\{d(x, y), p_\alpha(x, Qx), p_\alpha(y, Qy), \frac{p_\alpha(x, Qy) + p_\alpha(y, Qx)}{2}\right\}, \\ K(x, y) &= \min\left\{p_\alpha(x, Qx), p_\alpha(y, Qy), p_\alpha(x, Qy), p_\alpha(y, Qx)\right\}. \end{aligned}$$

Then there exists $x \in X$ such that x_α is a fuzzy fixed point of T.

If $\psi(t) = \theta t$ where $\theta \in (0, 1)$, $l = 0$ and $\beta(x, y) = 1$ for all $x, y \in X$ in Corollary 3.4, , then we have the following Corollary.

Corollary 3.8 *Let (X, d) be a complete metric linear space, $\alpha \in [0, 1]$ and T be fuzzy mapping from X to $\mathcal{W}_\alpha(X)$ such that*

$$D_\alpha(Qx, Qy) \leq k \max\left\{d(x, y), p_\alpha(x, Qx), p_\alpha(y, Qy), \frac{p_\alpha(x, Qy) + p_\alpha(y, Qx)}{2}\right\} \tag{3.14}$$

for all $x, y \in X$, where $k \in (0, 1)$. Then there exists $x \in X$ such that x_α is a fuzzy fixed point of Q.

Acknowledgements

This research is supported by Rajamangala University of Technology Rattanakosin Research and Development Institute, Rajamangala University of Technology Rattanakosin(RMUTR).

References

[1] L. A. Zadeh, *Fuzzy sets*, Informations and Control **8** (1965) 103–112.

[2] S. Heilpern, *Fuzzy mappings and fixed point theorems*, J. Math. Anal. Appl. **83** (1981) 566–569.

[3] S.B. Nadler Jr., *Multivalued Contraction Mapping*, Pacific Journal of Mathematics **30(2)** (1969) 475–488.

[4] V.D. Estruch and A. Vidal, *A note on fixed fuzzy points for fuzzy mappings*, Rend Istit Univ Trieste **32** (2001) 39–45.

[5] A. Azam and I. Beg, *Common fixed points of fuzzy maps*, Mathematical and Computer Modelling **49** (2009) 1331–1336.

[6] D. Turkoglu and B. E. Rhoades, *A fixed fuzzy point for fuzzy mapping in complete metric spaces*, Mathematical Communications **10** (2005) 115–121.

[7] S. Sedghi, N. Shobe and I. Altun, *A fixed fuzzy point for fuzzy mappings in complete metric space*, Mathematical Communications **13** (2008) 289–294.

[8] S. Phiangsungnoen, W. Sintunavarat and P. Kumam, *Common α-fuzzy fixed point theorems for fuzzy mappings via $\beta_{\mathcal{F}}$-admissible pair*, Journal of Intelligent and Fuzzy Systems **27(5)** (2014) 2463–2472.

[9] S. Phiangsungnoen, W. Sintunavarat and P. Kumam, *Fuzzy fixed point theorems in Hausdorff fuzzy metric spaces*, Journal of Inequalities and Applications **2014** (2014): 201.

[10] S. Phiangsungnoen, W. Sintunavarat and P. Kumam, *Fuzzy fixed point theorems for fuzzy mappings via β-admissible with applications*, Journal of Uncertainty Analysis and Applications **2014** (2014): 2: 20.

[11] M. Abbas and D. Turkoglu, *Fixed point theorem for a generalized contractive fuzzy mapping*, Journal of Intelligent and Fuzzy Systems **26(1)** (2014) 33–36.

[12] A. Azam, S. Hussain and M. Arshad, *Common fixed points of Chatterjea type fuzzy mappings on closed balls*, Neural Computing and Applications **21(1)** (2012) 313–317.

[13] J. H. Asl, S. Rezapour, and N. Shahzad, *On fixed points of α-ψ-contractive multifunctions*, Fixed Point Theory and Applications **2012** (2012): 212.

[14] B. Mohammadi, S. Rezapour and N. Shahzad, *Some results on fixed points of α-ψ-Ciric generalized multifunctions*, Fixed Point Theory and Applications **2013** (2013): 24.

[15] P. M. Pu and Y. M. Liu, *Fuzzy Topology. I. Neighborhood Structure of a Fuzzy Point and Moore-Smith Convergence*, J. Math. Anal. Appl. **76** (1980) 571–599.

[16] M. U. Ali, T. Kamran and E. Karapnar, *(α, ψ, ξ)-contractive multivalued mappings*, Fixed Point Theory and Applications **2014** (2014): 7.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

A game theoretical analysis in a rumor spreading model based on the SIR epidemic model

Alberto Pinto[1] **and José Martins**[2]

[1] *LIAAD-INESC TEC and Department of Mathematics, School of Technology and Management, Polytechnic Institute of Leiria, Portugal*

[2] *LIAAD-INESC TEC and Department of Mathematics, Faculty of Sciences, University of Porto, Portugal*

emails: aapinto@fc.up.pt, jmmartins@ipleiria.pt

Abstract

Based on the classical epidemiological SIR model, we propose a similar model to analyze the spreading of a false rumor in an homogeneous mixing population. The individuals of the population can be ignorants to a certain rumor, or they can either believe or unbelieve on the rumor, depending on the level of the knowledge about the rumor achieved. Hence, an individual can decide to search for real information, or not, in order to be informed and believe, or not, on the rumor. This search for information can have costs but can also be very advantageous to the individual. Hence, we introduce the expected learning payoff of an individual to find the his/her best learning strategy.

Key words: Rumor, SIR model, Nash

1 Introduction

The first mathematical model for rumor spreading was proposed by Daley and Kendall in 1964 (see [1]), and became known as the D-K model. In this model, the population is divided in three groups: ignorants - people who are ignorant to the rumor; spreaders - people who are actively spreading the rumor; stiflers - people who have heard the rumor, but are no longer interested in spreading it. Even before Daley and Kendall's work, Goffman and Newill (see [3]) published a paper generalizing the epidemics theory, with a clear analogy between the spreading of an infectious disease and the transmission of ideas. Since then, several authors developed the rumor spreading models, with new transitions capable of describing

 ISBN: 978-84-617-8694-7

different issues in the spreading process (see [2, 8, 10]). The analogy between epidemics and rumors states that: people can be *infective*, those who are host to the infectious material, as *believers and spreaders* of a certain rumor; people can be *susceptible*, those who can be infective given a contact with infectious material, as *ignorants*, those who can be believers given a contact with the rumor provided by a believer; and people can be *removals*, those who have been removed by death, immunization, hospitalization, etc., as *stiflers*, those who do not spread the rumor any more.

A rumor is defined as an information disseminated with neither a official confirmation nor an official refutation, but, typically, it has a negative impact on the society (see [10]). A rumor can appear in one of many forms: it can be a scientific result that as been published in a certain paper; it can be the idea that some food has cancer-causing compounds; it can be the idea that a bank is near the bankruptcy; etc.. Since the rumors have the power to provoke high changes in population behaviors, sometimes causing even panic for society, there are always some risks associated with them, especially when they are false. Nowadays, with the social media and all the technologies of information, a rumor can easily spread and reach larges scales of individuals. Hence, the study about how a rumor spreads, done to prevent or minimize its negative effects, is an extremely important task.

In this work, we study the spread of false rumors using a mathematical model based on the classical SIR epidemic model. Instead of being suscepitbles, infecteds or removed, in our model the individuals can be either *ignorants*, *believers* or *unbelievers* to a certain rumor. The ignorants are those who have not heard the rumor yet, while the believers are the individuals that already heard, believe and spread the rumor. The unbelievers are the individuals that do not believe on the rumor by one of the following two reasons: they believed before but no longer believe by absence of proofs; or, they are well informed and they do not believe on a false rumor because they know the reality. Obviously, the unbeliever individuals do not spread the rumor.

Depending on the belief on the rumor, people can have different attitudes. Individuals that believe on false rumors can easily take some wrong decisions, with possible negative consequences. Hence, being informed is a necessary condition to do not believe on false rumors and avoid wrong decisions. On the other hand, the access to information might have significant costs and sometimes not being informed is the best solution. Facing the benefits of being informed against the costs of accessing the information, we will make a game theoretical analysis (see [5]) about the individuals' decisions regarding the search, or not, for real information. We introduce the individual learning payoff, that can be used to compute the most profitable strategies.

 ISBN: 978-84-617-8694-7

2 The spreading model

Based on the classical SIR epidemiological model, we purpose a similar model to describe the spread of a rumor in a homogeneous population.

In the classical SIR epidemiological model, proposed originally by Kermack and McKendrik in 1927 (see [6]), the population is assumed to be divided in three distinct compartments: S - the susceptible individuals; I - the infecteds; and R - the individuals that have recovered from infection and are now immune. The reaction schemes for transitions of the SIR system are given by

$$\begin{aligned} S+I &\xrightarrow{\beta} I+I \\ I &\xrightarrow{\gamma} R \\ S,I,R &\xrightarrow{\mu} S \end{aligned}$$

where: β is the mean infection rate; γ is the mean recovery rate and so $1/\gamma$ is the mean infectious period; and μ is the mean birth and death rate, hence $1/\mu$ is the mean life expectancy at birth. The reaction schemes of the SIR model result in the following ODE system

$$\begin{aligned} \frac{dS}{dt} &= \mu - \beta SI - \mu S \\ \frac{dI}{dt} &= \beta SI - \mu I - \gamma I \\ \frac{dR}{dt} &= \gamma I - \mu R\,. \end{aligned}$$

In the following years, several authors used the SIR model to develop more complex epidemic models, in which temporary and/or partial immune protection were considered (see [4, 9]). The introduction of a vaccination transition from susceptible to recovered

$$S \xrightarrow{\nu} R$$

was also explored (see [4]). Besides the epidemiology, the SIR epidemic model has been used to describe the evolution of several phenomena in many other fields, like informatics and sociology. Based on the SIR model, we propose a rumor spreading model where individuals can be either *Ignorant*, *Believers* or *Unbelievers* on a certain rumor. This model will be designated by *IBU model*. A direct analogy between the SIR model and the IBU model can be done by taking the following identifications on the state variables:

- S - *Susceptibles* are the I - *Ignorants*;
- I - *Infecteds* are the B - *Believers* on the rumor;

 ISBN: 978-84-617-8694-7

- R - *Recovered* are the U - *Unbelievers* on the rumor.

We will make the following assumptions about the IBU rumor spreading model: (i) all individuals are distributed homogeneously in the population and have the same possibilities to access information, hence they all make decisions under the same conditions; (ii) the believer individuals are the active spreaders, i.e. they are the individuals that transmit the rumor to the ignorants; (iii) once a believer stops believing on the rumor and becomes an unbeliever, he/she will also stops transmitting the rumor. Hence, the unbelievers are not active spreaders. Similarly to the vaccination in the SIR epidemic model, the IBU model comprises a learning transition that puts the ignorants automatically unbelievers, because an individual becomes informed about the reality after learning and do not believe on false rumors.

To find the learning strategies that are more likely to be adopted by an ignorant individual we will define the Nash equilibria strategies (see [5, 7]). We start by assuming that: P is the probability that an individual will choose to be informed, i.e. it is the individual's learning strategy; and p is the proportion of individuals who will choose to learn. With this notation, P^* is a population Nash learning strategy if

$$\Delta_{P^*\to Q} = E(Q, P^*) - E(P^*, P^*) \leq 0 \ , \quad \text{for every strategy} \quad Q \in [0, 1],$$

where $E(P, p)$ denotes the learning expected payoff of an individual that learns with probability P. Hence, if P^* is a population Nash learning strategy then no single individual has the incentive to deviate his/her learning strategy from P^* to any other strategy $Q \neq P^*$.

The learning expected payoff will be studied depending on the benefits and the costs of learning but also depending on the cost of believing on a false rumor. Maximizing the learning expected payoff, we will characterize the most profitable learning strategies.

3 Conclusions and future work

Based on the SIR epidemic model, we propose a rumor spreading model to describe the spread of false rumors. In this model, individuals can be either ignorants, believers or unbelievers on the rumor. Depending on the costs of believing on false rumors, and depending on the benefits and costs of being informed, we introduce the individual learning expected payoff. This expected payoff can now be used to compute the Nash learning strategies that are more likely to be adopted. Future work can also include the characterization of the evolutionary stable learning strategies in the IBU rumor spreading model or, possibly, in models where the rumor do not need to be false.

 ISBN: 978-84-617-8694-7

Acknowledgements

The authors thank the financial support of LIAAD-INESC TEC and FCT Fundação para a Ciência e Tecnologia (Portuguese Foundation for Science and Technology) within project UID/EEA/50014/2013 and ERDF (European Regional Development Fund) through the COMPETE Program (operational program for competitiveness) and by National Funds through the FCT within Project "Dynamics, optimization and modelling", with reference PTDC/MAT-NAN/6890/2014 and Project "NanoSTIMA: Macro-to-Nano Human Sensing: Towards Integrated Multimodal Health Monitoring and Analytics/NORTE-01-0145-FEDER-000016" financed by the North Portugal Regional Operational Programme (NORTE 2020), under the PORTUGAL 2020 Partnership Agreement, and through the European Regional Development Fund (ERDF).

References

[1] D. J. Daley and D. G. Kendall, *Epidemics and rumours*, Nature **204** (1964) 1118.

[2] D. J. Daley and D.G. Kendall, *Stochastic rumours*, IMA Journal of Applied Mathematics **1** (1965) 4255.

[3] W. Goffman and V. A. Newill, *Generalization of epidemic theory: an application to the transmission of ideas*, Nature, **204** (1964) 225228.

[4] M.G.M. Gomes, L.J. White and G.F. Medley, *Infection, reinfection, and vaccination under suboptimal immune protection: epidemiological perspectives* J. Theor. Biol., **228** (2004) 539-549.

[5] J. Hofbauer and K. Sigmund, *Evolutionary Games and Population Dynamics*, Cambridge University Press, 1998.

[6] W.O. Kermack and A.G. McKendrick, *A contribution to the mathematical theory of epidemics*, Proc. Roy. Soc. Lond., **115** (1927) 700-721.

[7] J. Martins and A. Pinto, *Bistability of Evolutionary Stable Vaccination Strategies in the Reinfection SIRI Model*, Bull. Math. Biol., **79** (2017) 853-883.

[8] Y. Moreno, M. Nekovee and A. F. Pacheco, *Dynamics of rumor spreading in complex networks*, Phys. Rev. E **69** (2004) 066130.

[9] N. Stollenwerk, S. van Noort, J. Martins, M. Aguiar, F. Hilker, A. Pinto and M.G. Gomes, *A spatially stochastic epidemic model with partial immunization shows in mean field approximation the reinfection threshold*, J. Biol. Dyn., **4** (2010) 634649.

 ISBN: 978-84-617-8694-7

[10] N. Zhang, H. Huang, M. Duarte and J. Zhang, *Risk analysis for rumor propagation in metropolises based on improved 8-state ICSAR model and dynamic personal activity trajectories*, Physica A, **451** (2016) 403-419.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

High order in space and time discretization for the numerical solution of anisotropic wave equations

A. M. Portillo[1]

[1] *Departamento de Matemática Aplicada, Universidad de Valladolid, Spain*

emails: anapor@mat.uva.es

Abstract

Two-dimensional wave equation in anisotropic media, on a rectangular domain with initial conditions and periodic boundary conditions, is considered. The space discretization is done with second and forth order finite differences on a uniform grid, paying attention to the mixed derivative of the equation. For the time integration of the system of ordinary differential equations obtained, a fourth order exponential splitting method, which is a geometric integrator, is proposed. The stability condition for time step and space step ratio is deduced. Numerical experiments comparing the discrete energy of the semi-discrete problem with the energy of the continuous problem are showed. Experiments displaying the good behavior in the long time integration and the efficiency of the numerical solution are too provided.

Key words: Anisotropic, Mixed derivative term, Energy, Finite differences, Splitting method
MSC 2000: 65M12, 65M20

1 Introduction

We study the two dimensional time-dependent anisotropic and dispersive wave equation

$$\partial_{tt}u = \alpha_{11}\partial_{xx}u + 2\alpha_{12}\partial_{xy}u + \alpha_{22}\partial_{yy}u - s^2u, \tag{1}$$

in a rectangular domain $R = [a, b] \times [c, d]$, for the unknown $u(x, y, t)$, with initial-boundary conditions. We asume that the coefficients of the anisotropic wave equation (1) α_{ij} and s^2 are constant and

$$\alpha_{11} > 0,\ \alpha_{22} > 0,\ \alpha_{11}\alpha_{22} - \alpha_{12}^2 > 0, \tag{2}$$

 ISBN: 978-84-617-8694-7

so that in the steady state the equation is elliptic.

We impose periodic boundary conditions,

$$\begin{aligned} u(a,y,t) &= u(b,y,t), \qquad y \in [c,d], && (3)\\ \partial_x u(a,y,t) &= \partial_x u(b,y,t), \quad y \in [c,d], && (4)\\ u(x,c,t) &= u(x,d,t), \qquad x \in [a,b], && (5)\\ \partial_y u(x,c,t) &= \partial_y u(x,d,t), \quad x \in [a,b]. && (6) \end{aligned}$$

In addition, we consider the initial conditions,

$$u(x,y,0) = u_0(x,y), \quad \partial_t u(x,y,0) = v_0(x,y), \tag{7}$$

which satisfy periodic boundary conditions in R.

2 Energy of the continuous problem

An energy,

$$\begin{aligned} E(t) &= \frac{1}{2}\iint_R ((\partial_t u(x,y,t))^2 + \alpha_{11}(\partial_x u(x,y,t))^2 + 2\alpha_{12}\partial_x u(x,y,t)\partial_y u(x,y,t)\\ &+ \alpha_{22}(\partial_y u(x,y,t))^2 + s^2 u(x,y,t)^2)dxdy, \end{aligned}$$

can be introduced. The energy $E(t)$ is constant with time

$$\begin{aligned} E(t) = E(0) = \frac{1}{2}\iint_R (v_0(x,y)^2 + \alpha_{11}(\partial_x u_0(x,y))^2 + 2\alpha_{12}\partial_x u_0(x,y)\partial_y u_0(x,y)\\ +\alpha_{22}(\partial_y u_0(x,y))^2 + s^2 u_0(x,y)^2)dxdy, \end{aligned} \tag{8}$$

and, in this way, we can compute the energy of the problem calculating the initial energy through the initial conditions.

3 Spatial discretization

We start approximating the spatial derivatives in (1) by using finite differences. For the sake of simplicity, we consider the same size step in both directions x and y, that is, for a value of N, $h = \dfrac{b-a}{N}$ and $M = \dfrac{d-c}{h}$. Let $x_j = a + (j-1)h$, $j = 1, \ldots, N+1$, and $y_l = c + (l-1)h$, $l = 1, \ldots, M+1$, be the nodes of the spatial discretization. This produces a uniform grid in the computational domain and a matrix of unknowns $u_{jl}(t) = u(x_j, y_l, t)$.

 ISBN: 978-84-617-8694-7

3.1 Second order spatial discretization

Second order spatial derivatives in the direction x and in the direction y are approximated by second order central finite differences

$$\begin{aligned} \partial_{xx}\, u_{jl} &\approx \frac{u_{j-1,l} - 2u_{jl} + u_{j+1,l}}{h^2}, \\ \partial_{yy}\, u_{jl} &\approx \frac{u_{j,l-1} - 2u_{jl} + u_{j,l+1}}{h^2}, \end{aligned}$$

in stencil form

$$\frac{1}{h^2}\begin{pmatrix} 0 & 0 & 0 \\ 1 & -2 & 1 \\ 0 & 0 & 0 \end{pmatrix}, \quad \frac{1}{h^2}\begin{pmatrix} 0 & 1 & 0 \\ 0 & -2 & 0 \\ 0 & 1 & 0 \end{pmatrix},$$

respectively.

For second order mixed derivatives we consider second order approximation with stencil forms

$$\frac{1}{2h^2}\begin{pmatrix} -1 & 1 & 0 \\ 1 & -2 & 1 \\ 0 & 1 & -1 \end{pmatrix}, \quad \frac{1}{2h^2}\begin{pmatrix} 0 & -1 & 1 \\ -1 & 2 & -1 \\ 1 & -1 & 0 \end{pmatrix}. \tag{9}$$

The first stencil form should be used in the case $\alpha_{12} > 0$, whereas the second is suitable when $\alpha_{12} < 0$. Taking this into account, here mixed derivative is approximated as follows

$$\begin{aligned} &\text{if} \quad \alpha_{12} > 0, \\ &\qquad 2\partial_{xy}\, u_{jl} \approx \frac{u_{j-1,l} + u_{j+1,l} - 2u_{jl} + u_{j,l-1} + u_{j,l+1} - u_{j-1,l+1} - u_{j+1,l-1}}{h^2}, \\ &\text{if} \quad \alpha_{12} < 0, \\ &\qquad 2\partial_{xy}\, u_{jl} \approx \frac{-u_{j-1,l} - u_{j+1,l} + 2u_{jl} - u_{j,l-1} - u_{j,l+1} + u_{j-1,l-1} + u_{j+1,l+1}}{h^2}. \end{aligned}$$

Let it be $\mathbf{u}_j$ the approximations to the unknowns $(u(x_j, y_1), \ldots, u(x_j, y_{M+1}))^T$, for fixed x_j. Denoting $\mathbf{u}_h = [\mathbf{u}_1^T, \ldots, \mathbf{u}_{N+1}^T]^T$, we achieve the second order in time ODEs system

$$\frac{d^2\mathbf{u}_h}{dt^2} = A\mathbf{u}_h, \tag{10}$$

where the matrix of the problem is

$$A = \frac{1}{h^2}B - s^2 I, \tag{11}$$

I is the identity matrix of dimension $(N+1)(M+1)$ and matrix B comes from second order spatial derivatives approximation.

 ISBN: 978-84-617-8694-7

Lemma 1. *The eigenvalues of matrix B, for the coefficients α_{ij} meeting (2), satisfy*

$$\sigma(B) \subset [-4(\alpha_{11}+\alpha_{22}) - 8|\alpha_{12}|, 0].$$

Lemma 2. *The matrix*

$$A = \frac{1}{h^2}B - s^2 I$$

is symmetric negative definite.

Theorem 3. *The discrete energy*

$$E_h(t)(\mathbf{u},\mathbf{v}) \;=\; \frac{h^2}{2}(\mathbf{v}^T\mathbf{v} - \mathbf{u}^T A\mathbf{u}), \tag{12}$$

is conserved for $(\mathbf{u}_h, d\mathbf{u}_h/dt)$, being $\mathbf{u}_h$ the solution of (10).

3.2 Fourth order spatial discretization

Computational cost becomes especially important when the number of equations in the system increases. In order to approach equation (1) with periodic boundary conditions (3)-(6) and initial conditions (7), with higher accuracy, we consider fourth order approximation of the spatial derivatives.

Second order spatial derivatives in the direction x and in the direction y are approximated by fourth order central finite differences

$$\begin{aligned}
\partial_{xx}\, u_{jl} &\approx \frac{1}{h^2}(-\frac{1}{12}u_{j-2,l} + \frac{4}{3}u_{j-1,l} - \frac{5}{2}u_{jl} + \frac{4}{3}u_{j+1,l} - \frac{1}{12}u_{j+2,l}),\\
\partial_{yy}\, u_{jl} &\approx \frac{1}{h^2}(-\frac{1}{12}u_{j,l-2} + \frac{4}{3}u_{j,l-1} - \frac{5}{2}u_{jl} + \frac{4}{3}u_{j,l+1} - \frac{1}{12}u_{j,l+2}),
\end{aligned}$$

in stencil form

$$\frac{1}{12h^2}\begin{pmatrix} 0 & 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0 & 0\\ -1 & 16 & -30 & 16 & -1\\ 0 & 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}, \quad \frac{1}{12h^2}\begin{pmatrix} 0 & 0 & -1 & 0 & 0\\ 0 & 0 & 16 & 0 & 0\\ 0 & 0 & -30 & 0 & 0\\ 0 & 0 & 16 & 0 & 0\\ 0 & 0 & -1 & 0 & 0 \end{pmatrix},$$

respectively.

Mixed derivative are approximated by the following fourth order finite differences

$$\begin{aligned}
\partial_{xy}\, u_{jl} \approx \frac{1}{144h^2}\;& (u_{j-2,l-2} - 8u_{j-2,l-1} + 8u_{j-2,l+1} - u_{j-2,l+2}\\
&-8u_{j-1,l-2} + 64u_{j-1,l-1} - 64u_{j-1,l+1} + 8u_{j-1,l+2}\\
&8u_{j+1,l-2} - 64u_{j+1,l-1} + 64u_{j+1,l+1} - 8u_{j+1,l+2}\\
&-u_{j+2,l-2} + 8u_{j+2,l-1} - 8u_{j+2,l+1} + u_{j+2,l+2}),
\end{aligned}$$

 ISBN: 978-84-617-8694-7

in stencil form

$$\frac{1}{144h^2}\begin{pmatrix} -1 & 8 & 0 & -8 & 1 \\ 8 & -64 & 0 & 64 & -8 \\ 0 & 0 & 0 & 0 & 0 \\ -8 & 64 & 0 & -64 & 8 \\ 1 & -8 & 0 & 8 & -1 \end{pmatrix}.$$

4 Time discretization

We rewrite problem (10) as a first order system, naming $\mathbf{v}_h = [\frac{d}{dt}\mathbf{u}_1^T, \ldots, \frac{d}{dt}\mathbf{u}_{N+1}^T]^T$,

$$\frac{d}{dt}\begin{bmatrix} \mathbf{u}_h \\ \mathbf{v}_h \end{bmatrix} = \begin{bmatrix} 0 & I \\ A & 0 \end{bmatrix}\begin{bmatrix} \mathbf{u}_h \\ \mathbf{v}_h \end{bmatrix}, \tag{13}$$

where I is the identity matrix of dimension $(N+1)(M+1)$. We notice system (13) is a Hamiltonian problem.

4.1 Exponential splitting method

Denoting k the time step, we propose to approximate the exact solution of (13),

$$\begin{bmatrix} \mathbf{u}(t+k) \\ \mathbf{v}(t+k) \end{bmatrix} = \exp\left(k\begin{bmatrix} 0 & I \\ A & 0 \end{bmatrix}\right)\begin{bmatrix} \mathbf{u}(t) \\ \mathbf{v}(t) \end{bmatrix}, t \geq 0,$$

by using an exponential splitting method as time integrator. We split the matrix of the problem (13) in two parts

$$\begin{bmatrix} 0 & I \\ A & 0 \end{bmatrix} = \begin{bmatrix} 0 & I \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ A & 0 \end{bmatrix} = M_1 + M_2.$$

The intermediate problems

$$\frac{d}{dt}\begin{bmatrix} \mathbf{u}_h \\ \mathbf{v}_h \end{bmatrix} = M_i\begin{bmatrix} \mathbf{u}_h \\ \mathbf{v}_h \end{bmatrix}, \quad i = 1, 2,$$

can be solved exactly using that $M_i^2 = 0$ for $i = 1, 2$ and,

$$\exp(kM_1) = \begin{bmatrix} I & kI \\ 0 & I \end{bmatrix}, \quad \exp(kM_2) = \begin{bmatrix} I & 0 \\ kA & I \end{bmatrix}.$$

Then, the flows of these intermediate problems applied to $[\mathbf{u}, \mathbf{v}]^T$ are

$$\begin{aligned} \psi_k^{[1]}: \quad \exp(kM_1)\begin{bmatrix} \mathbf{u} \\ \mathbf{v} \end{bmatrix} &= \begin{bmatrix} \mathbf{u} + k\mathbf{v} \\ \mathbf{v} \end{bmatrix}, \\ \psi_k^{[2]}: \quad \exp(kM_2)\begin{bmatrix} \mathbf{u} \\ \mathbf{v} \end{bmatrix} &= \begin{bmatrix} \mathbf{u} \\ \mathbf{v} + kA\mathbf{u} \end{bmatrix}. \end{aligned}$$

 ISBN: 978-84-617-8694-7

To advance a step of size k in time, once we have solved exactly each intermediate step, it is necessary combining these solutions to obtain an approximation of the solution of (13). To do this, first we use the symmetric second order Strang splitting $\mathcal{S}^{[2]}$

$$\mathcal{S}_k^{[2]} = \psi_{k/2}^{[1]} \circ \psi_k^{[2]} \circ \psi_{k/2}^{[1]}, \tag{14}$$

and then, by composition of $\mathcal{S}^{[2]}$, we consider the fourth order integrator $\mathcal{S}^{[4]}$ [8, 9]

$$\mathcal{S}_k^{[4]} = \mathcal{S}_{\alpha k}^{[2]} \circ \mathcal{S}_{\beta k}^{[2]} \circ \mathcal{S}_{\alpha k}^{[2]}, \quad \text{with} \quad \alpha = \frac{1}{2 - 2^{1/3}}, \quad \beta = 1 - 2\alpha. \tag{15}$$

The advantage of composing exact solutions in this way is that geometric properties of the true flow are preserved. Symplectic time integrators [4, 7] not only provides better qualitative properties of the numerical solution, but also better accuracy when a long time computation is made.

It is possible to save some computational cost writing (15) as

$$\begin{aligned} \mathcal{S}_k^{[4]} &= \psi_{\alpha k/2}^{[1]} \circ \psi_{\alpha k}^{[2]} \circ \psi_{\alpha k/2}^{[1]} \circ \psi_{\beta k/2}^{[1]} \circ \psi_{\beta k}^{[2]} \circ \psi_{\beta k/2}^{[1]} \circ \psi_{\alpha k/2}^{[1]} \circ \psi_{\alpha k}^{[2]} \circ \psi_{\alpha k/2}^{[1]}, \\ &= \psi_{\alpha k/2}^{[1]} \circ \psi_{\alpha k}^{[2]} \circ \psi_{(\alpha+\beta) k/2}^{[1]} \circ \psi_{\beta k}^{[2]} \circ \psi_{(\alpha+\beta) k/2}^{[1]} \circ \psi_{\alpha k}^{[2]} \circ \psi_{\alpha k/2}^{[1]}. \end{aligned} \tag{16}$$

This splitting method is explicit and it is easy to implement. However, it is not unconditionally stable and the stability has to be studied.

Theorem 4. *The value of ω_* in the stability interval of (16) is*

$$\omega_* = \sqrt{\frac{-1 + \sqrt{1 + 1152\gamma}}{48\gamma}}, \quad \gamma = \frac{6 + 5 \cdot 2^{1/3} + 4 \cdot 2^{2/3}}{36}.$$

Since the eigenvalues of $k(-A)^{1/2} = \frac{k}{h}(-B + s^2h^2 I)^{1/2}$ must be in the *stability interval*, we obtain the stability condition

$$\frac{k}{h}\sqrt{4(\alpha_{11} + \alpha_{22}) + 8|\alpha_{12}| + s^2h^2} < \sqrt{\frac{-1 + \sqrt{1 + 1152\gamma}}{48\gamma}}.$$

This will be reached, for sh small enough, when

$$\frac{k}{h} < \frac{0.9711}{\sqrt{4(\alpha_{11} + \alpha_{22}) + 8|\alpha_{12}|}}. \tag{17}$$

 ISBN: 978-84-617-8694-7

5 Numerical experiments

In this Section we consider the problem described in Section 1 with initial conditions

$$u_0(x,y) = \begin{cases} \dfrac{(x+0.2)^3(0.2-x)^3(y+0.2)^3(0.2-y)^3}{(0.2)^{12}}, & -0.2 < x, y < 0.2, \\ 0, & \text{otherwise}, \end{cases}$$

and $v_0(x,y) = 0$, with compact support contained in the computational domain $[-1/4, 1/4] \times [-1/4, 1/4]$. The polynomial in u_0 is chosen so that $u_0 \in C^1([-1/4, 1/4] \times [-1/4, 1/4])$.

We set the dispersion coefficient $s^2 = 1$. For the numerical experiments we have selected three cases of coefficients α_{ij} from [3]. Table 1 displays these coefficients with the same notation used in [3]. We have numerically computed the eigenvalues of the corresponding

Run	α_{11}	α_{22}	α_{12}
1.2	0.875	0.625	-0.217
2.2	0.160	0.940	0.225
2.4	0.472	0.628	0.443

Table 1: The three cases of α_{ij} considered.

matrix B for Section 3.2, for the α_{ij} considered, and we can conclude that they are non positive real numbers.

Table 2 displays the ratio of stability between the time step and the space step. It can

Run	Second order FD	Fourth order FD
1.2	0.3491	0.3434
2.2	0.3900	0.4009
2.4	0.3445	0.3974

Table 2: Ratio of stability.

be seen in Table 2 that the stability condition for the splitting method is acceptable.

Now, we are going to compare the continuous energy (8) for the test problem with the discrete energy

$$E_h(t)(\mathbf{u}, \mathbf{v}) = \frac{h^2}{2}(\mathbf{v}^T\mathbf{v} - \mathbf{u}^T A\mathbf{u}),$$

of the semi-discrete problems. We denote by $E_{h,2}(t)$ the discrete energy where matrix A is the matrix obtained when second order finite differences are used, and $E_{h,4}(t)$ the

 ISBN: 978-84-617-8694-7

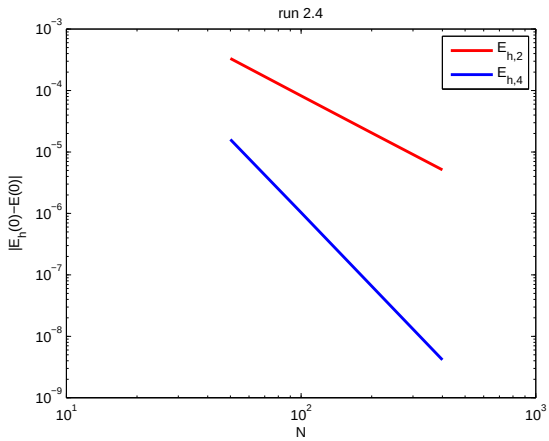


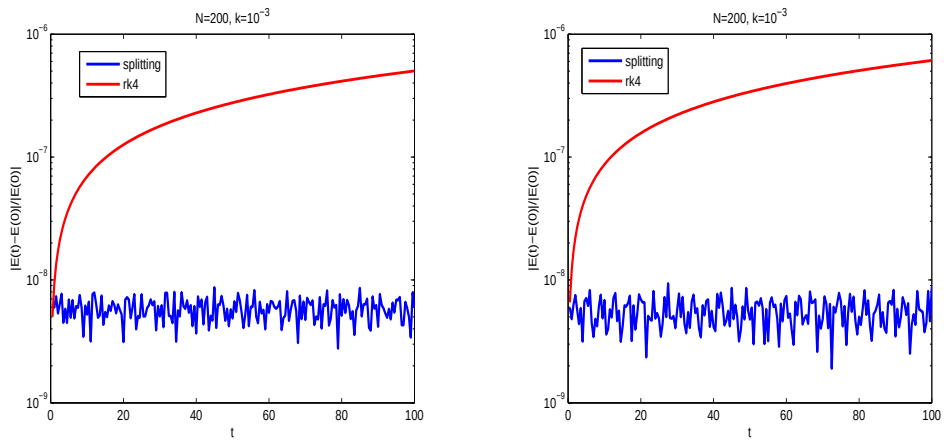
Figure 1: Energy error for the second order finite differences and the fourth-order finite differences for run 2.4.

discrete energy where matrix A is the matrix obtained when fourth order finite differences are considered. Energy error for the second order finite differences and the fourth-order finite differences for run 2.4 are shown in Figure 1. It can be appreciated the second and fourth order of the discretization of Subsections 3.1 and 3.2, respectively.

Finally, in the following experiments we compare the behavior of the splitting scheme and the fourth-order four-stage Runge-Kutta method when fourth order finite differences introduced in Section 3.2 and the energy norm $E_{h,4}(t)$ are considered. We measure the relative energy error $|E_{h,4}(t) - E_{h,4}(0)|/|E_{h,4}(0)|$. We set $N = M = 200$ and $k = 10^{-3}$. Figure 2 displays relative energy error for the exponential splitting integrator and the fourth-order four-stage Runge-Kutta method, for times from 0 to 100, for run 1.2 and run 2.2. The splitting method maintains the same size error throughout the interval of time $[0, 100]$. This agrees with the fact that scheme (16) is a geometric integrator. Whereas for the Runge-Kutta method the size of the error grows when the time increases.

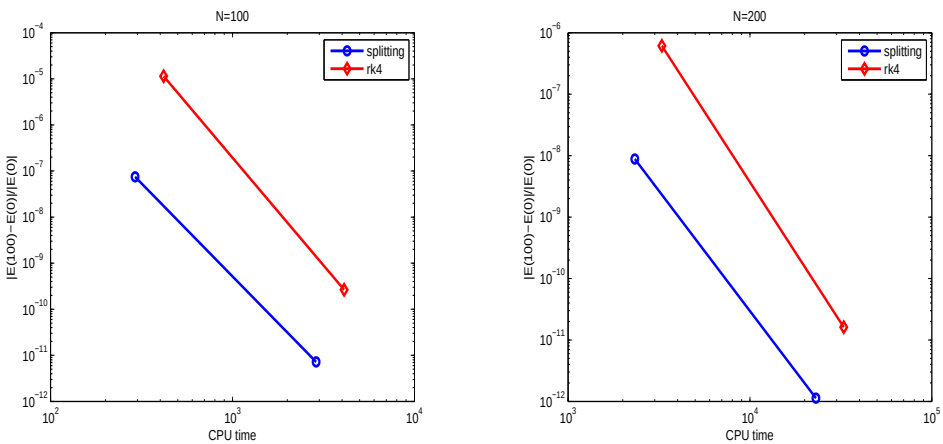
Lastly, we study the efficiency of the splitting scheme by comparing with the fourth-order four-stage Runge-Kutta method measuring the computational cost in terms of CPU time. For the exponential splitting integrator, if the last step in the composition (16) of $\mathcal{S}_k^{[4]}$ for one step and the first one in $\mathcal{S}_k^{[4]}$ for the next step are joined together, that is, $\psi_{\alpha k/2}^{[1]} \circ \psi_{\alpha k/2}^{[1]} = \psi_{\alpha k}^{[1]}$, only three times of step 1 are needed for each step in time. A similar analysis of the efficiency of the algorithms to the one done in [1] can be done here. Then, regarding the products required, for the Runge-Kutta method and the splitting method, the relation is four to three. It can be seen in Figure 3 that the splitting method is better than the Runge-Kutta method. For the same error the computational cost is smaller.

 ISBN: 978-84-617-8694-7



(a) run 1.2 (b) run 2.2

Figure 2: Relative energy error for the exponential splitting integrator and the fourth-order four-stage Runge-Kutta method, for run 1.2 and run 2.2.



(a) $N = 100$ (b) $N = 200$

Figure 3: Relative energy error at $T = 100$ versus CPU time for the exponential splitting integrator and the fourth-order four-stage Runge-Kutta method for run 2.2, for $N = 100$ and $N = 200$.

 ISBN: 978-84-617-8694-7

The numerical experiments confirm the good behavior in the long time integration and the efficiency of the splitting method considered.

Acknowledgements

This work has obtained financial support from project MTM2015-66837-P of Ministerio de Economía y Competitividad.

References

[1] I. Alonso-Mallo, A. M. Portillo, *Time exponential splitting technique for the KleinGordon equation with HagstromWarburton high-order absorbing boundary conditions*, J. Comput. Phys. **311** (2016) 196-212.

[2] S. Blanes, F. Casas, A. Murua, *On the Linear Stability of Splitting Methods*, Found. Comp. Math. **8** (2008) 357–393.

[3] E. Bécache, D. Givoli, T. Hagstrom, *High-order Absorbing Boundary Conditions for anisotropic and convective wave equations*, J. Comput. Phys. **229** (2010) 1099–1129.

[4] E. Hairer, G. Wanner, C. Lubich, *Geometric Numerical Integration*, Springer, 2006.

[5] R. I. McLachlan, R. Quispel, *Splitting methods*, Acta Numerica **11** (2002) 341–434.

[6] A. M. Portillo, *High order full discretization for anisotropic wave equations*, submitted.

[7] J. M. Sanz-Serna, M. P. Calvo, *Numerical Hamiltonian Problems*, Chapmand and Hall, London, 1994.

[8] M. Suzuki, *Fractal decomposition of exponential operators with applications to many-body theories and Monte Carlo simulations*, Phys. Lett. A **146** (1990) 319–323.

[9] H. Yoshida, *Construction of higher order symplectic integrators*, Phys. Lett. A **150** (1990) 262–268.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

On photosynthesis process with the interaction between two types of leaves

Anna Poskrobko[1] and Antoni Leon Dawidowicz[2]

[1] *Faculty of Computer Science, Bialystok University of Technology*

[2] *Faculty of Mathematics and Computer Science, Jagiellonian University*

emails: a.poskrobko@pb.edu.pl, Antoni.Leon.Dawidowicz@im.uj.edu.pl

Abstract

The paper deals with the description of the mathematical model of photosynthesis process in two interacting (peripheral and shaded) leaves. The bioenergetic phenomena is described by six equations for surfaces, thicknesses and levels of substrates of photosynthesis contents in two types of leaves. The well-posedness of the problem and the uniqueness of its solution are proved.

Key words: photosynthesis, differential equations
MSC 2000: 34A12, 92D25, 92D40

1 Introduction

A mathematical model of photosynthesis shows the relationships between the variables which influence the dynamics of photosynthesis taking place in green leaves. A general mathematical model of photosynthetic processes is presented in the work [4]. In paper [3] dedicated to the mathematical modelling of photosynthesis another model has been presented by the authors. It takes into account the positioning of a leaf in the conditions of full light access (in the peripheral part of the tree crown) and of a leaf shaded inside the tree crown. A completely shaded leaf can function in the tree crown on condition it is self-sufficient, i.e. capable of nourishing its tissues with photosynthetic products. Such a leaf, in case of higher light intensity and increasing day length in relation to night length, may also effectively provide the produced biomass for other plant tissues which do not participate in the process of photosynthesis (e.g. tissues of the root system or stem) [1, 8, 11, 12].

 ISBN: 978-84-617-8694-7

The process of photosynthesis also depends on the accessibility of water and mineral salts dissolved in it. In some papers authors put a considerable emphasis on bioenergetic processes working out mathematical models of plant bioenergetics (see [2, 5, 6, 7, 9, 10] and the references therein).

The mathematical model of photosynthesis presented in this work takes into consideration both the water balance differential model of a plant and the dissolved mineral salts which get to the tissues assimilating carbon dioxide together with water, being used in the course and for regulation of metabolic processes as well as for building the anatomical structures of the plant, i.e. its biomass.

Conduction of water and mineral salts to particular leaves in the tree crown is dependent on the local light conditions and on the positioning of a leaf in the topography of the tree crown, as well as on the distance between the leaf and the root system, which is connected with the water and energy balance of plant. The empiric data used in this paper come from the doctors dissertation [11] (in Polish language).

2 Variables of model

There are six basic variables in our model:

- $x_1(t)$ the surface of a peripheral leaf;
- $x_2(t)$ the surface of a shaded leaf;
- $y_1(t)$ the thickness of a peripheral leaf;
- $y_2(t)$ the thickness of a shaded leaf;
- $z_1(t)$ the mineral salt (substrates of photosynthesis) content in a peripheral leaf;
- $z_2(t)$ the mineral salt (substrates of photosynthesis) content in a shaded leaf.

Let φ be the degree of leaf insolation. It is clear that φ is a time function and is positive only at day time. Let $A(\tau)$ denotes any increasing function for small arguments and which equals zero for $\tau > \tau_0$ where τ_0 is the leaf insolation sufficient for photosynthesis process. The function A describes the influence of the leaf growth on the photosynthesis process.

3 Model assumptions

First of all, surface growth is assumed to be caused by photosynthesis which depends on the insolated leaf surface, volume of photosynthetic substrates and degree of insolation. It is believed here that in case of poor insolation the leaf builds up its surface and thickness, whilst in case of high insolation it only builds up its thickness (palisade and spongy

 ISBN: 978-84-617-8694-7

mesophylls). The decrease in surface caused by respiration is directly proportional to the biomass. Hence, the equation for x_1 will be the following

$$x_1'(t) = A(\varphi(t))x_1(t)z_1(t) - ax_1(t)y_1(t) \tag{1}$$

where coefficient a is a certain constant characterized by growth due to respiration. For the leaf shaded by the peripheral one the equation will look similar apart from the degree of insolation which will be lower due to the fact that part of the light will be shaded by the peripheral leaf. Therefore, the equation will be the following

$$x_2'(t) = A(\varphi(t) - \lambda x_1(t))x_2(t)z_2(t) - ax_2(t)y_2(t) \tag{2}$$

where λ is the coefficient of solar exposure reduction. It is obvious that the biomass will be directly proportional to the product of surface and thickness. Its growth will be directly proportional to insolation, surface and mineral salt content. Similarly, as in the case of surface, this growth is caused by respiration. Therefore, it can be put down as follows

$$(x_1y_1)'(t) = \varphi(t)x_1(t)z_1(t) - bx_1(t)y_1(t)$$

where the coefficient b describes the slowdown of the biomass caused by the respiration process. It should be noted, however, that leaf biomass cannot increase unlimitedly, which could be observed in equation (3). Therefore, its right part should be multiplied by the expression $J(y(t))$ where $J(\tau)$ equals 1 for arguments $\tau < C$ and 0 for $\tau > C + \eta$. Here η is sufficient small and C is the highest possible leaf thickness. We assume that J is differentiable on its domain. The equation in this case is the following

$$(x_1y_1)'(t) = (\varphi(t)x_1(t)z_1(t) - bx_1(t)y_1(t))J(y_1(t))$$

After applying Leibniz rule and making some transformations, and after inserting equation (1), the following equation can be obtained

$$y_1'(t) = (\varphi(t)z_1(t) - by_1(t))\, J(y_1(t)) - A(\varphi(t))y_1(t)z_1(t) + ay_1^2(t). \tag{3}$$

The equation for y_2 is analogical to the above, but we replace $\varphi(t)$ with $\varphi(t) - \lambda x_1(t)$

$$y_2'(t) = \big((\varphi(t) - \lambda x_1(t))^+ z_2(t) - by_2(t)\big)\, J(y_2(t)) - A(\varphi(t) - \lambda x_1(t))y_2(t)z_2(t) + ay_2^2(t). \tag{4}$$

What is left to be done is to put down equations for variables z_1 and z_2. For this purpose it should be noted that due to the fact that mineral salts are drawn by the same capillary, their total intake per time unit is a constant value. Let v_1 and v_2 denote, respectively, the intensity of mineral salt inflow to the peripheral and shaded leaves. It follows that the total consumption of mineral salt per unit time is constant value Q. Thus, the following assumption can be made

$$v_1 + v_2 = Q.$$

 ISBN: 978-84-617-8694-7

Moreover, the rate of v_1 and v_2 is directly proportional to particular biomasses and distances between the leaf and the root. In addition, the shaded leaf, being closer to the root, has a better access to the source of mineral salts. The following relationship can be observed:

$$v_1 = k\frac{x_1(t)y_1(t)}{x_2(t)y_2(t)}v_2. \tag{5}$$

The constant $k > 1$ denotes the coefficient characterizing access to mineral salts. Denoting

$$\frac{x_1(t)y_1(t)}{x_2(t)y_2(t)} = w(t)$$

we obtain

$$v_1(t) = \frac{Qkw(t)}{1+kw(t)}, \tag{6}$$

$$v_2(t) = \frac{Q}{1+kw(t)}. \tag{7}$$

To define the derivative of z_1 and z_2, the intensity of photosynthesis should be subtracted from v_1 and v_2. Thus, the following set of equations can be obtained:

$$\begin{aligned} z_1'(t) &= \frac{Qkw(t)}{1+kw(t)} - \gamma A(\varphi(t))x_1(t)z_1(t), \\ z_2'(t) &= \frac{Q}{1+kw(t)} - \gamma A(\varphi(t)-\lambda x_1(t))x_2(t)z_2(t) \end{aligned} \tag{8}$$

where γ is transformation coefficient expressing the consumption of mineral salt. Finally, we get the following system of the equations

$$\left\{\begin{array}{rcl} x_1'(t) &=& A(\varphi(t))x_1(t)z_1(t) - ax_1(t)y_1(t) \\ y_1'(t) &=& (\varphi(t)z_1(t) - by_1(t))\,J(x_1(t)y_1(t)) - A(\varphi(t))y_1(t)z_1(t) + ay_1^2(t) \\ z_1'(t) &=& \frac{Qkw(t)}{1+kw(t)} - \gamma A(\varphi(t))x_1(t)z_1(t) \\ x_2'(t) &=& A(\varphi(t)-\lambda x_1(t))x_1(t)z_2(t) - ax_2(t)y_2(t) \\ y_2'(t) &=& ((\varphi(t)-\lambda x_1(t))^+ z_2(t) - by_2(t))\,J(x_2(t)y_2(t)) \\ &-& A(\varphi(t)-\lambda x_1(t))y_2(t)z_2(t) + ay_2^2(t) \\ z_2'(t) &=& \frac{Q}{1+kw(t)} - \gamma A(\varphi(t)-\lambda x_1(t))x_2(t)z_2(t). \end{array}\right. \tag{9}$$

4 Existence and uniqueness of the solutions

Let us consider system (9) with the initial conditions

$$x_1(0) = x_1^0,\ y_1(0) = y_1^0,\ z_1(0) = z_1^0,\ x_2(0) = x_2^0,\ y_2(0) = y_2^0,\ z_2(0) = z_2^0. \tag{10}$$

 ISBN: 978-84-617-8694-7

Theorem 1. *Let* $x_i^0, y_i^0, z_i^0 > 0$ *for* $i = 1, 2$ *and* $y_1^0, y_2^0 \leqslant M$ *for some positive* M*. Assume that*

i) $A(\tau)$ *is an increasing function for small arguments and which equals zero for* $\tau > \tau_0$ *where* $\tau_0 > 0$ *is fixed;*

ii) φ *is positive bounded and Lipschitz function;*

iii) $J : [0, T] \to [0, 1]$ *is a differentiable function such that* $J(\tau)$ *equals* 1 *for arguments* $\tau < C$ *and* 0 *for* $\tau > C + \eta$*, where* $\eta, C > 0$ *are fixed.*

Then system of differential equations (9) *with initial conditions* (10) *has exactly one non-negative solution on the interval* $[0, T]$*.*

The assumptions of the theorem have justifiable biological interpretation. The proof of the theorem is based on the a priori estimations and the Banach fixed point theorem. But to avoid the necessity of the extension of the solution we use Bielecki's idea. In this section we will also attempt the discussion about the stability of the solution.

5 Computer simulation and comparison with empirical results

The authors have used the MAPLE 15 program to solve the system (9)–(10). The choice of coefficients is quite general because even within the same plant there are cells with very different functions among which there are cells whose mass is sometimes 50 000 times high-end than that of other cells. The graphs presented below (Figs. (1) and (2)) illustrate the solution and the most interesting result, that the peripheral leaf builds first of all its thickness and the shaded one its area. Obtained theoretical results confirm empirical results of measurement of a tree *Fagus silvatica L.* (Figs. (3) and (4)) growing in the south of Poland [11].

Acknowledgements

This work is supported by Bialystok University of Technology (Grant No. S/WI/1/2016) and founded by the resources for research by Ministry of Science and Higher Education.

References

[1] Y. G. Anderson, *Seasonal development in sun and shade leaves*, Ecology **36** (1955) 430–439.

[2] O. Björkman, P. Holmgren, *Adaptability of the Photosynthetic Apparatus to Light Intensity in Ecotypes from Exposed and Shaded Habitats*, Physiologia Plantarum **316** (1963) 889–914.

[3] A. L. Dawidowicz, A. Poskrobko, J. L. Zalasiński, *Mathematical Model of Photosynthesis Process in Leaf, The interaction between two leaves*, Proceedings of the XX National Conference Applications of Mathematics in Biology and Medicine, Łochów September 23–27 (2014) 29–34.

[4] A. L. Dawidowicz, A. Poskrobko, J. L. Zalasiński, *A Mathematical Model of the Bioenergetic Processes in Green Plants*, Mathematical Population Studies **21** (2014) 159-165.

[5] U. Foryś, Z. Szymańska, *Models of Interactions between Heterotrophic and Autotrophic Organisms*, Applicationes Mathematicae **316** (2009) 279–294.

[6] D. M. Gates, *Transpiration and Leaf Temperature*, Annual Review of Plant Physiology **19** (1968) 211–238.

[7] T. W. Goodwin, *Chemistry and biochemistry of plant pigments*, Academic Press, London, New York, 1965.

[8] K. Krebb, *Methoden der Pflanzenökologie*, Gustav Fischer Verlag, Jena, 1977.

[9] R. O. Slatyer, *Plant-water relationships*, Academic Press, London, New York, 1967.

[10] P. P. Szopa, M. J. Piotrowska, *Growth of Heterotrophe and Autotrophe Populations in an Isolated Terrestrial Environment*, Applicationes Mathematicae **38** (2011) 67–84.

[11] J. L. Zalasiński, *Wpływ lokalnych warunków świetlnych w koronie buka (Fagus silvatica L.) na kierunki zmian morfologicznych, anatomicznych i niektórych właściwoci optycznych liścia*, Rozprawa doktorska, Akademia Rolnicza w Krakowie, 1973.

[12] M. H. Zimmermann, C. L. Brown, *Trees: Structure and Function*, Springer, Berlin, Heidelberg, 1975.

 ISBN: 978-84-617-8694-7

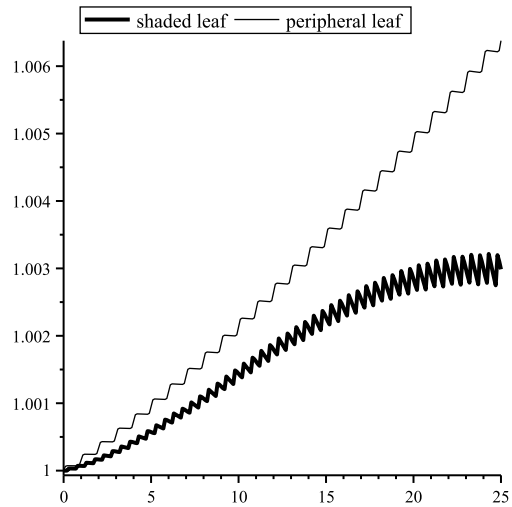


Figure 1: Comparison of area of peripheral and shaded leaf

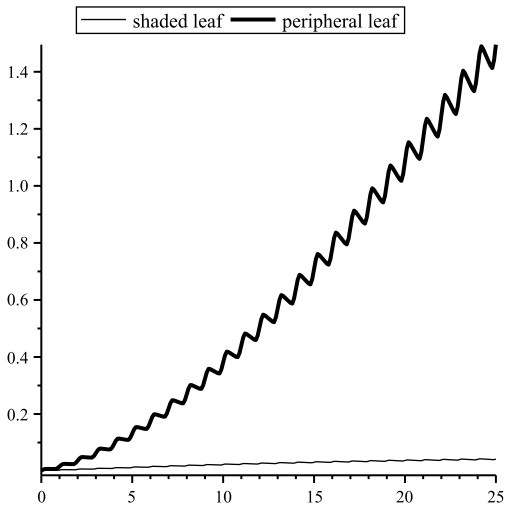


Figure 2: Comparison of thickness of peripheral and shaded leaf

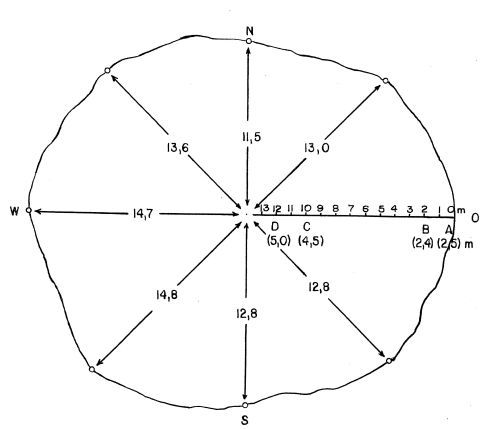


Figure 3: Cast of crown of researched tree with localization of measuring position ( ) - height from surface of ground (m)

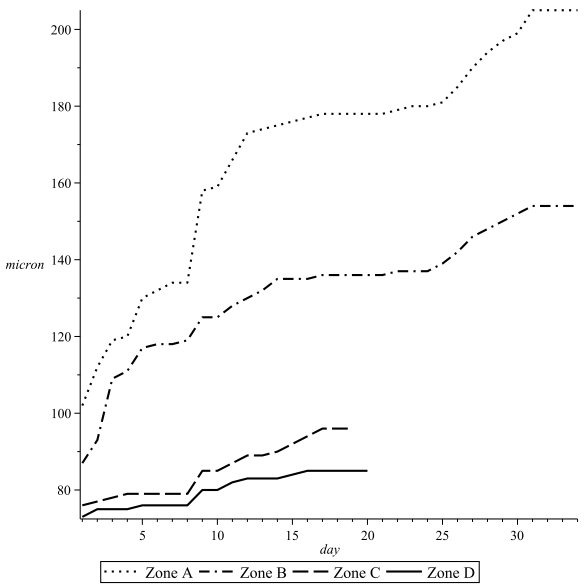


Figure 4: Empirical results in respect of zoning researched crown

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

Parametric Analysis of Active Flow Control using Steady Suction and Steady Blowing

Bhanu Prakash[1], Fernando Mellibovsky[1] and Josep M Bergada [2]

[1] *Department of Physics, Universitat Politècnica de Catalunya, Barcelona, Spain*

[2] *Department of Fluid Mechanics, Universitat Politècnica de Catalunya, Barcelona, Spain*

emails: bhanu2204@gmail.com, fernando.mellibovsky@upc.edu, josep.m.bergada@upc.edu

Abstract

Active Flow Control is implemented over NACA 2412 airfoil using Steady Suction and Steady Blowing techniques. The pre-stall angle of attack (AOA) 12^0 is studied at a high Reynolds number (Re_{chord}) of $3.1*10^6$ for two-dimensional, incompressible and steady flow conditions. A wide range of parametric value set is considered for Slot location (l_s), Velocity magnitude ratio (U_j/U_0), slot width (w) and angle of perturbation (β) using both steady suction and steady blowing independently. The numerical modeling is done using the corresponding solver in OpenFOAM, an open source CFD framework. The turbulence modeling is done using Reynolds Averaged Navier Stokes (RANS) equations, specifically k-k_l-ω model implementation in OpenFOAM. The impact of the parametric set on aerodynamic coefficients, lift (C_l) and drag(C_d), and flow separation is illustrated. Along with, the relevant boundary layer physics are explained.

Key words: Steady Suction, Steady Blowing, Aerodynamic Performance, Boundary layer analysis, High Reynolds number

1 Introduction

Flow Control provides a means for drag reduction, lift enhancement and/or noise reduction of aircraft operating at off-design flight conditions. Control Techniques may be classified as passive, active and reactive[5]. Passive control has been implemented successfully in transportation vehicles, however its counterproductive effects in its non-essential flight conditions has become a topic of interest for research community. To this objective, Active Flow Control (AFC) is a conceptually proven technology for enhancing aerodynamic performance of

both bluff and streamlined (at high AoA) bodies. It has become a rapidly evolving field in fluid dynamics for the past few decades[1, 4, 15] .

A common implementation of AFC consists in injecting/subtracting fluid in/from the nearwall region through small holes. The jets can be steady or time dependent classifying the type of actuation accordingly. Steady suction and steady blowing have been shown to provide lift augmentation (and drag reduction) in the presence of adverse pressure gradients by supressing or at least retarding flow separation([3, 2]). The effect of increasing Reynolds number is studied experimentally with steady suction being observed slightly more effective with increasing Reynolds number[6]. A 2D finite volume code is utilized to study the slat noise reduction using steady suction at high Reynolds number and is proven to be effective in this objective[8].

The impact of steady blowing in effective prevention of stall is studied at low Reynolds number with varying blowing ratios and free stream turbulence[10, 11]. Also, the dynamic stall control using strong steady blowing has resulted in increased lift coefficient, reduced hysteresis and aversion of stall[14]. Steady blowing is also implemented at the rear end of generic car models to observe an overall drag reduction but with larger massflow rates [9, 13]. It is already proven that oscillatory blowing and periodic blowing & suction provide in controlling flow separation or aerodynamic performance enhancement[7, 12]. However, due to its experimentation for different physical fluid flow conditions, there is a necessity for intensive parametric analysis of both steady suction and blowing and their boundary layer manipulation dynamics to better understand the possibility of their future consideration for engineering applications. To achieve this objective, high Reynolds number flow is chosen and analysed computationally using RANS turbulence model with assumed 2D fluid flow in order to lower the computational time requirement for the wider parametric set considered.

The section 2 is used to describe the numerical model. This is followed by the results obtained from the parametric analysis of slot location, width, perturbation angle and actuation velocity ratio for steady suction, and similarly for steady blowing without varying slot width. The final section is used to present the conclusions obtained.

2 Numerical Model

The two dimensional Navier Stokes (NS) Equations for incompressible, viscous, turbulent and steady flow is resolved using RANS Turbulence Modeling. Specifically, k-kl-Ω turubulence model is used. A finite volume method (FVM) based open-source solver, simpleFoam, within the OpenFOAM solvers framework is used. The gradient, divergence and laplacian terms in the NS equations are spatially discretized using standard finite volume discretisation of Gaussian integration with hybrid first/second order limiting scheme, limitedLinear, from cell centres to face centres.

The Semi-Implicit Method for Pressure-Linked Equations (SIMPLE) algorithm is used

 ISBN: 978-84-617-8694-7

for resolving pressure-velocity coupled equations. Hence, the post processed aerodynamic coefficients and analyzed flow physics is an average state of the fully transient flow at the very high Reynolds number considered. Pressure (p) equation is solved by Geometric Algebraic Multi Grid (GAMG) solver with 2 pressure corrections while Smooth Solver is used for Velocity field (U), Turbulent kinetic energy (k_t), Laminar kinetic energy (k_l) and Specific disspation rate (Ω) with Gauss Seidel smoother for both solvers. Final residuals of order 10^{-6} for pressure and 10^{-8} for U, k_t, k_t and Ω were consistently obtained throughout the simulations.

A parabolic computational domain is considered with the outlet placed at 24 C from trailing edge and inlet at 13 C upstream of airfoil leading edge with periodic boundary conditions in spanwise direction. The Reynolds number chosen is 3.1 $*10^6$ for NACA 2412 airfoil at an angle of attack 120. The inlet boundary velocity condition is fixed (46 m/s) with a Neumann boundary condition at outlet. For pressure, Neumann boundary conditions are defined at both inlet and outlet . The airfoil pressure and suction surface is defined as no slip walls. The initial values of k_t and Ω are computed at Inlet by considering a turbulent intensity of 0.06% and viscosity ratio of (μ_t/μ) 0.01. And k_l is defined zero at all boundaries with Neumann condition at outlet. The mesh independence is primarily checked by using $y^+ < 1$, with three different values 0.33, 0.24 and 0.16, and concluding that y+ = 0.24 with respect to computational time, as the lift and drag coefficients are close to experimental data for all three values. Active Flow Control, both steady suction and steady blowing, is implemented with a numerical boundary condition with an uniform flat profile constant in time.

3 Results

3.1 SteadySuction

3.1.1 Variation of Velocity ratio and Slot location

The skin friction coefficient for the baseline flow is plotted in Figure 3(b). It can be observed that the turbulent separation is occurring at around 0.815C. The slot location is varied only in the upstream by using this position as reference. The velocity magnitude of suction expressed in terms of ratio with free stream velocity (U_j/U_0) is varied for five values i.e., 0.2, 0.3, 0.5, 0.7 and 0.9 at four different slot locations 0.4C, 0.5C, 0.6C and 0.7C. The slot width and perturbation angle are maintained constant with values of 0.15%C and 90^0 respectively. The lift coefficient (C_l) and drag coefficient (C_d) location are plotted in Figure 2(a) and 2(b) respectively.

From Figure 2(a), there is a clear increase in lift coefficient with suction velocity and interestingly even with very low value of U_j/U_0 of 0.2, the lift obtained is higher than baseline case. For an overview, with the ratios of 0.2, 0.5 and 0.9 at the slot location of

0.6 C, the lift is enhanced by 5%, 7.2% and 8.9% respectively. The Pressure Coefficient (C_p) is plotted in Figure 3(a) which illustrates this trend in lift coefficient. With increasing velocity ratio, the pressure at suction side is decreased both upstream and downstream of the slot location followed by an increase near trailing edge even compared with baseline case. Similarly, an increase in pressure at the pressure side of airfoil can be observed with an increasing suction velocity. The interesting trend in C_p is seen just before the slot where the pressure distribution has a high impact with the flow suction. But immediately after the slot there is a slight increase in pressure with increasing velocity ratio of suction, which could be due to the presence of the reverse flow or vortex. However, the impact of suction is regained farther downstream of slot with higher suction rate reducing the pressure over airfoil's upper side.

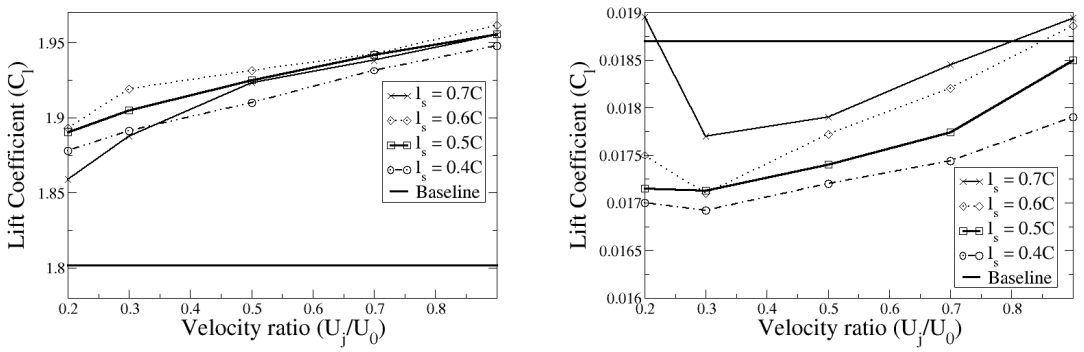
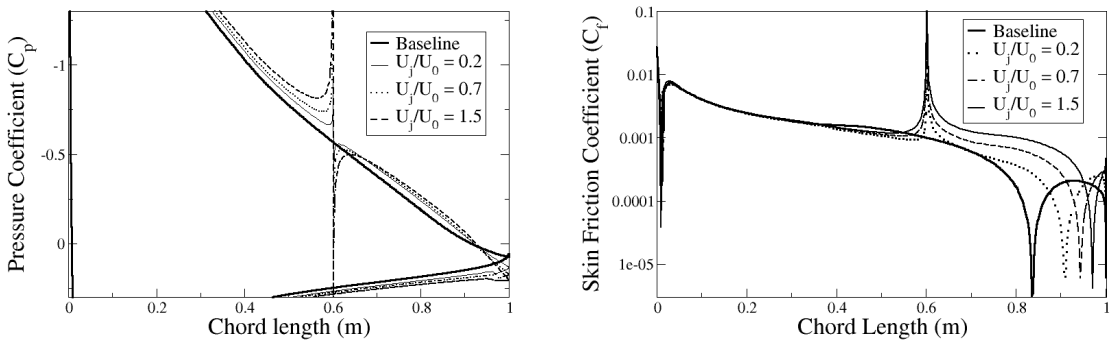


Figure 1: (a) Lift Coefficient (b) Drag Coefficient for varying velocity ratios and slot locations

The drag coefficient(C_d) has increased with increasing velocity ratio except from 0.2 to 0.3 as illustrated in Figure 2(b) although most of the ratios and slot locations contributed for a decrease when compared to the baseline case. Numerically, a drag decrease of 8.1%, 8% and 4.3% is obtained for the ratios of 0.2, 0.5 and 0.9 respectively at slot location of 0.4C. The skin friction coefficient plotted in Figure 3(b) shows higher value of C_f with increasing suction velocity after the actuated position specifically. Also the increasing separation delay with suction velocity ratio can be seen from C_f plot where the C_f value tends to zero closer to the trailing edge locations with increasing velocity ratio. In terms of slot location, suction placed farther away to the natural separation point, significantly reduced drag as compared to a closer location, whereas a critical location at 0.6C is observed for Cl which increased the lift to maximum followed by a decrease nearer and farther to the natural separation. This trend is also seen in Figure 2(a).

 ISBN: 978-84-617-8694-7



(a) (b)

Figure 2: (a) C_p and (b) C_f for varying velocity ratios

3.1.2 Variation of Slot width

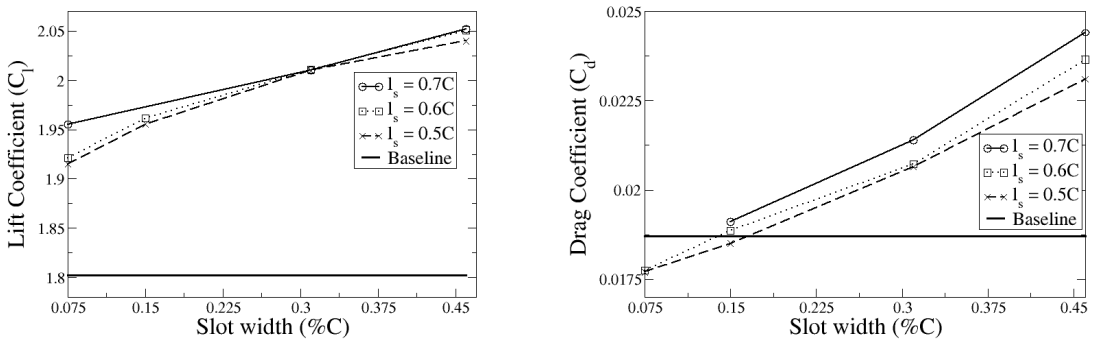
The width of the slot is varied for four parameters i.e., 0.075%C, 0.15%C, 0.31%C and 0.46%C at a constant U_j/U_0 and perturbation angle of 0.9 and 90^0 respectively. There is a clear increase in the Cl with increasing slot width as shown in Figure 4(a). The impact of slot location is low as compared to the slot width here. A drastic decrease in suction side pressure with higher slot widths was observed in the C_p. Also the C_f did show an increase in separation delay position with higher slot widths, and in the case of the slot located at 0.46%C, it is observed that there is an almost complete attachment of the boundary layer. Hence the very low flow separation is also a reason for higher lift increments.

However, the drag coefficient (C_d) in Figure 4(b) has also increased significantly by increasing the slot width. This is due to the increasing skin friction after the flow suction. The larger delayed separation is contributing to the higher C_f which probably is due the increase in the viscous component of the drag. The adverse effect here is the drag being even higher than the baseline case in most of the parametric set. The higher slot widths mean higher flow suction rates which applies to large suction velocity ratios (U_j/U_0) too. Hence, from both the velocity and slot width variation, it is observed that optimal flow suction rate is necessary in order to obtain higher lift to drag ratios.

3.1.3 Variation of Angle of Suction

The variation of angle of suction (β) has shown the trends in lift and drag coefficients as shown in the Figure 5 (a) and (b). The tangent to the location of variation on the suction side of airfoil is considered as a reference for zero degrees. The increasing values are in the clockwise rotation for this reference. The values of β are varied in the range of 30^0 to

 ISBN: 978-84-617-8694-7



(a) (b)

Figure 3: (a) C_l and (b) C_d with varying slot width and slot locations

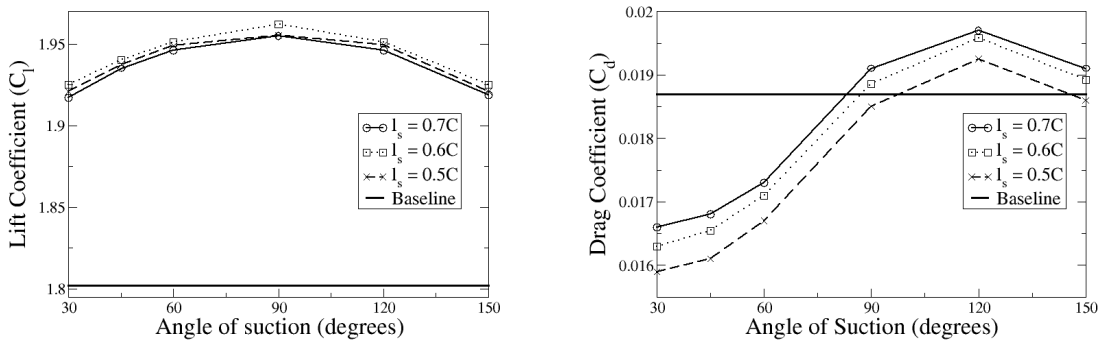
150^0. It is observed that 90^0 provided higher lift with a decreasing trend from it by either increasing or decreasing the angle of suction. Interestingly, there is a symmetry in the lift coefficient values at angles higher or lower magnitude by an equal magnitude, for example the pairs of 60^0 & 120^0 and 30^0 & 150^0. The parameter that is similar in these pairs is the vertical component of velocity that signifies the flow suction rate. As observed in the varation of velocity ratio and slot width, the impact of flow suction rate is very clear in the Figure 5 (a). The pressure coefficients also have shown similar trends supporting these conclusions with reduced pressure over the suction side. The drag is increased by increasing the angle of suction from 30^0 to 120^0 followed by a decrease from 120^0 to 150^0 which is likely to be associated with the varying horizontal component of the suction velocity combined with the viscous drag associated with flow attachment.

3.2 Steady Blowing

3.2.1 Variation of Slot location

With the same baseline case as reference, steady blowing is implemented by varying the slot location (l_s), velocity ratio (U_j/U_0) and blowing angle (β) and the trends in lift and drag are observed. The reliability on the values of the drag is low due to the highly transient nature of the flow after blowing location which resulted in a difference of 30% between the mean and the extremes values of C_d. However the C_l is not so significantly varying due to its higher magnitude and is considerably reliable. It has to be noted that the angle considered is in the counter clockwise direction with reference to the tangent to the airfoil at considered specific location.

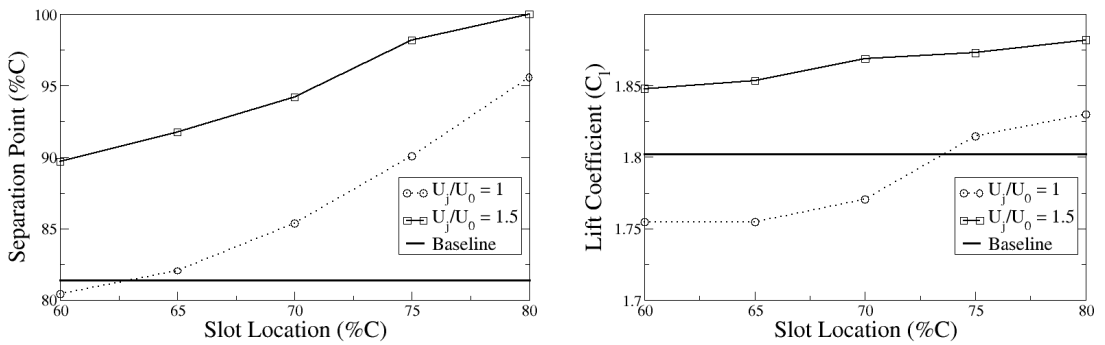
By varying the slot location(l_s) from 60%C, 65%C, 70%C, 75%C and 80%C, the lift is



(a) (b)

Figure 4: (a) C_l (b) C_d for varying Angles of suction and slot locations

increased when the slot location is closer to the natural point of separation (81%C) with reference to upstream position and it decreases as the slot is moved towards the leading edge(Figure 6(a)and 6(b)). It can also be observed that the increase in C_l is 1.55% as compared to the steady suction (8.9%) with similar magnitudes of velocity ratios (0.9 for suction and 1 for blowing). The slot width considered is 0.15%C and the angle is 30^0



(a) (b)

Figure 5: (a) Separation position (b) C_l for varying slot locations

 ISBN: 978-84-617-8694-7

3.2.2 Variation of Velocity ratio

Similary, by increasing the velocity ratio from 0.5 to 2, it can be observed from Figure 7 that the C_l is increased with U_j/U_0. To achieve performances in lift similar to suction, approximately twice the magnitudes of U_j/U_0 are necessary. It should also be noted that for lower blowing ratios, there is decrease in lift as compared to baseline case which suggests the dependency on the higher velocity magnitudes. The trend in slot location dependency can be observed here.

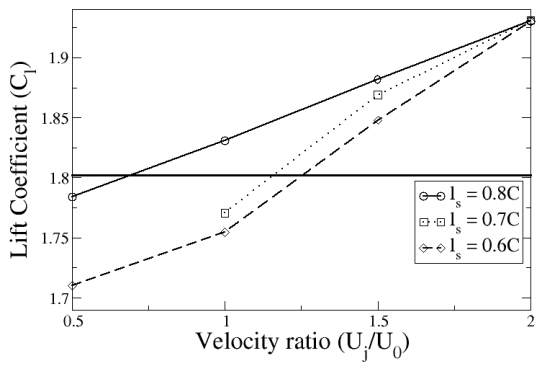


Figure 6: C_l for varying velocity ratios at different slot locations

3.2.3 Variation of Blowing Angle

The steady blowing is observed to be very sensitive to the blowing angle and any increase to the angle higher than 45^0 resulted in the divergence of numerical simulations. Hence the trends are presented from 0^0 (tangential) to 45^0. From Figure 8, it can be observed that 30^0 is the best value for blowing within the limited parametric set considered and an increase or decrease to this angle resulted in the reduction of C_l. For locations farther than 60%C from the leading edge, there is a continuous decrease with increasing angle. However, from the magnitudes of the C_l observed, these locations are of little interest.

4 Conclusions

The baseline flow is computed using RANS turbulence modeling,k-k_l-ω model, for Re_C = 3.1 *10^6 at pre-stall angle of attack 12^0. The flow is assumed to be 2D, steady and incompressible due to the sheer interest in the wide range of parametric set. A detailed parametric analysis is performed for Slot location (l_s), Velocity magnitude ratio (U_j/U_0),

 ISBN: 978-84-617-8694-7

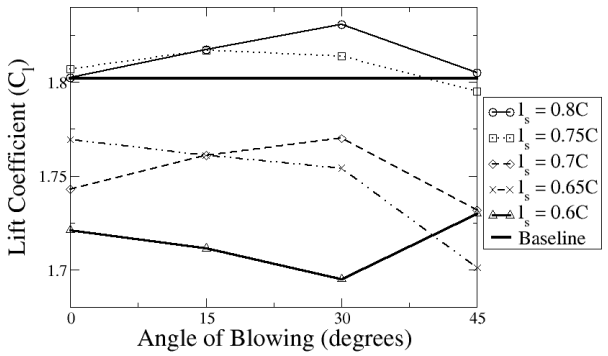


Figure 7: C_l for varying steady blowing angles at different slot locations

slot width (w) and angle of perturbation (β) using both steady suction and steady blowing independently. For steady suction, a critical slot location (60%C) is observed with natural separation of baseline flow occuring at 81%C with a decrease in aerodynamic performance with slots closer to leading and trailing edges, upstream to natural separation. An increasing velocity ratio U_j/U_0 resulted in increasing C_l and C_d. Similar trend is observed by varying slot width. However, suction angle of 90^0 is the best for suction. The dependency on net flow suction rate is explained using the observed trends, a higher rate implies better aerodynamic performance.

The similar parametric analysis is performed for steady blowing. The slot location closer to the natural separation point provided higher lift with a decreasing trend when the slot is moved towards the leading edge. Also, C_l increased with increasing velocity ratios. For a similar increase in C_l as compared to steady suction, the magnitudes of velocity ratio have to approximately twice for blowing as compared to suction. Also, the lower velocity ratios are detrimental to aerodynamic performance with respect to baseline case. With the limited parametric set considered for blowing angle, 30^0 is optimal with higher C_l and a decreasing trend is observed by either increasing or decreasing the angle.

References

[1] Louis N Cattafesta III and Mark Sheplak. Actuators for active flow control. *Annual Review of Fluid Mechanics*, 43:247–272, 2011.

[2] Chunmei Chen, Roman Seele, and Israel Wygnanski. Separation and circulation control on an elliptical airfoil by steady blowing. *AIAA journal*, 50(10):2235–2247, 2012.

 ISBN: 978-84-617-8694-7

[3] Chunmei Chen, Roman Seele, and Israel Wygnanski. Flow control on a thick airfoil using suction compared to blowing. *AIAA journal*, 51(6):1462–1472, 2013.

[4] S Scott Collis, Ronald D Joslin, Avi Seifert, and Vassilis Theofilis. Issues in active flow control: theory, control, simulation, and experiment. *Progress in Aerospace Sciences*, 40(4):237–289, 2004.

[5] Mohamed Gad-el Hak. *Flow control: passive, active, and reactive flow management*. Cambridge University Press, 2007.

[6] David Greenblatt, Keith Paschal, Chung-Sheng Yao, Jerome Harris, Norman Schaeffler, and Anthony Washburn. A separation control cfd validation test case. part 1: Baseline & steady suction. In *2nd AIAA Flow Control Conference*, page 2220, 2004.

[7] David Greenblatt and Israel J Wygnanski. The control of flow separation by periodic excitation. *Progress in aerospace Sciences*, 36(7):487–545, 2000.

[8] Thilo Knacke and Frank Thiele. Slat noise reduction using steady suction. In *46th AIAA Aerospace Sciences Meeting and Exhibit*, page 17, 2008.

[9] RP Littlewood and MA Passmore. Aerodynamic drag reduction of a simplified square-back vehicle using steady blowing. *Experiments in fluids*, 53(2):519–529, 2012.

[10] Brian R McAuliffe and Steen A Sjolander. Active flow control using steady blowing for a low-pressure turbine cascade. In *ASME Turbo Expo 2004: Power for Land, Sea, and Air*, pages 1223–1235. American Society of Mechanical Engineers, 2004.

[11] Hanns F Müller-Vahl, Christoph Strangfeld, Christian N Nayeri, Christian O Paschereit, and David Greenblatt. Control of thick airfoil, deep dynamic stall using steady blowing. *AIAA journal*, 53(2):277–295, 2014.

[12] A Seifert, T Bachar, D Koss, M Shepshelovich, and I Wygnanskil. Oscillatory blowing: a tool to delay boundary-layer separation. *AIAA journal*, 31(11):2052–2060, 1993.

[13] Erik Wassen and Frank Thiele. Drag reduction for a generic car model using steady blowing. In *4th Flow Control Conference*, page 3771, 2008.

[14] David Weaver, Kenneth McAlister, and Jin Tso. Control of vr-7 dynamic stall by strong steady blowing. *Journal of aircraft*, 41(6):1404–1413, 2004.

[15] Israel J Wygnanski. A century of active control of boundary layer separation: a personal view. In *IUTAM Symposium on One Hundred Years of Boundary Layer Research*, pages 155–165. Springer, 2006.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

Enhancing Molecular Shape Comparison by a Parallel Global Evolutionary Algorithm

S. Puertas-Martín[1], M. R. Ferrández[1], J. L. Redondo[1], H. Perez-Sanchez[2] and P. M. Ortigosa[1]

[1] *Department of Computer Science, University of Almería, Agrifood Campus of International Excellence, ceiA3, Almería, Spain*

[2] *Department of Computer Science, Universidad Católica San Antonio de Murcia (UCAM), Murcia, Spain*

emails: savinspm@ual.es, mrferrandez@ual.es, jlredondo@ual.es, hperez@ucam.edu, ortigosa@ual.es

Abstract

Virtual screening techniques provide predictions about drug bioactivity and toxicity and their activity in novel diseases. To achieve this aim, it is needed to incorporate descriptors to obtain specific information about the compounds according to the source database. In particular, we focus on the shape similarity descriptor. Recently, a new strategy for molecular shape comparison was proposed. It is an evolutionary optimization algorithm based on subpopulations which needs big population sizes to explore the search space deeply and then obtain good solutions. This translates directly into larger computational times and larger resource requirements. In view of this situation, in this work, the optimization algorithm has been parallelized. A computational study to analyze the new parallel method in terms of efficiency and effectiveness has been carried out. The use of several processors, and hence more computational resources, allows us to accelerate the molecular shape comparison procedure.

Key words: ligand based virtual screening, parallel computing, evolutionary algorithm

1 Introduction

Virtual Screening (VS) techniques allow us providing predictions about which chemical compounds might interact with a given protein target in some specified way and thus

achieving the desired biological function. VS techniques are mainly divided into Ligand Based Virtual Screening (LBVS) and Structure Based Virtual Screening (SBVS).

SBVS methods require detailed structural information about the target protein and cannot be applied in situations where this data is not available. Unfortunately, this issue is very common, for instance, in the case of membrane proteins such as GPCRs (G protein-coupled receptor), which are of the highest pharmacological relevance. In such cases, one can recur to LBVS methods, where only information about known ligands (actives and inactives, agonists and antagonists, etc) is exploited in order to predict new bioactive compounds against selected protein targets. LBVS methods will, therefore, consider all existing available information (structural, physico-chemical parameters, binding affinities, etc) about known active and inactive compounds, and this information will be referred to as molecular descriptors. There exist a large number of molecular descriptors or potentials used to compare molecules, as for example Electrostatic similarity, Atomic property fields, Aromatic potential, Desolvation potential, etc. In this work, we will focus on Shape similarity.

Molecular shape is a very important issue in computer-aided drug design. Several methods for comparing molecular shapes have been designed and numerous applications for virtual screening, scaffold hopping, and shape-feature based molecular alignment have been presented. However, due to new drug discovery paradigms, such as network pharmacology or system pharmacology, faster and more effective techniques for lead identification and optimization are required. Recently, an evolutionary algorithm, called Optipharm, has been proposed to this aim. It is a population-based algorithm which applies randomized operators over a population of candidate solutions to generate new points in the search domain. To be able to explore the search space deeply and then obtain good solutions, such an algorithm requires big population sizes. This translates into larger computational times. However, the sequential optimization method has been implemented with parallelism in mind, and hence, high performance computing can be easily incorporated to accelerate it. In this work, a master-slave strategy has been considered. The parallel version of Optipharm will be named Paral-Optipharm throughout this paper.

The rest of the paper is organized as follows. Section 2 shows how to measure the shape similarity between two molecules, which in turn is the objective function of both the sequential and parallel algorithms, Section 3 briefly describes the fundamentals of the parallel version. Finally, Section 4 shows some preliminary results and future works.

2 The problem: Shape Similarity

The shape of a molecule is an important concept in computer-aided drug design. That is supported by the importance of bindings have between ligands and receptor: if a set of different ligands bind to the same site receptor, it is possible that all of them possess a similar shape similarity [3]. Different models and applications have been developed in that

 ISBN: 978-84-617-8694-7

regard with the aim to get more accurate and efficient solutions.

Among the existing models, two of them have had more acceptance: Hard-Sphere method and Gaussian method. The first one is based on considering each atom as an sphere. The main problem with this method is its implementation, since although the analytical expression for the volume and its derivatives are known [1, 7], the formulas for the intersection of multiple spheres are not trivial. For this reason, the Gaussian method proposed by Good and Richards has been used more widely. Now each atom is represented by a gaussian function [2, 3]. In the original expression is recommended the inclusion of six terms, although in the practice, the first term is enough to get a good precision [6].

However, the Gaussian model does not consider the properties of the molecules to be evaluated. To include this information, Yan et al. add a weight factor to each atom in the evaluated molecules [8]. This consideration was based on the fact that not all the substances have the same density, and this point affects the overlap values and hence, the shape similarity scope.

In [8], the shape density of a molecule is expressed as a linear combination of weighted atomic Gaussian functions as:

$$G(r) = \sum_i w_i g_i(r) = \sum_i w_i p e^{-(\frac{3p\pi^{1/2}}{4\phi_i^3})(r-r_i)^2} \tag{1}$$

Therefore, the overlap volume of two molecules is given by:

$$V_{AB}^g = \sum_{i \in A, j \in B} w_i w_j v_{ij}^g \tag{2}$$

Because of the different size of the molecules, the value obtained in the previous expression is not comparable with a different couple of molecules. To solve that and get a normalized value, Tanimoto metric has been used:

$$S_{AB} = \frac{V_{AB}}{V_{AA} + V_{BB} - V_{AB}} \tag{3}$$

3 Paral-Optipharm: a parallel algorithm for molecular shape comparison

An evolutionary algorithm called Optipharm has been recently proposed to predict the shape similarity between two molecules. Such an algorithm inherits some ideas and fundamentals from UEGO [5]. In Algorithm 1, its main structure is depicted. Optipharm, as UEGO, works with a population of independent candidate solutions (species). This means that a single species can create a new offspring and evolve to the local or global optima without

 ISBN: 978-84-617-8694-7

Algorithm 1 Algorithm Optipharm

```
1 Init_species_list
2 Optimize_species(n_1)
3 FOR i = 2 to L
4     Determine R_i, new_i, n_i
5     Create_species(new_i)  # budget_per_species = new_i/length(species_list_i)
6     Select_species(R_i)
7     Optimize_species(n_i)  # budget_per_species = n_i/max_spec_num
```

participation of the remaining ones. Therefore, there exists an intrinsic parallelism that can be exploited by dividing the species among the available processing elements.

This work explores and evaluates a master-slave strategy applied to Optipharm, which leads to a parallel version named Paral-Optipharm. A master-slave technique is a "global parallel model", i.e. the management of the population is global and all the individuals in the population are considered when selection or creation procedures are carried out. Two kinds of processing elements can be distinguished, the *master* and the *slaves*. The master processor sees into making global decisions and delivering information among the slaves, which execute different tasks in a concurrent way.

In our particular master-slave (MS) model, the master processor executes Optipharm sequentially. The parallelism comes from the simultaneous evaluation of the candidate solutions in the Create_species function, and from the concurrent execution of the Optimize_species procedure. Therefore, new creation and optimization procedures have been designed to cope with the parallel model. These new procedures, called Create_species_paral and Optimize_species_paral, are described now:

- *Create_species_paral*: At each level i, the master obtains a new offspring of candidate solutions. The evaluation of the objective function is carried out in a parallel way. To this aim, the master processor divides the list of candidate solutions by the number of processors P and delivers the resulting sublists among all the processing elements (including itself). Each processing element receives a species sublist from the master processor and evaluate it.

  The master processor does not receive information from the slaves until it has finished its work (first synchronization point). When it does, it passes to a reception state, where it picks up the evaluated sublists sent by the slaves. Once the master has received all the information from the slaves, it updates the candidate solutions list.

  Following with the general structure of Optipharm, a selection procedure is carried out. This procedure is accomplished by the master, while the slaves stay idle (second synchronization point). If the list length is larger than the maximum allowed, then

 ISBN: 978-84-617-8694-7

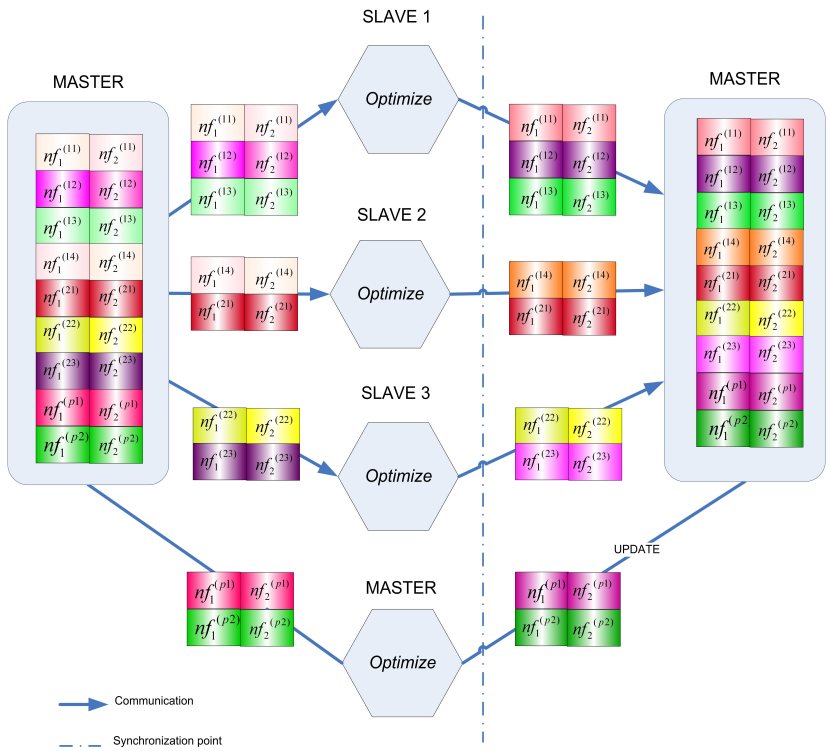


Figure 1: Master-slave strategy: delivering species for optimization.

the best candidate solutions in terms of the objective function, are selected.

- *Optimize_species_paral*: To perform the optimization process, the master divides the species list among all the processors (again including itself). In this case, each slave executes the Optimize_species procedure to every species in its sublist, i.e. a local optimization process is applied (see Figure 1). When the master finishes its work, it begins to receive the new species sublists from the slaves (third synchronization point).

Note that the synchronization points are imposed because the master is managing the whole species list.

4 Preliminary results and conclusions

In this work, a parallel global evolutionary algorithm for shape comparison between molecules has been presented. In particular, a master-slave method has been proposed. The perfor-

mance of the parallel method has been tested by using the well-known Directory of Useful Decoys (DUD) database [4]. It contains 40 different targets. Each target has associated a number of ligands and decoys. Decoys are molecules physically similar to the ligands but topologically distinct. In total, DUD is made up by 3961 ligands and 124413 decoys.

Results have shown that the efficiency of the parallel method highly depends on the two input molecules to be compared, i.e. the particular properties of the molecules may affect the speedup of the parallel strategy. Even so, the results are promising, and in some cases, efficiencies close or equal to the ideal one have been obtained.

As a future work, we will test the parallel method with different databases and develop other parallel versions based on different paradigms.

Acknowledgements

This work has been funded by grants from the Spanish Ministry of Economy and Competitiveness (TIN2015-66680-C2-1-R), Junta de Andalucía (P11-TIC7176 and P12-TIC301), Fundación Séneca–Agencia de Ciencia y Tecnología de la Región de Murcia under Projects 19419/PI/14 and 18946/JLI/13, and by the Nils Coordinated Mobility under grant 012-ABEL-CM-2014A, in part financed by the European Regional Development Fund (ERDF). Powered@NLHPC: This research was partially supported by the supercomputing infrastructure of the NLHPC (ECM-02). The authors also thankfully acknowledge the computer resources and the technical support provided by the Plataforma Andaluza de Bioinformática of the University of Málaga. This work was partially supported by the computing facilities of Extremadura Research Centre for Advanced Technologies (CETACIEMAT), funded by the European Regional Development Fund (ERDF). CETACIEMAT belongs to CIEMAT and the Government of Spain. Juana López Redondo is a fellow of the Spanish 'Ramón y Cajal' contract program, co-financed by the European Social Fund. Savíns Puertas Martín is a fellow of the Spanish 'Formación de profesorado universitario' program, financed by the Ministry of Education, Culture and Sport.

References

[1] Gibson, K. D. and Scheraga, H. A., *Exact Calculation of the Volume and Surface Area of Fused Hard-Sphere Molecules with Unequal Atomic Radii*, Molecular Physics **62** (1987) 1247–1265.

[2] Grant, J. A. and Pickup, B. T., *A Gaussian Description of Molecular Shape*, The Journal of Physical Chemistry, **99** (1995) 3503–3510.

 ISBN: 978-84-617-8694-7

[3] Grant, J. A., Gallardo, M. A. and Pickup, B. T., *A fast method of molecular shape comparison: A simple application of a Gaussian description of molecular shape*, Journal of Computational Chemistry, **17(14)** (1996) 1653–1666.

[4] Huang, N., Shoichet, B. K. and Irwin, J. J., *Benchmarking Sets for Molecular Docking*, Journal of Medicinal Chemistry, **49** (2006) 6789–6801.

[5] Jelásity, M., Ortigosa, P.M. and García, I., *UEGO, An Abstract Clustering Technique for Multimodal Global Optimization*, Journal of Heuristics **7 (3)** (2001) 215–233.

[6] Nicholls, A., MacCuish, N. E. and MacCuish, J. D., *Variable Selection and Model Validation of 2D and 3D Molecular Descriptors*, Journal of Computer-Aided Molecular Design, **18** (2004) 451–474.

[7] Richmond, T. J., *Solvent Accessible Surface Area and Excluded Volume in Proteins*, Journal of Molecular Biology, **178** (1984) 63–89.

[8] Yan, X., Li, J., Liu, Z., Zheng, M., Ge, H. and Xu, J., *Enhancing molecular shape comparison by weighted Gaussian functions*, Journal of Chemical Information and Modeling **53(8)** (2013) 1967–1978.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

Parallel SUMIS Soft Detector for MIMO Systems on Multicore

Carla Ramiro[1], M. Ángeles Simarro[2], Alberto Gonzalez[2] and Antonio M. Vidal[3]

[1] *Institute on Robotics and Information and Communication Technologies, Universidad de Valencia*

[2] *Institute of Telecommunications and Multimedia Applications, Universidad Politécnica de Valencia*

[3] *Department of Information Systems and Computation, Universidad Politécnica de Valencia*

emails: cramiro@dsic.upv.es, mdesiha@iteam.upv.es, agonzal@dcom.upv.es, avidal@dsic.upv.es

Abstract

The study of MIMO systems with large number of antennas and high-order constellations is currently generating considerable interest since provides significantly higher espectral efficiency. Unfortunately, increased complexity of the signal detection stage is the price to pay for large MIMO systems. The *Subspace Marginalization with Interference Suppression* (SUMIS) detector exhibits good performance with reduced complexity and its design allows massively parallel algorithmic implementations. This paper presents practical parallel approaches of the SUMIS detector for large MIMO systems: using multicore processors. Parallel approaches have been evaluated and compared in terms of performance and complexity with other detectors for different system parameters. Results show how these parallel versions allow to accelerate dramatically the detection, especially if very large systems and high order modulations are considered.

1 Introduction

Multiple-Input Multiple-Output (MIMO) systems provide significant capacity improvement using multiple antennas at both sides of digital communication systems. This capacity

 ISBN: 978-84-617-8694-7

increases with the minimum number of transmit and receive antennas [1]. This technology has become essential for wireless communications and has been incorporated into many communication wireless standards like IEEE 802.11n/ac (Wi-Fi)[2], wiMAX (4G)[3] and Long Term Evolution (4G)[4].

An emerging research area are so-called Large MIMO systems, often referred to as massive MIMO systems. It can be defined as those systems that use very large number of antennas, e.g., one hundred ore more [5], in contrast to conventional MIMO systems, which employ up to ten antennas. The price to pay is increased complexity and energy consumption at both ends. Particularly the MIMO detection problem is generally computationally very expensive to deal with. Thus, an adequate balance between efficiency and complexity is critical, especially in large MIMO systems [6] and large sizes constellation.

The optimal detector, which solves the MIMO detection problem optimally, computes the log-likelihood ratios (LLRs) values exactly and holds prohibitively high computational complexity, which grows with the size of the signal constellation and the number of antennas. In the above context, several detectors and exhibit different trade-offs between complexity and performance have been recently proposed. "*Single Tree Search*" (STS) [7] and "*Repeated Tree Search*" (RTS) [8] algorithms are the most common detectors which achieve the max-log approximation exactly. Both algorithms are based on the "*Sphere Decoder*" (SD) method and their computational complexity varies depending on the channel and noise realization. Sub-optimal max-log algorithms reduce the complexity at the expense of a certain performance loss. Examples of such detectors are: "*Soft Fixed Sphere Decoder*" (SFSD) [9], "*List Sphere Decoder*" (LSD) [10] or "*Soft Output k-Best*" [11], among others. On the other hand, "*Partial Marginalization*" (PM) [12] and SUMIS [13] algorithms represent an intermediate approach between the optimal detector and its max-log approximation. The SUMIS algorithm offers a good trade-off between exact and approximate computation of the LLR values and a given complexity. Even so, SUMIS detector can be the bottleneck for the overall system performance if large number of antennas or high order constellations are used. A reduction of the SUMIS detector complexity based on Box Optimization was presented in [14] but at the expense of a slight performance degradation.

This paper aims to reduce the computational cost of the SUMIS method, not only from a theoretical point of view, but through its scalable and versatile implementation for efficient processing thereof e.g. multicore processors. This allows to guarantee the SUMIS detection performance in large MIMO systems with higher throughput.

2 Background

2.1 System Model

Let us consider a complex-valued MIMO system model, using n_T transmit and n_R receive antennas with $n_T \leq n_R$. At the transmitter, the information bits are encoded, interleaved

 ISBN: 978-84-617-8694-7

and then mapped to symbols. Each symbol s_j is taken independently from the M-ary constellation Ω. The symbol contains $k = \log_2(M)$ encoded and interleaved bits. The corresponding bits are denoted by $s_{j,b}$, where the indices refer to the bth bit associated with the jth symbol. At each signaling period, the relation between the transmitted symbol vector, $\mathbf{s}_c \in \mathbb{C}^{n_T}$, and the associated received vector, $\mathbf{y}_c \in \mathbb{C}^{n_R}$, can be expressed as

$$\mathbf{y}_c = \mathbf{H}_c\mathbf{s}_c + \mathbf{v}_c, \tag{1}$$

where $\mathbf{H}_c \in \mathbb{C}^{n_R \times n_T}$denotes a fading channel matrix with independent elements $h_{j,i} \sim \mathcal{N}(0,1)$ and it is assumed to be perfectly known by the receiver. Vector $\mathbf{v}_c \sim \mathcal{N}(\mathbf{0}, \frac{N_o}{2}\mathbf{I})$ denotes an additive Gaussian noise (AWGN). Since separable complex-value constellation can be considered (quadrature amplitude modulation (M-QAM)), we can easily transform the $(n_R \times n_T)$-dimensional complex equation (1) into an equivalent $(2n_R \times 2n_T)$-dimensional real-valued representation as described in [15]:

$$\begin{bmatrix}\Re(\mathbf{y}_c)\\ \Im(\mathbf{y}_c)\end{bmatrix} = \begin{bmatrix}\Re(\mathbf{H}_c) & -\Im(\mathbf{H}_c)\\ \Im(\mathbf{H}_c) & \Re(\mathbf{H}_c)\end{bmatrix}\begin{bmatrix}\Re(\mathbf{s}_c)\\ \Im(\mathbf{s}_c)\end{bmatrix} + \begin{bmatrix}\Re(\mathbf{v}_c)\\ \Im(\mathbf{v}_c)\end{bmatrix} \tag{2}$$

where $\Re(\cdot)$ and $\Im(\cdot)$ denote the real and imaginary part of $(\cdot)$, respectively. Thereby, the real model can be described as $\mathbf{y} = \mathbf{H}\mathbf{s} + \mathbf{v}$.

Let $\sqrt{M}$-PAM with $P_M = \{-\sqrt{M}+1, \cdots, -1, 1, \cdots, \sqrt{M}-1\}$ be the real-valued representation of a M-QAM constellation where $\Omega = \{a+bj : a, b \in P_M\}$. As is explained in [13], this model offers some advantages for the SUMIS algorithm. For this reason, throughout the rest of this paper we assume the real-valued model represented in (2).

At the receiver, the demodulator computes soft information in form of LLR values for each of the encoded and interleaved bit $s_{j,b}$. This soft information expresses how likely is the hypothesis that the $s_{j,b}$ bit was equal to 1 or 0. Assuming equal a priori probabilities and using Bayes' theorem, it can be expressed as

$$L_{j,b} = \log \frac{\sum_{\mathbf{s}\epsilon\chi^1_{j,b}} \exp(-\frac{1}{N_0}\|\mathbf{y} - \mathbf{H}\mathbf{s}\|^2)}{\sum_{\mathbf{s}\epsilon\chi^0_{j,b}} \exp(-\frac{1}{N_0}\|\mathbf{y} - \mathbf{H}\mathbf{s}\|^2)}, \tag{3}$$

where $\chi^u_{j,b}$ denotes the set of possible transmitted vectors for which $s_{j,b}$ bit is equal to u. The computational complexity of (3) grows exponentially with n_T and polynomially with M. Thus, the exact MIMO detection scheme becomes prohibitive. The most common approach to cope with this limitation is the max-log approximation [10] where

$$L_{j,b} \approx \min_{\mathbf{s}\epsilon\chi^0_{j,b}} \frac{1}{N_o}\|\mathbf{y} - \mathbf{H}\mathbf{s}\|^2 - \min_{\mathbf{s}\epsilon\chi^1_{j,b}} \frac{1}{N_o}\|\mathbf{y} - \mathbf{H}\mathbf{s}\|^2. \tag{4}$$

However, the max-log approximation does not lead to a complexity reduction by itself. Eq (4) requires the computation of the same metrics than (3). Nonetheless, it can be

 ISBN: 978-84-617-8694-7

exploited to design low-complexity algorithms [7]–[11]. In other works, some authors propose an alternative to max-log approximation [16]–[13] and consider a new approach to the problem of computing (3). The basic idea is to define the following partitioning model, which is based in (1),

$$\mathbf{y} = \mathbf{H}\mathbf{s} + \mathbf{n} = [\overline{\mathbf{H}} \quad \widetilde{\mathbf{H}}] \quad [\overline{\mathbf{s}}^T \quad \widetilde{\mathbf{s}}^T]^T + \mathbf{v} = \overline{\mathbf{H}}\overline{\mathbf{s}} + \widetilde{\mathbf{H}}\widetilde{\mathbf{s}} + \mathbf{v} \tag{5}$$

where $\overline{\mathbf{H}} \in \mathbb{C}^{n_R \times n_s}$, $\widetilde{\mathbf{H}} \in \mathbb{C}^{n_R \times (n_T - n_s)}$, $\overline{\mathbf{s}} \in \Omega^{n_s}$ and $\widetilde{\mathbf{s}} \in \Omega^{n_T - n_s}$ for fixed $n_s \in 1, \cdots, n_T$.

The partitioned model carries intrinsically an optimal permutation of the columns of $\mathbf{H}$ that determines $\overline{\mathbf{H}}$ and $\widetilde{\mathbf{H}}$. It is important to note that this optimal permutation is difficult to find out and depends on the selected detection method.

2.2 SUMIS algorithm review

SUMIS [13] algorithm is composed by two main stages and employs the partitioning model (5). This partition uses a permutation based on $\mathbf{H}^T\mathbf{H}$ as it is explained in [13]. In Stage I, a first approximation to the LLRs values are computed and then, in Stage II, new refined LLRs values are calculated using these approximate values.

Here we give a brief revision of SUMIS algorithm [13]. It is important to note that in [13] referring to a symbol is equivalent to a bit, which is not the case here.

Stage I: The algorithm begins with the partitioned model (5) denoting the new model as

$$\overline{\mathbf{y}} = \overline{\mathbf{H}}\overline{\mathbf{s}} + \mathbf{e} \tag{6}$$

where $\mathbf{e} = \widetilde{\mathbf{H}}\widetilde{\mathbf{s}} + \mathbf{v}$ is a Gaussian stochastic vector $\mathbf{e} \sim \mathcal{N}(\mathbf{0}, \mathbf{Q})$ with $\mathbf{Q} = \widetilde{\mathbf{H}}\widetilde{\mathbf{H}}^T + \frac{N_o}{2}\mathbf{I}$. We compute the approximate $\lambda_{j,b}$ LLR using the next operator $\|\mathbf{x}\|^2_{\mathbf{Q}} \triangleq \mathbf{x}^T\mathbf{Q}^{-1}\mathbf{x}$, as

$$\lambda_{j,b} = \log \frac{\sum_{\overline{\mathbf{s}} \epsilon \chi^0_{j,b}} \exp(-\frac{1}{2}\|\mathbf{y} - \overline{\mathbf{H}}\overline{\mathbf{s}}\|^2_Q)}{\sum_{\overline{\mathbf{s}} \epsilon \chi^1_{j,b}} \exp(-\frac{1}{2}\|\mathbf{y} - \overline{\mathbf{H}}\overline{\mathbf{s}}\|^2_Q)}. \tag{7}$$

Stage I is performed for all bits $b = 1, \cdots, k$ in all symbols $j = 1, \cdots, n_T$.

Stage II: In the second stage, the LLR values are computed again over a new model. In this context, the interfering vector $\widetilde{\mathbf{s}}$ is suppressed in (6) and the purified model is given by

$$\mathbf{y}' \triangleq \mathbf{y} - \widetilde{\mathbf{H}}\mathbb{E}\{\widetilde{\mathbf{s}}|\mathbf{y}\} \approx \overline{\mathbf{H}}\overline{\mathbf{s}} + \mathbf{n}', \tag{8}$$

where $\mathbb{E}\{\widetilde{\mathbf{s}}|\mathbf{y}\}$ is the conditional expected value of vector $\widetilde{\mathbf{s}}$, and $n' \backsim \mathcal{N}(\mathbf{0}, \mathbf{Q}')$ with $\mathbf{Q}' \triangleq \widetilde{\mathbf{H}}\widetilde{\mathbf{\Upsilon}}\widetilde{\mathbf{H}}^T + \frac{N_o}{2}\mathbf{I}$. $\widetilde{\mathbf{\Upsilon}}$ is the conditional covariance matrix of $\widetilde{\mathbf{s}}$ and can be computed by

$$\widetilde{\mathbf{\Upsilon}} = \mathbb{E}\{\mathrm{diag}(\widetilde{\mathbf{s}})^2|\mathbf{y}\} - \mathbb{E}\{\mathrm{diag}(\widetilde{\mathbf{s}})|\mathbf{y}\}^2 \tag{9}$$

 ISBN: 978-84-617-8694-7

where diag(∇) gives a diagonal matrix with the elements of ∇ vector on its diagonal. Hence, the refined LLR values can be computed as

$$L_{j,b} \approx \log \frac{\sum_{\bar{\mathbf{s}} \epsilon \chi^0_{j,b}} \exp(-\frac{1}{2}\|\mathbf{y}' - \overline{\overline{\mathbf{H}}}\bar{\mathbf{s}}\|^2_{Q'})}{\sum_{\bar{\mathbf{s}} \epsilon \chi^1_{j,b}} \exp(-\frac{1}{2}\|\mathbf{y}' - \overline{\overline{\mathbf{H}}}\bar{\mathbf{s}}\|^2_{Q'})}. \quad (10)$$

Note that the processing per bit can be performed in parallel. This allows for massively parallel algorithmic implementations such as those presented in this paper.

3 Proposed Parallelization

The use of the last generation of High Performance Computing (HPC) systems such as multi-core CPUs has become attractive for the efficient implementation of parallel signal processing algorithms with high computational requirements, such as high throughput MIMO detectors [17] or fast LDPC. The implementation of advanced algorithms able to use both architectures is crucial in MIMO research, since it allows to fully exploit the capabilities of the modern machine architectures and to reduce the response time of computationally expensive problems. The programming challenge involves the developers to know in depth different programming languages and the features of the architecture. In this sense, high performance libraries become valuable tools for specialists of a particular field, since they ease the development of scientific codes. The SUMIS algorithm computes the $\lambda_{j,b}$ and $L_{j,b}$ values using (7) and (10) respectively, where the number of terms in the two summations over $\bar{\mathbf{s}}$ is M^{n_s}. This implies that SUMIS algorithm has to compute the terms $\exp(-\frac{1}{2}\|\mathbf{y} - \overline{\overline{\mathbf{H}}}\bar{\mathbf{s}}\|^2_Q)$, a number of times M^{n_s} to get the total $\lambda_{j,b}$ or $L_{j,b}$ values in each stage. That is why almost the 98% of the total time of the detector is consumed in these two steps. Therefore, the parallelization is focused in reduce the computational cost of (7) and (10).

For the multicore implementation we are using the Intel Math Kernel Library (MKL) [18], which is composed of several optimized math routines including BLAS, LAPACK, ScaLAPACK, sparse solvers, fast Fourier transforms, etc, and is optimized specifically for Intel processors.

4 Results

Using Monte Carlo simulations we evaluate the performance of SUMIS and the algorithms chosen for comparison in terms of Bit Error Rate and Mutual Information. The transmitted symbols are assumed to be independent and uniformly distributed. The transmitted bits are encoded using a 1/2 LDPC code of codeword size 648 bits, which is available from http://www.csl.cornell.edu/vstuder/software ldpc.html and implements a LDCPC code from the IEEE 802.11n wireless LAN standard. There is no iteration between

 ISBN: 978-84-617-8694-7

the detector and the decoder and the sum-product algorithm option has been chosen as the decoding option. A machine with two multicore Intel processors has been employed. Each multicore is an Intel Xeon CPU E5-2697 at 2.70 GHz with 12 cores per CPU. We measure the execution time of the detection, with different number of antennas and constellation sizes to evaluate the SUMIS efficiency of the proposed parallel prototypes.

It is interesting to note how parallel versions allow to boost the performance of the system with a speed comparable to a low-complexity linear detector such as MMSE.

Thus, we can detect signals with a similar and even higher throughput than the MMSE detector with much better BER.

5 Conclusion

In this paper, we have proposed and analyzed parallel SUMIS Soft Detector implementations. We have also studied its performance in terms of BER performance, mutual information, complexity and speedup. These measures have been compared with those of the state-of-art soft-output detectors such as FPFSD and MMSE detectors. Furthermore, the experiments have been realized considering very large MIMO systems reaching up to 200 transmitter/receiver antennas and up to 1024-QAM constellations. This comparison shows the robust behavior of the SUMIS algorithm. This detector behaves much more efficiently than the other detectors in terms of BER, achieving up to 5 dB improvement in SNR compared to FPFSD.

Acknowledgment

This work has been supported by the RACHEL project of the Programa Estatal de Investigación, Desarrollo e Innovación Orientada a los Retos de la Sociedad (TEC2013-47141-C4-4-R) and by the Generalitat Valenciana through the PROMETEO FASE II 2014/003 project.

References

[1] E. Bigilieri, R. Calderbank, A. Constantinides, and A. Goldsmith, *MIMO Wireless Communications*. Cambridge Univ. Press, 2010.

[2] *Specification Framework for TGac. IEEE 802.11-09/0992r21*, IEEE P802.11ac, Jan 2011.

[3] *Draft Amendment to IEEE Standard for Local and Metropolitan Area Networks Part 16: Air Interface for Fixed and Mobile Broadband Wireless Access Systems: Multihop Relay Specification*, IEEE P802.16j/D9, Feb 2009.

 ISBN: 978-84-617-8694-7

[4] *Overview of 3GPP Release 10 V0.0.8 (2010-09) [Online]*, online at http://www.3gpp.org/ftp/Information/WORKPLAN/DescriptionReleases/Rel-10 description 20100924.zip.

[5] E. G. Larsson, O. Edfors, F. Tufvesson, and T. L. Marzetta, "Massive MIMO for next generation wireless systems," *IEEE Communications Magazine*, vol. 52, no. 2, pp. 186–195, 2014.

[6] F. Rusek, D. Persson, B. Lau, E. G. Larsson, T. Marzetta, O. Edfors, and F. Tufvesson, "Scaling up MIMO: Opportunities and challenges with very large arrays," *IEEE Signal Proc. Magazine*, vol. 30, no. 1, pp. 40–60, Jan. 2013.

[7] C. Studer, A. Burg, and H. Bölcskei, "Soft-output sphere decoding: algorithms and VLSI implementation," *IEEE J. Sel. Areas Commun.*, vol. 26, no. 2, pp. 290–300, 2008.

[8] R. Wang and G. B. Giannakis, "Approaching MIMO channel capacity with reduced-complexity soft sphere decoding," *Wireless Communications and Networking Conference, 2004. WCNC. 2004 IEEE*, vol. 3, pp. 1620–1625, 2004.

[9] L. G. Barbero, T. Ratnarajah, and C. Cowan, "A low-complexity soft-MIMO detector based on the fixed-complexity sphere decoder," *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, pp. 2669–2672, 2008.

[10] B. Hochwald and S. ten Brink, "Achieving Near-Capacity on a Multiple-Antenna Channel," *IEEE Trans. Commun.*, vol. 51, pp. 389–399, 2003.

[11] X. Wu and J. Thompson, "A fixed-complexity soft-MIMO detector via parallel candidate adding scheme and its FPGA implementation," *Commun. Lett.*, vol. 15, no. 2, pp. 241–243, Feb. 2011.

[12] D. Persson and E. G. Larsson, "Partial Marginalization soft MIMO detection with higher order constellations," *IEEE Trans. on Signal Proccesing*, vol. 59, no. 1, pp. 453–458, Jan. 2011.

[13] M. Cîrkić and E. G. Larsson, "SUMIS: Near-optimal soft-in soft-out MIMO detection with low and fixed complexity," *Signal Processing, IEEE Transactions on*, vol. 62, no. 12, pp. 3084–3097, June 2014.

[14] M. Á. Simarro, V. M. García, F.-J. Martínez-Zaldívar, A. Gonzalez, and A. Vidal, "Complexity reduction of SUMIS MIMO soft detection based on box optimization for large systems," *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*, pp. 3821–3825, 2016.

 ISBN: 978-84-617-8694-7

[15] Z. Guo and P. Nilsson, "Algorithm and implementation of the K-best sphere decoding for MIMO detection," *IEEE Journal on Selected Areas in Communications*, vol. 24, no. 3, pp. 491–503, 2006.

[16] E. G. Larsson and J. Jaldén, "Fixed-complexity soft MIMO detection via partial marginalization," *IEEE Trans. on Signal Proccesing*, vol. 56, no. 1, pp. 3397–3407, Aug. 2008.

[17] D. Wu, J. Eilert, and D. Liu, "Implementation of a high-speed MIMO soft-output symbol detector for software defined radio," *Journal of Signal Processing Systems*, vol. 63, no. 1, pp. 27–37, 2011.

[18] *Intel MKL Reference Manual (2015) [Online]*, online at: https://software.intel.com/en-us/articles/mkl-reference-manual.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

Alternation Direction Implicit Method for the Aliev-Panfilov Monodomain Model

Zaynab Rammal[1] and Youssef Belhamadia[1]

[1] *Department of Mathematics and Statistics, American University of Sharjah, United Arab Emirates.*

emails: g00056172@aus.edu, ybelhamadia@aus.edu

Résumé

In this work, an Alternating Direction Implicit method (ADI) for solving monodomain model that describes the propagation of the electrical potential in cardiac tissue is presented. The proposed numerical algorithm is of second order in space and time and allows the use of large grids for accurate simulation of the cardiac electrical wave fronts. Two-dimensional numerical results are presented illustrating the advantages of the proposed method in terms of computational time and memory consumption.

Key words: Alternating Direction Implicit Method, Monodomain Model, Aliev-Panfilov Model.

1 Introduction

The bidomain model and its simplified version, the monodomain model, are considered the most used mathematical equations for simulating the cardiac electrical activity. The numerical solution of these models is a computationally challenging problem and is known that they require high space and time resolutions (see Clayton et al. [5]).

Many methods have been developed in the literature to decrease the computational cost of the simulations. Different approaches were implemented for time discretization, including semi-implicit methods (see Keener and Bogar [8], Franzone and Pavarino [10], and Bourgault and Ethier [9]), fully implicit methods (see Ying et al. [19], Dal et al. [7], and Belhamadia [2]), and operator splitting methods (see Schroll et al. [16], Spiteri and Ziaratgahi [17], and Sundnes et al. [18]). A variety of numerical methods were also used for space discretization using both uniform meshes (see Franzone et al. [11] and Scacchi [15]) and adapted meshes (see Whiteley [20], Cherry et al. [6], Belhamadia et al. [3, 4] and the references therein).

 ISBN: 978-84-608-6082-2

In this work, we suggest implementing the Alternating Direction Implicit Method (ADI) for solving the Aliev-Panfilov monodomain model. The proposed method is based on the original work of Peaceman and Rachford where the ADI method was proposed for the first time as a type of finite difference method for solving two dimensional parabolic and elliptic Partial differential equation. The ADI algorithm presented in this work is of se-cond order in space and time and has the advantage to use large grids for accurate and efficient simulation while reducing the computational time and the memory consumption. A comparison between the suggested scheme and the second order Crank-Nicolson-Adams-Bashforth semi-implicit scheme (CNAB) proposed in [9] is performed showing the accuracy of the suggested method.

This paper is organized as follows. Next section is devoted to the monodomain model, section 2 describes the ADI method and CNAB method, and section 3 presents two-dimensional numerical quantitative results.

2 Aliev-Panfilov Monodomain Model

A simplification of the so-called bidomain model, by assuming equal anisotropy, is the monodomain model. This model describes the cardiac electrical activity and has the form :

$$\begin{cases} \dfrac{\partial V}{\partial t} - \nabla \cdot (D \nabla V) = I_{ion}(V, W_n) \\ \dfrac{\partial W_n}{\partial t} = g(V, W_n) \end{cases} \tag{1}$$

Where V is the transmembrane potential and W_n is a set of cell-level variables such as ion concentrations and gating variables. In this work, we will use the Aliev-Panfilov ionic model [1] that contains only one gating variable but it still reproduces realistic shape of the cardiac action potential. This ionic model consists of the following nonlinear equations :

$$I_{ion}(V, W) = kV(V - \alpha)(1 - V) - VW, \tag{2}$$

$$g(V, W) = \left(\epsilon + \frac{\mu_1}{\mu_2 + V}\right)(-W - kV\,(V - \alpha - 1))\,. \tag{3}$$

All the parameters including the time t are dimensionless and will be determined in the numerical simulation section. The boundary conditions considered are homogeneous Neumann boundary conditions.

3 Numerical Method

In this work, we will consider an alternating direction implicit method (ADI) to solve the Aliev-Panfilov monodomain model. We limit our work to the two dimensional case. The

 ISBN: 978-84-608-6082-2

ADI method is a type of finite difference method and thus the continuous space domain should be discretized into a mesh with a finite number of grid points. To ensure the second order in space, we use the second-order central difference operators δ_x^2 and δ_y^2 for the second derivatives with respect to x and y respectively :

$$(\delta_x^2 V^n)_{i,j} = \frac{V_{i+1,j}^n - 2V_{i,j}^n + V_{i-1,j}^n}{(\Delta x)^2}, \qquad (\delta_y^2 V^n)_{i,j} = \frac{V_{i,j+1}^n - 2V_{i,j}^n + V_{i,j-1}^n}{(\Delta y)^2}. \tag{4}$$

To ensure the second order in time the following time-dependent algorithm will be used :

Step 1 : on the interval $[t_n, t_{n+\frac{1}{2}}]$

We first get the solutions $(V^{n+\frac{1}{2}}, W^{n+\frac{1}{2}})$ at the mid time $t_{n+\frac{1}{2}}$, based on the following equations :

$$\begin{cases} \dfrac{V^{n+\frac{1}{2}} - V^n}{\frac{\Delta t}{2}} &= D\delta_x^2 V^{n+\frac{1}{2}} + D\delta_y^2 V^n + \dfrac{3}{2} I_{ion}(V^n, W^n) - \dfrac{1}{2} I_{ion}(V^{n-1}, W^{n-1}) \\ \dfrac{W^{n+\frac{1}{2}} - W^n}{\frac{\Delta t}{2}} &= \dfrac{3}{2} g(V^n, W^n) - \dfrac{1}{2} g(V^{n-1}, W^{n-1}) \end{cases} \tag{5}$$

This steps applies an implicit update in the x-direction and explicit update in the y-direction on the time interval $[t_n, t_{n+\frac{1}{2}}]$.

Step 2 : on the interval $[t_{n+\frac{1}{2}}, t_{n+1}]$

The final solutions (V^{n+1}, W^{n+1}) at the time step t_{n+1} is then calculated using the following equations :

$$\begin{cases} \dfrac{V^{n+1} - V^{n+\frac{1}{2}}}{\frac{\Delta t}{2}} &= D\delta_x^2 V^{n+\frac{1}{2}} + D\delta_y^2 V^{n+1} + \dfrac{3}{2} I_{ion}(V^n, W^n) - \dfrac{1}{2} I_{ion}(V^{n-1}, W^{n-1}) \\ \dfrac{W^{n+1} - W^{n+\frac{1}{2}}}{\frac{\Delta t}{2}} &= \dfrac{3}{2} g(V^n, W^n) - \dfrac{1}{2} g(V^{n-1}, W^{n-1}) \end{cases} \tag{6}$$

In this step, the method applies an implicit update in the y-direction and explicit update in the x-direction on the time interval $[t_{n+\frac{1}{2}}, t_{n+1}]$.

The above algorithm is of second order is space and time (see Rammal [14]) and is based on the work of Peaceman and Rachford [13]. Since the monodomain model is combined with homegenous homogenous boundary conditions, and in order to keep the second-order accuracy, the central-difference approximation will be used to discretize these boundary

conditions. This implies the introduction of the virtual points by expanding the region with Δx to the left and right and with Δy to the bottom and top, and for more details the reader is referred to [14].

To verify the efficiency of the above-mentioned ADI method, a comparison with the second order Crank-Nicolson-Adams-Bashforth semi-implicit scheme (CNAB) will be presented. The CNAB method was proposed in [9] as an impressive scheme compared to different type of first, second, and third other schemes and this is why it was chosen for this study. The algorithm for CNAB method for solving the Aliev-Panfilov monodomain model takes the form :

$$\begin{cases} \dfrac{V^{n+1}-V^n}{\Delta t} &= \dfrac{1}{2}(D\delta_x^2 V^{n+1} + D\delta_y^2 V^{n+1}) + \dfrac{1}{2}(D\delta_x^2 V^n + D\delta_y^2 V^n) \\ & \quad + \dfrac{3}{2} I_{ion}(V^n, W^n) - \dfrac{1}{2} I_{ion}(V^{n-1}, W^{n-1}) \\ \dfrac{W^{n+1}-W^n}{\Delta t} &= \dfrac{3}{2} g(V^n, W^n) - \dfrac{1}{2} g(V^{n-1}, W^{n-1}) \end{cases} \tag{7}$$

This method is of second order in space and time and for more details the reader is refereed to [9].

4 Numerical Results

This section is devoted to a quantitative comparison of the solutions obtained with ADI and CNAB methods. The computational domain is $[0,\ 100] \times [0,\ 100]$. The initial transmembrane potential V and the recovery variable W are given by :

$$V(x,t) = \begin{cases} 1 & \text{if } \sqrt{(x-50)^2 + (y-50)^2} < 10 \\ 0 & \text{if } \sqrt{(x-50)^2 + (y-50)^2} \geq 10, \end{cases} \quad \text{and } W(x,t) = 0.$$

In the numerical simulations, the following physical parameters are used :

parameter	k	α	ϵ	μ_1	μ_2	D
value	8	0.15	0.002	0.2	0.3	1

The time evolution of the transmembrane potential is presented in Figure 1. As can be seen, the wave is initiated as a circular wave in the center of the computational domain and

 ISBN: 978-84-608-6082-2

it propagates in a similar manner till it reaches the boundaries, an expected behaviour of this model.

Now, to illustrate the efficiency of the ADI scheme applied to the Aliev-Panfilov monodomain model, a comparison with CNAB scheme is presented in terms of memory and run time. Both methods are tested using the same number of time iterations to obtain the same numerical solution, are executed using MATLAB, and are run using the same computer with Intel(R) Core(TM) i7 − 4790 CPU processor, and 8.00 GB installed memory functioning with a 64-bit operating system.

First, figure 2 shows a memory consumption required for both the CNAB and ADI methods. As can be seen in this figure, the memory consumption of the CNAB increases dramatically. However, the ADI keeps running even with large grids, and while doing so, the memory consumption remains considerably low compared to what is consumed by CNAB method.

Second, figure 3 shows the computational time required for both the CNAB and ADI methods. As can be seen, the time required to run CNAB with 400 space steps is close to that needed by the ADI to run with 1200 space steps. The rapid results provided by the ADI is a powerful advantage for simulating a model that aims to capture the behaviour of a sharp cardiac electrical wave.

5 Conclusion

A numerical algorithm based on the alternation direction implicit method for solving the Aliev-Panfilov monodomain model was presented. The efficiency of the this algorithm was provided though demonstrating the difference in the run-time and memory requirements between this method and Crank-Nicolson-Adams-Bashforth method. The numerical simulations has shown that the ADI algorithm not only speeds up the run-time but also reduces memory usage over the standard numerical methods used for solving the electrocardiology models.

Références

[1] R. Aliev and A. Panfilov, "A Simple Two-variable Model of Cardiac Excitation", *Chaos, Solirons and Fractals*, vol. 7, pp. 293-301, 1996.

[2] Y. Belhamadia, "A Time-Dependent Adaptive Remeshing for Electrical Waves of the Heart", *IEEE Transactions on Biomedical Engineering*, vol. 55, pp. 443-452, 2008.

[3] Y. Belhamadia, A. Fortin, and Y. Bourgault, "Towards Accurate Numerical Method for Monodomain Equations Using a Realistic Heart Geometry", *Mathematical Biosciences*, vol. 220, pp. 89-101, 2009.

 ISBN: 978-84-608-6082-2

[4] Y. Belhamadia, A. Fortin, and Y. Bourgault, "On the Performance of Anisotropic Mesh Adaptation for Spiral and Scroll Wave Turbulence Dynamics in Reaction-Diffusion Systems", *Journal of Computational and Applied Mathematics*, vol. 271, pp. 233-246, 2014.

[5] R.H. Clayton and O. Bernus and E.M. Cherry and H. Dierckx and F.H. Fenton and L. Mirabella and A.V. Panfilov and F.B. Sachse and G. Seemann and H. Zhang, "Models of cardiac tissue electrophysiology : Progress, challenges and open questions",*Progress in Biophysics and Molecular Biology*, vol. 104(1-3), pp. 22-48, 2011.

[6] E. Cherry, H. Greenside, and C. Henriquez, "Efficient Simulation of Three Dimensional Anisotropic Cardiac Tissue using an Adaptive Mesh Refinement Method", *Chaos*, vol. 13, pp. 853-865, 2003.

[7] H. Dal, S. Goktepe, M. Kaliske, and E. Kuhl, "A Fully Implicit Finite Element Method for Bidomain Models of Cardiac Electromechanics", *Computer Methods in Applied Mechanics and Engineering*, vol. 253, pp. 323-336, 2013.

[8] J. Keener and K. Bogar, "A Numerical Method for the Solution of the Bidomain Equations in Cardiac Tissue", *Chaos*, vol. 1, pp. 234-241, 1998.

[9] M. Ethier and Y. Bourgault, "Semi-Implicit Time-Discretization Schemes for the Bidomain Model", *Journal on Numerical Analysis*, vol. 46, pp. 2443-2468, 2008.

[10] P. Franzone and L. Pavarino, "A Parallel Solver for Reaction Diffusion Systems in Computational Electrocardiology", *Mathematical Models and Methods in Applied Sciences*, vol. 14, pp. 883-911, 2004.

[11] P. Franzone, L. Pavarino and S. Scacchi, "A Comparison of Coupled and Uncoupled Solvers for the Cardiac Bidomain Model", *Mathematical Modelling and Numerical Analysis*, vol. 47, pp. 1017-1035, 2013.

[12] E. Heidenreich, F. Gaspar, J. Ferrero and J. Rodriguez, "Compact Schemes for Anisotropic Reaction-Diffusion Equations with Adaptive Time Step", *International Journal for Numerical Methods in Engineering*, vol. 82, pp. 1022-1043, 2009.

[13] D. Peaceman and H. Rachford, "The Numerical Solution of Parabolic and Elliptic Differential Equations", *Journal of the Society for Industrial and Applied Mathematics*, vol. 3, pp. 28-41, 1955.

[14] Z. Rammal, "Alternating Direction Implicit Method for the Electro-Cardiology Models", Master Thesis, American University of Sharjah, 2017.

[15] S. Scacchi, "A Multilevel Hybrid Newton-Krylov-Schwarz Method for the Bidomain Model of Electrocardiology", *Computer Methods in Applied Mechanics and Engineering*, vol. 200, pp. 717-725, 2011.

[16] H. Schroll, G. Lines and A. Tveito, "On the Accuracy of Operator Splitting for the Monodomain Model of Electrophysiology", *International Journal of Computer Mathematics*, vol. 84, pp. 871-885, 2007.

 ISBN: 978-84-608-6082-2

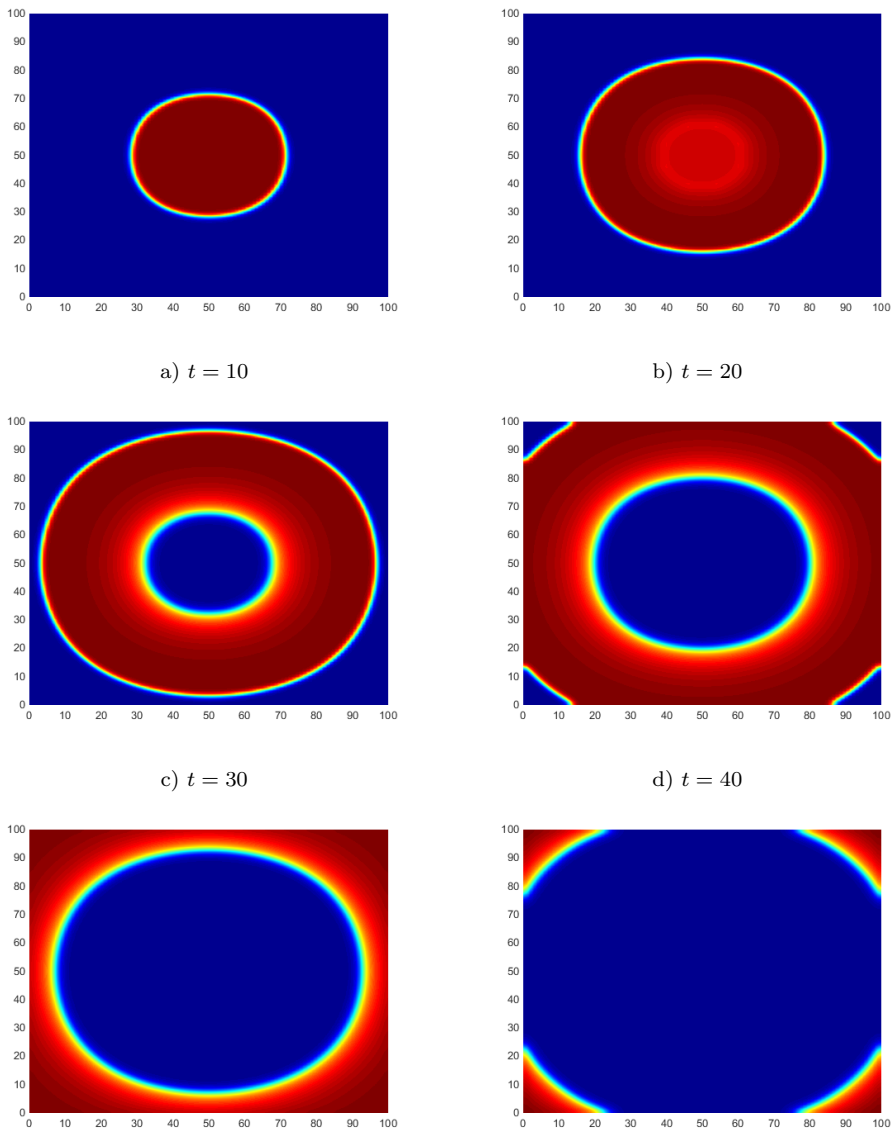
[17] R. Spiteri and S. Ziaratgahi, "Operator Splitting for the Bidomain Model Revisited", *Journal of Computational and Applied Mathematics*, vol. 296, pp. 550-563, 2016.

[18] J. Sundnes, G. Lines, and A. Tveito, "An Operator Splitting Method for Solving the Bidomain Equations Coupled to a Volume Conductor Model for the Torso", *Mathematical Biosciences*, vol. 194, pp. 233-248, 2005.

[19] W. Ying, D. Rose and C. Henriquez, "Efficient Fully Implicit Time Integration Methods for Modeling Cardiac Dynamics", *IEEE Transactions on Biomedical Engineering*, vol. 55, pp. 2701-2711, 2008.

[20] J. Whiteley, "Physiology Driven Adaptivity for the Numerical Solution of the Bidomain Equations", *Annals of Biomedical Engineering*, vol. 35, pp. 1510-1520, 2007.

 ISBN: 978-84-608-6082-2



a) $t = 10$ b) $t = 20$

c) $t = 30$ d) $t = 40$

e) $t = 50$ f) $t = 60$

Figure 1 – Evolution of transmembrane potential over time for the 2D Aliev-Panfilov monodomain model.

 ISBN: 978-84-608-6082-2

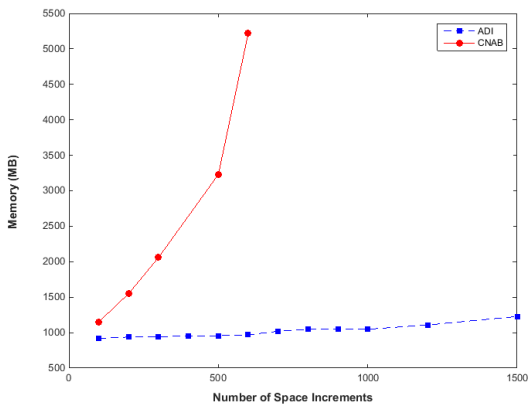


Figure 2 – Memory Consumption of the *ADI* and *CNAB* Methods

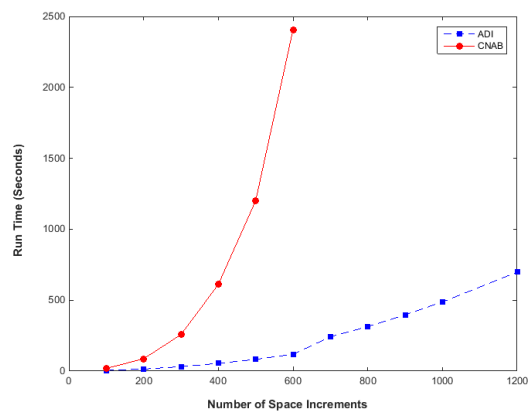


Figure 3 – Run Time of the *ADI* and *CNAB* Methods

 ISBN: 978-84-608-6082-2

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

Zero Forcing in Maximal Outerplanar Graphs

Sandra Ranilla-Cortina[1], Gregorio Hernández[1] and Jose Ranilla[2]

[1] *Departamento de Matemática Aplicada a las TIC, Universidad Politécnica de Madrid*

[2] *Departamento de Informática, Universidad de Oviedo*

emails: s.ranillac@alumnos.upm.es, gregorio@fi.upm.es, ranilla@uniovi.es

Abstract

In this work we study the zero forcing problem and some variants of it attending to the connectivity of the subgraph generated by the zero forcing set. A set Z of vertices of a graph G is a zero forcing set of G if iteratively adding to Z vertices from $V(G)\backslash Z$ that are the unique neighbor in $V(G)\backslash Z$ of some vertex in Z, results in the entire vertex set $V(G)$ of G. The zero forcing number $Z(G)$ of G is the minimum cardinality of a zero forcing set of G.
This study is restricted to maximal outerplanar graphs. We establish tight combinatorial bounds for zero forcing, total zero forcing and connected zero forcing for any n-vertex maximal outerplanar graph.

Key words: zero forcing, maximal outerplanar graphs
MSC 2000: AMS codes 05C69, 05C85

1 Introduction

Given a graph $G = (V, E)$ a zero forcing set is a set $Z \subseteq V$ such that every proper subset $\bar{Z} \subseteq V$ with $\bar{Z} \subseteq Z$ contains a vertex that has exactly one neighbour in $V\backslash\bar{Z}$. Equivalently, the Zero Forcing process is a discrete-time process in which we have a set of vertices of a graph G that is initially colored black, while the remaining vertices are colored white. At each time step, the color change rule is applied. Once a vertex has been changed to black, it remains black forever. The process terminates if all of the vertices are colored black in finite time. Otherwise, if there are some white vertices that are unable to be colored in finite time, then we may artificially halt the process. The zero forcing number $Z(G)$ of G is the minimum cardinality of a zero forcing set of G.

 ISBN: 978-84-617-8694-7

The zero forcing problem was introduced in [2] as an upper bound for the maximum nullity of real symmetric matrices whose nonzero pattern of off-diagonal entries is described by a given graph, and independently by mathematical physicists studying control of quantum systems. Furthermore, it is related with the power domination concept, which was first studied by Haynes et al. [3] for monitoring electric power systems. The relationship between both was stablished in [4].

Results on the zero forcing number for trees and unicyclic graphs were stablished in [5]. Also, Davila et al. studied this problem for graphs with a given girth in [6]. They proved that $Z(G) \geq \delta + (\delta - 2)(\delta - 3)$ where G is a graph with a given girth $g \geq 3$ and minimum degree $\delta \geq 2$. Moreover, results on the connected and total zero forcing were shown in [7] and [8] respectively.

The domination problem in outerplanar graphs has received special attention in recent years [1]. A graph is outerplanar if it has a crossing-free embedding in the plane such that all vertices are on the boundary of its outer face (the unbounded face). An outerplanar graph is maximal if it is not possible to add an edge such that the resulting graph is still outerplanar. A maximal outerplanar graph embedded in the plane corresponds to a triangulation of a polygon.

In this study we establish tight combinatorial bounds for the zero forcing number and the following zero forcing variants: connected and total. Our results are restricted to maximal outerplanar graphs.

2 Connected and Total Zero Forcing in MOPs

Let G be a graph of order n and Z a zero forcing set of G. By imposing different connectivity conditions to the subgraph generated by Z, $G[Z]$, we have some variants on zero forcing. In this work we will study the following ones: connected zero forcing, if $G[Z]$ is connected; total zero forcing, if $G[Z]$ does not contain isolated vertices. We denote the connected and total forcing number by $Z_C(G)$ and $Z_T(G)$ respectively.

A triangle face T is a separator triangle of G, maximal outerplanar graph, if it has no edges in the outer face. The number of separator triangles of G is denoted by t. A MOP is serpentine if $t = 0$. We denote with G_Δ to the subgraph formed by the separator triangles of G.

The dual tree of G (weak dual) is denoted by H. We describe as serpentine leaf to the

 ISBN: 978-84-617-8694-7

set of triangles corresponding to the only path in H that connects a leaf of H with a vertex u with $degree(u) > 2$. We denote the number of serpentine leafs by h, and the number of serpentine leafs with fan structure, called fan leafs, by h_F such as $h_F \leq h$. In addition, we denoted by c to the number of minimum paths in H between vertices with degree 3.

3 Results

According to the previous definitions we establish the following theorems.

Theorem 1. *If G is a serpentine graph, then*

$$Z(G) = Z_T(G) = Z_C(G) = 2$$

Theorem 2. *If G is a maximal outerplanar graph with $t = 1$ and $h_F \geq 0$, then*

1. *If $h_F \geq 2$, $Z(G) = Z_T(G) = 3$*
2. *If $h_F = 1$, $Z(G) = 3$ and $Z_T(G) = 4$*
3. *If $h_F = 0$, $Z(G) = Z_T(G) = 4$*

In addition, the zero forcing set of minimum cardinality is explicitly obtained for each maximal outerplanar graph.

Theorem 3. *If G is a maximal outerplanar graph with h serpentine leafs and h_F fan leafs, whose G_Δ is connected, then*

$$Z(G) \leq Z_T(G) \leq 2\lceil \tfrac{h}{2} \rceil$$

with equality if and only if $h_F = 0$.

In the extended version of this work, we explain how the zero forcing set of minimum cardinality is obtained when $h_F = 0$.

Now, take into account the above theorems, we may establish an upper bound on the zero forcing number and the total zero forcing number for any maximal outerplanar graph.

Theorem 4. *If G is a maximal outerplanar graph with h serpentine leafs and c serpentine paths, then*

$$Z(G) \leq Z_T(G) \leq 2\lceil \tfrac{h+c}{2} \rceil$$

and these bounds are tight when G_Δ is connected.

As future work we are working on an algorithm to find the minimum zero forcing set for any maximal outerplanar graph. New theorems and results will be presented in an extended version of this work.

 ISBN: 978-84-617-8694-7

Acknowledgements

This work has been partially supported by the Ministry of Economy and Competitiveness ("Ministerio de Economía y Competitividad") from Spain/FEDER under grant TEC2015-67387-C4-3-R.

References

[1] C. N. Campos, Y. Wakabayashi *On dominating set of maximal outerplanar graphs*, Discrete Applied Mathematics, **161** (2013), 330-335.

[2] AIM Special Work Group *Zero forcing sets and the minimum rank of graphs*, Linear Algebra and its Applications, **428** (7), (2008), 1628-1648.

[3] T. W. Haynes, S. Hedetniemi, M. A. Henning *Domination in Graphs Applied to Electric Power Networks*, SIAM Journal on Discrete Mathematics, **15** (2002), 519-529.

[4] K. F. Benson, D. Ferrero, M. Flagg, V. Furst, L. Hogben, V. Vasilevska, B. Wissman *Power domination and zero forcing*, arXiv:1510.02421, (2015).

[5] L. Eroh, C. X. Kang, E. Yi *A comparision between the Metric Dimension and Zero Forcing of Trees and Unicyclic Graph*, Acta Mathematica Sinica -English Series, **33** (6), (2017), 731-747.

[6] R. Davila, T. Kalinowski, S. Stephen *The Zero Forcing Number of Graphs with Given Girth*, arXiv:1611.06557, (2016).

[7] B. Brimkov, R. Davila *Characterizations of the Connected Forcing Number of a Graph*, arXiv:1611.07513, (2016).

[8] R. Davila, M. A. Henning *On the Total Forcing Number of a Graph*, arXiv:1702.06035, (2017).

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

Conservation laws and symmetries for a generalized Rosenau-RLW equation

E. Recio[1], T.M. Garrido[1], R. de la Rosa [1] and M.S. Bruzón[1]

[1] *Department of Mathematics, University of Cádiz*

emails: elena.recio@uca.es, tamara.garrido@uca.es, rafael.delarosa@uca.es, m.bruzon@uca.es

Abstract

This work considers a generalized family of Rosenau-RLW equations depending on an arbitrary function $f(u)$. This family of equations is interesting from the point of view of shallow water waves theory, because particular equations in the family have been shown to model wave interactions that are not possible to be described with the Korteweg-de Vries equation. We will present a complete classification of conservation laws for this dispersive wave equation and we will apply Lie symmetry analysis to the equation in order to find exact solutions.

Key words: Nonlinear partial differential equation, Lie symmetries, Conservation laws

1 Introduction

The Rosenau equation [8, 9], given by

$$u_t \pm (u + u^2)_x + u_{xxxxt} = 0, \tag{1}$$

is a fifth-order dispersive wave equation that provides a good model for a two-sided wave propagation, nonlinear wave-wave and wave-wall interactions. Some of these interactions cannot be described by the Korteweg-de Vries (KdV) equation $u_t - \alpha u u_x + u_{xxx} = 0$.

Another well-known third-order dispersive wave equation that improves the KdV equation for long waves is the Benjamin-Bona-Mahony or regularized long-wave (RLW) equation, given by

$$u_t + (u + u^2)_x - u_{xxt} = 0. \tag{2}$$

 ISBN: 978-84-617-8694-7

Recently, a viscous Rosenau equation, given by

$$u_t + au_x + \epsilon(u^4)_x - bu_{xx} - cu_{xxt} - du_{xxxxt} = 0, \tag{3}$$

with a, ϵ, b, c and d arbitrary constants, has been considered and its solitary wave solution has been studied numerically [7].

In this paper, we consider a more general family of Rosenau-RLW equations, that includes equations (1), (2) and (3), given by

$$\Delta \equiv u_t + au_x + (f(u))_x - bu_{xx} - cu_{xxt} - du_{xxxxt} = 0, \tag{4}$$

for $u(t, x)$ where the nonlinear advective term depends on an arbitrary function $f(u)$ and a, b, c and d are arbitrary constants. Since particular equations in the family (4) have been shown to describe some interactions that cannot be explained with the KdV equation, we are interested in finding exact solutions that model this kind of behaviour.

Local conservation laws are continuity equations that provide physical conserved quantities for all solutions. In addition, they can be used to check the accuracy of numerical methods. We will carry out a complete classification of local conservation laws for the Rosenau-RLW equation (4) by using a general method based on multipliers [1, 2, 3, 4, 5].

Symmetry methods applied to partial differential equations lead to reductions and exact invariant solutions. In particular, invariant solutions can be used to study analytical properties such as asymptotic behaviour and blow-up behavior and to check the accuracy of numerical integrators. A point symmetry classification of equation (4) is presented in terms of the function $f(u)$. The required theory can be found in [5, 6].

2 Conservation Laws

A local conservation law for the generalized Rosenau-RLW equation (4) is a space-time divergence expression

$$D_t T + D_x X = 0 \tag{5}$$

holding for all solutions $u(t, x)$ of equation (4), where the conserved density T and the flux X are functions of t, x, u and its derivatives. If a conserved density is a total x-derivative, $T = D_x\Psi$, when it is restricted to the solution space, then the conservation law (5) holds trivially, with the flux being a total t-derivative, $X = -D_t\Psi$, when it is restricted to the solution space. Such conservation laws are called locally trivial, and any conservation laws that differ by a locally trivial conservation law are called locally equivalent. The set of all admitted conservation laws forms a vector space, in which the set of locally trivial conservation laws is a subspace.

All nontrivial local conservation laws can be obtained by the multiplier method [6, 3, 4, 1]. A function Q of t, x u, and derivatives of u is a multiplier for a local conservation law iff

 ISBN: 978-84-617-8694-7

the product of Q and the equation (4) is a total divergence with respect to t and x. All multipliers can be determined by the condition that a function $h(t, x, u, u_x, u_t, u_{xx}, u_{tx}, u_{tt}, \ldots)$ is such a divergence iff $E_u(h) = 0$ holds identically, where E_u denotes the Euler operator. It is straightforward to show that multipliers are related to conserved densities by $Q = E_u(T)$. As a consequence, all locally equivalent conservation laws have the same multiplier, and hence there is a one-to-one correspondence between multipliers and conserved densities (up to local equivalence). This result holds more generally for any evolution equation [6, 3, 4, 1].

For dispersive nonlinear evolution equations, conservation laws of physical importance come from low-order multipliers [1]. For the Rosenau-RLW equation (4), the dependence of low-order multipliers Q in terms of u and derivatives of u is given by those variables that can be differentiated to obtain a leading derivative of equation (4). This yields

$$Q(t, x, u, u_t, u_x, u_{xt}, u_{xx}, u_{xxt}, u_{xxx}, u_{xxxt}, u_{xxxx}) \tag{6}$$

as the general form for a low-order multiplier. All low-order multipliers (6) can be found by solving the determining equation $E_u((u_t + au_x + (f(u))_x - bu_{xx} - cu_{xxt} - du_{xxxxt})Q) = 0$, yielding an overdetermined system in Q, f and the parameters a, b, c, d.

The complete classification of low-order multiplier and corresponding densities and fluxes will be presented in the talk.

3 Lie symmetries

A point symmetry of the generalized Rosenau-RLW equation (4) is a one-parameter Lie group of transformations on (t, x, u) generated by a vector field of the form

$$\mathbf{X} = \tau(t, x, u)\partial_t + \xi(t, x, u)\partial_x + \eta(t, x, u)\partial_u, \tag{7}$$

whose prolongation leaves invariant equation (4).

The condition for a vector field (7) to generate a point symmetry of equation (4) is given by

$$\mathrm{pr}^{(5)}\mathbf{X}(\Delta) = 0 \quad \text{when} \quad \Delta = 0, \tag{8}$$

where $\mathrm{pr}^{(5)}\mathbf{X}$ is the fifth prolongation of the vector field (7). The equation (8) splits with respect to the x and t derivatives of u giving an overdetermined linear system of equations for the infinitesimals $\xi(t, x, u)$, $\tau(t, x, u)$, $\eta(t, x, u)$, f and the parameters a, b, c, d. Solving this system we obtain:

Theorem 1 *Group Classification Theorem:*
The Lie point symmetry group of the generalized Rosenau-RLW equation (4) with $f(u)$ an arbitrary function, is determined by the generators

$$\mathbf{X}_1 = \partial_t, \quad \mathbf{X}_2 = \partial_x. \tag{9}$$

 ISBN: 978-84-617-8694-7

If $f(u) = \alpha(\beta - u)^n - a$, *the symmetry group is determined by generators* $\mathbf{X}_1$ *and* $\mathbf{X}_2$ *and*

$$\mathbf{X}_3 = t\partial_t + \frac{1}{n}(\beta - u)\partial_u. \tag{10}$$

Acknowledgements

The authors gratefully acknowledge the support and generosity of the Plan Propio de Investigación de la Universidad de Cádiz and they express their sincerest thanks to the Escuela Doctoral EDUCA.

References

[1] S. C. Anco, Generalization of Noether's theorem in modern form to non-variational partial differential equations, Recent progress and Modern Challenges in Applied Mathematics, Modeling and Computational Science, Fields Institute Communications, *Springer-Verlag, New York, 2017.*

[2] S. C. Anco and G. Bluman, Direct construction of conservation laws from field equations, Phys. Rev. Lett. **78** *(1997) 2869–2873.*

[3] S. C. Anco and G. Bluman, Direct constrution method for conservation laws of partial differential equations Part I: Examples of conservation law classifications, Euro. Jnl of Appl. Math. **13** *(2002) 545–566.*

[4] S. C. Anco and G. Bluman, Direct constrution method for conservation laws of partial differential equations Part II: General treatment, Euro. Jnl of Appl. Math. **13** *(2002) 567–585.*

[5] G. Bluman, A. Cheviakov and S. C. Anco, Applications of symmetry methods to partial differential equations, *Springer, New York, 2009.*

[6] P. J. Olver, Applications of Lie groups to differential equations, *Springer-Verlag, New York, 1986.*

[7] J. I. Ramos and C. M. García-López, Solitary Wave Formation from a Generalized Rosenau Equation, Math. Probl. Eng. **2016** *(2016), 17 pages, Article ID 4618364.*

[8] P. Rosenau, A quasi-continuous description of a nonlinear transmission line, Physica Scripta. **34** *(1986) no. 6B, 827–829.*

[9] P. Rosenau, Dynamics of dense discrete systems, Prog. Theor. Phys. **79** *(1988) no. 5, 1028–1042.*

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

On the first general Zagreb index

José M. Rodríguez[1], José L. Sánchez[2] and José M. Sigarreta[3]

[1] *Departamento de Matemáticas, Universidad Carlos III de Madrid*

[2] *Facultad de Matemáticas, Universidad Autónoma de Guerrero and Universidad de Holguín*

[3] *Facultad de Matemáticas, Universidad Autónoma de Guerrero*

emails: jomaro@math.uc3m.es, jlsanchezsantiesteban@gmail.com, josemariasigarretaalmira@hotmail.com

Abstract

The aim of this paper is to obtain new inequalities involving the first general Zagreb index, and characterize graphs which are extremal with respect to them. We also obtain inequalities involving the forgotten and second general Zagreb indices.

Key words: Graph invariant, Topological index, forgotten index, variable Zagreb index, general Zagreb index
MSC 2000: 05C07, 92E10

1 Introduction

A topological index is defined as a single number that represents a chemical structure in graph-theoretical terms, and that correlates with a molecular property; it is used to understand physicochemical properties of chemical compounds. Topological indices are interesting since they capture some of the properties of a molecule in a single number.

Although only about 1000 benzenoid hydrocarbons are known, the number of possible benzenoid hydrocarbons is huge. For instance, the number of possible benzenoid hydrocarbons with 35 benzene rings is 5851000265625801806530 [32]. Therefore, the modeling of their physico-chemical properties is very important in order to predict properties of currently unknown species.

 ISBN: 978-84-617-8694-7

Hundreds of topological indices have been introduced and studied, starting with the seminal work by Wiener [33] in which he used the sum of all shortest-path distances of a (molecular) graph for modeling physical properties of alkanes.

Topological indices based on end-vertex degrees of edges have been used over 40 years. Among them, several indices are recognized to be useful tools in chemical researches. Probably, the best known of such descriptors is the Randić connectivity index (R) [22]. There are more than thousand papers and a couple of books dealing with this index (see, e.g., [11], [14], [27], [28] and the references therein). During many years, scientists were trying to improve the predictive power of the Randić index. This led to the introduction of a large number of new topological descriptors resembling the original Randić index.

The reason for introducing a new index is to gain prediction of some property of molecules somewhat better than obtained by already presented indices. Therefore, a test study of predictive power of a new index must be done.

Two of the main successors of the Randić index are the first and second Zagreb indices, denoted by M_1 and M_2, respectively, and introduced by Gutman et al. in 1972 (see [12]). They are defined as

$$M_1(G) = \sum_{u \in V(G)} d_u^2, \qquad M_2(G) = \sum_{uv \in E(G)} d_u d_v,$$

where uv denotes the edge of the graph G connecting the vertices u and v, and d_u is the degree of the vertex u.

In the same paper, where Zagreb indices were introduced, the *forgotten topological index* (or *F-index*) is defined as

$$F(G) = \sum_{u \in V(G)} d_u^3.$$

Both the forgotten topological index and the first Zagreb index were employed in the formulas for total π-electron energy in [12], as a measure of branching extent of the carbon-atom skeleton of the underlying molecule. However, this index never got attention except recently, when Furtula et al. in [9] established some basic properties of the F-index and showed that its predictive ability is almost similar to that of first Zagreb index and for the entropy and acetic factor, both of them yield correlation coefficients greater than 0.95. Besides, [9] pointed out the importance of the F-index: it can be used to obtain a high accuracy of the prediction of logarithm of the octanol-water partition coefficient (see also [1]). Recently, this index has been studied for different graph operations [6]. The coindex version of the forgotten index is also introduced in [7]. The extremal trees with respect to the F-index have been investigated in [1]. Furthermore, [5] and [26] contain more lower and upper bounds for the forgotten index.

Miličević and Nikolić defined in [18] the *first and second variable Zagreb indices* as

$$^{\alpha}M_1(G) = \sum_{u \in V(G)} d_u^{2\alpha}, \qquad ^{\alpha}M_2(G) = \sum_{uv \in E(G)} (d_u d_v)^{\alpha},$$

 ISBN: 978-84-617-8694-7

with $\alpha \in \mathbb{R}$. In [16] and [4] the *first and second general Zagreb indices* are introduced as

$$M_1^\alpha(G) = \sum_{u \in V(G)} d_u^\alpha, \qquad M_2^\alpha(G) = \sum_{uv \in E(G)} (d_u d_v)^\alpha,$$

respectively. It is clear that these indices are equivalent to the previous ones, since ${}^\alpha M_1(G) = M_1^{2\alpha}(G)$ and ${}^\alpha M_2(G) = M_2^\alpha(G)$. We prefer to use $M_1^\alpha(G)$ instead of ${}^\alpha M_1(G)$ since the inequalities obtained in this paper become simpler with it.

Note that M_1^1 is $2m$, M_1^2 is the first Zagreb index M_1, M_1^{-1} is the inverse index $ID(G)$ [8], M_1^3 is the forgotten index $F(G)$, etc.; also, $M_2^{-1/2}$ is the usual Randić index, M_2^1 is the second Zagreb index M_2, M_2^{-1} is the modified Zagreb index [20], etc. Note that it is interesting to study M_1^α for $\alpha \neq 0, 1$, and M_2^α for $\alpha \neq 0$, since if G has n vertices and m vertices, then $M_1^0(G) = n$, $M_1^1(G) = 2m$ and $M_2^0(G) = m$.

The concept of the variable molecular descriptors was proposed as a new way of characterizing heteroatoms in molecules (see [23], [24]), but also to assess the structural differences (e.g., the relative role of carbon atoms of acyclic and cyclic parts in alkylcycloalkanes [25]). The idea behind the variable molecular descriptors is that the variables are determined during the regression so that the standard error of estimate for a studied property is as small as possible. The second variable Zagreb index is used in the structure-boiling point modeling of benzenoid hydrocarbons [21]. Various properties and relations of these indices are discussed in several papers (see, e.g., [3], [15], [17], [30], [34], [35]).

The aim of this paper is to obtain new inequalities involving the first general Zagreb index, and characterize graphs which are extremal with respect to them. We also obtain inequalities involving the forgotten and second general Zagreb indices.

Throughout this paper, $G = (V(G), E(G))$ denotes a (non-oriented) finite simple (without multiple edges and loops) nontrivial ($E(G) \neq \emptyset$) graph.

2 Some bounds for M_1^α

We start with some bounds for M_1^α involving different parameters.

Theorem 2.1. *Let G be a nontrivial graph with m edges, maximum degree Δ and minimum degree δ, and $\alpha \in \mathbb{R}$. Then*

$$2\Delta^{\alpha-1}m \leq M_1^\alpha(G) \leq 2\delta^{\alpha-1}m, \qquad \text{if } \alpha < 1,$$
$$2\delta^{\alpha-1}m \leq M_1^\alpha(G) \leq 2\Delta^{\alpha-1}m, \qquad \text{if } \alpha \geq 1,$$

and the equality holds in each inequality for some $\alpha \neq 1$ if and only if G is regular.

Theorem 2.2. *Let G be a nontrivial graph with n vertices and m edges, and $\alpha \in \mathbb{R}$. Then*

$$M_1^\alpha(G) \geq 2m\alpha + n(1-\alpha), \qquad \text{if } \alpha \leq 0 \text{ or } \alpha \geq 1,$$
$$M_1^\alpha(G) \leq 2m\alpha + n(1-\alpha), \qquad \text{if } 0 < \alpha < 1.$$

 ISBN: 978-84-617-8694-7

The equality holds in the inequality for some $\alpha \neq 0, 1$ if and only if G is a union of pairwise disjoint edges.

Next, we have inequalities relating two indices M_1^α and M_1^β.

Theorem 2.3. *Let G be a nontrivial graph with n vertices, maximum degree Δ and minimum degree δ, and $\alpha, \beta \in \mathbb{R}$. Then*

$$\begin{aligned} M_1^\alpha(G) &\leq \delta^{\alpha-\beta} M_1^\beta(G), && \text{if } \alpha \leq \beta, \\ M_1^\alpha(G) &\leq \Delta^{\alpha-\beta} M_1^\beta(G), && \text{if } \alpha \geq \beta, \\ M_1^\alpha(G) &\geq \frac{\Delta^{\alpha+\beta} n^2}{M_1^\beta(G)}, && \text{if } \alpha \leq -\beta, \\ M_1^\alpha(G) &\geq \frac{\delta^{\alpha+\beta} n^2}{M_1^\beta(G)}, && \text{if } \alpha \geq -\beta. \end{aligned}$$

The equality is attained in the lower bound with $(\alpha, \beta) \neq (0, 0)$ if and only if G is regular; if $\alpha = \beta = 0$, then the lower bound is attained for every graph. The equality holds in the upper bound for some $\alpha \neq \beta$ if and only if G is regular; if $\alpha = \beta$, then the upper bound is attained for every graph.

Proposition 2.4. *Let G be a nontrivial graph with n vertices, $s > 0$ and $\alpha \in \mathbb{R}$. Then*

$$2sn \leq s^2 M_1^\alpha(G) + M_1^{-\alpha}(G),$$

Theorem 2.5. *Let G be a nontrivial graph with n vertices, $\alpha \in \mathbb{R}$ and $\beta > 0$. Then*

$$n^{\beta+1} \leq M_1^{-\alpha\beta}(G) M_1^\alpha(G)^\beta,$$

and the equality is attained for some values $\alpha \neq 0$ and β if and only if G is regular.

The following result appears in [31].

Lemma 2.6. *If $\alpha \geq 1$ is an integer and $0 \leq x_1, \ldots, x_n \leq n-1$, then*

$$\Big(\sum_{j=1}^n x_j^\alpha \Big)^{1/\alpha} \leq (n-1)^{1-1/\alpha} \sum_{j=1}^n x_j^{1/\alpha}.$$

We have a generalization of this lemma which is interesting by itself.

Lemma 2.7. *Consider real numbers $0 < \beta \leq 1 \leq \alpha$, $\Delta > 0$ and $0 \leq x_1, \ldots, x_n \leq \Delta$. Then*

$$\Big(\sum_{j=1}^n x_j^\alpha \Big)^{1/\alpha} \leq \Delta^{1-\beta} \sum_{j=1}^n x_j^\beta.$$

 ISBN: 978-84-617-8694-7

This lemma allows to prove the following result.

Proposition 2.8. *Let G be a nontrivial graph with maximum degree Δ, and consider real numbers $0 < \beta \leq 1 \leq \alpha$. Then*

$$M_1^{\alpha}(G)^{1/\alpha} \leq \Delta^{1-\beta} M_1^{\beta}(G).$$

Theorem 2.9. *Let G be a nontrivial graph with n vertices, maximum degree Δ and minimum degree δ, and $\alpha \in \mathbb{R}$. Then*

$$\frac{2(\Delta\delta)^{\alpha/2}}{\Delta^{\alpha}+\delta^{\alpha}} \sqrt{nM_1^{2\alpha}(G)} \leq M_1^{\alpha}(G) \leq \sqrt{nM_1^{2\alpha}(G)}\,.$$

The lower bound is attained for every value of α if G is regular. The upper bound is attained for some $\alpha \neq 0$ if and only if G is regular.

Proposition 2.10. *Let G be a nontrivial graph with n edges, maximum degree Δ and minimum degree δ, and $\alpha \in \mathbb{R}$. Then*

$$M_1^{\alpha}(G) + (\Delta\delta)^{\alpha} M_1^{-\alpha}(G) \leq n(\Delta^{\alpha}+\delta^{\alpha}),$$

and the equality holds for some $\alpha \neq 0$ if and only if $d_u \in \{\Delta, \delta\}$ for every $u \in V(G)$.

Theorem 2.11. *Let G be a nontrivial graph with n vertices, and $\alpha, \beta \in \mathbb{R}$ with $\alpha > 0$. Then*

$$\begin{aligned} n + \alpha M_1^{\beta}(G) &\leq \left(M_1^{\alpha\beta}(G)^{1/\alpha} + n^{1/\alpha}\right)^{\alpha}, && \text{if } \alpha \geq 1, \\ n + \alpha M_1^{\beta}(G) &\geq \left(M_1^{\alpha\beta}(G)^{1/\alpha} + n^{1/\alpha}\right)^{\alpha}, && \text{if } 0 < \alpha < 1. \end{aligned}$$

Theorem 2.12. *Let G be a nontrivial graph with n vertices, and $\alpha, \beta \in \mathbb{R}$ with $\beta > 0$. Then*

$$\begin{aligned} M_1^{\alpha+\beta}(G) &\geq \frac{1}{n} M_1^{\alpha}(G) M_1^{\beta}(G), && \text{if } \alpha \geq 0, \\ M_1^{\alpha+\beta}(G) &\leq \frac{1}{n} M_1^{\alpha}(G) M_1^{\beta}(G), && \text{if } \alpha \leq 0, \end{aligned}$$

and the equality holds in the inequality for some $\alpha \neq 0$ if and only if G is regular.

Since $M_1^1(G) = 2m$, Theorem 2.12 has the following consequence.

Corollary 2.13. *Let G be a nontrivial graph with n vertices and m edges, and $\alpha \in \mathbb{R}$. Then*

$$\begin{aligned} M_1^{\alpha+1}(G) &\geq \frac{2m}{n} M_1^{\alpha}(G), && \text{if } \alpha \geq 0, \\ M_1^{\alpha+1}(G) &\leq \frac{2m}{n} M_1^{\alpha}(G), && \text{if } \alpha \leq 0, \end{aligned}$$

and the equality holds in the inequality for some $\alpha \neq 0$ if and only if G is regular.

 ISBN: 978-84-617-8694-7

3 Inequalities for M_1^α involving other topological indices

The previous results for the first general Zagreb index hold, in particular, for the forgotten index. Next, we obtain particular bounds for the forgotten index.

Theorem 3.1. *Let G be a nontrivial graph with maximum degree Δ and minimum degree δ. Then*

$$\frac{4M_2^2(G)}{\Delta^2} - 2M_2(G) \leq F(G) \leq \frac{4M_2^2(G)}{\delta^2} - 2M_2(G),$$

and each inequality is attained if and only if G is regular.

We also have some inequalities relating the first and second general Zagreb indices.

Theorem 3.2. *Let G be a nontrivial graph with maximum degree Δ and minimum degree δ, and $\alpha \in \mathbb{R}$. Then*

$$\begin{aligned} 2\Delta^{1-\alpha}M_2^{\alpha-1}(G) \leq M_1^\alpha(G) \leq 2\delta^{1-\alpha}M_2^{\alpha-1}(G), \qquad & \text{if } \alpha \geq 1, \\ 2\delta^{1-\alpha}M_2^{\alpha-1}(G) \leq M_1^\alpha(G) \leq 2\Delta^{1-\alpha}M_2^{\alpha-1}(G), \qquad & \text{if } \alpha \leq 1, \end{aligned}$$

and the equality holds in each inequality for some $\alpha \neq 1$ if and only if G is regular.

The *modified Narumi-Katayama index*

$$NK^*(G) = \prod_{u \in V(G)} d_u^{d_u} = \prod_{uv \in E(G)} d_u d_v$$

is introduced in [10], inspired in the Narumi-Katayama index defined in [19]. Finally, we present an inequality relating the modified Narumi-Katayama and the first general Zagreb indices.

Theorem 3.3. *Let G be a nontrivial graph with m edges, and $\alpha \in \mathbb{R}$. Then*

$$M_1^\alpha(G) \geq 2m\, NK^*(G)^{(\alpha-1)/(2m)},$$

and the equality holds for some $\alpha \neq 1$ if and only if G is regular.

Acknowledgements

This work has been partially supported by two grants from Ministerio de Economía y Competititvidad, Agencia Estatal de Investigacin (AEI) and Fondo Europeo de Desarrollo Regional (FEDER) (MTM 2016-78227-C2-1-P and MTM 2015-69323-REDT), Spain, and a grant from CONACYT (FOMIX-CONACyT-UAGro 249818), México.

 ISBN: 978-84-617-8694-7

References

[1] H. Abdo, D. Dimitrov, I. Gutman, On extremal trees with respect to the F-index, *Kuwait J. Sci.*, in press.

[2] C. Alsina, R. B. Nelsen, *When Less is More: Visualizing basic Inequalities*. Mathematical Association of America, Washington D.C., 2009.

[3] V. Andova, M. Petrusevski, Variable Zagreb Indices and Karamatas Inequality, *MATCH Commun. Math. Comput. Chem.* **65** (2011) 685–690.

[4] G. Britto Antony Xavier, E. Suresh and I. Gutman, Counting relations for general Zagreb indices, *Kragujevac J. Math.* **38** (2014) 95–103.

[5] Z. Che, Z. Chen, Lower and Upper Bounds of the Forgotten Topological Index, *MATCH Commun. Math. Comput. Chem.* **76** (2016) 635–648.

[6] N. De, S. M. A. Nayeem, A. Pal, F-index of some graph operations, *Discr. Math. Algor. Appl.* (2016), doi :10.1142/S1793830916500257.

[7] N. De, S. M. A. Nayeem, A. Pal, The F-coindex of some graph operations, SpringerPlus 2016, 5:221, doi: 10.1186/s40064-016-1864-7.

[8] S. Fajtlowicz, On conjectures of Graffiti–II, *Congr. Numer.* **60** (1987) 187–197.

[9] B. Furtula, I. Gutman, A forgotten topological index, *J. Math. Chem.* **53 (4)** (2015) 1184–1190.

[10] M. Ghorbani, M. Songhori, I. Gutman, Modified Narumi–Katayama index, *Kragujevac J. Sci.* **34** (2012) 57–64.

[11] I. Gutman, B. Furtula (Eds.), *Recent Results in the Theory of Randić Index*, Univ. Kragujevac, Kragujevac, 2008.

[12] I. Gutman, N. Trinajstić, Graph theory and molecular orbitals. Total π–electron energy of alternant hydrocarbons, *Chem. Phys. Lett.* **17** (1972) 535–538.

[13] G. H. Hardy, J. E. Littlewood, G. Polya, *Inequalities*, Cambridge Univ. Press, Cambridge, 1952.

[14] X. Li, I. Gutman, *Mathematical Aspects of Randić Type Molecular Structure Descriptors*, Univ. Kragujevac, Kragujevac, 2006.

[15] X. Li and H. Zhao, Trees with the first smallest and largest generalized topological indices, *MATCH Commun. Math. Comput. Chem.* **50** (2004) 57–62.

 ISBN: 978-84-617-8694-7

[16] X. Li and J. Zheng, A unified approach to the extremal trees for different indices, *MATCH Commun. Math. Comput. Chem.* **54** (2005) 195–208.

[17] M. Liu and B. Liu, Some properties of the first general Zagreb index, *Australas. J. Combin.* **47** (2010) 285–294.

[18] A. Miličević, S. Nikolić, On variable Zagreb indices, *Croat. Chem. Acta* **77** (2004) 97–101.

[19] H. Narumi, M. Katayama, Simple topological index. A newly devised index characterizing the topological nature of structural isomers of saturated hydrocarbons, *Mem. Fac. Engin. Hokkaido Univ.* **16** (1984) 209–214.

[20] S. Nikolić, G. Kovačević, A. Miličević, N. Trinajstić, The Zagreb Indices 30 years after, *Croat. Chem. Acta* **76** (2003) 113–124.

[21] S. Nikolić, A. Miličević, N. Trinajstić, A. Jurić, On Use of the Variable Zagreb ${}^{\nu}M_2$ Index in QSPR: Boiling Points of Benzenoid Hydrocarbons *Molecules* **9** (2004) 1208–1221.

[22] M. Randić, On characterization of molecular branching, *J. Am. Chem. Soc.* **97** (1975) 6609–6615.

[23] M. Randić, Novel graph theoretical approach to heteroatoms in QSAR, *Chemometrics Intel. Lab. Syst.* **10** (1991) 213–227.

[24] M. Randić, On computation of optimal parameters for multivariate analysis of structure-property relationship, *J. Chem. Inf. Comput. Sci.* **31** (1991) 970–980.

[25] M. Randić, D. Plavšić, N. Lerš, Variable connectivity index for cycle-containing structures, *J. Chem. Inf. Comput. Sci.* **41** (2001) 657–662.

[26] J. M. Rodríguez, J. M. Sigarreta, New Results on the Harmonic Index and Its Generalizations, *MATCH Commun. Math. Comput. Chem.*, in press.

[27] J. A. Rodríguez–Velázquez, J. M. Sigarreta, On the Randić index and condicional parameters of a graph, *MATCH Commun. Math. Comput. Chem.* **54** (2005) 403–416.

[28] J. A. Rodríguez–Velázquez, J. Tomás–Andreu, On the Randić index of polymeric networks modelled by generalized Sierpinski graphs, *MATCH Commun. Math. Comput. Chem.* **74** (2015) 145–160.

[29] J. M. Sigarreta, Bounds for the geometric–arithmetic index of a graph, *Miskolc Math. Notes* **16** (2015) 1199–1212.

 ISBN: 978-84-617-8694-7

[30] M. Singh, K. Ch. Das, S. Gupta, A. K. Madan, Refined variable Zagreb indices: highly discriminating topological descriptors for QSAR/QSPR, *Int. J. Chem. Modeling* **6(2-3)** 403–428.

[31] L. A. Székely, L. H. Clark, R. C. Entringer, An inequality for degree sequences, *Discr. Math.* **103** (1992) 293–300.

[32] M. Vöge, A. J. Guttmann, I. Jensen, On the number of benzenoid hydrocarbons, *J. Chem. Inf. Comput. Sci.* **42** (2002) 456–466.

[33] H. Wiener, Structural determination of paraffin boiling points, *J. Am. Chem. Soc.* **69** (1947) 17–20.

[34] S. Zhang, W. Wang and T. C. E. Cheng, Bicyclic graphs with the first three smallest and largest values of the first general Zagreb index, *MATCH Commun. Math. Comput. Chem.* **55** (2006) 579–592.

[35] H. Zhang and S. Zhang, Unicyclic graphs with the first three smallest and largest values of the first general Zagreb index, *MATCH Commun. Math. Comput. Chem.* **55** (2006) 427–438.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

A new LES model using non linear viscosity

José M. Rodríguez[1] and Raquel Taboada-Vázquez[1]

[1] *Department of Mathematics, Universidade da Coruña*

emails: jose.rodriguez.seijo@udc.es, raquel.taboada@udc.es

Abstract

We present in this paper a new Large Eddy Simulation (LES) model obtained filtering the Navier-Stokes equations, where the viscosity has been substituted by a non linear function of the strain rate tensor. This model is a generalization of the model introduced in [8]. Here we show some numerical results, comparing them with those of [2] and [4].

Key words: Large Eddy Simulation, nonlinear viscosity
MSC 2000: 35Q35, 35Q30, 76F65, 76D05

1 Introduction

Large Eddy Simulation (LES) models are widely used to simulate turbulent flows (see [1] for example). Usually, these models are presented as "averaged" or "filtered" versions of Navier–Stokes equations because one of the possible ways to derive them is to apply a filter operator to the Navier-Stokes equations, obtaining a new equation governing the behavior of the filtered velocity.

In [8] a new LES model was deduced by applying a filter, not to the Navier-Stokes equations, but to these generalized Navier–Stokes equations with a nonlinear effective viscosity

$$\rho_0 \left(\frac{\partial \mathbf{u}}{\partial t} + (\nabla \mathbf{u})\, \mathbf{u} \right) = \rho_0 \mathbf{f} + \nabla \cdot \mathbf{T}, \tag{1}$$

$$\nabla \cdot \mathbf{u} = 0, \tag{2}$$

where $\mathbf{u}$ is the velocity field, $\mathbf{f}$ is the acceleration due to external forces and the stress tensor $\mathbf{T}$ is given by

$$\mathbf{T} = -p\mathbf{I} + 2\mu_e(|\mathbf{D}|)\mathbf{D}, \qquad \mathbf{D} = \frac{1}{2}\left(\nabla \mathbf{u} + \nabla \mathbf{u}^T\right), \tag{3}$$

 ISBN: 978-84-617-8694-7

where p is the pressure and the *effective viscosity* μ_e, depending on the norm of the strain rate tensor $\mathbf{D}$, was chosen as

$$\mu_e(|\mathbf{D}|) = \mu_0 \left(1 + \lambda^2 |\mathbf{D}|^2\right)^{1/2}, \quad \lambda > 0 \tag{4}$$

The parameter λ was considered constant in [8].

In this previous work, we applied to the following Gaussian filter equations (1)-(2):

$$\bar{f}(t, \mathbf{x}) = \int_{-\infty}^{\infty} \int_{\mathbb{R}^3} G(s - t, \mathbf{y} - \mathbf{x}) f(s, \mathbf{y}) \, \mathrm{d}\mathbf{y} \, \mathrm{d}s \tag{5}$$

with

$$G(s, \mathbf{y}) = \frac{\gamma_T^{1/2} \gamma_L^{3/2}}{16\pi^2 \eta^4} \exp\left(-\frac{\gamma_T s^2 + \gamma_L |\mathbf{y}|^2}{4\eta^2}\right) \tag{6}$$

where $\eta > 0$ is a small parameter related to the size of the filter, and $\gamma_T > 0$, $\gamma_L > 0$ are parameters related to the shape of the filter.

The model thus obtained is as follows:

$$\begin{aligned} &\frac{\partial \bar{\mathbf{u}}}{\partial t} + (\nabla \bar{\mathbf{u}}) \, \bar{\mathbf{u}} + \frac{1}{\rho_0} \nabla \bar{p} = \bar{\mathbf{f}} + \nabla \cdot (\mathbf{S} - \tau), \\ &\nabla \cdot \bar{\mathbf{u}} = 0, \end{aligned} \tag{7}$$

where τ (the so called *subgrid-scale stress tensor*) is given by

$$\tau_{ij} = \overline{u_i u_j} - \bar{u}_i \bar{u}_j \tag{8}$$

$$= 2 \left[\frac{1}{\gamma_T} \frac{\partial \bar{u}_i}{\partial t} \frac{\partial \bar{u}_j}{\partial t} + \frac{1}{\gamma_L} \nabla \bar{u}_i \cdot \nabla \bar{u}_j\right] \eta^2 + O(\eta^4), \tag{9}$$

and $\mathbf{S}$ is given by

$$\begin{aligned} S_{ij} &= \frac{2}{\rho_0} \left(\overline{\mu_e(|\mathbf{D}|)\mathbf{D}}\right)_{ij} \\ &= 2\nu \Big[\left(1 + \lambda^2 K_1\right)^{1/2} + \lambda^2 \left(1 + \lambda^2 K_1\right)^{-1/2} K_2 \eta^2 \\ &\quad - \frac{\lambda^4}{4} \left(1 + \lambda^2 K_1\right)^{-3/2} K_3 \eta^2 \Big] \bar{D}_{ij} \\ &\quad + 2\nu\lambda^2 \left(1 + \lambda^2 K_1\right)^{-1/2} \hat{K}_{ij} \eta^2 + O(\eta^4) \end{aligned} \tag{10}$$

 ISBN: 978-84-617-8694-7

The notation introduced is:

$$\nu = \mu_0/\rho_0, \quad K_1 = |\bar{\mathbf{D}}|^2, \tag{11}$$

$$K_2 = \sum_{m,n=1}^{3} \left[\frac{1}{\gamma_T} \left(\frac{\partial \bar{D}_{mn}}{\partial t} \right) \left(\frac{\partial \bar{D}_{mn}}{\partial t} \right) + \frac{1}{\gamma_L} \nabla \bar{D}_{mn} \cdot \nabla \bar{D}_{mn} \right], \tag{12}$$

$$K_3 = \frac{1}{\gamma_T} \left(\frac{\partial K_1}{\partial t} \right)^2 + \frac{1}{\gamma_L} \nabla K_1 \cdot \nabla K_1, \tag{13}$$

$$\hat{K}_{ij} = \frac{1}{\gamma_T} \frac{\partial K_1}{\partial t} \frac{\partial \bar{D}_{ij}}{\partial t} + \frac{1}{\gamma_L} \nabla K_1 \cdot \nabla \bar{D}_{ij}. \tag{14}$$

So, a nonlinear viscosity was introduced in (4) as part of the model itself rather than using it as a procedure to close the subgrid-scale stress tensor in (8), where we have used the Clark approximation instead (see [3] and [9]).

This new model thus obtained reminds us of the dynamic procedure of Germano (see [6]).

2 New model when lambda is not constant

We are interested in investigating what happens when λ in (4) is allowed to depend on the time and spatial variables ($\lambda = \lambda(t, \mathbf{x})$). If we now apply the filter (5)-(6) to equations (1)-(2), but with non-constant λ, we obtain again (7)-(9) but now $\mathbf{S}$ is given by

$$\begin{aligned} S_{ij} = 2\nu \Big[& (1+K_1)^{1/2} + (1+K_1)^{-1/2} K_2 \eta^2 - \frac{1}{4} (1+K_1)^{-3/2} K_3 \eta^2 \\ & + 2\bar{\lambda} (1+K_1)^{-1/2} K_4 \eta^2 \Big] \bar{D}_{ij} \\ & + 2\nu (1+K_1)^{-1/2} \hat{K}_{ij} \eta^2 + O(\eta^4) \end{aligned} \tag{15}$$

where the notation used is:

$$\nu = \mu_0/\rho_0, \quad K_1 = \bar{\lambda}^2 |\bar{\mathbf{D}}|^2, \tag{16}$$

$$K_2 = \sum_{m,n=1}^{3} \left[\frac{1}{\gamma_T} \left(\frac{\partial (\bar{\lambda} \bar{D}_{mn})}{\partial t} \right) \left(\frac{\partial (\bar{\lambda} \bar{D}_{mn})}{\partial t} \right) + \frac{1}{\gamma_L} \nabla (\bar{\lambda} \bar{D}_{mn}) \cdot \nabla (\bar{\lambda} \bar{D}_{mn}) \right], \tag{17}$$

$$K_3 = \frac{1}{\gamma_T} \left(\frac{\partial K_1}{\partial t} \right)^2 + \frac{1}{\gamma_L} \nabla K_1 \cdot \nabla K_1, \tag{18}$$

$$K_4 = \sum_{m,n=1}^{3} \left[\frac{1}{\gamma_T} \frac{\partial \bar{\lambda}}{\partial t} \frac{\partial \bar{D}_{mn}}{\partial t} + \frac{1}{\gamma_L} \nabla \bar{\lambda} \cdot \nabla \bar{D}_{mn} \right] \bar{D}_{mn}, \tag{19}$$

$$\hat{K}_{ij} = \frac{1}{\gamma_T} \frac{\partial K_1}{\partial t} \frac{\partial \bar{D}_{ij}}{\partial t} + \frac{1}{\gamma_L} \nabla K_1 \cdot \nabla \bar{D}_{ij}. \tag{20}$$

 ISBN: 978-84-617-8694-7

3 Numerical results

With the aim of comparing the model proposed in [8] with the model we have just proposed, we have solved first equations (1)-(4) with constant and variable λ, and then we have compared the results obtained with Armaly et al. experimental measures (see [2]) and the numerical results presented by Chacón and Lewandowski with different turbulence models (see [4]) for the backward facing step test case. This test is widely used for turbulence validation, we want to examine the ability of our models to accurately compute the steady turbulent 2D backward step flow. The test consists in prescribing an inflow at a certain distance from the step, then the flow suddenly encounters the expansion of a wall, causing a flow separation, a large vortex is formed behind the step front. Armaly et. al. used to conduct their study a channel that had a height of 1.01 cm downstream the step and whose inlet was 0.52 cm (h) in height. The definition of the Reynolds number which they used is given by

$$Re = \frac{4hV_{max}}{3\nu} \tag{21}$$

where V_{max} is the maximum inlet velocity (the incoming flow is parabolic) and ν is the kinematic viscosity. Predictions of this flow, in a geometry equivalent to the case used for the experiments, were obtained by numerically solving the models proposed employing FreeFem++ (see [5]). The step length considered has been $l = 4h$ and the length of the computational domain 20 times the step height, in the same way as in [4].

When a variable λ is chosen, it is constant (λ_c) in the inner part of the domain and it goes linearly to zero at the boundaries (with limit layer thickness 10^{-3}).

Several numerical simulations have been performed for different values of λ (considered both constant and a function of $\mathbf{x}$) and different values of the Reynolds number. All the results have been obtained with a mesh width $h/20$ and a time step 0.01. The test stop for the time-stepping procedure has been set to

$$\frac{\|u_1^{n+1} - u_1^n\|_\infty}{\|u_1^{n+1}\|_\infty} < 10^{-5} \tag{22}$$

Re	X_r/h measured [2]	X_r/h computed [4]	X_r/h constant λ	X_r/h $\lambda = 0$
100	3	$2.76 - 2.86$	3.02	2.83
500	10	$9.18 - 10.61$	10.09	9.23
5000	6.7	$7.29 - 7.9$	6.51	14.85
6000	7	$7.58 - 8.29$	7.02	14.89
10000	8	$8.02 - 8.85$	7.96	–

Table 1: Reattachment points (X_r/h) for different Reynolds numbers

 ISBN: 978-84-617-8694-7

In table 1 we present the computed (or measured) length of the main vortex formed. This length has been calculated by X_r/h where X_r is the reattachment length.

We can conclude that our models are able to compute the 2D backward step flow, reobtaining the reattachment lengths measured by Armaly et al. as accurately as Chacón and Lewandowski's models do for different Reynolds numbers (corresponding to laminar and turbulent regimes). When we consider non-constant λ we achieve similar results for the main vortex and the predictions of the secondary vortex are generally improved.

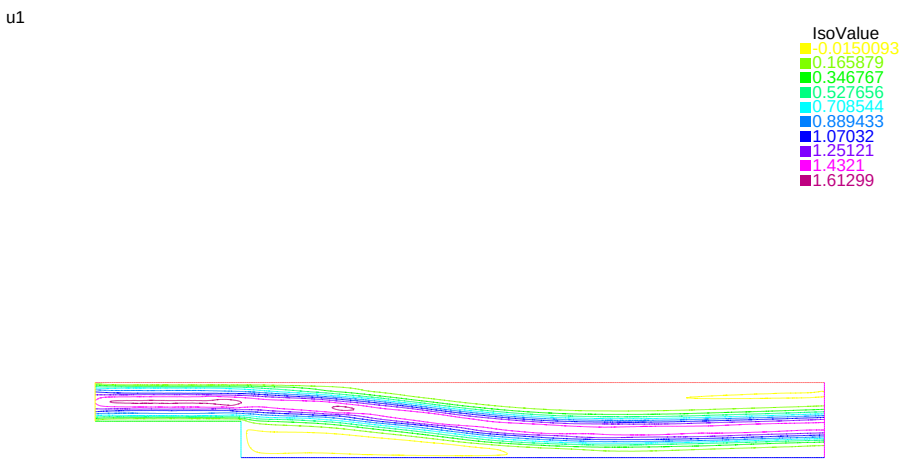


Figure 1: u_1 component of the velocity for $Re = 10000$, $\lambda = 0.1$

Figures 1 and 2 show the first component of the velocity map provided by the model considering in the first place constant and then non-constant λ, with $Re = 10000$. For this Reynolds number, with $\lambda = 0$ the scheme does not converge for the mesh width chosen.

The numerical simulations of the 'filtered' models are still ongoing, but we expect very similar results to the ones presented in [7] and [8].

 ISBN: 978-84-617-8694-7

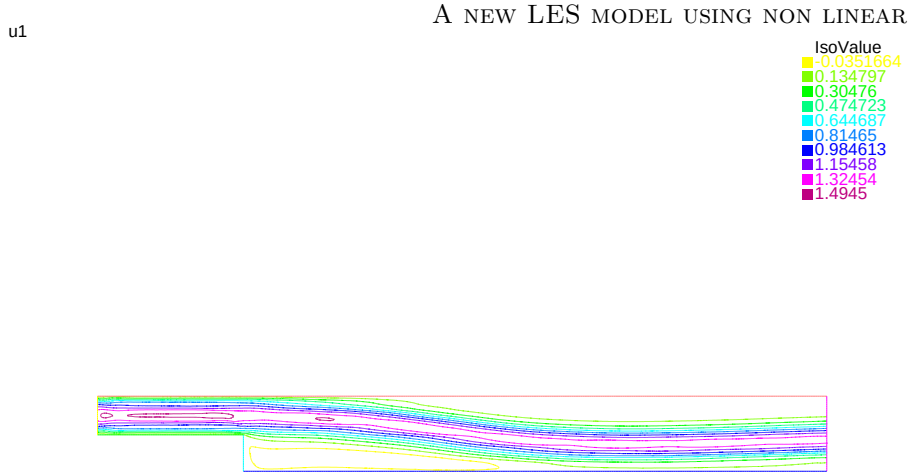


Figure 2: u_1 component of the velocity for $Re = 10000$, non-constant λ, $\lambda_c = 0.2$

Acknowledgements

This work has been partially supported by Ministerio de Economía y Competitividad (Spain) under grant MTM2016-78718-P with the participation of FEDER.

The numerical simulations have been performed using CESGA (Centro de Super Compuntación de Galicia)'s servers.

References

[1] C. D. Argyropoulos, N. C. Markatos, *Recent advances on the numerical modelling of turbulent flows*, Applied Mathematical Modelling **39** (2015), 693–732.

[2] B. F. Armaly, F. Durst, J. C. F. Pereira and B. Schöung, *Experimental and theoretical investigation of backward-facing step flow*, J. Fluid Mech. **127** (1983), 473–496.

[3] D. Carati, F. S. Winckelmans, H. Jeanmart, *On the modelling of the subgrid-scale and filtered-scale stress tensors in large-eddy simulation*, J. Fluid Mech. **441** (2001), 119–138.

[4] T. Chacón and R. Lewandowski, *Mathematical and Numerical Foundations of Turbulence Models and Applications*, Birkhäuser, New York, 2014.

 ISBN: 978-84-617-8694-7

[5] F. Hecht, *New development in FreeFem++*, J. Numer. Math. **20** no. 3-4 (2012), 251–265.; https://doi.org/10.1515/jnum-2012-0013

[6] M. Germano, U. Piomelli, P. Moin, W. H. Cabot, *A dynamic subgrid-scale eddy viscosity model*, Phys. Fluids A **3** (1991), 1760–1765.

[7] J. M. Rodríguez, R. Taboada-Vázquez, *Time-averaged shallow water model: asymptotic derivation and numerical validation*, Journal of Mathematical Analysis and Applications **428** no. 2 (2015) 930–950. doi.org/10.1016/j.jmaa.2015.03.050

[8] J. M. Rodríguez, R. Taboada-Vázquez, *A new LES model derived from generalized Navier-Stokes equations with nonlinear viscosity*, Computers and Mathematics with Applications **73** (2017) 294–303. doi.org/10.1016/j.camwa.2016.11.024

[9] B. Vreman, B. Geurts, H. Kuerten, *Large-Eddy Simulation of the Temporal Mixing Layer Using the Clark Model*, Theoret. Comput. Fluid Dynamics **8** (1996) 309–324.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

Analysis of time series from H264/AVC compressed domain for video summarization

L. Rodriguez-Benitez[1], J. Giralt[1], L. Jimenez[1] and J. Moreno-Garcia[2]

[1] *Information Systems and Technologies Department, University of Castilla-La Mancha. Escuela Superior de Informatica de Ciudad Real. (Spain)*

[2] *Information Systems and Technologies Department, University of Castilla-La Mancha. Escuela de Ingenieria Industrial de Toledo. (Spain)*

emails: luis.rodriguez@uclm.es, juan.giralt@uclm.es, luis.jimenez@uclm.es, juan.moreno@uclm.es

Abstract

In this paper a video summarization technique based on the analysis of a video time series is presented. Time series is a relevant research field with a great number of different applications. Furthermore, the representation of such series and the extraction of information from them can be made using different techniques like for example fuzzy logic [1]. In this proposal, the information composing such time series is obtained from the H264/AVC [2] compressed domain. H264/AVC is a video coding standard to compress and decompress a video stream that balances a high visual quality with an efficient file size. It combines a transformation of pixel values into a frequency-domain representation [3], typically used to encode still images, with motion compensation techniques specific of video compression techniques. This motion compensation information is stored in the motion vectors. More concretely, the field of motion vectors in a frame can be considered as the sparse and imprecise approximation of the optical flow field. So, we consider the encoding process using motion compensation and estimation a motion segmentation process allowing to obtain information about the motion patterns of the different elements or objects in the scene [4].

In order to generate summaries identifying global patterns of motion in video we propose the use of statistical measures obtained directly from the values of the motion vectors in every frame of the video. So, we establish a correspondence between every frame in the video and a class where such class is characterized by the statistical correlation between several motion vectors organized by groups, where the organization of groups is derived from the spatial position in the picture or frame of the motion vectors. Then, as the frame is divided into several areas or groups of motion vectors more

ISBN: 978-84-617-8694-7

than one correlation measure is needed and a combination of such measures allows the definitions of some frames as the class representative models. Working in compressed domain and using basic statistical measures drives to the generation of video summaries in an efficient way [5, 6, 7].

In our proposal, the first step is to obtain from an unique video a set of class representative models

$$Classes = \{(f, C_i),\ f \in CRM\} \tag{1}$$

where CRM is the set of frames considered as class representative models in the video and C_i represents the class identified in the frame f.

After that another video is taken as input and a summary is generated and it is a represented as a set of frames associated with a class previously identified and represented in *Classes*. A frame is associated with a class if the correlation with such class is greater than or equal to 0.9.

$$Summary = \{(f, C_i),\ f \in F\} \tag{2}$$

where C_i represents the class associated to the frame f and F is the total set of frames of the video. An example is shown in Table 1.

Table 1: Summary obtained from a video.

$Initial frame$	$Final frame$	C_i
1	37	1
38	38	2
39	103	1
104	105	2
106	107	3
109	110	2
111	131	3
132	133	2
134	174	3
175	182	2
183	283	3
284	291	2
292	330	1
331	463	4

Finally, the experimentation of this paper is driven to the comparison of the motion patterns of several videos captured from different vehicles driving the same distance in the same road. Summaries can be compared in order to identify different ranges of velocities or trajectories for each one of the vehicles with respect to the original video. From the experimentation results it can be concluded that classes are mainly associated to traffic events, like the presence of another cars in the scene, and to concrete variations in the road geometry where the different classes are associated to bends, left or right,

 ISBN: 978-84-617-8694-7

or to situations where the vehicle is moving in a straight line. For instance, Figure 1 shows different classes corresponding to these situations described above.

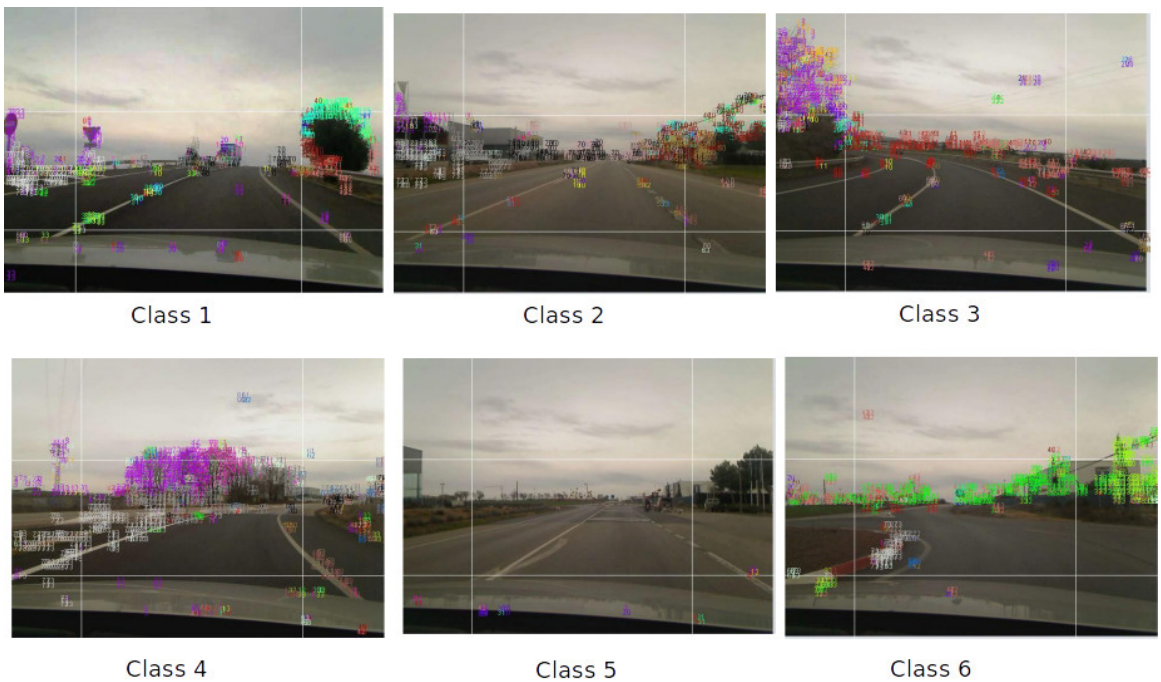


Figure 1: Frames corresponding with class representative models

Acknowledgements

Supported by the project TIN2015-64776-C3-3-R of the Science and Innovation Ministry of Spain, co-funded by the European Regional Development Fund (ERDF).

References

[1] A. Moreno-Garcia, J. Moreno-Garcia, L. Jimenez and L. Rodriguez-Benitez, *Time series representation using fuzzy logic*, Proceedings of the International Conference on Computational and Mathematical Methods in Science and Engineering. Rota (Spain) **3** (2016) 898–908.

[2] T. Wiegand, G.J. Sullivan, A.G. Bjontegaard and A.A. Luthra, *Overview of the H.264/AVC video coding standard*, IEEE Trans. Cir. and Sys. for Video Technol. **13** (2003) 1051–8215.

[3] R. Kordasiewicz and A. Shahram Shirani , *On Hardware Implementations Of DCT and Quantization Blocks for H.264/AVC*, J. VLSI Signal Process. Syst. **47** (2007) 93–102.

[4] J. Giralt, J. Moreno-Garcia, L.Jimnez-Linares, E. Castillo, L. Rodriguez-Benitez *A Fuzzy Representation of Vehicle Trajectories using Motion Data from H264/AVC Video.*, Proceedings of the International Conference on Computational and Mathematical Methods in Science and Engineering. Rota (Spain) **2** (2014) 635–646.

[5] J. Almeida, N.J. Leite and R.S. Torres , *Online video summarization on compressed domain*, J. Vis Commun Image R. **24** (2013) 729–738.

[6] J. Almeida, N.J. Leite and R.S. Torres, *VISON: VIdeo Summarization for ONline applications*, Pattern Recogn. Lett. **33** (2007) 397–409.

[7] S. Mei, G. Guan, Z. Wang, S. Wan, M. He and D. Feng, *Video summarization via minimum sparse reconstruction*, Pattern Recogn. **48** (2007) 522–533.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

A Distributed and Flexible Platform for Large-Scale Data Storage in HPC Systems

Cristina Rodríguez-Quintana[1], Antonio F. Díaz[1], Julio Ortega[1], Raúl H. Palacios [2], Juan José Escobar[1] and Fabricio Marcillo[3]

[1] *Department of Architecture and Computer Technology, University of Granada, Spain*

[2] *Autonomous University of the State of Hidalgo, Pachuca, México*

[3] *Faculty of Engineering Sciences, Quevedo State Technical University, Ecuador*

emails: crodriguez@ugr.es, afdiaz@ugr.es, jortega@ugr.es, raulhp@ugr.es, jjescobar@ugr.es, fmarcillo@correo.ugr.es

Abstract

The realm of HPC systems lies in sharing computational resources efficiently. Their challenge is to turn massively large data into valuable information and meaningful knowledge. To accomplish this, I/O subsystems have to provide scalable bandwidth and capacity in order to deliver on the increasing demand for their requests. Emerging technologies, new programming paradigms and virtualized environments need novel ways to offer optimised solutions to support heavy data flows in storage services.

In this paper, we propose a distributed storage layer on computer nodes that can be used as a robust data storage service to handle intensive I/O operations. Preliminary experiments show that our platform outperforms other distributed data storage solutions.

Key words: Data storage, high performance computing, supercomputers, communication networks, file systems

1 Introduction

HPC centres are service providers, where multiple specialized computational resources provide different services. Some of these resources are designed to run simulations and generate data. Others provide post-processing or analysis of the generated data, while some are responsible for visualization of the raw or reduced data. The common denominator in

 ISBN: 978-84-617-8694-7

all of these scenarios is data. Data must be shared among these systems in an efficient manner.

I/O is mainly used in scientific applications to store output from simulations for later analysis, for implementing algorithms that process more data that can fit in system memory and must page data to and from disk and for checkpointing to save the state of application in case of system failure.

Modern distributed storage systems employ techniques that can help improve application performance, alleviate I/O bandwidth bottleneck, mask failures, and ensure data availability.

It is needed to develop both the theoretical and practical aspects of building efficient and scalable distributed storage that will scale and be tolerant of the inevitable node failures and service partitioning that come with multiple nodes. Raicu et al. [10] proposes a distributed storage architecture that will make exascale computing more tractable, touching virtually all disciplines in high-end computing and fueling scientific discovery. Chen et al.[1] proposes a Parallel, Reliable and Scalable Storage Software Infrastructure (PASSI) to support the design and prototyping of next-generation active storage environment.

Some tools have been created to modeling large-scale HPC I/O workload. Snyder et al.[12] have defined a novel workload abstraction layer that may be used by diverse tools to regenerate and analyze I/O workloads. They implemented three workload generators based on distinct representations of I/O workloads: I/O traces, synthetic I/O kernels, and I/O characterizations. They used a simulation model of an HPC storage system to analyse and compare each IOWA workload generation technique in detail, and they used an I/O replay engine to evaluate large-scale workloads on a production HPC storage system.

Ghoshal et al.[7] present their results in benchmarking the I/O performance over different cloud and HPC platforms to identify the major bottlenecks in existing infrastructure.

Multiple failure are inevitable in large storage systems, and designers must consider to avoid lost of data. The core technology for protecting data from failures is erasure coding, which has a rich 50+ year history stemming communication systems, and as such, can be confusing to the storage systems community. Plank [9] presents a primer on erasure coding as it applies to storage systems, and summarizes some interesting research on erasure coding.

In our paper, we provide a short background of a high performance data storage system in Section 2. We show a design overview in Section 3. We present some preliminary results in Section 4 and we conclude the paper in Section 5.

2 Background

A high performance data storage system is composed of several basic elements: different communication layers, storage units, data manipulation algorithms and control software.

 ISBN: 978-84-617-8694-7

All these elements are continually evolving to push the limits of existing technology. There are several aspects to managing a distributed high-performance storage system:

- Try to take full advantage of all resources: CPUs, storage and networks with a fast asynchronous client/server communication layer.
- Define a reliable system to manage resources: There are several software solutions: etcd[5], Consul[8], Zookeeper.
- Support different topologies: create storage volumes composed of several units and group them according to their performance: i.e: SSD for fast storage and HDD for slow operations.
- Redundancy management: To define how many units are needed to use for storage.
- Dynamic storage management: The less used files go to slow units.
- Use erasure codes to optimise integrity.

3 Design Overview

A distributed high-performance storage system requires to control resources in a efficient way and demands low latencies and high data bandwidth. Thus this complex system have to be decomposed into small services that solve different requirements. On a previous publication, we have shown a metadata system applied to HPC [11] and how to improve data sharing on Infiniband Networks [4].

The proposed model, AbFS3_L3 (AbFS3 Low Level Layer), is not a generic filesystem but rather a flexible platform to store any kind of data, defining a low level layer for a unified data storage (ie. object storage, files or blocks).

The main subsystem are:

- Resource management and service discovery system: It detects all available resources and changes in configuration. It supervises that all components are properly working. It is based on etcd [5].
- Fast data transfer system: It is a fluent data transport which adapts to different conditions (real and virtualized networks).
- Data balancing and redistribution: It is responsible for deciding how to distribute data among servers, replace data into new locations and recover accidental.
- Data storage API: It is the service that clients can access and provides cache and block control to accelerate interoperability.

 ISBN: 978-84-617-8694-7

Clients has a reduced configuration in order to achieve a coherent image and they "learn" new configuration profiles as they demand resources and, therefore, they need not be notified of general changes.

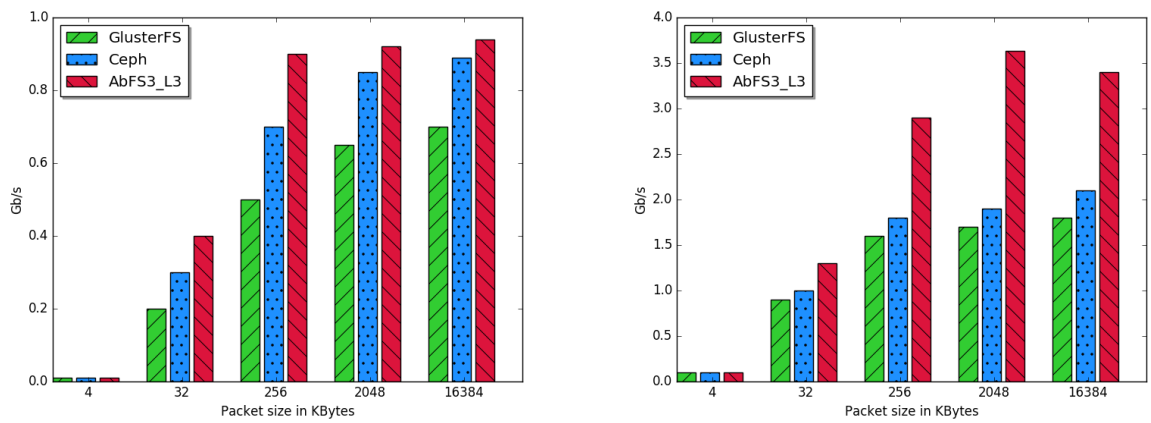
Clients can request operations to any server who will resolve the queries, and the server can "suggest" any other server to improve response times.

Administrator can select multiple configuration options to define redundancy, restrictions, migration policies or strict working conditions to support no single point of failure (NSPOF).

4 Results

The system has been tested in a cluster with 16 server nodes equipped with: Dual 2.27GHz Intel(R) Xeon(R) E5520 CPUs, 16GB RAM memory, 1TB Local Disk and OS CentOS 6.6. Our scenario has 4 servers and 10 clients. These nodes have two Broadcom Corporation NetXtreme II BCM5716 Gigabit Ethernet interfaces and a IBA7220 InfiniBand HCA.

We have compare our system with other implementations based on distributed filesytems, such as GlusterFS[3] and Ceph[2] with two different networks: Infiniband and Gigabit Ethernet.



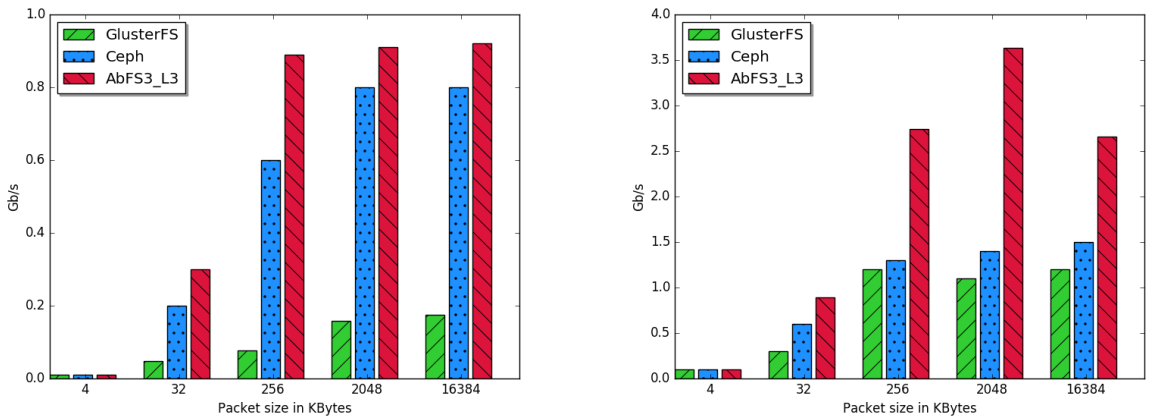
(a) with Gigabit Network

(b) with Infiniband Network

Figure 1: Read bandwidth obtained

We show the throughput with a variety of block sizes ranging from 4 KB to 16 MB. Figure 1(a) shows read bandwidth obtained from a Gigagit Network and Figure 1(b) from a Infiniband and, Figure 2(a) shows write bandwidth using Gigabit Network and Figure 2(b) using Infiniband.

 ISBN: 978-84-617-8694-7



(a) with Gigabit Network

(b) with Infiniband Network

Figure 2: Write bandwidth obtained

Both figures show that our proposed system outperform Ceph and GlusterFS in all conditions: reads, writes, with diverse block sizes and different networks.

5 Conclusions and future work

In this paper we have presented a flexible platform for data storage in HPC systems. This system can be used as a robust low level interface to store and retrieve data efficiently. Preliminary experiments show that our platform outperforms other distributed data storage solutions. As future work, we want to combine this platform with an optimised metadata support and a virtual filesytem layer to perform fast I/O based on file operations.

Acknowledgements

This work has been partially supported by European Union FEDER and the Spanish Ministry of Economy and Competitiveness TIN2015-67020-P and FPA2015-65150-C3-3-P.

References

[1] Hsing-bung Chen and Song Fu. PASSI: A parallel, reliable and scalable storage software infrastructure for active storage system and I/O environments. In *34th IEEE*

International Performance Computing and Communications Conference, IPCCC 2015, Nanjing, China, December 14-16, 2015, pages 1–8, 2015.

[2] Red Hat Inc. (Ceph Community). Ceph. `http://ceph.com`, 2017.

[3] Red Hat Inc. (Gluster Community). Glusterfs. `https://www.gluster.org`, 2017.

[4] Antonio F. Díaz, Julio Ortega, Andrés Ortiz, Godofredo Garay, and Alberto Prieto. Improving data sharing on Infiniband Networks. In *Proceedings of the International Conference on Computational and Mathematical Methods in Science and Engineering,* CMMSE, pages 1326–1334, Jul 2014.

[5] etcd Community. etcd. `https://github.com/coreos/etcd`, 2017.

[6] The Apache Software Fundation. Hadoop distributed file system (hdfs). `https://wiki.apache.org/hadoop/HDFS`, 2017.

[7] Devarshi Ghoshal, Richard Shane Canon, and Lavanya Ramakrishnan. I/o performance of virtualized cloud environments. In *Proceedings of the Second International Workshop on Data Intensive Computing in the Clouds*, DataCloud-SC '11, pages 71–80, New York, NY, USA, 2011. ACM.

[8] HashiCorp. Consul. `http://consul.io`, 2017.

[9] J. S. Plank. Erasure codes for storage systems: A brief primer. volume 38. Usenix Association, December 2013.

[10] Ioan Raicu, Ian T. Foster, and Pete Beckman. Making a case for distributed file systems at exascale. In *Proceedings of the Third International Workshop on Large-scale System and Application Performance*, LSAP '11, pages 11–18, New York, NY, USA, 2011. ACM.

[11] Cristina Rodríguez-Quintana, Antonio F. Díaz, Julio Ortega, Raúl H. Palacios, and Andrés Ortiz. Evaluating distributed metadata in HPC. In *Proceedings of the 15th International Conference on Computational and Mathematical Methods in Science and Engineering,* CMMSE, pages 965–972, 2015.

[12] Shane Snyder, Philip H. Carns, Robert Latham, Misbah Mubarak, Robert B. Ross, Christopher D. Carothers, Babak Behzad, Huong Vu Thanh Luu, Surendra Byna, and Prabhat. Techniques for modeling large-scale HPC I/O workloads. In *Proceedings of the 6th International Workshop on Performance Modeling, Benchmarking, and Simulation of High Performance Computing Systems, PMBS 2015, Austin, Texas, USA, November 15, 2015*, pages 5:1–5:11, 2015.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

Improving Energy Efficiency in Virtual Data Centers: A real-world case study

Julio Rodríguez-Soares[1], Alberto Cocaña-Fernández[1], Raquel Cortina[1], Luciano Sánchez[1] and José Ranilla[1]

[1] *Department of Computer Science, University of Oviedo*

emails: UO237347@uniovi.es, cocanaalberto@gmail.com, raquel@uniovi.es, luciano@uniovi.es, ranilla@uniovi.es

Abstract

Cloud computing era has posed important challenges in the energy and operational costs of data centers and, therefore, efforts to efficiently manage these computing infrastructures are essential. In this work, the power consumption of medium-scale Virtualized Data Centers operated by a multi-service IT provider is analysed. After modelling and profiling the workload, some experimental studies have been conducted to improve power savings through Virtual Machines allocation and migration. Our preliminary results are competitive compared to those obtained using commercial systems.

Key words: Data Center, Resource Management, Virtual Machine Migration, Energy efficiency

1 Introduction

Information and Communications Technology (ICT) has played a key role in the transformation of modern society by consistently delivering innovative products and services, increasing productivity and supporting economic growth. Nevertheless, ICT environmental impact is far from negligible as data centers alone are responsible for over 1% of the world's electricity consumption, and 14% of its carbon footprint [1]. Moreover, data center-related power consumptions continue to grow, and are estimated to increase by 53% between 2013 and 2020 according to [2]. Given so, substantial efforts have been done by the leading ICT service providers to maximise the energy efficiency of large-scale data centers in the pursuit of reducing power consumptions and operating costs. However, large-scale infrastructures

 ISBN: 978-84-617-8694-7

only represent 5% of the data centers' energy use, with the remaining 95% used by the far less efficient small and medium-scale ones [2]. As a result of this, efforts focused in improving the efficiency of multi-purpose small and medium-scale infrastructures with common workload patterns may yield greater overall power savings.

Virtualized Data Centers (VDC) are the cornerstone of most Infrastructure as a Service (IaaS) providers. In a VDC, applications are wrapped within Virtual Machines (VMs) representing their execution environment. VMs are then mapped to Physical Machines (PMs) by optimising a set of predefined objectives such as service quality, operating costs, reliability, network load, etc. Hence, energy optimisation techniques in these environments must tackle one or more layers of the VDC stack: Application, Virtual Machine and Physical Machine [3, 4, 5].

With regard to VMs, the problem of finding the best mapping of VMs onto the PMs is well-known and already has been extensively studied in the literature. Essentially, schedulers implementing the mapping algorithms are composed of two modules: allocation and migration. The allocation module performs the initial VM placement, while the migration module consolidates the VM load onto the minimum possible number of PMs. Recently, some authors stated that whenever the cluster workload is well-defined and exhibits a relatively low variability, migration can be also be addressed as a VM reallocation within decision intervals of several minutes.

In this work, the power consumption of medium-scale Virtualized Data Centers operated by a multi-service IT provider is analysed in order to build a representative environment to assess the results attained with a set of VM migration algorithms. In a full paper a Genetic Fuzzy Rule-Based System implementing the migration mechanism will be presented to improve the energy savings achieved with state-of-the-art commercial virtualization software.

2 Reference Data Center setup

ASAC Comunicaciones[1], an IT company established in 1996 offering a portfolio private-cloud services, was chosen as real-world case study to assess the performance of the proposed algorithms. Such provider was selected given that its mid-size TIER III-certified[2] data centers make an excellent representation of the vast amount of medium-scale service providers operating under reasonable quality standards, but with great room to improve power savings.

Figure 1 depicts briefly ASAC data centers architecture. Basically, computing resources are grouped in VDCs, which are then organised in clusters, each one comprised of multiple PMs hosting the VMs. The number of VDCs, clusters and PMs/VMs per cluster depends

[1] http://www.asac.as

[2] These certifications are granted by the Uptime institute according to the availability and redundancy of its physical infrastructure. Further information can be found on its website https://uptimeinstitute.com

 ISBN: 978-84-617-8694-7

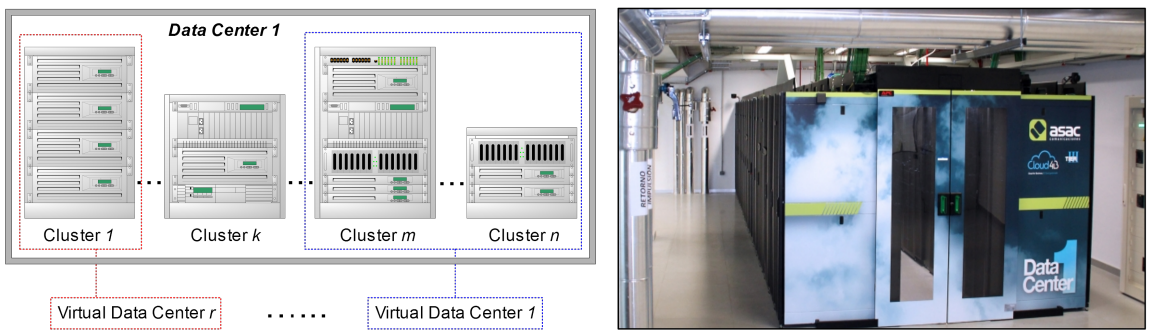


Figure 1: ASAC Data Center 1 architecture.

heavily on the applications deployed and their resource requirements (CPU, memory, network throughput, etc.), with an overall VMs/cluster ratio ranging from 9 to 1500 deployed over between 2 and 12 PMs. ASAC VDCs are built using VMware vSphere as the virtualization platform[3]. In particular, the experimental testbed consists of a single cluster with 12 PMs, each one with 32 physical cores, hosting a total of 200 VMs with 2-4 vCPUs. As for the workload, it was determined according to the trace logs of the last six months, providing quantitative and accurate data on resource consumptions regarding CPU and memory usage, network traffic, Input/Output Operations Per Second (IOPS), etc. After careful analysis of the recorded workload, two patterns were found:

1. **Stationary Workload**, featuring slight fluctuations in resource consumption but which does not change substantially over time. In our study three levels were defined: 1) *LowCharge* load following a uniform distribution with VMs using less than 40% of resources requested at startup, 2) *MedCharge* load, same as the former though rising VMs' occupancy threshold from 40% to 65%, and 3) *HighCharge* load with the occupancy threshold between VMs 65% and 100%.

2. **Dynamic Workload**, where VMs' occupancy evolves overtime with a certain daily seasonality.

3 Experimental results

Given the productive nature of ASAC data centers, experimentations were done using the well-known CloudSim framework [6, 7]. This framework was customized in order to suit the evolving nature of the workload described in the preceding section by implementing temporal series of *cloudlets* (the minimum unit of VM load), as well as including two additional VM

[3]https://www.vmware.com//products/vsphere.html

 ISBN: 978-84-617-8694-7

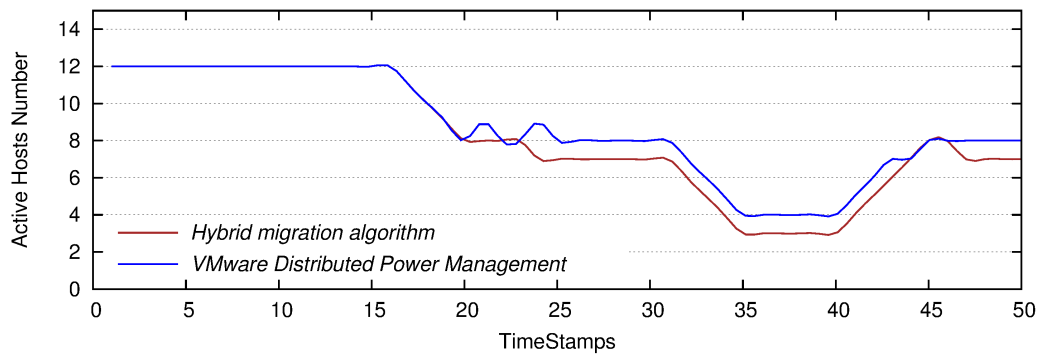
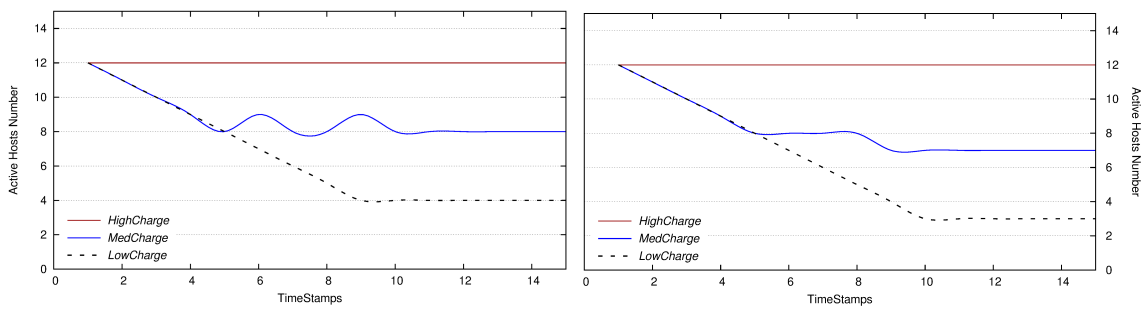


Figure 2: Results obtained with Dynamic Workload and both algorithms.



(a) (b)

Figure 3: Results obtained with (a) VMware Distributed Power Management and (b) Hybrid migration algorithm under three different scenarios of Stationary Workload.

allocation algorithms used as benchmark. The first one, called *VMware Distributed Power Management* within the figures, is a modified version of the median absolute deviation-based migration algorithm used by VMware's Distributed Power Management module, extending it to comply with predetermined occupancy intervals. In particular, the minimum and maximum occupancy thresholds used in the experiment were set at 40% and 80%, respectively. Migration rules were based on the CPU's utilization rate.

The second algorithm uses a weighted function with multiple criteria to adjust the migrations. This algorithm, called *Hybrid Migration Algorithm* (HMA), uses the same set of predefined thresholds as those used in the previous one. Preliminary results are depicted in Figures 2 and 3, and show that VM migrations with the proposed HMA attain greater energy savings by achieving a higher degree of consolidation without impact on service quality.

 ISBN: 978-84-617-8694-7

4 Future Work

Given that initial results show substantial room for improvements over commercial virtualization solutions in mid-size data centers, in the full paper a set of state-of-the-art solutions will be compared against our reactive decision-making mechanism presented in [8, 9]. This mechanism consists of a utility function implemented by means of a Genetic Fuzzy Ruled-Based System (GFRBS), elicited from workload records to optimise a weighed fitness involving multiple criteria such as service quality, energy savings, hardware reliability, etc. This utility function will be used to assess the value (utility) of every possible VM migration to a different PM, and also including the possibility of rejecting any migration.

In the same way, a proactive migration and allocation mechanism based on the one described in [10] for High Performance Computing clusters will be adapted, since actual workloads suggest that load in these infrastructures are rather predictable.

Finally, given the amount of service-related concerns regarding data centers operation, learning algorithms based on many-objective optimization techniques will be leveraged to cope with the growing dimensionality of Pareto fronts.

Acknowledgements

This work has been partially supported by the Ministry of Economy and Competitiveness ("Ministerio de Economía y Competitividad") from Spain/FEDER under grants TEC2015-67387-C4-3-R and TIN2014-56967-R and by the Regional Ministry of the Principality of Asturias under grant FC-15-GRUPIN14-073.

References

[1] "Smart 2020: Enabling the low carbon economy in the information age," Tech. Rep., The Climate Group and the Global e-Sustainability Initiative, 2008.

[2] P. Delforge and J. Whitney, "Issue Paper: Data Center Efficiency Assessment scaling up energy efficiency across the Data Center Industry: evaluating Key Drivers and Barriers," Tech. Rep., Natural Resources Defense Council (NRDC), 2014.

[3] Meera Vasudevan, Yu-Chu Tian, Maolin Tang, and Erhan Kozan, "Profile-based application assignment for greener and more energy-efficient data centers," *Future Generation Computer Systems*, vol. 67, pp. 94 – 108, 2017.

[4] C. Ghribi, M. Hadji, and D. Zeghlache, "Energy efficient vm scheduling for cloud data centers: Exact allocation and migration algorithms," in *2013 13th IEEE/ACM International Symposium on Cluster, Cloud, and Grid Computing*, May 2013, pp. 671–678.

 ISBN: 978-84-617-8694-7

[5] J. Xu and J. A. B. Fortes, "Multi-objective virtual machine placement in virtualized data center environments," in *Green Computing and Communications (GreenCom), 2010 IEEE/ACM Int'l Conference on Int'l Conference on Cyber, Physical and Social Computing (CPSCom)*, Dec 2010, pp. 179–188.

[6] Wei Zhao, Yong Peng, Feng Xie, and Zhonghua Dai, "Modeling and simulation of cloud computing: A review," in *Cloud Computing Congress (APCloudCC), 2012 IEEE Asia Pacific*. IEEE, 2012, pp. 20–24.

[7] Rodrigo N. Calheiros, Rajiv Ranjan, Anton Beloglazov, César A. F. De Rose, and Rajkumar Buyya, "Cloudsim: A toolkit for modeling and simulation of cloud computing environments and evaluation of resource provisioning algorithms," *Softw. Pract. Exper.*, vol. 41, no. 1, pp. 23–50, Jan. 2011.

[8] Alberto Cocaña-Fernández, Jose Ranilla, and Luciano Sánchez, "Energy-efficient allocation of computing node slots in hpc clusters through parameter learning and hybrid genetic fuzzy system modeling," *The Journal of Supercomputing*, vol. 71, no. 3, pp. 1163–1174, 2015.

[9] Alberto Cocaña Fernández, Luciano Sánchez, and José Ranilla, "Improving the eco-efficiency of high performance computing clusters using eecluster," *Energies*, vol. 9, no. 3, 2016.

[10] Alberto Cocaña Fernández, Luciano Sánchez, and José Ranilla, "Leveraging a predictive model of the workload for intelligent slot allocation schemes in energy-efficient hpc clusters," *Eng. Appl. Artif. Intell.*, vol. 48, no. C, pp. 95–105, Feb. 2016.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

Fixed point theorems by combining Jleli and Samet's, and Branciari's inequalities

Antonio Francisco Roldán López de Hierro[1] **and Naseer Shahzad**[2]

[1] *Department of Quantitative Methods for Economics and Business, University of Granada.*

[1] *PAIDI Research Group FQM-268, University of Jaén, Jaén, Spain.*

[2] *Operator Theory and Applications Research Group, Department of Mathematics, King Abdulaziz University.*

emails: aroldan@ugr.es, afroldan@ujaen.es, nshahzad@kau.edu.sa

Abstract

The aim of this work is to introduce a new class of generalized metric spaces (called *RS-spaces*) that unify and extend, at the same time, Branciari's generalized metric spaces and Jleli and Samet's generalized metric spaces. In order to show its great applicability, we present some fixed point theorems in the setting of RS-spaces that extend well-known results in this line of research.

Key words: Generalized metric space, Branciari metric space, Fixed point, Contractive mapping
MSC 2000: 46T99, 47H10, 47H09, 54H25.

1 Introduction

Fixed point theory is currently one of the most active branches of nonlinear analysis. In the last years, there have been introduced many fixed point results in the setting of natural extensions of metric spaces: quasimetric-spaces [1], Mustafa and Sims' generalized metric spaces [2], Czerwik's b-metric spaces [3], Hitzler and Seda's dislocated metric spaces [4], Nakano's modular spaces [5], Musielak and Orlicz's spaces [6], Bakhtin b-metric spaces [1], etc.

Very recently, two very general families of generalized metric spaces have attracted the attention of researchers. On the one hand, Branciari's generalized metric spaces were introduced in [7] in order to show some fixed point theorems. Although these spaces have metrically non-intuitive properties, these drawbacks have not been a limitation for developing fixed point theory in this environment (see [9, 10, 11, 12, 13]). On the other hand, Jleli and Samet [14] introduced a kind of generalized metric spaces which are not endowed with a proper triangle inequality: it was replaced by a weaker condition involving convergent sequences.

At a first sight, Branciari's spaces and Jleli and Samet's spaces seem to be incompatible: for instance, in the second kind of spaces, the limit of a convergent sequence is unique, and two points can be placed having infinite distance between them.

In this work, we introduce a new class of spaces, that we call *RS-spaces*, that are natural extensions of both Branciari's spaces and Jleli and Samet's spaces. We also show some fixed point results.

2 Preliminaries

Henceforth, $\mathbb{N} = \{0, 1, 2, \ldots\}$ stands for the set of all non-negative integer numbers, and let $\mathbb{N}^* = \mathbb{N} \setminus \{0\}$. From now on, X will denote a nonempty set and $T : X \to X$ will be a self-mapping.

Given a point $x_0 \in X$, the *Picard sequence of T based on x_0* is the sequence $\{x_n\}_{n\geq 0}$ given by $x_{n+1} = Tx_n$ for all $n \in \mathbb{N}$. In particular, $x_n = T^n x_0$ for all $n \in \mathbb{N}$, where T^n denotes the n^{th}-iterates of T (we assume that T^0 denotes the identity mapping on X). A Picard sequence satisfies $x_{n+m} = T^m x_n = T^n x_m$ for all $n, m \in \mathbb{N}$. The *orbit of x_0 by T* is the set $\mathcal{O}_T(x_0) = \{\, T^n x_0 : n \in \mathbb{N} \,\}$.

A *binary relation on X* is a nonempty subset $\mathcal{S}$ of the Cartesian product $X \times X$. For simplicity, we denote $x\mathcal{S}y$ if $(x, y) \in \mathcal{S}$. We say that x and y are *$\mathcal{S}$-comparable* if $x\mathcal{S}y$ or $y\mathcal{S}x$. A binary relation $\mathcal{S}$ on X is *reflexive* if $x\mathcal{S}x$ for all $x \in X$; it is *transitive* if $x\mathcal{S}z$ for all $x, y, z \in X$ such that $x\mathcal{S}y$ and $y\mathcal{S}z$; and it is *antisymmetric* if $x\mathcal{S}y$ and $y\mathcal{S}x$ imply $x = y$. Given a non-empty subset A of X, we will say that $\mathcal{S}$ is *transitive on A* if

$$x, y, z \in A, \quad x\mathcal{S}y, \quad y\mathcal{S}z \quad \Rightarrow \quad x\mathcal{S}z.$$

A sequence $\{x_n\} \subseteq X$ is *$\mathcal{S}$-nondecreasing* if $x_n\mathcal{S}x_{n+1}$ for all $n \in \mathbb{N}$.

A *preorder* (or a *quasiorder*) is a reflexive, transitive binary relation and a *partial order* is an antisymmetric preorder.

An *extended comparison function* (or, simply, a *comparison function*) is a function $\phi : [0, \infty] \to [0, \infty]$ such that

($\mathcal{P}_1$) ϕ is nondecreasing;

 ISBN: 978-84-617-8694-7

$(\mathcal{P}_2)$ for all $t \in (0,\infty)$, $\lim_{n\to\infty} \phi^n(t) = 0$.

Let $\mathcal{F}_{\text{com}}$ be the family of all (extended) comparison functions.

2.1 Branciari N-generalized metric spaces

The following notion was introduced by Branciari in [7].

Definition 1 *(Branciari [7]) Given $N \in \mathbb{N}^*$, a* Branciari N-generalized metric space *(for short, a* B_N-space*) is a pair (X,d), where X is a non-empty set and $d : X \times X \to [0,\infty)$ is a function such that the following properties hold:*

(B_1) $d(x,y) = 0$ *if, and only if,* $x = y$.

(B_2) $d(y,x) = d(x,y)$.

(B_3) $d(x,y) \leq d(x,u_1) + d(u_1,u_2) + d(u_2,u_3) + \ldots + d(u_{N-1},u_N) + d(u_N,y)$ *for any $x, u_1, u_2, \ldots, u_N, y \in X$ such that $x, u_1, u_2, \ldots, u_N, y$ are all different.*

If $N = 2$, then (X,d) is a Branciari generalized metric space *(for short, a* B-space*).*

A B_1-space is a metric space. However, if $N \geq 2$, it was proved that B_N-spaces can satisfy some properties that are not metrically desirable (see [8, 9]). For instance, in a B_N-space,

- there may exist convergent sequences that are not Cauchy sequences;
- there may exist convergent sequences with two different limits;
- the metric $d : X \times X \to [0,\infty)$ may not be a continuous function;
- there may exist open balls centered in different points that are never disjoint although their radius are arbitrarily small.

Surprisingly, in [11], Suzuki *et al.* proved that, for $N \geq 2$, only B_3-spaces have a compatible symmetric topology.

2.2 Jleli and Samet's generalized metric spaces

Henceforth, let $\mathcal{D} : X \times X \to [0,\infty]$ be a given mapping. For every $x \in X$, define the set

$$C(\mathcal{D}, X, x) = \left\{ \{x_n\} \subseteq X : \lim_{n\to\infty} \mathcal{D}(x_n, x) = 0 \right\}. \tag{1}$$

Generalized metric and generalized metric space are defined as follows.

 ISBN: 978-84-617-8694-7

Definition 2 *(Jleli and Samet [14], Definition 2.1) Let X be a nonempty set and let $\mathcal{D} : X \times X \to [0, \infty]$ be a function which satisfies:*

$(\mathcal{D}_1)$ $\mathcal{D}(x, y) = 0$ *implies* $x = y$;

$(\mathcal{D}_2)$ $\mathcal{D}(x, y) = \mathcal{D}(y, x)$ *for all* $x, y \in X$;

$(\mathcal{D}_3)$ *there exists* $C > 0$ *such that*

$$\text{if } x, y \in X \text{ and } \{x_n\} \in C(\mathcal{D}, X, x), \quad \text{then} \quad \mathcal{D}(x, y) \leq C \limsup_{n\to\infty} \mathcal{D}(x_n, y). \tag{2}$$

Then $\mathcal{D}$ is called a generalized metric *and the pair $(X, \mathcal{D})$ is called a* generalized metric space *(in the sense of Jleli and Samet; for short, a* JS-space*).*

Jleli and Samet presented in [14] a large list of abstract metric spaces that can be seen as particular cases of JS-spaces: metric spaces, b-metric spaces, Hitzler-Seda metric spaces and modular spaces with the Fatou property. Given a JS-space $(X, \mathcal{D})$ and a point $x \in X$, a sequence $\{x_n\} \subseteq X$ is said to be:

- $\mathcal{D}$*-convergent to* x if $\{x_n\} \in C(\mathcal{D}, X, x)$ (in such a case, we will write $\{x_n\} \overset{\mathcal{D}}{\to} x$);
- $\mathcal{D}$*-Cauchy* if $\lim_{n,m\to\infty} \mathcal{D}(x_n, x_m) = 0$.

A JS-space $(X, \mathcal{D})$ is *complete* if every $\mathcal{D}$-Cauchy sequence in X is $\mathcal{D}$-convergent.

Definition 3 *(Branciari [7]) A* rectangular metric on X *is a mapping $d : X \times X \to [0, \infty)$ satisfying the following properties:*

(RM_1) $d(x, y) = 0$ *if, and only if,* $x = y$;

(RM_2) $d(x, y) = d(y, x)$ *for all* $x, y \in X$;

(RM_3) $d(x, y) \leq d(x, u) + d(u, v) + d(v, y)$ *for all* $x, y \in X$ *and all distinct points* $u, v \in X \setminus \{x, y\}$.

In such a case, (X, d) is called a rectangular metric space *(in short* RM-space*).*

 ISBN: 978-84-617-8694-7

3 RS-generalized metric spaces

In this section we present the class of generalized metric spaces in which we are interested and we show that well-known abstract metric spaces belong to this new class.

Definition 4 (Roldán López de Hierro and Shahzad [16]) *An* RS-generalized metric space *(for short, an* RS-space*) is a pair $(X, \mathcal{D})$ where X is a non-empty set and $\mathcal{D} : X \times X \to [0, \infty]$ is a function such that the following properties are fulfilled:*

$(\mathcal{D}_1)$ *If $\mathcal{D}(x, y) = 0$ then $x = y$;*

$(\mathcal{D}_2)$ *$\mathcal{D}(x, y) = \mathcal{D}(y, x)$ for all $x, y \in X$;*

$(\mathcal{D}'_3)$ *there exists $C > 0$ such that if $x, y \in X$ are two points and $\{x_n\}$ is a $\mathcal{D}$-Cauchy infinite sequence in X such that $\{x_n\} \overset{\mathcal{D}}{\to} x$ then*

$$\mathcal{D}(x, y) \leq C \limsup_{n\to\infty} \mathcal{D}(x_n, y). \tag{3}$$

If X is endowed with a binary relation $\mathcal{S}$, then an RS-space *is a triple $(X, \mathcal{D}, \mathcal{S})$ satisfying $(\mathcal{D}_1)$, $(\mathcal{D}_2)$ and $(\mathcal{D}'_3)$ assuming that the sequence $\{x_n\}$ in $(\mathcal{D}'_3)$ is $\mathcal{S}$-nondecreasing.*

Let us show that the class of RS-spaces contains some important subclasses.

Lemma 5 *Every JS-space is an RS-space.*

Corollary 6 *Every b-dislocated metric is a JS-space and so it is an RS-space.*

Lemma 7 *Every B_N-space is an RS-space (where $\mathcal{D} = d$ and $C = 1$).*

4 Ćirić type fixed point theorems in the context of RS-generalized metric spaces

This section is dedicated to introduce, in the setting of RS-spaces, the main results of this manuscript inspired by the Ćirić type contractivity condition presented in [15].

Theorem 8 (Roldán López de Hierro and Shahzad [16]) *Let $(X, \mathcal{D}, \mathcal{S})$ be an $\mathcal{S}$-nondecreasing-complete RS-space with respect to a preorder $\mathcal{S}$ and let $T : X \to X$ be an $\mathcal{S}$-nondecreasing self-mapping. Let $x_0 \in X$ be a point such that $x_0 \mathcal{S} T x_0$ and $\delta_{n_0}(\mathcal{D}, T, x_0) < \infty$ for some $n_0 \in \mathbb{N}$. Suppose that there exists $\phi \in \mathcal{F}_{\text{com}}$ such that*

$$\mathcal{D}(Tx, Ty) \leq \phi(\max\{\mathcal{D}(x, y), \mathcal{D}(x, Tx), \mathcal{D}(y, Ty), \mathcal{D}(x, Ty), \mathcal{D}(y, Tx)\}) \tag{4}$$
$$\text{for all } x, y \in \mathcal{O}_T(x_0).$$

Additionally, assume that

 ISBN: 978-84-617-8694-7

(a) T is $\mathcal{S}$-nondecreasing-continuous.

Then the Picard sequence $\{x_n\}_{n\in\mathbb{N}}$ of T based on x_0 $\mathcal{D}$-converges to a fixed point ω of T. Furthermore, $\mathcal{D}(\omega,\omega)=0$ and

$$\mathcal{D}(x_n,\omega)\leq C\,\phi^{n-n_0}\left(\delta_{n_0}(\mathcal{D},T,x_0)\right)\quad \text{for all } n\in\mathbb{N} \text{ such that } n\geq n_0,$$

where $C=C_{X,\mathcal{D}}$ is the (lowest) constant for which $(X,\mathcal{D})$ satisfies property $(\mathcal{D}_3')$.

In addition to this, if condition (4) holds for all $x,y\in X$ such that $x\mathcal{S}y$, and ω' is another fixed point of T such that $\omega\mathcal{S}\omega'$, $\mathcal{D}(\omega,\omega')<\infty$ and $\mathcal{D}(\omega',\omega')<\infty$, then $\omega=\omega'$.

Acknowledgments

This article was funded by the Deanship of Scientific Research (DSR), King Abdulaziz University, Jeddah. N. Shahzad acknowledges with thanks DSR for financial support. A.F. Roldán López de Hierro is grateful to the Department of Quantitative Methods for Economics and Business of the University of Granada. The same author has been partially supported by Junta de Andalucía by project FQM-268 of the Andalusian CICYE.

References

[1] I.A. Bakhtin, *The contraction mapping principle in quasimetric spaces*, Funct. Anal., Unianowsk Gos. Ped. Inst. **30** (1989) 26–37.

[2] Z. Mustafa, B. Sims, *A new approach to generalized metric spaces*, J. Nonlinear Convex Anal. **7** (2006) 289–297

[3] S. Czerwik, S, *Contraction mappings in b-metric spaces*, Acta Math. Inform. Univ. Ostrav. **1** (1993) 5–11.

[4] P. Hitzler, A.K. Seda, *Dislocated topologies*, J. Electr. Eng. **51** (12) (2000) 3–7.

[5] H. Nakano, *Modular semi-ordered spaces*, Tokyo, Japan (1959).

[6] J. Musielak, W. Orlicz, *On modular spaces*, Stud. Math. **18** (1959) 49–65.

[7] A. Branciari, *A fixed point theorem of Banach-Caccioppoli type on a class of generalized metric spaces*, Publ. Math. Debrecen **57** (2000) 31–37.

[8] I.R. Sarma, J.M. Rao, S.S. Rao, *Contractions over generalized metric spaces*, J. Nonlinear Sci. Appl. **2** (3) (2009) 180–182.

 ISBN: 978-84-617-8694-7

[9] Z. Kadelburg, S. Radenović, *Fixed point results in generalized metric spaces without Hausdorff property*, Math. Sci. **125** (2014) 1–8.

[10] W.A. Kirk, N. Shahzad, *Generalized metrics and Caristi's theorem*, Fixed Point Theory Appl. **2013**, Article ID 2013:129, 9 pages (2013)

[11] T. Suzuki, B. Alamri, M. Kikkawa, *Only 3-generalized metric spaces have a compatible symmetric topology*, Open Math. **13** (2015) 510–517.

[12] T. Suzuki, B. Alamri, M. Kikkawa, *Edelsteins fixed point theorem in generalized metric spaces*, J. Nonlinear Convex Anal. **16** (11) (2015) 2301–2309.

[13] E. Karapinar, D. O'Regan, A.F. Roldán López de Hierro, N. Shahzad, *Fixed point theorems in new generalized metric spaces*, J. Fixed Point Theory Appl. **18** (3) (2016) 645–671.

[14] M. Jleli, B. Samet, *A generalized metric space and related fixed point theorems*, Fixed Point Theory Appl. **2015** (2015), Article ID 2015:61, 14 pages.

[15] L.B. Ćirić, *A generalization of Banach's contraction principle*, Proc. Am. Math. Soc. **45** (2) (1974) 267–273.

[16] A.F. Roldán López de Hierro, N. Shahzad, *Fixed point theorems by combining Jleli and Samet's, and Branciari's inequalities*, J. Nonlinear Sci. Appl. **9** (6) (2016) 3822–3849.

[17] W.A. Kirk, N. Shahzad, *Fixed point theory in distance spaces*. Springer, Cham, 2014.

[18] W.A. Kirk, N. Shahzad, *Correction: Generalized metrics and Caristi's theorem*, Fixed Point Theory Appl. **2014** (2014), Article ID 2014:177, 3 pages.

[19] N. Shahzad, E. Karapinar, A.F. Roldan Lopez de Hierro, *On some fixed point theorems under (α, ψ, ϕ)-contractivity conditions in metric spaces endowed with transitive binary relations*, Fixed Point Theory Appl. **2015** (2015), Article ID 2015:124, 24 pages.

[20] W.A. Kirk, N. Shahzad, *Fixed points and Cauchy sequences in semimetric spaces*, J. Fixed Point Theory Appl. **17** (3) (2015) 541–555.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

An approach for ranking fuzzy numbers using finite fuzzy numbers and its application in Economics

Antonio Francisco Roldán López de Hierro[1], Concepción Aguilar Peña[2], Antonio Márquez Montávez[3] and Concepción Roldán[4]

[1] *Department of Quantitative Methods for Economics and Business, University of Granada*

[2] *Department of Mathematics, University of Jaén*

[3] *Zaidín-Vergeles Institute, Granada*

[4] *Department of Statistics and Operations Research, University of Granada*

emails: aroldan@ugr.es, caguilar@ujaen.es, ma250962@correo.ugr.es, iroldan@ugr.es

Abstract

Ordering fuzzy subsets is important in decision-making, data analysis and socioeconomic systems. Since fuzzy numbers do not always yield a totally ordered set as real numbers do, an importan issue in operational fuzzy set theory is how to compare fuzzy numbers. In this paper we describe an approach to ranking fuzzy numbers with compact support using a finite image of them. The method is compared with other techniques by numerical examples. Finally we consider a real application to ranking consumer products.

Key words: Fuzzy number, Finite fuzzy number, Ranking

1 Introduction

In many applications, ranking of fuzzy numbers is an important tool in decision processes. It is well-known that fuzzy numbers do not form a natural linear order, like real numbers. Many approaches have been developed different methods for ranking fuzzy numbers [1, 4, 6, 13, 18, 21]. A commonly used technique is to construct maps to transform fuzzy numbers into real number so called defuzzification, for example considering the central gravity, the median or area measurements. Each defuzzification method provides a correspondence from

 ISBN: 978-84-617-8694-7

the set of all fuzzy numbers into the set of the real number. These real numbers are then compared to obtain a ranking. Some approaches produce different rankings for the same data. Then the results of the methods proposed in the literature are often conflict and there is yet no method that can always give a satisfactory solution to every situation.

In this paper we describe an approach to ranking fuzzy numbers with compact support using a finite image of them. In [16], Roldán *et al.* did a complete study of the image of a FN. In their manuscript, they considered FNs $\mathcal{A}$ whose image (or range) $\mathcal{A}(\mathbb{R})$ is a countable (or finite) subset of $\mathbb{I}$, and they called them *finite or discrete* FNs. This kind of FNs is interesting since, in many cases, the usual computation of FNs is only referred to certain data (a finite approximation of a FN) and, in practice, most of examples of fuzzy structures (probabilistic metric spaces, fuzzy metric spaces in several senses and intuitionistic fuzzy metric spaces) are constructed using these classes of FNs. Among other properties, they succeeded in proving that such family of FNs is closed under the usual operations between FNs, that is, if $\mathcal{A}$ and $\mathcal{B}$ are discrete FNs, them $\mathcal{A}+\mathcal{B}$, $\mathcal{A}-\mathcal{B}$, $\mathcal{A}\cdot\mathcal{B}$ and $\mathcal{A}/\mathcal{B}$ (if this last FN is well defined) also are discrete FNs.

2 Preliminaries

Let $\mathbb{R}$ denote the set of all real numbers and $\overline{\mathbb{R}} = [-\infty, \infty]$ the extended real line (for simplicity, $+\infty$ will be denoted as ∞). Henceforth, $x_0 \in \mathbb{R}$ will be a real number.

Let $f : X \to Y$ be a mapping. We say that f is a *finite* (respectively, *countable*) *mapping* if its image $\operatorname{Im} f$ is a finite (respectively, countable) subset of Y. We point out that we will use the term *countable* admitting the possibility that $\operatorname{Im} f$ is *finite*.

A *fuzzy set on* $\mathbb{R}$ is a map $\mathcal{A} : \mathbb{R} \to [0,1]$. A *fuzzy number on* $\mathbb{R}$ (hereinafter, *FN*) is a fuzzy set $\mathcal{A}$ on $\mathbb{R}$ that verifies the following properties:

(1) Normality: there exists a real number $x_0 \in \mathbb{R}$ such that $\mathcal{A}(x_0) = 1$.

(2) For all $\alpha \in]0,1]$, the set $\mathcal{A}_{[\alpha]} = \{x \in \mathbb{R} : \mathcal{A}(x) \geq \alpha\}$ is a closed subinterval of $\mathbb{R}$.

The set $\mathcal{A}_{[\alpha]}$ is known as the *α-level set* (or *α-cut*) of $\mathcal{A}$. The *kernel* of a FN $\mathcal{A}$ is $\ker \mathcal{A} = \mathcal{A}_{[1]}$ and its *support* is the closure $\operatorname{supp}(\mathcal{A}) = \overline{\{x \in \mathbb{R} : \mathcal{A}(x) > 0\}}$. Let $\mathcal{F}$ be the family of all FNs.

From now on, we will denote by FFN (respectively, by CFN) the family of all FNs whose image is finite (respectively, countable).

Regarding FNs, we refer the reader to Klir and Yuan [11], Dubois and Prade [7], Mizumoto and Tanaka [12], Wu and Ma [17], and Buckley and Jowers [2]. Clearly, $\mathbb{R}$ can be embedded in $\mathcal{F}$: if $r \in \mathbb{R}$ then $\hat{r} \in \mathcal{F}$ satisfies $\hat{r}(x) = 1$ if $x = r$ and $\hat{r}(x) = 0$ if $x \neq r$. A *triangular FN* $\mathcal{A} = (a/b/c)$ is defined by three real numbers $a < b < c$ where the graph of $\mathcal{A}$ is a triangle with base on the interval $[a, c]$ and vertex at $x = b$. *Trapezoidal FNs*

 ISBN: 978-84-617-8694-7

$\mathcal{A} = (a/b/c/d)$, where $a < b < c < d$, are similarly defined (see [2, 10, 11]). Any FN $\mathcal{A}$ can be extended to $\overline{\mathbb{R}}$ defining $\mathcal{A}(\pm\infty) = 0$. Hence we only consider FNs on $\overline{\mathbb{R}}$.

Let $\mathcal{A} \in \mathcal{F}$ be a FN. Notice that $\mathcal{A}$ is completely determined by its level sets (see [7]). For every $\alpha \in]0,1]$ let $\mathcal{A}_{[\alpha]} = [\underline{a}(\alpha), \overline{a}(\alpha)]$ (this interval is open on one side if $\underline{a}(\alpha) = -\infty$ or $\overline{a}(\alpha) = \infty$), so we may consider mappings $\underline{a}, \overline{a} :]0,1] \to \overline{\mathbb{R}}$ which determine the extremes of each level set (for simplicity, we will write $\overline{a}_\alpha = \overline{a}(\alpha)$ and $\underline{a}_\alpha = \underline{a}(\alpha)$ for all $\alpha \in]0,1]$). If $0 < \alpha \leq \beta \leq 1$ and $x_0 \in \ker \mathcal{A}$, then $\mathcal{A}_{[\beta]} \subseteq \mathcal{A}_{[\alpha]}$, so $\underline{a}_\alpha \leq \underline{a}_\beta \leq x_0 \leq \overline{a}_\beta \leq \overline{a}_\alpha$. In particular, $\underline{a}_\alpha \leq \underline{a}_1 \leq x_0 \leq \overline{a}_1 \leq \overline{a}_\alpha$, $\underline{a}(]0,1]) \subseteq [-\infty, \underline{a}_1] \subseteq [-\infty, x_0]$ and $\overline{a}(]0,1]) \subseteq [\overline{a}_1, \infty] \subseteq [x_0, \infty]$.

3 Finite fuzzy numbers

In this section we define FNs whose image is finite. In general, given two nonempty sets X and Y, we will say that a function $f : X \to Y$ is *finite* if the image of f is a finite subset of Y. The following notion is a key piece of the current study.

A FN $\mathcal{A}$ is *finite* if its image, $\mathcal{A}(\mathbb{R})$, is a finite subset of $\mathbb{I}$. If Λ is a finite subset of $\mathbb{I}$, we will denote by $\mathcal{F}_\Lambda$ the family of all finite FNs $\mathcal{A} \in \mathcal{F}$ such that $\mathcal{A}(\mathbb{R}) \subseteq \Lambda$.

An equivalent way to represent finite FNs by using a real number (that can be interpreted as its *center*) and a finite set of nonnegative real numbers (that we will call its *spreads*) is as follows.

Given $n \in \mathbb{N}$, $\{\alpha_i\}_{i=0}^n \subset \mathbb{I}$ verifying $0 = \alpha_0 < \alpha_1 < \alpha_2 < \ldots < \alpha_{n-1} < \alpha_n = 1$, $A^c \in \mathbb{R}$ and $A^m, A^{\ell,1}, A^{\ell,2}, \ldots, A^{\ell,n-1}, A^{r,1}, A^{r,2}, \ldots, A^{r,n-1} \in [0, \infty)$, there exists a unique finite FN $\mathcal{A}$ such that

$$\underline{a}_\alpha = \begin{cases} A^c - A^m, & \text{if } \alpha_{n-1} < \alpha \leq 1, \\ A^c - A^m - A^{\ell,n-1} - A^{\ell,n-2} - \ldots - A^{\ell,i+1}, & \text{if } \alpha_i < \alpha \leq \alpha_{i+1} \\ \qquad\qquad (\text{for some } i \in \{1, 2, \ldots, n-2\}), & \\ A^c - A^m - A^{\ell,n-1} - A^{\ell,n-2} - \ldots - A^{\ell,1}, & \text{if } 0 \leq \alpha \leq \alpha_1; \end{cases} \tag{1}$$

$$\overline{a}_\alpha = \begin{cases} A^c + A^m, & \text{if } \alpha_{n-1} < \alpha \leq 1, \\ A^c + A^m + A^{r,n-1} + A^{r,n-2} + \ldots + A^{r,i+1}, & \text{if } \alpha_i < \alpha \leq \alpha_{i+1} \\ \qquad\qquad (\text{for some } i \in \{1, 2, \ldots, n-2\}), & \\ A^c + A^m + A^{r,n-1} + A^{r,n-2} + \ldots + A^{r,1}, & \text{if } 0 \leq \alpha \leq \alpha_1. \end{cases} \tag{2}$$

The FN $\mathcal{A}$ is the unique finite FN whose center is A^c, whose central spread is A^m, whose left spreads are $\{A^{\ell,i}\}_{i=1}^{n-1}$, whose right spreads are $\{A^{r,i}\}_{i=1}^{n-1}$ and whose image is included in $\{\alpha_i\}_{i=0}^n$.

Then, a finite FN can be equivalently determined in the way

$$\mathcal{A} = \text{FN}\left(\{\alpha_i\}_{i=0}^n, A^c, A^m, \left\{A^{\ell,i}\right\}_{i=1}^{n-1}, \left\{A^{r,i}\right\}_{i=1}^{n-1}\right),$$

 ISBN: 978-84-617-8694-7

which has the advantage that all spreads are nonnegative, that is,

$$A^c \in \mathbb{R} \quad \text{but} \quad A^m, A^{\ell,i}, A^{r,i} \geq 0.$$

If we set

$$\Lambda = \{\alpha_0 = 0, \alpha_1, \ldots, \alpha_{n-1}, \alpha_n = 1\} \subset \mathbb{I},$$

then we will use the notation

$$\mathcal{A} = \mathrm{FFN}_\Lambda \left(A^c, A^m, \left\{A^{\ell,i}\right\}_{i=1}^{n-1}, \left\{A^{r,i}\right\}_{i=1}^{n-1} \right) \tag{3}$$

to describe any FN $\mathcal{F}$ in $\mathcal{F}_\Lambda$, where A^c is its center, and A^m, $\left\{A^{\ell,i}\right\}_{i=1}^{n-1}$ and $\left\{A^{r,i}\right\}_{i=1}^{n-1}$ are its (nonnegative) spreads.

4 Ranking FNs with compact support

It is well-known that there is a lack of universally acceptable total ordering between FNs. However, ranking of FNs is a useful tool to deal with decision-making problems. Therefore, many orderings have been proposed by several authors. Next, we show a novel method for ordering FNs using countable FNs. This method has similar properties to a specificity [19, 9].

Let $\Lambda = \{\alpha_i\}_{i\in I}$ be a countable (or finite) and strictly increasing sequence of real numbers such that $0 \leq \alpha_0 < \alpha_1 < \ldots < \alpha_i < \ldots \leq 1$. Given any FN $\mathcal{A}$ with compact support, we may consider the FN $\mathcal{A}^\Lambda$ given, for all $x \in \mathbb{R}$, by:

$$\mathcal{A}^\Lambda(x) = \begin{cases} 1, & \text{if } x \in \mathcal{A}_{[1]}, \\ \alpha_i, & \text{if } x \in \mathcal{A}_{[\alpha_i]} \setminus \mathcal{A}_{[\alpha_{i+1}]} \quad (i \geq 0), \\ 0, & \text{otherwise.} \end{cases}$$

Then $\mathcal{A}^\Lambda$ is a countable (or finite) FN, has compact support, its image is in $\{\alpha_i\}_{i\in I} \cup \{1\}$ and $\mathcal{A}^\Lambda_{[\alpha_i]} = \mathcal{A}_{[\alpha_i]}$ for all $i \in I$. Let $S_\Lambda(\mathcal{A})$ be the area of the histogram given by the plot of $\mathcal{A}^\Lambda$, i.e.,

$$S_\Lambda(\mathcal{A}) = \ell^{\mathcal{A}}_1 + \textstyle\sum_{i\geq 1} \alpha_i (\ell^{\mathcal{A}}_{\alpha_i} - \ell^{\mathcal{A}}_{\alpha_{i+1}}) \qquad \text{for all } \mathcal{A} \in \mathcal{F} \text{ with compact support,}$$

where $\ell^{\mathcal{A}}_{\alpha_i} = \overline{a}_{\alpha_i} - \underline{a}_{\alpha_i}$ is the length of the corresponding α_i-cut $\mathcal{A}_{[\alpha_i]} = [\underline{a}_{\alpha_i}, \overline{a}_{\alpha_i}]$. Notice that $S_\Lambda(\mathcal{A})$ can be interpreted as a Choquet integral [3, 5] corresponding to the Euclidean measure. As

$$\sum_{i\geq 1} \alpha_i (\ell^{\mathcal{A}}_{\alpha_i} - \ell^{\mathcal{A}}_{\alpha_{i+1}}) \leq \sum_{i\geq 1} \left[\ell^{\mathcal{A}}_{\alpha_i} - \ell^{\mathcal{A}}_{\alpha_{i+1}} \right] = \ell^{\mathcal{A}}_{\alpha_1} - \lim_i \ell^{\mathcal{A}}_{\alpha_i} \leq \ell^{\mathcal{A}}_{\alpha_1} \leq \mathrm{length}(\mathrm{supp}(\mathcal{A})) < \infty,$$

 ISBN: 978-84-617-8694-7

then the real number $S_\Lambda(\mathcal{A})$ is well-defined. This function allows us to induce an ordering (depending on Λ) on the set of all FNs with compact support given by

$$\mathcal{A} \prec_\Lambda \mathcal{B} \text{ if, and only if, } S_\Lambda(\mathcal{A}) < S_\Lambda(\mathcal{B}); \quad \mathcal{A} \sim_\Lambda \mathcal{B} \text{ if, and only if, } S_\Lambda(\mathcal{A}) = S_\Lambda(\mathcal{B}).$$

We will write $\mathcal{A} \preccurlyeq_\Lambda \mathcal{B}$ if $\mathcal{A} \prec_\Lambda \mathcal{B}$ or $\mathcal{A} \sim_\Lambda \mathcal{B}$. Some basic properties of the previous ordering are as follows.

- If $\mathcal{A} \preccurlyeq_\Lambda \mathcal{B}$ then $\mathcal{A}+\mathcal{C} \preccurlyeq_\Lambda \mathcal{B}+\mathcal{C}$ for all $\mathcal{C}$.
- If $\mathcal{A} \preccurlyeq_\Lambda \mathcal{B}$ then $r\mathcal{A} \preccurlyeq_\Lambda r\mathcal{B}$ for all $r \geq 0$.
- If $\mathcal{A}_{[\alpha]} = \mathcal{B}_{[\alpha]}$ for all $\alpha \in \Lambda$, then $\mathcal{A} \sim_\Lambda \mathcal{B}$.
- If $\mathcal{A} \leq \mathcal{B}$ (that is, $\mathcal{A}(x) \leq \mathcal{B}(x)$ for all $x \in \mathbb{R}$), then $\mathcal{A} \preccurlyeq_\Lambda \mathcal{B}$ for all Λ.
- If $\mathcal{A} \leq \mathcal{B}$ and $\mathcal{A} \sim_\Lambda \mathcal{B}$, then $\mathcal{A}_{[\alpha]} = \mathcal{B}_{[\alpha]}$ for all $\alpha \in \Lambda$.
- $\mathcal{A} \sim_\Lambda \mathcal{A}^\Lambda$.

Although $\mathcal{A} \mapsto S_\Lambda(\mathcal{A})$ does not define a measure of specificity (see [19, 9, 8, 20]), $S_\Lambda(\mathcal{A})$ can be seen as a measure of the imprecision associated to a FN because the property of the fourth bullet recalls one of its main characteristic: the specificity measure of a normal fuzzy set decreases when the membership degree of its elements increases. Therefore, the above process can be interpreted as a specificity-type method. Most authors use triangular or trapezoidal FNs in order to describe their ordering methods. In that cases, the following properties may be useful.

- If $n \in \mathbb{N}$ and $\Lambda_n = \{0, 1/n, 2/n, \ldots, (n-1)/n, 1\}$ then

$$S_{\Lambda_n}(\mathcal{A}) = \frac{n-1}{n}\, S(\mathcal{A}) + \frac{\ell_1^{\mathcal{A}}}{n} \qquad \text{for all trapezoidal FN } \mathcal{A},$$

  where $S(\mathcal{A})$ denotes the area under $\mathcal{A}$ and $\ell_1^{\mathcal{A}} = \ell(\ker \mathcal{A})$ is the length of its kernel.

- If $\mathcal{A}$ and $\mathcal{B}$ are trapezoidal FNs such that $\operatorname{supp}(\mathcal{A}) \subseteq \operatorname{supp}(\mathcal{B})$, $\ker(\mathcal{A}) \subseteq \ker(\mathcal{B})$ and $\mathcal{A} \sim_\Lambda \mathcal{B}$ for some $\Lambda \subset [0,1]$ with, at least, three points, then $\mathcal{A} = \mathcal{B}$.

5 Numerical examples

In [13], the authors proposed a method for ranking FNs using areas. In this work, they showed several examples and compared their results with the methods used by other authors. We illustrated our decision rule using the FNs considered in [13, Example 4].

$$\mathcal{A}_1 = (1/2/5) \text{ (triangular)}, \quad \mathcal{A}_2(x) = \begin{cases} [1-(x-2)^2]^{1/2}, & \text{if } 1 \leq x \leq 2, \\ [1-(x-2)^2/4]^{1/2}, & \text{if } 2 \leq x \leq 4, \\ 0, & \text{otherwise.} \end{cases}$$

 ISBN: 978-84-617-8694-7

Obviously, our ordering process depends on Λ. If $\Lambda = \{0, 0.1, 0.2, 1\}$, then $S_\Lambda(\mathcal{A}_1) = 0.68$ and $S_\Lambda(\mathcal{A}_2) = 0.592435$, so $\mathcal{A}_2 \prec_\Lambda \mathcal{A}_1$. If $n \in \mathbb{N}$ and $\Lambda_n = \{0, 1/n, 2/n, \ldots, (n-1)/n, 1\}$, then:

$$S_{\Lambda_n}(\mathcal{A}_1) = \frac{2(n-1)}{n}, \quad S_{\Lambda_n}(\mathcal{A}_2) = \frac{3}{n^2} \sum_{k=1}^{n-1} \left[k \left(\sqrt{n^2 - k^2} - \sqrt{n^2 - (k+1)^2} \right) \right],$$

so $\mathcal{A}_1 \prec_{\Lambda_n} \mathcal{A}_2$ for all $n \in \mathbb{N}$, $n \geq 2$.

In [1], Abbasbandy and Asady proposed a modification of the distance based approach (called the *sign distance* that depends on a real number $p \geq 1$) in order to overcome the shortcomings of previous techniques for ranking FNs. In their work, they did a complete comparison between their methodology and previous processes (Yager [18], Chen [4], Chu and Tsao [6], Yao and Wu [21] among others) using five subsets of triangular or trapezoidal FNs. Next, we use some of these subsets so as to compare their methodology with our technique.

The FNs $\mathcal{A} = (0.3/0.5/0.7)$, $\mathcal{B} = (0.3/0.5/0.8/0.9)$ and $\mathcal{C} = (0.3/0.5/0.9)$ were considered in [1] (Example 1, Set 3) and in [13] (Example 5, Set 3), producing the results showed in Table 1.

FNs	Yager	Chen	Chu-Tsao	Yao-Wu	Abbas.-Asady $p=1$	Abbas.-Asady $p=2$	Nejad-Mashinchi	Method with Λ_4
$\mathcal{A}$	0.5	0.375	0.25	0.5	1	0.7257	0.0714	0.15
$\mathcal{B}$	0.55	0.425	0.31526	0.625	1.25	0.9416	0.1387	0.4125
$\mathcal{C}$	0.625	0.55	0.27475	0.55	1.1	0.8165	0.1044	0.225
Result	$\mathcal{A} \prec \mathcal{B} \prec \mathcal{C}$	$\mathcal{A} \prec \mathcal{B} \prec \mathcal{C}$	$\mathcal{A} \prec \mathcal{C} \prec \mathcal{B}$	$\mathcal{A} \prec \mathcal{C} \prec \mathcal{B}$	$\mathcal{A} \prec \mathcal{C} \prec \mathcal{B}$	$\mathcal{A} \prec \mathcal{C} \prec \mathcal{B}$	$\mathcal{A} \prec \mathcal{C} \prec \mathcal{B}$	$\mathcal{A} \prec \mathcal{C} \prec \mathcal{B}$

Table 1: Comparison between fuzzy rankings, where $\mathcal{A} = (0.3/0.5/0.7)$, $\mathcal{B} = (0.3/0.5/0.8/0.9)$ and $\mathcal{C} = (0.3/0.5/0.9)$.

We emphasize that our results are reasonable and they do not coincide with other previous techniques.

Notice that ranking methods that verify the property in the fourth bullet could also be interpreted as a measure of the imprecision associated to a FN rather than an evaluation of its location. Thus, when there is no imprecision, this order does not permit us to distinguish between crisp FNs. In this sense, we point out that this method is in agreement with fuzzy set inclusion, but not with the ordering of real numbers. Comparing with the previous fuzzy rankings, this property could yield to a different ordering, as in the following example.

The FNs $\mathcal{A} = (0/0.4/0.7/0.8)$, $\mathcal{B} = (0.2/0.5/0.9)$ and $\mathcal{C} = (0.1/0.6/0.8)$ were also considered in [1] (Example 1, Set 4) and in [13] (Example 5, Set 4), producing the results showed in Table 2.

 ISBN: 978-84-617-8694-7

FNs	Yager	Chen	Chu-Tsao	Yao-Wu	Abbas.-Asady $p=1$	Abbas.-Asady $p=2$	Nejad-Mashinchi	Method with Λ_4
$\mathcal{A}$	0.45	0.52	0.24402	0.475	0.95	0.7853	0.2488	0.4875
$\mathcal{B}$	0.525	0.57	0.26243	0.525	1.05	0.7958	0.3631	0.2625
$\mathcal{C}$	0.55	0.625	0.2619	0.525	1.05	0.8386	0.3348	0.2625
Result	$\mathcal{A} \prec \mathcal{B} \prec \mathcal{C}$	$\mathcal{A} \prec \mathcal{B} \prec \mathcal{C}$	$\mathcal{A} \prec \mathcal{C} \prec \mathcal{B}$	$\mathcal{A} \prec \mathcal{B} \sim \mathcal{C}$	$\mathcal{A} \prec \mathcal{B} \sim \mathcal{C}$	$\mathcal{A} \prec \mathcal{B} \prec \mathcal{C}$	$\mathcal{A} \prec \mathcal{C} \prec \mathcal{B}$	$\mathcal{B} \sim \mathcal{C} \prec \mathcal{A}$

Table 2: Comparison between fuzzy rankings, where $\mathcal{A} = (0/0.4/0.7/0.8)$, $\mathcal{B} = (0.2/0.5/0.9)$ and $\mathcal{C} = (0.1/0.6/0.8)$.

Acknowledgements

This manuscript has been partially supported by Junta de Andalucía by projects FQM-268, FQM-245, FQM-265 of the Andalusian CICYE (University of Jaén). The first author is grateful to the Department of Quantitative Methods for Economics and Business of the University of Granada because of its support.

References

[1] S. Abbasbandy, B. Asady, Ranking of fuzzy numbers by sing distance, Inf. Sci. 176 (2006) 2405–2416.

[2] J.J. Buckley, L.J. Jowers, Monte Carlo Methods in Fuzzy Optimization, Studies in Fuzziness and Soft Computing, Volume 222, Springer, Berlin, 2008.

[3] L.M. de Campos, M.J. Bolaños, Characterization and comparison of Sugeno and Choquet integrals, Fuzzy Sets Syst. 52 (1992) 61-67.

[4] S. Chen, Ranking fuzzy numbers with maximizing set and minimizing set, Fuzzy Sets Syst. 17 (1985) 113–129.

[5] G. Choquet, Theory of capacities, Ann. Inst. Fourier 5 (1953) 131-295.

[6] T. Chu, C. Tsao, Ranking fuzzy numbers with an area between the centroid point and original point, Comput. Math. Appl. 43 (2002) 11–117.

[7] D. Dubois, H. Prade, Operations on fuzzy numbers, International Journal of System Sciences 9 (6) (1978) 613–626.

 ISBN: 978-84-617-8694-7

[8] D. Dubois, H. Prade, The principle of minimum specificity as a basis for evidential reasoning, in: B. Bouchon, R.R. Yager (Ed.), Uncertainty in Knowledge-Based Systems, Springer, Berlin, 1987, pp. 75–84.

[9] D. Dubois, H. Prade, A note on measures of specificity for fuzzy sets, Internat. J. General Systems 10 (1995) 279–283.

[10] P. Grzegorzewski, E. Mrówka, Trapezoidal approximations of fuzzy numbers, Fuzzy Sets Syst. 153 (2005) 115–135.

[11] G.J. Klir, B. Yuan. Fuzzy sets and fuzzy logic. Prentice Hall, 1995.

[12] M. Mizumoto, J. Tanaka, Some properties of fuzzy numbers. 153–164.

[13] A.M. Nejad, M. Mashinchi, Ranking fuzzy numbers based on the areas on the left and the right sides of fuzzy number, Comput. Math. Appl. 61 (2011) 431–442.

[14] A. Roldán, J. Martínez-Moreno, C. Roldán, On interrelationships between fuzzy metric structures, Iran. J. Fuzzy Syst. 10 (2) (2013) 133-150.

[15] C. Roldán, A. Roldán, J. Martínez-Moreno, A fuzzy regression model based on distances and random variables with crisp input and fuzzy output data: a case study in biomass production, Soft Computing 16 (5) (2012) 785–795.

[16] A. Roldán, J. Martínez-Moreno, C. Roldán, Some applications of the study of the image of a fuzzy number: Countable fuzzy numbers, operations, regression and a specificity-type ordering, Fuzzy Sets Syst. **257** (2014) 204-216.

[17] C.X. Wu, M. Ma, The Basic of Fuzzy Analysis, National Defence Industry Press, Beijing, 1991.

[18] R.R. Yager, A procedure for ordering fuzzy subsets of the unit interval, Inf. Sci. 24 (1981) 143–161.

[19] R.R. Yager, On the specificity of a possibility distribution, Fuzzy Sets Syst. 50 (3)(1992) 279–292.

[20] R.R. Yager, Measures of specificity over continuous spaces under similarity relations, Fuzzy Sets Syst. 159 (2008) 2193–2210.

[21] J. Yao, K. Wu, Ranking fuzzy numbers based on decomposition principle and signed distance, Fuzzy Sets Syst. 116 (2000) 275–288.

[22] L.A. Zadeh, Fuzzy set. Inform. Control 8 (1965) 338–353.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

Lie Symmetries for a generalized fourth order nonlinear wave equation

M. Rosa[1], J.C. Camacho[1], M.S. Bruzón[1] and M.L Gandarias[1]

[1] *Departamento de Matemáticas, University of Cádiz*

emails: maria.rosa@uca.es, jcarlos.camacho@uca.es, m.bruzon@uca.es, marialuz.gandarias@uca.es

Abstract

In this work, we consider a generalized fourth order nonlinear wave equation from the point of view of the theory of symmetry reductions in partial differential equations.

Key words: Symmetry reductions, partial differential equations

1 Introduction

The description of physical, biological and other process is frequently given by nonlinear partial differential equations (PDEs) and their solutions play an important role in the understanding of these process. A large number of publications has been done in this area and many methods have been derived for finding analytical solutions for integrable nonlinear PDEs. Recently the nonlinear wave equation

$$u_t + \alpha u^n u_x - \delta(u^m u_x)_x + u_{xx} + \sigma u_{xxx} + u_{xxxx} = 0 \tag{1}$$

has been considered in [5], where the author has derived exact solutions for equation (1) by using Painlevé property. The machinery of Lie group theory provides the systematic method to search for special group-invariant solutions. For PDEs with two independent variables, as is equation (2), a single group reduction transforms the PDE into ODEs, which are generally easier to solve than the original PDE. Most of the required theory and description of the method can be found in [6, 1, 2] and many recent papers using this method have been published in [3, 7, 4].

 ISBN: 978-84-617-8694-7

The aim of this work is to consider the following generalization of Eq. (1)

$$u_t + u_{xx} + cu_{xxx} + u_{xxxx} + f(u)u_x - g'(u)u_x^2 - g(u)u_{xx} = 0. \tag{2}$$

We apply Lie classical method to equation (2) in order to obtain exact solution, as well as, to derive conservation laws for these equations by using the multipliers method. We also will apply the multiplier method to the reduced ODEs in order to reduce the order directly.

2 Classical symmetries and reductions

Lie classical method is based on the determination of the symmetry group of a differential equation, i.e., the largest group of transformations acting on dependent and independent variables of the equation so that maps solutions of the equation into other solutions.
In order to apply Lie classical method to equation (2) we consider the one-parameter Lie group of infinitesimal transformations in (x, t, u), where ϵ is the group parameter. The symmetry group of equation (2) will be given by the set of vector fields of the form

$$\mathbf{v} = \xi(x,t,u)\partial_x + \tau(x,t,u)\partial_t + \eta(x,t,u)\partial_u. \tag{3}$$

Equation (2) admits a Lie point symmetry provided that

$$pr^{(4)}\mathbf{v}(\Delta) = 0 \qquad \text{when} \qquad \Delta = 0,$$

where $\Delta = u_t + u_{xx} + cu_{xxx} + u_{xxxx} + f(u)u_x + g'(u)u_x^2 - g(u)u_{xx}$ and $pr^{(4)}\mathbf{v}$ is the fourth prolongation of the vector field (3), we obtain a set of determining equations for the infinitesimals $\xi(x,t,u)$, $\tau(x,t,u)$ and $\eta(x,t,u)$. By solving the determining system we obtain the following results:

Case 1. The point symmetries admitted by equation 2 in the general case are generated by

$$\mathbf{v_1} = \partial_x, \qquad \mathbf{v_2} = \partial_t. \tag{4}$$

Case 2. For $f(u) = -\dfrac{\ln(f_1+u)}{f_0} + f_2$ and $g(u) = g_0$, with f_0, f_1, f_2 and g_0 arbitrary constants, besides $\mathbf{v_1}$ and $\mathbf{v_2}$ we obtain a new generator:

$$\mathbf{v_3} = t\partial_x - f_0(f_1+u)\partial_u. \tag{5}$$

Case 3. For $f(u) = f_0 + f_1(-g_1+u)^{\frac{3g_0}{2}}$, $g(u) = (g_1-u)^{g_0}g_2 + 1$ and $c = 0$, with f_0, f_1, f_2, g_0, g_1 and g_2 arbitrary constants, besides $\mathbf{v_1}$ and $\mathbf{v_2}$ we obtain a new generator:

$$\mathbf{v_4} = \left(t + \frac{x}{3f_0}\right)\partial_x + \frac{4t}{3f_0}\partial_t + \frac{2(g_1-u)}{f_0\, g_0}\partial_u. \tag{6}$$

 ISBN: 978-84-617-8694-7

Case 4. For $f(u) = f_0 u + f_1$, $g(u) = g_0$ and $c = 0$, with f_0, f_1, and g_0 arbitrary constants, besides $\mathbf{v_1}$ and $\mathbf{v_2}$ we obtain a new generator:

$$\mathbf{v_5} = t\partial_x + \frac{1}{f_0}\partial_u. \tag{7}$$

The corresponding generators of the optimal system of subalgebras for each case respectively are:

$$\langle \lambda\mathbf{v}_1 + \mathbf{v}_2\rangle,\ \langle \lambda\mathbf{v}_1 + \mathbf{v}_2,\quad \beta\mathbf{v}_2 + \mathbf{v}_3\rangle,\ \langle \lambda\mathbf{v}_1 + \mathbf{v}_2,\quad \mathbf{v}_4\rangle,\ \langle \lambda\mathbf{v}_1 + \mathbf{v}_2,\quad \beta\mathbf{v}_2 + \mathbf{v}_5\rangle$$

where $\lambda, \beta \in \mathbb{R}$ are arbitrary. In the following, reductions of Eq. (2) to ODE's are obtained using the generators of the optimal system.

Reduction 1: For case 1 and by using the generator $\lambda\mathbf{v}_1 + \mathbf{v}_2$ we obtain the similarity variable and similarity solution

$$z = x - \lambda t, \quad u = h(z) \tag{8}$$

and the ODE_1

$$h_{z\,z\,z\,z} + c\,h_{z\,z\,z} - g\,h_{z\,z} + h_{z\,z} - g_h\,(h_z)^2 + f\,h_z - \lambda\,h_z = 0. \tag{9}$$

Reduction 2: For case 2, by using the generator $\beta\mathbf{v}_2 + \mathbf{v}_3$ and setting $f_0 = 1$, $f_1 = f_2 = 0$, we obtain the similarity variable and similarity solution

$$z = \beta x - \tfrac{t^2}{2}, \quad u = h(z)e^{\frac{-t}{\beta}} \tag{10}$$

and the ODE_2

$$-\beta^5\,h_{z\,z\,z\,z} - \beta^4\,c\,h_{z\,z\,z} + \left(\beta^3\,g_0 - \beta^3\right)\,h_{z\,z} + \beta^2\,\log(h)\,h_z + h = 0. \tag{11}$$

Reduction 3: For case 3, by using the generator $\mathbf{v}_4$ and setting $f_0 = g_0 = 1$, $g_1 = 0$, we obtain the similarity variable and similarity solution

$$z = xt^{\frac{-1}{4}} - t^{\frac{3}{4}}, \quad u = \frac{h(z)}{\sqrt{t}} \tag{12}$$

and the ODE_3

$$-4\,h_{zzzz} - 4\,g_2\,h\,h_{zz} - 4\,g_2\,h_z^2 + z\,h_z - 4\,f_1\,h^{\frac{3}{2}}\,h_z + 2\,h = 0. \tag{13}$$

Reduction 4: For case 4 and by using the generator $\beta\mathbf{v}_2 + \mathbf{v}_5$, we obtain the similarity variable and similarity solution

$$z = \beta x - \frac{t^2}{2}, \quad u = \frac{t}{\beta f_0} - h(z) \tag{14}$$

and the ODE_4

$$-\beta^4\,h_{z\,z\,z\,z} + \beta^2\,g_0\,h_{z\,z} - \beta^2\,h_{z\,z} + \beta\,f_0\,h\,h_z - \beta\,f_1\,h_z + \frac{1}{\beta\,f_0} = 0. \tag{15}$$

 ISBN: 978-84-617-8694-7

3 Conclusions

In this work, we have obtained Lie symmetries of the fourth order nonlinear wave equation (2). We have derived the optimal system of one-dimensional subalgebras of the invariant equation and we have obtained reductions to ODE's. All the low order conservation laws will be achived by using the multiplier method. Moreover, we will search for exact solutions.

Acknowledgements

The support of Junta de Andalucía group FQM-201 is gratefully acknowledged.

References

[1] G. W. Bluman, A. Cheviakov and S. C. Anco, *Applications of Symmetry Methods to Partial Differential Equations*, Springer, New York, 2009.

[2] G. W. Bluman, S. Kumei, *Symmetries and Differential Equations, Applied Mathematical Sciences*, Springer-Verlag, New York, 1989.

[3] M. S. Bruzón, M. L. Gandarias, *Classical And Nonclassical Symmetries for the Krichever-Novikov Equation*, Theoretical and Mathematical Physics, **168(1)** (2011) 875-885.

[4] H. Liu, J. Li, Q. Zhang, *Lie symmetry analysis and exact explicit solutions for general Burgerséquation*, Journal of Computational and Applied Mathematics, **228(1)** (1996) 1-9.

[5] N. Kudryashov, *Painlevé analysis and exact solutions of the fourth-order equation for description of nonlinear waves*, Commun Nonlinear Sci Numer Simulat, **28** (2015) 1–9.

[6] P. Olver, *Applications of Lie groups to differential equations*, Springer-Verlag, 1993.

[7] R. Tracinà *Fundamental solution in classical elasticity via Lie group method*, Applied Mathematics and Computation, **218(9)** (2012) 5132-5139.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

Measuring distance between subsequences in temporal series, for pattern recognition using particle swarm optimization

Jesús Rosado[1] **and Juan Moreno-Garcia**[2]

[1] *Universidad de Castilla-La Mancha, EII de Toledo, Departamento de Matemáticas*

[2] *Universidad de Castilla-La Mancha, EII de Toledo, Departamento de Tecnología de la información y Sistemas*

emails: jesus.rosado@uclm.es, juan.moreno@uclm.es

Abstract

Particle Swarm Optimization (PSO) algorithms have been shown to be highly efficient in finding patterns which appear repeatedly in a temporal series. A key component in these algorithms is the notion of distance used to compare sequences inside the series. In this paper we study the behavior of a variant of the usual PSO algorithm, where stronger interaction between the agents of the swarm is taken into account, when we use some naive distances.

Key words: Time Series, Pattern Recognition, Particle Swarm Optimization.
MSC 2000: 62-04, 62M10

1 Introduction

Time series appear naturally in any situation where data can be collected. The analysis has been long used in economy, biology, physics, medicine, industry, sociology, etc. as a mean to understand all sort of phenomena from measurements taken at different times. A particularly important application of this analysis consists in making predictions about the evolution of the object of study. A key tool to do so, relays in the identification of patterns that repeat themselves through the series, which is, in itself, a relevant structural information about the data [10, 11]. By interpreting as a time series any sequence, despite being obtained in a different context, such as imaging, we can apply tools developed for time series analysis to an even broader scope [8].

 ISBN: 978-84-617-8694-7

In the literature we may find many different approaches to this subject. Typically, the proposed algorithms are restricted to the comparison of same-length intervals in the series, albeit we can find some references where this restriction does not apply [15, 17, 18, 20]. In the last decade, algorithms based on what is known as Particle Swarm Optimization (PSO) [2, 3] have become popular due to their high performance and applicability to very different contexts [15, 11, 5, 13].

The idea behind the particle swarm optimization algorithms is to compare *simultaneously* different pairs of subsequences within the time series, chosen randomly. These pairs are considered to be the position of an agent in the swarm. A measure of how good a given position -in the ambient space of our agents- is given. Then each agent checks how good its current position is and moves to a new one, taking into account the best positions found, by itself and all any of the other individuals in the swarm, to choose which direction it should follow.

The goal of this paper is, in the one hand, to review the structure of PSO algorithms and how it can be interpreted in the context of pattern recognition, as well as summarizing some recent related works. In the other hand, we want to stress the importance of choosing an adequate notion of distance to compare subsequences of the time series.

Section 2 is dedicated to the formulation of the problem. We will also briefly review the common structure of PSO algorithms in this context, as well as discussing some of the considerations that should be taken into account. Next, in section 3 we propose some modification to the usual algorithm based on general swarming algorithms [9, 12, 4], in order to accelerate the convergence, and study different options to compare pairs of subsequences in the temporal series. This paper is intended as a preliminary work to establish the starting ground upon which we can build; hence we will limit ourselves to consider a few naive notions of distance, to exemplify how much, even in this cases, the performance of the PSO algorithm is dependent on what it means to be close for two subsequences of series. We present the results of our simulations and compare the performance achieved with each distance in Section 4.

2 A particle swarm optimization algorithms for pattern recognition

As we said in the introduction the basic idea of a PSO algorithm is to generate a swarm of agents that move in an appropriate space so that each of them can evaluate the adequacy of the position it occupies according to a fitness measure and explore its surroundings to find the position where the optimal is achieved. To do so, it is useful to let the agents remember

 ISBN: 978-84-617-8694-7

the position that yielded the best outcome of this measure. With this idea in mind, an agent can be thought of as a data structure $\mathsf{a}_i=\{\mathsf{x}_i, \mathsf{v}_i, \mathsf{cv}_i, \mathsf{bv}_i, \mathsf{bvx}_i\}$, where x_i represents the current position of the agent, v_i the velocity with which it is moving, i. e. the rule according to which it will choose a new position to explore, cv_i the value of the fitness measure in its current position, bv_i the best value it has found and bx_i is the position corresponding to that best value. We should also think of the swarm as an entity of itself, consisting of the collection of all agents, to which we will associate the best value and position encountered by any agent, gbv and gbx.

Using this framework, all PSO algorithms share a common core consisting in a double loop such as the one described by Algorithm 1, where, following with the analogy with the laws of movement, we shall call *acceleration* to the rate of variation of the velocity of the agents.

Algorithm 1 Core of the PSO algorithm.

```
1: while "optimal is improving" do
2:     for ALL agents IN swarm do
3:         dv_i ← Acceleration(agent, Swarm)
4:     end for
5:     for ALL agents IN swarm do
6:         x_i ← ← x_i) + dt· v_i
7:         v_i ← v_i + dt· dv_i
8:         cv_i ← OptMeasure(x_i)
9:         if cv_i < bv_i then
10:            bv_i ← cv_i
11:            bx_i ← x_i
12:            if gbv > bv_i then
13:                gbv ← bv_i
14:                gbx ← bx_i
15:            end if
16:        end if
17:    end for
18: end while
```

In order to use it for our purpose, we need an appropriate notion of position and velocity. Since our aim is to compare subsequences of a time series, it is natural to identify the position of an agent with these subsequences. Following the ideas in [15] we will define the position to be a 4-tuple, $\mathsf{x}_i=[\mathsf{x}_i^0,\mathsf{l}_i^0,\mathsf{x}_i^1,\mathsf{l}_i^1]$, where the first and third elements denote the start of the subsequences to compare and the second and fourth, their respective lengths.

 ISBN: 978-84-617-8694-7

An easy way to codify this, is to identify x_i^0 and x_i^1 with indices within the series and, if we assume that the time series is evenly distributed, l_i^0 and l_i^1 with the number of indices that constitute each subsequence. Thus, we can think of an agent moving in a subset $\mathbb{N}^4$. This approach implies strict restrictions in some (potentially all) of steps 3, 4 and 5 of the PSO algorithm. Namely, since x must be an unsigned integer, some adjustments are in order: we may force dt and v, and hence dv, to also be integers or we can correct the final value of x at each step to meet the requirements. Additional checks are necessary to ensure that the indices remain within meaningful boundaries: x_i^0 and x_i^1 must be smaller than the size of the time series. In [15] it is also required that $\mathsf{x}_i^1+\mathsf{l}_i^1$ is smaller than the the size of the time series, so that the whole subsequence referenced is included in the reference time series, and that $\mathsf{x}_i^0+\mathsf{l}_i^0<\mathsf{x}_i^1$, to avoid overlap between the subsequences and thus, the trivial solution where one of the subsequences is contained in the other. We will start by following the same considerations, in order to compare with their results. Nevertheless, our aim is to achieve this behavior trough the acceleration, rather than verifying in each step that all conditions are satisfied.

Finally, we want to point out that the terminating condition for the main loop has been stated in such an ambiguous way deliberately. In general, one may consider that when there is no improvement in the global best value the loop should come to conclusion. Against this, can be argued that even if there is no change in the global optimal position for a few iterations, while there exist some agent whose position is improving, it could happen that eventually it gets to a better position than the best already found. It is also possible to defend the thesis that even if there is no improvement in any of the agents, it can be due to the randomness of the exploration. In any case, a compromise between these postures must be achieved, and the analysis of the particular optimality function and ambient space for the agents is of great importance in this task.

3 Variations in the Acceleration and Optimality functions

As we have seen in the previous section, one of the key steps of the PSO algorithm is the computation of the optimality function, since it is there where the general algorithm is fitted to solve a specific problem. Nevertheless, the way in which the information acquired with this computation is used to dictate the movement of the agents may also affect the performance of the PSO. In this section we will address this two aspects of the implementation of the PSO algorithm.

3.1 Movement of the agents in the swarm

In a usual PSO algorithm, the change in velocity is computed as a linear combination of the current velocity, the direction to the best position encountered by the agent and the

 ISBN: 978-84-617-8694-7

direction to the best position found by the swarm:

$$\mathsf{v}_i^{n+1}=\mathsf{A}\mathsf{v}_i^n + \mathsf{B}(\mathsf{x}_i^n\text{-}\mathsf{bvx}_i) + \mathsf{C}(\mathsf{x}_i^n\text{-}\mathsf{gvx})$$

and then, the position updated as

$$\mathsf{x}_i^{n+1}=\mathsf{x}_i^n+\mathsf{v}-i^{n+1}.$$

This formulation is reminiscent of the phenomenon it want to mimic, albeit it is not uncommon that the values of the weights are chosen either randomly, or *ad hoc*. Also, the independence of the step-size and the rate of change carries the potential for instability of the algorithm [6]. While this is not necessarily a problem, since PSO algorithms are usually intended to randomly explore the ambient space where the agents live, this approach may be missing some of the good features of swarm behavior, it could benefit from.

In Algorithm 1 we already propose a slight variation to the standard PSO algorithm so that it is more in accordance with the formulation of swarming models, maintaining the second order formulation of the movement [1, 7, 19]. The inclusion of the coefficient dt allows us to regulate how much the rate of change in position and velocity will affect the current values, and thus avoid jumping over a minimum of the optimality function. Also, we change the velocity through the computation of an acceleration, where the desired directions will be taken into account. This will provide us a more general framework where different interaction rules between the agents can be included, as well as some self-propulsion term which shall help as deal with the restrictions on the positions of the particle mentioned in the previous section. Furthermore, this will help us to properly balance the weight of all the desired effects on the rate of change of the velocity. We will define the acceleration as

$$\mathsf{dv}_i^{n+1} = \omega_{\mathsf{I}}\mathsf{INTERACTION}+\omega_{\mathsf{SP}}\mathsf{SELF\_PROPULSION}+\omega_{\mathsf{PB}}(\mathsf{x}_i^n\text{-}\mathsf{bvx}_i) + \omega_{\mathsf{gb}}(\mathsf{x}_i^n\text{-}\mathsf{gbx}),$$

with $\omega_{\mathsf{I}} + \omega_{\mathsf{SP}}\omega_{\mathsf{PB}}\omega_{\mathsf{GB}} = 1$. Strictly speaking, the two last terms should be included in the SELF_PROPULSION and the INTERACTION terms respectively, but we keep them a part due to their relevance. The SELF_PROPULSION term represents the preferred velocity of the agent, in absence of the rest of the swarm, and can be used to provide the agent with a natural desire to remain within the boundaries dictated by the series and avoid positions corresponding to overlapping subsequences. In this first approach, though, we will overlook this effect and set ω_{SP} to zero. With respect to the INTERACTION term, we will limit ourselves to a basic attraction effect, nudging the agents to converge in the same position. While this can be counterproductive, as it may lead to concentration in a local minimum of the optimality function, it also facilitates a thorough exploration of the whole ambient

 ISBN: 978-84-617-8694-7

space of the agents if they are suitably spread during the initialization.

Algorithm 2 Interaction rule based solely in attraction.

$$\mathrm{dv}_i \leftarrow \frac{1}{\mathsf{SWARM\_SIZE}} \sum_{i \neq j} (x_j - x_i)|x_i - x_j|^{\alpha}$$

3.2 Distances between time series subsequences

In the context of pattern recognition between the subsequences of time series, it is natural to adopt the distance between subsequences as a measure of how good the position of an agent is. However, although intuitively the concept of distance is very clear, when it comes to a precise definition, there are many notions of distance that can be used [16]. The choice of one over the other, may imply also a different way of understanding the agents and will yield drastically different results. At this moment, the reference distance to evaluate the efficiency of new measures is the *Dynamic Time Warping* (see [14]). Our objective at this stage is not to provide a similarity measure that outperforms the currently available ones, but rather to explore how the choice of a distance affects not only the performance of the algorithm, in terms of speed, but also the behavior of the agents themselves, favoring the exploration of different regions.

With this idea in mind we propose two simple similarity functions. The first one is defined in the setting given in Section 2:

Algorithm 3 Direct comparison similarity function.

function DCDistance(AGENT, TS) ▷ TS is the time series that we want to study
  $seq1$ =TS([x^0,x^0+l^0])
  $seq2$ =TS([x^1,x^0+l^1])
  D=l^1-l^0 ▷ Assuming $l^1 > l^0$
  for $k \leftarrow 0$ To D **do**
    $$dif_k = \sqrt{\frac{1}{l^0} \sum_{l=0}^{l^0-1} (seq1(l) - seq2(k+l))^2}$$
  end for
  return $\min_k\{dif_k\}$
end function

This measure is a direct extension of the usual euclidean distance for vectors. We allow to compare sequences of different lengths by comparing the smaller one with each of the

 ISBN: 978-84-617-8694-7

subsequences of the same length in the larger one, without penalizing the proportion of the larger one that we are not able to compare. To compensate the different sizes of the sequences that every agent is comparing, we divide the sum of squared differences by the amount of terms in that sum.

We propose a second measure, which requires a small modification in the way we understand the agents, or rather, the space where they will move. An agent will still be defined by its position, velocity, current value, best value and best position, but we will understand the position as a subset of $\mathbb{R}^4$. From now on, x^0, x^1 shall denote any time between the first and the last time stamp of the series and l^0, l^1 will be the lengths of time intervals starting at x^0, x^1 respectively. By disengaging the agents from the actual points in the series we gain freedom of movement, at the price of needing to locate the right points in the series to evaluate the optimality of each position.

Algorithm 4 Interpolation similarity function.

function IDistance(AGENT, TS)
  $L \leftarrow \mathrm{l}^1/\mathrm{l}^0$
  $seq1x$ = Time-stamps of the series such that $[\mathrm{x}^0,\mathrm{x}^0+\mathrm{l}_0] \subseteq [seq1x(0), seq1x(end)]$, shifted to the origin.
  $seq2x$ = Time-stamps of the series such that $[\mathrm{x}^1,\mathrm{x}^1+\mathrm{l}_1] \subseteq [seq2x(0), seq2x(end)]$, shifted to the origin and scaled by L.
  $seq1v$ = original values of the series corresponding to $seq1x$.
  $seq2v$ = original values of the series corresponding to $seq2x$ scaled by L.
  $dif \leftarrow 0$.
  for $i \leftarrow 1$ To LENGTH($seq1x$) **do** ▷ We assume that $seq1x$ is shorter than $seq2x$
    $p \leftarrow seq1x(i)$
    $v \leftarrow$ INTERPOLATE(pa, pb)
      ▷ pa and pb being the pairs (x_i2p, v_i2p), (x_i2n, v_i2n), where $i2p$ and $i2n$ is the index of the elements in $seq2x$ such that $x_i2p < p < x_i2n$.
    $dif \leftarrow dif + |seq1v(i) - v|$
  end for
  return $\dfrac{dif}{\mathrm{LENGTH}(seq1x)}$
end function

 ISBN: 978-84-617-8694-7

4 Results

In order to study the behavior of the method that we present in this work, we have implemented an program using the Python language, which allows the use of different objective functions. In this first version, we have used as objective function the distances DCDistance and IDistance detailed in Section 3.2.

IDistance is more directly inspired by [15]. We also encounter the overlapping issues that are mentioned there. Due to the lack of space, we omit a detailed description of the tests that we have tried and present directly the results that we have obtained together with the conclusions:

1. The algorithm always converges.

2. As we could expect, when one of the subsequences is contained in the other, the results of the PSO algorithm is better, since we are comparing a subsequence with a portion of itself.

3. Whenever the intersection of the two subsequences is not empty, the agents find that the optimal position lies in the direction that increases the overlapping. The smaller subsequence is *absorved* by the larger one. Some authors advise against allowing this to happen (see for instance [15]). As we mentioned in the introduction, to enforce this restriction a control must be kept through all the execution of the code, since, if the subsequences are close enough, an overlapping may happen, even if initially there are no intersections.

4. Occasionally, the agents show a *static* behavior. By this we mean that the position of the agent only changes very slightly, i. e., the subsequences that it is comparing remain the same. This happens when the random initialization generates an agent in a position corresponding to big overlapping.

5. When the agents are instantiated in positions corresponding to far enough subsequences, the PSO algorithm terminates before reaching and acceptable optimal. We want to point out that this happens because there has been no improvement in the best value of any of the agents, which indicates that this distance should not be used as a measure of optimality.

As an example of test, we present the results obtained with a swarm of five agents. We have generated the agents with random positions. Four of them correspond to subsequences which do not overlap, while the fifth shows a strong overlapping.

 ISBN: 978-84-617-8694-7

Table 1: Test of the PSO algorithm with DCDistance.

Iter		P_1	P_2	P_3	P_4	P_5
0	xb_1	$[489, 199, 966, 119]$	$[689, 83, 918, 190]$	$[860, 182, 445, 135]$	$[200, 118, 78, 154]$	$[342, 217, 963, 96]$
	bv_1	100.3341	49.9639	60.3542	0.5226	77.5270
5	xb_5	$[484, 195, 954, 119]$	$[689, 83, 918, 190]$	$[853, 179, 441, 135]$	$[200, 118, 78, 154]$	$[338, 213, 951, 96]$
	bv	94.1980	49.9639	59.9998	0.5226	70.1617
10	xb_{10}	$[472, 190, 922, 119]$	$[689, 83, 918, 190]$	$[837, 175, 431, 135]$	$[200, 118, 78, 154]$	$[331, 208, 919, 96]$
	bv_{10}	75.7245	49.9639	59.6598	0.5226	49.0530
15	xb_{15}	$[458, 186, 883, 119]$	$[689, 83, 918, 190]$	$[837, 175, 431, 135]$	$[200, 118, 78, 154]$	$[327, 206, 902, 96]$
	bv_{15}	53.2119	49.9639	59.6595	0.5226	39.16991

This results in a fast convergence of the algorithm, needing just fifteen iterations. Table 1 shows the results. We show an iteration in every five. In each iteration we will expose the best positions and best values obtained by each agent. This table allows us to study the path of descent to the optimal position followed by the agents.

Next, focus on the results obtained with the second distance function studied.

1. This function does not strengthen the preference of the agents for positions corresponding to overlapping between subsequences. This is a consequence of comparing the subsequences as a whole.

2. In the same line as the previous remark, it is rare to find an overlapping, and when one occur, the position does not benefit from it through this distance.

3. The battery of tests is insufficient to assert the good performance of the PSO algorithm using this distance as the objective function. Nevertheless, we have usully obtained an optimal value close to 0.5, which is encouraging.

As with the distance DCDistance, we want to present a test done using IDistance as the objective function. We describe the results in the following table, which is structured in the same way as Table 1:

We show an iteration in every seven. The test converges in fifty-seven iterations. We want to stress that, contrary to what happen using DCDistance as objective function, even if the starting position is bad, all agents will explore the ambient space, and find a path to a position comparable with the best position of the swarm.

 ISBN: 978-84-617-8694-7

Table 2: Test of the PSO algorithm with IDistance.

Iter		P_1	P_2	P_3	P_4	P_5
0	xb_0	$[752, 193, 195, 121]$	$[493, 77, 140, 59]$	$[301, 224, 812, 130]$	$[718, 93, 585, 73]$	$[414, 187, 98, 131]$
	bv_0	7.9779	0.6635	7.2596	9.8150	0.6703
7	xb_7	$[752, 193, 195, 121]$	$[488, 76, 135, 58]$	$[301, 224, 812, 130]$	$[709, 92, 572, 72]$	$[414, 181, 98, 125]$
	bv_7	7.9779	0.6604	7.2595	8.2968	0.6231
14	xb_{14}	$[718, 179, 181, 109]$	$[481, 76, 128, 58]$	$[301, 224, 812, 130]$	$[688, 92, 538, 72]$	$[414, 179, 98, 123]$
	bv_{14}	7.9165	0.5473	7.2596	5.5782	0.6130
21	xb_{21}	$[686, 171, 170, 102]$	$[479, 76, 126, 58]$	$[318, 198, 676, 109]$	$[658, 92, 487, 72]$	$[414, 167, 98, 111]$
	bv_{21}	7.3441	0.5118	5.8909	3.8156	0.5864
28	xb_{28}	$[645, 157, 156, 95]$	$[479, 76, 126, 58]$	$[336, 182, 585, 102]$	$[658, 92, 487, 72]$	$[414, 159, 98, 107]$
	bv_{28}	5.9943	0.5118	3.3099	3.81565	0.5260
35	xb_{35}	$[597, 143, 142, 88]$	$[460, 76, 107, 58]$	$[359, 161, 475, 95]$	$[579, 92, 347, 72]$	$[414, 159, 98, 107]$
	bv_{35}	3.7798	0.4515	0.5862	3.0046	0.5260
42	xb_{42}	$[542, 123, 131, 81]$	$[453, 78, 106, 58]$	$[371, 152, 424, 90]$	$[531, 92, 265, 72]$	$[428, 127, 101, 81]$
	bv_{42}	2.3167	0.4241	0.3987	1.1148	0.46137
49	xb_{49}	$[480, 109, 128, 74]$	$[448, 83, 111, 58]$	$[371, 152, 424, 90]$	$[476, 92, 188, 72]$	$[431, 111, 116, 65]$
	bv_{49}	0.7149	0.4133	0.3987	0.5776	0.4536
57	xb_{57}	$[417, 95, 128, 67]$	$[442, 86, 117, 58]$	$[371, 152, 424, 90]$	$[428, 92, 122, 72]$	$[431, 103, 128, 60]$
	bv_{57}	0.4280	0.4077	0.3987	0.4174	0.4004

5 Conclusions

We have modified the usual PSO algorithm to allow a larger interaction between the agents and applied it to the study of patterns within time series. We have considered two (naive) similarity functions and confirmed how strongly the behavior of the agents depends on how the distance between subsequences is understood. The distance DCDistance presents serious issues concerning convergence; it shows a strong tendency to select as optimal the trivial solutions where there is a strong overlapping between subsequences. The distance IDistance shows promising results. From the tests done, the evident trivial optimal states have a negligibly small basin of attraction.

Acknowledgements

Supported by the project TIN2015-64776-C3-3-R of the Science and Innovation Ministry of Spain, co-funded by the European Regional Development Fund (ERDF).

References

[1] I. Aoki, *A simulation study on the schooling mechanism in fish.*

Bull. Japanese Society of Scienti

c Fisheries, **48(8)**, (1982) 1081-1088.

[2] A. Banks, J. Vincent, and C. Anyakoha, *A review of particle swarm optimization. Part I: background and development.* Nat. Comput. **6(4)**, (2007) 467484.

[3] J. Barrera, and C. Coello-Coello, *A reviewof particle swarm optimization methods used for multimodal optimization.* C.P. Lim, L.C. Jain, S. Dehuri (Eds.), Innovations in Swarm Intelligence, Studies in Computational Intelligence, (2009) **248**.

[4] J. A. Cañizo, J. A. Carrillo, and J. Rosado, *Collective behavior of animals: swarming and complex patterns*, Arbor: Ciencia, pensamiento y cultura **746** (2010) 1035–1049.

[5] Y.-Y. Hong, A. A. Beltran Jr., and A. C. Paglinawan, *A Chaos-Enhanced Particle Swarm Optimization with Adaptive Parameters and Its Application in Maximum Power Point Tracking*, **2016** (2016.) Mathematical Problems in Engineering.

[6] R. L. Burden and J. D. Faires, *Numerical Analysis* Cengage Learning, Boston, 2016.

[7] F. Cucker and S. Smale. *Emergent behavior in ocks. IEEE Transactions on automatic control*, (2007) **52(5):852**.

[8] I. S. Mitzev and N. H. Younan, *Time Series Shapelets: Training Time Improvement Based on Particle Swarm Optimization.* International Journal of Machine Learning and Computing, **5(4)** (2015).

[9] S. Motsch, and E. Tadmor, *A new model for self-organized dynamics and its flocking behavior.* Journal of Statistical Physics, **144(5)** (2011) 923947.

[10] A. Mueen,*Enumeration of time series motifs of all lengths.* Proceedings of the IEEE International Conference on Data Mining (ICDM), (2013) 547556.

[11] A. Mueen, *Time series motif discovery: dimensions and applications.* WIREs Data Min. Knowl. Discov. **4(2)** (2014) 152159.

[12] M. R. DOrsogna, Y.-L. Chuang, A. L. Bertozzi, and L. Chayes, *Self-propelled particles with soft-core interactions. patterns, stability, and collapse.* Phys. Rev. Lett. **96** (2006) 104302.

[13] K. E. Parsopoulos, M. N. Vrahatis, *Particle Swarm Optimization and Intelligence: Advances and Applications*, IGI Global, Hershey, USA, (2010).

 ISBN: 978-84-617-8694-7

[14] C. A. Ratanamahatana, and E. Keogh, *Everything you know about dynamic time warping is wrong.* Proceedings of the ACM SIGKDD Workshop on Mining Temporal and Sequential Data. (2004) 2225.

[15] J. Serrá, and J. L. Arcos, *Particle swarm optimizatio for time series motif discovery.* Knowledge-Based Systems, **92** (2016).

[16] J. Serr, and J. L. Arcos, *An empirical evaluation of similarity measures for time series classification.* Knowladge-Based Systems, **67** (2014) 305314.

[17] Y. Tanaka, K. Iwamoto and K. Uehara, *Discovery of time-series motif from multidimensional data based on MDL principle.* Mach. Learn. **58** (2005) 69300.

[18] H. Tang, and S. S. Liao, *Discovering original motifs with different lengths from time series.* Knowl. Based Syst., **21** (2008) 666671.

[19] T. Vicsek and A. Zefeiris, *Collective motion. Physics Reprints*, **517** (2012) 71-140.

[20] D. Yankov, E. Keogh, J. Medina, B. Chiu, and V. Zordan, V., *Detecting time series motifs under uniform scaling.* Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD) (2007) 844853.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference on Computational and Mathematical Methods in Science and Engineering, CMMSE 2017 4–8 July, 2017.

Automatic generation of textual reports from thermal comfort data by using a statistical procedure

Clemente Rubio-Manzano[1], Carlos Rubio-Bellido[2], Alexis Perez-Fargallo[3], Jesus Pulido-Arcas[3] and Alejandro Martínez-Rocamora[3]

[1] *Department of Information System, University of the Bio-Bio*

[2] *Department of Building Construction II, Higher Technical School of Building Engineering, University of Sevilla*

[3] *Department of Building Science, Faculty of Architecture, Construction and Design, University of the Bio-Bio*

emails: clrubio@ubiobio.cl, carlosrubio@us.es, aperez@ubiobio.cl, jpulido@ubiobio.cl, amartinez@ubiobio.cl

Abstract

In this paper, a statistical procedure for automatically generating textual reports from thermal comfort data is proposed. These reports provide design engineers with useful information to keep adequate comfort level conditions for occupants in buildings. In order to show and explore the possibilities of this procedure, we have implemented a software prototype called YADY for Thermal Comfort.

Key words: Summarization of data, Thermal Comfort Data, Textual Reports, Statistical Procedure, Software Application

1 Introduction

Thermal comfort in buildings has become a main research topic in recent years, since it allows designers to maintain adequate comfort level conditions through passive measures, hence reducing energy consumption [3, 7]. To this end, designers usually employ software applications to simulate a building model in order to obtain comfort predictions. However, the amount of data generated by such applications is extremely large and its correct interpretation becomes a complex task [1, 4]. In addition, the data generated by simulation software is used to identify certain behavioral patterns and to predict future trends.

 ISBN: 978-84-617-8694-7

In this paper, a method to overcome these difficulties on interpreting comfort simulation results is proposed. The use of linguistic data summaries has been identified as a possible solution, which can provide HVAC (heating, ventilation, and air conditioning) design engineers with a better and easier understanding of the results obtained. For example, data summaries have been employed for automatically generating advice for saving energy at home [2].

In particular, an automatic generation of textual reports as a translation of the data provided by the simulation software is implemented in order to provide design engineers with knowledge about particular situations that occur in the building in a particular time span (i.e. annually, monthly, daily, and other personalized time spans). Below, examples of the reports generated are shown:

- **Annual Report:** "*During this year (8760 hours), the building was in a comfort situation with a ratio of 60%; 10% of the time, the building was in a discomfort situation (180 hours over comfort temperatures and 240 under comfort temperatures). The rest of the time (30%), it was in a slightly cold situation during 430 hours and in a slightly warm situation during 608 hours. The months with the highest ratio of comfort situations were December, January, February and March with an average of 80%; the months with the lowest ratio of comfort situations were June, July and August with an average of 50%. In the rest of the months, the average of comfort situations, was 60%*".
- **Monthly Report:** "*In January, the number of comfort situations detected had an acceptable ratio of 75%. There were slightly cold situations covering 20 hours and slightly hot situations for 30 hours. The best days were the 16th, 18th, 19th and 20th, with a ratio of 90% in acceptable conditions. The worst days were the 1st, 2nd and 3rd, with a 60% ratio of discomfort situations. Cold conditions are usually produced between 2am and 9am, with an outdoor average temperature of* $6°C$*. Warm conditions are usually produced between 3pm and 7pm, with an average outdoor temperature of* $32°C$".
- **Daily Report:** "*For the day 07/07, comfort conditions were detected with a 75% ratio; there were cold situations over 2 hours and a little bit of warm for 1 hour. Discomfort situations were usually detected between 2am and 9am, with an average exterior temperature of* $6°C$".
- **Personalized Report:** "*In January, comfort conditions were maintained with a ratio of 75%. Some cold situations were detected covering 20 hours, while a few warm situations were detected covering a total of 30 hours*".

A statistical procedure for the linguistic summarization of thermal comfort data is here proposed. The report, which is conveyed in a linguistic style, can be easily included as a narrative component in other applications with the same objective. The structure of the paper is as follows. In Section 2, the procedure to automatically generate linguistic reports from thermal comfort data is presented. Section 4 gives details about the implementation and experimentation. Finally, in Section 5 conclusions and future lines of work are drawn

 ISBN: 978-84-617-8694-7

from the present study.

2 Procedure for automatically generating linguistic reports from thermal comfort data

The proposed procedure aims to determine which is the comfort situation of the building (cold, sightly cold, comfort, sightly warm, or warm)in a particular period of time (annually, monthly, daily). To this end, the adaptive thermal comfort model defined in the standard EN 15251 [5] is employed, where a set of variables is used in order to infer a particular situation. These variables are called comfort metrics, and are the following:

- Daily Average Temperature (DAT)
- Weighted Average Temperature (WAT)
- Operative Temperature (OP_EN)
- Upper Limit of Category I (Upper L1)
- Lower Limit of Category I (Lower L1)
- Upper Limit of Category III (Upper L3)
- Lower Limit of Category III (Lower L3).)

The procedure starts reading the data exported by the simulation tools or importing the data from an intermediate file (XLS) (see Table 1). The input variables are: Date-Hour, Indoor Operative Temperature, and Outdoor Air Temperature.

Table 1: Input data exported, table T

Id	Date-Hour	Indoor Operative Temperature	Outdoor Air Temperature
1	01/01/2002 1:00	23,97483	13,35
2	01/01/2002 2:00	23,34624	12,2
3	01/01/2002 3:00	22,8341	11,4
...	...	...	...
8760	31/12/2002 24.00	24,84875	13,8

After that, Algorithms 1, 2, 3 and 4 (see Appendix) have been designed and implemented in order to compute the values for each comfort metric based on EN 15251 (see Table in Figure 1).

 ISBN: 978-84-617-8694-7

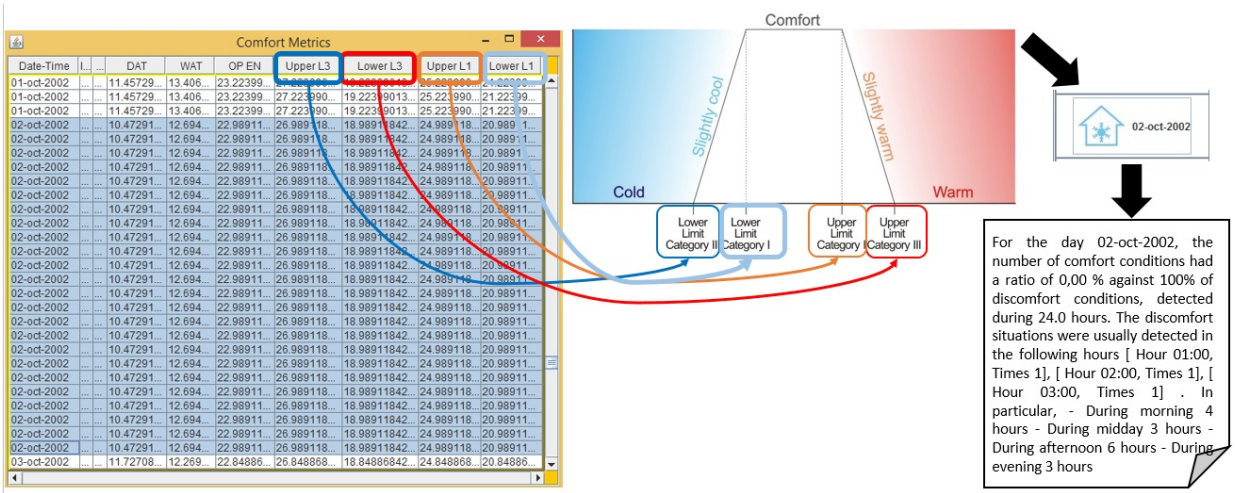


Figure 1: Data generated by means of the comfort model and the computation of a particular comfort situation based on the categories proposed in the model EN 15251

The table shown in Figure 1 contains the values for each comfort metric. A new database containing only two attributes is then created: $R_i = \{\text{date, current situation of comfort }\}$. Each row has the form: $R_i : D_i\{cold \mid slightly\_cold \mid comfort \mid slightly\_warm \mid warm\}$, where only one value can appear in each row, that is, the symbol "|" is acting as the conjunction "or". The reduced database has the following form:

```
R 1 : D 1 [cold | slightly cold | comfort | slightly warm | warm]
R 2 : D 2 [cold | slightly cold | comfort | slightly warm | warm]
...
R N : D N [cold | slightly cold | comfort | slightly warm | warm]
```

However, the ratio of comfort, discomfort and borderline (slightly cold and warm) situations have to be computed and arranged by day, month and year. From the existing database, the total number of cold, slightly cold, comfort, slightly warm and warm situations can be computed by counting the frequency of occurrence in each R_i. Therefore, the ratio is obtained as the summation of the number of a particular figure over the total number of situations detected, which are defined as follows: RATIO_COLD, RATIO_SIGHTLY_COLD, RATIO_COMFORT, RATIO_SIGHTLY_WARM, and RATIO_WARM. These ratios represent the percentage for each situation during a given year.

Now, as each D_i has the form "day-month-year" a sub database that represents either a year, a month or a day can be generated. For example, given "$D_i = d - m - y$", a sub database for "$m - y$", that is, for a particular month can also be generated:

```
R1: [m-y] [cold | slightly cold | comfort | slightly warm | warm]
```

 ISBN: 978-84-617-8694-7

```
R2: [m-y] [cold | slightly cold | comfort | slightly warm | warm]
...
Rk: [m-y] [cold | slightly cold | comfort | slightly warm | warm]
```

In this case, the ratio of comfort, discomfort and borderline (slightly cold and warm) situations are computed for a particular month (m). In the same way, the ratio can also be obtained as the summation of the number of a particular time span over the number total of situations that were detected in a month. These ratios represent the percentage for each situation during a particular month. Additionally, the linguistic report is also able to convey information about an important matter: the best and the worst months during a given year.

3 Reports Template

In this research, four types of reports are generated, namely annual, monthly, daily and personalized. Only first one will be hereby explained in detail, since the rest them are analogously created.

The Annual Report Template is based on the pattern explained in the introduction. This report template has been created in order to provide designers with the most relevant details about the comfort and discomfort situations during a particular year. The process of automatically creating an annual report requires a link to the annual report template that contains the ratio of comfort situations ($RATIO\_COMFORT$), the ratio of discomfort situations ($RATIO\_COLD + RATIO\_WARM$), and the ratio of both slightly cold and slightly warm situations ($RATIO\_SIGHTLY\_COLD + RATIO\_SIGHTLY\_WARM$).

Additionally, the number of hours where the temperature was above the comfort temperature (O), the number of hours where the temperature was below the comfort temperature (U), the number of hours where the temperature was slightly cold (SC) and slightly warm (SW), and the best M_i and worst M_j months should be linked by the report template module. Hence, the report template is defined as follows:

During the year [Ye], the building was in a [COMFORT | DISCOMFORT] situations with a ratio of [X] per-cent, only [Y] per-cent of the time it was in a [COMFORT | DISCOMFORT]** situations ([O] hours over the comfort temperature, [M] hours under the comfort temperatures). The rest of the time it was in a [SLIGHTLY WARM | SLIGHTLY COLD]* during [SC | SW]* and [SLIGHTLY WARM | SLIGHTLY COLD]* during [SC | SW]* hours. The rest of the ratios were the following: [Ratio Cold; Ratio Sightly Cold; Ratio Sightly Warm; Ratio Warm]. The best month was M_i with a comfort ratio of [B]. The worst month was M_j with a discomfort ratio of [W]*

Note that, X1= [a—b] * X2= [a—b] ** means that if X1=a then X2=b, and if X1=b then X2=a. The annual report template could also be restricted to query about a particular

 ISBN: 978-84-617-8694-7

season: “During the [winter | spring | autumn | summer], the situation of the building was [cold | slightly cold | comfort | slightly warm | warm]”.

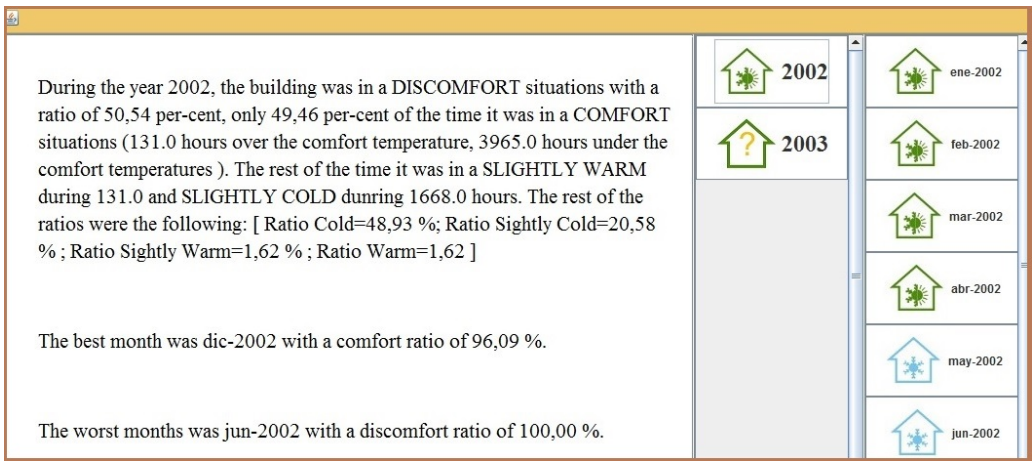


Figure 2: Automatically data generated annual Linguistic report

4 Implementation and Experimentation

YADY (”Your actions define you”) is a new technology whose objective is to get automated behavior recognition from the actions performed by entities interacting in a dynamic environment. This environment is usually modeled by physical parameters, which can be captured by using sensors, electronic devices, biometric tools and so on. This kind of environment is modified by its entities. Well-established metrics can be used to measure these changes. Next, a relationship can be established between these different entities by using linguistic terms. Linguistic terms allow us to express these numerical metrics and relationships in words. This automated behaviour recognition can be applied to many systems, depending on the application-context. As a first approach, YADY technology has been applied to improve players’ experience in computer games (Rubio-Manzano and Trivino 2016).

In this paper, this technology has been adapted and applied to automatically generate linguistic descriptions in the scientific field of thermal comfort. The behavior of a building with regards to the thermal comfort can be inferred by analyzing the data captured by the simulation tool. The procedure presented in the previous sections has been implemented in the application “YADY Thermal Comfort Tool”, which can be downloaded using the URL: `http://youractionsdefineyou.com/yady_tc/`. This software tool has been implemented by using the Java programming language.

As a result, the experiment generates three reports, one for each type. From an input

 ISBN: 978-84-617-8694-7

file, the user can generate all desired reports for a particular period of time, just by indicating the granularity: year, month or day.

The user interface for annual reports allows users to visualize an automatically generated report for a particular year. On the right side of the screen, users can see if discomfort situations occurred, and also click on it for more details about this phenomenon (Figure 2). This interface works in an analogous way for monthly and daily reports.

5 Conclusions and future work

In this study, a procedure for automatically generating textual reports from thermal comfort data has been presented. A software application based on this procedure, which is able to generate building's comfort use reports from comfort metrics, has been developed. This new resource provides designers with useful information regarding adaptive thermal comfort assessment.

This work is a first approximation towards a broader approach, and much work remains to be done in this direction. The proposed model is simple and effective in showing all the potential of natural language as a feedback mechanism for building designers.

Research on how this procedure could be implemented by using Linguistic Description of Complex Phenomena could represent an interesting future line of work. Therefore, a comparison between crisp and fuzzy paradigms is an important challenges for us.

Appendix: Algorithmms

Algorithm 1: Algorithm for the creation of the database for comfort situations

```
1 begin
2 |  To create a List L of elements (Date-Hour,Indoor-Operative-Temperature,Outdoor-Air-Temperature,
  |    DAT,WAT,OP_EN,Upper L1,Lower L1,Upper L3,Lower L3) by reading the input file
3 |  To compute_Daily_Average_Temperature(L)
4 |  To compute_Weighted_Average(L);
5 |  To compute_Operative_Temperature_EN_C1_C3(L);
6 |  To create T2 by using L
7 |  To remove T and L
8 |  return T2
9 end
```

Algorithm 2: Algorithm for computing Daily Average Temperature (DAT)

```
1 begin
2 |  foreach element i of L do
3 |  |  L[i].DAT = L[i].DAT/24.0;
4 |  end
5 end
```

 ISBN: 978-84-617-8694-7

Algorithm 3: Algorithm for computing the Weighted Average Temperature (WAT)

```
1  begin
2      cont=0;
3      foreach element i of L do
4          if cont==23 then
5              cont=0;
6              if i-168 > 0 then
7                  L[i].WAT=L[i-24].DAT;
8                  L[i].WAT+=(L[i-48].DAT)*0.8;
9                  L[i].WAT+=(L[i-72].DAT)*0.6;
10                 L[i].WAT+=(L[i-96].DAT)*0.5;
11                 L[i].WAT+=(L[i-120].DAT)*0.3;
12                 L[i].WAT+=(L[i-144].DAT)*0.6;
13                 L[i].WAT+=(L[i-168].DAT)*0.2;
14                 L[i].WAT= L[i].WAT/3.8;
15             end
16         end
17         Else ++cont;
18     end
19     value=0;
20     foreach element i of L do
21         if L[i].WAT > 0.0 then
22             for j to 24 do
23                 if i+j < size(L) then
24                     L[i+j].WAT=L[i].WAT;
25                 end
26                 i+=23;
27             end
28         end
29     end
30 end
```

Algorithm 4: Algorithm for Computing Operative Temperature, Upper-Lower Limit Category I, Upper-Lower Category III based on EN (OT_EN)

```
1  begin
2      foreach element i of L do
3          L[i].OT_EN=(0.33*L[i].WAT)+18.8;
4          L[i].Upper_Limit_3=L[i].OT_EN+4.0;
5          L[i].Lower_Limit_3=L[i].OT_EN-4.0;
6          L[i].Upper_Limit_1=L[i].OT_EN+2.0;
7          L[i].Lower_Limit_1=L[i].OT_EN-2.0;
8      end
9  end
```

Acknowledgements

This work has been done by the research group SOMOS (Software-Modelling-Science) funded by the Research

Agency and Graduate School of Management of the University of the Bío-Bío, in collaboration with the Sustainable Architecture and Construction Research Group (GACS) at the University of the Bío-Bío. We would like to say that this paper is part of the FONDECYT research project 3160806 "Study of the feasible energy improvement standard for social housing in fuel poverty by means of post occupational adaptive comfort assessment and its progressive implementation" funded by the Chilean National Commission for Research in Science and Technology.

References

[1] Chowdhury, A.A., Rasul, M.G., Khan, M.M.K. *Thermal-comfort analysis and simulation for various low-energy cooling-technologies applied to an office building in a subtropical climate.* Applied Energy 85(6):449-462 (2008)

[2] P. Conde-Clemente, J.M. Alonso and G. Trivino, *Toward automatic generation of linguistic advice for saving energy at home*, Soft Computing, 1-15 (2016).

[3] Feist, W., Schnieders, J., Dorer, V., Haas, A. *Re-inventing air heating: Convenient and comfortable within the frame of the Passive House concept.* Energy and Buildings 37(11 SPEC. ISS.):1186-1203 (2005)

[4] Karlsson, F., Rohdin, P., Persson, M.-L. *Measured and predicted energy demand of a low energy building: Important aspects when using building energy simulation.* Building Services Engineering Research and Technology 28(3):223-235 ((2007)

[5] Nicol, F., and Wilson, M. *An overview of the European Standard EN 15251. In pro-ceedings of Conference: Adapting to Change: New Thinking on Comfort.* Cumberland Lodge, Windsor, UK (Vol. 911), (2010, April).

[6] C. Rubio-Manzano, G. Trivino, *Improving player experience in computer games by using players' behavior analysis and linguistic descriptions*, Int. J. Hum.-Comput. Stud., 95 (2016), pp. 27-38.

[7] Omer, A.M. *Renewable building energy systems and passive human comfort solutions.* Renewable and Sustainable Energy Reviews 12(6):1562-1587 (2008)

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

Randomized response estimation in multiple frames surveys

María del Mar Rueda[1], Beatriz Cobo[1] and Pier Francesco Perri[2]

[1] *Department of Statistics and O. R., University of Granada, Spain*

[2] *Department of Economics, Statistics and Finance, University of Calabria, Italy*

emails: mrueda@ugr.es, beacr@ugr.es, pierfrancesco.perri@unical.it

Abstract

Surveys usually include sensitive topics as such as gambling, alcoholism, sexual behavior, domestic violence which characteristics are difficult to estimate using standard survey techniques because of the tendency of respondents to hold information in such settings. On the other hand, multiple frame surveys are becoming a widely used method to decrease bias due to undercoverage of the target population. In this work, we consider statistical techniques for handling sensitive data coming from a multiple frame survey using complex sampling designs. Our aim is to estimate the mean of undesirable behaviors when data are obtained by using a randomized response technique. Some estimators are constructed and their properties theoretically investigated. We also derive variance estimators.

Key words: complex surveys, randomized response techniques, multiple frames, calibration.
MSC 2000: AMS 62D05

1 Introduction

In socioeconomic or biomedical studies, very often the researcher has to gather information relating to highly sensitive issues. In these situations, posing direct questions to the respondents may procure untruthful responses or even refuse to respond because of social stigma or fear about threat of disclosure.. Such systematic response errors lead to social-desirability bias in estimates of the sensitive behaviors of interest, underestimating socially undesirable activities.

To overcome these problems, methods such as the randomized response (RR) technique (RRT) may be used to collect more reliable data, protect respondent's confidentiality and

 ISBN: 978-84-617-8694-7

avoid unacceptable rate of nonresponse. In the RRT, respondents use a randomization device to generate a probabilistic relationship between their answers and the true values of the sensitive characteristic. The RRT has been applied in surveys covering a variety of sensitive topics like racism, drug use, abortion, delinquency or AIDS.

The RRT was originated by Warner (1965) who proposed a data collection procedure that allows researchers to obtain sensitive information while guaranteeing privacy to respondents. Warner's study generated a rapidly-expanding body of research literature on alternative techniques for eliciting suitable RR schemes in order to estimate proportions, means or totals. A good review of different RR procedures is given in Chaudhuri and Christofides (2013) and Chaudhuri et al. (2016).

Traditionally, surveys have been carried out using three main methods of data collection: face-to-face interviews, mail surveys and telephone interviews. Over the last 20 years, the picture has changed sharply. Telephone surveys have become a popular mode of data collection, especially following the creation and development of computer-assisted telephone interviewing (CATI) systems. However, telephone surveys also present some drawbacks with regard to coverage, due to the absence of a telephone in some households and the generalized use of mobile phones, which are sometimes replacing fixed (land) lines entirely. The potential for coverage error as a result of the exponential growth of the cell phone-only population has led to the development of dual-frame surveys. Surveys where data are collected from three sampling frames are also used in practice. The popularity of multiple frame surveys has increased among scientific community along last years and now they are widely used both in statistical agencies and in private organizations. In the near future, importance of three frame surveys is expected to grow with the use of the internet for data collection (Lohr, 2010). Indeed, it is very likely that dual frame surveys consisting of a cell and a landline frame evolve to three frame surveys incorporating a third frame of web users. However, there are very few studies that address the problem of estimating sensitive behaviors from two frames. Recently Rueda et al. (2015) proposed some dual frame estimators for proportions and mean when the data are obtained by using the RRT. The aim of this paper, thus, is to propose new estimation techniques for sensitive parameters when data come from more than two frames.

2 Some generalities of estimation in multiple frames

We will employ the notation used in Mecatti (2007). Let U be a finite population composed of N units labeled from 1 to N, $U = \{1, ..., k, ..., N\}$ and let $A_1, \ldots, A_q, \ldots, A_Q$ be a collection of $Q \geq 2$ overlapping frames of sizes $N_1, \ldots, N_q, \ldots, N_Q$, all of them can be incomplete but it is assumed that overall they cover the entire target population U. With Q frames, there are $2^Q - 1$ possible distinct domains. Let the index sets K be the subsets of the range of the frame index $q = 1, \ldots, Q$. For every index set $K \subseteq \{1, \ldots, q, \ldots, Q\}$ a

 ISBN: 978-84-617-8694-7

domain is defined as the set $D_K = (\cap_{q\in K} A_q) \bigcap (\cap_{q\notin K} A_q^c)$, where ^{c} denotes the complement of a set (that is, D_K is the subset of units that are covered by all the frames A_q, $q \in K$, and by these frames only). Let y be a sensitive variable to study which cannot be observed directly. The objective is to estimate the population mean of y that is

$$\overline{Y} = \frac{1}{N} \sum_{k\in U} y_k, \tag{1}$$

where y_k is the value of the sensitive character for the k-th unit. This mean can be rewritten as follows

$$\overline{Y} = \frac{1}{N} \sum_{q=1}^{Q} \sum_{k\in A_q} \frac{y_k}{m_k}, \tag{2}$$

where m_k indicates the multiplicity of the k-th unit, i.e. the number of frames the unit k is included. Multiplicities m_k are needed in (2) to weight values y_k, otherwise, those units belonging to more than one frame would count more than once in the overall sum.

Let s_q be a sample drawn from frame A_q under a particular sampling design, independently for $q = 1, \ldots, Q$ and let $\pi_k(q)$ and $\pi_{kl}(q)$ be the first and second order inclusion probabilities under this sampling design, respectively. Let $d_k(q) = 1/\pi_k(q)$ be the sampling weight for units in frame q. Moreover, let n_q be the size of sample s_q and $s = \cup_q s_q$.

Lohr and Rao (2006) formulated the multiple frame extension of some of the estimators originally proposed for the dual frame case, as the one proposed by Hartley (1962, 1974) or by Fuller and Burmeister (1972). Although the optimal version of these estimators is asymptotically efficient, it is not internally consistent since a different set of weights is used for each response variable. Moreover, it is often unstable in small or moderate samples with more than two frames because the optimal estimated parameters involved in the computation of the estimators are functions of large estimated covariances matrices. Lohr and Rao (2006) also followed the so called single frame approach used by Kalton and Anderson (1986) to proposed a single frame estimator in a multiple frame context. This estimator is in the form:

$$\hat{\overline{Y}}_{KA} = \frac{1}{N} \sum_{k\in s} y_k d_k^{KA} \tag{3}$$

with $d_k^{KA} = \bar{\pi}_k^{-1}$, where $\bar{\pi}_k = \sum_{q'\ni k} \pi_k(q')$ where $q' \ni k$ denotes the sum for the frames containing to the unit k. To compute this estimator it is necessary to know not only the number of frames each unit belongs to, but also the specific frames the unit is included in. This can be an important drawback particularly if misclassification issues are present.

Lohr and Rao (2006) also proposed the following pseudo-maximum likelihood estimator for the multiple frame context:

$$\hat{\overline{Y}}_{PML} = \frac{1}{N} \sum_{k\in s} y_k d_k^{PML}(q), \tag{4}$$

 ISBN: 978-84-617-8694-7

where the weights d_k^{PML} can be defined as

$$d_k^{PML}(q) = d_k(q) f(q) \sum_{K:q\in K} \frac{\hat{N}_K \delta_k(K)}{\sum_{j\in K} f(j)\hat{N}_K(j)}$$

with $f(q) = \frac{1}{deff_z(q)} \frac{n_q}{N_q}$, being $deff_z(q)$ the design effect for variable z in the q-th frame. Values $\hat{N}_K(q)$ can be computed as $\hat{N}_K(q) = \sum_{k\in s_q} d_k(q)\delta_k(K)$, with $\delta_k(K)$ the indicator variable for domain K that takes the value 1 whether unit k belongs to domain K and 0 otherwise. The estimated domain sizes $\hat{N}_K$ are the solution of a system of non linear equations. The pseudo maximum likelihood is consistent and usually works well in practical situations but it is complex to compute for a general sampling design, since numerical procedures are required to obtain the values $\hat{N}_K$.

Mecatti (2007) also considered a single frame approach and proposed the following estimator

$$\hat{\bar{Y}}_M = \frac{1}{N} \sum_{k\in s} y_k d_k^M, \tag{5}$$

with $d_k^M = d_k/m_k$. The previous estimator, often called single frame multiplicity estimator, only requires the knowledge of the multiplicity of each unit, no matter which these frames are. This estimator can be adjusted using a raking ratio approach to get a single frame raking ratio multiplicity estimator where a new set of weights, resulting from an iterative procedure, is used.

Singh and Mecatti (2011) proposed a composite multiplicity estimator, which generalizes the single frame multiplicity estimator. This estimator can be written as

$$\hat{\bar{Y}}_{CM} = \frac{1}{N} \sum_{k\in s} y_k d_k^{CM} \tag{6}$$

where

$$d_k^{CM} = \frac{\lambda_k d_k + (1-\lambda_k) d_k^{KA}}{m_k}$$

with

$$\lambda_k = \frac{\sum_{q'\ni k}(1 - \bar{\pi}_k/\pi_k(q'))\pi_k(q')(1-\pi_k(q'))}{\sum_{q'\ni k}(1 - \frac{\bar{\pi}_k^2}{\pi_k(q')^2} - \frac{2\bar{\pi}_k}{\pi_k(q')})\pi_k(q')(1-\pi_k(q'))}.$$

Usually, additional information about auxiliary variables is available in surveys. Rao and Wu (2010) followed a single frame multiplicity based approach to extend the pseudo empirical likelihood estimator for the mean of a variable to the multiple frame setting. Calibration is also a well-known technique to deal with auxiliary information in estimation. Ranalli et al. (2016) proposed different calibration estimators for the dual frame case, which can be easily extended to the multiple frame context. Rueda et al. (2017) proposed model-assisted estimators for population proportions in multiple frames.

 ISBN: 978-84-617-8694-7

3 Randomized response techniques for multiple frames

We assume that the variable under study y cannot be observed directly and that in each frame it is possible to use a different RR procedure to collect data on it. In order to consider a wide variety of RR devices, we consider the unified approach given by Arnab (1996) according to the individuals in the sample s_q use a generic RR model denoted by RR_q. For each $k \in s_q$ the RR_q induces a random variable Z_{qk} so that the revised randomized response R_{qk} is an unbiased estimation of y_{qk}, the real value of the sensitive variable for the k-th unit in s_q. We consider RR_q, with $q = 1, ...Q$, to be independent randomized devices such that the respective revised randomized responses R_{qk} satisfy the conditions (see Arnab, 2004) $E_R(R_{qk}) = y_k$, $V_R(R_{qk}) = \sigma^2_{qk}$, $C_R(R_{qk}, R_{qj}) = 0$, where E_R, V_R and C_R denote the expectation, variance and covariance operators with respect to the RR mechanism.

Using the idea of Mecatti (2007) we propose the multiplicity estimator

$$\hat{\bar{Y}}_{RM} = \frac{1}{N} \sum_{q=1}^{Q} \sum_{k \in s_q} R_{qk} d_k^M(q), \tag{7}$$

with $d_k^M(q) = d_k(q)/m_k$.

4 Properties of the proposed estimator

In this section we describe the main properties of the proposed estimator.

Theorem 1. The multiplicity estimator $\hat{\bar{Y}}_{RM}$ is an unbiased estimator of the population mean $\overline{Y}$

Proof.

Let E_d and V_d denote the expectation and variance operators for any sampling design d. Hence, we have:

$$E(\hat{\bar{Y}}_{RM}) = \frac{1}{N} E_d E_R \left(\sum_{q=1}^{Q} \sum_{k \in s_q} R_{qk} d_k^M(q) \right) =$$

$$\frac{1}{N} E_d \left(\sum_{q=1}^{Q} \sum_{k \in s_q} E_R(R_{qk}) d_k^M(q) \right) = \frac{1}{N} E_d \left(\sum_{q=1}^{Q} \sum_{k \in s_q} y_{qk} d_k^M(q) \right) =$$

$$\frac{1}{N} \sum_{q=1}^{Q} E_d \left(\sum_{k \in s_q} y_{qk} d_k^M(q) \right) = \frac{1}{N} \left(\sum_{q=1}^{Q} \sum_{k \in A_q} y_{qk} \right) = \overline{Y}$$

Then, $\hat{\bar{Y}}_{RM}$ is an unbiased estimator of the population mean.

 ISBN: 978-84-617-8694-7

Theorem 2.
The variance of $\hat{\hat{Y}}_{RM}$ *is given by*

$$V(\hat{\hat{Y}}_{RM}) = \frac{1}{N^2}\left(\sum_{q=1}^{Q}\sum_{k\in A_q}\sigma_{qk}^2 d_k^M(q) + \sum_{q=1}^{Q}\sum_{k\in A_q}\sum_{j\in A_q} y_k y_j (d_k^M(q) d_j^M(q)\pi_{kj}(q) - 1)\right). \quad (8)$$

Proof. By using the properties of the Horvitz-Thompson estimator (Särndal et al. 1992), we have:

$$V(\hat{\hat{Y}}_{RM}) = E_d V_R(\hat{\hat{Y}}_{RM}) + V_d E_R(\hat{\hat{Y}}_{RM}) =$$

$$\frac{1}{N^2}\left(E_d\left(\sum_{q=1}^{Q}\sum_{k\in s_q} V_R(R_{qk})(d_k^M(q))^2\right) + V_d\left(\sum_{q=1}^{Q}\sum_{k\in s_q} E_R(R_{qk}) d_k^M(q)\right)\right) =$$

$$\frac{1}{N^2}\left(\sum_{q=1}^{Q}\sum_{k\in A_q} V_R(R_{qk})(d_k^M(q))^2\pi_k(q) + \sum_{q=1}^{Q}\sum_{k\in A_q}\sum_{j\in A_q} y_k y_j (d_k^M(q) d_j^M(q)\pi_{kj}(q) - 1)\right) =$$

$$\frac{1}{N^2}\left(\sum_{q=1}^{Q}\sum_{k\in A_q}\sigma_{qk}^2 d_k^M(q) + \sum_{q=1}^{Q}\sum_{k\in A_q}\sum_{j\in A_q} y_k y_j (d_k^M(q) d_j^M(q)\pi_{kj}(q) - 1)\right).$$

Hence the proof.

The variance of the estimator is composed of two terms which are common to all of the RR models. The second addendum depends on the sampling designs d_q and the y_k values in each frame. It denotes the variance of the Horvitz-Thompson estimator computed on the true values of y. The first term depends on the sampling designs and also on the randomization mechanism used in each frame. It represents the cost to pay, in terms of efficiency, to increase respondents' confidentiality.

In the formulation of the proposed estimators it is assumed that the population size N is known and Horvitz-Thompson estimator is used as baseline. Alternatively, one can consider Hájek-type estimators substituting N with $\hat{N}$, an unbiased design-based estimator of N. This is a special case of ratio estimator, and it can be more efficient than Horvitz-Thompson type estimators because the sample size in overlapping domains is not fixed. The estimators are thus approximately unbiased under certain conditions on the weights (see Lohr, 2009).

 ISBN: 978-84-617-8694-7

5 Variance estimation

Let $\hat{\sigma}^2_{qk}$ be an unbiased estimator of σ^2_{qk}. Hence, it can be proved that an analytical unbiased estimator of $V(\hat{\bar{Y}}_{RM})$ is given by:

$$\hat{V}(\hat{\bar{Y}}_{RM}) = \frac{1}{N^2}\left(\sum_{q=1}^{Q}\sum_{k\in s_q}\hat{\sigma}^2_{qk}d^M_k(q) + \sum_{q=1}^{Q}\sum_{k\in s_q}\sum_{j\in s_q} R_k R_j d^M_k(q) d^M_j(q)\left(1-\frac{\pi_k(q)\pi_j(q)}{\pi_{kj}(q)}\right)\right).$$

To define the estimator $\hat{V}(\hat{\bar{Y}}_{RM})$ we need $\pi_{kj}(q)$ for all units in each frame. In some common sampling designs (as cluster sampling with probability proportional to size) these probabilities are unknown or can be equal to 0 for some sampling units i, j. A simple alternative is the use of with replacement variance estimators (see Särndal et al.(1992), p. 99) or replicated sampling methods (see Wolter (2007) for a detailed description of these techniques in finite population sampling). Quenouille (1949) introduced the jackknife method to estimate the bias of an estimator by deleting one datum each time from the original data set and recalculating the estimator based on the rest of the data. In survey sampling it is usual to use jackknife techniques due to their simplicity and because they are implemented in general purpose software packages, such as R.

The last terms in (8),

$$\frac{1}{N^2}\sum_{q=1}^{Q}\sum_{k\in A_q}\sum_{j\in A_q} y_k y_j (d^M_k(q) d^M_j(q)\pi_{kj}(q) - 1) = V_{HT}$$

is a Horvitz-Thompson variance estimator with weights $d^M_k(q)$. If we consider a non-stratified design, the jackknife estimator for V_{HT} may be given by

$$v_J(\hat{V}(\hat{\bar{Y}}_{RM})) = \sum_{q=1}^{Q}\frac{n_q-1}{n_q}\sum_{i\in s_q}(\hat{\bar{Y}}_{RM}(i)(q) - \bar{Y}_{RM}(q))^2, \tag{9}$$

where $\hat{\bar{Y}}_{RM}(i)(q)$ is the value taken by estimator $\hat{\bar{Y}}_{RM}$ after eliminating unit i from s_q and $\bar{Y}_{RM}(q)$ is the average of $\hat{\bar{Y}}_{RM}(i)(q)$ values.

It is known that the jackknife variance estimator is asymptotically design unbiased for the variance of Horvitz-Thompson estimator (Wolter, 2007). So for the large sample size n we have $E(v_J(\hat{V}(\hat{\bar{Y}}_{RM}))) = V_{HT}$. Thus the adapted jackknife estimator

$$v_J(\hat{V}(\hat{\bar{Y}}_{RM}))^* = \sum_{q=1}^{Q}\frac{n_q-1}{n_q}\sum_{i\in s_q}(\hat{\bar{Y}}_{RM}(i)(q) - \bar{Y}_{RM}(q))^2 + \frac{1}{N^2}\left(\sum_{q=1}^{Q}\sum_{k\in A_q}\hat{\sigma}^2_{qk}d^M_k(q)\right)$$

is an unbiased estimator for the variance of the proposed estimator.

 ISBN: 978-84-617-8694-7

6 Estimation with auxiliary information

Usually, population level information about auxiliary variables is available in surveys. Let $\mathbf{x}_q = (x_{q1}, x_{q2}, \ldots, x_{qp_q})'$ be a set of p_q auxiliary variables observed in the q-th frame, so that the vector $\mathbf{x}_{qk} = (x_{q1k}, x_{q2k}, \ldots, x_{qp_qk})'$ includes the values taken by the variables $\mathbf{x_q}$ on the k-th unit in frame A_q. That is, we consider the case of complete auxiliary information. In addition, we consider the more general case in which auxiliary variables may differ in each frame, i.e. $\mathbf{x_q} \neq \mathbf{x_r}$, for $q, r = 1, \ldots, Q, q \neq r$. For the sample selected from frame A_q, the values of the variables $\{y_k, \mathbf{x}_{qk}\}$ are observed.

Calibration (Deville and Särndal, 1992) is also a well-known technique to deal with auxiliary information in estimation. Some works link RR models and calibration techniques together (Tracy and Singh (1999); Diana and Perri (2012)).

Ranalli et al. (2016) proposed different calibration estimators for the dual frame case. Using this idea we define a calibration estimator for randomized response to the multiple frame context. A calibration estimator in the case of several sampling frames can be defined as

$$\hat{\bar{Y}}_{CAL} = \frac{1}{N} \sum_{q=1}^{Q} \sum_{k \in s_q} d_k^{CAL}(q) R_k, \tag{10}$$

where $d_k^{CAL}(q)$ are such that they minimize $\sum_{q=1}^{Q} \sum_{k \in s_q} G\left(d_k^{CAL}(q), d_k^M(q)\right)$, where $G(\cdot, \cdot)$ is a particular distance function, subject to

$$\sum_{q=1}^{Q} \sum_{k \in s_q} d_k^{CAL}(q) \delta_k(A_q) = N_q, \quad q = 1, ..., Q$$

$$\sum_{q=1}^{Q} \sum_{k \in s_q} d_k^{CAL}(q) \mathbf{x}_{qk} \delta_k(A_q) = \mathbf{t}_{xq}, \quad q = 1, ..., Q,$$

where $\delta_k(A_q)$ is the indicator variable that takes value 1 if unit k is in frame q and zero otherwise, and $\mathbf{t}_{xq}$ are the population totals of $\mathbf{x_q}$.

The proposed model calibration estimator eliminates overestimation issues by several means. We consider $d_k^M(q)$ as the starting weights for the calibration and using the indicator variables $\delta_k(A_q)$, the calibration constraints ensure adjustment of the multiplicity issues by benchmarking all information on units from frame A_q included in the sample, irrespective of the frame they were originally selected from. Therefore, again, multiplicity is accounted for automatically by the constraints. The properties of this estimator can be derived from the properties of the calibrated estimators in multiple frames (see Ranalli et al. (2016))

 ISBN: 978-84-617-8694-7

7 Conclusions

In this paper, we present a new procedure aimed at determining a population sensitive mean by using a randomized response model when data are obtained from several frames. We introduce a way of combining estimates from the different frames. In practice, a different sampling procedure might feasibly be applied for each frame, or even no randomization at all (i.e., direct response) for a particular frame. The use of RR techniques has advantages but also drawbacks (the variance of estimates is increased by the randomization and individual response patterns cannot be interpreted directly, due to the observation of randomized responses, nor can individuals or groups of individuals be compared). Nevertheless, by making combined use of RR and direct answering in the sample, information that is both more valid and more reliable can be obtained.

Acknowledgements

This study was partially supported by Ministerio de Economía y Competitividad (grant MTM2015-63609-R and FPU grant program, Spain).

References

[1] R. Arnab, *Randomized response trials: A unified approach for qualitative data*, Communications in Statistics - Theory and Methods, **25** (1996) 1173-1183.

[2] R. Arnab, *Optional randomized response techniques for complex survey designs*, Biometrical Journal, **46** (2004) 114-124.

[3] A. Chaudhuri and T. Christofides, *Indirect Questioning in Sample Surveys*, Springer. 2013.

[4] A. Chaudhuri, T.C. Christofides and C.R. Rao, *Data Gathering, Analysis and Protection of Privacy Through Randomized Response Techniques: Qualitative and Quantitative Human Traits*, Handbook of Statistics, vol. 34. Elsevier. 2016.

[5] G. Diana and P.F. Perri, *A calibration-based approach to sensitive data: A simulation study*, Journal of Applied Statistics, **39** (2012) 3-65.

[6] J.C. Deville and C.E. Särndal, *Calibration estimators in survey sampling*, Journal of the American Statistical Association, **87** (1992) 376-382.

 ISBN: 978-84-617-8694-7

[7] W.A. FULLER AND L.F. BURMEISTER, *Estimators for samples selected from two overlapping frames*, Proceedings of the American Statistical Association, Social Statistics Section, (1972) 245-249.

[8] H.O. HARTLEY, *Multiple frame surveys*, Proceedings of the American Statistical Association, Social Statistics Sections, (1962) 203-206.

[9] H.O. HARTLEY, *Multiple frame methodology and selected applications*, Sankhya, Series C, **36** (1974) 99-118.

[10] G. KALTON AND D.W. ANDERSON, *Sampling rare populations*, Journal of the Royal Statistical Society, Series A, **149** (1986) 65-82.

[11] S. LOHR, *Multiple Frame Surveys*, In: Rao, C.R., Pfeffermann, D. (eds.) Handbook of Statistics, Vol. 29A, Sample Surveys: Design, Methods and Applications, pp. 71-88. North Holland, Amsterdam. 2009.

[12] S. LOHR, *Dual frame surveys: recent developments and challenges*, 45th Scientific Meeting of the Italian Statistical Society, Padua, Italy, June (2010) 16-18.

[13] S. LOHR AND J.N.K. RAO, *Estimation in multiple-frame surveys*, Journal of the American Statistical Association, **101** (2006) 1019-1030.

[14] F. MECATTI, *A single frame multiplicity estimator for multiple frame surveys*, Survey Methodology , **33** (2007) 151-157.

[15] M. QUENOUILLE, *Approximation tests of correlation in time series*, Journal of the Royal Statistical Society, Series B, **11** (1949) 18-84.

[16] M.G. RANALLI, A. ARCOS, M. RUEDA AND A. TEODORO, *Calibration estimation in dual-frame surveys*, Statistical Methods & Applications, **25** (2016) 321-349.

[17] J.N.K. RAO AND C. WU, *Pseudo empirical likelihood inference for multiple frame surveys*, Journal of the American Statistical Association, **105** (2010) 1494-1503.

[18] M. RUEDA, A. ARCOS AND B. COBO, *Use of randomized response techniques when data are obtained from two frames*, Applied Mathematics & Information Sciences, **9** (2015) 389-399.

[19] M. RUEDA, A. ARCOS, D. MOLINA AND M.G. RANALLI *Estimation techniques for discrete response variables in multiple frame surveys with complex sampling designs*, International Statistical Review. In press (2017).

[20] C.E. SÄRNDAL, B.SWENSON, J.WRETMAN, *Model Assisted Survey Sampling*. New York: Springer-Verlag. 1992.

 ISBN: 978-84-617-8694-7

[21] A.C. Singh and F. Mecatti, *Generalized multiplicity-adjusted Horvitz-Thompson estimation as a unified approach to multiple frame surveys*. Journal of Official Statistics, **27** (2011) 1-19.

[22] T.S. Tracy and S. Singh, *Calibration estimators in randomized response surveys* Metron **LVII** (1999) 47-68.

[23] S.L. Warner *Randomized response: A survey technique for eliminating evasive answer bias*, Journal of the American Statistical Association, **60** (1965) 63-69.

[24] K.M. Wolter, *Introduction to Variance Estimation*, 2nd Edition, Springer. 2007.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

Deep Learning for Variable-Length Handwritten Word Prediction

Victoria Ruiz[1], Jorge Sueiras[1], Angel Sanchez[1] and Jose F. Velez[1]

[1] *Department of Computer Sciences, Universidad Rey Juan Carlos*

emails: victoria.ruiz.parrado@urjc.es, jorge.sueiras@urjc.es, angel.sanchez@urjc.es, jose.velez@urjc.es

Abstract

Nowadays, Deep Learning is one of the most popular techniques which is used in several fields like handwriting text recognition. This paper presents the first steps for our propose for a handwritten text recognition system. Our system is based in Convolutional Neural Networks and it is trained using an on-demand scheme to recognize the existing characters from the IAM dataset. We show that using these training samples, it is not necessary segment or normalize the input images. Average accuracy recognition results were 85.5% for characters and 35.1% for words, respectively. The accuracy distribution depends on the lengths of the sequences. Moreover, in more than 48.8% of wrongly predicted numbers there was only one or two character errors.

Key words: Deep Learning, Handwriting Recognition, Convolutional Neural Network.

1 Introduction

"Aoccdrnig to a rscheearch at Cmabrigde Uinervtisy, it deosn't mttaer in waht oredr the ltteers in a wrod are, the olny iprmoetnt tihng is taht the frist and lsat ltteer be at the rghit pclae. The rset can be a toatl mses and you can sitll raed it wouthit porbelm. Tihs is bcuseae the huamn mnid deos not raed ervey lteter by istlef, but the wrod as a wlohe." Or rather, "According to a researcher (sic) at Cambridge University, it doesn't matter in what order the letters in a word are, the only important thing is that the first and last letter be at the right place. The rest can be a total mess and you can still read it without problem. This is because the human mind does not read every letter by itself but the word as a whole". Although this sentence have been circulated on the Internet since 2003, there are a

 ISBN: 978-84-617-8694-7

few studies that support this idea (or at least a part of it) ([13], [14]). The goal of this paper is to build the first steps of a handwriting recognition system which reads one word as a whole and recognizes it. In this system, the order of the letters are not relevant at all as the previous sentence stated. That means that, differently from most of the handwritten text recognition systems, we will not sequence the input images into their component characters using a sliding window. The first step of or recognition system is identifying the different characters being part of a word, and this task is in what we focus our paper.

We will work with handwriting text images. This can be done with the off-line approach or with the on-line approach, where in the last case the system uses dynamic information captured while someone is writing. These two approaches are the main ones for the automatic handwriting recognition process [12]. On the other hand, reading the word as a whole corresponds to the holistic approach, while in the segmented-based approach isolated characters are extracted to perform their classification ([18], [9]).

The segmentation phase has been one of the most important stages in the recognition systems. Many techniques aim to predict isolated characters and concatenate them, instead of predicting the whole words. A good segmentation reduces the recognition errors, so that many authors have focused their research in it [9]. Most recent techniques handle the whole word without segmentation [2], but sequence modeling with a sliding window is necessary in this case. Moreover, it is a frequent practice to normalize the text images to reduce their variability in order to achieve better recognition results.

Despite of the fact that there are many improvements in these recognition systems (especially with the appearence of Deep Learning techniques), there are still several problems to be solved. The possible high length-variability of text words in senteces, and the image noise presence must also be taken into account. In particular, this length variability in words may difficult the character segmentation task, and the presence of noise can cause that the system wrongly isolates some characters. Other approaches are non segmented-based, but instead, these use a sliding window to sequence the characters. Other non segmented-based systems, as [4], [5] and [6], use different techniques like Hidden Markov Models (HMM), Neural Netwoks(NN), Hybrid NN/HMM and Convolutional Networks (CNN) as classifiers. Moreover, to our knowledege, there are not techniques which handle the image as a whole without performing segmentation or using sliding windows.

The presence of high text variability and of noise can be reduced by using normalization techniques such as skew correction, baseline extraction, slant normalization, size normalization and contour smoothing [1]. In our system, instead of normalizing the images, we apply some types of random transformations to increase the sample size as in [16]. With this method we generate a "more realistic and almost infinite sample".

On one hand, for the above referred normalization-based methods, after that step a segmentation algorithm is applied in order to get isolated characters. One of the most used ones to extract the word characters consists on splitting the words into connected

 ISBN: 978-84-617-8694-7

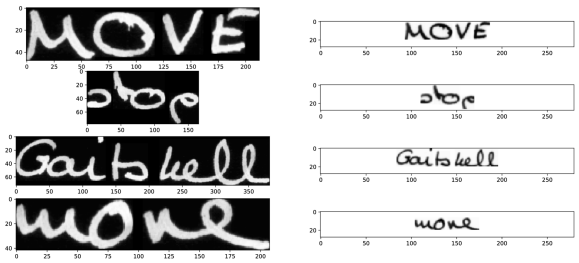


Figure 1: Examples IAM normalization

components which are classified as characters [9]. On the other hand, as our method does neither need from normalization nor segmentation, the text and noise variability conditions will be higher and this makes it more difficult the recognition task itself as it happens in many practical situations.

In this paper, we describe the first steps of a new deep learning approach for handwritten word recognition. Our method does not need from any normalization and segmentation or sliding windows stages, and it has been successfully tested on IAM dataset. The proposed system learns from an on-demand basis for generating infinite samples from IAM database, instead of using a standard fixed-length database.

The paper is organized as follows. Section 2 describes the generation of our experimental training examples from IAM dataset. Section 3 outlines the system architecture for the considered problem. Our experimental results are presented and explained in Section 4. Finally, the last Section summarizes the conclusions of this work.

2 Handwritten character database

We use IAM database to train and test our system [10]. This dataset is formed by 115,320 images, from which 97,704 are used of training and 17,616 are of tests, respectively. All of these samples are black and white images, centered and non size-normalized. For training and testing our system, we normalized the size of the images to 28×280 pixels. The relative position between characters is important to differentiate them. For example, the characters "l" and "e" sometimes just differ in their height. For that reason, we have extracted the upper and lower baselines of each word, and centered and reescaled them in order to fit to the new normalized baseline. Figure 1 shows some examples of IAM normalized images. Finally, the infinite sample was the result of moving and transforming the normalized images.

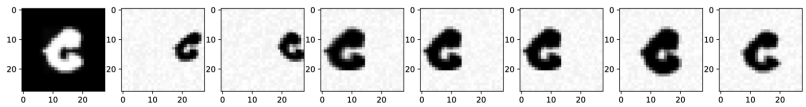


Figure 2: Example of the image transformations: (a) original image, (b) affine transformation to the right, (c) affine transformation to the left, (d), (e) and (f) respective transforamtions with increase of 15% of the heigth, width, and both, (g) dilated image, and (h) eroded one.

2.1 Transformations applied to images

In order to add more variability and get an infinite sample, some transformations are applied to the generated images. The generator can choose whether or not the following transformations are applied (see Figure 2):

1. Affine transformation to the right or to the left in order to rotate the image.
2. Increase the scale of the digfits by width or by height.
3. Morphological erosion.
4. Morphological dilation.
5. Translate the position of letters.

At the end of these transformations, a random background (with pixels between 240-255 graylevel values) is added to the image in order to avoid the noise which is produced when a transformation is applied. Note that it is possible to use more than one transformation in the same generated sequence. By the other hand, in the *Move* transformation, the sequence will be displaced to the right randomdly or using a fixed scalar.

3 Architecture proposal

The proposed model is based on the VGG architecture [17], and its goal is to predict the existing characters in an text image. This architecture is composed by stacks of convolutional layers which are followed by dense layers. The convolutional layers extract the letters features while the dense layers summarize that information to identify which characters are present in the image. Note that, as the convolution functions are invariant to rotations and translations, the positions and frequencies of the letters are not important at all. On the other hand, the target of that model is a binary vector

 ISBN: 978-84-617-8694-7

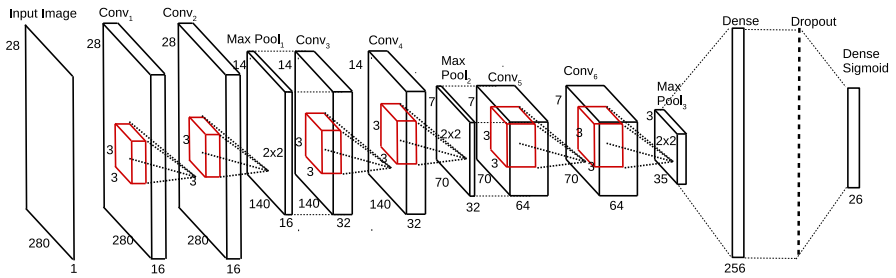


Figure 3: CNN Model Architecture.

$\mathbf{y} \in M^{26}$ where $\mathbf{y} = (y_0, \cdots y_{25})$ $/y_i = 1$ if class i is in the image. Uppercase and lowercase letters have been considered as representing the same label or class. Figure 3 shows the general model architecture of our system.

As it is shown, this architecture is composed by three blocks of convolutional layers stacked. In each stack, the kernel sizes are increased and the layers are followed by a sub-sampling (in particular, max-pooling). The last two layers are dense ones. In the first dense layer, a dropout of 60% was used as regularization method [15]. All the convolutions layers have a stride of 3×3 (which is equivalent to have a stride of 5×5 due to there are two consecutive convolution layers) and zero-padding. Also, all the neurons use ReLU [7] as activation function.

As pointed out before, we have used the real words from the IAM database to train our model. Real words are more challenging to train than isolated characters due to, in most times, there is not space between letters and some characters are deformed. Next section explains the results obtained by this recognition system.

4 Experiments

For training our model we used a Stochastic Gradient Descent (SGD) as optimizator. This type of optimizator works with batches of training examples, by decreasing the computation cost with a fast convervenge [3]. Moreover, we use a learning rate of 0.1. On the other hand, the model was trained during 500 epochs.

To measure the model goodness, we define the Character Accuracy Rate (CAR) and the Sequence Accuracy Rate (SAR), which measure how many correct predicted letters there are in all the tested words, and how many tested words have all their letters well predicted respectively:

$$CAR = \frac{\text{N}^{\text{o}} \text{ of correct test characters}}{\text{Total Number of test characters}}$$

 ISBN: 978-84-617-8694-7

	CAR	SAR
IAM examples	85.5%	35.1%

Table 1: Summary of recognition results of our system for isolated letters (CAR column) and variable-length numbers (SAR column)

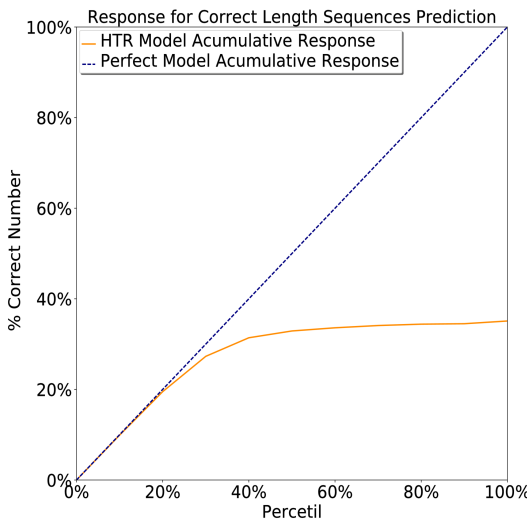


Figure 4: Accuracy rate for proposed model vs perfect model

$$SAR = \frac{\text{N}^{\text{o}} \text{ test sequence where all their characters are correct}}{\text{Total Number of test sequence}}$$

We tested these measures with 1000 test examples. As it is shown on Table 1, we have achieved an 85.5% of CAR and a 35.1% of SAR, respectively:

Figure 4 shows the accuracy rate by percentils of our model and compares it with the perfect model, i.e. when all the examples are well predicted). This graph have been built by sorting the predictions by their confidence. Note that these confidences have been calculated by summing the probabilties of each predicted letters. It is noticed that, if the 10% of the predictions with highest confidence is chosen, all of the tested examples are well predicted.

Figure 5 (a) shows the sequences length distribution for the correct predicted sequences. It's observed that there are more accuracy in words with lenght 2 or 3. Note that these lengths are the most frequent ones in the lengths distribution, as it is shown in Figure 5

 ISBN: 978-84-617-8694-7

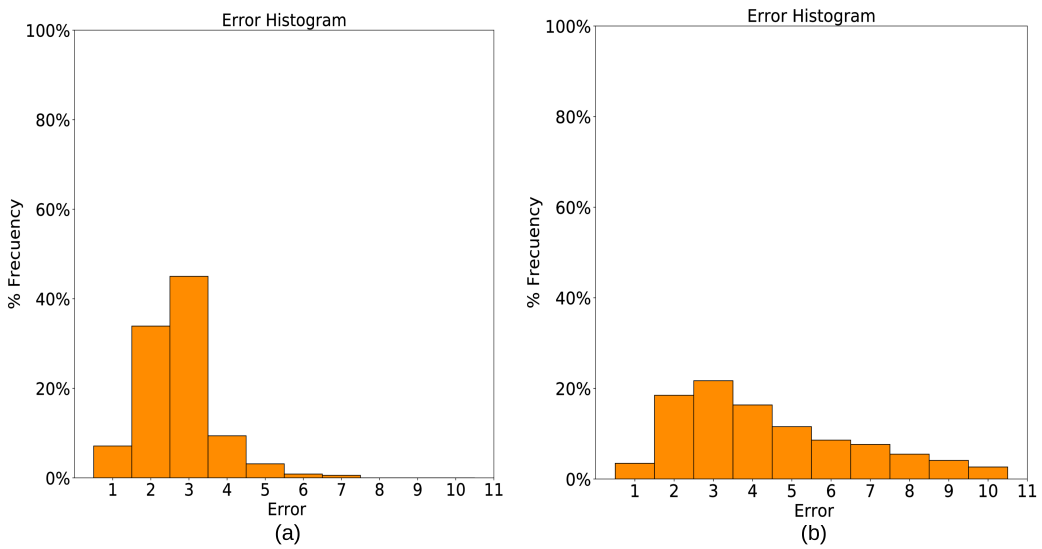


Figure 5: Recognition results: (a)frequency of well classified IAM test sequences by their number length and (b) frequency of IAM test words by their number length

(b).

Finally, we have also studied how much the system predicted wrongly by using the edit distribution [8] and the correct sequences' length distributions. Figure 6 shows that 35.1% of the tested words have 0 errors, 31.7% have 1 or 2 errors, and 33.2% have more than 2 errors. That means that the 48.8% of the wrong predictions have 1 or 2 errors.

5 Conclusions

This paper described the first stages of a new automatic system for handwriting recognition. This system aims to predict words as a whole without using neither segmentation nor sliding windows. This paper mainly focused on the task of identifying the existing letters in the handwritten word images. Our system was trained with an on-demand scheme using IAM database. As a result, we achieved a 85.5% in Character Accuracy Rate and 35.1% of Sequence Accuracy Rate. Moroever, the accuracy is higher in real words of length between 2 or 3. Finally, the 48.8% of the wrong predictions only have 1 or 2 errors. In the future, we will improve this model in order to predict complete handwritten words, and not just their component letters.

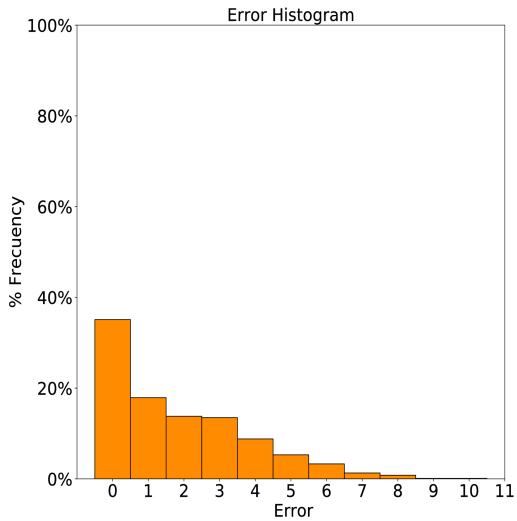


Figure 6: Frequency of tested numbers by their edit distance for IAM test words

Acknowledgements

This work was funded by the Spanish Ministry of Economy and Competitiveness under grant number TIN2014-57458-R.

References

[1] Arica,N. and Yarman-Vural,F.T. , *An Overview of Character Recognition Focused on Off-Line Handwriting*, Journal Systems, MAN, and CYBERNECTICS **31(2)** (2011)

[2] Bluche, T.;Ney,H.;Kermorvant,C. , *Feature extraction with convolutional neural networks for handwritten word recognition*, In 12th International Conference on Document Analysis and Recognition (2013)

[3] Bottou,L., *Large-Scale Machine Learning with Stochastic Gradient Descent*, In: COMPSTAT 2010 (2010)

[4] Ciresan,D.; Meier,U. and Schmidhuber,J. , *Multi-column Deep Neural Networks for Image Classification*, CVPR 2012, (2012) 3642–3649

[5] Graves,A and Schmidhuber,J., *Offline Handwriting Recognition with Multidimensional Recurrent Neural Networks*, In Advances in neural information processing systems (NIPS) **21** (2009) 545–552.

 ISBN: 978-84-617-8694-7

[6] Koerich,A.L;Sabourin,R. and Suen,C.Y., *Large Vocavulary off-line handwriting recognition: A survey*, Pattern Analysis and Applications **6(2)** (2003) 97-121.

[7] Krizhesvky,A.,Sutskever,I.,Hinton G.E. , *ImageNet Classification with Deep Convolutional Neural Networks.*, In 26th Advances in Neural Information Processing Systems (NIPS) (2012) 1097–1105.

[8] Levenshtein,V.I., *Binary codes capable of correcting delections, inserctions and reversals.*, Soviet physics doklady **10** (1966) 707

[9] Lu, Yi and Sridhar,M., *Character Segmentation in Handwritten words-An overview*, Pattern Recognition **29** (1995) 77–96

[10] Marti,U.-V and Bunke, H, *The IAM-database: an English sentence database for offfline handwriting recognitio*, International Journal on Document Analysis and Recognition **5(1)** (2002) 39–46

[11] Patel,M and Thakkar,S.P., *Handwritten Character Recognition in English: A Survey.*, International Journal of Advanced Research in Computer and Communication Engineering **4(2)** (2015)

[12] Plamondon,R. and Srihari,S. N. , *On-line and Off-line Handwriting Recognition: A Comprehensive Survey.*, Patter Analysis and Machine Intelligence, IEEE Transactions on **22(1)** (2000) 63–84.

[13] Rayner, K., White, S.J., Johnson, R.L., and Liversedge, S.P., *Raeding wrods with jubmled lettres: There is a cost*, Psychological Science **17** (2006) 192–193

[14] Rawlinson,G, *The Significance of Letter Position in Word Recognition*, PhD Thesis, Nottingham University

[15] Srivastava,N., Hinton,G., Krizhesvsky,A., Sutskever,I., Salakhutdinov,R, *Dropout: a simple way to prevent neural networks from overfitting*, J. Machine Learning Res. **15** (2014) 1929–1958

[16] Sueiras,J., Ruiz.V., Sanchez,A., Velez, J. , *Using a Synthetic Character Database for Training Deep Learning Models Applied to Offline Handwritten Recognition*, In: 16th ISDA (2017)

[17] Symonian,K. and Zisserman,A., *Very Deep Convolutional Networks for Large Scale Image Recognition*, In 3rd International Conference on Learning Representation (ICLR)

[18] Vinayakumar, R. and Paul, V., *A survey on Recognition and Analysis of Handwrittten Document*, IJCSET 2016 (2016)

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

Deep Learning for Digit Sequence Length Estimation

Victoria Ruiz[1], Jorge Sueiras[1], Angel Sanchez[1] and Jose F. Velez[1]

[1] *Department of Computer Sciences, Universidad Rey Juan Carlos*

emails: victoria.ruiz@urjc.es, jorge.sueiras@urjc.es, angel.sanchez@urjc.es, jose.velez@urjc.es

Abstract

In this short paper we show a deep learning architecture for recognizing the handwritten numbers which are presented to it, as an off-line scanned continuous digit sequence. The presented architecture is composed by a Convolutional Neural Network, a Long Short-Term Memory and a Multilayer Perceptron. The inputs to this system consist in digit sequences that have been previously generated using the MNIST database. This paper also describes the data augmentation performed for training the proposed architecture. The obtained results, which are combined with the features extracted by another Convolutional Neural Network, could be used to predict the numbers presented to the system.

Key words: Deep Learning, Handwriting Recognition, Convolutional Neural Network, Long Short-Term Memory.

1 Introduction

Off-line continuous handwritten recognition continues being an open research area. There are no commercial or free system which allows handwritten recognition with comparable accuracy to a human being. However, recent deep learning approaches are making important progresses for this task[1][3].

In the literature, we can see that there are two classical ways to recognize a word [4]. The first one consists in splitting the word in their component characters. This means that some type of segmentation process is needed. The method does not need from a dictionary, and so it can predict words that the system has not seen before. The second way consists in learning each word as a whole. In this case, the system usually needs to be trained with many examples of each possible word. Our group is working with an intermediate approach.

 ISBN: 978-84-617-8694-7

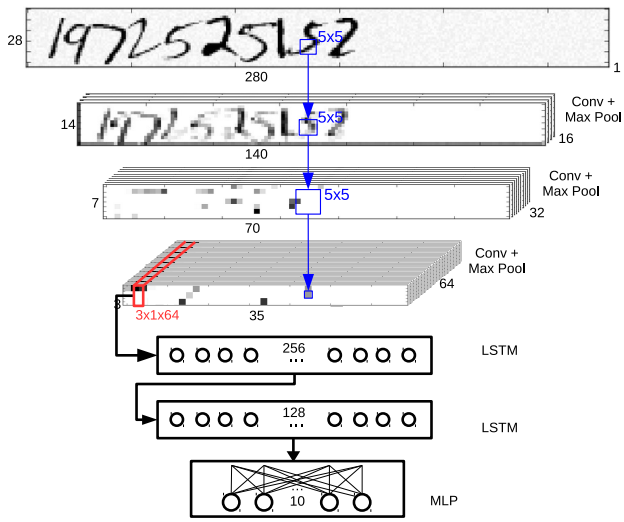


Figure 1: Proposed architecture for number recognition.

We recognize each word as a whole, but we do not need one example word of each type. That is because our recognition method is based in some features such as the components characters and the number of characters in the word. However, unlike segmentation-based systems, our proposal does not need to know exactly where a character ends and when it begins the next one.

In this short paper we present a part of our word recognition system, which is able ti determine the number of characters contained in a word. This part is based in a deep learning architecture that combines a Convolutional Neural Network (CNN), a Long Short Term Memory Network (LSTM) and a Multilayer Perceptron (MLP).

2 Architecture proposal

Our proposal consists in a CNN followed by a LSTM network and a MLP (see Figure 1). The CNN is devoted to extract the features that characterize a given digit sequence. The LSTM, whose input is the output of the last convolutional layer in the CNN, is targeted to obtain the length of the digit sequence. Finally, after the LSTM, it appears a fully-connected MLP with 10 output units corresponding to each possible digit.

 ISBN: 978-84-617-8694-7

	Number of errors		
	0	1	2
Number of cases	868	131	1
Percentage(%)	86.8	15	0.1

Table 1: Results obtained when the system predicts the length of the test digit sequences.

3 Experiments

This network is trained in two stages. In a first stage, the CNN is trained to predict the digits presented at its input. In this first stage, a MLP was concatenated to the CNN. When this network was trained, then the MLP layers were removed, and the LSTM was disposed after the CNN. Finally, a MLP was placed after the LSTM. When this architecture was completed, it was trained as a whole.

The network was trained using the MNIST handwritten digit database [2]. At each training step, a random digit sequence was generated from the training set. In addition skew and dilate transformations were applied to thos sequence in order to increase its variability.

To test our system, random digit sequences were generated from the test set. Table 1 (a) shows that the length of each sequence is predicted correctly in the 86.8% of the cases, and the error is only of one unit in the rest of the cases.

4 Conclusions

In this paper a deep learning model for length digit sequence estimation has been proposed. As future work, we propose to train this model with handwritten continuous text. Then, it will be used to decide the number of time steps of a recurrent model, trained to recognize a continuous handwritten word.

Acknowledgements

This work was funded by the Spanish Ministry of Economy and Competitiveness under grant number TIN2014-57458-R.

References

[1] Bluche, T. and Ney, H. and Kermorvant, C., *Feature extraction with convolutional neural networks for handwritten word recognition*, Int. Conf. on Doc. An. and Recog. (2013) 285–289.

[2] Deng,L. , *The MNIST Database of Handwritten Digit Images for Machine Learning Research*, IEEE Signal Processing Magazine (2012) 141–142.

[3] Graves,A and Schmidhuber,J., *Offline Handwriting Recognition with Multidimensional Recurrent Neural Networks*, Adv. in Neural Info. proc. Sys. **21** (2009) 545–552.

[4] Patel,M and Thakkar,S.P., *Handwritten Character Recognition in English: A Survey*, Int. Journal of Adv. Research in Computer and Comm. Eng. **4(2)** (2015) 345–350.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

Existence of solutions for a nonlinear simply supported beam equation

Lorena Saavedra[1]

[1] *Instituto de Matemáticas, Universidade de Santiago de Compostela*

emails: lorena.saavedra@usc.es

Abstract

This paper is devoted to prove the existence of one or multiple solutions for a family of nonlinear fourth order boundary value problems.

We use several fixed point theorems, previously developed by the authors, to prove the existence of solutions of some simply supported beam problems.

To finish the work, a particular case is studied and the existence of multiple solutions is proved for two different particular nonlinear functions.

Key words: Green's functions, fixed point theorems, nonlinear boundary value problems

MSC 2000: 34B15, 34B27, 34B18, 47H10

1 Introduction

The aim of this paper is to ensure the existence of one or multiple positive solutions for the fourth order problem:

$$\begin{cases} T_4[M]\,u(t) = f(t,u(t))\,, & t \in I \equiv [a,b]\,, \\ u(a) = u''(a) = u(b) = u''(b) = 0\,, \end{cases} \tag{1}$$

where $T_4[M]\,u(t) = u^{(4)}(t) + p_1(t)\,u^{(3)}(t) + p_2(t)\,u''(t) + M\,u(t)$, $p_1 \in C^3(I)$ and $p_2 \in C^2(I)$.

The simply supported beam problem (1) has been widely developed in the literature. For instance, in [2, 6, 7, 8], it is studied the existence of unique constant sign solutions of the following linear problem

$$\begin{cases} u^{(4)}(t) + c(t)\,u(t) = f(t)\,, & t \in [0,1]\,, \\ u(0) = u''(0) = u(1) = u''(1) = 0\,, \end{cases}$$

 ISBN: 978-84-617-8694-7

where $c,\ f \in C([0,1])$.

In [5], there are proved some results which ensure the existence of one or multiple solutions for a n^{th}-order nonlinear differential equation coupled with the so-called $(k, n-k)$ boundary conditions, which are defined by:

$$u(a) = u'(a) = \cdots = u^{(k-1)}(a) = 0\,, \quad u(b) = u'(b) = \cdots = u^{(n-k-1)}(b) = 0\,.$$

This existence of solution needs as hypothesis that the related Green's function, $g_M(t,s)$, satisfies the following suitable property:

(Pg_1) There exist Φ, k_1 and k_2 continuous functions on I such that $\Phi(s) > 0$ for all $s \in (a,b)$, $0 < k_1(t) \leqslant k_2(t)$ for all $t \in (a,b)$ and

$$\Phi(s)\, k_1(t) \leqslant g_M(t,s) \leqslant \Phi(s)\, k_2(t)\,, \quad \forall (t,s) \in I \times I\,.$$

From the constant sign Green's function characterization related to the operator $T_4[M]$, previously defined, in the set

$$X = \left\{ u \in C^4(I) \ \mid \ u(a) = u''(a) = u(b) = u''(b) \right\}, \tag{2}$$

obtained in [4] by means of spectral theory, it can be characterized the parameter set for which the related Green's function of $T_4[M]$ in X verifies either the property (Pg_1) or the analogous for the negative case:

(Ng_1) There exist Φ, k_1 and k_2 continuous functions on I such that $\Phi(s) > 0$ for all $s \in (a,b)$, $0 > k_1(t) \geqslant k_2(t)$ for all $t \in (a,b)$ and

$$\Phi(s)\, k_1(t) \geqslant g_M(t,s) \geqslant \Phi(s)\, k_2(t)\,, \quad \forall (t,s) \in I \times I\,.$$

Remark 1.1. *It is clear that if the Green's function related to problem* (1) *satisfies the property* (Ng_1)*, then the related Green's function to problem*

$$\begin{cases} T_4[M]\, u(t) + f(t, u(t)) = 0\,, & t \in I \equiv [a,b]\,, \\ u(a) = u''(a) = u(b) = u''(b) = 0\,, \end{cases}$$

satisfies the property (Pg_1)*, just by taking* $\tilde{k}_1(t) = -k_2(t)$ *and* $\tilde{k}_2(t) = -k_1(t)$.

Thus, in order to obtain the existence results from the ones proved in [5], we need to impose conditions on $-f(t,u)$ *instead of* $f(t,u)$.

In a preliminary section we transform our problem into a fixed point problem. Then, we show the results collected in [5] which ensure the existence of one or multiple fixed points under suitable hypotheses. To finish this section, it is characterized the parameter set for which the related function of problem (1) satisfies either the property (Pg_1) or (Ng_1).

In the main section of this paper, it is proved the existence of solution for a particular simply supported beam problem by applying the previous results.

 ISBN: 978-84-617-8694-7

2 Fixed point formulation and preliminary results

First, let us denote $g_M(t,s)$, as the related Green's function of problem (1).

It is well-known that the solutions of (1) are determined by the fixed points of the following integral operator:

$$\mathcal{L}[M]\,u(t) = \int_a^b g_M(t,s)\,f(s,u(s))\,ds\,. \tag{3}$$

Let us consider the Banach space $\mathcal{B} = C(I)$ coupled with the norm:

$$\|u\|_\infty = \max_{t\in I} |u(t)|\,.$$

Let us assume that $g_M(t,s)$ satisfies the property (Pg_1). Consider the subinterval $I_1 = [a_1, b_1] \subset I$, such that $k_1(t) > 0$ for all $t \in I_1$ and we denote:

$$K_1 = \|k_1\|_\infty\,, \quad K_2 = \|k_2\|_\infty \ \text{ and } \ m_1 = \min_{t\in I_1} k_1(t)\,. \tag{4}$$

In the sequel, we enunciate the results, collected in [5], where some sufficient conditions to ensure the existence of one or multiple fixed points of $\mathcal{L}$ are shown.

Let us consider the following conditions on f:

(H_1) $\exists\, p > 0$ such that $f(t,u) \leqslant \dfrac{p}{K_2 \int_a^b \Phi(s)\,ds}\,, \quad \forall\, t \in I\,,\ \forall\, u \in [0,p]\,.$

(H_2) $\exists\, q > 0$ such that $f(t,u) \geqslant \dfrac{K_2\, u}{K_1 \int_{a_1}^{b_1} k_1(s)\,\Phi(s)\,ds}\,, \quad \forall\, t \in I_1\,,\ \forall\, u \in \left[\frac{m_1}{K_2}\,q, q\right].$

Then, we have the following result.

Theorem 2.1. *Suppose that there exist two positive numbers $p \neq q$ such that condition (H_1) is satisfied with respect to p and condition (H_2) is satisfied with respect to q. Then, provided that the integral kernel g_M satisfies (Pg_1), operator $\mathcal{L}$, defined in (3), has a fixed point, $u \in \mathcal{B}$, such that $\|u\|_\infty$ lies between p and q.*

Now, for the existence of at least two fixed points, we have the next theorem.

Theorem 2.2. *Suppose that there exist positive numbers p, q and r such that $0 < p < q < r$, and assume that function f satisfies the following conditions:*

(i) $f(t,u) \geqslant \dfrac{u}{m_1 \int_{a_1}^{b_1} \Phi(s)\,ds}\ \forall\, t \in I_1$ *and* $u \in \left[r, \frac{K_2}{m_1}\,r\right]$, *being the inequality strict at* $u = r$,

(ii) $f(t,u) \leqslant \dfrac{q}{K_2 \int_a^b \Phi(s)\,ds}\ \forall\, t \in I$ *and* $u \in \left[0, \frac{K_2}{m_1}\,q\right]$, *being the inequality strict at* $u = q$,

 ISBN: 978-84-617-8694-7

(*iii*) $f(t,u) > \dfrac{K_2\, u}{K_1 \int_{a_1}^{b_1} k_1(s)\, \Phi(s)\, ds}$ *for all* $t \in I_1$ *and* $u \in \left[\dfrac{m_1}{K_2}\, p, p\right]$.

Then, if g_M *satisfies* (Pg_1)*, the operator* $\mathcal{L}$ *has at least two fixed points,* u_1 *and* u_2*, such that* $p < \|u_1\|_\infty\,, \quad \max_{t\in I_1} u_1(t) < q < \max_{t\in I_1} u_2(t)\,, \quad \min_{t\in I_1} u_2(t) < r\,.$

Finally, we have the next result which ensures the existence of, at least, three fixed points.

Theorem 2.3. *Let* p*,* q *and* r *be positive numbers satisfying* $0 < p < q < \dfrac{K_2}{m_1} q \leqslant r\,.$

Assume, moreover, that the function f *satisfies the following conditions:*

(*a*) $f(t,u) \leqslant \dfrac{r}{K_2 \int_a^b \Phi(s)\, ds}$ *for all* $t \in I$ *and* $u \in [0,r]$,

(*b*) $f(t,u) < \dfrac{p}{K_2 \int_a^b \Phi(s)\, ds}$ *for all* $t \in I$ *and* $u \in [0,p]$,

(*c*) $f(t,u) \geqslant \dfrac{u}{m_1 \int_{a_1}^{b_1} \Phi(s)\, ds}$ $\forall\, t \in I_1$ *and* $u \in \left[q, \frac{K_2}{m_1} q\right]$*, being the inequality strict for* $u = q$.

Then, if g_M *satisfies* (Pg_1)*, the operator* $\mathcal{L}$ *has at least three fixed points* u_1*,* u_2*,* u_3 *such that* $\|u_i\|_\infty \leqslant r$ *for* $i = 1,2,3$ *and* $\max_{t\in I_1} u_1(t) < p < \min_{t\in I_1} u_2(t)\,,\ \max_{t\in I_1} u_2(t) < q < \min_{t\in I_1} u_3(t)\,.$

Now, following the characterization of the constant sign Green's function for problem (1), given in [4], we obtain a characterization of the parameter set where g_M satisfies either the condition (Pg_1) or (Ng_1).

First, we need to introduce a concept about the maximum number of zeros for the related second order linear differential equation:

$$u''(t) + p_1(t)\, u'(t) + p_2(t)\, u(t) = 0\,, \quad t \in I\,, \tag{5}$$

collected in [3].

Definition 2.4. *The second order linear differential equation* (5) *is said to be disconjugate on* I *if every non trivial solution has, at most, a simple zero on* I.

Remark 2.5. *The second order equation* $u''(t) = 0\,,$ *is disconjugate on every real interval,* I.

Now, we have the following result.

Theorem 2.6. *If the second order linear differential equation* (5) *is disconjugate on* I*, then the following properties are fulfilled:*

 ISBN: 978-84-617-8694-7

- *$g_M(t,s)$, the related Green's function of problem* (1)*, satisfies condition* (Pg_1) *if, and only if,* $M \in (-\lambda_1, -\lambda_2]$*, where*
  - $\lambda_1 > 0$ *is the least positive eigenvalue of* $T_4[0]$ *in* X*.*
  - $\lambda_2 < 0$ *is the maximum between:*
    - $\lambda_2' < 0$*, the biggest negative eigenvalue of* $T_4[0]$ *in*
    $$X_1 = \{u \in C^4(I) \ \mid \ u(a) = u(b) = u'(b) = u''(b) = 0\}\,.$$
    - $\lambda_2'' < 0$*, the biggest negative eigenvalue of* $T_4[0]$ *in*
    $$X_3 = \{u \in C^4(I) \ \mid \ u(a) = u'(a) = u''(a) = u(b) = 0\}\,.$$
- *$g_M(t,s)$ satisfies condition* (Ng_1) *if, and only if,* $M \in [-\lambda_3, -\lambda_1)$*, where:*
  - $\lambda_3 > 0$ *is the minimum between:*
    - $\lambda_3' > 0$*, the least positive eigenvalue of* $T_4[0]$ *in*
    $$U_1 = \{u \in C^4(I) \ \mid \ u(a) = u''(a) = u(b) = u'(b) = 0\}\,.$$
    - $\lambda_3'' > 0$*, the least positive eigenvalue of* $T_4[0]$ *in*
    $$U_2 = \{u \in C^4(I) \ \mid \ u(a) = u'(a) = u(b) = u''(b) = 0\}\,.$$

Proof. From the proof of [4, Theorem 6.1], the following inequalities are fulfilled for $M \in (-\lambda_1, -\lambda_2]$.

$$\begin{aligned}
&(+\infty >) \qquad g_M(t,s) > 0\,, \quad \forall\,(t,s) \in (a,b)\times(a,b)\,,\\
&(+\infty >)\ \frac{\partial g_M(t,s)}{\partial t}_{|t=a} > 0\,, \quad \forall\, s \in (a,b)\,,\\
&(+\infty >)\ \frac{\partial g_M(t,s)}{\partial s}_{|s=a} > 0\,, \quad \forall\, t \in (a,b)\,,\\
&(-\infty <)\ \frac{\partial g_M(t,s)}{\partial t}_{|t=b} < 0\,, \quad \forall\, s \in (a,b)\,,\\
&(-\infty <)\ \frac{\partial g_M(t,s)}{\partial s}_{|s=b} < 0\,, \quad \forall\, t \in (a,b)\,.
\end{aligned}$$

Consider

$$\Phi(s) = s\,(1-s) > 0\,, \quad \forall\, s \in (a,b)\,, \tag{6}$$

and we have that

$$u(t,s) = \frac{g_M(t,s)}{\Phi(s)} > 0\,, \quad \forall\,(t,s) \in (a,b)\times(a,b)\,.$$

 ISBN: 978-84-617-8694-7

Moreover, we obtain the following real limits:

$$\lim_{s\to a^+}\frac{g_M(t,s)}{\Phi(s)}=\frac{\frac{\partial g_M(t,s)}{\partial s}_{|s=a}}{b-a}>0\,,\quad\forall\,t\in(a,b)\,,$$
$$\lim_{s\to b^-}\frac{g_M(t,s)}{\Phi(s)}=\frac{\frac{\partial g_M(t,s)}{\partial s}_{|s=b}}{-(b-a)}>0\,,\quad\forall\,t\in(a,b)\,.$$

So, we define $\tilde{u}(t,s)$, as the continuous extension of $u(t,s)$ to $(a,b)\times I$, and taking

$$\begin{aligned}k_1(t)&:=\min_{s\in I}\tilde{u}(t,s)>0\,,\quad\forall\,t\in(a,b)\,,\\ k_2(t)&:=\max_{s\in I}\tilde{u}(t,s)>0\,,\quad\forall\,t\in(a,b)\,,\end{aligned}\tag{7}$$

clearly, property (Pg_1) is fulfilled by g_M if $M\in(-\lambda_1,-\lambda_2]$.

On the other hand, from the proof of [4, Theorem 6.1], it is deduced that if $M\in[-\lambda_3,-\lambda_1)$, then:

$$\begin{aligned}(-\infty<)\qquad g_M(t,s)&<0\,,\quad\forall\,(t,s)\in(a,b)\times(a,b)\,,\\ (-\infty<)\,\frac{\partial g_M(t,s)}{\partial t}_{|t=a}&<0\,,\quad\forall\,s\in(a,b)\,,\\ (-\infty<)\,\frac{\partial g_M(t,s)}{\partial s}_{|s=a}&<0\,,\quad\forall\,t\in(a,b)\,,\\ (+\infty>)\,\frac{\partial g_M(t,s)}{\partial t}_{|t=b}&>0\,,\quad\forall\,s\in(a,b)\,,\\ (+\infty>)\,\frac{\partial g_M(t,s)}{\partial s}_{|s=b}&>0\,,\quad\forall\,t\in(a,b)\,.\end{aligned}$$

Thus, proceeding as in the positive case, we obtain that (Ng_1) is fulfilled by taking $\Phi(s)$ defined in (6) and

$$\begin{aligned}k_1(t)&:=\max_{s\in I}\tilde{u}(t,s)<0\,,\quad\forall\,t\in(a,b)\,,\\ k_2(t)&:=\min_{s\in I}\tilde{u}(t,s)<0\,,\quad\forall\,t\in(a,b)\,,\end{aligned}$$

Finally, in [4, Theorem 6.1], it is proved that if $M\notin[-\lambda_3,-\lambda_2]$, then $g_M(t,s)$ oscillates on the square $I\times I$. Thus, neither property (Pg_1) or (Ng_1) can be fulfilled. □

Remark 2.7. *Realize that from Theorems 2.1, 2.2 and 2.3, we prove the existence of non-negative solutions for problem* (1)*, under the hypothesis that $g_M(t,s)$ satisfies property (Pg_1).*

Proceeding analogously, by imposing property (Ng_1) on $g_M(t,s)$ instead of (Pg_1), we can prove the existence of non-positive solutions for problem (1) *by repeating the arguments done in [5], with the same assumptions on f, choosing $m_1=-\max_{t\in I_1}k_1(t)$ and K_1, K_2 as in* (4).

 ISBN: 978-84-617-8694-7

3 Existence results for particular simply supported beam problems

From Remark 2.5, Theorem 2.6 can be applied to the fourth order operator

$$T_4^0[M]\,u(t) = u^{(4)}(t) + M\,u(t)\,, \quad t \in I\,. \tag{8}$$

Let us choose $I = [0,1]$, and, in [4], there are obtained the different eigenvalues of operator $T_4^0[0] = \frac{d^4}{dt^4}$.

- $\lambda_1 = \pi^4$ is the least positive eigenvalue of $T_4^0[0]$ in X.
- $\lambda_2 = -m_1^4 \approxeq -5.55^4$, where m_1 is the least positive solution of $\tan\left(\frac{m}{\sqrt{2}}\right) = \tanh\left(\frac{m}{\sqrt{2}}\right)$, is the biggest negative eigenvalue of $T_4^0[0]$ in X_1 and X_3.
- $\lambda_3 = m_2^4 \approxeq 3.927^4$, with m_2 the least positive solution of $\tan(m) = \tanh(m)$, is the least positive eigenvalue of $T_4^0[0]$ in U_1 and U_2.

So, if we denote $g_M^0(t,s)$ as the related Green's function of problem (1), where $T_4[M] = T_4^0[M]$, from Theorem 2.6, we can conclude:

- $g_M^0(t,s)$ satisfies property (Pg_1) if, and only if, M belongs to the interval $(-\pi^4, 5.55^4]$.
- $g_M^0(t,s)$ satisfies property (Ng_1) if, and only if, M is in the interval $[-3.927, -\pi^4)$.

Remark 3.1. *For $M \in (-\lambda_1, -\lambda_2] \approxeq (-\pi^4, 5.55^4]$, since the property (Pg_1) is fulfilled by the related Greens function, we are able to obtain the continuous functions k_1 and k_2 associated to the problem*

$$\begin{cases} u^{(4)}(t) + M\,u(t) = f(t,u(t))\,, & t \in [0,1]\,, \\ u(0) = u''(0) = u(1) = u''(1) = 0\,. \end{cases}$$

The construction of these functions is described on the proof of Theorem 2.6. The difficulty remains in the expression of the related Green's function.

Sometimes, it is really difficult to find the exact functions k_1 and k_2 as they are defined in (7). However, we can consider some approximations, ℓ_1 and ℓ_2, such that $\ell_1(t) \leqslant k_1(t)$ and $\ell_2(t) \geqslant k_2(t)\,, \quad \forall\, t \in [0,1]\,.$

Clearly, the bounds obtained with ℓ_1 and ℓ_2 are still valid for Theorems 2.1, 2.2 and 2.3.

Remark 3.2. *Once that property (Pg_1) is proved, as a consequence of Theorem 2.1 the existence of at least one positive solution for problem (1), with $T_4[M] = T_4^0[M]$, can be proved*

 ISBN: 978-84-617-8694-7

if f is either sub-linear $\left(\lim_{u\to 0^+}\dfrac{f(t,u)}{u}=+\infty \text{ and } \lim_{u\to+\infty}\dfrac{f(t,u)}{u}=0\right)$ *or super-linear* $\left(\lim_{u\to 0^+}\dfrac{f(t,u)}{u}=0 \text{ and } \lim_{u\to+\infty}\dfrac{f(t,u)}{u}=+\infty\right)$ *without obtaining k_1 and k_2.*

Remark 3.3. *Using Remarks 1.1 and 2.7, from property (Ng_1) similar conclusions than those in the previous remarks can be obtained for $M\in[-\lambda_3,-\lambda_1)\approxeq[-3.927^4,-\pi^4)$.*

In the sequel we will obtain the different bounds and results for the particular case when $M=0$.

That is, we want to prove the existence of one or multiple positive solutions of the problem:

$$\begin{cases} u^{(4)}(t)=f(t,u(t)), & t\in[0,1], \\ u(0)=u''(0)=u(1)=u''(1)=0. \end{cases} \tag{9}$$

Φ is given in (6). Now, let us obtain the correspondent k_1 and k_2.

We have to calculate the related Green's function. By means of the Mathematica program developed in [1], we obtain:

$$g_0^0(t,s)=\begin{cases} \frac{1}{6}\,s\,(1-t)\left((2-t)\,t-s^2\right), & 0\leqslant s\leqslant t\leqslant 1, \\ \\ \frac{1}{6}\,t\,(1-s)\left((2-s)\,s-t^2\right), & 0<t<s\leqslant 1. \end{cases}$$

Thus, clearly, for this case

$$\tilde{u}(t,s)=\begin{cases} \frac{1}{6}\cdot\frac{1-t}{1-s}\left((2-t)\,t-s^2\right), & 0\leqslant s\leqslant t<1,\ t\neq 0 \\ \\ \frac{1}{6}\cdot\frac{t}{s}\left((2-s)\,s-t^2\right), & 0<t<s\leqslant 1. \end{cases}$$

So, we have

$$\frac{\partial\tilde{u}(t,s)}{\partial s}=\begin{cases} \frac{1-t}{6}\cdot\frac{(s-t)\,(s+t-2)}{(1-s)^2}, & 0\leqslant s\leqslant t<1,\ t\neq 0 \\ \\ \frac{t}{6}\cdot\frac{(t-s)\,(t+s)}{s^2}, & 0<t<s\leqslant 1. \end{cases}$$

Hence, for $s\in[0,t]$, $\tilde{u}(t,s)$ is non-decreasing as a function of s and, for $s\in[t,1]$, $\tilde{u}(t,s)$ is non-increasing as a function of s.

Therefore, we obtain $k_2(t)=\tilde{u}(t,t)=\dfrac{t\,(1-t)}{6}$, for every $t\in[0,1]$ and, we have, $k_1(t)=\min\{\tilde{u}(t,0),\tilde{u}(t,1)\}=\min\left\{\dfrac{1}{6}\,t\,(1-t)\,(2-t),\ \dfrac{1}{6}\,t\,(1-t^2)\right\}$, or, which is the same,

$$k_1(t)=\begin{cases} \frac{1}{6}\,t\,(1-t^2), & 0\leqslant t\leqslant\frac{1}{2}, \\ \\ \frac{1}{6}\,t\,(1-t)\,(2-t), & \frac{1}{2}<t\leqslant 1. \end{cases}$$

 ISBN: 978-84-617-8694-7

Directly, we obtain

$$K_1 = \|k_1\|_\infty = k_1\left(\frac{1}{2}\right) = \frac{1}{16} \quad \text{and} \quad K_2 = \|k_2\|_\infty = k_2\left(\frac{1}{2}\right) = \frac{1}{12}\,.$$

Moreover, since k_1 is a symmetric function with respect to $t = \dfrac{1}{2}$, we can choose

$$I_1 = [a_1, b_1] = \left[\frac{1}{2} - c, \frac{1}{2} + c\right], \quad c \in \left(0, \frac{1}{2}\right).$$

Clearly, $m_1 = k_1\left(\frac{1}{2} - c\right) = \frac{1}{48}\,\left(1 - 4c^2\right)\,(3 - 2c)\,.$

Now, let us obtain the bounds which appear in Theorems 2.1, 2.2 and 2.3, for the different values of $c \in \left(0, \frac{1}{2}\right)$.

Φ is defined in (6), so

$$\int_0^1 \Phi(s)\,ds = \frac{1}{6}, \quad \int_{\frac{1}{2}-c}^{\frac{1}{2}+c} \Phi(s)\,ds = \frac{1}{6}\,c\,\left(3 - 4\,c^2\right)\,,$$

and,

$$\int_{\frac{1}{2}-c}^{\frac{1}{2}+c} k_1(s)\,\Phi(s)\,ds = \frac{c\,(45 - 15\,c - 120\,c^2 + 60\,c^3 + 144\,c^4 - 80\,c^5)}{1444}\,.$$

Thus, we obtain:

$$\frac{p}{K_2\,\int_0^1 \Phi(s)\,ds} = 72\,p\,, \tag{10}$$

$$\frac{K_2\,u}{K_1\,\int_{\frac{1}{2}-c}^{\frac{1}{2}+c} k_1(s)\,\Phi(s)\,ds} = \frac{1083\,u}{c\,(45 - 15\,c - 120\,c^2 + 60\,c^3 + 144\,c^4 - 80\,c^5)}\,, \tag{11}$$

$$\frac{u}{m_1\,\int_{\frac{1}{2}-c}^{\frac{1}{2}+c} \Phi(s)\,ds} = \frac{288\,u}{c\,(3 - 4\,c^2)\,(3 - 2\,c)\,(1 - 4\,c^2)}\,. \tag{12}$$

Studying the behavior of (11) and (12) for $c \in \left(0, \frac{1}{2}\right)$, it can be easily seen that (11) is decreasing with c and (12) attains its minimum on the interval $\left[\frac{1}{5}, \frac{3}{10}\right]$.

We want to make this bounds the least possible, but we cannot minimize them together. We choose $c = \frac{3}{10}$, then $I_1 = \left[\frac{1}{5}, \frac{4}{5}\right]$ and we obtain the bounds (11) and (12) for this particular case:

$$\frac{K_2\,u}{K_1\,\int_{\frac{1}{5}}^{\frac{4}{5}} k_1(s)\,\Phi(s)\,ds} = \frac{1600000\,u}{8073}\,, \tag{13}$$

$$\frac{u}{m_1\,\int_{\frac{1}{5}}^{\frac{4}{5}} \Phi(s)\,ds} = \frac{15625\,u}{66}\,. \tag{14}$$

Thus, we can rewrite (H_1) and (H_2) for this case as follows.

 ISBN: 978-84-617-8694-7

(H_1) $\exists\, p>0$ such that $f(t,u)\leqslant 72\,p\,,\quad \forall\, t\in[0,1]\,,\ \forall\, u\in[0,p]\,.$

(H_2) $\exists\, q>0$ such that $f(t,u)\geqslant \dfrac{1600000\,u}{8073}\,,\quad \forall\, t\in\left[\frac{1}{5},\frac{4}{5}\right]\,,\ \forall\, u\in\left[\frac{48}{125}\,q,q\right].$

Using (10), (13) and (14), we rewrite Theorem 2.2 for this case.

Theorem 3.4. *Suppose that there exist positive numbers p, q and r such that $0<p<q<r$, and assume that function f satisfies the following conditions:*

(i) $f(t,u)\geqslant \dfrac{15625\,u}{66}$ $\forall\, t\in\left[\frac{1}{5},\frac{4}{5}\right]$ *and* $u\in\left[r,\frac{125}{48}\,r\right]$, *being the inequality strict at* $u=r$,

(ii) $f(t,u)\leqslant 72\,q$ $\forall\, t\in[0,1]$ *and* $u\in\left[0,\frac{125}{48}\,q\right]$, *being the inequality strict at* $u=q$,

(iii) $f(t,u)>\dfrac{1600000\,u}{8073}$ $\forall\, t\in\left[\frac{1}{5},\frac{4}{5}\right]$ *and* $u\in\left[\frac{48}{125}\,p,p\right]$.

Then, the problem (9) *has at least two positive solution, u_1 and u_2, such that*

$$p<\|u_1\|_\infty\,,\quad \max_{t\in I_1}u_1(t)<q<\max_{t\in I_1}u_2(t)\,,\quad \min_{t\in I_1}u_2(t)<r\,.$$

Now, consider the following continuous function

$$f(t,u)=\begin{cases}\dfrac{11390625}{512}\,t\,(1-t)\,u\,, & u\in\left[0,\frac{8}{225}\right],\\[2ex] \dfrac{t\,(1-t)}{u^2}\,, & u\in\left(\frac{8}{225},4\right),\\[2ex] t\,(1-t)\left(\dfrac{56607479}{15250000}u^3-\dfrac{3621925531}{244000000}u^2\right), & u\geqslant 4\,.\end{cases}\tag{15}$$

Let us choose $p=\dfrac{5}{54}$, $q=3$ and $r=\dfrac{111}{5}$. So, we have:

(i) For $u\leqslant r$, $f(t,u)\geqslant \dfrac{5\,f\left(\frac{1}{5},\frac{111}{5}\right)\,u}{111}=\dfrac{9149203310439\,u}{38125000000}\geqslant\dfrac{15625\,u}{66}$ for all $t\in\left[\dfrac{1}{5},\dfrac{4}{5}\right]$,

(ii)

$$\begin{aligned}f(t,u)&\leqslant \max\left\{\max_{t\in[0,1]}f\left(t,\frac{8}{225}\right),\max_{t\in[0,1]}f\left(t,\frac{125}{16}\right)\right\}=\max\left\{\frac{50625}{256},216\right\}\\&=216=3\cdot 72\,,\end{aligned}$$

for all $t\in[0,1]$ and $u\in\left[0,\dfrac{125}{16}\right]$, being the inequality strict for $u<\frac{125}{16}$. In particular, for $u=3$.

 ISBN: 978-84-617-8694-7

(iii) $f(t,u) = \dfrac{t\,(1-t)}{u^3}u \geqslant \dfrac{4}{25}\cdot\dfrac{157464}{125}\,u > \dfrac{1600000\,u}{8073}$ for all $t \in \left[\dfrac{1}{5},\dfrac{4}{5}\right]$ and $u \in \left[\dfrac{8}{225},\dfrac{5}{54}\right]$.

So, we can ensure the existence of at least two positive solutions for problem (9) with f defined in (15).

Finally, we have the next result which warrants the existence of three solutions by using (10) and (14).

Theorem 3.5. *Let p, q and r be positive numbers satisfying the relation:*

$$0 < p < q < \frac{125}{48}q \leqslant r\,.$$

Assume, moreover, that the function f satisfies the following conditions:

(a) $f(t,u) \leqslant 72\,r$ *for all* $t \in [0,1]$ *and* $u \in [0,r]$,

(b) $f(t,u) < 72\,p$ *for all* $t \in [0,1]$ *and* $u \in [0,p]$,

(c) $f(t,u) \geqslant \dfrac{15625\,u}{66}$ $\forall\, t \in \left[\frac{1}{5},\frac{4}{5}\right]$ *and* $u \in \left[q,\frac{125}{48}q\right]$, *being the inequality strict for* $u = q$.

Then, the problem (9) *has at least three solutions,* u_1, u_2, u_3 *such that* $\|u_i\|_\infty \leqslant r$ *for* $i = 1,2,3$ *and*

$$\max_{t\in I_1} u_1(t) < p < \max_{t\in I_1} u_2(t)\,, \quad \min_{t\in I_1} u_2(t) < q < \min_{t\in I_1} u_3(t)\,.$$

Consider the continuous function:

$$f(t,u) = \begin{cases} 1400\,t\,(1-t)\,u^2, & u \geqslant \frac{7}{2}\,, \\ \\ 17150\,t\,(1-t)\,, & u > \frac{7}{2}\,. \end{cases} \tag{16}$$

Let us choose $r = 60$, $p = \frac{1}{5}$ and $q = \frac{5}{4}$, we have:

(a) $f(t,u) \leqslant \dfrac{17150}{4} < 60\cdot 72$ for all $t \in [0,1]$ and $u \in [0,60]$,

(b) $f(t,u) \leqslant \dfrac{1400}{4}\cdot\dfrac{1}{5^2} = 14 < 72\cdot\dfrac{1}{5}$ for all $t \in [0,1]$ and $u \in \left[0,\frac{1}{5}\right]$,

(c) $f(t,u) \geqslant \dfrac{1400\cdot 4}{5^2}\cdot\dfrac{5}{4}\,u = 280\,u > \dfrac{15625\,u}{66}$ for all $t \in \left[\dfrac{1}{5},\dfrac{4}{5}\right]$ and $u \in \left[\dfrac{5}{4},\dfrac{625}{192}\right]$.

Thus, by using the previous result, we conclude that problem (9) has at least three solutions for f defined in (16).

 ISBN: 978-84-617-8694-7

Acknowledgements

This work has been supported by FPU scholarship, Ministerio de Educación, Cultura y Deporte, Spain.

This work has been partially supported by Ministerio de EconomÃa y Competitividad, Spain and FEDER, project MTM2013-43014-P.

This work has been partially supported by the Agencia Estatal de Investigación (AEI) of Spain under grant MTM2016-75140-P, co-financed by the European Community fund FEDER.

References

[1] A. Cabada, J.A. Cid, B. Máquez-Villamarín, *Computation of Green's functions for boundary value problems with Mathematica*, Applied Mathematics and Computation **219** (2012) 1919-1936.

[2] A. Cabada, J.A. Cid, L. Sanchez, *Positivity and lower and upper solutions for fourth order boundary value problems*, Nonlinear Anal. **67** (2007) 1599-1612.

[3] W. A. Coppel, *Disconjugacy*, Lecture Notes in Mathematics, Vol. 220. Springer-Verlag, Berlin-New York, 1971.

[4] A. Cabada, L. Saavedra, *Constant sign Green's function for simply supported beam equation*, Advances in differential equations, **22** (2017) 403-432

[5] A. Cabada, L. Saavedra *Existence of solutions for* n^{th}*-order nonlinear differential boundary value problems by means of new fixed point theorems*, arXiv:1703.09115

[6] P. Drabek, G. HolubovÃ¡ *On the maximum and antimaximum principles for the beam equation* Appl. Math. Lett. **56** (2016) 29-33.

[7] P. Drabek, G. HolubovÃ¡ *Positive and negative solutions of one-dimensional beam equation*, Appl. Math. Lett. **51** (2016), 1-7.

[8] J. Schröder, *Operator inequalities* in: Mathematics in Science and Engineering, vol. 147, Academic Press, nc., New York-London, 1980.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

Note on resonant problems

Felix Sadyrbaev[1]

[1] *Institute of Mathematics and Computer Science, University of Latvia, Latvia*

emails: felix@latnet.lv

Abstract

The second order boundary value problem with the Dirichlet boundary conditions is essentially resonant if it has k-resonant solution.

Key words: resonant boundary value problem
MSC 2000: 34B15

1 Introduction

The second order boundary value problem (BVP in short)

$$x'' + p(t)x' + q(t)x = \varphi(t, x, x'), \tag{1}$$

$$x(a) = A, \quad x(b) = B, \tag{2}$$

where all coefficients and functions are continuous and the right side φ is bounded (in modulus) is solvable if the homogeneous problem

$$x'' + p(t)x' + q(t)x = 0, \tag{3}$$

$$x(a) = 0, \quad x(b) = 0 \tag{4}$$

has only the trivial solution ([1]). If φ is not bounded various approaches and techniques can be used to reduce the problem to the quasi-linear form with bounded right hand side.

In case the homogeneous problem (3), (4) has a nontrivial solution the problem (1), (2) is called *resonant*. Resonant problems may have no solutions even for the simple case $\varphi = \varphi(t)$. The Fredholm alternative provides criteria for solvability in this case.

In this note we wish to emphasize that formally resonant problems mostly are not essentially resonant.

 ISBN: 978-84-617-8694-7

2 Definitions

To make this idea more precise we will move to the final statement in several easy steps.

First, let us distinguish between linear expressions of the form $(l_2x)(t) := x'' + p(t)x' + q(t)x$. We will say that *the type of* $(l_2x)(t)$ *is* k if a solution $x(t)$ of the respective Cauchy problem

$$x'' + p(t)x' + q(t)x = 0, \quad x(a) = 0, \quad x'(a) = 1 \tag{5}$$

has exactly k zeros in the interval (a, b) and $x(b) \neq 0$.

If $x(b) = 0$ we will call the linear expression $(l_2x)(t)$ by $(k+1)$-resonant (the $x(t)$ in (5) has k internal zeros and additional zero at $t = b$).

In a similar way we classify possible solutions of the problem (1), (2). We will say that a solution $\xi(t)$ of the problem (1), (2) is of type n if the respective linear differential expression

$$y'' + p(t)y' + y(t)x - \varphi_x(t, \xi(t), \xi'(t))y - \varphi_{x'}(t, \xi(t), \xi'(t))y' \tag{6}$$

is of type n. The latter means that a solution of the Cauchy problem

$$y'' + p(t)y' + y(t)x = \varphi_x(t, \xi(t), \xi'(t))y + \varphi_{x'}(t, \xi(t), \xi'(t))y', \tag{7}$$

$$y(a) = 0, \quad y'(a) = 1 \tag{8}$$

has exactly n zeros in (a, b) and $y(b) \neq 0$. If $y(b) = 0$ we will say that $\xi(t)$ *is* $(n+1)$*-resonant solution.*

We wish to show that essentially resonant BVP is a problem that has i-resonant solution for some i.

All other "resonant" problems can be reduced to a quasi-linear form (1), (2) with a bounded right side.

Remark. It was shown in [2] that a quasi-linear problem in the form (1), (2) with linear part (in the left hand side) of the type k has a solution $x(t)$ that is either of type k, or k-resonant, or $(k + 1)$-resonant.

3 Observations

Proposition 3.1 *Suppose a resonant problem (1), (2), where φ is continuous in all arguments and continuously differentiable with respect to x and x', has a solution $\xi(t)$ of type n. Then the problem (1), (2) can be reformulated in a quasi-linear form*

$$x'' + P(t)x' + Q(t)x = \tilde{\varphi}(t, x, x'), \quad x(a) = A, \quad x(b) = B, \tag{9}$$

where the linear part $(L_2x)(t) := x'' + P(t)x' + Q(t)x$ is of type n.

ISBN: 978-84-617-8694-7

Proof. Consider $x'' + p(t)x' + q(t)x = \varphi(t, x, x')$ and $\xi''(t) + p(t)\xi'(t) + q(t)\xi(t) = \varphi(t, \xi(t), \xi'(t))$, where $\xi(t)$ is a solution of type n of the problem (1), (2). Then

$$\begin{aligned}(x-\xi)'' + p(t)(x-\xi)' \quad &+q(t)(x-\xi(t)) = \varphi(t,x,x') - \varphi(t,\xi,\xi') \\ &= \varphi_x(t,\xi,\xi')(x-\xi) + \varphi_{x'}(t,\xi,\xi')(x-\xi)' + r(t,x,x')\end{aligned} \tag{10}$$

or

$$\begin{aligned}x'' + p(t)x \quad &+q(t)x - \varphi_x x - \varphi_{x'}x' \\ &= \xi'' + p(t)\xi' + q(t)x - \varphi_x\xi - \varphi_{x'}\xi' + r(t,x,x')\end{aligned} \tag{11}$$

or, finally,

$$(L_2x)(t) = h(t) + \tilde{r}(t, x, x'),$$

where
$(L_2x)(t) = x'' + (p(t) - \varphi_{x'}(t,\xi,\xi'))x' + (q(t) - \varphi_x(t,\xi,\xi'))x$,
$h(t) = \xi''(t) + (p(t) - \varphi_{x'}(t,\xi,\xi'))\xi'(t) + (q(t) - \varphi_x(t,\xi,\xi'))\xi(t)$,
$\tilde{r}(t,x,x')$ is C^1 smooth bounded function that coincides with $r(t,x,x')$ in some vicinity of $(t, \xi(t), \xi'(t),\ \ t \in [a,b]$.

The linear differential expression $(L_2x)(t)$ is of type n since a solution $\xi(t)$ is supposed to be of type n. □

We are in a position now to make more precise the "vague" term *essentially resonant BVP* used earlier.

If the second order BVP cannot be reduced to a quasi-linear form around each of its solutions (if any) then we will say that this problem is *essentially resonant.*

4 Conclusions

We arrived therefore at the following result.

Theorem 4.1 *The problem (1), (2) is essentially resonant if for some k it has a k-resonant solution. Otherwise either it has no solutions or it can be reduced to a quasi-linear problem around any of existing solutions.*

References

[1] R. Conti, *Equazioni differenziali ordinarie quasilineari con condizioni lineari*, Ann. mat. pura ed appl. **57** (1962), 49 - 67.

[2] I. Yermachenko and F. Sadyrbaev, *Types of solutions and multiplicity results for two-point nonlinear boundary value problems*, Nonlinear Analysis **63** (2005), e1725–e1735.

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

Wave propagation through linear viscoelastic media using the Generalized Finite Difference Method

E. Salete[1], M. Ureña[1], J.J. Benito[1], F. Ureña[2], A. Muelas[1] and L. Gavete[3]

[1] *Departamento de Construcción y Fabricación, Universidad Nacional de Educación a Distancia (UNED)*

[2] *Departamento de Matemática Aplicada, Universidad de Castilla-La Mancha (UCLM)*

[3] *Departamento de Matemática Aplicada a los Recursos Naturales, Universidad Politécnica de Madrid (UPM)*

emails: esalete@ind.uned.es, miguelurenya@gmail.com, jbenito@ind.uned.es, fuprieto@terra.es, amuelas@ind.uned.es, lu.gavete@upm.es

Abstract

Formulation for P and SV wave propagation through viscoelastic media, described by the Kelvin–Voigt model, using the generalized finite difference method is proposed. The stability of the proposed scheme is analyzed and a criterion to guarantee the stability is obtained. The star dispersion for both P and SV waves and group velocity are also analyzed.

A numerical example of a wave traveling on a bidimensional half-plane, initiated perpendicular to its free surface, is developed and solved. The obtained results are compared with its analytical solution.

Key words: Generalized finite difference method, Wave propagation, Viscoelasticity
MSC 2000: 65M06, 65M12

1 Introduction

Finite differences methods are numerical methods widely used for solving wave propagation problems. Traditionally, these methods are based on a mesh (discretization of the domain), but in the past decades meshless methods have been developed.

Generalized finite differences method (GFDM) is one of these finite differences methods that do not require a base mesh. It is based on moving least squares. The method approximates the value of the unkown function at a series of points within the domain (usually refered to as nodes). The derivatives of the function are linearized using a Taylor expansion, at each of these points. And therefore the problem is reduced to solving a linear system of equations. The current formulation of the method was presented by Liska and Orkisz[1] in 1998 and Benito et al. [2] provided in 2001, explicit formulae for the resolution of the problem.

The application of GFDM to wave propagation problems was introduced by Ureña et. al. [3]. They studied the star dispersion and stability for different type of waves [4] traveling through elastic domains. They also analyzed the simulation of absorbing boundary conditions, and obtained schemes for them, using perfectly matched layers (PML) [5].

The present paper proposes a formulation for including viscoelastic materials described by the Kelvin–Voigt model as the media through which the wave travels, considering the energy attenuation of the wave.

A scheme for GFDM is proposed and its stability is analyzed. A criterion to guarantee the stability is obtained. The star dispersion for both P and SV waves and group velocity are also analyzed.

Finally, a numerical example is proposed and solved. A wave is applied on the free surface of a bidimensional half-plane, perpendicular to its free surface. The obtained results are compared to its analytical solution.

2 Linear viscoelastic material

Certain materials behave in such a way that cannot be considered purely elastic or purely viscuous, but exhibit an intermediate behavior between both. In these viscoelastic materials the strains depend on time even without external forces applied. Stresses (represented by the tensor σ_{ij}) depend not only on strains but also on the velocity of deformation. Different models can be used to represent these relationships; in this article the Kelvin-Voigt model is used, which can be expressed for a bidimensional media as:

 ISBN: 978-84-617-8694-7

$$\begin{cases} \sigma_{xx} = \lambda\left(\dfrac{\partial u}{\partial x} + \dfrac{\partial w}{\partial z}\right) + \lambda t_r \dfrac{\partial}{\partial t}\left(\dfrac{\partial u}{\partial x} + \dfrac{\partial w}{\partial z}\right) + 2\mu\dfrac{\partial u}{\partial x} + 2\mu t_r \dfrac{\partial}{\partial t}\dfrac{\partial u}{\partial x} \\ \sigma_{zz} = \lambda\left(\dfrac{\partial u}{\partial x} + \dfrac{\partial w}{\partial z}\right) + \lambda t_r \dfrac{\partial}{\partial t}\left(\dfrac{\partial u}{\partial x} + \dfrac{\partial w}{\partial z}\right) + 2\mu\dfrac{\partial w}{\partial z} + 2\mu t_r \dfrac{\partial}{\partial t}\dfrac{\partial w}{\partial z} \\ \sigma_{xz} = \mu\left(\dfrac{\partial u}{\partial z} + \dfrac{\partial w}{\partial x}\right) + \mu t_r \dfrac{\partial}{\partial t}\left(\dfrac{\partial u}{\partial z} + \dfrac{\partial w}{\partial x}\right) \end{cases} \tag{1}$$

Where t_r is the relaxation time, λ the first Lamé parameter and μ the shear modulus.

$$\lambda = \frac{E\nu}{(1+\nu)(1-2\nu)}, \mu = G = \frac{E}{2(1+\nu)}, t_r = \frac{\eta}{E} \tag{2}$$

With ν as the Poisson ratio, E the elasticity modulus and η is the viscosity of the material.

3 Equations of motion

The general equations of motion of P-SV waves through a bidimensional (x-z plane) domain $\Omega \subset \mathbb{R}^2$ can be expressed, in absence of body forces, as:

$$\begin{cases} \rho\dfrac{\partial^2 U}{\partial t^2} = \dfrac{\partial \sigma_{xx}}{\partial x} + \dfrac{\partial \sigma_{xz}}{\partial z} \\ \rho\dfrac{\partial^2 W}{\partial t^2} = \dfrac{\partial \sigma_{zx}}{\partial x} + \dfrac{\partial \sigma_{zz}}{\partial z} \end{cases} \tag{3}$$

with the appropriate boundary conditions for each case.

U and W represent the displacements in two orthogonal directions and ρ is the densitiy of the medium.

By using the Lamé expressions to include the Kelvin-Voigt viscoelastic model into Eq.(3) the following expression in displacements is obtained.

$$\begin{cases} U_{tt} = & \alpha^2 U_{xx} + \beta^2 U_{zz} + (\alpha^2 - \beta^2) W_{xz} + t_r\Big(\alpha^2 U_{xx} + \beta^2 U_{zz} + \\ & +(\alpha^2 - \beta^2) W_{xz}\Big)_t \\ W_{tt} = & \alpha^2 W_{zz} + \beta^2 W_{xx} + (\alpha^2 - \beta^2) U_{xz} + t_r\Big(\alpha^2 W_{zz} + \beta^2 W_{xx} + \\ & +(\alpha^2 - \beta^2) U_{xz}\Big)_t \end{cases} \tag{4}$$

 ISBN: 978-84-617-8694-7

where α and β are the velocities of the P and SV waves respectively:

$$\begin{cases} \alpha = \sqrt{\dfrac{\lambda + 2\mu}{\rho}} \\ \beta = \sqrt{\dfrac{\mu}{\rho}} \end{cases} \tag{5}$$

4 Generalized finite differences method

4.1 Generalized finite differences scheme

The application of GFDM to media with elastic materials is completely developed in [6]. The methodology is breifly summarized hereafter.

A cloud of nodes is set within the domain Ω. In order to approximate the derivatives of the displacements by explicit linear expressions, a star is formed at each of these node. The star is made up by the node itself (center node) and a series of sourrounding nodes. The selection criteria for the nodes that form the star has already been studied by J.J. Benito et al. [2]

Let us consider a discretization D of the domain Ω, the inner node $(\mathbf{x}_0, t_n) \in D$ and the function ψ twice differentiable. Denoting the number of nodes of a star by N and the coefficients both the central node and the other nodes by m_{0ij} and m_{kij}, respectively, where $i, j \in \{x, z\}$ and $k \in \{1, ...N\}$, the linearization of each partial derivative is given by

$$\psi^n_{0,ij} = -m_{0ij}\psi^n_0 + \sum_{k=1}^{N} m_{kij}\psi^n_k \tag{6}$$

The temporal approximations are given by the following classical finite differences for first and second order, respectively

$$\psi^n_{0,t} = \frac{3\psi^n_0 - 4\psi^{n-1}_0 + \psi^{n-2}_0}{2\triangle t} \tag{7}$$

$$\psi^n_{0,tt} = \frac{\psi^{n-1}_0 - 2\psi^n_0 + \psi^{n+1}_0}{\triangle t^2} \tag{8}$$

Substituting (6), (7) and (8) in (4) and arranging terms, the scheme for P-SV waves in viscoelastic media is obtained

$$\begin{cases} u^{n+1}_0 = \sum\limits_{k=0}^{2}\left[A^{-k}_{u_0}u^{n-k}_0 + A^{-k}_{w_0}w^{n-k}_0 + \sum\limits_{j=1}^{N}\left(A^{-k}_{u_j}u^{n-k}_j + A^{-k}_{w_j}w^{n-k}_j\right)\right] \\ w^{n+1}_0 = \sum\limits_{k=0}^{2}\left[B^{-k}_{w_0}w^{n-k}_0 + B^{-k}_{u_0}u^{n-k}_0 + \sum\limits_{j=1}^{N}\left(B^{-k}_{w_j}w^{n-k}_j + B^{-k}_{u_j}u^{n-k}_j\right)\right] \end{cases} \tag{9}$$

 ISBN: 978-84-617-8694-7

where the coefficients are

$$\begin{cases} A_{u_0}^0 = 2 - (\triangle t^2 + 1.5t_r\triangle t)(\alpha^2 m_{0xx} + \beta^2 m_{0zz}) \\ A_{u_0}^{-1} = -1 + 2t_r\triangle t(\alpha^2 m_{0xx} + \beta^2 m_{0zz}) \\ A_{u_0}^{-2} = -0.5t_r\triangle t(\alpha^2 m_{0xx} + \beta^2 m_{0zz}) \\ A_{w_0}^0 = B_{u_0}^0 = -(\triangle t^2 + 1.5t_r\triangle t)(\alpha^2 + \beta^2)m_{0xz} \\ A_{w_0}^{-1} = B_{u_0}^{-1} = 2t_r\triangle t(\alpha^2 + \beta^2)m_{0xz} \\ A_{w_0}^{-2} = B_{u_0}^{-2} = -0.5t_r\triangle t(\alpha^2 + \beta^2)m_{0xz} \\ A_{u_j}^0 = (\triangle t^2 + 1.5t_r\triangle t)(\alpha^2 m_{jxx} + \beta^2 m_{jzz}) \\ A_{u_j}^{-1} = -2t_r\triangle t(\alpha^2 m_{jxx} + \beta^2 m_{jzz}) \\ A_{u_j}^{-2} = 0.5t_r\triangle t(\alpha^2 m_{jxx} + \beta^2 m_{jzz}) \\ A_{w_j}^0 = B_{u_j}^0 = (\triangle t^2 + 1.5t_r\triangle t)(\alpha^2 + \beta^2)m_{jxz} \\ A_{w_j}^{-1} = B_{u_j}^{-1} = -2t_r\triangle t(\alpha^2 + \beta^2)m_{jxz} \\ A_{w_j}^{-2} = B_{u_j}^{-2} = 0.5t_r\triangle t(\alpha^2 + \beta^2)m_{jxz} \\ B_{w_0}^0 = 2 - (\triangle t^2 + 1.5t_r\triangle t)(\alpha^2 m_{0zz} + \beta^2 m_{0xx}) \\ B_{w_0}^{-1} = -1 + 2t_r\triangle t(\alpha^2 m_{0zz} + \beta^2 m_{0xx}) \\ B_{w_0}^{-2} = -0.5t_r\triangle t(\alpha^2 m_{0zz} + \beta^2 m_{0xx}) \\ B_{w_j}^0 = (\triangle t^2 + 1.5t_r\triangle t)(\alpha^2 m_{jzz} + \beta^2 m_{jxx}) \\ B_{w_j}^{-1} = -2t_r\triangle t(\alpha^2 m_{jzz} + \beta^2 m_{jxx}) \\ B_{w_j}^{-2} = 0.5t_r\triangle t(\alpha^2 m_{jzz} + \beta^2 m_{jxx}) \end{cases} \tag{10}$$

4.2 Stability

We apply the Von Neumann method in the following harmonic waves

$$\begin{cases} u = a\exp(i(\omega t - \mathbf{k}^T\mathbf{x})) \\ w = b\exp(i(\omega t - \mathbf{k}^T\mathbf{x})) \end{cases} \tag{11}$$

where a and b are amplitudes, ω the angular frequency, $\mathbf{k} = (k_x, k_z)$ the wavenumber vector and $\mathbf{x} = (x, z) \in D$. For each star and time and denoting $\xi = \exp(i\omega\triangle t)$,

$$\begin{cases} u_0^n = a\xi^n\exp(-i\mathbf{k}^T\mathbf{x}_0) \\ u_j^n = a\xi^n\exp(-i\mathbf{k}^T\mathbf{x}_j) \\ w_0^n = b\xi^n\exp(-i\mathbf{k}^T\mathbf{x}_0) \\ w_j^n = b\xi^n\exp(-i\mathbf{k}^T\mathbf{x}_j) \end{cases} \tag{12}$$

 ISBN: 978-84-617-8694-7

We substitute (12) in (9), we denote $\mathbf{h}_j = \mathbf{x}_j - \mathbf{x}_0$ and we take into account that $m_{0xx} = \sum_{j=1}^{N} m_{jxx}$, $m_{0zz} = \sum_{j=1}^{N} m_{jzz}$ and $m_{0xz} = \sum_{j=1}^{N} m_{jxz}$. Then dividing by $\xi^n \exp(-i\mathbf{k}^T\mathbf{x}_0)$ and arranging terms, the following is obtained

$$\begin{cases} a\Gamma = b\Phi \Rightarrow \frac{a}{b} = \frac{\Phi}{\bar{\Gamma}} \\ \\ b\bar{\Gamma} = a\Phi \Rightarrow \frac{a}{b} = \frac{\bar{\Gamma}}{\Phi} \end{cases} \Rightarrow \Gamma\bar{\Gamma} = \Phi^2 \tag{13}$$

where

$$\begin{cases} \Gamma = \xi - \sum\limits_{k=0}^{2} \xi^{-k} \left(A_{u_0}^{-k} + \sum\limits_{j=1}^{N} A_{u_j}^{-k} \exp(-i\mathbf{k}^T\mathbf{h}_j) \right) \\ \bar{\Gamma} = \xi - \sum\limits_{k=0}^{2} \xi^{-k} \left(B_{w_0}^{-k} + \sum\limits_{j=1}^{N} B_{w_j}^{-k} \exp(-i\mathbf{k}^T\mathbf{h}_j) \right) \\ \Phi = \sum\limits_{k=0}^{2} \xi^{-k} \sum\limits_{j=1}^{N} A_{w_j}^{-k} (-1 + \exp(-i\mathbf{k}^T\mathbf{h}_j)) \end{cases} \tag{14}$$

The following notations are considered

$$\begin{cases} P = \sum\limits_{j=1}^{N} (\alpha^2 m_{jxx} + \beta^2 m_{jzz})(1 - \exp(-i\mathbf{k}^T\mathbf{h}_j)) \\ Q = \sum\limits_{j=1}^{N} (\alpha^2 m_{jzz} + \beta^2 m_{jxx})(1 - \exp(-i\mathbf{k}^T\mathbf{h}_j)) \\ R = \sum\limits_{j=1}^{N} (\alpha^2 - \beta^2) m_{jxz} (1 - \exp(-i\mathbf{k}^T\mathbf{h}_j)) \\ \chi = \dfrac{2\triangle t}{t_r} \cdot \dfrac{-\xi^3 + 2\xi^2 - \xi}{p\xi^2 - 4\xi + 1} \end{cases} \tag{15}$$

where $p = 3 + 2\frac{\triangle t}{t_r}$.

Operating, the following second order equation in χ is obtained

$$\begin{aligned} \Gamma\bar{\Gamma} = \Phi^2 \Rightarrow (\chi - \triangle t^2 P)(\chi - \triangle t^2 Q) = (\triangle t^2 R)^2 \Rightarrow \\ \Rightarrow \chi^2 - \triangle t^2 (P+Q)\chi + \triangle t^4 (PQ - R^2) = 0 \end{aligned} \tag{16}$$

and the solution is

$$\chi = \frac{\triangle t^2}{2} \left(P + Q + \sqrt{(P-Q)^2 + 4R^2} \right) \tag{17}$$

Denoting $\Upsilon = P + Q + \sqrt{(P-Q)^2 + 4R^2}$

 ISBN: 978-84-617-8694-7

$$\chi = \frac{\triangle t^2}{2}\Upsilon \tag{18}$$

Taking into account that $|\xi| \leq 1$ and using conservative criteria, a stability criterion is achieved

$$\triangle t \leq -2t_r + 2\sqrt{t_r^2 + \frac{2}{(\alpha^2+\beta^2)[(m_{0xx}+m_{0zz}) + \sqrt{(m_{0xx}+m_{0zz})^2 + 4m_{0xz}^2}]}} \tag{19}$$

4.3 Star dispersion for P and SV waves

From (18) and calling $\Theta = \frac{t_r \triangle t}{4}\Upsilon$, the following equation is obtained

$$\xi^3 + (p\Theta - 2)\xi^2 + (1 - 4\Theta)\xi + \Theta = 0 \tag{20}$$

Denoting

$$\begin{cases} \triangle_0 = (p\Theta - 2)^2 + 12\Theta - 3 \\ \triangle_1 = 2(p\Theta - 2)^3 + 36p\Theta^2 - (9p + 45)\Theta + 18 \\ \triangle_2 = \sqrt[3]{\dfrac{\triangle_1 + \sqrt{\triangle_1^2 - 4\triangle_0^3}}{2}} \end{cases} \tag{21}$$

the solution of (20) is

$$\xi_k = -\frac{1}{3}\left(p\Theta - 2 + \zeta^k \triangle_2 + \frac{\triangle_0}{\zeta^k \triangle_2}\right) \tag{22}$$

being $\zeta = -\frac{1}{2} + \frac{\sqrt{3}}{2}i$ and $k \in \{0, 1, 2\}$.

As $\xi = \exp(i\omega\triangle t)$, isolating ω_k we obtain $\omega_k = \frac{Arg(\xi_k)}{\triangle t}$ and therefore

$$\omega_{GFD} = \min_{k\in\{0,1,2\}} |\omega_k - \omega| = \frac{Arg(\xi)}{\triangle t} \tag{23}$$

This way, the velocity ratios for both waves are

$$\begin{cases} \dfrac{\alpha_{GFD}}{\alpha} = \dfrac{\lambda_P Arg(\xi)}{2\pi\alpha\triangle t} \\ \dfrac{\beta_{GFD}}{\beta} = \dfrac{\lambda_{SV} Arg(\xi)}{2\pi\beta\triangle t} \end{cases} \tag{24}$$

where λ_P and λ_{SV} are the wavelengths of P and SV waves, respectively.

 ISBN: 978-84-617-8694-7

4.4 Star dispersion for group velocity

First of all, we consider the following derivatives

$$\begin{cases} P_{\mathbf{k}} = i\sum_{j=1}^{N}(\alpha^2 m_{jxx} + \beta^2 m_{jzz})\exp(-i\mathbf{k}^T\mathbf{h}_j)\mathbf{h}_j \\ Q_{\mathbf{k}} = i\sum_{j=1}^{N}(\alpha^2 m_{jzz} + \beta^2 m_{jxx})\exp(-i\mathbf{k}^T\mathbf{h}_j)\mathbf{h}_j \\ R_{\mathbf{k}} = i\sum_{j=1}^{N}(\alpha^2 - \beta^2) m_{jxz}\exp(-i\mathbf{k}^T\mathbf{h}_j)\mathbf{h}_j \\ \Upsilon_{\mathbf{k}} = P_{\mathbf{k}} + Q_{\mathbf{k}} + \dfrac{(P-Q)(P_{\mathbf{k}} - Q_{\mathbf{k}}) + 4RR_{\mathbf{k}}}{\sqrt{(P-Q)^2 + 4R^2}} \\ \Theta_{\mathbf{k}} = \dfrac{t_r \triangle t}{4}\Upsilon_{\mathbf{k}} \\ \triangle_{0,\mathbf{k}} = (2p^2\Theta - 4p + 12)\Theta_{\mathbf{k}} \\ \triangle_{1,\mathbf{k}} = [6p(p\Theta - 2)^2 + 72p\Theta - 9p - 45]\Theta_{\mathbf{k}} \\ \triangle_{2,\mathbf{k}} = \dfrac{1}{6}\sqrt[3]{\left(\dfrac{2}{\triangle_1 + \sqrt{\triangle_1^2 - 4\triangle_0^3}}\right)^2}\left(\triangle_{1,\mathbf{k}} + \dfrac{\triangle_1\triangle_{1,\mathbf{k}} - 6\triangle_0\triangle_{0,\mathbf{k}}}{\sqrt{\triangle_1^2 - 4\triangle_0^3}}\right) \\ \xi_{\mathbf{k}} = -\dfrac{1}{3}\left(\Theta_+\zeta^k\triangle_{2,\mathbf{k}} + \dfrac{\triangle_{0,\mathbf{k}}\zeta^k\triangle_2 - \triangle_0\zeta^k\triangle_{2,\mathbf{k}}}{\zeta^{2k}\triangle_2^2}\right) = Re(\xi_{\mathbf{k}}) + iIm(\xi_{\mathbf{k}}) \end{cases} \tag{25}$$

where the value of k in the last case is the same that the value chosen in (23).

The group velocity is $v^g = \nabla_{\mathbf{k}}\omega = \dfrac{Im(\xi_{\mathbf{k}})Re(\xi) - Im(\xi)Re(\xi_{\mathbf{k}})}{\triangle t}$

5 Numerical example

The previously developed scheme is applied to a numerical example.

The domain is the $z > 0$ subspace with a boundary condition at $y = 0$ defined as a free surface with a normal wave applied, moving through the x direction with a velocity v:

$$\begin{cases} \sigma_{xz}(z=0) = 0 \\ \sigma_{zz}(z=0) = -p_0 \cdot exp(iC(x - v\cdot t)) \end{cases} \tag{26}$$

The previous boundary condition only uses the real part of the expresion, as the stress (σ_{zz}) is a real quantity. The real part of all quantities should be taken to obtain physically meaningful results.

 ISBN: 978-84-617-8694-7

The problem is solved and its analytical solution is obtained in [7].

A cloud of 19076 nodes is generated for a 5.00 m x 1.50 m domain and the GFDM is applied to solve the problem for this particular case. The maximum distance between nodes is 0.0283 m. The criterion for the selection of star nodes is the quadrant criterion (see [2]).

Hysteretic damping is considered, assuming that the product $\omega \cdot t_r$ is constant, and therefore the hysteretic damping ratio is $h = \omega t_r/2$, being $\omega = C \cdot v$ the frequency of the loading.

By using a null Poisson's ratio all the material properties can be expressed as a function of β and the results can therefore be normalized.

Applying the stability criterion, the maximum time step can be obtained: $\Delta t_{max} = 6.67 \cdot 10^{-5}$. A time step of $\Delta t = 4.20 \cdot 10^{-5} < \Delta t_{max}$ is used, for 5000 calculation steps.

Figure 1 represents movement amplitud versus velocity. Dimensionless variables have been used, showing the ratio of the dynamic amplitude to the static amplitude (for $v = 0$) and the ratio of the velocity of the wave v to the velocity of the shear waves β. Different curves have been obtained for different damping ratios. Dashed lines show the results obtained using the GFDM and continuous lines are used to represent the analyitical solution.

As it can be seen, the obtained results using GFDM match very accurately the analytical solution. The maximum deviation from the analytical solution is obtained for $h = 0.3$ and its value is $|1 - \mathrm{w}_{GFDM}/\mathrm{w}_{Analytical}| = 4.5\%$.

6 Conclusions

Formulation for P and SV wave propagation through viscoelastic media, described by the Kelvin–Voigt model, using the generalized finite difference method has been proposed.

The stability of the scheme is analyzed, obtaining a criterion valid to guarantee this stability.

Also, the star dispersion for both P and SV waves and group velocity are analyzed, and expressions for the dispersions are obtained.

A numerical example has been developed and compared with its analytical solution, showing the validity of the presented method and its application.

 ISBN: 978-84-617-8694-7

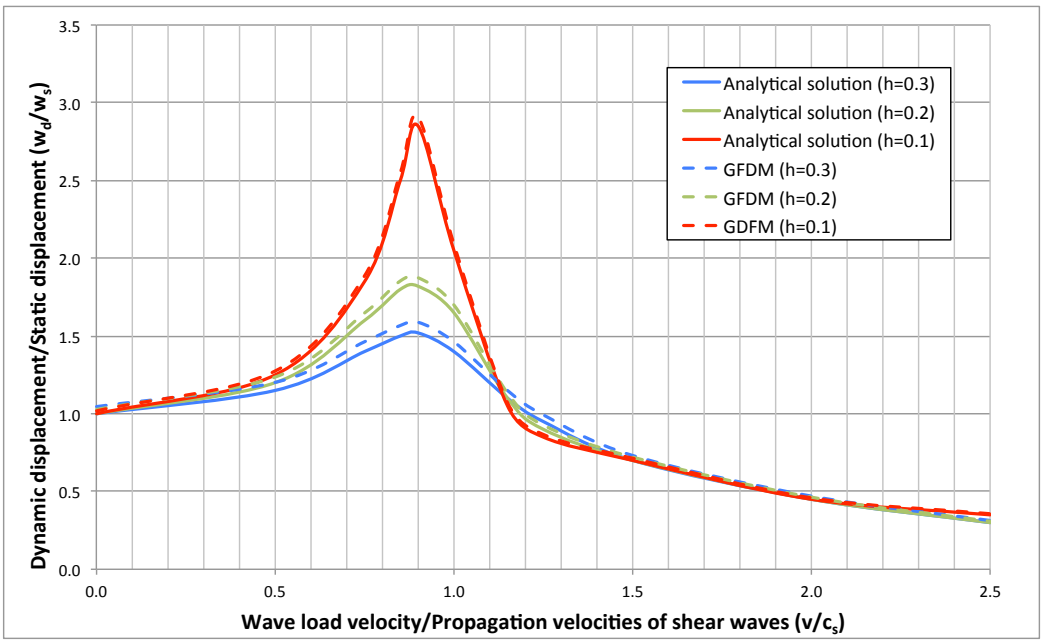


Figure 1: Movement amplitude vs velocity

ISBN: 978-84-617-8694-7

Acknowledgements

The authors acknowledge the support of the *Escuela Técnica Superior de Ingenieros Industriales (UNED)* of Spain, project Ref: 2016-IFC02.

References

[1] J. Orkisz. *Finite Difference Method (Part, III) in handbook of Computational Solid Mechanics*. M. Kleiber (Ed.) Springer-Verlag, Berlin 1998.

[2] F. Ureña, J.J. Benito, and L. Gavete. Influence several factors in the generalized finite difference method. *Applied Mathematical Modeling*, 25:10391053, 2001.

[3] F. Ureña, J.J. Benito, E. Salete, and L. Gavete. A note on the application of the generalized finite difference method to seismic wave propagation in 2-D. *Journal of Computational and Applied Mathematics*, 236(12):3016–3025, 2012.

[4] J.J. Benito, F. Ureña, L. Gavete, E. Salete, and A. Muelas. A GFDM with PML for seismic wave equations in heterogeneous media. *Journal of Computational and Applied Mathematics*, 252:40–51, 2013.

[5] J.J. Benito, F. Ureña, E. Salete, A. Muelas, L. Gavete, and R. Galindo. Wave propagation in soils problems using the Generalized Finite Difference Method. *Soil Dynamics and Earthquake Engineering*, 79, Part A:190–198, 2015.

[6] M. Ureña, J.J. Benito, F. Ureña, E. Salete, and L. Gavete. Application of generalised finite differences method to reflection and transmission problems in seismic SH waves propagation. *Mathematical methods in the Applied Sciences*, 2016.

[7] A. Verruijt. *Soil Dynamics*. Delft University of Technology, 1994, 2008.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

Efficient Parallel Implementation of Active-Set Newton Algorithm for Non-Negative Sparse Representations

P. San Juan[1], Tuomas Virtanen[2], V. M. Garcia-Molla[1] and A. M. Vidal[1]

[1] *Department of Information Systems and Computing, Universitat Politècnica de València*

[2] *Department of Signal Processing, ampere Uni- versity of Technology*

emails: p.sanjuan@upv.es, tuomas.virtanen@tut.fi, vmgarcia@dsic.upv.es, avidal@dsic.upv.es

Abstract

This paper presents an efficient parallel implementation of the Active-Set Newton Algorithm (ASNA) for overcomplete non-negative representations of audio, which performance has been proved in previous works against other state of the art methods. The implementation presented in the paper has been developed in C, using parallel programming techniques to obtain a better performance in multicore architectures than the original MATLAB implementation.

Key words: non-negative matrix factorization, Newton algorithm, convex optimization, sparse representation, multicore, parallel computing

1 Introduction

The Active-Set Newton Algorithm (ASNA) presented in [1, 2] is an algorithm designed to minimize the Kullback-Leibler divergence for non-negative decomposition, which have been widely used in audio processing, for example on source separation [3], automatic music transcription [4], and sound event detection [5]. The KL divergence between vectors x and $\acute{x}$ is defined as

$$KL(x||\acute{x}) = \sum_i d(x_i, \acute{x_i})$$

where function d corresponds to

$$d(p,q) = \begin{cases} p\ log(p/q) - p + q) & p > 0 \text{ and } q > 0 \\ q & p = 0 \\ \infty & p > 0 \text{ and } q = 0 \end{cases}$$

 ISBN: 978-84-617-8694-7

Observation vector is modelled as $x \approx v = Bw$ subject to $(w, B, x, v) \geq 0$ where $x \in \Re^f$ is the original observation, $v \in \Re^f$ is the approximation obtained form the decomposition, $w \in \Re^n$ are the the non-negative weights and $B \in \Re^{f \times n}$ is the precomputed dictionary, which contains n atoms.

In the original paper [1], ASNA algorithm showed its advantages against some state of the art algorithms like the expectation-maximization update rules [6] and the projected gradient algorithm [7, pp. 267-268] for the overcomplete case. Due to this, we decided to improve the existing MATLAB implementation in order to obtain an efficient parallel version suitable for shared memory multicore machines.

The structure of the paper is as follows. In Section 2 ASNA algorithm and its existing implementation is explained. In Section 3 the developed implementations are presented. Then, Section 4 shows some preliminary results and a brief comparison between implementations. Finally, in Section 5 the results and the ongoing work are discussed.

2 ASNA algorithm

The main principle of the ASNA algorithm is that it estimates and updates a set of "active" atoms that have non-zero weights. The active set is initialized with a single atom which alone gives the smallest divergence. Then, it finds the most promising atom not in the active set by identifying the atom whose weight derivative is the smallest, and adds it to the active set. The weights of the atoms in the active set are estimated using the Newton method where the step size is chosen to ensure non-negativity of the weights. Atoms whose weights become zero or negative are removed from the active set. The algorithm iterates until a convergence criterion is achieved or a maximum number of iterations given by the user are reached. A detailed view of the algorithm can is found in [1, Sec. III].

The existing implementation programmed in MATLAB, that can be found in [8], uses a more general model than the one shown in the original algorithm [1, pp. 5]. The extended model can work with multiple observations at time, becoming $X \approx V = WB$ subject to $(W, B, X, V) \geq 0$ where the rows of $X, V \in \Re^{o \times f}$ are the observations and the rows of $W \in \Re^{o \times n}$ are the non-negative weights of each corresponding observation. Note that the model is transposed. That model gets some advantages from the fact of computing multiple observations at a time.

A brief pseudocode of that implementation can be found in Algorithm 1. In that implementation the weights in the active set are represented by the nonzero elements in a sparse weight matrix W_A and the active atoms in the dictionary are represented by S_A.

 ISBN: 978-84-617-8694-7

Algorithm 1 Original ASNA implementation algorithm

Require: $X \in \Re^{o \times f}$ $B \in \Re^{n \times f}$.

return $W \in \Re^{o \times n}$
Normalize each dictionary atom to unity norm
Pre compute operations for the gradient computations
(some matrix operations in the gradient computation only depend on the entry data and can be precomputed for speed)
Initialize active set for each observation
(Active atoms have values in W_A and not active are 0)
for $i = 1$ **to** $maximunnumberofiterations$ **do**
  Find active atoms
  Compute $V = W_A S_A$
  $R = X/V$ (element wise)
  if $i \ mod \ 2 = 0$ **then**
    Compute gradient w.r.t all weights
    if $i \ mod \ 10 = 0$ **then**
      Check convergence for non converged observations
      Remove converged observations from the computations
      if all observation have converged **then**
        Scale back W and exit
      end if
    end if
    Mark as 0 the gradient of the already active weights
    Add the atom with the minimum gradient of each observation
    to the active set, adding a small number to W_A
  end if
  Compute $R2 = X/V^2$ (element wise)
  Find the indexes of the active atoms
  Compute sparse product $Rcov = RS^T$
  for each observation not converged t **do**
    Find the active atoms of t
    if all gradients computed **then**
      Get $grad$ from the already computed gradients
    else
      Compute gradients w.r.t active atoms of t ($grad$)
    end if
    Compute Hessian GG
    Solve $GG \times searchDir = grad$
    Compute step size
    Update weights in W_A. If a weight becomes negative is removed.
  end for
end for

 ISBN: 978-84-617-8694-7

3 Proposed algorithms

The first step was to improve the existing MATLAB implementation, then we implemented a parallel version of the algorithm in C programming language using HPC mathematical libraries. Finally, we implemented a parallel version of the algorithm using threading with OPENMP. In order to save space, the full algorithm for each proposed implementation is not written in the paper and only the most important changes will be explained. The source code of all proposed implementations can be found in [9].

3.1 Improved MATLAB implementation

The improved MATLAB implementation has some modifications that affect positively to the performance of the algorithm.

The first change was transposing the problem, most of the operations in the original implementation were made row-wise while MATLAB uses a column-wise memory arrangement. Transposing the problem allows the algorithm to do its operations column-wise taking advantage of MATLAB's memory arrangement. The second modification was changing some conditionals that were checking the existence of a variable containing all gradients to boolean variables, what caused a surprising improve in the performance. Then the sparse product function in line 23 was reworked to use both matrices in column-wise order and the system of equations solving in line 32 was solved directly using the Cholesky decomposition instead of using the default MATLAB solver. Finally, some minor tweaks and structural changes were done to improve performance and code readability.

3.2 C implementation

The authors chose the C programming language because it is much more efficient than MATLAB and has a better performance, despite being more difficult for the developer. The C implementation uses the Intel Math Kernel Library (MKL) which implements the BLAS and LAPACK computational interfaces in a very efficient way for Intel architectures.

The implementation is based on the improved MATLAB implementation and uses all improvements explained in Section 3.1. In this implementation the weight matrix is stored in memory as a full matrix, and the atoms in the active set are controlled by a double linked list of "atoms" for each observation. Each "atom" contains a link to the adjacent active atoms and the index of that atom in the full matrix in memory. Using this strategy the algorithm still can compute the sparse products in lines 7 and 23 without the need of finding the active atoms each time (lines 6 and 22), reducing the computation time needed for the sparse products. When removing active atoms in line 34 the atom should be removed from the atom list of observation t.

The second main improvement is that the sparse product on line 23, the computation of

 ISBN: 978-84-617-8694-7

$R2$ (line 21), the computation of the gradient (line 29) and the computation of the Hessian (line 31) have been mixed in a single computation. All these operations use the same data, so mixing the computations in the proper way diminishes the number of memory accesses and operations.

Finally, the system of linear equations in line 32 has been solved by mean of the LAPACK functions DPOTRF and DPOTRS. The first function computes the Cholesky factorization of a symmetric and positive definite matrix, while the second function uses the factor computed by DPOTRF to solve a triangular system of linear equations. Note that the function DPOTRF is threaded inside the MKL library, that means that in a multicore architecture it will benefit from the multiple cores increasing the algorithm performance. This function is one of the most costly parts of the algorithm, and this is why we do not call sequential to the non-parallel implementation.

3.3 Parallel C implementation

The parallel implementation of the ASNA algorithm takes advantage of the data independence between all the observations. Due to this, all observations can be processed in parallel. For the parallel implementation we used the OpenMP pragma"parallel for" for all loops which iterate along the observations. These loops correspond to lines 4, 7, 18, 19 and 24 . The schedule chosen is dynamic because during the iterative progression of the algorithm the already converged observations are removed from the computations, so the thread that tries to compute an already converged observation will skip it. The dynamic scheduling improves the performance for unbalanced load situations like that.

As said in Section 3.2 the DPOTRF functions is already threaded inside the library, but in the parallel version is going to be called sequentially for each observation. That fact will impact the speedup between both versions.

4 Experimental results

The experimental environment consists of a multicore machine with two Intel Xeon E5-2697 V2 (2,7GHz) with 12 physical cores each and 128 GB RAM. By the software side, the machine has MATLAB R2015b and the Intel parallel studio 2017 (contains icc v17.0.1 and MKL v2017) installed. All the test were executed using the 24 cores available.

In all proposed versions the KL divergence value obtained is the same, and equal to the KL divergence obtained by the original MATLAB implementation. Due to this, we are not going to evaluate the KL divergence value in this paper.

To keep the same structure of the experiments as in the original ASNA paper [1], all implementations were tested with three different dictionary sizes (100, 1000 and 10000) until convergence was achieved. Table 1 shows the execution times of all versions for all

 ISBN: 978-84-617-8694-7

three dictionary sizes and 130 observations at a time. The times had been obtained by averaging 100 different executions.

	Dictionary size		
ASNA algorithm implementation	100	1000	10000
Original MATLAB implementation	1,99	6,23	22,92
Improved MATLAB implementation	1,03	3,93	13,80
C implementation	0,34	1,64	5,96
Parallel C implementation	0,03	0,32	1,75

Table 1: Execution time of each ASNA implementation for different dictionary sizes computing 130 observations (seconds).

5 Discussion

The experimental results show a big improvement in the performance of the algorithms by using the proposed versions. Furthermore, with the parallel version, we can compute the whole model for the biggest dictionary analysed in less time than the needed using the smallest dictionary in the original ASNA version. If only one observation needs to be computed, due to the internal parallelism of the MKL library , the algorithm will still benefit from the multicore architecture.

Non-negative sparse representations have recently been used in many audio processing problems. However, their use in practical applications has been so far limited because of their high computational complexity. In paper we show that that the computational complexity of state-of-the-art ASNA algorithm, which itself is significantly faster than the established EM algorithm, can be reduced by more than 10 times. This makes the algorithm appealing for real-time applications such as speech enhancement.

In future works deeper tests will be done using bigger dictionaries. Furthermore, due to the trivial parallelism of the multiple observations model a GPU version of the algorithm can be implemented to speed up the process even more.

Acknowledgements

This work has been partially supported by Programa de FPU del MECD, by MINCO and FEDER from Spain, under the projects TEC2015-67387- C4-1-R, and by project PROMETEO FASE II 2014/003 of Generalitat Valenciana. The authors want to thank Dr. Konstantinos Drossos for some very useful mind changing discussions. This work has been conducted in Laboratory of Signal Processing, Tampere University of Technology.

References

[1] T. Virtanen J. Gemmeke and B. Raj, *Active-set Newton algo- rithm for overcomplete non-negative representations of audio*, IEEE Transactions on Audio, Speech, and Language Process- ing , vol. 21, no. 11, 2013.

[2] T. Virtanen J. Gemmeke and B. Raj, *Active-set newton algorithm for non-negative sparse coding of audio*, Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on. IEEE, 2014. p. 3092-3096.

[3] B. Raj and P. Smaragdis, *Latent variable decomposition of spectrograms for single channel speaker separation*, Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA 2005), New Paltz, Ny, October 2005.

[4] N. Bertin R. Badeau and E. Vincent, *Enforcing Harmonicity and Smoothness in Bayesian Non-Negative Matrix Factorization Applied to Polyphonic Music Transcription*, IEEE Transactions on Audio, Speech, and Language Processing, vol. 18, no. 3, pp. 538-549, March 2010.

[5] O. Dikmen and A. Mesaros, *Sound Event Detection Using Non-negative Dictionaries Learned From Annotated Overlapping Events*, IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA 2013), New Paltz, NY, 2013.

[6] A. T. Cemgil, *Bayesian inference for nonnegative matrix factorisation models*, Computational Intelligence and Neuroscience, 2009, vol. 2009.

[7] A. Cichocki R. Zdunek A. H. Phan and S. Amari,, *Nonnegative Matrix and Tensor Factorizations.*, Wiley 2009.

[8] Tuomas Virtanen, *Original MATLAB implementation of ASNA algorithm*, http://www.cs.tut.fi/ tuomasv/software.html.

[9] P. San Juan, *Efficient implementations of ASNA algorithm*, https://P.SanJuan@gitlab.com/P.SanJuan/ASNA.git.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

Binary classification based on a quantitative variable – an accuracy comparison by simulation

Rui Santos[1], João Paulo Martins[1], Miguel Felgueiras[2] and Liliana Ferreira[3]

[1] *School of Technology and Management, Polytechnic Institute of Leiria, CEAUL – Center of Statistics and Applications of the University of Lisbon*

[2] *School of Technology and Management and CIGS, Polytechnic Institute of Leiria, CEAUL – Center of Statistics and Applications of the University of Lisbon*

[3] *School of Technology and Management, Polytechnic Institute of Leiria, CMAFCIO – Center for Mathematics, Fundamental Applications and Operations Research*

emails: rui.santos@ipleiria.pt, jpmartins@ipleiria.pt, mfelg@ipleiria.pt, liliana.ferreira@ipleiria.pt

Abstract

The accuracy of a binary classification test based on a quantitative variable is usually measured by the total or the partial area under the receiver operating characteristic (ROC) curve. However, in practice a single cut-off point is used and, therefore, only the accuracy at that specific point is needed. Two different ways of determining the optimal cut-off point are the maximization of the Youden's index and the minimization of the distance between this point and the ideal point, which corresponds to the absence of misclassification. In this work a simulation study is performed in order to compare these measures and highlight the advantages and disadvantages of each procedure.

Key words: Cut-off point, ROC curve, Sensitivity, Specificity, Youden's index.
MSC 2000: 62P10

1 Introduction

A diagnostic test is performed to achieve a binary classification (e.g. healthy versus infected) based on the observed value of a quantitative variable. Thus, if this observed value exceeds a given threshold (the cut-off point), the individual is classified as infected (a positive result).

 ISBN: 978-84-617-8694-7

Otherwise, he is classified as healthy (a negative result). Unfortunately, almost all tests may result in misclassification: the occurrence of false negatives and false positive results. The usual applied measures to evaluate the classification accuracy are the specificity φ_e, which corresponds to the probability of getting a negative result from a healthy individual, and the sensitivity φ_s that is defined as the probability of getting a positive result from an infected individual. Hence, these probabilities depend on the value considered for the cut-off point. Furthermore, when applying the same classification test those measures are inversely correlated, since increasing one of them implies a decrease in the other.

2 The receiver operating characteristic (ROC) curve

The receiver operating characteristic (ROC) curve allows to observe the value of the sensitivity and the specificity when the cut-off point ranges from one end, in which all individuals are classified as infected, to the other extreme where all individuals are classified as healthy (i.e., at every possible cut-off point). Thus, the ROC curve represents the set of points with coordinates $(1-\varphi_e, \varphi_s)$. This curve enables the visualization of the optimal cut-off point in a test, the evaluation of the test's accuracy, as well as the comparison of different tests' performances [3, 4, 9, 16].

The area under the ROC curve (AUC) is usually applied to measure the accuracy of the binary classification method [3, 4, 9, 10, 16, 18]. Nevertheless, usually only a specific cut-off point is used, instead of all its possible values. Therefore, methodologies based in the partial area under the ROC curve (pAUC) were also developed to measure the test's accuracy [1, 7, 8, 15]. Hence, in order to only use the values of interest, those measures use the area under the ROC curve over a range of high specificity (or sensitivity) values to assess the diagnostic accuracy [5].

In the context of compound tests, in [13, 14] was proposed the use of the probability ϕ, which verifies $\varphi_s = \varphi_e = \phi$ for some cut-off point, to measure the test's performance. In the use of count distributions, this value may not exist, therefore, the distance between φ_s and φ_e shall be minimized and $\phi = \frac{\varphi_s + \varphi_e}{2}$. In fact, this measure can be seen as the use of a specific point on the ROC curve, the result of its intersection with the straight line $\varphi_s = \varphi_e$.

3 The optimal cut-off point

The choice of the optimal cut-off point is a decision that depends on several factors, such as the severity of the infection, the risk of not diagnosing the infection, the side effects of the treatment, among others. Hence, it may be important to decide between having a larger sensitivity or a larger specificity.

Nevertheless, considering the absence of clinical factors that led to the choice of one of these measures over the other, the optimal cut-off point can be set by the optimization of

 ISBN: 978-84-617-8694-7

some criterion. Hence, it can be the value that maximizes the Youden's index $\varphi_e + \varphi_s - 1$ [2, 6, 12, 17], which also corresponds to the maximization of the sum $\varphi_e + \varphi_s$, or the value that minimizes the distance to the perfect test $\sqrt{(\varphi_e - 1)^2 + (\varphi_s - 1)^2}$ (minimum distance) [6, 11]. This last procedure will provide the point on the ROC curve that is the closest to the ideal case $(0, 1)$.

In fact, in the application of each test, the focus is the evaluation of the accuracy for a single cut-off point, which shall be the best one for our purposes. Thus, it is indeed critical to realize if a greater value in the previously mentioned accuracy's measures guarantee that the test will generate a better performance at the specific cut-off point used in the application of the diagnostic test.

4 An accuracy comparison by simulation

A simulation study was performed using the R software in order to compare the differences between the area and the partial area under the ROC curve as well as the index ϕ. Moreover, those results were compared with the obtained accuracy in the application of a specific cut-off point, namely when this point is determined by the maximum Youden's index or by the minimum distance to the ideal case. For the test design, different distributions for the characterization of the infected and healthy individuals were considered in the simulations, both discrete and continuous, as well as diverse sample sizes. The main goal is to assess the association between those accuracy measures and, therefore, to assess whether those measures are able to evaluate the same criterion of accuracy. Moreover, the advantages and disadvantages of each measure will be highlighted.

Acknowledgements

Funded by FCT - Fundação Nacional para a Ciência e Tecnologia, Portugal, through the projects UID/MAT/00006/2013 and UID/MAT/04561/2013.

References

[1] L. E. Dodd, M. S. Pepe, *Partial AUC estimation and regression*, Biometrics **59** (2003) 614–623.

[2] R. Fluss, D. Faraggi, B. Reiser, *Estimation of the Youden Index and its associated cutoff point*, Biom J **47** (2005) 458–72.

[3] K. Hajian-Tilaki, *Receiver Operating Characteristic (ROC) Curve Analysis for Medical Diagnostic Test Evaluation*, Caspian J Intern Med **4** (2013) 627–635.

 ISBN: 978-84-617-8694-7

[4] J. A. Hanley, J. J. McNeil, *The meaning and use of the area under a receiver operating characteristic (ROC) curve*, Radiology **143** (1982) 29–36.

[5] Y. Jiang, C. E. Metz, R. M. Nishikawa, *A receiver operating characteristic partial area index for highly sensitive diagnostic tests*, Radiology **201** (1996) 745–750.

[6] X. Liu, *Classification accuracy and cut pointselection*, Stat Med **31** (2012) 2676–2686.

[7] H. Ma, A. Bandos, H. Rockette, D. Gur, *On use of partial area under the ROC curve for evaluation of diagnostic performance*, Stat Med **32** (2013) 3449–3458.

[8] H. Ma, A. Bandos, D. Gur, *On the use of partial area under the ROC curve for comparison of two diagnostic tests*, Biom J **57** (2015) 304–320.

[9] C. E. Metz, *ROC analysis in medical imaging: a tutorial review of the literature*, Radiol Phys Technol **1** (2008) 2–12.

[10] M. S. Pepe, *Statistical Evaluation of Medical Tests for Classification and Prediction*, Oxford University Press, Oxford, 2003.

[11] N. J. Perkins, E. F. Schisterman, *The inconsistency of "optimal" cut-points using two ROC based criteria*, Am J Epidemiol **163** (2006) 670–675.

[12] D. Powers, *Evaluation: From Precision, Recall and F-Score to ROC, Informedness, Markedness & Correlation*, J Mach Learn Tech **2** (2011) 37–63.

[13] R. Santos, J. P. Martins, M. Felgueiras, *An Overview of Quantitative Continuous Compound Tests*, In J. P. Bourguignon *et al.* (Eds.): Dynamics, Games and Science, CIM Series in Mathematical Sciences **1** (2015) 627–641.

[14] R. Santos, M. Felgueiras, J. P. Martins, *Discrete Compound Tests and Dorfman's Methodology in the Presence of Misclassification*, In C. P. Kitsos *et al.* (eds.): Theory and Practice of Risk Assessment, Springer Proceedings in Mathematics & Statistics **136** (2015) 85–98.

[15] S. D. Walter, *The partial area under the summary ROC curve*, Stat Med **24** (2005) 2025–2040.

[16] E. Witten, *Receiver Operating Characteristic (ROC) Curve Analysis for Medical Diagnostic Test Evaluation*, Caspian J Intern Med **4** (2013) 627–635.

[17] W. J. Youden, *Index for rating diagnostic tests*, Cancer **3** (1950) 32–35.

[18] X. H. Zhou, N. A. Obuchowski, D. K. McClish, *Statistical Methods in Diagnostic Medicine*, Wiley & Sons, New York, 2002.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

New tool to teach advanced mathematics

Íñigo Sarría[1], Á. A. Magreñán[1] and L. Orcos[1]

[1] *Escuela Superior de Ingeniería y Tecnología, Universidad Internacional de La Rioja*

emails: inigo.sarria@unir.net, alberto.magrenan@unir.net, lara.orcos@unir.net

Abstract

In this work, a new tool, based on *The Convergence Plane* and other graphical and complex tools, that allows to study real and complex dynamics of iterative methods is presented. This tool can be used, inter alia, to find the elements of a family that have good convergence properties and discard the bad ones or to see how the basins of attraction change along the elements of the family. The uses and results obtained will be commented to show the applicability of the tool.

Key words: Mathematics, e-learning, tool
MSC 2000: 65D10, 65D99, 65G99, 90C30

1 Introduction

In this work we are concerned with the problem of teaching advanced mathematics. The study of iterative methods for solving nonlinear equations is generally complicated for students. To give rise to the behavior of the iterates of a method starting on different initial points we will study the dynamical behavior of an iterative method. The dynamical properties related to an iterative method applied to polynomials give important information about its stability and reliability. Most of one-point iterative methods applied to polynomials generate a rational function. Consequently, we will focus our attention on studying rational functions. One of the main interests in this work is the study of the parameter spaces associated to families of iterative methods, which allows us to distinguish between the good and bad methods of the family in terms of its numerical properties.

 ISBN: 978-84-617-8694-7

2 Basic dynamical concepts

Firstly, we present some dynamical concepts of complex dynamics (see [1, 2, 3, 4, 5, 6, 7] and the references therein for more information). One of the most common problems in Mathematics is solving a nonlinear equation $f(z) = 0$, with $f : \mathbb{C} \to \mathbb{C}$. The solutions of theses equations cannot be solved in a direct way, except in very special cases. That is why most of the methods for solving these equations are iterative. From now on we define a rational function $R : \hat{\mathbb{C}} \to \hat{\mathbb{C}}$, where $\hat{\mathbb{C}}$ is the Riemann sphere. Notice that $R(z) = \frac{P(z)}{Q(z)}$ where $P(z)$, $Q(z)$ are polynomials with complex coefficients without common factors. Moreover, the degree of a rational function $R(z)$ is defined as the highest degree of $P(z)$, $Q(z)$.

We will analyze the phase plane of the map R by classifying the starting points from the asymptotic behavior of their orbits.

Definition 1 *The* orbit of a point $z_0 \in \hat{\mathbb{C}}$ *is defined as*

$$\mathcal{O}(z_0) = \{z_0, R(z_0), R^2(z_0), ..., R^n(z_0), ...\}.$$

In the dynamical study of rational functions, one of the most commonly found problems is to study the behavior of the orbits of a point $z_0 \in \hat{\mathbb{C}}$. If the orbit converges to some value it will be a fixed point of the rational function $R(z)$.

Definition 2 *A point $z_0 \in \hat{\mathcal{C}}$, is called a* fixed point *of $R(z)$ if it verifies that $R(z) = z$.*

There exist different types of fixed points depending on its associated multiplier $\mu = |R'(z_0)|$.

Definition 3 *Taking the associated multiplier into account a fixed point z_0 is called:*

- superattractor *if* $|R'(z_0)| = 0$
- attractor *if* $|R'(z_0)| < 1$
- repulsor *if* $|R'(z_0)| > 1$
- parabolic *if* $|R'(z_0)| = 1$.

The study of ∞ as a fixed point is a little bit different.

Definition 4 *∞ is a fixed point of a rational function $R(z)$ if and only if $z = 0$ is a fixed point of the function:*

$$F : z \to \frac{1}{R(\frac{1}{z})}.$$

Moreover, if ∞ is a fixed point of $R(z)$, the associated multiplier of it is $\mu = F'(0)$

 ISBN: 978-84-617-8694-7

Consequently, ∞ can be attractor or even superattractor. The fixed points of a rational function are special cases of periodic points which are defined as follows.

Definition 5 *A point $z_0 \in \hat{\mathbb{C}}$, is called a periodic point of period p, if $R^p(z_0) = z_0$ and $R^n(z_0) \neq z_0$ for each $n < p$.*

Notice that the orbit associated to a periodic point z_0 of period n has only n different terms.

Definition 6 *The orbit associated to a periodic point of period n is called an n-cycle.*

The multiplier associated to an n-cycle is the same for every point of the cycle

$$|(R^n)'(z_0)| = \cdots = |(R^n)'(z_n)| = |R'(z_0)||R'(z_1)| \cdots |R'(z_n)|.$$

As the fixed points the cycles can be classified by means of the value of the multiplier as

- *superattractor* if $|R'(z_0)| = 0$
- *attractor* if $|R'(z_0)| < 1$
- *repulsor* if $|R'(z_0)| > 1$
- and *parabolic* if $|R'(z_0)| = 1$.

Due to the form of the method, some fixed points can be different to the roots of the polynomial. These points are called *strange fixed points*.

Another important concept in the study of iterative methods is the notion of basin of attraction.

Definition 7 The basin of attraction *of an attracting fixed point α is defined as:*

$$\mathcal{A}(\alpha) = \{z_0 \in \hat{\mathbb{C}} \ : \ R^n(z_0) \to \alpha, \ n \to \infty\}.$$

On the other hand, it is also important the notion of critical point defined as follows.

Definition 8 *Let be $R(z)$ a rational function of degree d. A point $w \in \hat{\mathbb{C}}$ for which the cardinality of $R^{-1}(w)$ is lower than d, is called critical value of $R(z)$. A point $z \in R^{-1}(w)$ which is a root of $R(z) - w$ of multiplicity greater than 1, is called a critical point of $R(z)$. Moreover, ∞ is a critical point of $R(z)$ if 0 is a critical point of $F(z)$. The multiplicity of ∞ as critical point of $R(z)$ is the same as the multiplicity of 0 as critical point of $F(z)$.*

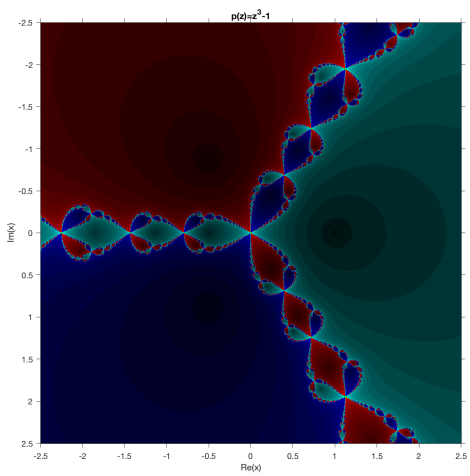
Remark 1 *A point z is a critical point of a holomorphic function $p(z)$ if $p'(z) = 0$.*

 ISBN: 978-84-617-8694-7

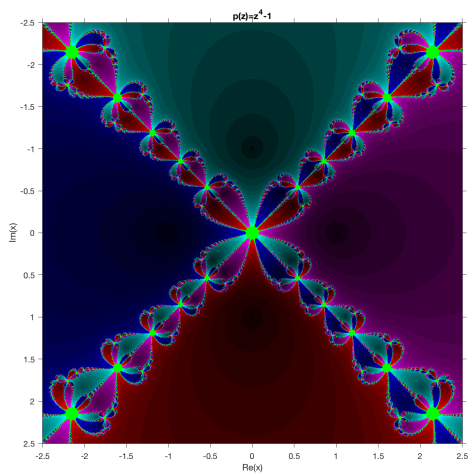
3 Graphical tool

Using all the preceding results we have developed a new tool in which we study the dynamical behavior of different iterative methods applied to polynomials:

- The authors need to introduce the degree of the polynomial, the coefficients and the definition of the iterative method the want to study.
- The program, using connections with the programs MATLAB and MATHEMATICA, gives the polynomial roots, the fixed points and its behavior and the critical points.
- Moreover, the program draws the basins of attraction in a region (which must be introduced previously by the author), as appears in Figure 1.
- On the other hand, in case of studying a family of methods or a family of polynomials, the program also draws the parameter planes.
- Even more, the program gives the possibility to study real and complex dynamics.



$p(z) = z^3 - 1$



$p(z) = z^4 - 1$

Figure 1: Dynamical planes for Newton's method applied to different polynomials.

References

[1] A. F. Beardon *Iteration of rational functions*, Springer-Verlag, New York, 1991.

[2] F. I. Chicharro, A. Cordero, J. R. Torregrosa. Drawing dynamical and parameters planes of iterative families and methods. *The Scientific World Journal*, **Volume 2013** (2013), Article ID 780153, 11 pages.

[3] Á. A. Magreñán *Estudio de la dinámica del método de Newton amortiguado (PhD Thesis)*, Servicio de Publicaciones, Universidad de La Rioja, 2013.

[4] Á. A. Magreñán Different anomalies in a Jarratt family of iterative root-finding methods, *Appl. Math. Comput.*, **233** (2014), 29–38.

[5] J. Milnor: *Dynamics in one complex variable: Introductory lectures. Third edition*, Princeton University Press, Princeton, New Jersey, 2006.

[6] J. F. Traub: *Iterative methods for the solution of equations*, Prentice Hall, Englewood Cliffs, New Jersey, 1964.

[7] J. L. Varona: Graphic and numerical comparison between iterative methods, *The Math. Intel.*, **24** 1 (2002), 37–46.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

A generalized strong Borwein-Preiss variational principle in a complete metric space

Thidaporn Seangwattana[1] **and Somyot Plubtieng**[2]

[1] *Faculty of Science Energy and Environment, King Mongkut's University of Technology North Bangkok, Rayong Campus (KMUTNB), Rayong 21120, Thailand*

[2] *Department of Mathematics, Faculty of Science, Naresuan University, Phitsanulok 65000, Thailand*

emails: thidaporn.s@sciee.kmutnb.ac.th, somyotp@nu.ac.th

Abstract

The purpose of this paper is a generalized strong variant of the Borwein-Preiss variational principle (see in [11]) for which a strong minimality holds in case the divergent series of positive numbers underlying a complete metric space, a generalized distance and a special class of perturbation by following an ideal of Kruger, et al. [9]. Furthermore, our main result extends and strengthens the strong Ekeland and Borwein - Preiss variational principles on a complete metric space in Georgive [6] and Li and Shi [10], respectively.

Key words: Banach spaces complete metric spaces Borwein-Preiss variational principle strong minimizers

1 Introduction

The Ekeland variational principle (for short, EVP) was discovered by Ekeland in 1974 [4]. There were many generalized and equivalent forms of the Ekeland's variational principle in [7, 16, 5, 12, 13, 17, 14]. One of the generalized Ekeland's variational principles is "the Borwein-Preiss variational principle" which was introduced by Borwein and Preiss in 1987 [1]. This principle is an important tool in finite dimensional nonsmooth analysis.

A counterpart of the principle subsequently found by Deville, Godefroy and Zizler in 1993 [3] but their result gives no information about the location of the strong minimizer, and offers no way to identify explicitly a perturbation g in [3]. Later, Li and Shi [10] extended

 ISBN: 978-84-617-8694-7

the Borwein-Preiss variational principle in case the divergent series of positive numbers to a complete metric space, and gave a simpler proof of Theorem 2.5. Very recently, Kruger et al. [9] refine and slightly strengthen the metric space version of the Borwein-Preiss variational principle due to Li and Shi. Moreover, they also clarify the assumptions and conclusions of [10] and [2] and streamline the proofs.

On the other hand, Loewen and Wang [11] constructed in the Banach space setting a special class of perturbations subsuming those used in Theorem 2.5, established strong minimality in the analogue of (1), and gave its applications. Their result is called "a strong variant of the Borwein-Preiss variational principle".

The purpose of this paper is a generalized strong variant of the Borwein-Preiss variational principle (see in [11]) for which a strong minimality holds in case the divergent series of positive numbers underlying a complete metric spacee, a generalized distance and a special class of perturbation by following an ideal of Kruger, et al. [9]. Furthermore, our main result extends and strengthens the strong Ekeland and Borwein - Preiss variational principles on a complete metric space in Georgive [6] and Li and Shi [10], respectively.

2 Preliminaries

We note that X is a complete metric space, E is a Banach space, $\mathbb{N}$ is the set of numbers, and $\mathbb{R}$ is the set of reals.

Definition 2.1 A function $\omega : X \times X \to [0, \infty)$ is called a w-distance on X if the following are satisfied:

1.) $\omega(x, z) \leq \omega(x, y) + \omega(y, z)$ for any $x, y, z \in X$;
2.) for any $x \in X$, $\omega(x, \cdot) : X \to [0, \infty)$ is lower semicontinuous;
3.) for any $\epsilon > 0$, there exists $\delta > 0$ such that $\omega(z, x) \leq \delta$ and $\omega(z, y) \leq \delta$ imply $d(x, y) \leq \epsilon$.

Definition 2.2 [15] Let $f : X \to \mathbb{R} \cup \{+\infty\}$.

a.) A point $\bar{x} \in X$ is said to be a strict minimizer of f if $f(\bar{x}) < f(x)$ for each $x \in X, x \neq \bar{x}$.
b.) A point $\bar{x} \in X$ is said to be a strong minimizer of f if $f(\bar{x}) = \inf_{x \in X} f(x)$ and each minimizing sequence for f is convergent to $\bar{x}$.

It is clear that each strong minimizer of f is also a strict minimizer. But the converse is false. For example, the point $\bar{x} = 0$ is a strict but not a strong minimizer of the function $f(x) := x^2 e^x$ on $\mathbb{R}$; notice that each sequence $\{x_n\}$ tending to $+\infty$ as $n \to \infty$ is a minimizing sequence for f. Recall that $diam(A) := sup\{\omega(x, y) | x, y \in A\}$ denotes the diameter of the set $A \subset X$.

Remark 2.3 [15] For $\epsilon > 0$ let

$$\Sigma_\epsilon(f) := \{x \in X | f(x) < \inf_X f + \epsilon\}.$$

 ISBN: 978-84-617-8694-7

The functional f has a strong minimizer on X if and only if

$$\inf\{diam(\Sigma_\epsilon(f))|\epsilon > 0\} = 0.$$

Theorem 2.4 Let X be a nonempty complete metric space and $\omega : X \times X \to [0, \infty)$ be a w-distance on X and $\omega(x, x) = 0$ for all $x \in X$. Let C_n be a decreasing sequence of nonempty closed sets with $\sup\{\omega(x, y)|\, x, y \in C_n\} \to 0$. Then the intersection of the C_n contains exactly one point:

$$\bigcap_{n=1}^{\infty} C_n = \{x\}$$

for some $x \in X$.

Proof For each k, let $x_k \in C_k$. Since $\sup\{\omega(x, y)|\, x, y \in C_n\} \to 0$ the sequence $\{x_k\}$ is Cauchy sequence. Since X is complete and $\{x_k\}$ is a Cauchy, then we have $x_k \to x$ for some $x \in X$. For any n, the subsequence $\{x_{n+k}\} = \{x, x_{n+2}, \ldots\} \to x$. Since $\{C_n\}$ is decreasing, this subsequence is inside the closed set C_n, so $x \in C_n$. Therefore $x \in \bigcap_{n=1}^{\infty} C_n$. If $x, y \in \bigcap_{n=1}^{\infty} C_n$, then $\omega(x, y) \leq \sup\{\omega(x, y)|\, x, y \in C_n\}$. Since $\sup\{\omega(x, y)|\, x, y \in C_n\} \to 0$, we get $\omega(x, y) = 0$. Thus $x = y$. Hence $\bigcap_{n=1}^{\infty} C_n = \{x\}$

Theorem 2.5 [1] Let E be a Banach space and $f : E \to \mathbb{R} \cup \{+\infty\}$ be a lower semicontinuous function bounded from below. Given $\epsilon > 0, \lambda > 0, p \geq 1$, and $z \in X$ satisfying

$$f(z) \;\leq\; \inf_{x \in E} f(x) + \epsilon.$$

Then, there exist a sequence $\mu_n > 0$, with $\sum_{n=1}^{\infty} \mu_n = 1$, and a point $v \in E$, expressible as the (norm-) limit of some sequence $\{v_n\}$, such that

$$f(x) + \frac{\epsilon}{\lambda} \triangle_p(x) > f(v) + \frac{\epsilon}{\lambda} \triangle_p(v) \quad \forall x \in E \setminus \{v\}, \tag{1}$$

where $\triangle_p(x) := \sum_{n=1}^{\infty} \mu_n \|x - v_n\|^p$. Moreover, $\|x_0 - v\| \leq \lambda$ and $f(v) \leq \epsilon + \inf_{x \in E} f(x)$. When E is a smooth space and $p > 1$, the perturbation function involved in (1) of the above theorem is smooth.

Remark 2.6 When $p = 1$, Theorem 2 essentially recaptures The Ekeland variational principle since $\triangle_1(x) - \triangle_1(v) \leq \|x - v\|$.

Theorem 2.7 [A strong variant of the Borwein-Preiss variational principle][11] Let E be a Banach space, a function $f : E \to \mathbb{R} \cup \{+\infty\}$ be a lower semicontinuous, $\epsilon > 0$, and $x_0 \in E$ be such that

$$f(x_0) < \inf_{x \in E} f(x) + \epsilon.$$

Assume that $\{\mu_n\}_{i=0}^{\infty} \in (0, 1)$ is a decreasing sequence with $\sum \mu_n < +\infty$, and $\rho : E \to [0, +\infty)$ is a continuous function satisfying

$$\rho(0) = 0, \;\; and \;\; \eta := \sup\{\|x\| : \bar{\rho}(x) < 1\} < +\infty. \tag{2}$$

 ISBN: 978-84-617-8694-7

Then, there exist a sequence $\{v_n\}$ in E and a function ρ_∞ such that $\lim_{i\to\infty} v_n = v \in X$ satisfies

(i) $\bar{\rho}(x_0 - v) < 1$,
(ii) $f(v) + \epsilon\rho_\infty(v) \leq f(x_0)$, and
(iii) v is a strong minimizer of $f + \epsilon\rho_\infty$.

3 A generalized strong Borwein-Preiss variational principle in a complete metric space

Now, we introduce a perturbation by shifting and scaling a given continuous function $\rho : \omega(X \times X) \to [0, +\infty)$ on which satisfies for every $x \in X$, and $\lambda > 0$

$$\rho(\omega(x, x)) = 0 \ \textit{and} \ \eta := \sup_{(x,v)\in(X\times X)} \{\omega(x, v) : \rho(\omega(x, v)) < \lambda\} < +\infty, \tag{3}$$

and set

$$\rho_\infty(x, \{v_i\}) := \sum_{i=0}^{\infty} \mu_i(i + 1)\rho(\omega(x, v_i)) \tag{4}$$

where for scalars $\mu_i \in (0, 2)$ and $\{v_i\}_{i=0}^{\infty} \subset X$.

The following theorem is a generalized strong Borwein-Preiss variational principle in case the series $\sum_{i=0}^{\infty} \mu_i$ is diverges underlying a complete metric space, generalized distance and special class of perturbation.

Theorem 3.1 Let (X, d) be a complete metric space, ω be a w-distance with $\omega(x, x) = 0$ such that $t \cdot \omega(X \times X) \subseteq \omega(X \times X)$ for all $t \in \mathbb{N}$, a function $f : X \to \mathbb{R} \cup \{+\infty\}$ be a lower semicontinuous, $\bar{x} \in X$, and $\epsilon > 0$ be such that

$$f(\bar{x}) < \inf_{x\in X} f(x) + \epsilon.$$

Suppose that $\rho : \omega(X \times X) \to [0, +\infty)$ is a continuous function define as in (3). Then for any $\lambda > 0$ and decreasing sequence $\{\mu_i\}_{i=0}^{\infty}$ in $(0, 2)$, there exist a sequence $\{v_i\}$ in X and a function ρ_∞ of the form (4) such that $\lim_{i\to\infty} v_i := v \in X$ satisfies

(i) $\rho(\omega(\bar{x}, v)) < \lambda/\mu_0, \quad (i + 1)\rho(\omega(v, v_i)) < \lambda/2^i\mu_0 \ (i = 1, 2, \ldots)$;
(ii) the series $\sum_{i=0}^{\infty} \mu_i(i + 1)\rho(\omega(v, v_i))$ is convergent and

$$f(v) + \epsilon\rho_\infty(v, \{v_i\}) \leq f(\bar{x}), \tag{5}$$

and v is a strong minimizer of $f + \epsilon\rho_\infty$ on X.
In particular case, if there exists $N \in \mathbb{N}$ such that $\mu_i = 0$ for all $i \geq N$, then

$$f(v) + \epsilon \sum_{i=0}^{N-1} \mu_i(i + 1)\rho(\omega(v, v_i)) \leq f(\bar{x}), \tag{6}$$

 ISBN: 978-84-617-8694-7

and v is a strong minimizer of $f + \epsilon\rho_{N-1}$ on X.

Proof Define sequences $\{v_i\}_{i=0}^{\infty}, \{f_i\}_{i=0}^{\infty}, \{S_i\}_{i=0}^{\infty}$, and $\{\mu_i\}_{i=0}^{\infty}$ is decreasing sequence, inductively starting with $v_0 = \bar{x}, f_0 = f$, and $f_1(x) := f_0(x) + \epsilon(\mu_0 + 1)\rho(\omega(x, v_0))$. By our assumption, we have $f_0(v_0) < \inf_{x\in X} f_0(x) + \epsilon$. Note that $\inf_X f_1 \leq f_1(v_0) = f_0(v_0)$. If $\inf_X f_1 = f_1(v_0) = f_0(v_0)$, we let $v_1 := v_0$. If $\inf_X f_1 < f_1(v_0)$, it follows by the definition of the infimum that there exists $v_1 \in X$ such that

$$\begin{aligned} f_1(v_1) &< \inf_X f_1 + \frac{\lambda(\mu_1 + \frac{1}{2})}{2^2(\mu_0 + 1)}(f_0(v_0) - \inf_X f_1) \\ &= \frac{\lambda(\mu_1 + \frac{1}{2})}{2^2(\mu_0 + 1)} f_0(v_0) + (1 - \frac{\lambda(\mu_1 + \frac{1}{2})}{2^2(\mu_0 + 1)}) \inf_X f_1 \\ &< f_0(v_0). \end{aligned}$$

Set

$$S_0 := \{x \in X | f_1(x) \leq f_1(v_1) + \frac{\lambda(\mu_0 + 1)\epsilon}{2^1(\mu_0 + 1)}\}. \tag{7}$$

Obviously, $v_1 \in S_0$. Since the function f_1 is lower semicontinuous, we get S_0 is closed. Thus, by induction, we have

$$\inf_X f_{i+1} \leq f_{i+1}(v_i) = f_i(v_i), \tag{8}$$

and

$$f_{i+1}(x) := f_i(x) + \epsilon\bar{\mu}_i(i + 1)\rho(\omega(x, v_i)), \forall x \in X, \tag{9}$$

where $\bar{\mu}_i := \mu_i + \frac{1}{i+1}$ for any $i = 0, 1, 2, \ldots$. Therefore, v_{i+1} can be chosen satisfying

$$f_{i+1}(v_{i+1}) \leq \frac{\lambda\bar{\mu}_{i+1}}{2^{i+2}\bar{\mu}_0} f_i(v_i) + (1 - \frac{\lambda\bar{\mu}_{i+1}}{2^{i+2}\bar{\mu}_0}) \inf_X f_{i+1} \leq f_i(v_i), \tag{10}$$

and

$$S_i := \{x \in X | f_{i+1}(x) \leq f_{i+1}(v_{i+1}) + \frac{\lambda\bar{\mu}_i\epsilon}{2^{i+1}\bar{\mu}_0}\}. \tag{11}$$

Since f_{i+1} is lower semicontinuous, the set S_i is closed. Moreover, S_i is nonempty as $v_{i+1} \in S_i$. Since $\bar{\mu}_{i+1}$ is a strictly decreasing sequence, $f_{i+1} \geq f_i$, it follows by (10) that

$$f_{i+1}(v_{i+1}) - \inf_X f_{i+1} \leq \frac{\lambda\bar{\mu}_{i+1}}{2^{i+2}\bar{\mu}_0}(f_i(v_i) - \inf_X f_{i+1}) \leq f_i(v_i) - \inf_X f_i. \tag{12}$$

We have $S_i \subseteq S_{i-1}$ for all $i = 1, 2, \ldots$. In fact, if $x \in S_i$, then we note by (10) that

$$f_i(x) \leq f_{i+1}(x) \leq f_{i+1}(v_{i+1}) + \frac{\lambda\bar{\mu}_i\epsilon}{2^{i+1}\bar{\mu}_0} \leq f_i(v_i) + \frac{\lambda\bar{\mu}_{i-1}\epsilon}{2^{i+1}\bar{\mu}_0} \tag{13}$$

 ISBN: 978-84-617-8694-7

and therefore $x \in S_{i-1}$. We will show that $diam(S_i) \to 0$ as $i \to \infty$. Since $f_{i-1} \leq f_i$, it follows by (10) (with i replaced by $i-1$) that

$$\begin{aligned} f_i(v_i) - \inf_X f_i &\leq \frac{\lambda\bar{\mu}_i}{2^{i+1}\bar{\mu}_0}(f_{i-1}(v_{i-1}) - \inf_X f_i) \\ &\leq \frac{\lambda\bar{\mu}_i}{2^{i+1}\bar{\mu}_0}(f_{i-1}(v_{i-1}) - \inf_X f_{i-1}) < \frac{\lambda\bar{\mu}_i\epsilon}{2^{i+1}\bar{\mu}_0}. \end{aligned} \tag{14}$$

The last $<$ follows from (12) and $f_0(v_0) - \inf_X f_0 = f(\bar{x}) - \inf_X f < \epsilon$. Now, let $x \in S_i$. By the definitions of S_i and f_{i+1}, we obtain

$$\begin{aligned} \bar{\mu}_i\epsilon(i+1)\rho(\omega(x, v_i)) &\leq f_{i+1}(v_{i+1}) - f_i(x) + \frac{\lambda\bar{\mu}_i\epsilon}{2^{i+1}\bar{\mu}_0} \\ &\leq f_{i+1}(v_{i+1}) - \inf_X f_i + \frac{\lambda\bar{\mu}_i\epsilon}{2^{i+1}\bar{\mu}_0}. \end{aligned} \tag{15}$$

By (15) with $f_{i+1}(v_{i+1}) \leq f_i(v_i)$ and (14), it can conclude that

$$(i+1)\rho(\omega(x, v_i)) < \frac{\lambda}{2^i\bar{\mu}_0} < \lambda, \quad \forall i = 0, 1, 2, \ldots \tag{16}$$

The hypothesis (3) therefore implies

$$(i+1)\omega(x, v_i) \leq \eta \Rightarrow \omega(x, v_i) \leq \frac{\eta}{i+1} \tag{17}$$

and so $diam(S_i) \leq \frac{2\eta}{i+1} \to 0$ as $i \to \infty$. By theorem 2, we note that $\bigcap_{i=0}^{\infty} S_i$ contains exactly one point, say v. For each i, we have $v_{i+1} \in S_i$ and $v \in S_i$. Hence $\omega(v_{i+1}, v) \to 0$ as $i \to \infty$. Moreover, by setting $x = v$ in (16), we have $\rho(\omega(\bar{x}, v)) < \frac{\lambda}{\bar{\mu}_0}$ when $i = 0$ and $(i+1)\rho(\omega(v, v_i)) < \frac{\lambda}{2^i\bar{\mu}_0}$ for every $i = 1, 2, \ldots$. Therefore (i) holds.

We now define

$$\tilde{S}_i := \{x \in X | f_{i+1}(x) \leq f_{i+1}(v_{i+1})\} \quad \forall i = 0, 1, 2, \ldots. \tag{18}$$

(ii) Since $f_{i+1} \geq f_i$ and $f_{i+1}(v_{i+1}) \leq f_i(v_i)$, we have $\tilde{S}_i \subseteq \tilde{S}_{i-1}$. Moreover, we note that each $\tilde{S}_i$ is a nonempty closed subset of S_i. By theorem 2, we have $\bigcap_{i=0}^{\infty} \tilde{S}_i = \{v\}$. This together with $f_{i+1}(v_{i+1}) \leq f_i(v_i)$ implies

$$f_n(v) \leq f_n(v_n) \leq f_i(v_i) \leq f_0(v_0) = f(\bar{x}), \quad \forall n > i. \tag{19}$$

From (9), we have

$$f_n(x) = f(x) + \epsilon\sum_{j=0}^{n-1} \bar{\mu}_j(j+1)\rho(\omega(x, v_j)), \quad x \in X. \tag{20}$$

 ISBN: 978-84-617-8694-7

According to (19) and (20), we get

$$\begin{aligned}\sum_{j=0}^{n}\mu_j(j+1)\rho(\omega(v,v_j)) &< \sum_{j=0}^{n-1}\bar{\mu}_j(j+1)\rho(\omega(v,v_j))\\ &= \frac{f_n(v)-f(v)}{\epsilon}\\ &\leq \frac{f(\bar{x})-f(v)}{\epsilon},\end{aligned} \tag{21}$$

for every $n\in\mathbb{N}$. This implies that $\sum_{j=0}^{\infty}\mu_j(j+1)\rho(\omega(v,v_j))$ is convergent. By the definition of ρ_∞ and (20), we conclude that

$$f(v)+\epsilon\rho_\infty(v,\{v_j\})=\lim_{n\to\infty}f_n(v)\leq f_i(v_i)\leq f(\bar{x}). \tag{22}$$

Hence, the condition (5) holds. Let $\tilde{f}:=f+\epsilon\rho_\infty$.

We are going to show that $\Sigma_{\frac{\epsilon\lambda\bar{\mu}_i}{2^{i+1}\bar{\mu}_0}}(\tilde{f})\subseteq S_i$ for each i. Suppose that $x\in\Sigma_{\frac{\epsilon\lambda\bar{\mu}_i}{2^{i+1}\bar{\mu}_0}}(\tilde{f})$. From (20) and (22), we have

$$f_{i+1}(x)\leq\tilde{f}(x)\leq\tilde{f}(v)+\frac{\epsilon\lambda\bar{\mu}_i}{2^{i+1}\bar{\mu}_0}\leq f_{i+1}(v_{i+1})+\frac{\epsilon\lambda\bar{\mu}_i}{2^{i+1}\bar{\mu}_0}. \tag{23}$$

This implies that $x\in S_i$ and therefore $\Sigma_{\frac{\epsilon\lambda\bar{\mu}_i}{2^{i+1}\bar{\mu}_0}})(\tilde{f})\subseteq S_i$ for each i. Moreover, it follows from $diam(S_i)\to 0$ as $i\to 0$ and $\Sigma_{\frac{\epsilon\lambda\bar{\mu}_i}{2^{i+1}\bar{\mu}_0}}(\tilde{f})\subseteq S_i$ that

$$\lim_{i\to\infty}diam(\Sigma_{\frac{\epsilon\lambda\bar{\mu}_i}{2^{i+1}\bar{\mu}_0}}(\tilde{f}))=0 \tag{24}$$

Since the sequence $\bar{\mu}_i$ is strictly decreasing, it follows that the sequence of the closed sets $\Sigma_{\frac{\epsilon\lambda\bar{\mu}_i}{2^{i+1}\bar{\mu}_0}}(\tilde{f})$ is strictly decreasing. Hence, by the Cantor's intersection theorem, $\bigcap_{i=0}^{\infty}\Sigma_{\frac{\epsilon\lambda\bar{\mu}_i}{2^{i+1}\bar{\mu}_0}}(\tilde{f})=\{v\}$. Moreover, it follows from remark 2 that v is a strong minimizer of $\tilde{f}$. Therefore, we can conclude that (ii) is true. In particular case, if there exists $N\in\mathbb{N}$ such that $\mu_i=0$ for all $i\geq N$, then

$$f(v)+\epsilon\sum_{i=0}^{N-1}\mu_i(i+1)\rho(\omega(v,v_i))\leq f(\bar{x}), \tag{25}$$

and v is a strong minimizer of $f+\epsilon\rho_{N-1}$ on X.

When reduced to Banach spaces, it extends and strengthens the strong variant of Borwein-Preiss variational principle on Banach spaces in Loewen and Wang (Theorem 2.7).

Corollary 3.2 Let E be a Banach space, a function $f:E\to\mathbb{R}\cup\{+\infty\}$ be a lower semicontinuous, $\epsilon>0$, and $\bar{x}\in E$ be such that

$$f(\bar{x})<\inf_{x\in E}f(x)+\epsilon.$$

 ISBN: 978-84-617-8694-7

Assume that for any $\lambda > 0, \{\mu_i\}_{i=0}^{\infty} \in (0,2)$ is a decreasing sequence, and $\bar{\rho} : E \to [0, +\infty)$ is a continuous function satisfying

$$\bar{\rho}(0) = 0, \ \ and \ \ \eta := \sup\{\|x\| : \bar{\rho}(x) < \lambda\} < +\infty. \tag{26}$$

Then, there exist a sequence $\{v_i\}$ in E and a function ρ_∞ such that $\lim_{i\to\infty} v_i := v \in X$ satisfies

(i) $\bar{\rho}(\bar{x} - v) < \lambda/\mu_0, \quad (i+1)\bar{\rho}(v - v_i) < \lambda/2^i\mu_0 \ (i = 1, 2, \ldots)$;

(ii) the series $\sum_{i=0}^{\infty} \mu_i \bar{\rho}((i+1)v - v_i)$ is convergent and

$$f(v) + \epsilon\rho_\infty(v) \le f(\bar{x}), \tag{27}$$

and v is a strong minimizer of $f + \epsilon\rho_\infty$ on E.

In particular case, if there exists $N \in IN$ such that $\mu_i = 0$ for all $i \ge N$, then

$$f(v) + \epsilon \sum_{i=0}^{N-1} \mu_i \bar{\rho}((i+1)v - v_i) \le f(\bar{x}), \tag{28}$$

and v is a strong minimizer of $f + \epsilon\rho_{N-1}$ on E.

Proof Setting

$$\rho(\omega(x,y)) := \bar{\rho}(x - y), \quad and \quad \eta := \sup_{(x,v)\in X\times X} \{\omega(x,v) \mid \rho(\omega(x,v)) < \lambda\} \tag{29}$$

for any $x, y \in E$. Note that

$$\begin{aligned} \sup_{(x,v)\in X\times X} \{\omega(x,v) \mid \rho(\omega(x,v)) < \lambda\} &= \sup_{(x,-v)\in X\times X} \{\|x - v\| \mid \|x - v\| < \lambda\} \\ &= \sup_{u\in X} \{\|u\| \mid \bar{\rho}(u) < \lambda\} \\ &= \eta < +\infty. \end{aligned}$$

Then, by Theorem 3.1, we conclude that (i) and (ii) hold.

Next, we give an example for illustrating our main result (Theorem 3.1).

Example 3.4 Let $X = (-\infty, -1] \cup [1, +\infty)$ and $f : X \to \mathbb{R}^+$ be a lower semicontinuous function defined by

$$f(a) := a^2, \ \forall a \in X. \tag{30}$$

Let $\epsilon = \frac{1}{2}, \ \lambda = 1$, and $\bar{x} \in X$. Assume that $\rho : \omega(X \times X) \to [0, +\infty)$ is a continuous function defined by

$$\rho(\omega(b,c)) := \omega(b,c) = || \ b \mid - c \mid \quad forall \ \ b, c \in X. \tag{31}$$

By Theorem 3, for any decreasing sequence $\{\mu_i\}_{i=0}^{\infty}$ in $(0,2)$, there exists a sequence $\{v_i\}_{i=0}^{\infty} \in X$ which $v_i = \sqrt{\bar{x}^2 - \frac{1}{2}} + \frac{\lambda}{2^i\mu_0(i+1)^2}$ converges to some $v = \sqrt{\bar{x}^2 - \frac{1}{2}}$ in X such that

$$\rho(\omega(\bar{x}, v)) = \omega(\bar{x}, \sqrt{\bar{x}^2 - \frac{1}{2}}) = || \ \bar{x} \mid - \sqrt{\bar{x}^2 - \frac{1}{2}} \ | < 1. \tag{32}$$

 ISBN: 978-84-617-8694-7

Suppose that $\mu_i = \frac{1}{i+1}$ for any $i = 0, 1, 2, \ldots$, we have

$$(i+1)\rho(\omega(v, v_i)) = \omega(v, v_i) = (i+1) \mid -\frac{1}{2^i(i+1)^2} \mid < \frac{1}{2^i}. \tag{33}$$

Next, we consider

$$\begin{aligned} f(v) + \epsilon\rho_\infty(v, \{v_i\}) &= (\sqrt{\bar{x}^2 - \frac{1}{2}})^2 + \epsilon \sum_{i=0}^{\infty} \frac{1}{(i+1)}(i+1)\rho(\omega(v, v_i)) \\ &\leq \bar{x}^2 - \frac{1}{2} + \frac{1}{2} = f(\bar{x}). \end{aligned}$$

Let $\tilde{f} := f + \epsilon\rho_\infty$. For every $\bar{x} \in X$, we get

$$f(v) + \epsilon\rho_\infty(v, \{v_i\}) \leq f(\bar{x}) + \epsilon\rho_\infty(\bar{x}, \{v_i\}). \tag{34}$$

Thus $\inf_X \tilde{f} = \tilde{f}(v)$. From (34), we obtain that for any $x \in X$ with $x \neq v$,

$$\begin{aligned} f(v) &< f(x) + \epsilon[\rho_\infty(x, \{v_i\}) - \rho_\infty(v, \{v_i\})] \\ &= f(x) + \epsilon \sum_{i=0}^{\infty} \mu_i(i+1)(\omega(x, v_i) - \omega(v, v_i)) \\ &= f(x) + \epsilon \sum_{i=0}^{\infty} \mu_i(i+1)(\mid\mid x \mid -v_i \mid - \mid\mid v \mid -v_i \mid) \\ &\leq f(x) + \epsilon \sum_{i=0}^{\infty} \mu_i(i+1) \mid\mid x \mid -v \mid \\ &= f(x) + \epsilon \sum_{i=0}^{\infty} \mu_i(i+1)\omega(x, v). \end{aligned} \tag{35}$$

Since $\lim_{i\to\infty} \tilde{f}(v_i) = f(v)$ and inequality (35), it follows that $\tilde{f}(v_i)$ converges to $\tilde{f}(v)$ as $i \to \infty$. By Theorem 3, we have $v_i \to v$ as $i \to \infty$. Hence, we can conclude that a point v is a strong minimizer of $\tilde{f}$. In particular case, if there exists $N \in \mathbb{N}$ such that $\mu_i = 0$ for all $i \geq N$, then

$$\begin{aligned} f(v) + \epsilon \sum_{i=0}^{N-1} \mu_i(i+1)\rho(\omega(v, v_i)) &= (\sqrt{\bar{x}^2 - \frac{1}{2}})^2 + \epsilon \sum_{i=0}^{N-1} \frac{1}{(i+1)}(i+1)\rho(\omega(v, v_i)) \\ &\leq \bar{x}^2 - \frac{1}{2} + \frac{1}{2} = f(\bar{x}). \end{aligned}$$

and v is a strong minimizer of $f + \epsilon\rho_{N-1}$ on X.

 ISBN: 978-84-617-8694-7

Remark 3.5 In the result of Theorem 3.1, it can cover the Borwein-Preiss smooth $\epsilon-$variational principle in [10] by setting

$$p \geq 1,\ \epsilon := \frac{\epsilon}{\lambda^p},\quad \mu_i := \frac{1}{2^{i+1}(i+1)}\quad and\quad \rho(\omega(a,b)) := \omega(a,b)^p \tag{36}$$

with $\omega(x,x) = 0$.

The next corollary gives some direct consequences of the result of Theorem 3.1 when there exists $N \in \mathbb{N}$ such that $\bar{\mu}_i = 0$ for any $i \geq N$

Corollary 3.6 Suppose all the assumptions of Theorem 3.1 and (36) are satisfied. Then there exist v and $\{v_i\}_{i=0}^{\infty}$ are a point and a sequence guaranteed by Theorem 3.1 such that

$$f(v) + \frac{\epsilon}{\lambda^p} \oplus_p (v) \leq f(\bar{x}) \tag{37}$$

and

$$f(v) + \frac{\epsilon}{\lambda^p} \oplus_p (v) < f(x) + \frac{\epsilon}{\lambda^p} \oplus_p (x), \quad \forall\, x \in X \setminus \{v\}, \tag{38}$$

where $\oplus_p(x) = \sum_{i=0}^{N-1} \frac{1}{2^{i+1}}\omega(x,v_i)^p$ and $\omega(x,x) = 0$.

Remark 3.7 Applying the result of Theorem 3.1 by setting

$$\epsilon := \frac{\epsilon}{\lambda},\ \rho(\omega(a,b)) := \omega(a,b),\ and\ \ \mu_i := \frac{1}{2^{i+1}(i+1)}. \tag{39}$$

Looking at the result of Theorem 3.1 (If v is a strong minimizer of $f + \epsilon\rho_\infty$ on X, then v is a strict minimizer of $f + \epsilon\rho_\infty$ on X), we get that for every $x \neq v$

$$\begin{aligned} f(v) &< f(x) + \epsilon[\rho_\infty(x,\{v_i\}) - \rho_\infty(v,\{v_i\})] \\ &= f(x) + \frac{\epsilon}{\lambda}\sum_{i=0}^{\infty}\frac{1}{2^{i+1}}(\omega(x,v_i) - \omega(v,v_i)) \\ &\leq f(x) + \frac{\epsilon}{\lambda}(|\ x - v_i\ | - |\ v - v_i\ |) \\ &\leq f(x) + \frac{\epsilon}{\lambda}(|\ x - v\ |) \\ &= f(x) + \frac{\epsilon}{\lambda}\omega(x,v). \end{aligned} \tag{40}$$

For any sequence v_i satisfying $\lim_{i\to\infty}(f(v_i) + \frac{\epsilon}{\lambda}\omega(v,v_i)) = f(v)$ and inequality (40), it implies that $f(v_i) + \epsilon\rho_\infty(v_i,\{v_i\}) \to f(v) + \epsilon\rho_\infty(v,\{v_i\})$ as $i \to \infty$. By Theorem 3.1, we have $v_i \to v$ as $i \to \infty$. Therefore, v is a strong minimizer for $f + \frac{\epsilon}{\lambda}\omega(\cdot,v)$.

We can establish a corollary which extends and strengthens a result of Georgiev in a complete metric space [6].

Corollary 3.8 Let (X,d) be a complete metric space, $f : X \to \mathbb{R} \cup \{+\infty\}$ be a lower semicontinuous bounded below, $\epsilon > 0$, and $z \in X$. Suppose that

$$f(z) < \inf_{\bar{x}\in X} f(\bar{x}) + \epsilon.$$

 ISBN: 978-84-617-8694-7

Then for every $\lambda > 0$, there exists $v \in \mathbf{B}_\lambda(z)$ is a strong minimizer for the function $x \mapsto f(x) + \frac{\epsilon}{\lambda}\omega(x, v)$.

Acknowledgments The authors would like to thank Naresuan University and King Mongkut's University of Technology North Bangkok. We thank the anonymous referee for the careful reading of the paper and a valuable suggestion improving the presentation.

References

[1] J. Borwein, D. Preiss, *A smooth variational principle with applications to subdifferentiability and to differentiability of convex functions*, Trans. Amer. Math. Soc. **303** (1987), 517–527.

[2] J. Borwein, Q. Zhu *Techniques of Variational Analysis*, Springer Berlin Heidelberg, NewYork, 2005.

[3] R. Deville, G. Godefrory, V. Zizler *Smoothness and renorming in Banach spaces, Pitman monographs surveys*, Pure Appl. Math. **64** (1993), 126–139.

[4] I. Ekeland *On the variational principle*, J. Math. Anal. Appl. **47** (1974), 324–353

[5] C. Farkas, A. *É.* Molnár *A Generalized Variational Principle and Its Application to Equilibrium Problems*, Journal of Optimization Theory and Applications **156(2)** (2013), 213–231

[6] P. Georgirv *The storng Ekeland variational principle, the strong drop theorem and applications*, J. Math. Anal. Appl. **131** (1988), 1–21

[7] T. X. D. Ha *Some Variants of the Ekeland Variational Principle for a Set-Valued Map*, Journal of Optimization Theory and Applications **124(1)** (2005), 187–206

[8] O. Kada, T. Suzuki, W. Takahashi, *Nonconvex minimization theorems and fixed point theorems in complete metric spaces*, Math. Japon. **44** (1996), 381–391

[9] A.Kruger, S. Plubtieng, T. Seangwattana, *Borwein-Preiss variational principle*, J. Math. Anal. Appl. **435** (2016), 1183–1193

[10] Y. Li, S. Shi, *A generalization of Ekeland's ϵ-variational principle and its Borwein-Preiss smooth variant*, J. Math. Anal. Appl. **246** (2000), 308–319

[11] P. Loewen, X. Wang, *A generalized variational principle*, Canad. J. Math. **53(6)** (2001), 1174–1193

 ISBN: 978-84-617-8694-7

[12] J. -P. Penot, *The drop theorem, the petal theorem and Ekeland's variational principle*, Nonlinear Analysis: Theory, Methods & Applications **10(9)** (1986), 813–822

[13] S. Plubtieng, T. Seangwattana, *Generalizations of the strong Ekeland varaitional principle with a generalized distance in complete metric spaces*, Journal of Inequalities and Applications **120** (2013), doi:10.1186/1029-242X-2013-120

[14] S. Plubtieng, T. Seangwattana, *The Borwein-Preiss variational principle for non-convex countable systems of equilibrium problems*, Journal of Nonlinear Science and Applications **9** (2016), 2224–2232

[15] W. Schirotzek, *Nonsmooth Analysis*, Springer-Verlag Berlin Heidelberg, (2007)

[16] J. S. Ume, *Variational Principles, Minimization Theorems, and Fixed-Point Theorems on Generalized Metric Spaces*, Journal of Optimization Theory and Applications **118(3)** (2003), 619–633

[17] Z. Wu, *Equivalent formulations of Ekeland's variational principle* Nonlinear Analysis: Theory, Methods & Applications **55(5)** (2003), 609–615

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

Polyhedron Over-approximation for Complexity Reduction in Static Analysis

Yassamine Seladji[1] and Zheng Qu[2]

[1] *Department of Computer Science, University of Tlemcen*

[2] *Department of Mathematics, The University of Hong Kong*

emails: yassamine.seladji@mail.univ-tlemcen.dz, zheng.qu@maths.hku.hk

Abstract

Polyhedron representation is widely used in the field of static analysis by abstract interpretation, to express the invariant of program. These invariants are used to verify the safety of programs. The used of the polyhedron invariant makes the analysis very expressiveness but also very time consuming. The idea is to find a good trade off between the expressiveness and the time execution. For that, we propose in this article an optimisation method to over-approximate the polyhedron invariant by minimizing the loss of precision.

Key words: Optimisation problem, Polyhedral representation, Static Analysis.

1 Introduction

Static analysis aims at automatically analysing computer program behaviours, without actually executing programs. As a representative example, the numerical safety of a program can be verified by computing the set of values that the variables of the program can reach during its execution, taking into account all possible program inputs. The set is known as the *invariant* of the program. The goal is to be able to declare at the end of the analysis **that nothing bad, caused by the numerical variables, will happen**, like for example **No division by zero** or also **No overflow**.

Abstract interpretation is a widely-used and efficient approach in static analysis. The main advantage of this approach lies in the fact that it allows designing sound static analyser based on the notion of *abstract domains*. The latter can be seen as a mathematical representation of the program behaviours, which must be easily manipulated and stored in

 ISBN: 978-84-617-8694-7

computer memory. Once the type of the abstract domains is fixed, a sequence of (abstract) semantics is used to interprete the program. The semantics must be chosen in accordance with the abstract domains in order to easily manipulate them. The program behaviour, i.e., the set of values that all the program variables can possibly reach, can then be over-approximated by the least fixed-point of the monotone map on the abstract domain defined by the (abstract) semantics of the program [5, 3]. Using Tarski's theorem [15], this fixed-point can be computed as the limit of the iterates of the monotone map starting from the least element.

Unfortunately, due to the over-approximation, an affirmative answer to the numerical safety is not always possible because the over-approximation can possibly add some false alarms. Therefore, in the case where the static analysis returns bugs, it is impossible to know if these bugs are false alarms or not. So, the analysis answer will **be We don't know**. The important point to note here is that the quality of the analysis depends on the quality of the chosen abstract domain. However, the quality of an abstract domain is defined by two parameters: the amount of approximation added to the corresponding concrete element and the efficiency of its operators (the semantics). In other words, this quality is controlled by both the complexity and accuracy of the computation.

A common example of abstract domain is a polyhedral over-approximation of the set of values that program variables can take. For example, the abstract domain of a one-variable (x) program can be an interval $[\min x, \max x]$ where $\min x$ ($\max x$) is the minimal (maximal) value that the variable x can take, though not all the values in $[\min x, \max x]$ are reached in the program. More generally, polyhedron abstract domain [4] is defined by a conjunction of linear constraints of the form

$$\bigcap_{i=1}^{m} \{x \in \mathbb{R}^n : \langle a_i, x\rangle \leqslant b_i\}, \tag{1}$$

where $\langle \cdot, \cdot \rangle$ denotes the standard inner product, $x \in \mathbb{R}^n$ represents the n variables of the program and $\{a_1, \cdots, a_m\} \subset \mathbb{R}^n$ and $\{b_1, \cdots, b_m\}$ fully define the abstract domain.

On one hand, the polyhedral abstract domain returns invariants which express a large set of properties. On the other hand, this expressiveness makes the analysis more complex and expensive: it has worst-case exponential space and time complexity. A lot of efforts have been made to find a good trade-off between expressiveness and efficiency. In [2], a fast version of the polyhedral abstract domain is defined, to decrease the complexity of the analysis. A decomposed version of the polyhedron in used instead of the complete one.
To reduce the expressiveness of the analysis new domains have been developed, which allow to express only a certain kind of linear relations between variables. These domains are known as the weakly relational abstract domains [8, 10, 12, 7, 14]: for example, the octagon abstract domain [10] encodes relations of the form $\pm x \pm y \leqslant c$ for $c \in \mathbb{R}$. In [14], authors define an abstract domain based on supporting functions. The corresponding analysis uses templates,

 ISBN: 978-84-617-8694-7

which represent a set of directions uniformly distributed on the unit sphere. More precisely, the vectors $\{a_i\}$ in (1) are chosen from the unit sphere and the scalars $\{b_i\}$ are determined by the analysis. Indeed, the analysis in [14] computes the templates abstraction of the least fixed point obtained using the polyhedral analysis, which means that the accuracy of this analysis depends strongly on the chosen templates. The authors in [13] propose a novel technique to define a relevant template using the principal components analysis. They combine this technique with the abstract domain defined in [14] to improve its accuracy. To be efficient, the cardinality of the obtained template should be as big as possible. The bottleneck is that the number of constraints m can be so large that the computation and analysis of the resulting polyhedron (1) can be very expensive. Moreover, the number of the linear constraints increases during the analysis, while a minimal representation should be maintained to perform the analysis. For that, the redundancy of each constraint is checked by solving a linear programming problem, which makes the analysis very time consuming. In [9], authors present method to decrease the complexity of the redundancies elimination process by reducing the number of LP problem solved. For that, the method of raytracing is used.

In this paper, we propose a method to reduce the number of linear constraints used to describe the polyhedron while trying to minimize the loss of accuracy. More precisely, given a polyhedral invariant in the form of (1), we want to find $\{i_1, \cdots, i_k\} \subset \{1, \cdots, m\}$ to over-approximate the polyhedral invariant (1) as follows:

$$\bigcap_{i=1}^{m} \{x \in \mathbb{R}^n : \langle a_i, x\rangle \leqslant b_i\} \simeq \bigcap_{j=1}^{k} \{x \in \mathbb{R}^n : \langle a_{i_j}, x\rangle \leqslant b_{i_j}\}. \tag{2}$$

Here k is a number smaller than m. In principle it should be chosen in accordance with the memory limit. The idea is to delete those constraints which constribute the least to the computation of the polyhedron. For this purpose, we randomly generate witness points on the boundary of the polyhedron and define distance functions between the witness points and the halfspaces $\{x \in \mathbb{R}^n : \langle a_i, x\rangle \leqslant b_i\}$. We then formulate a k-median problem, where k is the number of inequalities that we want to maintain, and solve the k-median problem by Jain and Vazirani's approximation algorithm. The output of the optimization algorithm is directly related to k halfspaces, the intersection of which is the overapproximation of the original polyhedron. Afterwards, this method is used as a substitution method of the redundancies elimination process in the case of static analysis by abstract interpretation, to increase its efficiency.

 ISBN: 978-84-617-8694-7

2 Polyhedron over-approximation

2.1 Problem formulation

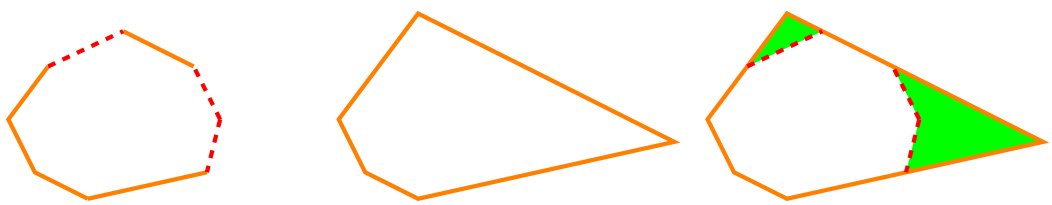
Let $\{a_1, \ldots, a_m\} \subset \mathbb{R}^n$, $\{b_1, \ldots, b_m\} \subset \mathbb{R}$ and $P \subset \mathbb{R}^n$ be the polyhedron defined as the intersection of m halfspaces:

$$P := \bigcap_{i=1}^{m} \{x \in \mathbb{R}^n : \langle a_i, x\rangle \leqslant b_i\}, \tag{3}$$

where $\langle \cdot, \cdot \rangle$ denotes the standard inner product in Euclidean space. We assume throughout the paper that P is bounded and has nonempty interior. In addition, we assume that $\|a_i\| := \sqrt{\langle a_i, a_i\rangle} = 1$ for all $i \in [m]$. Fix some $k \in [m] := \{1, \cdots, m\}$. Let $S \subset [m]$ be a subset of cardinality k. Associated with S we define a new polyhedron $Q(S)$:

$$Q(S) := \bigcap_{i \in S} \{x \in \mathbb{R}^n : \langle a_i, x\rangle \leqslant b_i\}, \tag{4}$$

obtained by removing all the linear constraints not in S. Then $Q(S)$ gives an overapproximation of P, defined as the intersection of k halfspaces. We give a simple illustration in Figure 2.1. The polytope $Q(S)$ shown in Figure 1(b) is obtained by removing the three dashed faces of the polygon P given in Figure 1(a).



(a) Polygon P (b) Polygon $Q(S)$ (c) Polygon $Q(S)$

Figure 1: Example

The approximation error can be measured through the volume difference between $Q(S)$ and P. More precisely, let $\mathrm{vol}(A)$ denote the volume of a polyhedron A, then the volume difference between $Q(S)$ and P is given by $\mathrm{vol}(Q(S)) - \mathrm{vol}(P)$. In the example of Figure 2.1, the volume difference between the polygon $Q(S)$ and P is the volume of the region filled in green in Figure 1(c). Given two subsets S_1 and S_2 of cardinality k, $Q(S_1)$ is preferrable to $Q(S_2)$ if $\mathrm{vol}(Q(S_1)) \leqslant \mathrm{vol}(Q(S_2))$. The "best" approximation of P is then the polyhedron with minimum volume associated with a subset with cardinality k. In other words, we need to solve the following combinatorial optimization problem:

$$\min_{\substack{S \subset [m] \\ |S| = k}} \mathrm{vol}(Q(S)) - \mathrm{vol}(P). \tag{5}$$

 ISBN: 978-84-617-8694-7

Although the solution of optimization problem (5) returns the best overapproximation in terms of minimum volume difference, we will not attempt to solve it directly. In fact, even the volume computation is known to be a #-complete problem [6]. Instead, we will formulate in the following a discrete approximation of (5), which can be solved to some extent by polynomial-time algorithms.

2.2 Distance functions

For each $j \in [m]$ we define the hyperplane

$$H_j := \{x \in \mathbb{R}^n : \langle a_j, x\rangle = b_j\}.$$

The boundary of P is the set of points in P which intersect with at least one hyperplane, i.e.,

$$\mathrm{bd}(P) = \bigcup_{i=1}^{m} \left(P \bigcap H_i\right).$$

For any $x \in \mathrm{bd}(P)$ and $j \in [m]$, define the projection operator:

$$\Pi_j(x) := \arg\min\{\|x - y\| : y \in H_j\}.$$

That is, $\Pi_j(x)$ is the point on the hyperplane H_j which is closest to x. Now we define the projective distance of x to the hyperplane H_j as follows:

$$p_j(x) := \|x - \Pi_j(x)\|.$$

It is easy to check that:

$$p_j(x) = b_j - \langle a_j, x\rangle, \;\; j \in [m], x \in \mathrm{bd}(P).$$

The relative interior of $\mathrm{bd}(P)$ is the set of points in P which intersects with exactly one hyperplane, i.e.,

$$\mathrm{ri}(\mathrm{bd}(P)) = \bigcup_{i=1}^{m} \left(P \bigcap H_i \backslash \left(\bigcup_{j \neq i} H_j \right) \right).$$

Therefore, for $x \in \mathrm{ri}(\mathrm{bd}(P))$, there is a unique index $i \in [m]$ such that $x \in H_i$. For any $j \in [m]$, we then define the inverse projection operator as follows:

$$\Pi_j^{-1}(x) := \{y \in H_j : \Pi_i(y) = x\}.$$

That is, $\Pi_j^{-1}(x)$ is the point in H_j whose projection onto the hyperplane H_i is x. For any $x \in \mathrm{ri}(\mathrm{bd}(P))$ and $j \in [m]$, we define the inverse projective distance $\delta_j(x)$ as follows.

$$\delta_j(x) := \|x - \Pi_j^{-1}(x)\|.$$

 ISBN: 978-84-617-8694-7

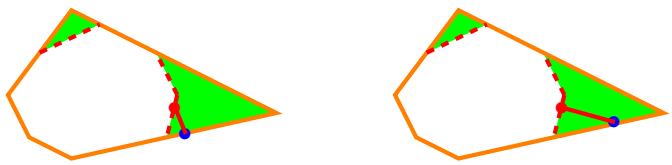
It can be checked that:

$$\delta_j(x) := \frac{b_j - \langle a_j, x\rangle}{\max\left(\langle a_i, a_j\rangle, 0\right)}, \quad j \in [m], x \in \mathrm{ri}(\mathrm{bd}(P)) \cap H_i. \tag{6}$$

Since $\|a_i\| = 1$ for any $i \in [m]$, we clearly have

$$\delta_j(x) \geqslant p_j(x) \geqslant 0, \quad \forall x \in \mathrm{ri}(\mathrm{bd}(P)), \quad j \in [m]. \tag{7}$$

In Figure 2.2 we give an illustration of the projective distance $p.(\cdot)$ and the inverse projective distance $\delta.(\cdot)$. Note that the blue point in Figure 2(a) is the projection of the red point onto the hyperplane which contains the blue point. The red point in Figure 2(b) is the projection of the blue point onto the hyperplane which contains the red point.



(a) blue=Π_j(red) (b) blue=Π_j^{-1}(red)

Figure 2: Illustration of projection and inverse projection operators.

2.3 Volume difference approximation

The normal cone to P at $x \in P$ can be written as [11, Theorem 6.14]:

$$N_P(x) = \left\{ \sum_{i=1}^{m} y_i a_i : y_i \geqslant 0, y_i(\langle a_i, x\rangle - b_i) = 0, \quad \forall i \in [m] \right\}. \tag{8}$$

As an illustration, the normal cone of the polygon P in Figure 1(a) is the region filled with blue lines in Figure 3(a). Now define

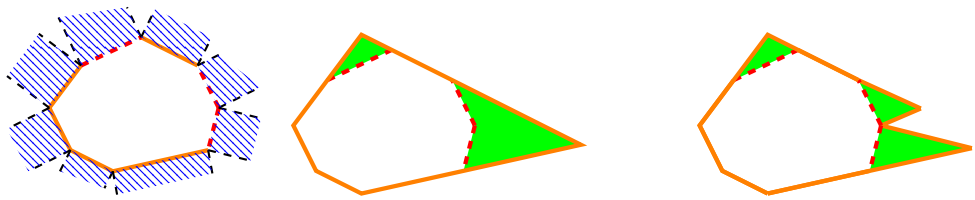
$$O(S) = Q(S) \backslash \left(\bigcup_{x \in \mathcal{F}_{n-2}(P)} \{x + N_P(x)\} \right),$$

where $\mathcal{F}_{n-2}(P)$ denotes the set of faces with dimension $n-2$. We give in Figure 3(c) an illustration of $O(S)$.

Proposition 1. *We have*

$$\mathrm{vol}(O(S) \backslash P) = \int_{\mathrm{ri}(\mathrm{bd}(P))} \min_{j \in S} \delta_j(x) dx \tag{9}$$

 ISBN: 978-84-617-8694-7



(a) $N(P)$ (b) $Q(S)$ (c) $O(S)$

Figure 3: Illustration of the normal cones and $O(S)$.

Remark 1. *The integration in* (9) *can be decomposed into the sum of integrations over faces of* P*, each of which is reduced to a standard integration on a compact domain in* $\mathbb{R}^{n-1}$*.*

The proof of Proposition 1 is omitted due to the page limit. We propose to approximate problem (5) as follows:

$$\min_{\substack{S\subset[m]\\|S|=k}} \int_{\mathrm{bd}(P)} \min_{j\in S} d(j,x)dx \tag{10}$$

where the integral $d(j,x)$ can be replaced by the projective distance $p_j(x)$ or the inverse projective distance $\delta_j(x)$. Note that by (7), we have:

$$\int_{\mathrm{bd}(P)} \min_{j\in S} p_j(x)dx \leqslant \int_{\mathrm{ri}(\mathrm{bd}(P))} \min_{j\in S} \delta_j(x)dx \leqslant \mathrm{vol}(Q(S)\backslash P). \tag{11}$$

2.4 Integration approximation

To obtain an approximation of the integration in (10), we propose to generate a discrete set X of "representative" points and approximate as follows:

$$\sum_{x\in X} \min_{j\in[m]} d_j(x) \simeq \int_{\mathrm{bd}(P)} \min_{j\in[m]} d_j(x) \tag{12}$$

A natural choice of X appears to be a subset of points uniformly distributed over $\mathrm{bd}(P)$. For this purpose, we apply the *running shake and bake* algorithm introduced in [1].

2.5 k-median problem

Let X be a discrete subset of P and $d(\cdot,\cdot) : [m] \times \mathbb{R}^n \to \mathbb{R}_+$ be a distance function. We propose to solve the following discrete approximation problem of (5).

$$\min_{\substack{S\subset[m]\\|S|=k}} \sum_{x\in X} \min_{j\in S} d(j,x). \tag{13}$$

 ISBN: 978-84-617-8694-7

In view of the previous discussion, we know that if X approximates the uniform distribution over $\mathrm{bd}(P)$ and $d(\cdot,\cdot)$ is either the projection distance $p_\cdot(\cdot)$ or the inverse projection distance $\delta_\cdot(\cdot)$, then (13) provides an approximation of the original problem (5).

We recognize in (13) the k-median problem. Indeed, here we identify the discrete set X as cities and the hyperplanes $H_1, \cdots, H_m$ as facilities and the distance between a city $x \in X$ and a facility H_j is $d(j, x)$. We then apply the algorithm of Jain and Vazirani , designed for approximately solving k-median problems in polynomial time.

3 Numerical Experiments

To show the efficiency of our optimisation method, we use it in the field of static analysis by abstract interpretation to improve the analysis process. Especially, when the analysis is performed using the polyhedra abstract domain, this is known as the polyhedral analysis. Our optimisation method is applied to decrease the complexity of the polyhedral analysis, by over-approximating the obtained polyhedron fixed-point.

We include the optimisation method in our static analyser, then we apply it to analyse several programs. The used programs contain a number of stable linear systems and digital filters, known to be hard to analyse using the polyhedral analysis. The experimentations are done on $2.4GH_Z$ Intel Core2 Duo laptop, with $8Gb$ of RAM. To show the quality of the obtained over-approximation, we used Matlab [1] to compute the volume of the obtained polyhedra. The results are summarised in Table 1. Table 1 shows the volume of the initial polyhedron, it is the polyhedron obtained using the polyhedral analysis. The volume of the over-approximated polyhedron is given using two metrics the projective distance and the inverse projective one. It is the polyhedron obtained using the polyhedral analysis mixed with the optimisation method. Note that, N represents the number of constraints of the initial polyhedron. And $K \leqslant N$ is the number of constraints of the over-approximated polyhedron. We have also that $|V|$ is the number of programs variables, we use its to show the scalability of our method. The results of Table 1 show that the optimisation method, in the most of cases, does not add a big amount of over-approximation. To illustrate the precision of our analysis, we display on Figure 4 and Figure 5 the result given in Table 1 for, respectively, the example `filter2` and `Harmonic_oscillator`. Note that, figures are given in dimension 2. The results of the polyhedral analysis, given by the red polyhedra, and the one obtained using the polyhedral analysis with the optimisation method, is given by the blue polyhedra. Note that, the red polyhedra are contained into the blue polyhedra, this shows the quality of the invariant we compute.

[1] https://mathworks.com/products/matlab.html

 ISBN: 978-84-617-8694-7

```
double input() {
  double u = 10.0;
  double l = 0.0;
  return (rand()/RAND_MAX)*(u-l)+l;
}

void main() {
  double xn,yn,ynm1,ynm2;
  xn=xnm1=xnm2=yn=ynm1=ynm2=0;
  int i=0;
  while (i<4000) {
    yn=xn+0.5*ynm1-0.45*ynm2;
        ynm2=ynm1; ynm1=yn;
        xn = input();
}}
```

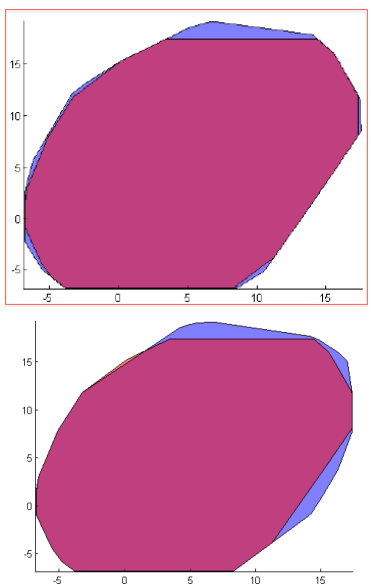


Figure 4: (Left)The body of the program called *Filter2*.(right top) The red polyhedron is the post fixed point obtained using polyhedral analysis and the blue polyhedron is the one obtained using the optimisation method with the projective distance. (right bottom) The red polyhedron is the post fixed point obtained using polyhedral analysis and the green polyhedron is the one obtained using the optimisation method with the inverse projective distance.

 ISBN: 978-84-617-8694-7

```
double input() {
  double u = 1.0;
  double l =  0.0;
  return (rand()/RAND_MAX)*(u-l)+l;
}

void main() {
  double x, xn, y,yn;
  xn=input();
  yn=input();
  x=y=0;
  int i=0;
  while (i<10000){
        x = 0.95 *xn + 0.09975 * yn;
        y = -0.1 * xn + 0.95 * yn;
        xn = x;
        yn = y;
}}
```

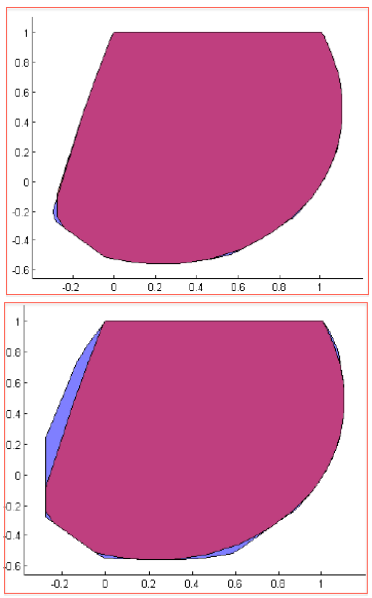


Figure 5: (Left)The body of the program called *Harmonic oscillator*.(right top) The red polyhedron is the post fixed point obtained using polyhedral analysis and the blue polyhedron is the one obtained using the optimisation method with the projective distance.(right bottom) The red polyhedron is the post fixed point obtained using polyhedral analysis and the blue polyhedron is the one obtained using the optimisation method with the inverse projective distance.

 ISBN: 978-84-617-8694-7

Table 1: Table of volumes of the obtained polyhedra.

Program		The initial polyhedron		The over-approximated polyhedron		
Name	$\|V\|$	N	volume	K	Inverse Projective	Projective
`filter2`	4	332	35.91	221	39.03	36.62
`Linear_quadratic_gaussian`	7	398	628.14	265	851.149	685.78
`Observer_based_controller`	10	500	10796.84	333	11389.06	14710.72
`Butterworth_low_pass_filter`	9	542	3293.69	361	4434.05	4286.32
`Dampened_oscillator`	4	332	31.78066	221	36.6896	41.74
`Harmonic_oscillator`	6	332	243.07	221	272.71	300.58

References

[1] J. F. McDonald A. H. G. Rinnooy Kan H. E. Romeijn R. L. Smith J. Telgen C. G. E. Boender, R. J. Caron and A. C. F. Vorst. Shake-and-bake algorithms for generating uniform points on the boundary of bounded polyhedr. *Operations Research*, 39(6):945–954, 1991.

[2] Giuseppe Castagna and Andrew D. Gordon, editors. *Proceedings of the 44th ACM SIGPLAN Symposium on Principles of Programming Languages, POPL 2017, Paris, France, January 18-20, 2017*, 2017.

[3] P. Cousot and R. Cousot. Comparing the Galois connection and widening/narrowing approaches to abstract interpretation, invited paper. In *PLILP*, 1992.

[4] P. Cousot and N. Halbwachs. Automatic discovery of linear restraints among variables of a program. In *Symposium on Principles of Programming Languages*, 1978.

[5] Patrick Cousot and Radhia Cousot. Abstract interpretation: a unified lattice model for static analysis of programs by construction or approximation of fixpoints. In *Symposium on Principles of Programming Languages*, 1977.

[6] M. E. Dyer and A. M. Frieze. On the complexity of computing the volume of a polyhedron. *SIAM Journal on Computing*, 17(5):967–974, 1988.

[7] Eric Goubault, Sylvie Putot, and Franck Védrine. Modular static analysis with zonotopes. In *Static Analysis - 19th International Symposium, SAS 2012, Deauville, France. Proceedings*, 2012.

 ISBN: 978-84-617-8694-7

[8] Vincent Laviron and Francesco Logozzo. Subpolyhedra: a family of numerical abstract domains for the (more) scalable inference of linear inequalities. *STTT*, 2011.

[9] Alexandre Maréchal and Michaël Périn. Efficient elimination of redundancies in polyhedra by raytracing. In *Verification, Model Checking, and Abstract Interpretation - 18th International Conference, VMCAI 2017, Paris, France, Proceedings*, 2017.

[10] Antoine Miné. The octagon abstract domain. *Higher-Order and Symbolic Computation*, 2006.

[11] R. Tyrrell Rockafellar and Roger J.-B. Wets. *Variational analysis*, volume 317 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, Berlin, 1998.

[12] Sriram Sankaranarayanan, Henny B. Sipma, and Zohar Manna. Scalable analysis of linear systems using mathematical programming. In *VMCAI*, 2005.

[13] Yassamine Seladji. Finding relevant templates via the principal component analysis. In *Verification, Model Checking, and Abstract Interpretation - 18th International Conference, VMCAI 2017, Paris, France, Proceedings*, 2017.

[14] Yassamine Seladji and Olivier Bouissou. Numerical abstract domain using support functions. In *NASA Formal Methods, 5th International Symposium, NFM 2013, Moffett Field, CA, USA. Proceedings*, 2013.

[15] A. Tarski. A lattice-theoretical fixpoint theorem and its applications. *Pacific Journal of Mathematics*, 1955.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

Acceptance Tail Method for Sampling from Unimodal Distributions

Efraim Shmerling[1]

[1] *Department Mathematics, Ariel University,Science Park,44837 Ariel,Israel*

emails: efraimsh@ariel.ac.il

Abstract

A universal method for generating continuous random variables (rv's) with unbounded range and infinite-valued discrete rv's is presented.The validity of the method is proved . It is shown in some detail how the method can be implemented for generating unimodal continuous rv's. Theorem that enables one estimate the efficiency of the unimodal version of the presented method is formulated.

Key words: random veriable generation,ziggurat algorithm, unimodal distribution

1 Introduction

A new efficient method named acceptance tail (AT) for generating rv's is presented in the paper. Different versions of the AT method have been developed: a discrete version, a multivariate continuous version, a multimodal univariate continuous version and the univariate unimodal continuous version which will be called the $UMAT$ method in further discussion. The development of these versions shows that the AT method like the "classical" acceptance rejection (AR) method extends to all types of rv's except multivariate rv's of high order. In Section 2 the general AT method is presented. In Section 3 which is the most important part of the paper the $UMAT$ method is described in some detail. For the case where the probability density function (pdf) is monotonically decreaing or is a symmetric unimodal function, a very efficient in terms of generation time ziggurat algorithm has been developed by Marsaglia and Tsang (see [1,2]), and became the standard generator for Gaussian rv's in many platforms. However the ziggurat algorithm has not been implemented for distributions other than Gaussian and exponential, primarily due to complicated and time-expensive setup and the fact that is does not extend to nonmonotonic and nonsymmetric

 ISBN: 978-84-617-8694-7

distributions. In section 3 of the paper it is shown how the presented UMAT method overcomes these drawbacks of the ziggurat algorithm while retaining its main advantage which is high computational speed.

2 DESCRIPTION OF THE GENERAL AT METHOD

In order to describe the general AT method we introduce the following notations. Let X designate the "target" rv . Let the term: "density region" denote in the case where X is a continuous rv the region under the graph of the pdf f(x) of X , and the union of vertical intervals which extend from $(x_i; 0)$ to $(x_i; p(x_i)), x_i \in R, i = \overline{1, \infty}$ in the case where X is a discrete rv. Here p(x) designates the probability mass function of X. Let H designate in the case where X is a continuous rv a bounded part of R such that the integral $\int_H f(x)dx = 1-\delta$ is very close to one , and in the discrete case a finite subset of R such that the sum

$$\sum_{i:x_i \in H} p(x_i)$$

is very close to one. Let T designate the complement of R with respect to H. Let C designate the "head" of the density region which is the part of the density region over H, and let D designate the "tail" of the density region which is the part of the density region over T. Let V(A) designate the volume of a domain A in the case where $A \subset R^k, k \geq 2$, square of A in the case where $A \subset R^2$, and the total length of the "vertical" intervals included in A in the case where A is a union of "vertical" intervals.

The AT method is based on covering C by a set of m instrumental units L_i, $1 \leq i \leq m$ which are "vertical" intervals with common length $1/n$ in the discrete case, rectangles with common area $1/n$ in the univariate continuous case or hyperrectangles with common volume $1/n$ in the multivariate case, $n \geq m$. An algorithm for defining the instrumental units is a part of any specific version of the AT method. Besides defining the set of instrumental units, an algorithm for generating the "tail" rv G needs to be chosen at the preliminary stage of AT. The pdf $g(x)$ of G is defined by the formula: $g(x) = f(x)/\delta$, $x \in T$ in the case where X is continuous, and the pmf $g(x)$ of G is defined by the formula : $g(x_i) = p(x_i)/\delta$, $x_i \in T$ for the case where X is discrete. Note that the efficiency of the algorithm for generating G does not affect significantly the efficiency of the AT method since the probability that the generation of G will be required in order to obtain a generated value of X is very low as shown below. In order to define the set of instrumental units respective tables need to be initialized at the setup stage of the AT method. Provided that such tables have been initialized and the algorithm for generating G chosen, the generation of X looks as follows.

1. Select randomly an integer i in the range $1 : n$ (thus one of the covering elements L_i is chosen).

 ISBN: 978-84-617-8694-7

2. If $i \leq m$, generate a point $P = (P1; P2)$ uniformly distributed in L_i; otherwise go to step 4.

3. If P belongs to the density region of X, set $X = P1$.

4. Generate G and set $X = G$.

Obviously, by the presented method a point P is chosen from the density region of X and $P1$ accepted as a generated value of X. (At step 4 a point P is selected from D by the AR method, and P1 is accepted as a generated value of X). Thus in order to prove the validity of the AT method it is enough to show that for any domain A which belongs to the density region of X the probability P(A) that the selected by the AT method point P belongs to A equals V(A). In order ro prove the assertion, let's first consider the case where

$$A \subset \widehat{L_i} = L_i \bigcap C, 1 \leq i \leq m$$

for some $i, 1 \leq i \leq m$. In this case we have

$$P(A) = P(B_i) * P(A|B_i) = (1/n) * (V(A)/(1/n)) = V(A),$$

where the event B_i is selecting the integer i at step 1. Now let's consider the case where $A \subset D$. In this case in order to show that $P(A) = V(A)$ it is enough to show that the probability $P(D)$ that step 4 will be executed in generating a sampled value of X by the AT method equals $V(D) = q$. Obviously, P(D) can be expressed as the sum $(n-m)/n + \sum_1^m P_i(D)$, where the first term is the probability that an integer larger than m will be chosen at step 1, and $P_i(D)$ is the probability that a given integer i less or equal to m will be chosen at step 1 and the point P generated at step 2 will not belong to $\widehat{L_i}$. The probability $P_i(D)$ equals $(1/n) * ((1/n - V(\widehat{L_i}))/(1/n)) = 1/n - V(\widehat{L_i})$. It follows that the sum $\sum_1^m P_i(D)$ equals $\sum_1^m (\frac{1}{n} - V(\widehat{L_i})) = \frac{m}{n} - \sum_1^m V(\widehat{L_i}) = \frac{m}{n} - V(C) = \frac{m}{n} - (1-q)$, therefore P(A)=P(D). Q. E. D.

3 THE UNIVARIATE UNIMODAL VERSION OF THE AT METHOD

Let's consider the case where the pdf $f(x)$ of the "target" rv X is monotonically decreasing with mode $M = 0$. n horizontal rectangular layers $L_i, 1 \leq i \leq n$ with common area $1/n$ need to be obtained according to the procedure described below, a layer i extends from 0 to x_i horizontally and from y_i to y_{i-1} vertically, $x_0 = 0, y_0 = f(0), x_{1-1} \leq x_i, y_i < y_{i-1}, 1 \leq i \leq n$. The obtained instrumental units define H as $H = \{x : 0 \leq x \leq x_n\}$.

At the setup stage of the $UMAT$ method 2 tables: $\{x_i\}_{i=1}^n, \{y_i\}_{i=1}^n$ have to be generated. Let G designate the "tail" rv with pdf $g(x) = f(x)/\delta, x > x_n$. Then the generation stage of the $UMAT$ method looks as follows.

 ISBN: 978-84-617-8694-7

1. Select randomly an integer i in the range $1 : n$.
2. Generate U_0 uniformly distributed in $(0, 1)$ and calculate $s = U_0 * x_i$.
3. If $s < x_i$, set $X = s$.
4. Generate U_1 uniformly distributed in $(0, 1)$ independently of U_0.
5. If $U_1 < (f(s) - y_i)/(y_{i-1} - y_i)$, set $X = s$;
6. generate G and set $X = G$.

The initialization of the tables required by the $UMAT$ method looks as follows. First the values of $x_i\,, y_i\; i = 1, 2, ...$ are calculated recursively utilizing the equalities:

$$(f(x) - y_{i-1}) \cdot x_i = \frac{1}{n},\; y_i = f(x_i),\; x_0 = 0,\; y_0 = f(0). \tag{1}$$

The consecutive values of x_i are calculated until the point x_{i^*}, for which the condition $S(i^*) = \sum_{j=1}^{i^*}(x_j - x_{j-1}) \cdot f(x_{j-1}) > 1$ is satisfied, is obtained. Since $S(i^* - 1) < 1$ and $S(i^*) > 1$, the following inequality holds: $x_{i^*-1} < d = (n - i^* + 1)/(n \cdot f(x_{i^*-1})) < x_{i^*}$. We set $x_i = d, i^* \le i \le n$, and we set $y_i = f(x_i), 1 \le i < i^*$, $y_i = f(x_{i^*-1}/(n-i^*+1) \cdot (n-i), i^* \le i \le n$.

Note that the generation of the tables requires calculating a root of the equation (1) less than n times. Ziggurat method requires the calculation of the common area A of the layers utilizing an iteration method (such as the bisection method), and at each iteration the equation $f(x) = y_i$, where $y_i = y_{i-1} + A/x_{i-1}$ has to be solved for each i, $1 < i \le n$, i. e. $n - 1$ times, from which follows that the setup time required by the $UMAT$ method is many times shorter than the setup time of the ziggurat. The main advantage of the $UMAT$ method is its computational speed which follows from the fact that the probability $P^*(n)$ that only the first 3 steps of the algorithm will be executed in order to obtain a generated value of the "target" rv X is close to one for sufficiently large n. In order to estimate $P^*(n)$ we introduce the following notation. Let $E_i, 1 \le i \le i^*$ designate the rectangular part of L_i bounded on the left by the line $x = x_i$, and let $S(A)$ designate the area of an arbitrary region A in R^2. Then $P^*(n)$ can be calculated via the following formula: $P^*(n) = 1 - S(U_2^{i^*} E_i) - 1/n$. The following theorem proves that $P^*(n)$ tends to one as n tends to infinity and enables one to estimate the rate of convegence.

Theorem 1

If the derivative of $f(x)$ is bounded : $\mid f^{'}(x) \mid \le M, x > 0$, then for any $\delta > 0$ and any n which satisfies the inequality

$$\frac{2.5 + ln(\sqrt{Mn} \cdot x_\delta)}{n} < \delta$$

 ISBN: 978-84-617-8694-7

the following inequalities hold: $x_n > x_\delta$ and $S(U_1^{i^*} E_i) < 2\delta$. Here x_δ desinates the point on the x-axis for which the equality $\int_{x_\delta}^{\infty} f(x)dx = \delta$ is satisfied.

Utilizing the theorem, the proof of which will be presented in the full version of the paper, we have shown that for gamma distributions and many other widely used distributions the rate of convergence of $P^*(n)$ to one with growth of n is very high, therefore the UMAT method can be implemented very successfully in sampling from these distributions.

References

[1] L. Devroye, *Non-Uniform Random Variate Generation*,(Springer-Verlag,New York,1986),Ch.9(3). generators", Appl. Stat. 28,290-295 (1979).

[2] G. Marsaglia and W.W. Tsang, "The ziggurat method for generating random variables", Journal of Statistical Software 5,17 (2000),www.jstatsoft.org.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

Selection and optimized sizing methodology of logistics associated with hydrogen systems

Juan Antonio Sicilia[1], Alberto Fraile[2], Emilio Larrodé[2], Lara Orcos[1] and J. Javier Rainer[1]

[1] *Universidad Internacional de La Rioja, Avda de La Paz, 137, 26006 Logroño, La Rioja, Spain,*

[2] *Engineering and Architecture Center, Zaragoza University, María Luna 3, 50018 Zaragoza, Spain,*

emails: juanantonio.sicilia@unir.net, afrailep@unizar.es, elarrode@unizar.es, lara.orcos@unir.net, javier.rainer@unir.net

Abstract

The incorporation of hydrogen into transport systems means not only the incorporation of vehicles with new technologies, but the creation of a new infrastructure that encompasses production, storage, transport and distribution, spearheaded by the refueling hydrogen stations. For this reason, it is almost essential to develop a methodology for selection and optimized sizing of the logistics associated with hydrogen systems. Given this situation, this study has focused on the need to create a methodology that will help to solve the dimensioning of the hydrogen refueling facilities and the logistics involved.

Key words: Hydrogen, Logistics, Fueling station, Sizing methodology, Optimization

1 Introduction

Transportation is an activity derived from other activities taking place in a given geographical area, a country, a region, a city or a neighborhood [9]. Its most evident manifestation is the urban traffic, that is to say, the people and vehicles circulation by the public spaces of the city. Starting from this definition, a series of determinants are presented as responsible for the evolution that has led to the urban transport to the point of transition in which it is at present. On the one hand, transport has to adapt to the available resources and the topology of cities, but also to a series of problems that are the result of uncontrolled growth; congestion, pollution and limitation of available resources makes inevitably the need for new alternative transport systems [1].

The incorporation of hydrogen into transport systems means not only the incorporation of vehicles with new technologies, but the creation of a new infrastructure that encompasses production, storage, transport and distribution, spearheaded by the refueling hydrogen stations. The extent to which emission reduction will occur depends on the mix of technologies that make up the hydrogen supply chain [2, 10, 13]. For this reason, it is almost essential to develop a methodology for selection and optimized sizing of the logistics associated with hydrogen systems at two levels. On the one hand, it is necessary to develop a model that could define the characteristics of hydrogen stations automatically and also in a simple and agile way. On the other hand, it is required a complete simulation model that has as objective the

 ISBN: 978-84-617-8694-7

detailed sizing of the hydrogen stations according to their particularities, and whose functional bases are embodied in this study.

Given this opportunity, the present work presents a methodology that tries to give solution to the sizing of the hydrogen refueling facilities and the logistics involved.

2 Hydrogen refueling facilities

There are different types of hydrogen stations from the point of view of hydrogen supply, taking into account both technical and economic viability of this product. In the future, the location of the hydrogen stations will influence in the technical solution adopted for the supply: hydrogen stations located in remote locations, large stations in rural areas, hydrogen stations located in the environs of the cities or hydrogen stations located in main roads.

3 Methodology

The methodology proposed in this work is formed by:

- A model that defines the characteristics of hydrogen-type stations from an input data and that has been denominated *Pattern of Use*.
- A calculation model that allows the complete sizing of a hydrogen station depending on the particularities of each case.

3.1 Pattern of Use

The Pattern of Use is a model that defines the basic characteristics of a hydrogen-type station from basic input data: the type and the number of vehicles that are expected to be use by the facility. The characteristics it defines are: the number of dispensers required, the daily supply flow, the type of supply and the approximate cost of the infrastructure. In this way, it is obtained a pre-design of the installation that will be used as the first level of the calculation model object of the present study.

After analysing the documentation on the most representative projects and experiences on the subject, there was no shared trend in the structure and technologies of the hydrogen refueling facilities used in these projects [3–8, 11]. In addition, as the vast majority are demonstration projects, these stations are highly oversized, consequently, modelling by means of direct assignment of the type of hydrogen station according to the experience was not advisable.

As a consequence of this previous analysis, a formulation to make the model has been used in order to determine the number of dispensers, the delivery flow, limited by the number of dispensers and an approximate cost, in a generic way.

The objective is to provide a hydrogen station pre-sizing and an approximate cost, as close to reality as possible, starting from a single initial data, which are the vehicles (buses and passenger cars) that are expected to use the facilities on a daily basis.

As a prerequisite to perform the calculations, it is needed to know:

- The average capacity of hydrogen deposits with which the vehicles count.
- The different models, both electrolyzers and reformers.
- The estimated time that the hydrogen production equipment can remain in operation.

 ISBN: 978-84-617-8694-7

- The characteristics of the hydrogen delivery process of the dispenser. It is especially important the rate at which hydrogen is supplied.
- The hours in which the station remains operational.
- The time it takes to perform the refueling ($T_{refueling}$) of a vehicle and which is calculated by the following formulas:

$$T_{refueling}(min) = T_{approximation}(min) + T_{preparation}(min) + T_{exit}(min) + T_{supply}(min) \quad (1)$$

$$T_{supply}(min) = T_{unitary}(min/(kgH_2)) * Tank\_Capacity(kgH_2) \quad (2)$$

The starting data to be taken into account in the model are as follows: Propulsion; Number of vehicles; Number of dispensers; Refueling capacity; Type of supplying; In situ supply amount; Exterior supply amount; Storage capacity; Costs.

3.2 Calculation Model

The calculation model will allow to determine what should be the main characteristics that a hydrogen supply station has to face, depending on the needs to cover and the external conditions that surround it. The purpose of the calculation model is to achieve a criterion that allows the sizing and the correct placing of a hydrogen supply station, taking into account economic, technical and environmental aspects.

The global process of development of the calculation model is divided into four phases, each of them subdivided into several activities: Analysis of requirements and specifications of the calculation model; Design of the calculation model; Validation of the calculation model; Feedback of the calculation model.

To facilitate the compression of the operation of this model, Figure 1 shows the general scheme of the calculation model, which helps to understand the mechanism used to decide the configuration of the hydrogen station infrastructure. This modeling allows to know the environment and particularities in which the new infrastructure will be located, and will help to determine the size, characteristics and location of hydrogen refueling stations network that minimizes the cost functions and maximizes those of operation and functionality.

Firstly, it is necessary to determine the number of hydrogen stations which is needed to satisfy the total hydrogen demand.

The $Simulation_A$ process takes place if there is no hydrogen station with in situ production or hydrogen production plant near the hydrogen station. In this case, it is simulated that only one production plant will supply the n hydrogen stations, that is, it will take care of the production of all the demand, a result obtained in the process of calculating the number of hydrogen stations. The case of $Simulation_B$, it is studied when there is a hydrogen station plant with in situ production or a production plant near the hydrogen station, but which is not able to supply all the new demand (the total demand calculated in the process of calculating the number of hydrogen stations) plus the previous demand.

Finally, once the simulation process has been completed and the best result has been selected according to the objectives, the last process, called location and adjustment is started. This process is the most complex, since it is necessary to consider a greater number of parameters considering all the alternatives that take place in the previous processes.

In this process, it is firstly determined the location of the in situ production hydrogen stations, and afterwards the distribution ones.

 ISBN: 978-84-617-8694-7

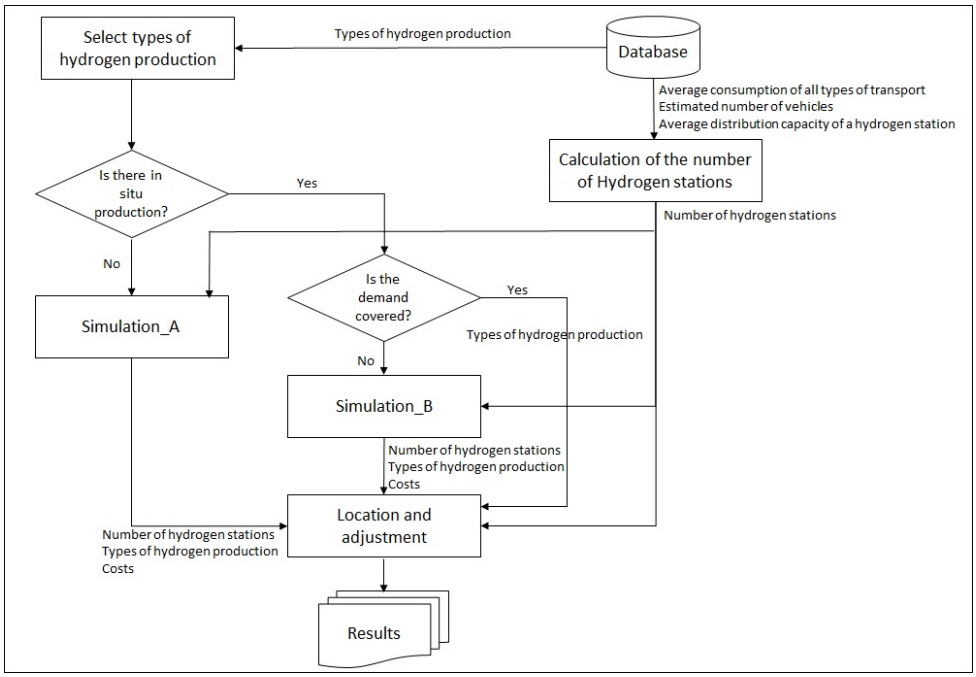


Figure 1: General scheme of a calculation model.

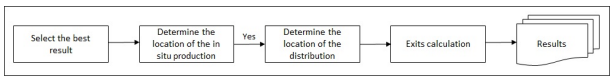


Figure 2: Location and adjustment

4 Conclusions

This study has focused on the need to create a methodology that will help to solve the dimensioning of the hydrogen refueling facilities and the logistics involved. The purpose of the realization of an optimized design of a new distribution infrastructure is to determine the size, characteristics and location of the network of hydrogen refueling stations that minimizes cost functions and maximizes those of operability and functionality.

An essential requirement previous to the determination of the optimum the optimum dimension of the hydrogen refueling stations is the daily hydrogen demand, which will have to be met. This future demand can be determined from demand forecasts of the fleet of public or private vehicles that will use hydrogen in a specific time horizon and the study and design of routes to be followed by vehicles.

When designing the methodology, the point at which it has placed special emphasis, due to its high sensitivity in both the technical and economic aspects, is the type of hydrogen supply to the station. The type of supply depends on four basic variables: cost, demand, location of facilities and security of supply.

References

[1] E. Alcantara, Análisis de la movilidad urbana. Espacio, medio ambiente y equidad, Corporación Andina de Fomento, Bogotá, Colombia, 2010.

[2] J. Andrews, B. Shabani, Re-envisioning the role of hydrogen in a sustainable energy economy. International Journal of Hydrogen Energy, 37(2) (2012), 1184-1203.

[3] Clean Energy Partnership, CEP sees Berlin on way to hydrogen metropolis. Fuel Cells Bulletin, 2006 (5) (2006).

[4] CUTE: Clean Urban Transport for Europe - Vision, Teamwork and Technology, European Commission's 5th Framework Research Programme, 2007.

[5] A. Dada, B. Boyd, L. Law, D. Semczyszyn, NRC/UBC fuelling station with intelligent compression. Towards a greener world: hydrogen and fuel cells 2004 conference and trade show Conference proceedings. Canadian Hydrogen Association, Toronto, Canada, 2004.

[6] C. Gruber, R. Wurster, Hydrogen-fueled buses: the Bavarian fuel cell bus project, Munich, Deutschland, 2002.

[7] JHFC project, Japan Hydrogen & Fuel Cell Demonstration Project. Ministry of Economy, Trade and Industry, Japan, 2010.

[8] M.H. Maack, K.D. Nielsen, H.T. Torfason, S.O. Sverrisson, K. Benediktsson, ECTOS: Ecological City Transport System - Assessment and evaluation of socio-economic factors, Rykjavik, Iceland, 2004.

[9] H.C. Manheim, Improving Organizational Effectiveness in a Changing World, 1984.

[10] S. Nocera, F. Cavallaro, The competitiveness of alternative transport fuels for CO_2 emissions. Transport Policy, 50 (2016), 1-14.

[11] STEP: Sustainable Transport Energy for Perth, Government of Western Australia's Department for Planning and Infrastructure, Perth, Australia, 2005.

[12] S.D. Stephens-Romero, T.M. Brown, M. Carreras-Sospedra, J.E. Kang, J. Brouwer, D. Dabdub, W.W. Recker, G. Scott Samuelsen, Projecting full build-out environmental impacts and roll-out strategies associated with viable hydrogen fueling infrastructure strategies. International Journal of Hydrogen Energy, 36 (22) (2011), 14309–14323.

[13] S. Stephens-Romero, G.S. Samuelsen, Demonstration of a novel assessment methodology for hydrogen infrastructure deployment. International Journal of Hydrogen Energy, 34(2) (2009), 628-641.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

Remarks and observations on (q-) Bernstein Basis functions

Yilmaz Simsek[1]

[1] *Department of Mathematics Faculty of Sciences, Akdeniz University*

emails: ysimsek@akdeniz.edu.tr

Abstract

In this paper, we investigate the q-Bernstein bases functions with their generating functions. We give many properties of these functions. We also give recurrence relations and derivative formulas for these functions. Moreover, we give some comments and observations on the q-Bernstein basis functions and the Bernstein basis functions. Finally, we simulate the results by their plots of these functions

Key words: q-Bernstein bases, Generating functions, q-exponential functions, Combinatorial sums.
MSC 2000: 65Qxx, 33Dxx, 65D17, 26Cxx, 30C10.

1 Introduction

The q-calculus related to the q-Bernstein basis functions have many applications in many various different areas. Especially, the Bernstein basis functions are of important role in order to construct the Bézier type curves and their usage are very common in the car design and other relates industrial design and Computer Aided Geometric Design (CAGD). For more detailed information about the properties and their application of the Bernstein basis functions, the reader can consult the following references ([1]-[14]).

Let $x \in [0,1]$. The classical Bernstein basis functions are given by

$$B_k^n(x) = \binom{n}{k} x^k (1-x)^{n-k}, \qquad k = 0, \ldots, n. \tag{1.1}$$

The q-Bernstein basis functions are defined by

$$B_k^n(x;q) = \begin{bmatrix} n \\ k \end{bmatrix}_q x^k \prod_{j=0}^{n-k-1} (1-q^j x), \qquad k = 0, \ldots, n \tag{1.2}$$

where $x \in [0,1]$.

Some q-calculus notations are given as follows:

 ISBN: 978-84-617-8694-7

Let $\begin{bmatrix} n \\ k \end{bmatrix}_q$ denote the q-binomial coefficient defined by

$$\begin{bmatrix} n \\ k \end{bmatrix}_q = \frac{[n]_q!}{[k]_q![n-k]_q!}, \tag{1.3}$$

$k = 0, \ldots, n$, and $q \neq 1$,

$$[k]_q! = [1]_q[2]_q \cdots [k]_q, \tag{1.4}$$

$$[0]_q! := 1,$$

and for $k \in \mathbb{Z}^+$, $[k]_q$ stands for the q-integer given by

$$[k]_q = 1 + q + \cdots + q^{k-1} = \begin{cases} \frac{1-q^k}{1-q}, & q \neq 1, \\ k, & q = 1. \end{cases} \tag{1.5}$$

Relations between $[k]_q$ and $[k]_{1/q}$ are given as follows:

$$[k]_q = q^{k-1}[k]_{1/q} \tag{1.6}$$

and

$$[k]_q! = q^{\binom{k}{2}}[k]_{1/q}!. \tag{1.7}$$

In [3], the q-exponential function $\mathcal{E}_q(x)$ is defined as follows

$$\mathcal{E}_q(x) = \sum_{n=0}^{\infty} \frac{x^n}{[n]_q!}. \tag{1.8}$$

In the work of Goldman et al. [8], we observe that if one uses the ratio test in this series for $\mathcal{E}_q(x)$ converges and therefore $\mathcal{E}_q(x)$ is well defined for all

$$|x| < \frac{1}{|1-q|}$$

if $|q| < 1$, and for all $x \in \mathbb{C}$ if $|q| > 1$ or $q = 1$. Therefore for any fixed value of $q > 0$, there is an interval in which the series for $\mathcal{E}_q(x)$ converges.

By using the q-exponential function $\mathcal{E}_q(x)$, the generating functions for the q-Bernstein basis functions is given by [8]:

$$G_k(x, t; q) = \sum_{n=k}^{\infty} B_k^n(x; q) \frac{t^n}{[n]_q!}, \tag{1.9}$$

where

$$\mathcal{E}_q(xt)G_k(x, t; q) = \frac{(xt)^k}{[k]_q!}\mathcal{E}_q(t).$$

The q-Bernstein operator is given as follows (*cf.* [10]; and also see [9]):

$$B_{n+1,q}(f)(x) = \sum_{k=0}^{n+1} \begin{bmatrix} n+1 \\ k \end{bmatrix} \Delta^k f_0 x^k \tag{1.10}$$

 ISBN: 978-84-617-8694-7

where

$$f_i = f\left(\frac{[i]_q}{[n+1]_q}\right), \Delta^0 f_i = f_i, \Delta^{k+1} f_i = \Delta^k f_{i+1} - q^k \Delta^k f_i,$$

$$\Delta^k f_i = \sum_{j=0}^{k} (-1)^j \, q^{\frac{j(j-1)}{2}} \begin{bmatrix} k \\ j \end{bmatrix}_q f\left(\frac{[i+k-j]_q}{[n+1]_q}\right),$$

and also

$$B_{n,q}(f)(x) = \sum_{i=0}^{n} f\left(\frac{[i]_q}{[n]_q}\right) B_i^n(x;q). \tag{1.11}$$

It is time to give some basic information about the q-calculus.

In [1], [8], the discrete q-derivative of a function $f(x)$ is given as follows: for $q \neq 1$

$$D_{q,x} f(x) = \frac{f(qx) - f(x)}{(q-1)x}.$$

Observe that when $f'(x)$ exists,

$$f'(x) = \lim_{q \to 1} D_{q,x} f(x).$$

The q-integral of a function $f(x)$ is given as follows (*cf.* [1], [8]):

$$\int_0^t f(x) \, d_q x = t(1-q) \sum_{j=0}^{\infty} f(tq^j) q^j, \tag{1.12}$$

provided that the infinite sum converges.

From the above definitions, one can easily get the following formulas:

$$\int_0^t x^n \, d_q x = \frac{t^{n+1}}{[n+1]_q}, \tag{1.13}$$

and

$$\int_0^t x^n \, d_{1/q} x = \frac{t^{n+1}}{[n+1]_{1/q}} = \frac{q^n t^{n+1}}{[n+1]_q}. \tag{1.14}$$

In order to give our results, we need the following Theorems:

Theorem 1.1 (*cf.* [8])**.**

$$\sum_{k=j}^{n} \begin{bmatrix} k \\ j \end{bmatrix}_q B_k^n(x;q) = \begin{bmatrix} n \\ j \end{bmatrix}_q x^j. \tag{1.15}$$

Theorem 1.2 (*cf.* [8], [7])**.** *If* $|q| < 1$*, and* $k = n$ *then*

$$\int_0^1 B_k^n(x;q) \, d_q x = \frac{1}{[n+1]_q}$$

and if $|q| < 1$*, and* $k = 0, \ldots, n-1$ *then*

$$\int_0^1 B_k^n(x;q) \, d_q x = \frac{q^{k+1}}{[n+1]_q}. \tag{1.16}$$

Theorem 1.3 (*cf.* [8])**.** *If* $|q| > 1$ *and* $k = 0, \ldots, n$*, then*

$$\int_0^1 B_k^n(x;q) \, d_{1/q} x = \frac{q^k}{[n+1]_q}. \tag{1.17}$$

 ISBN: 978-84-617-8694-7

2 Main Results

In this section, by using q-calculus methods and q-Bernstein basis functions, we derive some combinatorial sums.

By applying q-integral to both side of the equation (1.15) from 0 to 1 and combining with (1.13) and (1.16), we arrive at the following theorem:

Theorem 2.1. *Let $k = 0, 1, \ldots, n-1$. Then, we have*

$$\sum_{k=j}^{n} \begin{bmatrix} k \\ j \end{bmatrix}_q q^{k+1} = \begin{bmatrix} n \\ j \end{bmatrix}_q \frac{[n+1]_q}{[j+1]_q}. \tag{2.1}$$

Remark 2.2. *If $q \to 1$, the equation (2.1) reduce to*

$$\sum_{k=j}^{n} \binom{k}{j} = \binom{n}{j} \frac{n+1}{j+1}$$

which is given by the author in [12, Theorem 5.5].

Since

$$[n+1]_q = [n+1]_q [n]_q!,$$

equation (2.1) reduces to the following identity:

Corollary 2.3. *Let $k = 0, 1, \ldots, n-1$. Then, we have*

$$\sum_{k=j}^{n} \begin{bmatrix} k \\ j \end{bmatrix}_q q^{k+1} = \begin{bmatrix} n+1 \\ j+1 \end{bmatrix}_q. \tag{2.2}$$

By applying q-integral to both side of the equation (1.15) from 0 to 1 and combining with (1.14) and (1.17), we arrive at the following theorem:

Theorem 2.4. *Let $k = 0, 1, \ldots, n-1$. Then, we have*

$$\sum_{k=j}^{n} \begin{bmatrix} k \\ j \end{bmatrix}_q q^{k} = q^j \begin{bmatrix} n \\ j \end{bmatrix}_q \frac{[n+1]_q}{[j+1]_q}. \tag{2.3}$$

Remark 2.5. *If $q \to 1$, the equation (2.3) reduce to the equation (2.2).*

Similarly, we have the following identity:

Corollary 2.6. *Let $k = 0, 1, \ldots, n-1$. Then, we have*

$$\sum_{k=j}^{n} \begin{bmatrix} k \\ j \end{bmatrix}_q q^{k} = q^j \begin{bmatrix} n+1 \\ j+1 \end{bmatrix}_q. \tag{2.4}$$

By applying q-integral to both side of the equations (1.10) and (1.11) from 0 to 1 respectively, and combining with (1.14) and (1.17), we arrive at the following lemmas:

 ISBN: 978-84-617-8694-7

Lemma 2.7.

$$\int_0^1 B_{n+1,q}(f)(x)\, d_{1/q}x = \sum_{k=0}^{n+1} \begin{bmatrix} n+1 \\ k \end{bmatrix}_q \Delta^k f_0 \frac{q^k}{[k+1]_q}. \tag{2.5}$$

Lemma 2.8.

$$\int_0^1 B_{n+1,q}(f)(x)\, d_{1/q}x = \sum_{i=0}^{n+1} f\left(\frac{[i]_q}{[n+1]_q}\right) \frac{q^k}{[n+2]_q}. \tag{2.6}$$

Combining (2.5) with (2.6)' we get the following theorem:

Theorem 2.9.

$$\sum_{k=0}^{n+1} \begin{bmatrix} n+1 \\ k \end{bmatrix}_q \Delta^k f_0 \frac{q^k}{[k+1]_q} = \sum_{i=0}^{n+1} f\left(\frac{[i]_q}{[n+1]_q}\right) \frac{q^k}{[n+2]_q}. \tag{2.7}$$

3 Remarks and Observations

By combining (1.2), (1.8) and (1.9), one may obtain not only some fundamental properties of the generating functions for the q-Bernstein basis functions and their functional equations, but also recurrence relations and derivative formulas for these functions. Moreover, one also may give some comments and observations on the q-Bernstein basis functions and the Bernstein basis functions. And also, one may simulate the results by their plots of these functions.

Acknowledgements

This paper was supported by the Scientific Research Project Administration of Akdeniz University. This paper is dedicated to the 70th birthday of Professor Wolfgang Sproessig.

References

[1] G. E. ANDREWS, R. A. ASKEY AND R. ROY, *Special Functions*, Cambridge University Press, Cambridge, 1999.

[2] M. ACIKGOZ AND S. ARACI, *On the generating function for Bernstein polynomials*, Amer. Inst. Phys. Conf. Proc. **CP1281** (2010) pp 4.

[3] H. EXTON, *q-Hypergeometric Functions and Applications*, New York: Halstead Press, Chichester: Ellis Horwood, 1983.

[4] R. T. FAROUKI, *The Bernstein polynomials basis: a centennial retrospective*, Comput. Aided Geom. Des. **29** (2012) 379–419.

[5] R. GOLDMAN, *Pyramid Algorithms, A Dynamic Programming Approach to Curves and Surfaces for Geometric Modeling*, The Morgan Kaufmann Series in Computer Graphics and Geometric Modeling, Elsevier Science, 2003.

 ISBN: 978-84-617-8694-7

[6] R. GOLDMAN, P. Simeonov, *Formulas and algorithms for quantum differentiation of quantum Bernstein bases and quantum Bézier curves based on quantum blossoming*, Graphical Models, **74**(6)(2012) 326–334.

[7] T. Kim, *Some Identities on the q-Integral Representation of the Product of Several q-Bernstein-Type Polynomials*, Abstr. Appl. Anal. **2011**, Article ID 634675 (2011) 11 pages.

[8] R. Goldman, P. Simeonov and Y. Simsek, *Generating functions for the q-Bernstein bases*, SIAM J. Discrete Math. **28**(3) (2014) 1009–1025.

[9] N. I. Mahmudov and P. Sabancigil, *q-Parametric Bleimann Butzer and Hahn Operators*, J. Inequalities Appl. **2008** Article ID 816367, (2008) 15 pages.

[10] G. M. Phillips, *Bernstein polynomials based on the q-integers*, The heritage of P. L. Chebyshev: a Festschrift in honor of the 70th birthday of T. J. Rivlin, Annals of Numerical Math. **1-4** (1997) 511–518.

[11] G. M. Phillips, *A survey of results on the q-Bernstein polynomials,* IMA J. Numer. Analysis **30** (2010) 277–288.

[12] Y. Simsek, *Analysis of the Bernstein basis functions: an approach to combinatorial sums involving binomial coefficients and Catalan numbers,* Math. Meth. Appl. Sci. **38** (2015) 3007–3021.

[13] Y. Simsek, *Generating functions for the Bernstein type polynomials: a new approach to deriving identities and applications for the polynomials,* Hacet. J. Math. Stat. **43** (1) (2014) 1–14.

[14] Y. Simsek, *Functional equations from generating functions: a novel approach to deriving identities for the Bernstein basis functions,* Fixed Point Theory Appl. **2013** 1:80 (2013), 13 pages.

[15] Y. Simsek, M. Acikgoz, *A new generating function of (q-) Bernstein-type polynomials and their interpolation function*, Abstr. Appl. Anal. **2010** Article ID 769095 (2010), 12 pages.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

Remarks on common fixed point results in C^*-algebra-valued metric spaces

Kamonrat Sombut[1], Tanusri Senapati[2], Poom Kumam[3], Lakshmi Kanta Dey[4] and Phatiphat Thounthong[5]

[1] *Department of Mathematics and Computer Science, Faculty of Science and Technology, Rajamangala University of Technology Thanyaburi (RMUTT), Rungsit-Nakorn Nayok Rd., Klong 6, Thanyaburi, Pathumthani 12110, Thailand*

[2] *Department of Mathematics, National Institute of Technology, Durgapur, West Bengal, India.*

[3] *Department of Mathematics, King Mongkut's University of Technology Thonburi (KMUTT), 126 Pracha Uthit Rd., Bang Mod, Thung Khru, Bangkok 10140, Thailand.*

[4] *Department of Mathematics, National Institute of Technology Durgapur, West Bengal, India.*

[5] *Department of Teacher Training in Electrical Engineering, Faculty of Technical Education, King Mongkuts University of Technology North Bangkok (KMUTNB), Wongsawang, Bangsue, Bangkok 10800, Thailand.*

emails: kamonrat_s@rmutt.ac.th, senapati.tanusri@gmail.com, poom.kum@kmutt.ac.th, lakshmikdey@yahoo.co.in, phtt@kmutnb.ac.th

Abstract

n this short note, we study the article of Xin et al. [J. Nonlinear Sci. Appl., (9) 2016] and unexpectedly notice that the common fixed point result of this article do not produce any new result in literature. In fact the main results of this article coincide with some consequences of previous published results.

Key words: C^*-algebra, common fixed point
MSC 2000: 47H10, 54H25.

ISBN: 978-84-617-8694-7

1 Introduction

In 2014, Ma et al. [4] introduced the concept of C^*-algebra-valued metric space and presented some fixed point results for mappings satisfying contractive or expansive conditions in this space. Many researchers have already done their research in this structure but surprisingly Kadelburg and Radenović [1] and Alsulami et al. [5] observed that the all these results in this structure can be directly deduced as consequences of different fixed point results in standard metric and other related structures of metric counterpart. Recently, Xin et al. [7] presented common fixed point results on C^*-algebra-valued metric spaces. They established the following result.

Theorem 1.1 *Let $(X, \mathcal{A}, d)$ be a complete C^*-algebra-valued metric space. Suppose that two mappings $T, S : X \to X$ satisfy*

$$d(Tx, Sy) \preceq a^* d(x, y)a$$

for any $x, y \in X$ and $a \in \mathcal{A}$ with $||a|| < 1$. Then T and S have a unique common fixed point in X.

In this article we show that actually the mapping T is identical with S i.e. $Tx = Sx$ for all $x \in X$. Hence the Theorem 1.1 coincides with the result of Ma et al. [4]. Also, the authors of [7] proved the following theorem as a corollary.

Theorem 1.2 *Let $(X, \mathcal{A}, d)$ be a complete C^*-algebra-valued metric space. Suppose that the mapping $T : X \to X$ satisfies*

$$d(T^m x, T^n y) \preceq a^* d(x, y)a$$

for any $x, y \in X$; $a \in \mathcal{A}$ with $||a|| < 1$ and m, n are any positive integers. Then T has a unique fixed point in X.

In next section, we show that every point $x \in X$, $T^n x$ is a fixed point of T whenever $m > n$.

In 2007, Huang and Zhang [6] introduced the concept of cone metric spaces. Later on, Radenović and Kadelburg [2] showed that every cone metric space (X, d) with a normal solid cone and normal constant $K = 1$ is identical with standard metric space. Hence, the common fixed point results in cone metric spaces presented by Abbas and Jungck [3] also hold if we consider the underlying space as standard metric space. Here, we consider the common fixed point results of Abbas and Jungck [3] in the context of standard metric spaces.

Theorem 1.3 *Let (X, d) be a standard metric space. Suppose that the mappings $T, S : X \to X$ satisfy either of the following conditions:*

 ISBN: 978-84-617-8694-7

(C1) $d(Tx,Ty) \leq kd(Sx,Sy)$,

(C2) $d(Tx,Ty) \leq k[d(Tx,Sx)+d(Ty,Sy)]$,

(C3) $d(Tx,Ty) \leq k[d(Tx,Sy)+d(Ty,Sx)]$

for any $x,y \in X$ *and* $k \in [0,1)$ *for* $(C1)$ *contraction and* $k \in [0,\frac{1}{2})$ *for rest of the contractions. If* $R(T) \subset R(S)$ *and* $R(S)$ *is complete in* X*, then* S *and* T *have unique point of coincidence in* X*. Also, if* S *and* T *are weakly compatible, then there exists a unique common fixed point of* S *and* T *in* X*.*

2 Main result

In this section, we show that the results presented by Xin et al. [7] do not produce any new idea in literature.

Theorem 2.1 *Let* $(X,\mathcal{A},d)$ *be a* C^**-algebra-valued metric space. Suppose that two mappings* $T,S : X \to X$ *satisfy*

$$d(Tx,Sy) \preceq a^*d(x,y)a$$

for any $x,y \in X$ *and* $a \in \mathcal{A}$ *with* $||a|| < 1$*. Then* $Tx = Sx$ *for all* $x \in X$*.*

Proof. Let $x \in X$. Then from the hypothesis of the theorem we have

$$d(Tx,Sx) \preceq a^*d(x,x)a \quad \Rightarrow d(Tx,Sx) \preceq \theta.$$

Thus we have that for all $x \in X$, $Tx = Sx$. Hence, S and T are identical.

Remark 2.2 *From the above theorem we observe that Theorem 1.1 does not give anything new and it coincides with the fixed point result in* C^**-algebra-valued metric space [4, p.4, Theorem 2.1]. On the other hand Kadelburg and Radenović [1] and Alsulami et al. [5] independently proved that fixed point results in this space are the direct consequences of metric fixed point results. Hence this result contributes nothing new.*

Theorem 2.3 *Let* $(X,\mathcal{A},d)$ *be a complete* C^**-algebra-valued metric space. Suppose that the mapping* $T : X \to X$ *satisfies*

$$d(T^mx,T^ny) \preceq a^*d(x,y)a$$

for any $x,y \in X$*;* $a \in \mathcal{A}$ *with* $||a|| < 1$ *and* m,n *are any positive integers. Then every point* $x \in X$*,* T^nx *is a fixed point of* T *whenever* $m > n$*.*

 ISBN: 978-84-617-8694-7

Proof. As previous case, for all $x \in X$ and for $m > n$, we obtain

$$\begin{aligned} d(T^m x, T^n x) \preceq a^* d(x,x)a \quad &\Rightarrow \quad d(T^m x, T^n x) \preceq \theta \\ &\Rightarrow \quad T^m x = T^n x \\ &\Rightarrow \quad T^{m-n}(T^n x) = T^n x, \end{aligned}$$

which implies that $T^n x$ is fixed point of T for every $x \in X$ and for all $n \in \mathbb{N}$.

Next, we pick up the following result from [7].

Theorem 2.4 *Let $(X, \mathcal{A}, d)$ be a complete C^*-algebra-valued metric space. Suppose that two mappings $T, S : X \to X$ satisfy*

$$d(Tx, Ty) \preceq a^* d(Sx, Sy)a$$

for any $x, y \in X$ and $a \in \mathcal{A}$ with $||a|| < 1$. If $R(T) \subset R(S)$ and $R(S)$ is complete in X, then S and T have unique point of coincidence in X. Also, if S and T are weakly compatible, then there exists a unique common fixed point of S and T in X.

Now, we prove the following result.

Theorem 2.5 *Theorem 2.4 is equivalent with Theorem 1.3.*
Proof. In Theorem 2.4, if we consider $\mathcal{A} = \mathbb{R}$, absotule value as norm and $a^* = a$ for involution then we obtain

$$d(Tx, Ty) \leq a^2 d(Sx, Sy)$$

for all $x, y \in X$ and $a^2 \in [0, 1)$. Then clearly the contraction principle of the Theorem 2.4 coincides with the contraction $(C1)$ in Theorem 1.3.

On the other hand, we consider the hypotheses of Theorem 2.4. Then by choosing $\tilde{d}(x, y) = ||d(x, y)||$, obviously, one can get

$$\tilde{d}(Tx, Ty) = ||d(Tx, Ty)|| \leq ||a^* d(Sx, Sy)a|| \leq ||a||^2 \tilde{d}(x, y)$$

where $||a||^2 \in [0, 1)$. Hence by Theorem 1.3 with $(C1)$ contraction principle, S and T have unique coincidence point. Also, if the mappings are weakly compatible then they have unique common fixed point in X.

Remark 2.6 *In a similar fashion, one can see that the Theorem 2.9 and Theorem 2.10 in Xin et al. [7] are equivalent to Theorem 1.3 with $(C2)$ and $(C3)$ contraction principles respectively.*

 ISBN: 978-84-617-8694-7

Acknowledgements

This project was partially supported by the Theoretical and Computational Science (TaCS) Center under Computational and Applied Science for Smart Innovation Cluster (CLASSIC), Faculty of Science, KMUTT. Moreover, this research work was financially supported by King Mongkut's University of Technology North Bangkok. Contract No. KMUTNB-60-ART-084. The first author was supported by Rajamangala University of Technology Thanyaburi (RMUTT) for financial support The second named author would like to express her sincere thanks to DST-INSPIRE, New Delhi, India for their financial support under INSPIRE fellowship scheme.

References

[1] Z. Kadelburg, S. Radenović, Fixed point results in C^*-algebra-valued metric spaces are direct consequences of their standard metric counterparts. Fixed point Theory Appl., **53, 2016.**

[2] S. Radenović, Z. Kadelburg, Quasi-contractions on symmetric and cone symmetric spaces, Banach J. Math. Anal. **5 (1), 2011, 38-50.**

[3] M. Abbas, G. Jungck, Common fixed point results for noncommuting mappings without continuity in cone metric spaces, J. Math. Anal. Appl. **341, 2008, 416-420.**

[4] Z.H. Ma, L.N. Jiang, H. Sun, C^*-algebra-valued metric spaces and related fixed point theorems, Fixed Point Theory Appl. **206, 2014.**

[5] H.H. Alsulami, R.P. Agarwal, E. Karapinar, F. Khojasteh, A short note on C^*-valued contraction mappings, J. Inequalities Appl. **50, 2016.**

[6] L.G. Huang, X. Zhang, Cone metric spaces and fixed point theorems for contractive mappings, J. Mat. Anal. Appl. **332(2), 2007, 1468-1476.**

[7] Q. Xin, L. Jiang, Z. Ma, Common fixed point theorems in C^*-algebra-valued metric spaces. J. Nonlinear Sci. Appl., **9, 2016.**

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

Effective parameters, likelihoods and Bayesian model selection in application to epidemiological models: from SHAR to effective SIR models

Nico Stollenwerk[1], Raquel Filipe[1], Luís Mateus[1], Peyman Ghaffari[1], Bob Kooi[2], Scott Halstead[3] and Maíra Aguiar[1,4]

[1] *Centro de Matemática e Aplicações Fundamentais e Investigação Operacional, Universidade de Lisboa, Portugal*

[2] *Department of Earth and Life Sciences, Vrije Universiteit, Amsterdam, The Netherlands*

[3] *Department of Preventive Medicine and Biometrics, Uniformed Services University of the Health Sciences, Bethesda, Maryland, USA*

[4] *Centro de Matemática e Aplicações, Universidade Nova de Lisboa, Portugal*

emails: nico@ptmat.fc.ul.pt, raquel.m.filipe@gmail.com, luisgam1@yahoo.com, pgsaid@fc.ul.pt, bob.kooi@vu.nl, halsteads@erols.com, maira@ptmat.fc.ul.pt

Abstract

We derive stochastic versions of epidemiological models with severe and asymptomatic infection, which have previously been described as differential equation systems, and fit data output of these models with simpler SIR type models to obtain numerically effective infection rates from likelihoods. The results can be compared with the previously stated simpler analytical arguments for the relation between effective infection rate in the SIR models and more complex parameter combinations in the SHAR models via comparing stationary states. The tool box can be extended to be used to perform numerically Bayesian model comparison via the Bayes factor, as described in previous simpler epidemiological models, like the linear infection model or the Poisson model, in which all steps towards the Bayes factor could be performed analytically. The role of asymptomatic infection contributing to the force of infections is of special interest e.g. in the dengue fever epidemiology, as well for primary infections, which is often mild or asymptomatic, as well as for secondary infections, where the majority of hospitalizations occur but also not always leading to severe disease.

Key words: asymptomatic versus severe infection, stochastic processes, numerical likelihood functions, Bayes factor, dengue fever

ISBN: 978-84-617-8694-7

1 Introduction

In many diseases the infection is not always severe or even symptomatic. Hence the number of notified cases can differ from the number of infectious individuals. Then one of the key questions is in how far asymptomatic infected are infectious to susceptibles. We derive stochastic versions of epidemiological models with severe and asymptomatic infection, which have previously been described as differential equation systems [1]. Here we derive stochastic versions of such processes and fit data output of these models with simpler SIR type models to obtain numerically effective infection rates. The results can be compared with the previously stated simpler analytical arguments for the relation between effective infection rate in the SIR models and more complex parameter combinations in the SHAR models [1]. Namely, the relation between effective infection rate in the SIR system and the parameters of the SHAR model is confirmed by numerically evaluating the likelihoods of initial infection rate of the SHAR model and the effective infection rate of the SIR system, even outside stationarity of the respective processes, from which the analytical results were obtained in [1].

The tool box can be extended to be used to perform numerically Bayesian model comparison via the Bayes factor. We described this in previous simpler epidemiological models, in which all steps towards the Bayes factor could be performed analytically [2]. Similar approaches as we use here for the numerical calulation of likelihoods and Bayes factors have been used e.g. in [3, 4]. In our case study it turns out that the SHAR model, from which the toy data were generated, appears to be not much more likely than the effective SIR model in terms of Bayes factors. For alternative approaches of model selection, based on the Kolmogorov-Smirnov test, see e.g. [5]. However, such approaches are not as easily generalizable as the Bayes factor, where simply probabilities are given to models based on data.

The present study of such initial models is well applicable to the situation of dengue fever infections in which severity of disease is associated with antibody dependent enhancement (ADE) since a long time [6, 7]. And recently field studies tried to clarify the question if asymptomatically infected are equally infectious as severe cases with their high viral load [8]. These authors come to the conclusion that asymptomatically infected can infect as much mosquitoes as severely infected, and there is the expectation that due to their mobility, asymptomatically infected might even contribute more to the force of infection than severe cases. However, the best notification data of dengue infections from Thailand use only severe cases being hospitalized [9], which means that in relatively simple models which already can describe the empirical data well we would expect an "masked" effective infection rate as suitable rather than a "bare" infection rate, however this might be defined. Many present models of dengue fever have so complicated structures that any systematic analysis in terms of Bayes factors would most likely prefer the simpler models due to the limited number of empirical data. However, such effective models are the best tools for realistic predictions

 ISBN: 978-84-617-8694-7

of future disease levels [10], and even more in analysis of control measures like vaccination and its impact [11, 12, 13]. In the present study we concentrate on rather simplified models like the SHAR and SIR model without any serotype interactions, which would be needed to describe the empirical data more accurately [14, 15, 16]. Since such multi-strain models show not only oscillations into stationary states, and eventually stabilized by noise as we observe here in our simple models, but chaotic attractors, we will leave the study of such models with the present techniques for future research. For chaotic systems up to now mostly random walks toward the likelihood maximum are in technical reach, see [17] including further references. Hence future studies on such complex systems with mutli-strain interaction have to be performed.

2 The SHAR model

The SHAR model with severe infection or hospitalized cases H and mild or asymptomatic cases A in an otherwise SIR-type epidemiological model, see [18] for more information on the notations and standart analysis techniques, is given by

$$\begin{aligned} \frac{d}{dt}S &= \alpha R - \frac{\beta}{N}S(A+\phi H) \\ \frac{d}{dt}H &= \eta\frac{\beta}{N}S(A+\phi H) - \gamma H \\ \frac{d}{dt}A &= (1-\eta)\frac{\beta}{N}S(A+\phi H) - \gamma A \\ \frac{d}{dt}R &= \gamma(A+H) - \alpha R \end{aligned} \tag{1}$$

and $R = N - S - A - H$. Here η is the ratio of infection leading to severe disease cases and ϕ is the change of infectivity of severe cases versus the "natural" infectivity via asymptomatic cases.

The role of asymptomatic infection becomes more recognized recently, since many pathogens appear as only accidentally pathogenic, see e.g. [18, 19] initially for bacterial meningitis [20, 21, 22], with fluctuations similar to crossing a vaccination threshold [23]. Similarly, in dengue fever the virus is completely asymptomatic in its animal reservoir, monkeys, and only turns occasionally pathogenic in humans, mostly in secondary infections. The SHAR model is an oversimplified model for dengue fever in the sense that higher viral load due to antibody dependent enhancement, ADE [6, 7], or lower infectivity due to hospitalization gives a ϕ different from unity, and the majority of primary infections is asymptomatic or mild (sub-clinical) with ratio $(1-\eta)$, but here we do not consider primary versus secondary infection nor any further serotype interaction as done in [14, 16].

 ISBN: 978-84-617-8694-7

2.1 The stochastic SHAR model

We now want to generate "toy data" via a stochastic version of the SHAR model via the Gillespie algorithm, see [18] including further references, with transitions given by the following reaction scheme

$$\begin{array}{ccc} S+A & \xrightarrow{\eta\beta} & H+A \\ \text{„} & \xrightarrow{(1-\eta)\beta} & A+A \\ S+H & \xrightarrow{\phi\eta\beta} & H+H \\ \text{„} & \xrightarrow{\phi(1-\eta)\beta} & A+H \\ H & \xrightarrow{\gamma} & R \\ A & \xrightarrow{\gamma} & R \\ R & \xrightarrow{\alpha} & S \end{array} \tag{2}$$

giving the master equation for the dynamics of the probabilities

$$\begin{aligned} \frac{d}{dt}p(S,H,A,t) &= \eta\frac{\beta}{N}(S+1)A\; p(S+1,H-1,A,t) \\ &\quad +(1-\eta)\frac{\beta}{N}(S+1)(A-1)\; p(S+1,H,A-1,t) \\ &\quad +\eta\frac{b}{N}(S+1)\phi(H-1)\; p(S+1,H-1,A,t) \\ &\quad +(1-\eta)\frac{\beta}{N}(S+1)\phi H\; p(S+1,H,A-1,t) \\ &\quad +\gamma(A+1)\; p(S,H,A+1,t) \\ &\quad +\gamma(H+1)p(S,H+1,A,t) \\ &\quad +\alpha(N-(S-1)-H-A)\; p(S-1,H,A,t) \\ &\quad -\left(\frac{\beta}{N}S(A+\phi H)+\gamma(A+H)+\alpha(N-S-H-A)\right)\; p(S,H,A,t) \quad . \end{aligned} \tag{3}$$

The master equation can be written in a generic form using densities $x_1 := S/N$, $x_2 := H/N$ and $x_3 := A/N$ instead of the total numbers of individuals in the population classes S, H and A, hence state vector $\underline{x} := (x_1, x_2, x_3)^{tr}$, as

$$\frac{d}{dt}\; p(\underline{x},t) \;=\; \sum_{j=1}^{n}\Big(\; Nw_j(\underline{x}+\Delta\underline{x}_j)\cdot p(\underline{x}+\Delta\underline{x}_j,t) \;\; - Nw_j(\underline{x})\cdot p(\underline{x},t)\Big) \tag{4}$$

 ISBN: 978-84-617-8694-7

with $n = 5$ different transitions and small deviation from state $\underline{x}$ as $\Delta\underline{x}_j := \frac{1}{N} \cdot \underline{r}_j$. For the SHAR model we have explicitly the following transitions $w_j(\underline{x})$ and its vectors $\underline{r}_j$ given by

$$\begin{array}{lll} w_1(\underline{x}) = \eta\beta x_1(x_3 + \phi x_2) & , & \underline{r}_1 = (1,-1,0)^{tr} \\ w_2(\underline{x}) = (1-\eta)\beta x_1(x_3 + \phi x_2) & , & \underline{r}_2 = (1,0,-1)^{tr} \\ w_3(\underline{x}) = \gamma x_2 & , & \underline{r}_3 = (0,1,0)^{tr} \\ w_4(\underline{x}) = \gamma x_3 & , & \underline{r}_4 = (0,0,1)^{tr} \\ w_5(\underline{x}) = \alpha(1 - x_1 - x_2 - x_3) & , & \underline{r}_5 = (-1,0,0)^{tr} \quad . \end{array} \tag{5}$$

With these $w_j(\underline{x})$ and $\underline{r}_j$ specified we also can express the mean field ODE system and in Kramers-Moyal approximation of the master equation to a Fokker-Planck equation a stochastic differential equation system, as will be shown below. From the SHAR model given as master equation we can now simulate realizations of the stochastic process via the Gillespie algorithm, and in this way obtain a "toy data set"

$$\underline{D} := (H_1, H_2, ..., H_{n_d}) \tag{6}$$

of n_d data points of hospital cases H_ν at times t_ν. On this data set we can then test statistical methods to fit parameters of various models including model comparison in a Bayesian framework. We will now try to describe the output of the SHAR model $\underline{D}$ via an effective simpler model, the SIR model with new infection rate $\tilde{\beta}$.

3 The effective SIR model

An effective SIR model tries to explain the observed number of infected via a simple infection class I without any further distinction into more complex classes like asymtomatically infected, severe cases or any other possible mechanisms like primary or secondary infection etc. The SIR model is given by

$$\begin{aligned} \frac{d}{dt}S &= \alpha R - \frac{\tilde{\beta}}{N} SI \\ \frac{d}{dt}I &= \frac{\tilde{\beta}}{N} SI - \gamma I \\ \frac{d}{dt}R &= \gamma I - \alpha R \end{aligned} \tag{7}$$

with an initially unknown infection rate $\tilde{\beta}$ and to be determined either from data as we will do below, or as was done in [1], by assuming that the effective SIR system has the same stationary state value I^* as a more complex model, here the SHAR model with observed severe infected H^*.

 ISBN: 978-84-617-8694-7

3.1 Effective infection rate

In [1] we obtained for the effective infection rate $\tilde{\beta}$ as a function of the parameters of the more complex model, the SHAR model, the expression

$$\tilde{\beta} = \frac{\gamma}{1 - \eta\left(1 - \frac{1}{1+(\phi-1)\eta} \cdot \frac{\gamma}{\beta}\right)} \tag{8}$$

from assuming the condition that the stationary state I^* of the effective SIR model is the same as the stationary state value of the observed severe cases H^* in the SHAR model, hence $I^*(\tilde{\beta}) = H^*(\beta, \eta, \phi)$. It turned out that also outside stationarity the deterministic simulations of the SHAR model and the effective SIR model with the above determined effective infection rate $\tilde{\beta}$ agree quite well, when using otherwise the same parameters and initial conditions, as far as the SIR system has corresponding parameters in the SHAR model, see Fig. 1.

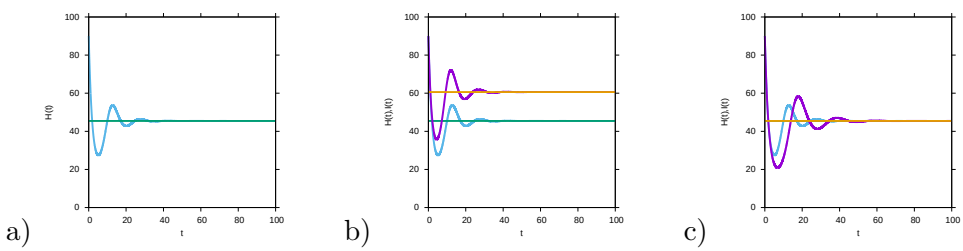


Figure 1: *Comparison between deterministic SHAR model and SIR model, also outside the stationary state. a) The trajectory of the severe diseased cases $H(t)$ of the SHAR model oscillating into the stationary state H^* after a transient. b) In comparison also the number of infected $I(t)$ from the SIR model with still $\tilde{\beta} = \beta$, in comparison to $H(t)$ from the SHAR model. c) Now we use the effective infection rate in the effective SIR model, hence $\tilde{\beta} = \tilde{\beta}(\alpha, \beta, \gamma, \eta)$ from $I^* = H^*$, Eq. (8).*

Parameters for the initial study of the deterministic matching are given as follows: For the SHAR model we use $N = 1000$, $\gamma = 1$, $\beta = 3 \cdot \gamma$, $\alpha = 0.1$, $\eta = 3/4$, and for the moment $\phi = 1$. Initial conditions are $S_0 = 200$, $A_0 = 30$, $H_0 = \frac{\eta}{1-\eta} A_0$, a relation which holds in stationarity $H^* = \frac{\eta}{1-\eta} A^*$, and $R_0 = N - S_0 - A_0 - H_0$. For the SIR system we use the same parameters if not otherwise stated, hence initially $\tilde{\beta} = \beta$, see Fig. 1 b), and as initial conditions $S_{0,SIR} = (\beta/\tilde{\beta}) S_{0,SHAR}$, $I_0 = H_0$ and $R_0 = N - S_0 - I_0$. In Fig. 1 c) we use the analytically calculated effective infection rate Eq. (8), which turns with the present $\eta = 3/4$ etc. out to be $\tilde{\beta} = 2 \cdot \gamma$. In Fig. 2 b) we also change the intial condition in the SHAR model to $A_0 = 100$ to observe the effect of initial conditions far away from the equilibrium, and still $I(t)$ is quite well comparable with $H(t)$ outside the condition $I^* = H^*$.

 ISBN: 978-84-617-8694-7

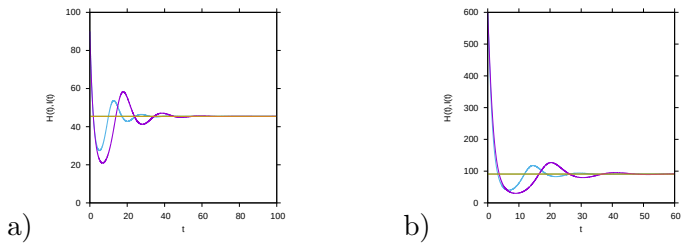


Figure 2: *The comparison of deterministic SHAR and SIR models for varying initial conditions. a)* $A_0 = 30$ *and b)* $A_0 = 100$ *and* $N = 2000$*. Even transients far out of equilibrium of the SHAR model are matched quite well via the transient of the effective SIR model, though the effective parameters are determined only by the stationary state condition* $I^* = H^*$*.*

Next we look at the master equation formulation of the SHAR model to simulate toy data. For a comparison of the stochastic version as compared to the deterministic see Fig. 3. The stochastic fluctuations of the SHAR model itself go well beyond the differences between the SHAR model and the effective SIR in their deterministic versions. In the present study we want to determin the effective infection rate $\tilde{\beta}$ by comparing stochastic simulations of the effective SIR model with the data $\underline{D}$ obtained from a stochastic simulation of the SHAR model. For this we vary values of $\tilde{\beta}$ and count the number of simulations being close to the data, and the $\tilde{\beta}$ value with maximal number of simulations close to the data is the best estimator of $\tilde{\beta}$ for this data set. Therefore, we now need the stochastic version of the effective SIR model, and best a version which is quick in simulation time, since we need many simulations with varying $\tilde{\beta}$ values.

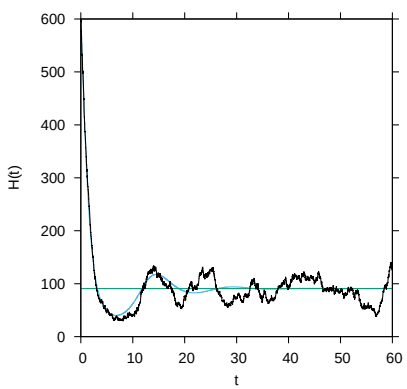


Figure 3: *Stochastic realization via the Gillespie algorithm from the master equation of the SHAR model in comparison with the deterministic mean field system.*

 ISBN: 978-84-617-8694-7

3.2 Stochastic versions of the effective SIR model

The master equation for the effective SIR model is given by the reaction scheme

$$\begin{array}{rcl} S+I & \xrightarrow{\tilde{\beta}} & I+I \\ I & \xrightarrow{\gamma} & R \\ R & \xrightarrow{\alpha} & S \end{array} \tag{9}$$

and explicitly for the dynamics of the probabilities

$$\begin{aligned} \frac{d}{dt}p(S,I,t) &= \frac{\tilde{\beta}}{N}(S+1)(I-1)\ p(S+1,I-1,t) \\ &\quad +\gamma(I+1)p(S,I+1,t) \\ &\quad +\alpha(N-(S-1)-I)\ p(S-1,I,t) \\ &\quad -\left(\frac{\tilde{\beta}}{N}SI+\gamma I+\alpha(N-S-I)\right)\ p(S,I,t) \quad . \end{aligned} \tag{10}$$

For the master equation in densities $x_1 := S/N$ and $x_2 := I/N$ we have now transitions $w_j(\underline{x})$ and its vectors $\underline{r}_j$ given by

$$\begin{array}{lll} w_1(\underline{x}) = \tilde{\beta}x_1x_2 & , & \underline{r}_1 = (1,-1)^{tr} \\ w_2(\underline{x}) = \gamma x_2 & , & \underline{r}_2 = (0,1)^{tr} \\ w_3(\underline{x}) = \alpha(1-x_1-x_2) & , & \underline{r}_3 = (-1,0)^{tr} \quad . \end{array} \tag{11}$$

We could use now simulations of the master equation directly to measure the effective infection rate $\tilde{\beta}$, but for large population sizes N the simulations become very slow. So we will use the Kramers-Moyal expansion to obtain a Fokker-Planck equation as approximation of the master equation, for which the corresponding stochastic differential equation system can be simulated much faster.

3.3 Fokker-Planck approximation of the effective SIR model

From the master equation in densities, Eq. (4), now for the SIR system, we use Taylor's expansion

$$w_j(\underline{x}+\Delta\underline{x}_j)\cdot p(\underline{x}+\Delta\underline{x}_j,t) = \sum_{\nu=0}^{\infty}\frac{1}{\nu!}\left(\Delta\underline{x}_j\cdot\nabla_{\underline{x}}\right)^{\nu} w_j(\underline{x})\ p(\underline{x},t) \tag{12}$$

 ISBN: 978-84-617-8694-7

giving to second order in $1/N$ a Fokker-Planck equation

$$\begin{aligned} \frac{\partial}{\partial t}\, p(\underline{x},t) &= -\nabla_{\underline{x}} \left(\sum_{j=1}^{n} (-\underline{r}_j \cdot w_j(\underline{x}))\, p(\underline{x},t) \right) \\ &\quad + \frac{\sigma^2}{2} \sum_{j=1}^{n} (\underline{r}_j \cdot \nabla_{\underline{x}})^2 w_j(\underline{x})\, p(\underline{x},t) \end{aligned} \tag{13}$$

with

$$\nabla_{\underline{x}} = \begin{pmatrix} \frac{\partial}{\partial x_1} \\ \frac{\partial}{\partial x_2} \end{pmatrix} = \partial_{\underline{x}} \tag{14}$$

or in different notation

$$\frac{\partial}{\partial t}\, p(\underline{x},t) = -\partial_{\underline{x}} \left(\underline{f}(\underline{x})\, p(\underline{x},t) \right) + \frac{\sigma^2}{2} \overrightarrow{\partial}_{\underline{x}} \left(G^2(\underline{x})\, p(\underline{x},t) \right) \overleftarrow{\partial}_{\underline{x}} \tag{15}$$

using simply a quadratic form $\overrightarrow{\partial}_{\underline{x}} \left(G^2(\underline{x})\, p(\underline{x},t) \right) \overleftarrow{\partial}_{\underline{x}}$ here with

$$\overrightarrow{\partial}_{\underline{x}} \left(G^2\, p \right) \overleftarrow{\partial}_{\underline{x}} = \left(\frac{\partial}{\partial x_1}, \frac{\partial}{\partial x_2} \right) \cdot \begin{pmatrix} g_{11} & g_{12} \\ g_{21} & g_{22} \end{pmatrix}^2 p(\underline{x},t) \cdot \begin{pmatrix} \overleftarrow{\frac{\partial}{\partial x_1}} \\ \\ \overleftarrow{\frac{\partial}{\partial x_2}} \end{pmatrix} \tag{16}$$

and

$$\begin{aligned} \underline{f}(\underline{x}) &= \sum_{j=1}^{n} \underline{f}_j(\underline{x}) = \sum_{j=1}^{n} (-\underline{r}_j \cdot w_j(\underline{x})) \\ G^2(\underline{x}) &= \sum_{j=1}^{n} G_j^2(\underline{x}) = \sum_{j=1}^{n} \underline{r}_j \cdot \underline{r}_j^{tr}\, w_j(\underline{x}) \quad . \end{aligned} \tag{17}$$

The Fokker-Planck equation gives a stochastic differential equation system with $\sigma = 1/\sqrt{N}$ and in the XYZ case the two dimensional Gaussian normal noise vector $\underline{\varepsilon}(t) = (\varepsilon_{x_1}(t), \varepsilon_{x_2}(t))^{tr}$ as

$$\frac{d}{dt}\underline{x} = \underline{f}(\underline{x}) + \sigma G(\underline{x}) \cdot \underline{\varepsilon}(t) \tag{18}$$

and using matrix square root from eigenvalue-eigenvector decomposition $G^2(\underline{x}) = T \Lambda T^{-1}$ as $G(\underline{x}) = T \sqrt{\Lambda} T^{tr}$ to be numerically implemented easily, and much faster than the Gillespie algorithm for the master equation.

 ISBN: 978-84-617-8694-7

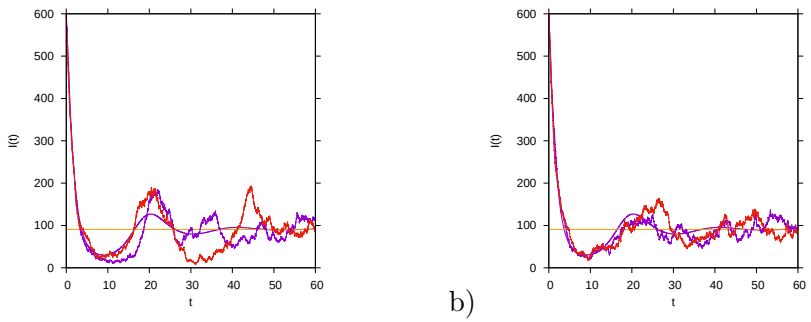


Figure 4: *Stochastic realizations via the Gillespie algorithm from the master equation (pink) and from the stochastic differential equation system (red), of the SIR model in comparison with the deterministic mean field system. Shown are two different stochastic realizations, but otherwise same parameters and initial conditions.*

Now we can vary $\tilde{\beta}$ and compare in an η-ball method [24, 10] the simulations of the SIR model with the data vector $\underline{D}$, which gives the likelihood $p(\underline{D}|\tilde{\beta})$. The η as a radius of a vicinity around the data vector $\underline{D}$ should not be confused with one of the parameters in the SHAR model, which is also named η, a confusion easily avoided by the very different contexts in which they appear. In case of doubt, we will call η_{SHAR} the model parameter of the SHAR model, and η_1 the η-ball radius around the data, used to evaluate simulations from model M_1, and η_2 from model M_2.

4 Data and analysis, results and conclusions

From the SHAR model we now generate data sets, like given in Fig. 5, hoping that the data already contain so much information about the system to find back parameters with not too large variance. Especially, we take a transient behaviour and a long period in stationarity covered by the time interval of the data. The data are taken by the slow Gillespie algorithm.

Then we compare simulations of the SHAR model, see Fig. 6 a), with the data, where now the simulations are performed with the much faster stochastic differential equations obtained by the Fokker-Planck approximation of the master equation with its slow Gillespie algorithm for simulations. We finde accurately the best estimate to be where expected, at $\beta_1 = 3 \cdot \gamma$, and values like $\beta_1 = 2 \cdot \gamma$ or $\beta_1 = 4 \cdot \gamma$ have already quite lower likelihood, this in spite of using different algorithms for data generation and model fitting.

The next result is, when fitting the effective SIR model to the data generated from the SHAR model, again using the stochastic differential equation approach, we obtain a best estimate from the maximum of the likelihood for the effective infection rate $\beta_2 = \tilde{\beta} = 2 \cdot \gamma$, in good agreement with out analytical result, Eq. (8). This is a non-trivial result, since

 ISBN: 978-84-617-8694-7

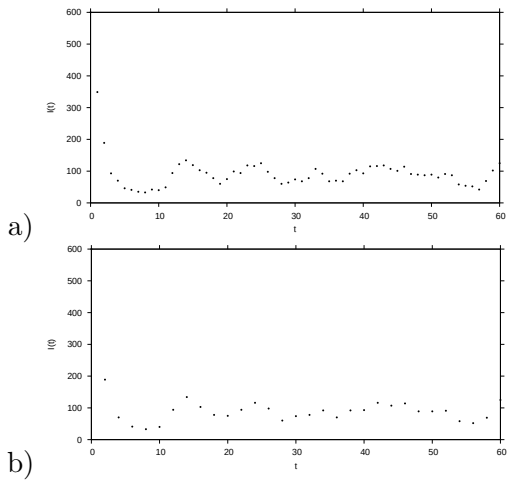


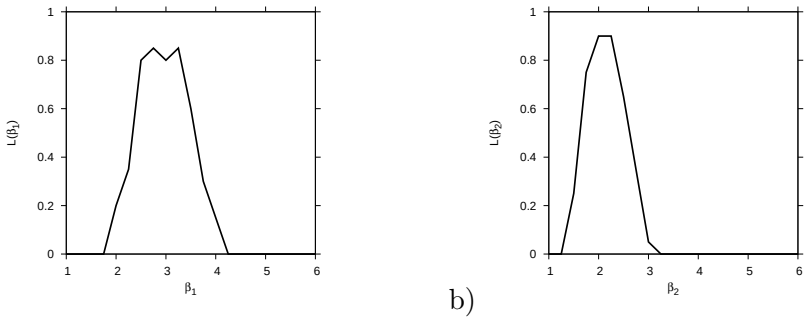
Figure 5: *Data taken from the SHAR model with different sampling times, reflecting different knowledge about the fluctuations in the system. a) Sampling time is one unit, here in units of recovery times since we set $\gamma = 1$, in b) sampling time are two units.*

we obtained the analytic expression via comparing stationary states $I^* = H^*$ only, but now obtain the same relation from simulated data outside stationarity with a long and oscillating transient and continued stochastically stabilized oscillations in stationarity. Also remarkable is that for $\beta_2 = \tilde{\beta}$ the value of $3 \cdot \gamma$, which is the one used in the SHAR model, can be statistically excluded.

Finally, we investigate the Bayes factor $k = p(M_1|\underline{D})/p(M_2|\underline{D})$ for a data set $\underline{D}$ generated from M_1, the SHAR model. Fo more details on the Bayes factor and analytic results in simple cases see [2]. A Bayes factor of $k = 1$ would indicate equal probability of both models M_1, the SHAR model, and M_2, the effective SIR model, whereas $k >> 1$ would give much higher probability of model M_1 over model M_2, that the data could lead us to reject model M_2 statistically. In our case, considering the Bayes factor as

$$k = \int\int p(M_1, \beta_1, \eta_{SHAR}|\underline{D})\, d\beta_1 d\eta_{SHAR} \Big/ \int p(M_2, \beta_2|\underline{D})\, d\beta_2 \tag{19}$$

due to the fact that we have to estimate in the SHAR model M_1 not only β_1 but also η_{SHAR} while in the SIR model M_2 we only need β_2 to estimate (neglecting for the moment all other possible insecurities) first results, still in low numerical resolution, give a Bayes factor of $k = 127/81 = 1.57$ for 360 runs of each of the models and distance to data $\eta = 450$, hence the two models have about equal probability to describe the toy data set Fig. 5 a). But future studies have to be performed to investigate this point in more detail, including the question of ensembles of realizations and their Bayes factors, hence the fluctuations of the Bayes factor, as could be performed in the simpler analytically treatable case study [2].



a) b)

Figure 6: *Likelihood measurement for a) model M_1 and b) model M_2. For model M_1 the best estimate is around $\beta_1 = 3$ as to be expected, since we set in generating the toy data set $\beta = 3\gamma$ and $\gamma = 1$. For model M_2 we obtain a best estimate around $\beta_2 = 2$, which agrees well with the analytically obtained effective $\tilde{\beta} = 2\gamma$, when using $\eta = 3/4$ in generating the toy data set.*

Further applications could be in ecology on Rosenzweig-MacArthur type models [25], and there on the Hudson Bay company data for hares and lynx furs, due to expected limit cycles rather than chaotic attractors as in dengue fever cases and its data from e.g. Thailand.

Acknowledgements

This work has been supported by the European Union under FP7 in the project DENFREE and by FCT, Portugal, including an FCT-DAAD exchange grant.

References

[1] Raquel Filipe, Nico Stollenwerk, Luís Mateus, Peyman Ghaffari, Scott Halstead & Maíra Aguiar (2016) Effective infection rate in SIR-type models from models with symptomatic and asymptomatic infection. *Proceedings of the 16th International Conference on Mathematical Methods in Science and Engineering - CMMSE 2015, Cádiz, Spain*, ISBN: 978-84-608-6082-2, edited by Jesus Vigo et al.

[2] Mateus, L., Stollenwerk, N., & Zambrini, J.C. (2013) Stochastic Models in Population Biology: From Dynamic Noise to Bayesian Description and Model Comparison for Given Data Sets, *Int. Journal. Computer Math.* **90**, 2161–2173.

[3] Tina Toni, David Welch, Natalja Strelkowa, Andreas Ipsen & Michael P. H. Stumpf (2009) Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems *J. R. Soc. Interface*, **6**, 187–202.

[4] Libo Sun, Chihoon Lee & Jennifer A. Hoeting (2015) Parameter inference and model selection in deterministic and stochastic dynamical models via approximate Bayesian computation: modeling a wildlife epidemic, *arXiv:1409.7715v2, 8 Sept. 2015, submitted to "Environmetrics"*.

[5] Stollenwerk, N., Drepper, F., & Siegel, H. (2001) Testing nonlinear stochastic models on phytoplankton biomass time series, *Ecological Modelling* **144**, 261–277.

[6] Halstead, S.B. (1982) Immune enhancement of viral infection. *Progress in Allergy*, **31**, 301–364.

[7] Halstead, S.B. (2003) Neutralization and antibody-dependent enhancement of dengue viruses. *Advances in Virus Research*, **60**, 421–467.

[8] Veasna Duong, Louis Lambrechts, Richard E. Paul, Sowath Ly, Rath Srey Laya, Kanya C. Long, Rekol Huy, Arnaud Tarantola, Thomas W. Scott, Anavaj Sakuntabhai, and Philippe Buchy (2015) Asymptomatic humans transmit dengue virus to mosquitoes, *Proc. Nat. Acad. Science* **112**, 14688–14693.

[9] Aguiar, M., Paul, R., Sakuntabhai, A., & Stollenwerk, N. (2014) Are we modeling the correct data set? Minimizing false predictions for dengue fever in Thailand, *Epidemiology and Infection*, **142**, 2447–2459.

[10] Stollenwerk, N., Mateus, L., Rocha, F., Skwara, U., Ghaffari, P., & Aguiar, M. (2015). Prediction and predictability in population biology: Noise and chaos, *Math. Model. Nat. Phenom.*, **10**, 142–164.

[11] Aguiar, M., Stollenwerk, N., & Halstead, S. (2016) The impact of the newly licensed dengue vaccine in endemic countries, *accepted for publication in "PLOS Neclected Tropical Diseases", published online December 21, 2016.*

[12] Aguiar, M., Stollenwerk, N., & Halstead, S. (2016) The risks behind Dengvaxia recommendation, *The Lancet Infectious Diseases*, **16**, 882–883.

[13] Aguiar, M., Halstead, S., & Stollenwerk, N. (2016) Consider stopping dengvaxia administration without immunological screening, *accepted for publication in "Expert Review of Vaccines", published online December 23, 2016*, http://dx.doi.org/10.1080/14760584.2017.1276831.

[14] Aguiar, M., Ballesteros, S., Kooi, B.W., & Stollenwerk, N. (2011) The role of seasonality and import in a minimalistic multi-strain dengue model capturing differences between primary and secondary infections: complex dynamics and its implications for data analysis, *Journal of Theoretical Biology*, **289**, 181–196.

 ISBN: 978-84-617-8694-7

[15] Aguiar, M., Stollenwerk, N. & Kooi, W. B. (2012). Scaling of stochasticity in dengue hemorrhagic fever epidemics. *Math. Model. Nat. Phenom.*, **7**, 1–11.

[16] Aguiar, M., Kooi, W. B., Rocha, F., Ghaffari, P. & Stollenwerk, N. (2013). How much complexity is needed to describe the fluctuations observed in dengue hemorrhagic fever incidence data? *Ecological Complexity*, **16**, 31–40.

[17] Stollenwerk, N., Aguiar, M., Ballesteros, S., Boto, J., Kooi, W. B., & Mateus, L. (2012). Dynamic noise, chaos and parameter estimation in population biology, *Interface, Focus*,**2**, 156–169.

[18] Stollenwerk, N., & Jansen, V. (2011) *Population Biology and Criticality: From critical birth–death processes to self-organized criticality in mutation pathogen systems* (Imperial College Press, World Scientific, London).

[19] Peyman Ghaffari, Vincent Jansen & Nico Stollenwerk (2011) Evolution towards critical fluctuations in a system of accidental pathogens, *Numerical Analysis and Applied Mathematics ICNAAM 2011, Chalkidiki, AIP Conf. Proc.*, **1389** (2011), 1263–1266; doi: 10.1063/1.3637837 Copyright 2011 American Institute of Physics 978-0-7354-0956-9).

[20] Stollenwerk, N., & Jansen, V.A.A. (2003) Meningitis, pathogenicity near criticality: the epidemiology of meningococcal disease as a model for accidental pathogens. *Journal of Theoretical Biology* **222**, 347–359.

[21] Stollenwerk, N., & Jansen, V.A.A. (2003) Evolution towards criticality in an epidemiological model for meningococcal disease. *Physics Letters A* **317**, 87–96.

[22] Stollenwerk, N., Maiden, M.C.J., & Jansen, V.A.A. (2004) Diversity in pathogenicity can cause outbreaks of menigococcal disease, *Proc. Natl. Acad. Sci. USA* **101**, 10229–10234.

[23] Jansen, V.A.A., Stollenwerk, N., Jensen, H.J., Ramsay, M.E., Edmunds, W.J., & Rhodes, C.J. (2003) Measles outbreaks in a population with declining vaccine uptake, *Science* **301**, 804.

[24] Stollenwerk, N., & Briggs, K.M. (2000) Master equation solution of a plant disease model, *Physics Letters* **A 274**, 84–91.

[25] Stollenwerk, N, Fuentes Sommer, P., Mateus, L., Kooi, B., & Aguiar, M. (2016) Hopf and torus bifurcations into chaos in mathematical population biology, *accepted for publication in "Ecological Complexity"*.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

New numerical methods for PDE models related to pricing and expected lifetime of an extraction project

María Suárez-Taboada[1] **and Carlos Vázquez**[1]

[1] *Dept. of Mathematics, University of A Coruña, Spain*

emails: mariasuarez@udc.es, carlosv@udc.es

Abstract

Mining companies face several uncertainties when they develop a mining project. There are many factors which provide randomness, such as the risk of political instability and the market fluctuations. These uncertainties will be reflected in the stock price of the commodity to be obtained, for example. In fact, variations of the price can be translated into the mine close down because it is no more profitable or just the opposite, the mining project can turn into a very profitable business. Other uncertain factors, such as the size of the resource inside the mine that can be extracted or the ore-grade of the mineral, also influence the value of the mining project and the probability of completing the extraction project.

In order to pose suitable mathematical models to obtain the value of the mining project, the probability of completing the extraction project and the expected life time, in [4] a methodology starting from Feynman-Kac theorem is followed. This methodology relates a expression of the problem solution in terms of expectations and the equivalent PDE formulation. Also the stochastic dynamics of the driving stochastic factors have to be chosen. For the commodity price we assume a geometric Brownian motion, while for the remaining resource we consider a constant extraction rate. In this way, the value of the mine with the option to close down is the solution of a complementarity problem associated to a PDE. When the ore-grade is assumed to be constant, the spatial variables of the PDE (which correspond to the considered stochastic factors) are the commodity price and the resource size. Therefore, the pricing problem involves a free boundary that represents the optimal abandonment boundary, which is part of the solution of the PDE problem. Moreover, the probability of project completion and the expected lifetime are the respective solutions of additional final-boundary value problems associated to the same differential operator that governs the mine value, although with different final and boundary conditions. We note that the resulting differential operator is highly

 ISBN: 978-84-617-8694-7

degenerated, as only the second order derivative with respect to the commodity price is not zero. In [4], the authors propose a first order characteristics method for time discretization combined with finite differences in space. For the complementarity problem, the propose a projected relaxation method. Also we note that some of the boundary conditions are not clear and properly justified.

In the present work, we propose a more rigorous statement of the appropriate boundary conditions. First, we follow the classical theory in [6] to determine the boundaries where boundary conditions are required. Secondly, in order to pose appropriate Dirichlet boundary conditions on some of these boundaries we apply some ideas already used in [2] for companies valuation models.

Next, taking into account the convection-dominating terms of the PDEs governing the mathematical models, we propose a higher order characteristics Crank-Nicholson method to discretize the time derivative jointly with the first order spatial derivative (i.e. the material derivative). For the spatial discretization related to the price and the remaining resource variables, we propose the use of piecewise quadratic Lagrange finite elements, so that the resulting fully discretization method falls into the framework of the so called Lagrange-Galerkin methods [1]. Additionally, in order to deal with the nonlinearity associated to the complementarity problem governing the mine pricing, the augmented Lagrangian active set method proposed from [5] is used. This set of numerical technique has been already successfully used in other problems related to pricing of pension plans and financial derivatives, such as in [3, 7, 8].

Finally, some numerical results will be shown in order to illustrate the performance of the proposed boundary conditions and numerical methods. More precisely, first an academic test with analytical solution illustrates the performance of the method. Secondly, a real mine example allows to obtain the expected behaviour for the mine value, the probability of completing the problem and the expected lifetime.

Key words: resources valuation problems, investment under uncertainty, complementarity problems, characteristics methods, finite elements, augmented Lagrangian active set methods

References

[1] A. Bermúdez, M. R. Nogueiras and C. Vázquez, *Numerical analysis of convection-diffusion-reaction problems with higher order characteristics finite elements. Part II: Fully discretized scheme and quadrature formulas*, SIAM Journal on Numerical Analysis, **44** (2006) 1854-1876.

[2] D. Castillo, A.M. Ferreiro, J.A. García-Rodríguez and C. Vázquez, *Numerical methods to solve PDE models for pricing business companies in different*

 ISBN: 978-84-617-8694-7

regimes and implementation in GPUs, Applied Mathematics and Computation, **219** (2013) 11233-11257.

[3] M.C. Calvo-Garrido, A. Pascucci and C. Vázquez, *Mathematical analysis and numerical methods for pricing pension plans allowing early retirement*, SIAM Journal of Applied Mathematics, **73** (2013) 17471767.

[4] G.W. Evatt, P.V. Johnson, P.W. Duck, S.D. Howell, J. Moriarty, *The expected lifetime of an extraction project* , Proceedings of the Royal Society, **467** (2011) 244-263.

[5] T.Kärkkäinen, K.Kunisch, P.Tarvainen, *Augmented Lagrangian active set methods for obstacle problems*, J. Optim. Theory Appl., **119** (2003) 499-533.

[6] O.A. Oleinik, E. V. Radkevic, *Second order equations with nonnegative characterisitc form*, A.M.S. And Plenum Press (1973).

[7] A. Pascucci, M. Suárez-Taboada and C. Vázquez, *Mathematical analysis and numerical methods for a PDE model governing ratchet-cap pricing problems*, Mathematical Models and Methods in Applied Sciences, **21**(2011) 1479-1498.

[8] A. Pascucci, M. Suárez-Taboada and C. Vázquez, *Mathematical analysis and numerical methods for a PDE model of a stock loan pricing problem*, Journal of Mathematical Analysis and Applications, **403** (2013) 38-53.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference on Computational and Mathematical Methods in Science and Engineering, CMMSE 2017 4–8 July, 2017.

NACA 2412 performance modification via using AFC.

Thorsten Summ[1], Bhanu Prakash[2], Josep M Bergada[3], Andreas Wierschem[1] and Fernando Mellibovsky[2]

[1] *Institute of Fluid Mechanics, Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU), Germany*

[2] *Physics Department, University Polytechnic of Catalunya, UPC, Spain*

[3] *Fluid Mechanics Department, University Polytechnic of Catalunya, UPC, Spain*

emails: thorsten.summ@gmx.de, bhanu2204@gmail.com, josep.m.bergada@upc.edu, andreas.wierschem@fau.de, fernando.mellibovsky@upc.edu

Abstract

The NACA 2412 profile was numerically studied via employing 2D-DNS and implementing Active Flow Control (AFC), the Reynolds number considered was 6757, being the angle of attack of 8°. Initially, the basic flow without implementing AFC was considered, the point in which the boundary layer separates as well as the y+ value along the profile length were evaluated. A single groove location, just before the separation point, was considered, periodic forcing was employed to both modify the location of the separation point and change the separation area where vortices are present. This was undertaken resulting in a reduction of the drag coefficient while increasing the lift. Via studying a set of frequencies and amplitudes linked with the AFC periodic actuation, it was obtained the optimum set of parameters to minimize the drag while maximizing the lift.

Key words: Active Flow Control, AFC, Periodic forcing, NACA 2412, Lift and Drag coefficients, DNS.

1 Introduction

Active Flow Control (AFC) is quite a recent technology used to modify fluid performance around bluff bodies. The essential idea behind AFC is to interact with the boundary layer in order to promote or delay flow separation, via doing so the downstream vortex

shedding amplitude and frequency can be modified and therefore lift and drag coefficients can be consequently affected. There are several ways to implement AFC in a bluff body, via using plasma actuators, via employing steady suction, steady blowing, periodic forcing and implementing zero net mass flow actuaors (ZNMFA) or using fluidic actuators (FA). Plasma actuation appears at the moment to have some difficulties in sufficiently interacting with the incoming flow [1, 2], then it is simply ionizing the lower parts of the boundary layer and therefore produces a narrow influence on the overal flow performance. Steady suction/blowing, are clearly giving the required energy to the boundary layer to generate a drastic modification of it, these two technologies can be clearly implemented in the vast majority of bluff bodies, obtaining drastic effects on vortex shedding, such effect deeply depends on the parameter defined as velocity ratio [3, 4, 5]. Probably, the only drawback of steady blowing/suction is the energy required to perform the required boundary layer delay, quite often relatively high velocity ratios are required and therefore the energy required for the AFC implementation is higher than the saved one when considering its beneficial effects. The technology which appears to be more efficient in overcoming such drawback is the use of periodic forcing. Periodic forcing interacts directly with the natural flow instabilities, such as the Kelvin-Helmholtz instability, shear layer flapping and ring-shaped wake structures [6]. Thus, employing a relatively small amount of energy, can dramatically affect the boundary layer. Nevertheless, in order to be able to interact efficiently with the natural shear layer instabilities, it is important to locate the groove, in which fluid is sucked in or blown out, precisely. Compared with the location of the groove whenever steady sucking is employed, which is used to modify the boundary layer separation point being located whether upstream or downstream and even at some distance of it [7], the use of periodic forcing as well as steady blowing, require a groove location near the separation point. To obtain optimum efficiency, the groove needs to be located just upstream of the boundary layer separation point whenever no AFC is employed. These main guide lines were followed in the present research. In what follows the numerical procedure followed for the numerical simulation will be outlined.

2 Numerical Model and boundary conditions

In any numerical simulation, the first step to be accomplished is to check the validity of the mesh employed. To do so, performing a grid independence test is always a good procedure, comparing the results obtained with experimental ones or previous researchers simulations, falls as well into the standard procedure, for the present case, the simulation will be validated via checking the y+ parameter. Whenever the maximum value of this parameter is smaller than one, it can be concluded that there are enough cells inside the laminar sub boundary layer, and therefore, the shear stresses nearby the walls are meant to be properly simulated. Three different meshes were evaluated in the present simulation, when a mesh

 ISBN: 978-84-617-8694-7

with 93686 cells was employed, the maximum y+ was 1, whenever the mesh was refined to 107166 cells the value of maximum y+ decreased to 0.666, and when using 120646 cells, the maximum y+ value decreased to 0.333. Although not presented in this paper, the dynamic lift coefficient was plotted for the three meshes employed and it was observed, that the results obtained when using the two most refined meshes were much closer than the dynamic result obtained when using the coarse mesh. Notice that via performing the previous comparison, in reality the authors make sure that the Strouhal number obtained from the simulation is also properly determined. As a conclusion of this preliminary study it was decided to use the mesh with 107166 cells to perform the rest of the simulations.

Regarding the boundary conditions used in the simulations, table 1 summarizes them. The fluid used for the simulation was air at standard conditions, its density was of $\rho = 1.225 kg/m^3$, being its absolute viscosity of $\mu = 1.48 * 10^{-5} m^2/s$, the chord of the profile was 1 m, the angle of attack and Reynolds number were kept constant, its values were respectively of $\alpha = 8^\circ$ and Re = 6757, fluid was considered as incompressible.

Table 1: Boundary conditions applied for the URANS simulation model. No-slip condition applied on the wing surface and constant velocity condition on the inlet boundary. Front and back planes are empty due to two-dimensional setup.

Boundaries	Parameter	Boundary conditions	Value
Inlet	Velocity	Dirichlet	$U_\infty = 0.1 m/s$
	Pressure	Neumann	$\frac{\partial p}{\partial n} = 0$
Outlet	Velocity	Neumann	$\frac{\partial u}{\partial n} = 0$
	Pressure	Dirichlet	p=0 Pa
Profile surface	Velocity	Dirichlet	u=0
	Pressure	Neumann	$\frac{\partial p}{\partial n} = 0$
Front and Back	Velocity	–	–
	Pressure	–	–

In what follows, initially the basic flow without implementing AFC will be considered, the next step will be to evaluate the wing profile performance whenever AFC is considered.

 ISBN: 978-84-617-8694-7

3 Results

3.1 Baseline case

In order to obtain the location where the boundary layer separates from the body, the wall shear stress for the baseline flow was determined and presented in figure 1. The separation point is defined as the point where the wall shear stress becomes zero, from figure 1 it can be observed that the laminar bubble separation is occurring at around 0.15 % chord. But figure 1 presents another two points where shear stresses suddenly drop, these points are located at about 60% and 81% of the chord. During the simulation it was noticed an small fluctuation of these minimum stress locations, the fluctuation was due to the boundary layer motion triggered by the flapping mechanism and vortex shedding. In reality, after the first separation point an elongated vortex, which turns at clockwise direction, appears, this first elongated vortex is called the laminar separation bubble, the location at which this initial vortex ends, and a second much smaller counterclockwise vortex begins, is where the shear stresses reach the second minimum. As presented in figure 2, this second vortex triggers a third one which begins at the third minimum stress point and ends at the end on the chord. This third vortex is also an elongated one and turns clockwise. It must be clarified that the shape, location and even the number of vortices appearing onto the wing profile, keep changing versus time which indicates the existence of vortex shedding. Vortex shedding is mostly linked with the appearance and separation of the second and third vortices, the laminar separation bubble, although fluctuates, is not being shed downstream.

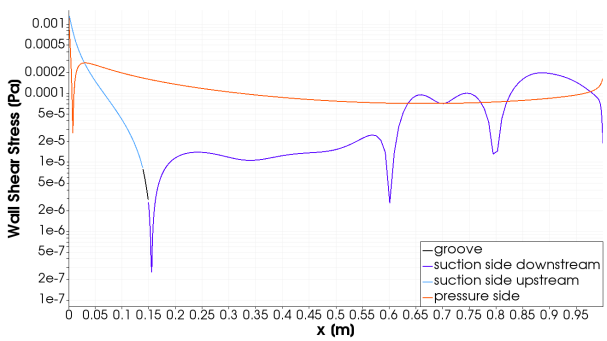


Figure 1: Wall shear stress vs. chord length of baseline case. The different colors show pressure and suction side of the wing as well as the location where flow control will be applied. Global minimum at 15 % of the chord represents flow separation point.

Among the information to be extracted from the baseline case, the vortex shedding frequency is particularly relevant, to obtain such frequency the fast fourier transform was applied to the temporal lift and drag coefficents, the result is presented in figure 3. It is

 ISBN: 978-84-617-8694-7

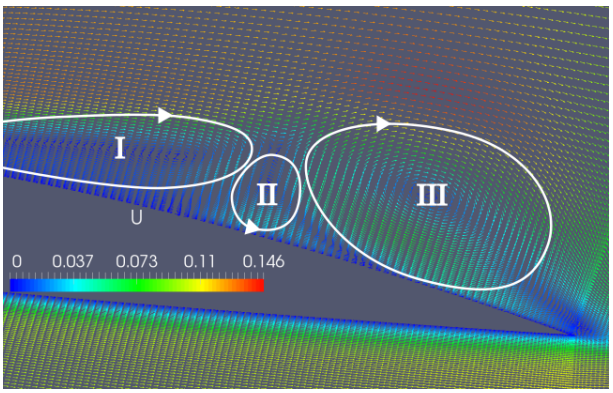


Figure 2: Main vortices appearing onto the wing surface shown in a velocity vector plot. Direction of rotation is shown by white arrows.

noticed that regardless of the coefficient chosen, the main frequency, which is the vortex shedding frequency, remains constant at 0.11 Hz. Accordng to previous studies [8, 9], among others, the frequency to be used whem employing periodic forcing to effectively interact with the unsteady shear layer, shall be around the natural vortex shedding frequency. A second parameter to be defined is the groove location, also based on previous studies [8, 9], it was found that whenever periodic forcing is employed, the most appropriate location for the groove is slightly upstream of the separation point, for the present simulation, the groove was located 1% of the chord (1 cm) upstream of the separation point. The groove width was as well 1% of the chord and the flow was injected perpendicular to the solid boundary.

3.2 Periodic forcing, frequency modification

In this section the effect of modifying the periodic forcing frequency will be explored. Five different non-dimensional frequencies were evaluated, F^+ 0.1; 0.5; 1; 5 and 10, $F^+ = 1$ characterizes the natural vortex shedding frequency obtained from the baseline case. The momentum coefficient was maintained constant and equal to $C_\mu = 1 * 10^{-3}$ for all these cases. The angle of attack $\alpha = 8^\circ$ was maintained constant as well during the entire study. In order to properly quantify the wing profile performance, temporal averaged lift and drag coefficients were recorded for each case and compared with the homologous obtained in the baseline case. A parameter defined as variation of performance, which compares the lift and drag coefficients increments versus the lift and drag coefficients obtained in the baseline case, see equation 1, was also employed to accurately determine the Active Flow Control advantages.

 ISBN: 978-84-617-8694-7

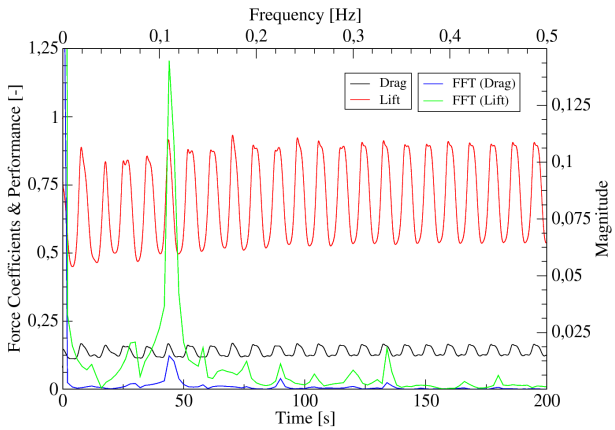


Figure 3: Fast Fourier Transformation (FFT) of lift coefficient C_L and drag coefficient C_D temporal signals. Main frequency at 0.11 Hz indicated by a significant peak of the green and blue curve.

$$\Psi = \frac{\Delta \frac{C_L}{C_D}}{\frac{C_{L_0}}{C_{D_0}}} \tag{1}$$

C_{L_0} and C_{D_0} represent the temporal averaged lift and drag coefficients from the baseline case.

Table 2 presents the lift and drag coefficients variation versus the respective baseline case values, the variation of the performance parameter, defined in equation 1 is also introduced. From this table it can clearly be seen that whenever the periodic forcing frequency is the same as the natural vortex shedding one, the wing performance is optimal, having the maximum lift and minimum drag, these results clearly agree with many of the former researchers undertaken, see for example [9, 10]. It is also interesting to observe, that regardless of the frequency used, a drag decrease is generated, although such decrease is particularly relevant at periodic forcing frequencies equal to the vortex shedding one or even half of it.

3.3 Periodic forcing, amplitude modification

Once obtained the most appropriate frequency to be employed, the next step was to evaluate the effect of the periodic forcing amplitude. For these new runs, the forcing frequency was maintained constant and equal to the vortex shedding one, five different forcing amplitudes were considered, which once non-dimensionalized in the form of the momentum coefficient, gave the following values $C_\mu = 1 * 10^{-4}; 5 * 10^{-4}; 1 * 10^{-3}; 5 * 10^{-3}; 1 * 10^{-2}$. The results

 ISBN: 978-84-617-8694-7

Table 2: Variation of wing parameters when modifying periodic forcing frequency resp. non-dimensional frequency. Relative change in performance, lift and drag compared to the values of baseline simulation.

$C_\mu = 1 * 10^{-3}$	Ψ	$\frac{\Delta C_L}{C_{L_0}}$	$\frac{\Delta C_D}{C_{D_0}}$
$F^+ = 0.1$	16.8%	8.3%	-2.4%
$F^+ = 0.5$	57.4%	12.2%	-26.2%
$F^+ = 1$	80.3%	12.8%	-35.2%
$F^+ = 5$	-2.5%	-9.8%	-5.8%
$F^+ = 10$	-0.8%	-1.9%	-0.6%

obtained from this second set of simulations are presented in table 3, where the same comparison paramenters used in table 2 were evaluated. When looking at table 3, it is very interesting to realize that all momentum coefficients used, generate a very high decrease on the drag coefficient, the maximum increase in lift it is obtained when using the minimum momentum coefficient, it is also interesting to notice that once a certain forcing amplitude was overcome, the lift coefficient drops to negative values and keeps decreasing as the forcing amplitude increases, this particular effect is very likely to be caused due to boundary layer tripping, the periodic forcing amplitude is higher than the boundary layer thickness at the injection point, therefore the energy associated to the pulsating flow is released outside the boundary layer, promoting stall.

Table 3: Variation of wing parameters when modifying periodic forcing velocity amplitude resp. momentum coefficient. Relative change in performance, lift and drag compared to the values of baseline simulation.

$F^+ = 1.0$	Ψ	$\frac{\Delta C_L}{C_{L_0}}$	$\frac{\Delta C_D}{C_{D_0}}$
$C_\mu = 1 * 10^{-4}$	66.9%	24.4%	-22.9%
$C_\mu = 5 * 10^{-4}$	73%	12%	-33%
$C_\mu = 1 * 10^{-3}$	80.3%	12.8%	-35.2%
$C_\mu = 5 * 10^{-3}$	61.8%	-2.5%	-37.5%
$C_\mu = 1 * 10^{-2}$	50.8%	-11.9%	-39.9%

 ISBN: 978-84-617-8694-7

3.4 Combining frequency and amplitude

In the present subsection, the combination of all results obtained from the previous cases will be introduced. In tables 4, 5 and 6, the values of the wing performance parameter Ψ, the lift gain $\frac{\Delta C_L}{C_{L_0}}$ and the drag gain $\frac{\Delta C_D}{C_{D_0}}$, will be respectively presented as a function of all the non dimensional frequencies and momentum coefficients evaluated. From these tables it is observed that the optimum wing performance increase $\Psi = 80.3\%$ it is obtained for the following set of conditions $F^+ = 1$ and $C_\mu = 1 * 10^{-3}$, lift and drag increase are respectively of 12.8% and -35.2%. Observing table 6, it is realized that the conditions required to obtain minimum drag are $F^+ = 1$ and $C_\mu = 1 * 10^{-2}$, if on the other hand, lift is to be maximized, according to table 5, the optimum set of parameters are $F^+ = 1$ and $C_\mu = 1 * 10^{-4}$. The conclusion is that employing a frequency matching the natural vortex shedding one, deeply affects the downstream flow performance, the variation of the forcing amplitude will generate weather a maximum lift, for small amplitudes, or a minimum drag, for big amplitudes, or even an optimum wing performance parameter, obtained for medium amplitudes. Whenever the periodic forcing frequency is half the vortex shedding one, there still appear interesting increases in lift and decreases in drag, therefore good performance parameters are to be found. The rest of the frequencies employed have a much smaller effect on the flow performance.

Table 4: Variation of wing performance when modifying non dimensional frequency and momentum coefficient. Maxima in performance increase arise when $F^+ = 1$. A significant drop can be seen when $F^+ > 1$.

Ψ	$C_\mu = 1 * 10^{-4}$	$C_\mu = 5 * 10^{-4}$	$C_\mu = 1 * 10^{-3}$	$C_\mu = 5 * 10^{-3}$	$C_\mu = 1 * 10^{-2}$
$F^+ = 0.1$	-6.6%	4%	16.8%	28.4%	26.7%
$F^+ = 0.5$	14.7%	28.1%	57.4%	68.5%	64.2%
$F^+ = 1.0$	66.9%	73%	80.3%	61.8%	50.8%
$F^+ = 5.0$	0.9%	-0.7%	-2.5%	1.4%	6.2%
$F^+ = 10.0$	0%	-0.1%	-0.8%	-4%	-5.1%

In order to visualize the advantages obtained when employing periodic forcing and using the optimum set of parameters, figure 4 compares the velocity distribution and pressure field around the NACA profile for the baseline case and for the optimum forcing conditions previously set. Clearly a drastic reduction of the number of vortices and its dimension is to be spotted whenever AFC is employed, notice as well that the pressure onto the upper side of the profile suffers an interesting reduction when periodic forcing is used.

 ISBN: 978-84-617-8694-7

Table 5: Variation of lift coefficient when modifying non dimensional frequency and momentum coefficient.

$\frac{\Delta C_L}{C_{L_0}}$	$C_\mu = 1 * 10^{-4}$	$C_\mu = 5 * 10^{-4}$	$C_\mu = 1 * 10^{-3}$	$C_\mu = 5 * 10^{-3}$	$C_\mu = 1 * 10^{-2}$
$F^+ = 0.1$	-21.8%	-5.4%	8.3%	12.5%	7.1%
$F^+ = 0.5$	17.3%	5.9%	12.2%	7.6%	5.5%
$F^+ = 1.0$	24.4%	12%	12.8%	-2.5%	-11.9%
$F^+ = 5.0$	-0.7%	-4.8%	-9.8%	-18.4%	-19.9%
$F^+ = 10.0$	-2.7%	-2.1%	-1.9%	-10.3%	-14.6%

Table 6: Variation of drag coefficient when modifying non-dimensional frequency and momentum coefficient.

$\frac{\Delta C_D}{C_{D_0}}$	$C_\mu = 1 * 10^{-4}$	$C_\mu = 5 * 10^{-4}$	$C_\mu = 1 * 10^{-3}$	$C_\mu = 5 * 10^{-3}$	$C_\mu = 1 * 10^{-2}$
$F^+ = 0.1$	-9.3%	-4.2%	-2.4%	-3.6%	-5%
$F^+ = 0.5$	5%	-15.3%	-26.2%	-33.9%	-33.7%
$F^+ = 1.0$	-22.9%	-33%	-35.2%	-37.5%	-39.9%
$F^+ = 5.0$	0.2%	-3%	-5.8%	-17.4%	-22%
$F^+ = 10.0$	-0.4%	-0.2%	-0.6%	-5.4%	-9%

4 Conclusions

A 2D-DNS simulation over the NACA 2412 profile at Re=6757 and for an angle of attack of 8° was undertaken. A baseline case was performed to obtain the natural vortex shedding flow, especially the natural vortex shedding frequency. The baseline case results were compared with the ones obtained when applying Active Flow Control and via periodic forcing. A set of five different frequencies and amplitudes were evaluated obtaining the optimum conditions required to maximize lift, minimize drag and to maximize the wing performance parameter. The optimum conditions were obtained whenever the forcing frequency was of the same order as the natural vortex shedding frequency. When maintaining the optimum forcing frequency and modifying the forcing amplitude, maximum lift, or minimum drag or maximum wing performance parameter could be obtained. Whenever the pulsating frequency employed was 50% of the natural one, the wing performance, lift increase and drag decrease were also quite relevant.

 ISBN: 978-84-617-8694-7

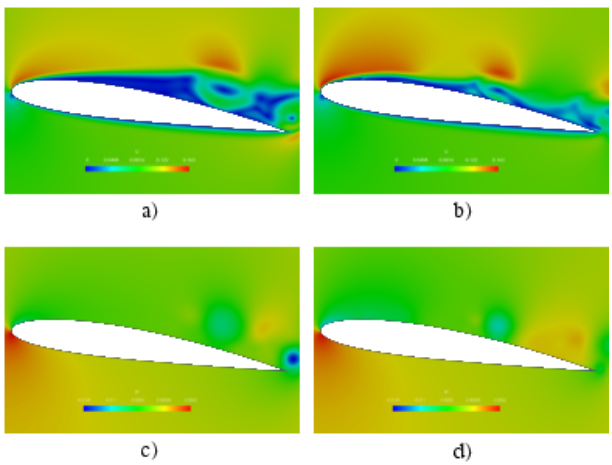


Figure 4: Velocity and pressure fields around the wing profile under study, comparison between the baseline case, figures a) and c), and the optimum periodic forcing case $F^+ = 1$; $C_\mu = 1 * 10^{-3}$, figures b) and d)

Acknowledgements

This work has been partially supported by an European ERASMUS+ scholarship, Which allowed Mr Thorsten Summ to undertake a project at the Universitat Politcnica de Catalunya, Barcelona Spain.

References

[1] Louis N Cattafesta III and Mark Sheplak. Actuators for active flow control. *Annual Review of Fluid Mechanics*, 43:247–272, 2011.

[2] Nicolò Fabbiane, Shervin Bagheri, and Dan S Henningson. Energy efficiency and performance limitations of linear adaptive control for transition delay. *Journal of Fluid Mechanics*, 810:60–81, 2017.

[3] Hanns Friedrich Müller-Vahl, Christian Navid Nayeri, Christian Oliver Paschereit, and David Greenblatt. Dynamic stall control via adaptive blowing. *Renewable Energy*, 97:47–64, 2016.

[4] R Radespiel, M Burnazzi, M Casper, and P Scholz. Active flow control for high lift with steady blowing. *The Aeronautical Journal*, 120(1223):171–200, 2016.

 ISBN: 978-84-617-8694-7

[5] Alice B Thompson, Dmitri Tseluiko, and Demetrios T Papageorgiou. Falling liquid films with blowing and suction. *Journal of Fluid Mechanics*, 787:292–330, 2016.

[6] Pierric Joseph, Xavier Amandolèse, and Jean-Luc Aider. Drag reduction on the 25 slant angle ahmed reference body using pulsed jets. *Experiments in fluids*, 52(5):1169–1185, 2012.

[7] Bhanu Prakash Reddy Samala. Numerical investigation of active flow control applied to an airfoil leading edge. Master's thesis, Universitat Politecnica de Catalunya, 2015.

[8] Sebastian D Goodfellow, Serhiy Yarusevych, and Pierre E Sullivan. Momentum coefficient as a parameter for aerodynamic flow control with synthetic jets. *AIAA journal*, 51(3):623–631, 2012.

[9] David Greenblatt and Israel J Wygnanski. The control of flow separation by periodic excitation. *Progress in aerospace Sciences*, 36(7):487–545, 2000.

[10] Michael Amitay and Ari Glezer. Role of actuation frequency in controlled flow reattachment over a stalled airfoil. *AIAA journal*, 40(2):209–216, 2002.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

Padé Approximants to Conicality Based Probabilistic Evolution Theory (PREVTH) Solutions: Two Classical Particles Systems Interacting via Central Forces

Elif TATAROĞLU[1] **and Metin DEMİRALP**[1]

[1] *Computational Science and Engineering Department, Informatics Institute, İstanbul Technical University*

emails: elif.istanbuuul@gmail.com, metin.demiralp@gmail.com

Abstract

This work focuses on the construction of various Padé approximants to conicality based Probabilistic Evolution Theory (PREVTH) solutions. PREVTH has been developed for conicality based explicit ODE(s) in Demiralp group studies in the last decade. Two classical particles system interacting via central forces has been taken as the targets. Paper is in conceptual level. Confirmative implementations will be given during the presentations in the conference.

Key words: PREVTH, Padé approximants, Two particle systems.

1 Introduction

A two particle system in Classical Mechanics has totally twelve unknown temporal function such that the positions of two particles in three dimensional physical Cartesian space can be described by totally six unknown temporally varying functions while the velocities of those two particles are also described by six temporally varying unknown functions. The motion's instances are defined by a single scalar independent variable, t, which is called "time". The equations of motion for this system is defined through equations amongst the partial derivative of a system function describing the total energy of the system. This function depends on position and momentum unknowns of the particles and does not change its explicit time dependence as long as the system is not under any external influence. The resulting equations are first order ODEs on the unknowns positions and momenta hence they are composed of twelve ODEs. Even though this may be considered quite complicated

 ISBN: 978-84-617-8694-7

the mass center positions and their relevant momentum unknowns (totally six unknowns) can be separated out as long as the other unknowns are organized as relative coordinates (the coordinates of one particle relative to the other). The equations of mass center describe the motion of a free particle under no external influence and they can be solved analytically. We do not intend to give those solutions and their interpretation.

Now by considering the following differences between positions as new relative position unknowns and by defining the following distance function to facilitate the analysis of relative motion

$$x_j(t) \equiv q_j(t) - q_{j+3}(t), \quad j = 1, 2, 3; \qquad r(t) \equiv \sqrt{x_1(t)^2 + x_2(t)^2 + x_3(t)^2} \tag{1}$$

we can construct the following equations of relative motion

$$\frac{dx_j(t)}{dt} = \frac{1}{\mu}\pi_j(t), \qquad \frac{d\pi_j(t)}{dt} = -V'\left(r(t)\right)\frac{x_j(t)}{r(t)}, \qquad j = 1, 2, 3 \tag{2}$$

where $1/\mu = 1/m_1 + 1/m_2$ and the relative positions and momenta are denoted by indexed $x(t)$ and $\pi(t)$s respectively. The Hamiltonian of the system characterized by relative motion can be given as follows

$$H\left(\boldsymbol{\pi}(t), r(t)\right) = \frac{1}{2\mu}\left(\pi_1(t)^2 + \pi_2(t)^2 + \pi_3(t)^2\right) + V\left(r(t)\right) \tag{3}$$

where V stands for the potential function of the considered system.

If the initial position and momentum vectors accompanying to (2) are linearly dependent then it is possible to prove that the solution vectors of (2) remain linearly dependent at every instant of the motion. In other words, the motion happens to limited on a straight line during the whole evolution including the initial moment. We do not intend to give further details for this case.

Abovementioned situation is a quite limited case. However, there is a more general issue dictating us that the system's motion happens to stay in the plane where the initial position and momentum vectors reside at the all time instances of the evolution. For this case we need to assume that the initial position and momentum vectors are linearly independent. We do not also intend to give further details of this issue since curios readers can get information even from some undergarduate textbooks.

Above analysis implies that the two particle system's motion can be described in the plane where the initial position and momentum vectors reside. This means there will be just four unknowns and four ODEs with appropriate initial conditions. We can write these planar equations of motions as follows

$$\begin{aligned}
&\frac{d\widetilde{x}_1}{dt}(t) = \frac{1}{\mu}\widetilde{\pi}_1(t), \qquad \widetilde{x}_1(0) = x_{0,1}, \qquad \frac{d\widetilde{\pi}_1}{dt}(t) = -\frac{V'\left(r(t)\right)}{r(t)}\widetilde{x}_1(t), \qquad \widetilde{\pi}_1(0) = \pi_{0,1} \\
&\frac{d\widetilde{x}_2}{dt}(t) = \frac{1}{\mu}\widetilde{\pi}_2(t), \qquad \widetilde{x}_2(0) = x_{0,2}, \qquad \frac{d\widetilde{\pi}_2}{dt}(t) = -\frac{V'\left(r(t)\right)}{r(t)}\widetilde{x}_1(t), \qquad \widetilde{\pi}_2(0) = \pi_{0,2}
\end{aligned} \tag{4}$$

 ISBN: 978-84-617-8694-7

$$r(t) = \sqrt{\widetilde{x}_1(t)^2 + \widetilde{x}_2(t)^2} \tag{5}$$

The sole dependence of the potential on a single distance function reminds us to use the circular (polar) coordinates. We can write

$$\widetilde{x}_1(t) = r(t)\cos\vartheta(t), \qquad \widetilde{x}_2(t) = r(t)\sin\vartheta(t) \tag{6}$$

$$r(t) = \sqrt{\widetilde{x}_1(t)^2 + \widetilde{x}_2(t)^2}, \qquad \vartheta(t) = \arctan\left(\frac{\widetilde{x}_2(t)}{\widetilde{x}_1(t)}\right) \tag{7}$$

$$\mathbf{e}_r(t) \equiv \frac{\nabla r(t)}{\|\nabla r(t)\|} = \begin{bmatrix} \cos\vartheta(t) \\ \sin\vartheta(t) \end{bmatrix}, \qquad \mathbf{e}_\vartheta(t) \equiv \frac{\nabla\vartheta(t)}{\|\nabla\vartheta(t)\|} = \begin{bmatrix} -\sin\vartheta(t) \\ \cos\vartheta(t) \end{bmatrix} \tag{8}$$

The new defined unit vectors vary in time so they have non-vanishing temporal derivatives. We can write,

$$\frac{d\mathbf{e}_r(t)}{dt} = \vartheta'(t)\mathbf{e}_\vartheta(t), \qquad \frac{d\mathbf{e}_\vartheta(t)}{dt} = -\vartheta'(t)\mathbf{e}_r(t) \tag{9}$$

The angular momentum square can be written as follows,

$$\|\widetilde{\mathbf{x}}(t) \times \widetilde{\boldsymbol{\pi}}(t)\|^2 = \mu\left(1 + \tan\vartheta(t)^2\right) r(t)^2\cos\vartheta(t)^2\vartheta'(t) = \mu r(t)^2\vartheta'(t) \tag{10}$$

$$p_\vartheta(t) \equiv \mu r(t)^2\vartheta'(t) \equiv \mu c_1 \tag{11}$$

where we have used the angular momentum conservation law.

We do not intend to get into the further details of angular momentum issue. However, we can use the angular momentum conservation law here in construction of the equations of motion in circular motion. Now we can write all the equations of unknowns as follows.

$$r'(t) = \frac{p_r(t)}{\mu}, \qquad r(0) = r_0 \tag{12}$$

$$p_r'(t) = \mu r(t)\vartheta'(t)^2 - V'(r(t)), \qquad p_r(0) = p_{r,0} \tag{13}$$

$$\vartheta'(t) = \frac{c_1}{r(t)^2}, \qquad \vartheta(0) = \vartheta_0 \tag{14}$$

$$p_\vartheta' = 0, \qquad p_\vartheta(0) = \mu c_1 \tag{15}$$

 ISBN: 978-84-617-8694-7

Radial momentum and angular momentum give us the below equations.

$$r'(t) = \frac{p_r(t)}{\mu}, \qquad r(0) = r_0 \tag{16}$$

$$p_r'(t) = \frac{\mu c_1^2}{r(t)^3} - V'(r(t)), \qquad p_r(0) = p_{r,0} \tag{17}$$

$$\vartheta'(t) = \frac{c_1}{r(t)^2}, \qquad \vartheta(0) = \vartheta_0 \tag{18}$$

$$p_\vartheta' = 0, \qquad p_\vartheta(0) = \mu c_1 \tag{19}$$

We need to solve these equations. To do so we take certain elimination steps and get

$$\frac{dr}{dt} = \pm\sqrt{\frac{2}{\mu}}\sqrt{E - \frac{\mu c_1^2}{2}\frac{1}{r^2} - V(r)} \tag{20}$$

This can be solved by using certain well known approaches as long as the structure of the potential permits us. We do not intend to give further details since they are well known for someones having sufficiently strong background.

On the other hand, instead of the solution of this equation one can seek a relation between the polar coordinate evolutions. We can give the result without repeating the derivation stages of the trajectory as follows

$$\frac{1}{r(\vartheta)} = \left(\frac{1}{r_0} - \frac{\nu}{\mu c_1^2}\right)\cos(\vartheta - \vartheta_0) - \frac{p_{r,0}}{\mu c_1}\sin(\vartheta - \vartheta_0) + \frac{\nu}{\mu c_1^2} \tag{21}$$

where the function $V(r)$ has been specified as the potential function of a system interacted via a central force. Hence we have used the equality $V(r) = -\nu/r$ where ν is the product of the system particles' masses and univariate gravitational constant ($\nu \equiv gm_1m_2$). All entities having a zero in their indices correspond to an initial value. c_1 is the constant specifying angular momentum conservation while μ stands for the previously defined reduced mass of the two particles system.

This formula can be put into more amenable form by defining

$$\overline{c} \equiv \frac{\nu}{\mu c_1^2}, \qquad 1 - \frac{\mu c_1^2}{\nu r_0} \equiv \varepsilon\cos\phi, \qquad \frac{c_1 p_{r,0}}{\nu} \equiv \varepsilon\sin\phi \tag{22}$$

and writing

$$\varepsilon = \sqrt{\left(1 - \frac{\mu c_1^2}{\nu r_0}\right)^2 + \frac{c_1^2 p_{r,0}^2}{\nu^2}}, \qquad \phi = \tan^{-1}\left(\frac{c_1 p_{r,0} r_0}{\nu r_0 - \mu c_1^2}\right) \tag{23}$$

 ISBN: 978-84-617-8694-7

By using these definitions, we can rewrite circular motion trajectory in a more simple (compact) form as follows.

$$\frac{1}{r\left(\vartheta\right)} = \overline{c}\left[\,1 - \varepsilon\cos\left(\vartheta - \vartheta_0 - \phi\right)\right] \tag{24}$$

In this formula ε is expected to be positive without exceeding 1. This means that the following inequalities hold

$$\varepsilon \leq 1 \Longrightarrow \varepsilon^2 \leq 1 \implies \left(1 - \frac{\mu c_1^2}{\nu r_0}\right)^2 + \frac{c_1^2 p_{r,0}^2}{\nu^2} \leq 1 \Longrightarrow \frac{\mu^2 c_1^2}{\nu^2 r_0^2} - \left(\frac{2\mu}{\nu r_0} - \frac{p_{r,0}^2}{\nu^2}\right) \leq 0 \tag{25}$$

$$-\sqrt{2 - \frac{r_0 p_{r,0}^2}{\mu\nu}} \leq c_1 \leq \sqrt{2 - \frac{r_0 p_{r,0}^2}{\mu\nu}} \Longrightarrow -\sqrt{2\mu^2 - \frac{\mu r_0 p_{r,0}^2}{\nu}} \leq p_{\vartheta,0} \leq \sqrt{2\mu^2 - \frac{\mu r_0 p_{r,0}^2}{\nu}} \tag{26}$$

We find this level of recalling information sufficient for this moment.

2 Constructing Probabilistic Evolution Theory Equations (PREVTH)

Now we take the following two ODEs as the target of PREVTH

$$\begin{aligned} r'(t) &= \frac{p_r(t)}{\mu}, \qquad r(0) = r_0 \\ p_r'(t) &= \frac{p_{\vartheta,0}^2}{\mu}\frac{1}{r(t)^3} - \frac{\nu}{r(t)^2}, \qquad p_r(0) = p_{r,0} \end{aligned} \tag{27}$$

Even though PREVTH expects conicality (second degree multinomiality) at the right hand side of target ODEs [8], these ODEs do not have such type of right hand sides. However it is possible to use space extension [5–7] herein.

$$u_1(t) \equiv \frac{1}{r(t)}, \quad u_2(t) \equiv \frac{1}{r(t)^2}, \quad u_3(t) \equiv \frac{p_r(t)}{r(t)}, \quad u_4(t) \equiv p_r(t)^2 \tag{28}$$

These new unknowns are added to two original ODEs one by one. First we add $u_1(t)$ as a new unknown temporal function. This addition realized by starting with the differentiation of both sides of first equation in (28) this gives an ODE whose right hand side contains the first derivative of previously existing unknown function $r(t)$. This derivative can be replaced by its equivalent in the previously existing ODE for $r(t)$. This replacement enables us to get a right hand side depending on $r(t)$, $p_r(t)$, $u_1(t)$. However conicality still has not been obtained and urges us to define the new additional unknown $u_2(t)$. Then $u_2(t)$ can be

similarly added to the obtained three ODEs for $r(t)$, $p_r(t)$, $u_1(t)$. The right hand sides of the obtained four ODEs for $r(t)$, $p_r(t)$, $u_1(t)$, $u_2(t)$ still do not have conicality. This urges us to continue in the same way for adding $u_3(t)$ and finally $u_4(t)$. At that addition (space extension) we arrive at

$$\begin{array}{ll}
r'(t) = \frac{1}{\mu} p_r(t), & r(0) = r_0 \\
p_r'(t) = \frac{p_{\vartheta,0}^2}{\mu} u_1(t) u_2(t) - \nu u_2(t), & p_r(0) = p_{r,0} \\
u_1'(t) = -\frac{1}{\mu} u_2(t) p_r(t), & u_1(0) = \frac{1}{r_0} \\
u_2'(t) = -\frac{2}{\mu} u_2(t) u_3(t), & u_2(0) = \frac{1}{r_0^2} \\
u_3'(t) = -\frac{1}{\mu} u_2(t) u_4(t) + \frac{1}{\mu} p_{\vartheta,0}^2 u_2(t)^2 + \nu u_1(t) u_2(t), & u_3(0) = \frac{p_{r,0}}{r_0} \\
u_4'(t) = \frac{2 p_{\vartheta,0}^2}{\mu} u_2(t) u_3(t) - 2\nu u_1(t) u_3(t), & u_4(0) = p_r^2
\end{array} \tag{29}$$

which are in conical structure and appropriate for the application of PREVTH [9]. We can now rewrite these equations in the following concise form

$$\dot{\mathbf{x}}(t) = \mathbf{F}_0 + \mathbf{F}_1 \mathbf{x}(t) + \mathbf{F}_2 \mathbf{x}(t)^{\otimes 2} \tag{30}$$

where

$$\mathbf{x}(t) = \left[\, r(t) \;\; p_r(t) \;\; u_1(t) \;\; u_2(t) \;\; u_3(t) \;\; u_4(t) \, \right]^T \tag{31}$$

For our two particle system $\mathbf{F}_0 = \mathbf{0}_6$ and $\mathbf{F}_1$ is composed of two outer products from six dimensional Cartesian space standard unit vectors while $\mathbf{F}_2$ is a rectangular matrix of 6×36 type. We do not intend to present the structure of $\mathbf{F}_2$ explicitly since we do not find it necessary.

3 Constancy Adding Space Extension (CASE)

The conical structure obtained in the previous section is not the ultimate form on which PREVTH is applied. The desired form is again conical but the constant vector is desired to be vanished while the desired form of linear term coefficient is proportional to the identity matrix whose number of rows and columns is identical to the element number of the system vector. These desired forms can be produced by using very specific Constancy Adding Space Extension (CASE) which adds just a constant function as a new unknown. We can write the following equalities

$$\mathbf{x}_{aug}(t) \equiv \left[\, \mathbf{x}(t)^T \;\; a \, \right]^T, \qquad \mathbf{x}_{aug}(t)^{\otimes 2} = \mathbf{P}^{-1} \left[\mathbf{x}^{\otimes 2^T} \;\; a\mathbf{x}^T \;\; a\mathbf{x}^T \;\; a^2 \right]^T \tag{32}$$

where the unitary matrix $\mathbf{P}$ stands for an appropriate permutation matrix of 49×49. These equalities are desired to take us to the following ODEs with specific conicality

$$\dot{\mathbf{x}}_{aug}(t) = \beta \mathbf{x}_{aug} + \mathbf{F}_{aug} \mathbf{x}_{aug}^{\otimes 2} = \beta \mathbf{x}_{aug} + \mathbf{G} \left[\mathbf{x}^{\otimes 2^T} \;\; a\mathbf{x}^T \;\; a\mathbf{x}^T \;\; a^2 \right]^T, \qquad \mathbf{G} \equiv \mathbf{F}_{aug} \mathbf{P}^{-1} \tag{33}$$

 ISBN: 978-84-617-8694-7

where β stands for a scalar which is arbitrary at this moment while augmented rectangular matrices, $\mathbf{F}_{aug}$, and $\mathbf{G}$, denote 7×49 type. $\mathbf{G}$'s blocks in the partitioning in accordance with the Kronecker square of the augmented system vector, $\mathbf{x}^{\otimes 2}$ should have the following equivalents to get harmony with the CASE

$$\begin{aligned} &\mathbf{G}_{1,1} \equiv \mathbf{F}_2, \qquad \beta \mathbf{I}_6 + a\mathbf{G}_{1,2} + a\mathbf{G}_{1,3} = \mathbf{F}_1, \qquad \mathbf{G}_{1,4} = \frac{1}{a^2}\mathbf{F}_0 \\ &\mathbf{G}_{2,1} = \mathbf{G}_{2,2} = \mathbf{G}_{2,3} = 0, \qquad G_{2,4} = \frac{\beta}{a} \end{aligned} \tag{34}$$

4 Kronecker Power Series Solution via Probabilistic Evolution Theory (PREVTH)

The first part of (33) can be solved by the following Kronecker power series

$$\mathbf{x}_{aug}(t) = \mathrm{e}^{\beta t} \sum_{j=0}^{\infty} \frac{u^j}{j!} \mathbf{T}_j \mathbf{a}^{\otimes (j+1)} \tag{35}$$

where $\mathbf{T}_j$ stands for the jth Telescope Matrix which is defined as the product of some Monocular Matrices in the following equalities

$$\mathbf{T}_j \equiv \prod_{k=1} \mathbf{M}_k, \qquad \mathbf{M}_k \equiv \sum_{\ell=0}^{k-1} \mathbf{I}_7^{\otimes \ell} \otimes \mathbf{F} \otimes \mathbf{I}^{\otimes (k-1-\ell)} \tag{36}$$

In the above Kronecker power solution the initial vector has been symbolized by $\mathbf{a}$. That Kronecker power solution can be reexpressed as follows

$$\mathbf{x}_{aug}(t) = \mathrm{e}^{\beta t} \sum_{j=0}^{\infty} \frac{u^j}{j!} \mathbf{v}_j(\mathbf{a}), \quad \mathbf{v}_j(\mathbf{a}) \equiv \mathbf{T}_j \mathbf{a}^{\otimes j} \tag{37}$$

where the vectors, $\mathbf{v}_j(\mathbf{a})$s, satisfy the following recursion as we have recently proven.

$$\mathbf{v}_j(\mathbf{a}) = \sum_{k=0}^{j-1} \binom{j-1}{k} \lfloor \mathbf{F}_{aug}, \mathbf{v}_k \rceil \mathbf{v}_{j-k-1}, \quad j = 1, 2, 3, ...; \quad \mathbf{v}_0(\mathbf{a}) = \mathbf{a} \tag{38}$$

where we have used the squarification and SquTelMat concepts which are defined through the following equalities

$$\mathbf{F} \equiv \left[\mathbf{F}^{(1)} \quad ... \quad \mathbf{F}^{(n+1)}, \right], \qquad \lfloor \mathbf{F}, \boldsymbol{\xi}_k \rceil = \sum_{i=1}^{n+1} \left(\mathbf{e}_i^T \boldsymbol{\xi}_k \right) \mathbf{F}^{(i)} \tag{39}$$

 ISBN: 978-84-617-8694-7

where a linear combination of the square blocks is formed using the vector elements as coefficients. Now, truncations from (37) may be used in order to approximate the solution of the initial value problem.

After the Constancy Adding Space Extension (CASE) application we have told above, we obtain the non-zero elements of $\mathbf{F}_{aug}$ as follows

$$\begin{aligned}\mathbf{F}_{aug} &= \frac{1}{2}\mathbf{e}_1\mathbf{e}_{28}^T + \frac{1}{2}\mathbf{e}_1\mathbf{e}_{46}^T - \frac{1}{2}\mathbf{e}_2\mathbf{e}_{18}^T - \frac{1}{2}\mathbf{e}_2\mathbf{e}_{24}^T - \mathbf{e}_3\mathbf{e}_{20}^T - \mathbf{e}_3\mathbf{e}_{38}^T + \frac{1}{2}\mathbf{e}_4\mathbf{e}_{10}^T + \frac{1}{2}\mathbf{e}_4\mathbf{e}_{16}^T \\ &\quad -\mathbf{e}_4\mathbf{e}_{18}^T - \mathbf{e}_4\mathbf{e}_{24}^T - 2\mathbf{e}_5\mathbf{e}_{19}^T + \mathbf{e}_5\mathbf{e}_{20}^T - 2\mathbf{e}_5\mathbf{e}_{31}^T + \mathbf{e}_5\mathbf{e}_{38}^T + \mathbf{e}_6\mathbf{e}_{17}^T - \frac{1}{2}\mathbf{e}_6\mathbf{e}_{19}^T \\ &\quad -\mathbf{e}_6\mathbf{e}_{20}^T - \mathbf{e}_6\mathbf{e}_{31}^T \end{aligned} \tag{40}$$

5 Padé Approximants of PREVTH Solution Components

The main purpose of this paper is to construct Certain Padé approximants to the PREVTH solutions mentioned in the previous section. There are various alternatives to construct Padé approximants: (i) Each element of PREVTH solution vector can be individually approximated by certain Padé approximants; (ii) The norm of the PREVTH vector solution can be Padé approximated as a whole approximations; (iii) The inner product of the PREVTH solution vector with another given (mostly constant) vector can also Padé approximated. The examples can be populated as desired.

We do not give any detail for the Padé approximant construction since it is a quite well-known issue. Beyond that we do not intend to give any symbolical or numerical results even though we have results of many implementation for confirmative purposes. This is because of the time and paper size limitations. We are going to realize a sufficiently comprehensive implementative during the presentation. However, we can now say that the results are very parallel to what we have expected in the convergence properties.

6 Concluding Remarks

As we can see from the last equation, after applying PREVTH, we have 6 ODEs which have at most second degree multinomials on the right hand side and these are all solvable. In this work, we have shown that the structure of the two particles celestial mechanical systems allow us to apply Probabilistic Evolution Theory(PREVTH). We are going to give more plots and implementation during our presentation in the conference. In the main part sections of this paper we have taken β as zero. This is because the norm suppression produces zero value for β. The other important point is the existence of just a single position variable (radial variable) at then. This facilitates the analysis pretty much.

 ISBN: 978-84-617-8694-7

References

[1] Elif Tataroğlu, Metin Demiralp, *An implementative application of probabilistic evolution theory: A case study for two particles celestical mechanical system*, AIP Conference Proceedings, 1798(1):020160, 2017.

[2] Coşar Gözükırmızı, Elif Tataroğlu, *Squarification of Telescope Matrices in the Probabilistic Evolution Theoretical Approach to the Two Particle Classical Mechanics as an Illustrative Implementation*, AIP Conference Proceedings, 1798(1):020062, 2017.

[3] Elif Tataroğlu, Metin Demiralp, *Solutions for the Case of Spectrally Separable Kernel Matrices in the Probabilistic Evolution Theory (PREVTH)*, International Journal of Mathematical and Computational Methods, Proceedings of the 10th WSEAS International Conference on Applied Mathematics, Simulation, Modelling(ASM '16), İstanbul, Turkey, April 15–17, (2016) 201–206.

[4] M. Demiralp and B. Tunga, *A probabilistic evolution approach trilogy, part 3: Temporal variation of state variable expectation values from Liouville equation perspective*, J. Math. Chem. Apr. (2013), 1198–1210.

[5] Coşar Gözükırmızı, Metin Demiralp, *Probabilistic evolution approach for the solution of explicit autonomous ordinary differential equations. Part 2: Kernel separability, space extension, and, series solution via telescopic matricess*, J. Math. Chem. **52(3)** Mar. 2014.

[6] Metin Demiralp and N. A. Baykara, *A probabilistic evolution approach trilogy, part 2: spectral issues for block triangular evolution matrix, singularities, space extension*, J. Math. Chem. Apr. (2013) 1187–1197.

[7] Muzaffer Ayvaz, Metin Demiralp, *Space Extension Strategies for Probabilistic Evolution Approach: Classical Symmetric Quartic Anharmonic Oscillator*, The Proceedings of the WSEAS 13th International Conference on System Theory and Scientific Computation (ISTASC'13), 6–8 August 2013, Valencia, Spain (2013) 81–86.

[8] Melike Ebru Kirkin, Coşar Gözükırmızı, *Probabilistic Evolution Theory for ODE sets with second degree multinomial right hand side functions: Certain reductive cases*, in Proceedings of 11th International Conference of Computational Methods in Sciences and Engineering, Atina, Greece, Mar. 2015.

[9] Metin Demiralp, *A probabilistic evolution approach trilogy, part 1: quantum expectation value evolutions, block triangularity and conicality, truncation approximants and their convergence*, J. Math. Chem. **51(4)** Apr. (2013) 1170–1186.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

Verification of positive definiteness using approximate inverse matrices of computed Cholesky factors

Takeshi Terao[1] and Katsuhisa Ozaki[2]

[1] *Graduate School of Engineering and Science, Shibaura Institute of Technology*

[2] *Department of Mathematical Sciences, Shibaura Institute of Technology*

emails: nb17105@shibaura-it.ac.jp, ozaki@sic.shibaura-it.ac.jp

Abstract

The aim of this paper is to verify the positive definiteness for real symmetric matrices. An excellent method for generating computer-assisted proofs of the positive definiteness was proposed by Rump. In this paper, we propose verification methods using computed Cholesky factors and their approximate inverse matrices. Positive definiteness can be verified for a wide range of problems by the proposed method.

Key words: computer-assisted proof, positive definiteness, numerical liner algebra
MSC 2000: 65F15, 65G50

1 Introduction and Preceding Study

The aim of this paper is to verify the positive definiteness for real symmetric matrices. A verification method has already been proposed by Rump [2]. The Rump's method is quick and but can fail to verify positive definiteness when the condition number of the given matrix is high. Herein, we propose verification methods for positive definiteness that work large scale or ill-conditioned matrices.

Here, we introduce the notation used in this paper and briefly review Rump's method [2]. Let $\mathbb{F}$ be the set of floating-point numbers as defined by the IEEE Std. 754 [1]. $fl(\cdot), fl_{\bigtriangledown}(\cdot)$ and $fl_{\bigtriangleup}(\cdot)$ indicate that all operations inside the parentheses are evaluated using floating-point arithmetic with rounding to the nearest, rounding downward and upward mode, respectively. For simplicity, we assume that neither overflow nor underflow occurs in $fl(\cdot)$ because underflow can be handled by adding a constant to the arguments presented in this paper. Let u be the unit roundoff, e.g. $u = 2^{-53}$ for binary64 in the IEEE standard. For

 ISBN: 978-84-617-8694-7

$x, y \in \mathbb{R}^n$, the notation $|x|$ is defined $|x| = (|x_1|, |x_2|, \ldots, |x_n|)^T$ and $x < y$ means $x_i < y_i$ for all i. This notation can be extended to matrices in a straightforward manner.

Let $\bar{\rho} = fl_{\triangle}(\sum_{j=1}^{n} \gamma_{j+1} a_{jj})$, $\gamma_n = nu/(1-nu)$, $nu < 1$ and $B = fl_{\triangledown}(A - \bar{\rho} I)$, where I is the identity matrix. Rump showed that if Cholesky decomposition for the matrix B can be successfully computed using numerical computation, the matrix A is positive definite. We call this method T0.

2 The Proposed Method

In this section, we first introduce lemmas concerning residuals, eigenvalues and norms.

Lemma 1 (Theorem 4.4 in [3]) *Let $A = A^T \in \mathbb{F}^{n\times n}$ and assume that Cholesky decomposition has been successfully completed using numerical computation. Then*

$$\hat{R}^T \hat{R} - A = \Delta A, \quad |\Delta A| \leq (n+1)u|\hat{R}|^T|\hat{R}|, \tag{1}$$

is satisfied, where $\hat{R} \in \mathbb{F}^{n\times n}$ is a factor computed by Cholesky decomposition.

Lemma 2 (Corollary 8.1.6 in [4]) *If A and $A+E$ are n-by-n symmetric matrices, then*

$$|\lambda_k(A+E) - \lambda_k(A)| \leq ||E||_2$$

for $1 \leq k \leq n$, where $\lambda_k(A)$ denotes the kth eigenvalue of the matrix A.

Lemma 3 (Lemma 2.3.3 in [4]) *If $F \in \mathbb{R}^{n\times n}$ and $||F|| < 1$, then $I+F$ is nonsingular and*

$$||(I-F)^{-1}|| \leq \frac{1}{1-||F||}.$$

Next, we propose a theorem that places a lower bound on the minimum eigenvalue.

Theorem 4 *Let $A = A^T$ and introduce $\hat{R}$ and $\hat{X} \in \mathbb{R}^{n\times n}$. Set $\Delta X := \hat{X}\hat{R} - I \in \mathbb{R}^{n\times n}$. If there exist matrices $\hat{R}$ and $\hat{X}$ such that $||\Delta X||_2 < 1$, then*

$$\lambda_{\min}(A) \geq \frac{(1-||\Delta X||_2)^2}{||\hat{X}\hat{X}^T||_2} - ||\Delta A||_2 \quad \textit{where} \quad \Delta A = \hat{R}^T\hat{R} - A. \tag{2}$$

Proof. From Lemma 2, we obtain

$$\lambda_{\min}(A) \geq \lambda_{\min}(\hat{R}^T\hat{R}) - ||\Delta A||_2 = ||(\hat{R}^T\hat{R})^{-1}||_2^{-1} - ||\Delta A||_2. \tag{3}$$

From the assumption $||\Delta X||_2 < 1$ and Lemma 3, $\hat{R}$ and $I + \Delta X$ are nonsingular and we obtain

$$\hat{R}^{-1}\hat{R}^{-T} \;=\; (I+\Delta X)^{-1}\hat{X}\hat{X}^T(I+(\Delta X)^T)^{-1}.$$

 ISBN: 978-84-617-8694-7

From this and Lemma 3, we can derive

$$||(\hat{R}^T\hat{R})^{-1}||_2 \leq \frac{||\hat{X}\hat{X}^T||_2}{(1-||\Delta X||_2)^2}. \tag{4}$$

We obtain (2) by substituting (4) into (3). □

We set $\hat{R}$ and $\hat{X}$ to be the computed Cholesky factors and an approximate inverse matrix of $\hat{R}$, respectively. The computational cost of calculating $\hat{R}$ and $\hat{X}$ is $\frac{1}{3}n^3 + O(n^2)$ flops.

We will prove $\lambda_{\min}(A) > 0$ by computing the upper bounds for $||\hat{X}\hat{X}^T||_2$ and $||\Delta A||_2$ in Theorem 4. First, we propose a fast method (T1). If $S = S^T \in \mathbb{R}^{n\times n}$, then $||S||_2 \leq ||S||_\infty$. The norm $||S||_\infty$ can be computed as follows:

$$||S||_\infty = ||\ |S|e\ ||_\infty, \quad e = (1, 1, \dots 1)^T \in \mathbb{F}^n. \tag{5}$$

We can obtain upper bounds for $||\hat{X}\hat{X}^T||_2$ and $||\Delta A||_2$ as follows:

$$||\hat{X}\hat{X}^T||_2 \leq fl_\triangle(||\ |\hat{X}|(|\hat{X}^T|e)||_\infty), \quad ||\Delta A||_2 \leq fl_\triangle((n+1)u||\ |\hat{R}^T|(|\hat{R}|e)||_\infty). \tag{6}$$

The total cost of proving $\lambda_{\min}(A) > 0$ is thus $\frac{2}{3}n^3 + O(n^2)$ flops.

Next, we introduce methods that use the enclosure property of matrix multiplication. Upper bounds for $||\hat{X}\hat{X}^T||_2$ and $||\Delta A||_2$ can be obtained as follows:

$$||\hat{X}\hat{X}^T||_2 \leq ||\max(|fl_\triangledown(\hat{X}\hat{X}^T)|, |fl_\triangle(\hat{X}\hat{X}^T)|)||_\infty, \tag{7}$$

$$||\Delta A||_2 \leq ||\max(|fl_\triangledown(\hat{R}^T\hat{R} - A)|, |fl_\triangle(\hat{R}^T\hat{R} - A)|)||_\infty. \tag{8}$$

The computational cost of calculating $\hat{X}\hat{X}^T$ and $\hat{R}^T\hat{R}$ is $\frac{1}{4}n^3 + O(n^2)$ flops, because $\hat{X}$ and $\hat{R}$ are triangular matrices, and $\hat{X}\hat{X}^T$ and $\hat{R}^T\hat{R}$ are symmetric. The computational cost of (7) and (8) is $\frac{1}{2}n^3 + O(n^2)$ flops. We now define methods T2, T3 and T4:

- T2: (7) and (6) are used as upper bounds for $||\hat{X}^T\hat{X}||_\infty$ and $||\Delta A||_\infty$.
- T3: (6) and (8) are used as upper bounds for $||\hat{X}^T\hat{X}||_\infty$ and $||\Delta A||_\infty$.
- T4: (7) and (8) are used as upper bounds for $||\hat{X}^T\hat{X}||_\infty$ and $||\Delta A||_\infty$.

The norm $||\Delta X||_2$ can be bounded as follows:

$$|\Delta X| \leq nu|\hat{X}||\hat{R}|, \quad ||\Delta X||_2 \leq \sqrt{||\Delta X||_1||\Delta X||_\infty}.$$

This computational cost of this is $O(n^2)$ flops using (5).

Table 1 shows the computational cost for these norms using each method. The cost of Cholesky decomposition is normalised to one in the Ratio column, in other words, the

 ISBN: 978-84-617-8694-7

Table 1: The computational cost of calculating $||\hat{X}\hat{X}^T||_\infty$ and $||\Delta A||_\infty$ (flops)

Method	$||\hat{X}\hat{X}^T||_\infty$	$||\Delta A||_\infty$	Ratio
T1	$O(n^2)$	$O(n^2)$	2
T2	$n^3/2 + O(n^2)$	$O(n^2)$	3.5
T3	$O(n^2)$	$n^3/2 + O(n^2)$	3.5
T4	$n^3/2 + O(n^2)$	$n^3/2 + O(n^2)$	5

cost of T0 is defined to be one. Next, we check that positive definiteness is guaranteed for more ill-conditioned matrices using our methods. Let $A = A^T \in \mathbb{F}^{n\times n}$. The matrix A was generated in MATLAB by $A =$ `gallery('randsvd'`$, n, -c, 3, n, n, 1)$, where c is the expected condition number of A. Table 2 shows the maximum condition number c for each method. The methods could verify the positive definiteness up to these condition numbers. These numerical results indicate that our methods can cover more ill-conditioned matrices.

Table 2: The maximum condition numbers for which each method can verify the positive definiteness of matrices of size n

n	T0	T1	T2	T3	T4
5000	5.01e+07	7.94e+07	1.99e+09	3.98e+10	7.94e+11
10000	5.01e+06	7.94e+06	3.16e+08	1.00e+10	3.16e+11
20000	3.98e+05	7.94e+05	5.01e+07	2.51e+09	1.00e+11
30000	3.98e+04	2.51e+05	1.99e+07	7.94e+08	6.30e+10

The application of the proposed method to validate the solutions of linear systems will be shown in the presentation.

References

[1] *IEEE Standard for Floating-Point Arithmetic*, 754–2008, (2008).

[2] S.M. Rump, *Verification of positive definiteness*, BIT Numer. Math., **46** (2006), 433 – 452.

[3] S.M. Rump and C.-P. Jeannerod, *Improved backward error bounds for LU and Cholesky factorizations*, SIAM. J. Matrix Anal. & Appl., **35** (2014), 684 – 698.

[4] G. H. Golub and C. F. Van Loan, *Matrix computations, 4th edition*, The Johns Hopkins University Press, (2013).

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

Inversion of infinite reduced Hessenberg matrices and operators

Venancio Tomeo[1]

[1] *Department of Algebra, Faculty of Statistics, Complutense University of Madrid, Spain*

emails: tomeo@ucm.es

Abstract

Inversion of infinite invertible unreduced Hessenberg matrices has been studied in [1], with applications to infinite invertible unreduced tridiagonal matrices. The case of infinite invertible reduced Hessenberg matrices and infinite reduced tridiagonal matrices, when the number of zeros on the subdiagonal is finite or infinite, is now studied for a complete landscape of inversion of such matrices. Reduced Hessenberg matrices are regarded finally as bounded linear operators

Key words: Hessenberg matrix, tridiagonal matrix, inverse matrix, reduced matrix.

1 Introduction

A characterization of the finite nonsingular unreduced Hessenberg matrices is related with the distinguished rank structure of their inverse matrix [2, 3, 4]. The inverses of such matrices are rank one perturbation, $UV + T$, of triangular matrices. The inverses of real or complex infinite Hessenberg matrices have a similar structure and the method for inverting infinite Hessenberg matrices is based on the structure properties. In particular, classical inverses of general tridiagonal matrices can be generated through recurrence relations.

Some background about inversion of infinite matrices and their applications can be found in the literature; see e.g. [5, 6] and the references given there. Hessenberg matrices appears in signal processing, time series, birth-dead processes and orthogonal polynomial as matrix representation of the multiplication by z operator. They arise in numerical linear algebra, as a result of the application of Givens or Householder orthogonal transformations for a general matrix, when solving the eigenvalue problem.

 ISBN: 978-84-617-8694-7

In this paper we present the inversion of infinite invertible reduced Hessenberg matrices and the inversion of infinite invertible reduced tridiagonal, as particular case, when the number of zeros in subdiagonal is finite or infinite. Finally, reduced Hessenberg matrices are regarded as bounded linear operator in a Hilbert space.

The material is organized as follow. In Section 1, we recall some basic result about inverses of finite Hessenberg matrices and infinite unreduced Hessenberg matrices and tridiagonal matrices. In Section 2 we study the inversion of infinite reduced Hessenberg matrices when the number of zeros in subdiagonal is finite and infinite. Section 3 contains a short remark regarding the inversion of bounded linear operators. Some illustrative examples are also given.

1.1 Unreduced Hessenberg matrices of finite order

We recall some lemmas related with finite upper Hessenberg matrices and tridiagonal matrices. A matrix $H = (h_{ij})_{i,j=1}^{n}$ is upper Hessenberg if its elements h_{ij} satisfy $h_{ij} = 0$ for $i \geq j + 2$. We also recall that an order n Hessenberg matrix H is an unreduced upper Hessenberg matrix if it has nonzero subdiagonal entries, $h_{i+1,i} \neq 0$, $i = 1, 2, ..., n - 1$. Next lemma is well-known, [1, 2, 3].

Lemma 1. *A nonsingular matrix $H = (h_{ij})_{i,j=1}^{n}$ is unreduced upper Hessenberg if and only if its inverse matrix has the structure $B = UV + T$, being U a column matrix with nonzero n-th component, V is a row matrix with nonzero 1-st component. T is strictly upper triangular having null entries on its main diagonal and nonzero entries on the superdiagonal, $t_{i,i+1} = \frac{1}{h_{i+1,i}} \neq 0, 1 \leq i \leq n - 1$.*

Some numerical methods for inverting finite Hessenberg matrices and tridiagonal matrices are available, see e.g. [2, 7, 8, 9] and references given there.

1.2 Reduced Hessenberg matrices of finite order

If we have a nonsingular upper Hessenberg matrix having a zero entry on the subdiagonal, we can decompose the matrix in blocks and using the Schur complement. For a nonsingular reduced Hessenberg matrix H with $h_{k+1,k} = 0$, we consider

$$H = \left(\begin{array}{c|c} H_{11} & H_{12} \\ \hline 0 & H_{22} \end{array}\right) \qquad \text{and} \qquad B = \left(\begin{array}{c|c} B_{11} & B_{12} \\ \hline 0 & B_{22} \end{array}\right),$$

where $B_{11} = H_{11}^{-1}$, $B_{22} = H_{22}^{-1}$ and $B_{12} = -B_{11}H_{12}B_{22}$. The matrix B is the inverse of H and both matrices have the same block upper triangular structure.

If matrix H is a nonsingular upper Hessenberg matrix having a finite number of zero entries on the subdiagonal, we can to proceed in the same way.

 ISBN: 978-84-617-8694-7

1.3 Inverses of infinite unreduced Hessenberg matrices

Given a real or complex infinite-dimensional matrix $A = (a_{ij})$, the matrix $B = (b_{ij})$ is a classical inverse of A if both matrices A and B satisfy $AB = BA = I$, with I the identity matrix. The matrix A may have (or may not have) classical inverse. Alternatively, A may have two classical inverses, and then infinitely many classical inverses; see e.g. [1, 6, 10].

An infinite matrix $H = (h_{ij})$ is an upper Hessenberg matrix if $h_{ij} = 0$ for $i \geq j + 2$. An infinite Hessenberg matrix H is an unreduced upper Hessenberg matrix if its subdiagonal entries are nonzero, $h_{i+1,i} \neq 0$, and $i = 1, 2, ...$

Conditions for an infinite invertible matrix H be an unreduced Hessenberg matrix are given in Corollary 1 of [1], that we recall below as Lemma 2, and applications to unreduced tridiagonal matrices. The submatrix H_n is the left principal one of order n. The submatrix $H^{(i)}_{n-i}$ is the right principal one of order $n - i$, which begins in the $i + 1$-th row and column and finishes in the n-th row and column.

Lemma 2. *Let $H = (h_{ij})$ be an infinite invertible matrix and assume that the limits*

$$\lim_n \frac{|H^{(i)}_{n-i}|}{|H_n|} = \xi_i,$$

for $i \geq 1$ are finite and nonzero. Then H is an unreduced upper Hessenberg matrix if and only if its classical inverse matrix $B = (b_{ij})$ has the form $B = UV + T$, where $U = (u_1, u_2, ...)^T$ is a column vector with nonzero components, $V = (v_1, v_2, ...)$ is a row vector with nonzero first component, and T is strictly upper triangular, having zeros on its main diagonal, and nonzero entries $t_{i,i+1} = h^{-1}_{i+1,i}$, $i \geq 1$, on its superdiagonal.

2 Inverses of infinite reduced Hessenberg matrices

2.1 Infinite reduced Hessenberg matrices with a finite number of zeros in the subdiagonal

When an infinite reduced Hessenberg matrix H has at least a null entry on its subdiagonal, we can evaluate its classical inverse in a similar way as the known method of the Schur complement for a matrix of finite order.

Proposition 1. *Let H be an infinite invertible upper Hessenberg matrix with only a zero entry on its subdiagonal. Then, its classical inverse matrix can be calculate using a block matrix procedure. If H is*

$$H = \left(\begin{array}{c|c} H_{11} & H_{12} \\ \hline 0 & H_{22} \end{array} \right),$$

 ISBN: 978-84-617-8694-7

a classical inverse matrix is

$$B = \left(\begin{array}{c|c} H_{11}^{-1} & -H_{11}^{-1} H_{12} H_{22}^{-1} \\ \hline 0 & H_{22}^{-1} \end{array} \right),$$

where H_{11} is a finite nonsingular unreduced upper Hessenberg matrix, H_{22} is an infinite invertible unreduced upper Hessenberg matrix. There are infinitely many classical inverses of H.

Proof. As H_{11} is a finite matrix and H_{12} has a finite number of rows, products satisfy the associative property and products HB and BH give obviously I. □

The case of finitely many zeros on the subdiagonal follows from the previous Proposition.

Proposition 2. *Let H be an infinite invertible upper Hessenberg matrix with a finite number of zero entries on the subdiagonal. Then its classical inverse matrices can be obtained using a block matrix procedure in a similar way as Proposition 1.*

Proof. We consider the block matrix

$$H = \left(\begin{array}{c|c} H_{11} & H_{12} \\ \hline 0 & H_{22} \end{array} \right),$$

where all zero entries in the subdiagonal are in H_{11}. The inverse of the finite matrix H_{11} is given in Subsection 1.2. Then H_{22} is an unreduced Hessenberg matrix and the inverse of matrix H is given by the formula of Proposition 1. □

Obviously, H has infinitely many classical inverses because H_{22} is an unreduced infinite Hessenberg matrix and, by Proposition 1, has infinitely many classical inverses.

Example 1. *The next matrix H has zero entries in $h_{23} = h_{54} = h_{76}$ and nonzero in all other entries on its subdiagonal. Its inverse matrix is B :*

$$H = \left(\begin{array}{cc|cc|cc|c} 1 & 0 & 1 & 0 & 1 & 0 & \\ 1 & 1 & 0 & 1 & 0 & 1 & H_{14} \\ \hline & & 1 & 0 & 1 & 0 & \\ & & 1 & 1 & 0 & 1 & H_{24} \\ \hline & & & & 1 & 0 & \\ & & & & 1 & 1 & H_{34} \\ \hline & & & & & & H_{44} \end{array} \right), \; B = \left(\begin{array}{cc|cc|cc|c} 1 & 0 & & & & & \\ -1 & 1 & B_{12} & & B_{13} & & B_{14} \\ \hline & & 1 & 0 & & & \\ & & -1 & 1 & B_{23} & & B_{24} \\ \hline & & & & 1 & 0 & \\ & & & & -1 & 1 & B_{34} \\ \hline & & & & & & B_{44} \end{array} \right),$$

 ISBN: 978-84-617-8694-7

where $B_{44}=H_{44}^{-1}$ *and*

$$\begin{aligned}
B_{12} &= -H_{11}^{-1}H_{12}H_{22}^{-1}=-\begin{pmatrix}1&0\\-1&1\end{pmatrix}\begin{pmatrix}1&0\\0&1\end{pmatrix}\begin{pmatrix}1&0\\-1&1\end{pmatrix}=\begin{pmatrix}-1&0\\2&-1\end{pmatrix},\\
B_{23} &= -H_{22}^{-1}H_{23}H_{33}^{-1}=-\begin{pmatrix}1&0\\-1&1\end{pmatrix}\begin{pmatrix}1&0\\0&1\end{pmatrix}\begin{pmatrix}1&0\\-1&1\end{pmatrix}=\begin{pmatrix}-1&0\\2&1\end{pmatrix},\\
B_{13} &= -H_{11}^{-1}(H_{13}-H_{12}H_{22}^{-1}H_{23})H_{33}^{-1},\\
B_{34} &= -H_{33}^{-1}H_{34}H_{44}^{-1},\\
B_{24} &= -H_{22}^{-1}(H_{24}-H_{23}H_{33}^{-1}H_{34})H_{44}^{-1},\\
B_{14} &= -H_{11}^{-1}(H_{14}-H_{13}H_{33}^{-1}H_{34}-H_{12}H_{22}^{-1}[H_{24}-H_{23}H_{33}^{-1}H_{34}])H_{44}^{-1}.
\end{aligned}$$

Different blocks can be obtained in each block-column from diagonal to top. A general formula for B_{ij} is shown in next theorem. We observe that H_{44} is unreduced and then it has not a unique inverse matrix.

The case of infinitely many zeros on the subdiagonal of H is a little different and is studied in next subsection.

2.2 Inverses of infinite Hessenberg matrices with an infinite number of zeros in the subdiagonal

Theorem 1. *A reduced invertible upper Hessenberg matrix* $H=(h_{ij})$ *with infinitely many zeros on its subdiagonal can be decomposed in infinite nonsingular blocks and the inverse matrix* $B=(b_{ij})$ *has the same block structure in a similar way to the finite case. The classical inverse matrix is unique. Different blocks of* B *can be obtained by the next recurrence relations*

$$\begin{aligned}
B_{ii} &= H_{ii}^{-1}\\
B_{ij} &= -H_{ii}^{-1}(H_{i,i+1}B_{i+1,j}+H_{i,i+2}B_{i+2,j}+\cdots+H_{ij}B_{jj}),\quad i<j.
\end{aligned}$$

Proof. Matrices H and B can be decomposed in an infinite number of finite blocks in the form

$$H=\begin{pmatrix}H_{11}&H_{12}&H_{13}&H_{14}&\cdots\\&H_{22}&H_{23}&H_{24}&\cdots\\&&H_{33}&H_{34}&\cdots\\&&&H_{44}&\cdots\\&&&&\ddots\end{pmatrix},\qquad B=\begin{pmatrix}B_{11}&B_{12}&B_{13}&B_{14}&\cdots\\&B_{22}&B_{23}&B_{24}&\cdots\\&&B_{33}&B_{34}&\cdots\\&&&B_{44}&\cdots\\&&&&\ddots\end{pmatrix}.$$

If we multiply the block-row i-th of H by the block-column i-th of B, or vice versa, we have $H_{ii}B_{ii}=B_{ii}H_{ii}=I$. Then, $B_{ii}=H_{ii}^{-1}$. If we multiply the i-th block-row of H by the j-th

 ISBN: 978-84-617-8694-7

block-column of B, with $i < j$, we obtain

$$H_{ii}B_{ij} + H_{i,i+1}B_{i+1,j} + \cdots + H_{i,j}B_{jj} = 0,$$

then, we have

$$B_{ij} = -H_{ii}^{-1}(H_{i,i+1}B_{i+1,j} + H_{i,i+2}B_{i+2,j} + \cdots + H_{ij}B_{jj}).$$

If we multiply the i-th block-row of H by the j-th block-column of B, with $i > j$, we obtain obviously 0. In a similar way if we multiply block-rows of B by block-columns of H. Each block-column, B_{ij}, is evaluated from the blocks placed under it. □

In the case that matrix H has not two nonzero consecutive entries in subdiagonal, that is, $d_i \cdot d_{i+1} = 0, \forall i$, the inverse matrix is also an upper Hessenberg matrix. We have a necessary and sufficient condition for the inverse of an invertible upper Hessenberg let be also an upper Hessenberg matrix. In a similar way we have the condition for an infinite invertible tridiagonal matrix let be also a tridiagonal matrix. There are the next theorem and the next corollary.

Theorem 2. *A necessary and sufficient condition for that the infinite invertible upper Hessenberg matrix $H = (h_{ij})$ has a classical inverse matrix that be also an upper Hessenberg matrix, is that verify $h_{i+1,i} \cdot h_{i+2,i+1} = 0$ for all $i = 1, 2, ...$ In this case, the classic inverse matrix is unique.*

Proof. Necessary condition: If an infinite matrix H has nonzero entries on its subdiagonal, it is an unreduced matrix, and then, if it is invertible, its classical inverse matrix $B = (b_{ij})$ has the form $B = UV + T$, as we have shown in Lemma 2. Hence, its classical inverse matrix is a full matrix.

If an infinite matrix H has a zero entry in $[n+1, n]$, we consider matrix H divided in blocks as

$$H = \left(\begin{array}{c|c} H_{11} & H_{12} \\ \hline 0 & H_{22} \end{array}\right),$$

where H_{11} is an $n \times n$ finite invertible matrix and H_{22} is an infinite invertible matrix. Entries of H_{11}^{-1} are given by expression (7) of [11], that is

$$(H^{-1})_{ij} = \frac{(-1)^{i+j}\left(\prod_{k=2}^{i} h_{k,k-1}\right)\left(\det H_{j-1} \det H_{n-i}^{(i)} - \det H_{j-i-1}^{(i)} \det H_{11}\right)}{\left(\prod_{k=2}^{j} h_{k,k-1}\right) \det H_{11}},$$

and in the case $i \geq j$, by expression (8) of [11], we have

$$(H^{-1})_{ij} = (-1)^{i+j} \frac{\left(\prod_{k=0}^{i-j-1} h_{i-k,i-k-1}\right) \det H_{j-1} \det H_{n-i}^{(i)}}{\det H_{11}},$$

 ISBN: 978-84-617-8694-7

where determinants of H_{j-1}, $H_{n-i}^{(i)}$ and H_{11} are nonzero because matrix H is invertible. Then, as H_{11}^{-1} is an upper Hessenberg matrix, for $i > j+1$, we have $(H_{11}^{-1})_{ij} = 0$, that is $\prod_{k=0}^{i-j-1} h_{i-k,i-k-1} = 0$. In particular, for $i = j+2$ we have $h_{i,i-1}h_{i-1,i-2} = 0$, $\forall i = 1, 2, ..., n$.

We can repeat the argument if matrix H_{22} plays the rol of matrix H.

Sufficient condition: We need to prove that entries $[i, j]$ in the inverse matrix H_{11}^{-1}, with $i > j+1$, are zero. If infinite matrix H has a zero entry in $[n+1, n]$, we consider the previous division in blocks for H, and the entries of H_{11}^{-1} are given by expression (8) of [11], and then, finite matrix H_{11} has the entry $[i, j]$, with $i > j+1$, given by previous expression and $(H_{11}^{-1})_{ij} = 0$ because one entry $h_{i-k,i-k-1}$ is zero. This is valid for all entries below the subdiagonal of H^{-1}.

Unicity follows because all the blocks are of finite order. □

Theorem 2 can be applied, in a similar way, to infinite invertible lower Hessenberg matrices.

Corollary 1. *A necessary and sufficient condition for the infinite invertible tridiagonal matrix $T = \{a_i, b_i, c_i\}$ has a classical inverse matrix that be also a tridiagonal matrix, is that verify $a_i \cdot a_{i+1} = 0$ and $c_i \cdot c_{i+1} = 0$ for all $i = 1, 2, ...$ This classic inverse matrix is unique.*

Proof. Tridiagonal matrix T is upper and lower Hessenberg matrix, then with condition in the establishment the inverse matrix T^{-1}, by Theorem 2, is both upper and lower Hessenberg matrix and thus T^{-1} is also tridiagonal. □

Example 2. *Matrix H_1 is an infinite reduced upper Hessenberg matrix and it is invertible. Matrix H_2 is an infinite reduced invertible tridiagonal matrix*

$$H_1 = \left(\begin{array}{ccc|ccc|c} 0 & 0 & 1 & & & & \\ 1 & 0 & 0 & & & & \\ 0 & 1 & 0 & & & & \\ \hline & & & 0 & 0 & 1 & \\ & & & 1 & 0 & 0 & \\ & & & 0 & 1 & 0 & \\ \hline & & & & & & \ddots \end{array}\right), \qquad H_2 = \left(\begin{array}{cc|cc|cc|c} 0 & 1 & & & & & \\ 1 & 0 & & & & & \\ \hline & & 0 & 1 & & & \\ & & 1 & 0 & & & \\ \hline & & & & 0 & 1 & \\ & & & & 1 & 0 & \\ \hline & & & & & & \ddots \end{array}\right),$$

an inverse of H_2 is itself.

 ISBN: 978-84-617-8694-7

Example 3. *Tridiagonal matrix H_3 does not verify the conditions of Corollary 1, it is also a Hessenberg matrix with conditions of Theorem 2:*

$$H_3 = \left(\begin{array}{cc|ccc} 0 & 1 & & & \\ 1 & 0 & 1 & & \\ \hline & & 0 & 1 & \\ & & 1 & 0 & 1 \\ \hline & & & & 0 \quad 1 \\ & & & & 1 \quad 0 \\ \hline & & & & & \ddots \end{array}\right).$$

This matrix is invertible because as lower Hessenberg has nonzero entries in superdiagonal. That is, an unreduced Hessenberg matrix. Its unique classical inverse matrix is

$$B_3 = \begin{pmatrix} 0 & 1 & & & & & & \\ 1 & 0 & 1 & & & & & \\ & & 0 & 1 & & & & \\ & & 1 & 0 & 1 & & & \\ & & & & 0 & 1 & & \\ & & & & 1 & 0 & 1 & \\ & & & & & & 0 & \ddots \\ & & & & & & \ddots & \ddots \end{pmatrix}^{-1} = \begin{pmatrix} 0 & 1 & 0 & -1 & 0 & 1 & 0 & \cdots \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & \cdots \\ & 0 & 0 & 1 & 0 & -1 & 0 & \cdots \\ & & 1 & 0 & 0 & 0 & 0 & \cdots \\ & & & 0 & 0 & 1 & 0 & \cdots \\ & & & & 1 & 0 & 0 & \cdots \\ & & & & & 0 & 0 & \ddots \\ & & & & & & \ddots & \ddots \end{pmatrix}.$$

Example 4. *Matrix H_4 is a tridiagonal matrix and it verifies the conditions of Corollary 1. Then, its unique classical inverse matrix is B_4 :*

$$H_4^{-1} = \left(\begin{array}{cc|cc|cc|c} 1 & 0 & & & & & \\ \frac{1}{3} & 1 & & & & & \\ \hline & 0 & 1 & 0 & & & \\ & & \frac{1}{3} & 1 & & & \\ \hline & & & 0 & 1 & 0 & \\ & & & & \frac{1}{3} & 1 & \\ \hline & & & & & & \ddots \end{array}\right)^{-1} = B_4 = \left(\begin{array}{cc|cc|cc|c} 1 & 0 & & & & & \\ \frac{-1}{3} & 1 & & & & & \\ \hline & & 1 & 0 & & & \\ & & \frac{-1}{3} & 1 & & & \\ \hline & & & & 1 & 0 & \\ & & & & \frac{-1}{3} & 1 & \\ \hline & & & & & & \ddots \end{array}\right).$$

 ISBN: 978-84-617-8694-7

Example 5. *Matrix H_5 is an upper Hessenberg matrix and it verifies the conditions of Theorem 2. The unique classical inverse matrix is B_5 :*

$$H_5^{-1}=\left(\begin{array}{cc|cc|cc|cc|c} 1 & & 1 & 0 & 1 & 0 & 1 & 0 & \cdots \\ 1 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & \cdots \\ \hline & 0 & 1 & & 1 & 0 & 1 & 0 & \cdots \\ & & 1 & 1 & 0 & 1 & 0 & 1 & \cdots \\ \hline & & & 0 & 1 & & 1 & 0 & \cdots \\ & & & & 1 & 1 & 0 & 1 & \cdots \\ \hline & & & & & 0 & 1 & & \cdots \\ & & & & & & 1 & 1 & \cdots \\ \hline & & & & & & & & \ddots \end{array}\right)^{-1} = B_5 = \left(\begin{array}{cc|cc|cc|cc|c} 1 & & 1 & 0 & 1 & 0 & 1 & 0 & \cdots \\ -1 & 1 & -2 & 1 & -4 & 1 & -8 & 1 & \cdots \\ \hline & & 1 & & 1 & 0 & 1 & 0 & \cdots \\ & & -1 & 1 & -2 & 1 & -4 & 1 & \cdots \\ \hline & & & & 1 & & 1 & 0 & \cdots \\ & & & & -1 & 1 & -2 & 1 & \cdots \\ \hline & & & & & & 1 & & \cdots \\ & & & & & & -1 & 1 & \cdots \\ \hline & & & & & & & & \ddots \end{array}\right).$$

Example 6. *Matrix H_6 is a Hessenberg matrix with conditions of Theorem 2. The unique classical inverse matrix is*

$$H_6^{-1}=\left(\begin{array}{cc|cc|cc|cc|c} 1 & & \frac{1}{4} & 0 & \frac{1}{16} & 0 & \frac{1}{64} & 0 & \cdots \\ a & 1 & 0 & \frac{1}{4} & 0 & \frac{1}{16} & 0 & \frac{1}{64} & \cdots \\ \hline & 0 & 1 & & \frac{1}{4} & 0 & \frac{1}{16} & 0 & \cdots \\ & & a & 1 & 0 & \frac{1}{4} & 0 & \frac{1}{16} & \cdots \\ \hline & & & 0 & 1 & & \frac{1}{4} & 0 & \cdots \\ & & & & a & 1 & 0 & \frac{1}{4} & \cdots \\ \hline & & & & & 0 & 1 & & \cdots \\ & & & & & & a & 1 & \cdots \\ \hline & & & & & & & & \ddots \end{array}\right)^{-1} = \left(\begin{array}{cc|cc|cc|cc|c} 1 & & \frac{-1}{4} & 0 & & & & & \\ -a & 1 & \frac{a}{2} & \frac{-1}{4} & \frac{-a}{16} & & & & \\ \hline & 0 & 1 & & \frac{-1}{4} & 0 & & & \\ & & -a & 1 & \frac{a}{2} & \frac{-1}{4} & \frac{-a}{16} & & \\ \hline & & & 0 & 1 & & \frac{-1}{4} & 0 & \\ & & & & -a & 1 & \frac{a}{2} & \frac{-1}{4} & \ddots \\ \hline & & & & & 0 & 1 & & \ddots \\ & & & & & & -a & 1 & \ddots \\ \hline & & & & & & & \ddots & \ddots \end{array}\right).$$

3 Reduced Hessenberg matrices related with bounded linear operators

In some cases, Hessenberg matrices can be regarded as bounded linear operators on ℓ^2. We recall here some basic features of the matrix representation of invertible linear operators. A bounded operator T between Hilbert spaces, for example from ℓ^2 to itself, is *invertible* if there exists an operator T^{-1}, obviously bounded, such that $T^{-1}Tx = x$ and $TT^{-1}y = y$ for every $x, y \in \ell^2$. The operator T^{-1} is called the inverse operator of T, and it is unique. A useful method for inverting bounded linear operators is the (iterative) power series method given by next lemma; see e.g. Theorem 8.2 from [5].

 ISBN: 978-84-617-8694-7

Lemma 3. *Let H be a matrix representation of a bounded operator on ℓ^2 that satisfies $||A|| < 1$, where $A = I - H$. Then H is invertible, and its inverse H^{-1},*

$$H^{-1} = \sum_{j=0}^{\infty} A^j = I + A + A^2 + A^3 + \cdots ,$$

is a matrix representation of its inverse operator defined on ℓ^2.

Example 7. *The infinite Hessenberg matrix H_6 from Example 6 satisfies $H_6 = I - A$, where A is the matrix*

$$A = \left(\begin{array}{cc|cc|cc|cc|c} 0 & & \frac{-1}{4} & 0 & \frac{-1}{16} & 0 & \frac{-1}{64} & 0 & \cdots \\ -a & 0 & 0 & \frac{-1}{4} & 0 & \frac{-1}{16} & 0 & \frac{-1}{64} & \cdots \\ \hline & 0 & 0 & & \frac{-1}{4} & 0 & \frac{-1}{16} & 0 & \cdots \\ & & -a & 0 & 0 & \frac{-1}{4} & 0 & \frac{-1}{16} & \cdots \\ \hline & & & 0 & 0 & & \frac{-1}{4} & 0 & \cdots \\ & & & & -a & 0 & 0 & \frac{-1}{4} & \cdots \\ \hline & & & & & 0 & 0 & & \cdots \\ & & & & & & -a & 0 & \cdots \\ \hline & & & & & & & \vdots & \ddots \end{array}\right).$$

If $|a| < \frac{2}{3}$, then is $||A|| < 1$. Hence H_6, as a matrix representation of a bounded linear operator in ℓ^2, is invertible. The matrix B_6 from Example 6 is a matrix representation of its inverse operator.

The infinite Hessenberg matrix H_5 from Example 5 satisfies also $H_5 = I - A$, where A is the matrix

$$A = \left(\begin{array}{cc|cc|cc|cc|c} 0 & 0 & -1 & 0 & -1 & 0 & -1 & 0 & \cdots \\ -1 & 0 & 0 & -1 & 0 & -1 & 0 & -1 & \cdots \\ \hline & 0 & 0 & 0 & -1 & 0 & -1 & 0 & \cdots \\ & & -1 & 0 & 0 & -1 & 0 & -1 & \cdots \\ \hline & & & 0 & 0 & 0 & -1 & 0 & \cdots \\ & & & & -1 & 0 & 0 & -1 & \cdots \\ \hline & & & & & 0 & 0 & 0 & \cdots \\ & & & & & & -1 & 0 & \cdots \\ \hline & & & & & & & & \ddots \end{array}\right).$$

In this case, H_5 is not a bounded linear operator in ℓ^2. The matrix B_5 from Example 5 is the unique classical inverse matrix of H_5, but is not the representation of a bounded operator.

 ISBN: 978-84-617-8694-7

The infinite tridiagonal matrix H_4 from Example 4 satisfies also $H_4 = I - A$, where A is the matrix

$$A = \left(\begin{array}{cc|cc|cc|c} 0 & 0 & & & & & \\ \frac{-1}{3} & 0 & & & & & \\ \hline & & 0 & 0 & & & \\ & & \frac{-1}{3} & 0 & & & \\ \hline & & & & 0 & 0 & \\ & & & & \frac{-1}{3} & 0 & \\ \hline & & & & & & \ddots \end{array}\right),$$

with norm $||A|| < 1$. Hence H_4, as a matrix representation of a bounded linear operator in ℓ^2, is invertible. The matrix B_4 from Example 4 is a matrix representation of its inverse operator.

Recall that a matrix H can have infinitely many classical inverses. Nevertheless, if H is a matrix representation of an invertible bounded operator, it has a unique inverse. Such an inverse is also a matrix representation of its inverse bounded linear operator, as those given in Example 7.

Some reduced tridiagonal matrices with specific invertible blocks can be inverted by blocks, as we shown in next example.

Example 8. *If we take $0 < a, b < 1$, matrix T_1 verifies $||I - T_1|| < 1$ and T_1 is an invertible bounded operator with*

$$T_1^{-1} = \left(\begin{array}{cc|cc|c} 1 & b & & & \\ a & 1 & & & \\ \hline & & 1 & b & \\ & & a & 1 & \ddots \\ & & & \ddots & \ddots \end{array}\right)^{-1} = \left(\begin{array}{cc|cc|c} \frac{1}{1-ab} & \frac{-b}{1-ab} & & & \\ \frac{-a}{1-ab} & \frac{1}{1-ab} & & & \\ \hline & & \frac{1}{1-ab} & \frac{-b}{1-ab} & \\ & & \frac{-a}{1-ab} & \frac{1}{1-ab} & \ddots \\ & & & \ddots & \ddots \end{array}\right).$$

The same happens with matrix T_2 if $0 < a, b < \frac{1}{2}\cdot$ The inverse operator is

$$T_2^{-1} = \left(\begin{array}{cc|cc|cc} 1 & 0 & & & & \\ a & 1 & b & & & \\ \hline & & 1 & 0 & & \\ & & a & 1 & b & \\ \hline & & & & 1 & \ddots \\ & & & & \ddots & \ddots \end{array}\right)^{-1} = \left(\begin{array}{cc|cc|cc} 1 & 0 & & & & \\ -a & 1 & -b & & & \\ \hline & & 1 & 0 & & \\ & & -a & 1 & -b & \\ \hline & & & & 1 & \ddots \\ & & & & \ddots & \ddots \end{array}\right).$$

 ISBN: 978-84-617-8694-7

Acknowledgements

This work has been supported by the Spanish Science Ministry, under the project MTM2016-80582-R (AEI/FEDER,U.E.).

References

[1] J. Abderramán, V. Tomeo, E. Torrano, *On inverses of infinite Hessenberg matrices,* J. Comp. Appl. Math. **275** (2015) 356-365.
[2] Y. Ikebe, *On inverses of Hessenberg matrices,* Linear Algebra Appl., **24** (1979), 93-97.
[3] D. K. Faddeev, *Properties of a matrix, inverse of a Hessenberg matrix,* Journal Mathematical Sciences, **24** (1984), 118-120.
[4] R. Vandebril, M. Van Barel, N. Mastronardi, Matrix Computation and Semiseparable Matrices, Volume I: Linear Systems, Johns Hopkins University Press, Baltimore, Maryland, USA, 2008.
[5] I. Gohberg, S. Goldberg, M. A. Kaashoek, Basic Classes of Linear Operators, Birkäuser Verlag, Basel, 2003.
[6] P. N. Sivakumar. K. C. Sivakumar, *A review of infinite matrices and their applications,* Linear Algebra Appl. **430** (2009) 976-998.
[7] B. Bukhberger, G. A. Emel'yanenko, *Methods of inverting tridiagonal matrices,* Comput. Math. Math. Phys. URSS **13** (1973), 10-20.
[8] J. Abderramán, M. Rachidi, V. Tomeo, *Non-symbolic algorithms for the inversion of tridiagonal matrices,* J. Comp. Appl. Math. **252** (2013) 3-11.
[9] J. Abderramán, M. Rachidi, V. Tomeo, *On new algorithms for inverting Hessenberg matrices,* J. Comp. Appl. Math. **252** (2013) 12-20.
[10] R. G. Cooke, Infinite matrices & sequence spaces, Dover Publications, New York, U.S.A. 1955.
[11] J. Abderramán, V. Tomeo, *On the closed representation for the inverses of Hessenberg matrices,* J. Comp. Appl. Math. **236** (2012) 2962-2970.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference on Computational and Mathematical Methods in Science and Engineering, CMMSE 2017 4–8 July, 2017.

Continuous and discrete time models for the bovine Babesiosis disease.

Deccy Y. Trejos[1] and Jose C. Valverde[2]

[1] *Faculty of Science and Education, District University Francisco José of Caldas, Bogotá-Colombia*

[2] *Department of Mathematics, University of Castilla-La Mancha*

emails: dytrejosa@udistrital.edu.co, jose.valverde@uclm.es

Abstract

In this work, we compare continuous and discrete models for the propagation of bovine Babesiosis disease among bovine and ticks populations. From our continuous-time model based on ordinary differential equations, we derive its discretization both ways: by direct formulation of the model and by using the forward Euler scheme. We show that the reproductive parameter $\mathcal{R}_0$ plays an important role in the existence and stability of the equilibrium points for the different versions. Finally, we conclude from some numerical simulations that the dynamics do not change when varying the value of the step size till a threshold value corresponding to the Euler discretization.

Key words: Bovine Babesiosis, continuous-time model, discrete-time model, forward Euler Scheme, Local and global stability.

1 Extended abstract

The bovine Babesiosis is a parasitic zoonotic disease transmitted by ticks and is produced by protozoan as *Babesia bovis* and *Babesia bigemina*. The bovine Babesiosis cause great economic losses to cattle farmers in tropical and subtropical climates around the world.

The first mathematical continuous deterministic model known for the spread of bovine Babesiosis was proposed by Aranda et al. [1] and consists of five ordinary differential equations. In 2014, Friedman and Yakubu [5] included the factor of dispersion for the bovine population and ticks converting the model in [1] into a system of partial differential equations and introducing a parameter $\mathcal{P}$, a *proliferation index*, which plays the same role as the *basic reproduction number*, $\mathcal{R}_0$, in our work [1]. Carvalho et al. [4] incorporated the memory effect in the model. They changed normal derivative by the Caputo derivative. Also in this sense, Zafar et al. [10] study a fractional-order scheme version of our model. Other works take into account factors such as the juvenile stage of the cattle population [9] or the effects of seasonal changes [3]. On the other hand, in [8], the authors use a computational multistage algorithm modified for approximating solutions of the model in [1] in a sequence of (time) intervals.

 ISBN: 978-84-617-8694-7

For our purposes, we assume an homogeneous-mixing for the disease dynamics. That is, all the populations have same rates of disease-causing contacts. The bovine population $N_B(t)$ is split into three subpopulations, namely, susceptible $\bar{S}_B(t)$; infected $\bar{I}_B(t)$; and controlled $\bar{C}_B(t)$, i.e., treated against Babesiosis. Ticks population, $N_T(t)$, is only divided into two subpopulations susceptible $\bar{S}_T(t)$ and infected $\bar{I}_T(t)$. The susceptible cattle may become infected due to ticks bites infected by the parasite at a rate β_B and susceptible ticks may become infected after a blood meal on infected bovine at a rate β_T. Parameters λ_B and α_B are the fractions of infected bovines which are controlled and of controlled bovines which return to be susceptible to parasite, respectively. We consider, in the tick population, vertical transmission of bovine Babesiosis with probability $1-p$. The birth and death rates are considered equal in each population, being denoted by μ_B for the bovine population and μ_T for the tick population.

Under the above assumptions, we obtained the continuous model described by the following system of first order equations [1]:

$$\begin{cases} \bar{S}_B(t) = \mu_B(\bar{S}_B(t)+\bar{C}_B(t)) + \alpha_B\bar{C}_B(t) - \mu_B\bar{S}_B(t) - \beta_B\bar{S}_B(t)\dfrac{\bar{I}_T(t)}{N_T(t)}, \\ \bar{I}_B(t) = \mu_B\bar{I}_B + \beta_B\bar{S}_B(t)\dfrac{\bar{I}_T(t)}{N_T(t)} - \mu_B\bar{I}_B - \lambda_B\bar{I}_B(t), \\ \bar{C}_B(t) = \lambda_B\bar{I}_B(t) - (\mu_B+\alpha_B)\,\bar{C}_B(t), \\ \bar{S}_T(t) = \mu_T(\bar{S}_T + p\bar{I}_T) - \beta_T\bar{S}_T(t)\dfrac{\bar{I}_B(t)}{N_B(t)} - \mu_T\bar{S}_T(t), \\ \bar{I}_T(t) = \beta_T\bar{S}_T(t)\dfrac{\bar{I}_B(t)}{N_B(t)} + (1-p)\mu_T\bar{I}_T(t) - \mu_T\bar{I}_T(t), \end{cases} \tag{1}$$

For simplicity, we considered the bovine and tick populations as constants, i.e., $N_B'(t) = 0$ and $N_T'(t) = 0$. Also, we introduced the following proportions

$$S_B(t) = \frac{\bar{S}_B(t)}{N_B(t)},\ I_B(t) = \frac{\bar{I}_B(t)}{N_B(t)},\ C_B(t) = \frac{\bar{C}_B(t)}{N_B(t)},\ S_T(t) = \frac{\bar{S}_T(t)}{N_T(t)},\ I_T(t) = \frac{\bar{I}_T(t)}{N_T(t)},$$

and equalities $C_B(t) = 1 - S_B(t) - I_B(t)$ and $S_T(t) = 1 - I_T(t)$. The model (1) is reduced to a normalized system with three differential equations of first order:

$$\begin{cases} S_B'(t) = (\mu_B+\alpha_B)\,(1 - S_B(t) - I_B(t)) - \beta_B I_T(t) S_B(t) \\ I_B'(t) = \beta_B S_B(t) I_T(t) - \lambda_B I_B(t), \\ I_T'(t) = \beta_T\,(1 - I_T(t))\,I_B(t) - \mu_T p I_T(t). \end{cases} \tag{2}$$

Here, all the parameters involved are considered non-negative.

For our direct discrete-time epidemic model [2], we assume that the population in the $(t+1)-th$ generation is a function of the $t-th$ generation with $t \in \mathbb{N}$. With similar procedures as above, we directly obtain the following system of difference equations:

$$\begin{cases} S_B(t+1) = S_B(t) + (\mu_B+\alpha_B)\,(1 - S_B(t) - I_B(t)) - \beta_B I_T(t) S_B(t) \\ I_B(t+1) = I_B(t) + \beta_B S_B(t) I_T(t) - \lambda_B I_B(t), \\ I_T(t+1) = I_T(t) + \beta_T\,(1 - I_T(t))\,I_B(t) - \mu_T p I_T(t). \end{cases} \tag{3}$$

On the other hand, using discretization schemes, we obtain a more complex discrete model. In particular, we use the forward Euler scheme (see [6] and [7]).

 ISBN: 978-84-617-8694-7

Specifically, we choose a time step size $h > 0$, for any $t \geq 0$. Taking into account the above assumptions, we get the system:

$$\begin{cases} S_B(t+1) = [1 - h(\mu_B + \alpha_B + \beta_B I_T(t))] S_B(t) + h(\mu_B + \alpha_B)(1 - I_B(t)) \\ I_B(t+1) = [1 - h\lambda_B] I_B(t) + h\beta_B S_B(t) I_T(t), \\ I_T(t+1) = [1 - h(\mu_T p + \beta_T I_B(t))] I_T(t) + h\beta_T I_B(t), \end{cases} \tag{4}$$

for all initial conditions $S_B(t_0) > 0$, $I_B(t_0) > 0$ and $I_T(t_0) > 0$ satisfying $S_B(t_0) + I_B(t_0) = 1$ and $I_T(t_0) > 0$ with $t_0 \in \mathbb{N}$.

We assume as the state space of systems (3) and (4) the set

$$\Omega = \left\{(S_B, I_B, I_t) \in \mathbb{R}^3_+ : 0 \leq S_B + I_B \leq 1, 0 \leq I_T \leq 1\right\}$$

which is a positive invariant set.

If the step size $h = 1$ in the model (4), we obtain the model (3) above. The model (3) is epidemiologically meaningful if, and only if, the parameters involved in modeling verifies the following constraints

$$1 - (\mu_B + \alpha_B) \geq 0, \quad 1 - \beta_B \geq 0, \quad 1 - \lambda_B \geq 0, \quad 1 - \beta_T \geq 0, \quad 1 - \mu_T p \geq 0.$$

These constraints are fundamental in order to prove results similar to the continuous system (2).

The mathematical results related to model (4) are similar to models (2) and (3) for $0 < h \leq 1$, i.e, the equilibrium points existence and stability depend only on the threshold parameter:

$$\mathcal{R}_0 = \frac{\beta_B \beta_T}{\lambda_B \mu_T p}.$$

When $\mathcal{R}_0 < 1$, the system has also the two equilibria, disease-free equilibrium asymptotically stable in Ω and endemic equilibrium unstable, being the last one in the outside of Ω. If the parameter value $\mathcal{R}_0 = 1$ the system undergoes a transcritical bifurcation.

Nevertheless, when $h > 1$, exists a theshold value, h^*, such that the disease-free equilibrium is asymptotically stable into Ω, for $h < h^*$, where

$$h^* = \min\left\{\frac{1}{\beta_B}, \frac{1}{\beta_T}, \frac{1}{\lambda_B}, \frac{1}{\mu_B + \alpha_B}, \frac{1}{\mu_T p}\right\}.$$

If $h > h^*$, the desease-free equilibrium is unstable.

Besides, we are able to confirm the absence of contradictions between the different versions. These results have been validated by numerical simulations.

Acknowledgements

Deccy Y. Trejos has performed this work within the specific educational cooperation agreement for fellowships in doctoral programs and short research stays for professors with doctorates between the Carolina Foundation, Spain and the University Francisco José of Caldas, Colombia (CSU Resolution 038 of 2016). Jose C. Valverde thanks FEDER OP2014-2020 of Castilla-La Mancha (Spain) and the Ministry of Economy and Competitiveness of Spain for their support for this work under the grants GI20173946 and MTM2014-51891-P, respectively.

 ISBN: 978-84-617-8694-7

References

[1] D. F. Aranda, D. Y. Trejos, J. C. Valverde and R. J. Villanueva, *A mathematical model for Babesiosis disease in bovine and tick populations*, Math. Meth. Appl. Sci. **35** (2012) 249–256.

[2] D. F. Aranda, D. Y. Trejos, J. C. Valverde, *A discrete epidemic model for bovine Babesiosis and tick populations*, (to appear).

[3] L. Bouzid and O. Belhamiti, *Effect of seasonal changes on predictive model of bovine babesiosis transmission*, Int. J. Model. Simul. Sci. Comput. **0, 1750030** (2017) [17 pages],

[4] J. P. Carvalho, L. C. Cardoso, E. Monteiro and N. H. Lemes, *A fractional-order epidemic model for bovine Babesiosis disease and tick populations*, Abstract Appl. Anal. **2015** (2015) [10 pages].

[5] A. Friedman and A. Yakubu, *A Bovine Babesiosis model with dispersion*, Bull. Math. Biol. **76** (2014) 98–135.

[6] Z. Hu, Z. Teng and H. Jiang, *Stability analysis in a class of discrete SIRS epidemic models* Nonlinear Anal. Real World Appl. **18** (2012) 2017–2033.

[7] Z. Hu, Z. Teng and Z. Long., *Stability and bifurcation analysis in a discrete SIR epidemic model* Math. Comput. Simul. **97** (2014) 80–93.

[8] H. Pourbashash, *Global analysis of the Babesiosis disease in bovine and tick populations model and numerical simulation with multistage modified sinc method*, Iran. J. Sci. Technol. Trans. A Sc. (in press)

[9] C. M. Saad-Roy, Z. Shuai and P. Driessche, *Models of bovine Babesiosis including juvenile cattle*, Bull. Math. Biol. **77** (2015) 514–547.

[10] Z. U. A. Zafar, K. Rehan and M. Mushtaq, *A Bovine Babesiosis model with dispersion*, Adv. Differ. Equ. **86** (2017) 1–19.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

An improved h-adaptive method with different applications for the generalised finite differences method in 2D and 3D

M. Ureña[1], J.J. Benito[1], F. Ureña[2], A. García[1], L. Gavete[3] and L. Benito[4]

[1] *Departamento de Construcción y Fabricación, Universidad Nacional de Educación a Distancia (UNED)*

[2] *Departamento de Matemática Aplicada, Universidad de Castilla-La Mancha (UCLM)*

[3] *Departamento de Matemática Aplicada a los Recursos Naturales, Universidad Politécnica de Madrid (UPM)*

[4] *–, UBICCA SL*

emails: miguelurenya@gmail.com, jbenito@ind.uned.es, fuprieto@terra.es, angelochurri@gmail.com, lu.gavete@upm.es, luisbenitonunez@gmail.com

Abstract

The generalised finite differences method allows the use of irregular clouds of nodes. The optimal values of the key parameters of the method vary depending on how the nodes in the cloud are distributed and this may be specially complicated in 3D. Therefore, we establish two criteria to allow the automation of the selection process of the key parameters. These criteria depends on two discrete functions, one of them penalises distances and the other penalises imbalances.

We propose an improved and more versatile h-adaptive method that allows adding, moving and removing nodes. In order to decide on which nodes to act we use an indicator of the error a posteriori. This h-adaptive method gives more accuracy results than those presented for the generalised finite differences method so far and, in addition, using fewer nodes.

Key words: Meshless method, Generalised finite differences, Adaptive method
MSC 2000: 65M06

1 Introduction

The generalised finite differences method, from here onwards GFDM, is a meshless method that allows, unlike the classical finite differences method, to solve partial differential equations over domains where a regular grid is not possible.

An overview of advances in h-adaptive methods with the GFDM so far may be as follows. In [1] an error indicator is defined and compared to the error proposed by Orkisz in [2] with similar results but lower computational cost. Moreover, in [3], a quality index for the error indicator is used in order to analyse its efficiency. Also in [1] two parameters associated with the adaptive process arise, a limit for the distance to which a new node can be placed (dpa) and a threshold of the error to be reached. Both parameters should be reduced little by little in several steps. In [4] the GFDM and the EFG method are compared, obtaining better results in the first one. In [4] it is also showed a case where a strategic irregularity provides results almost as accuracy as those provided by a finer discretisation. The GFDM and classical finite differences are compared in [5], obtained better results in the first one. Moreover, it is proposed an adaptive method based on adding a node in the barycentre of the triangles with higher area formed in each star. In [6] an index of irregularity of a cloud of nodes is defined. The quadtree method is applied in cloud of nodes with sharp variations of gradient in [7].

In this paper, we make the following contributions. Firstly, we define two penalty functions based on distances and imbalances in each star and, in this way, we establish two criteria to choose the number of nodes, the criterion of selection and the potential weighting function. This is specially important in 3D. Secondly, all adaptive methods analysed in previous works have been put together in an improved method with new functionalities such as the ability of moving and removing nodes. Our aim is to reach error reductions as low as possible with the fewest number of nodes and for that reason the quadtree method is not included in this paper.

2 Theoretical background

Given a domain $D \subset \mathbb{R}^n$, $n = 2, 3$, the aim is to solve a problem defined by the second order linear partial differential equation

$$\mathcal{L}_2\big(U(\mathbf{x})\big) = f(\mathbf{x}) \qquad \forall \mathbf{x} \in \Omega = int(D) \tag{1}$$

with boundary condition

$$\mathcal{L}_1\big(U(\mathbf{x})\big) = g(\mathbf{x}) \qquad \forall \mathbf{x} \in \Gamma = fr(D) \tag{2}$$

where $\mathbf{x} \in D$, U is the unknown function and f and g are continuous functions.

 ISBN: 978-84-617-8694-7

In order to solve the problem, a discretisation M of the domain is considered and for each star in the discretisation, it is defined the function given by the sum of the weighted quadratics errors in Taylor series for terms over third order

$$B(\mathbf{D}_u) = \sum_{i=1}^{N} \left(u_0 - u_i + \boldsymbol{\varepsilon}_i^T \mathbf{D}_u \right)^2 w_i^2 \tag{3}$$

where N is the number of nodes of the star, $w_i = w(\mathbf{x}_0, \mathbf{x}_i)$ is the weighting function and $u_0 = u(\mathbf{x}_0)$ and $u_i = u(\mathbf{x}_i)$ are the approximations of $U_0 = U(\mathbf{x}_0)$ and $U_i = U(\mathbf{x}_i)$, respectively. In both 2D and 3D, the vectors $\mathbf{D}_u$ and $\boldsymbol{\varepsilon}_i$ are

$$\mathbf{D}_u = \begin{cases} \left(\dfrac{\partial u_0}{\partial x} \quad \dfrac{\partial u_0}{\partial y} \quad \dfrac{\partial^2 u_0}{\partial x^2} \quad \dfrac{\partial^2 u_0}{\partial x \partial y} \quad \dfrac{\partial^2 u_0}{\partial y^2} \right)^T \\ \left(\dfrac{\partial u_0}{\partial x} \quad \dfrac{\partial u_0}{\partial y} \quad \dfrac{\partial u_0}{\partial z} \quad \dfrac{\partial^2 u_0}{\partial x^2} \quad \dfrac{\partial^2 u_0}{\partial y^2} \quad \dfrac{\partial^2 u_0}{\partial z^2} \quad \dfrac{\partial^2 u_0}{\partial x \partial y} \quad \dfrac{\partial^2 u_0}{\partial x \partial z} \quad \dfrac{\partial^2 u_0}{\partial y \partial z} \right)^T \end{cases} \tag{4}$$

$$\boldsymbol{\varepsilon}_i = \begin{cases} \left(h_i \quad k_i \quad \dfrac{h_i^2}{2} \quad h_i k_i \quad \dfrac{k_i^2}{2} \right)^T \\ \left(h_i \quad k_i \quad l_i \quad \dfrac{h_i^2}{2} \quad \dfrac{k_i^2}{2} \quad \dfrac{l_i^2}{2} \quad h_i k_i \quad h_i l_i \quad k_i l_i \right)^T \end{cases} \tag{5}$$

In order to minimise the error, this function is minimised and the following system is obtained

$$A\mathbf{D}_u = \mathbf{b} \tag{6}$$

where

$$A = \sum_{i=1}^{N} w_i^2 \boldsymbol{\varepsilon}_i \boldsymbol{\varepsilon}_i^T \tag{7}$$

$$\mathbf{b} = \sum_{i=1}^{N} w_i^2 (u_i - u_0) \boldsymbol{\varepsilon}_i \tag{8}$$

Solving for $\mathbf{D}_u$, the approximated values of the partial derivatives are obtained and, substituting these values in the original equation (1), the *equation of the star* is attained

$$-\lambda_0 u_0 + \sum_{i=1}^{N} \lambda_i u_i = f \tag{9}$$

The numerical values in each inner node are obtained solving the linear system of equations formed by all equations of the star. For more details about this, see [8] and [9].

 ISBN: 978-84-617-8694-7

2.1 Criteria for selecting the key parameters of the stars

Definition 1 (Radius of the star) *The radius of a star is defined as the average distance between the central node and the rest of the nodes of the star and is denoted by* $\overline{sr}$.

Definition 2 (Distance penalty function of a star) *The distance penalty function of a star,* $f_d : \Omega \longrightarrow [0,1]$*, is defined as the numbers of nodes outside the interval* $[\overline{sr} - 2\sigma_{sr}, \overline{sr} + 2\sigma_{sr}]$ *divided by the number of nodes of the star, being* σ_{sr} *the standard deviation of the distances between the central node and the rest of the nodes.*

Given a star with N nodes, the number of nodes by octant (quadrant) is the integer part of $N/8$ ($N/4$) and it is denoted by NO.

Definition 3 (Imbalance penalty function of a star) *The imbalance penalty function of a star,* $f_b : \Omega \longrightarrow [0,1]$*, is defined as the sum of the number of missing or extra nodes by octant (quadrant) divided by the maximum possible penalty in the star.*

The maximum possible penalty is $N + 6NO$ in 3D and $N + 2NO$ in 2D.

The index of irregularity of a cloud of nodes (IIC) defined in [6] gives an insight of how irregular the cloud is but it does not provide information about the choice of the number of nodes, the criterion of selection and the exponent of the potential weighting function. Although we know that the optimal parameters are 18 nodes with the distance criterion and 24 nodes with the octant criterion and, with regard to the weighting function, the potential function give better results than exponential ones, we want to know which the best choice is in each case and, for this purpose, we apply the following empirical criteria, as it will be seen in the examples in section 3.2.

If a cloud of nodes has some star whose balance penalty function is greater than 0.5, then 24-stars formed by the octant criterion are chosen. Otherwise, 18-stars formed by the distance criterion are chosen.

If a cloud of nodes has more than 25% of its stars with positive values of the distance penalty function, then low exponents in the potential weighting functions are chosen. Otherwise, higher exponents may be chosen.

Although the penalty functions will be used in 2D and 3D, these criteria are established only for 3D cases because that is where they really are needed.

2.2 H-adaptive method

Two algorithms are distinguished in the h-adaptive method, one that allows the addition of nodes and another that allows movement and elimination. In both algorithms, a node will be processed if the estimation of the error in that node is above a given value. The errors are calculated with the indicator obtained in [1].

 ISBN: 978-84-617-8694-7

Addition algorithm. In 2D case, we use the algorithm defined in [5] that adds in each selected node a new node in the barycentre of the triangles with higher area. In 3D case and for each star, the algorithm adds nodes halfway between the selected node and the central node. To do this, the algorithm handles two options, the first one selects the farthest node and adds a single node and the second one selects the nodes whose distance is greater than the radius of the star and adds the corresponding nodes. In all cases, a new node is added if the distance with the rest of the nodes is less than $\alpha \cdot dpa$, $\alpha \in (0, 1]$. *dpa* is the minimum distance between whatever two nodes in the cloud.

Movement algorithm. The next algorithm will be implemented jointly with the previous and may be applied in any domain. For each star $E(\mathbf{x}_0) = \{\mathbf{x}_0; \mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N\}$, we will consider the connections between the central node and the other nodes of the star as vectors whose starting point is the central node. The algorithm modifies the position of the central node of the star. The new position of the central node will be determined by the ending point of the resultant vector obtained by adding the N vectors, each of which will be weighted by the estimation of the error (e_i) on that node, namely, the linear combination

$$\overrightarrow{u_0} = \varphi_E(e_1)\overrightarrow{u_1} + \varphi_E(e_2)\overrightarrow{u_2} + ... + \varphi_E(e_N)\overrightarrow{u_N} \tag{10}$$

where $\overrightarrow{u_0} = \overrightarrow{\mathbf{x}_0\overline{\mathbf{x}}_0}$ with $\overline{\mathbf{x}}_0$ the new position of the central node, $\overrightarrow{u_i} = \overrightarrow{\mathbf{x}_0\mathbf{x}_i}, \forall i = 1, ..., N$ and φ_E the weighting function for the star E defined in terms of the estimation of the error is as follows

$$\varphi_E(e_i) = \frac{\frac{1}{\kappa + e_i}}{S_E} \tag{11}$$

where $\kappa > 0$ is a constant to avoid problems with null values such as in Dirichlet boundary and

$$S_E = \sum_{i=1}^{N} \frac{1}{\kappa + e_i} \tag{12}$$

In order to implement the algorithm, the following constraints are taken into account:

C1. If the distance between the new position of the central node and the old one is greater than $\mu \cdot dmm$, with $0 < \mu < 1$ and dmm the minimum distance between the central node and the other nodes of the star, then the new position of the central node is in the direction of the resultant vector but at distance of $\mu \cdot dmm$.

C2. If the distance from the new position of the central node to any node of the discretisation is less than $\alpha \cdot dpa$ the node is removed.

This movement algorithm helps to reduce the global error because of the reduction of the estimation of the error in the node itself as a consequence of improving the distribution of the nodes and because it produces a node repulsion effect in areas with close nodes and large values of the error, making spaces which can be occupied by new nodes in the adding step. In addition, it may be that, due to specific errors in the initial configuration of the cloud, the initial *dpa* distance is smaller than expected and then it is possible to establish a

 ISBN: 978-84-617-8694-7

distance *dpa* higher than the one obtained automatically in order to remove the nodes with the highest estimation of the error.

It is important to notice that both addition and movement algorithm must be used with stars formed by the octant criterion since the cloud of nodes will become more irregular in each step. For this reason, and whenever the committed error when using stars formed by the distance criterion is less than the committed error when using stars formed by the octant criterion, we show that error using the distance criterion in the examples and verify that the adaptive method reduces it.

In summary, we have three algorithms, one in 2D and two in 3D, to add nodes (addition algorithm), an algorithm to move nodes (movement algorithm) and an adaptive method that brings them together.

3 Numerical results

The error is evaluated using the following global error formula:

$$\text{Global Error } (\%) = \frac{1}{|e_{max}|}\sqrt{\frac{\sum_{i=1}^{NI} e_i^2}{NI}} \times 100 \tag{13}$$

where e_{max} is the maximum value of the error indicator and NI is the number of inner nodes.

In order to compare the addition algorithm and the adaptive method, four partial differential equations are solved in four irregular clouds of nodes. The movement algorithm will be implemented alternately with the addition algorithm. Moreover, in both algorithms, a node will be processed if the estimation of the error in that node is above the average of the estimation of the error in the whole cloud plus twice standard deviation. In all cases the values for the different parameters are $\alpha = 0.5$, $\mu = 2/3$ and $\kappa = max(e_1, \ldots, e_N)$, being e_i, $i = 1, \ldots, N$, the estimation of the error in each node of the star.

In 3D cases it is difficult to visualize the distribution of nodes but, despite this, we show both the initial and final clouds of nodes so that the difference can be appreciated.

3.1 Comparison between the addition algorithm and the adaptive method

Example 1. We consider the following equation over an arbitrary irregular domain contained in $D = [0, 3] \times [0, 2]$

$$\frac{\partial^2 U}{\partial x^2} + \frac{\partial^2 U}{\partial y^2} = 0 \tag{14}$$

with Dirichlet boundary conditions and exact solution

$$U(x, y, z) = \ln(x^2 + y^2) \tag{15}$$

 ISBN: 978-84-617-8694-7

where 8-stars formed with the quadrant criterion and weighting function $w(\mathbf{x}_0, \mathbf{x}_i) = \|\mathbf{x}_i - \mathbf{x}_0\|^{-4}$ have been used in each node.

Eight steps have been taken with the addition algorithm (barycentre method) and eight steps have been taken with the adaptive method. The error value is improved by adaptive method even using less nodes than the addition algorithm, as can be seen in table 1. Fig. 1 shows the initial cloud (55 nodes) and the final cloud (120 nodes) when adaptive method has been applied.

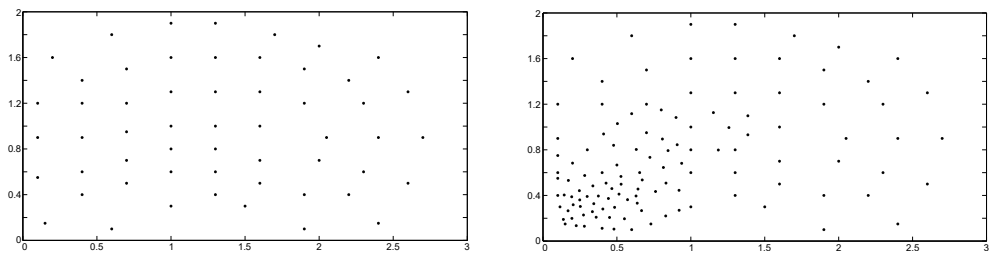


Figure 1: Cloud of nodes, initial (55 nodes) and final (120 nodes), applying the adaptive method in example 1

Table 1: Initial and final errors and number of nodes

Example 1	Initial error (nodes)	Final error (nodes)
Addition algorithm	$6.16 \cdot 10^{-1}\%$ (55)	$3.58 \cdot 10^{-2}\%$ (138)
Adaptive method	$6.16 \cdot 10^{-1}\%$ (55)	$3.29 \cdot 10^{-2}\%$ (120)

Example 2. We consider the following equation over an arbitrary irregular domain contained in $D = [0, 1]^3$

$$\frac{\partial^2 U}{\partial x^2} + \frac{\partial^2 U}{\partial y^2} + \frac{\partial^2 U}{\partial z^2} = \frac{20}{x^2 + y^2 + z^2} \tag{16}$$

with Dirichlet boundary conditions and exact solution

$$U(x, y, z) = 10 \ln(x^2 + y^2 + z^2) \tag{17}$$

where 24-stars formed with the octant criterion and weighting function $w(\mathbf{x}_0, \mathbf{x}_i) = \|\mathbf{x}_i - \mathbf{x}_0\|^{-6}$ have been used in each node.

Three steps have been taken with the second option of the addition algorithm and two steps have been taken with the adaptive method. The estimation of the initial error using

 ISBN: 978-84-617-8694-7

a star with 18 nodes formed by the distance criterion is $7,37 \cdot 10^{-2}\%$ and this value is improved by the adaptive method as can be seen in table 2. Fig. 2 shows the initial cloud (216 nodes) and the final cloud (250 nodes) when the adaptive method has been applied.

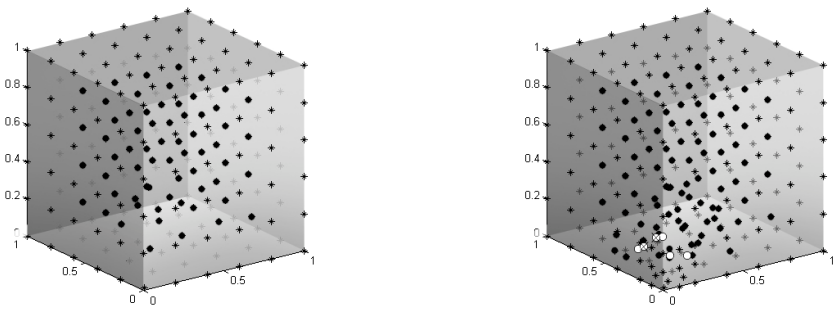


Figure 2: Cloud of nodes, initial (216 nodes) and final (250 nodes), applying the adaptive method in example 2. Legend: $\star$ boundary node, $\triangle$ inner node initial position, $\circ$ inner node final position, $\otimes$ removed node.

Table 2: Initial and final errors and number of nodes

Example 2	Initial error (nodes)	Final error (nodes)
Addition algorithm	$1,30 \cdot 10^{-1}\%$ (216)	$2,76 \cdot 10^{-2}\%$ (263)
Adaptive method	$1,30 \cdot 10^{-1}\%$ (216)	$2,54 \cdot 10^{-2}\%$ (250)

Example 3. We consider the following equation over an arbitrary irregular domain contained in $D = [0,1]^3$

$$\frac{\partial^2 U}{\partial x^2} + \frac{\partial^2 U}{\partial y^2} + \frac{\partial^2 U}{\partial z^2} = 0 \tag{18}$$

with Dirichlet boundary conditions and exact solution

$$U(x,y,z) = e^x \sin y + e^y \sin z + e^z \sin x \tag{19}$$

where 24-stars formed with the octant criterion and weighting function $w(\mathbf{x}_0, \mathbf{x}_i) = \|\mathbf{x}_i - \mathbf{x}_0\|^{-4}$ have been used in each node.

A single step has been taken with both algorithms. Fig. 3 shows the initial cloud of nodes (136 nodes) and the final one (137 nodes) obtained when the adaptive method has been applied. How the error decreases with the first option of the addition algorithm and with the adaptive method can be seen in table 3.

 ISBN: 978-84-617-8694-7

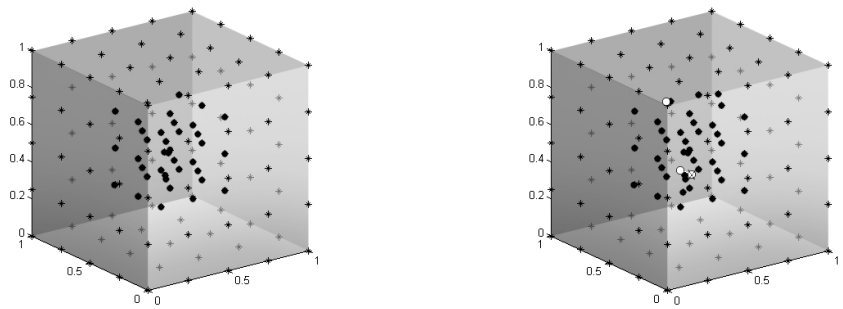


Figure 3: Clouds of nodes, initial (136 nodes) and final (137 nodes), applying the adaptive method in example 3. Legend: $\star$ boundary node, $\triangle$ inner node initial position, $\circ$ inner node final position, $\otimes$ removed node.

Table 3: Initial and final errors and number of nodes

Example 3	Initial error (nodes)	Final error (nodes)
Addition algorithm	$7,47 \cdot 10^{-3}\%$ (136)	$7,11 \cdot 10^{-3}\%$ (138)
Adaptive method	$7,47 \cdot 10^{-3}\%$ (136)	$4,39 \cdot 10^{-3}\%$ (137)

Example 4. We consider the following equation over an arbitrary irregular domain, in particular, the sphere with centre the origin and radius 0.7.

$$\frac{\partial^2 U}{\partial x^2} + \frac{\partial^2 U}{\partial y^2} + 2\frac{\partial^2 U}{\partial z^2} = 0 \tag{20}$$

with Dirichlet boundary conditions and exact solution

$$U(x, y, z) = e^z \sin x \cos y \tag{21}$$

where 24-stars formed with the octant criterion and weighting function $w(\mathbf{x}_0, \mathbf{x}_i) = \|\mathbf{x}_i - \mathbf{x}_0\|^{-4}$ have been used in each node.

Three steps have been taken with the first option of the addition algorithm and two steps have been taken with the adaptive method. Fig. 4 shows the initial cloud of nodes (575 nodes) and the final one (582 nodes) obtained when the adaptive method has been applied. Table 4 shows how the error decreases when the addition algorithm and the adaptive method are applied.

 ISBN: 978-84-617-8694-7

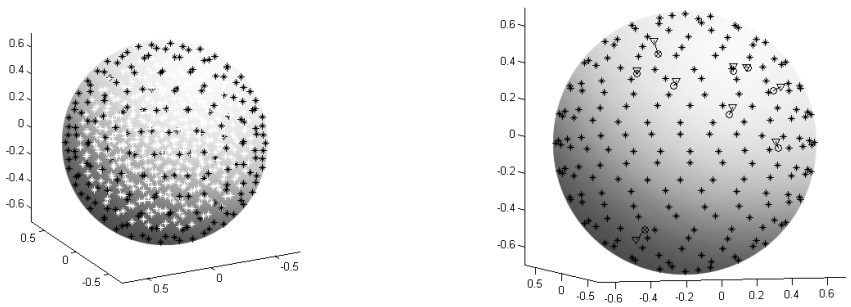


Figure 4: Clouds of nodes, initial (575 nodes) and final (582 nodes), applying the adaptive method in example 4. Legend: ($\star$ boundary node, $\triangle$ inner node initial position, $\circ$ inner node final position, $\otimes$ removed node.)

Table 4: Initial and final errors and number of nodes

Example 4	Initial error (nodes)	Final error (nodes)
Addition algorithm	$2,60 \cdot 10^{-2}\%$ (575)	$2,41 \cdot 10^{-2}\%$ (584)
Adaptive method	$2,60 \cdot 10^{-2}\%$ (575)	$1,94 \cdot 10^{-2}\%$ (582)

3.2 Selection of the key parameters of the stars in 3D

In order to check the most favourable choice of the parameters of the stars (number of nodes, criterion of selection and exponent of the weighting function), we use the 3D examples from the previous section. The best choice for the parameters has the error in bold in Tables 6 and 7 and it is in accordance with the established criteria in this paper as it is showed in Table 5. In Table 6 we vary the number of nodes and the criterion of selection using the optimal exponent, according with the distance penalty criterion, in each case. In Table 7 we vary the exponent of the weighting function using the optimal number of nodes and selection criterion, according with the imbalance penalty criterion, in each case.

Table 5: Criteria for selecting the key parameters of the stars

Penalty	Example 2	Example 3	Example 4
Imbalance	0	7	0
Distance	6	19	174

 ISBN: 978-84-617-8694-7

Table 6: Errors depending on the number of nodes and the criterion of selection (DC = Distance criterion and OC = Octant criterion)

Error	Example 2	Example 3	Example 4
18 and DC	$\mathbf{7.37 \cdot 10^{-2}}$	$4.29 \cdot 10^{-1}$	$\mathbf{2.16 \cdot 10^{-2}}$
24 and OC	$1.30 \cdot 10^{-1}$	$\mathbf{7.47 \cdot 10^{-3}}$	$2.60 \cdot 10^{-2}$

Table 7: Errors depending on the exponent of the weighting function

Error	Example 2	Example 3	Example 4
Exponent 4	$1.20 \cdot 10^{-1}$	$8.34 \cdot 10^{-3}$	$\mathbf{2.16 \cdot 10^{-2}}$
Exponent 6	$\mathbf{7.37 \cdot 10^{-2}}$	$\mathbf{7.47 \cdot 10^{-3}}$	$2.50 \cdot 10^{-2}$

4 Conclusions

Two penalty functions for the stars are defined, one of them for the distances and the other one for the imbalance. As consequence, two criteria are established. These criteria allow an automatic selection of the key parameters of the method in 3D. Several examples have showed the appropriate behaviour of these criteria. The minimum distance *dpa* has been redefined in order to give more insight when the addition algorithm is applied.

An improved h-adaptive method with more functionalities has been introduced. This method, apart from adding nodes, has the ability of moving and removing them. A comparison between the addition algorithms and the improved h-adaptive method has been performed. The best results has been obtained with the last one, which highlights the importance of moving and removing nodes and their application together to addition. With regard to the two options that we have handled in the addition algorithm in 3D, the first one works best when the error is distributed in larger regions while the second option works best when the error is more concentrated, as in the studied cases.

Both the automatic selection of the key parameters of the method and the h-adaptive method are very important in 3D where it is very difficult to work with irregular clouds of nodes that are very useful in many problems.

To conclude, we must point out that the clouds of nodes used in the cases presented in this paper have been made to extreme irregular situations and, although this kind of irregularity does not make sense in real cases, we have used it to apply the method under very adverse conditions.

 ISBN: 978-84-617-8694-7

Acknowledgements

The authors acknowledge the support of the Escuela Técnica Superior de Ingenieros Industriales (UNED) of Spain, project Ref: 2017-ICF02.

References

[1] Benito, J.J., Ureña, F., Gavete, L., Alvarez, R., 2003. An h-adaptive method in the generalized finite differences. Computer Methods in Applied Mechanics and Engineering 192, 735759.

[2] Orkisz, J., 1998. Meshless finite difference method II. Adaptative approach. Computation Mechanics, CIMNE.

[3] Ureña, F., Benito, J.J., Alvarez, R., Gavete, L., 2005. Computational Error Approximation and H-Adaptive Algorithm for the 3-D Generalized Finite Difference Method. International Journal for Computational Methods in Engineering Science and Mechanics 6, 3139.

[4] Gavete, L., Gavete, M.L., Benito, J.J., 2003. Improvements of generalized finite difference method and comparison with other meshless method. Applied Mathematical Modelling 27, 831847.

[5] Benito, J., Ureña, F., Gavete, L., Alonso, B., 2009. Application of the Generalized Finite Difference Method to improve the approximated solution of pdes. Computer Modelling in Engineering & Sciences 38, 3958.

[6] Gavete, L., Benito, J.J., Ureña, F., 2016. Generalized finite differences for solving 3D elliptic and parabolic equations. Applied Mathematical Modelling 40, 955965.

[7] Gavete, L., Gavete, M.L., Ureña, F., Benito, J.J., 2015. An Approach to Refinement of Irregular Clouds of Points Using Generalized Finite Differences. Mathematical Problems in Engineering 2015, 9.

[8] Benito, J.J., Ureña, F., Gavete, L., 2001. Influence of several factors in the generalized finite difference method. Applied Mathematical Modelling 25, 10391053.

[9] Ureña, M., Benito, J.J., Ureña, F., Salete, E., Gavete, L., 2017. Application of generalised finite differences method to reflection and transmission problems in seismic SH waves propagation. Mathematical Methods in the Applied Sciences. doi:10.1002/mma.4268

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

Existence and uniqueness of solutions for a nonlocal fractional boundary value problem

Min Wang[1]

[1] *Department of Mathematics, Rowan University, Glassboro, NJ 08028, USA*

emails: wangmin@rowan.edu

Abstract

We study a nonlinear fractional boundary value problem with nonlocal boundary conditions. An associated Green's function is constructed as a series of functions by the perturbation approach. Criteria for the existence and uniqueness of solutions are obtained based on it.

Key words: Green's function, nonlocal boundary value problem, fractional calculus,
MSC 2000: primary 34B15; secondary 34B10

1 Introduction

In this paper, we study the boundary value problem (BVP) consisting of the fractional differential equation

$$-D_{0+}^{\alpha}u + a(t)u = w(t)f(u,s),\ 0 < t < 1,\ n-1 < \alpha < n,\ n \in \mathbb{N},\ n \geq 3, \tag{1}$$

and the nonlocal boundary condition (BC)

$$u^{(k)}(0) = 0,\ k = 0, 1, \ldots, n-1,\ u(1) = \int_0^1 u(s)dA(s), \tag{2}$$

where the following assumptions are satisfied:

(i) $a \in C[0,1]$, $w \in L[0,1]$ with $w(t) \not\equiv 0$ a.e. on $[0,1]$, and $f \in C(\mathbb{R} \times [0,1], \mathbb{R})$.

(ii) $A : [0,1] \to \mathbb{R}$ is a function of bounded variation and $\int_0^1 u(s)dA(s)$ denotes the Riemann-Stieltjes integral of u with respect to A.

 ISBN: 978-84-617-8694-7

(iii) $D^{\alpha}_{0+}h$ is the α-th Riemann-Liouville fractional derivative of h for $h : [0,1] \to \mathbb{R}$ defined by

$$D^{\alpha}_{0+}h(t) = \frac{1}{\Gamma(l-\alpha)}\frac{d^l}{dt^l}\int_0^t (t-s)^{l-\alpha-1}h(s)ds,\ l = \lfloor \alpha \rfloor + 1,$$

provided the right-hand side exists with Γ the Gamma function.

Remark 1.1 It is easy to see that the Riemann-Stieltjes integral in BC (2)

$$u(1) = \int_0^1 u(s)dA(s)$$

covers the multi-point and integral BCs as special cases.

Fractional differential equations have extensive applications in various fields of science and engineering. Many phenomena in viscoelasticity, electrochemistry, control theory, porous media, electromagnetism, and other fields, can be modeled by fractional differential equations. We refer to the reader [10, 13] and references therein for some applications.

The existence of solutions is an essential problem for BVPs of fractional differential equations. It has been studied by many authors, see [2–9, 11, 14] and references therein. Due to certain special properties of fractional calculus, critical point theory can only be applied to study equations involving both the left and right Riemann-Liouville fractional derivatives; see for example [1]. To the best of our knowledge, if only the left (or right) Riemann-Liouville fractional derivatives are involved, the only feasible approach to study the existence of solutions of a BVP is to convert the problem to an integral equation and use various techniques to find the fixed points.

The special case of BVP (1), (2) with $a(t) \equiv 0$ on $[0,1]$ has been studied by Zhang and Han [16] and Tan, Cheng, and Zhang [12]. Henderson and Luca [8] further studied a system of coupled nonlocal fractional BVPs. But the general case of BVP (1), (2) with $a \not\equiv 0$ has not been considered in the literature. It is notable that due to the unusual feature of the fractional calculus, even for a linear fractional BVP, it is not easy to find the solution when the equation contains multiple terms involving u or its derivative. For instance, to the best of our knowledge, there is no results on the solutions of the BVP involving the equation

$$-D^{\alpha}_{0+}u + a(t)u = h(t),$$

and the BC (2).

In this paper, we first utilize the perturbation approach developed in [4–7] to derive the Green's function for the BVP involving the equation

$$-D^{\alpha}_{0+}u + a(t)u = 0, \tag{3}$$

and BC (2) and then study the nonlinear BVP (1), (2) by the fixed point theory.

This paper is organized as follows: After this introduction, our main results are stated in Section 2. All the proofs are given in Section 3.

 ISBN: 978-84-617-8694-7

2 Main results

We first consider the Green's function for BVP (3), (2). The following notations are needed. Let $\Lambda \in \mathbb{R}$ and $G_0 \in C([0,1]\times[0,1],\mathbb{R})$ be defined by

$$\Lambda = \int_0^1 t^{\alpha-1} dA(t)$$

and

$$G_0(t,s) = \begin{cases} \dfrac{[t(1-s)]^{\alpha-1} - (t-s)^{\alpha-1}}{\Gamma(\alpha)}, & 0 \le s \le t \le 1, \\ \dfrac{[t(1-s)]^{\alpha-1}}{\Gamma(\alpha)}, & 0 \le t \le s \le 1. \end{cases}$$

Then define $\mathcal{G}_A : [0,1] \to \mathbb{R}$ by

$$\mathcal{G}_A(s) = \int_0^1 G_0(t,s) dA(t).$$

Throughout this paper, we assume

(H1) $0 \le \Lambda < 1$ and $\mathcal{G}_A(s) \ge 0$ on $[0,1]$.

Define

$$H_0(t,s) = \frac{t^{\alpha-1}}{1-\Lambda}\mathcal{G}_A(s) + G_0(t,s) \tag{4}$$

and

$$\overline{H}_0 = \frac{\max_{s\in[0,1]} |\mathcal{G}_A(s)|}{1-\Lambda} + \frac{1}{\Gamma(\alpha-1)}. \tag{5}$$

Remark 2.1 It is known that $H_0 \in C([0,1]\times[0,1],\mathbb{R})$ is the Green's function for BVP (3), (2) with $a \equiv 0$ and $H_0(t,s) \ge 0$ on $[0,1]\times[0,1]$ when (H1) holds; the reader is referred to [12, 16] for more properties of H_0.

Let $H : [0,1]\times[0,1] \to \mathbb{R}$ be defined by

$$H(t,s) = \sum_{n=0}^{\infty} (-1)^n H_n(t,s) \tag{6}$$

with H_0 defined above and $H_n : [0,1]\times[0,1] \to \mathbb{R}$, $n = 1, 2, \ldots$, defined by

$$H_n(t,s) = \int_0^1 a(\tau) H_0(t,\tau) H_{n-1}(\tau,s) d\tau, \quad n \ge 1.$$

We also need the following assumption:

 ISBN: 978-84-617-8694-7

(H2) $\overline{a} := \max_{t\in[0,1]} |a(t)| < \overline{H}_0^{-1}$ with $\overline{H}_0$ defined by (5).

Then we have the following result.

Theorem 2.1 *Assume (H1) and (H2) hold. Then H defined by* (6) *as a series of functions is uniformly convergent for $(t,s) \in [0,1]\times[0,1]$ and continuous on $[0,1]\times[0,1]$. Furthermore, H is the Green's function for BVP* (3), (2) *with*

$$|H(t,s)| \le \frac{\overline{H}_0}{1-\overline{a}\overline{H}_0}$$

on $[0,1]\times[0,1]$, where $\overline{H}_0$ is defined by (5).

With Theorem 2.1, we are ready to study the nonlinear BVP (1), (2). Let

$$U = \max_{t\in[0,1]} \int_0^1 |H(t,s)w(s)|ds, \tag{7}$$

where H is defined by (6). For any $u \in C[0,1]$, define $\|u\| = \max_{t\in[0,1]} |u(t)|$.

Theorem 2.2 *Assume (H1) and (H2) hold. If there exist $M > 0$ and $\kappa \in [0, U^{-1})$ such that*

$$\max_{(x,t)\in[-M,M]\times[0,1]} |f(x,t)| \le M/U, \tag{8}$$

and for any x_1, x_2 with $|x_i| \le M$, $i = 1,2$,

$$|f(x_1,t) - f(x_2,t)| \le \kappa|x_1 - x_2|, \qquad t \in [0,1], \tag{9}$$

then

(a) BVP (1), (2) *has a unique solution $u \in C[0,1]$ with $\|u\| \le M$.*

(b) For any $u_0 \in C[0,1]$ with $\|u_0\| \le M$, the sequence $\{u_n\}$ defined by

$$u_{n+1} = \int_0^1 H(t,s)w(s)f(u_n(s),s)ds, \qquad n = 0,1,2,\ldots,$$

satisfies $\|u_n - u\| \to 0$ as $n \to 0$.

The following result on the existence of solutions is obtained when a weaker condition than Theorem 2.2 is used.

Theorem 2.3 *Assume (H1) and (H2) hold. If there exist $M > 0$ and $\kappa \in [0, U^{-1})$ such that* (8) *holds, then BVP* (1), (2) *has at least one solution $u \in C[0,1]$ with $\|u\| \le M$.*

 ISBN: 978-84-617-8694-7

3 Proofs

The following lemma on the spectral theory in Banach spaces will be used to prove Theorem 2.1. See [15, page 795, items 57b and 57d] for the detail.

Lemma 3.1 *Let X be a Banach space, $\mathcal{A} : X \to X$ a linear operator with the operator norm $\|\mathcal{A}\|$ and the spectral radius $r(\mathcal{A})$ of $\mathcal{A}$. Then*

(a) $r(\mathcal{A}) \leq \|\mathcal{A}\|$;

(b) if $r(\mathcal{A}) < 1$, then $(\mathcal{I} - \mathcal{A})^{-1}$ exists and $(\mathcal{I} - \mathcal{A})^{-1} = \sum_{n=0}^{\infty} \mathcal{A}^n$, where $\mathcal{I}$ stands for the identity operator.

The following lemma is used to estimate the bounds of the Green's function H; see [12, Lemma 2.4] for details.

Lemma 3.2 *Assume (H1) holds. Then for any $(t, s) \in [0, 1] \times [0, 1]$,*

$$0 \leq \frac{t^{\alpha-1}\mathcal{G}_A(s)}{1-\Lambda} \leq H_0(t, s) \leq \overline{H}_0 t^{\alpha-1}, \tag{10}$$

where H_0 and $\overline{H}_0$ are defined by (4) and (5).

In the sequel, we let $X = C[0, 1]$ be the Banach space with the standard maximum norm.

Proof of Theorem 2.1. The proof is in the same spirit of the proofs of Theorem 2.1 and Lemma 3.3 in [5] using Lemmas 3.1, 3.2, and the fact

$$\left| \int_0^1 a(\tau) H_0(t, \tau) d\tau \right| \leq \int_0^1 \overline{a}\overline{H}_0 t^{\alpha-1} d\tau \leq \overline{a}\overline{H}_0 < 1.$$

We omit the details. □

Proof of Theorem 2.2. For any $u \in X$, define $T : X \to X$ by

$$(Tu)(t) = \int_0^1 H(t, s) w(s) f(u(s), s) ds, \tag{11}$$

where H is defined by (6). It is easy to see that T is completely continuous and u is a fixed point of T if and only if u is a solution of BVP (1), (2).

 ISBN: 978-84-617-8694-7

Let $K = \{u \in X \mid \|u\| \leq M\}$. For any $u \in K$ and $t \in [0,1]$,

$$|(Tu)(t)| = |\int_0^1 H(t,s)w(s)f(u(s),s)ds| \leq \int_0^1 |H(t,s)w(s)||f(u(s),s)|ds.$$

By (7) and (8),

$$|(Tu)(t)| \leq \int_0^1 |H(t,s)w(s)|\frac{M}{U}ds = M.$$

Therefore $\|Tu\| \leq M$, i.e. $TK \subset K$.

For any u_1 and $u_2 \in K$ and $t \in [0,1]$,

$$\begin{aligned}|(Tu_1)(t) - (Tu_2)(t)| =& |\int_0^1 H(t,s)w(s)\left[f(u_1(s),s)ds - f(u_2(s),s)\right]ds| \\ \leq& \int_0^1 |H(t,s)w(s)||f(u_1(s),s) - f(u_2(s),s)|ds.\end{aligned}$$

By (9),

$$\begin{aligned}|(Tu_1)(t) - (Tu_2)(t)| \leq& \int_0^1 |H(t,s)w(s)|\kappa|u_1(s) - u_2(s)|ds \\ \leq& U\kappa\|u_1 - u_2\| < \|u_1 - u_2\|.\end{aligned}$$

Hence $\|Tu_1 - Tu_2\| < \|u_1 - u_2\|$. Then parts (a) and (b) follow from the Banach Fixed Point Theorem. □

Theorem 2.3 is proved by Schauder's fixed point theorem. We omit the details.

References

[1] C. Bai, Infinitely many solutions for a perturbed nonlinear fractional boundary-value problem. *Electron. J. Differential Equations* **2013**, No. 136, 12 pp.

[2] M. Feng, X. Zhang, and W. Ge, New existence results for higher-order nonlinear fractional differential equation with integral boundary conditions, *Bound. Value Probl.* (2011), Art. ID 720702, 20 pp.

[3] C. Goodrich, Coercive nonlocal elements in fractional differential equations, *Positivity* (2017) **21**, 377–39.

[4] J. R. Graef, L. Kong, Q. Kong, and M. Wang, Fractional boundary value problems with integral boundary conditions, *Appl. Anal.* **92** (2013), 2008–2020.

 ISBN: 978-84-617-8694-7

[5] J. R. Graef, L. Kong, Q. Kong, and M. Wang, Existence and uniqueness of solutions for a fractional boundary value problem with Dirichlet boundary condition, *Electron. J. Qual. Theory Differ. Equ.* **2013** No. 55, 11 pp.

[6] J. R. Graef, L. Kong, Q. Kong, and M. Wang, A fractional boundary value problem with Dirichlet boundary condition, *Commun. Appl. Anal.*, **19** (2015), 497 – 504.

[7] J. R. Graef, L. Kong, Q. Kong, and M. Wang, On a fractional boundary value problem with a perturbation term, *J. Appl. Anal. Comput.* **7** (2017), 57–66.

[8] J. Henderson and R. Luca, Positive solutions for a system of semipositone coupled fractional boundary value problems, *Bound. Value Probl.* (2016) 2016:61.

[9] J. Henderson and R. Luca, Existence of positive solutions for a singular fractional boundary value problem, *Nonlinear Anal. Model. Control.*, **22** (2017), 99–114.

[10] R. Hilfer, *Applications of Fractional Calculus in Physics*, World Scientific, Singapore, 2000.

[11] Q. Kong and M. Wang,Positive solutions of nonlinear fractional boundary value problems with Dirichlet boundary conditions, *J. Qual. Theory Differ. Equ.* **No. 17** (2012), 1–13.

[12] J. Tan, C. Cheng, and X. Zhang, Positive solutions of fractional differential equation nonlocal boundary value problems, *Adv. Difference Equ.* (2015), 2015:256.

[13] V. Tarasov, *Fractional Dynamics: Applications of Fractional Calculus to Dynamics of Particles, Fields and Media*, Springer-Verlag, New York, 2011.

[14] L. Yang and H. Chen, Unique positive solutions for fractional differential equation boundary value problems, *Appl. Math. Lett.* **23** (2010), 1095–1098.

[15] E. Zeidler, *Nonlinear Functional Analysis and its Applications I: Fixed-Point Theorems*, Springer-Verlag, New York, 1986.

[16] X. Zhang, and Y. Han, Existence and uniqueness of positive solutions for higher order nonlocal fractional differential equations, *Appl. Math. Lett.* **25** (2012), 555–560.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

Lattice Boltzmann Method for Flow and Heat Transfer in Periodic Systems

Zimeng Wang[1] **and Junfeng Zhang**[1]

[1] *Bharti School of Engineering, Laurentian University, Sudbury, ON P3E 2C6, Canada*

emails: zwang4@laurentian.ca, jzhang@laurentian.ca

Abstract

Flow and heat transfer in periodic structures are often encountered in many natural and industrial situations. In this work, the periodic features of fully developed periodic thermal flows have been implemented in the lattice Boltzmann method (LBM) for flow and heat transfer simulations. The unique particular dynamics in LBM is utilized and two numerical approaches, namely the distribution modification (DM) approach and the source term (ST) approach, are proposed. These methods can work with periodic thermal flows under either constant wall temperature (CWT) or surface heat flux (SHF) boundary conditions. Several example simulations are conducted, including flows through flat and wavy channels and flows through a square array with circular cylinders. Results are compared to analytical solutions, previous studies, and our own LBM calculations using different simulation techniques; and good agreement has been observed. These simple however representative simulations demonstrate the accuracy and usefulness of our proposed LBM methods for future thermal periodic flow simulations.

Key words: Lattice Boltzmann Method, Heat Transfer, Mass Transfer, Periodic Flow, Boundary Condition

1 Introduction

Periodic structures are often encountered in heat exchangers and other heat transfer systems, such as wavy or grooved pipes, fin-pin cold plates, and cross-flow heat exchangers [9, 10]. When fluid property changes neglected, identical flow field and similar temperature distributions can be observed in consecutive periodic modules after some distance from the entrance. The flow is then called fully developed in both flow and thermal fields [13, 6]. Numerous studies have been conducted on this topic; among them, existing simulations

 ISBN: 978-84-617-8694-7

mainly used traditional numerical techniques such as the finite-difference and finite-volume methods [13, 11, 15, 6, 1, 2].

Over the past two decades, the lattice Boltzmann method (LBM) has experienced significant development. In addition to various flow systems, LBM has also been successfully adopted to study other processes and phenomena, such as heat and mass transfer and electric and magnetic fields [7, 16, 19]. Unlike other traditional numerical schemes such as the finite-element, finite-difference, and finite-volume methods, where the governing equations of macroscopic properties are discretized mathematically, LBM works with a set of density distributions at each lattice node, and the evolution of these density distributions follows a simple collision-propagation process consecutively. Interestingly, macroscopic equations (such as the continuity and momentum equations for fluid dynamics, the convection-diffusion equation for heat and mass transfer, and the Poisson equation for electric fields) can be correctly recovered from the density distribution dynamics via mathematical analysis [7, 16]. Studies and applications have shown that LBM has some advantages over other methods in simulating multiphase flows, incorporating complex boundary geometries and moving boundaries, and implementing for parallel computation. Moreover, the particulate nature of LBM density distributions provides great convenience for applying periodic and symmetric (also including the free-slip boundary condition in fluid flows and the adiabatic boundary condition in heat transfer) boundary conditions along a lattice grid line, by simply *recycling* or *reflecting* the density distributions that cross the domain boundaries [16, 19]. Both periodic and symmetric boundary treatments have been frequently used in LBM flow simulations; however, this technical merit has not been recognized for LBM simulations of heat transfer processes in periodic flows yet.

In this paper, we extend the pressure periodic boundary method by Zhang and Kwok [20] to fully developed periodic thermal flows with constant wall temperature (CWT) or surface heat flux (SHF) boundaries. The similarity features of temperature field in periodic modules in such systems are first discussed; and then the double distribution LBM method for heat transfer is briefly outlined for readers′ convenience. Two different numerical approaches, the distribution modification (DM) and the source term (ST) approaches, are developed to incorporate these similarity features in LBM simulations. At last, several validation and demonstration simulations are performed to illustrate the correctness, accuracy, and usefulness of our proposed methods in LBM simulations of periodic thermal flows.

2 Theory and Methods

In this section we first describe in detail the periodic features of flow and temperature in fully developed periodic flows for both the CWT and SHF conditions, and then provide an outline of the LBM algorithm we use in this study. Such information is well documented in the literature; and we re-present these materials here for the completeness of this paper and

 ISBN: 978-84-617-8694-7

for the convenience of the following discussions of our new periodic boundary treatments.

2.1 Fully Developed Periodic Thermal Flows

The flow and temperature fields are governed by the following continuity (Eq. 1), momentum (Eq. 2), and energy (Eq. 3) equations:

$$\frac{\partial \rho}{\partial t} + \rho \nabla \cdot \mathbf{u} = 0 \quad , \tag{1}$$

$$\frac{\partial \mathbf{u}}{\partial t} + \nabla \cdot (\mathbf{u}\mathbf{u}) = -\frac{\nabla P}{\rho} + \nu \nabla^2 \mathbf{u} \quad , \tag{2}$$

$$\frac{\partial T}{\partial t} + \mathbf{u} \cdot \nabla T = \alpha \nabla^2 T \quad , \tag{3}$$

where $\mathbf{u}$ is the flow velocity, P is the pressure, T is the temperature, ρ is the fluid density, ν is the kinematic fluid viscosity, α is the fluid thermal diffusivity, and t is time. Here we have neglected the viscous dissipation term in the energy equation as in typically heat transfer research. When the flow is fully developed along the periodic passage, the velocity $\mathbf{u}$ becomes identical at locations of the same relative position in each periodic module (we will call them image locations hereafter); however, the pressure decreases for a certain amount over a module [13, 15]. Patankar et al. [13] then split the fluid pressure P into two components $P(x, y) = -\frac{\Delta P_L}{L} x + \widetilde{P}(x, y)$, and the momentum equation Eq. (2) is rewritten to

$$\frac{\partial \mathbf{u}}{\partial t} + \nabla \cdot (\mathbf{u}\mathbf{u}) = -\frac{\nabla \widetilde{P}}{\rho} + \nu \nabla^2 \mathbf{u} + \frac{\Delta P_L}{\rho L} \quad . \tag{4}$$

Now the flow system can be solved with an appropriate numerical method under a perfect periodic boundary condition across a module.

As for the temperature field, the periodic features depend on the boundary conditions imposed on the walls. For fully developed periodic thermal flows to be established, the solid surfaces must have a uniform, constant wall temperature (CWT) T_w, or they can have specified surface heat flux (SHF). For the latter situation, the heat flux could be uniform or varying over the surface within one module, but it must have the same distribution for all units. For the CWT systems, usually we first shift the temperature field by the wall temperature T_w to a reduced temperature $\theta = T - T_w$, and the energy equation Eq. (3) becomes

$$\frac{\partial \theta}{\partial t} + \mathbf{u} \cdot \nabla \theta = \alpha \nabla^2 \theta \quad ; \tag{5}$$

with the wall boundary condition for θ as $\theta(\Omega) = 0$ (here Ω denotes the wall surface). The periodic relationship for θ among modules is expressed as

$$\theta(x \pm mL, y) = e^{-\lambda_L (\pm mL)} \theta(x, y) \quad , \tag{6}$$

 ISBN: 978-84-617-8694-7

where λ_L is the decaying rate that describes the overall temperature variation in the streamwise direction [15]. Stalio and Piller [15] then introduced a normalized temperature $\bar{\theta} = \theta / e^{-\lambda_L x}$; and the energy equation for $\bar{\theta}$ becomes

$$\frac{\partial \bar{\theta}}{\partial t} + \mathbf{u} \cdot \nabla \bar{\theta} = \alpha \nabla^2 \bar{\theta} + (\alpha \lambda_L^2 + \lambda_L u_x)\bar{\theta} - 2\alpha \lambda_L \frac{\partial \bar{\theta}}{\partial x} \tag{7}$$

with an exact periodic boundary condition. In Eq. (7), u_x represents the x component of the flow velocity vector $\mathbf{u}$. The decaying rate λ_L can be determined from the energy balance as

$$\lambda_L = -\ln \left[1 - \alpha \int_\Omega \left(\frac{\partial \theta}{\partial n} \right) ds \Big/ \int_{in} \left(u_x \theta + \alpha \frac{\partial \theta}{\partial x} \right)_{in} dy \right] / L \quad . \tag{8}$$

Compared to the calculation method in Ref. [15], our method does not require volumetric integration over the entire simulation domain, which could be computational expensive especially in three-dimensional simulations.

For SHF systems, the periodic relationship for temperature is given as

$$T(x \pm mL, y) = T(x, y) \pm m \Delta T_L \quad ; \tag{9}$$

Here ΔT_L is the temperature change over a periodic unit and it is constant along the flow. Accordingly, Patankar et al. [13] defined a reduced temperature $\widetilde{T}(x, y) = T(x, y) - \frac{\Delta T_L}{L} x$ to achieve a perfect periodic boundary condition for $\widetilde{T}$. The energy equation should be rewritten correspondingly to

$$\frac{\partial \widetilde{T}}{\partial t} + \mathbf{u} \cdot \nabla \widetilde{T} = \alpha \nabla^2 \widetilde{T} - \frac{u_x \Delta T_L}{L} \quad . \tag{10}$$

The temperature change ΔT_L can be relatively easily found from the energy conservation principle:

$$\Delta T_L = \frac{\int_\Omega q ds}{\rho c \int_{in} u_{x,in} dy} \quad , \tag{11}$$

i.e., the temperature change equals the total thermal energy addition via the surface divided by the product of flow rate and volumetric heat capacity (ρc). Here q is the local heat flux entering the fluid domain via the boundary walls.

2.2 Double Distribution Thermal LBM Model

Next we describe the double-distribution lattice Bhatnagar-Gross-Krook (LBGK) model for thermal flows [8], although other LBM models are available in the literature [16, 7]. Here two sets of density distribution functions are employed: one as f_i for the fluid dynamics

 ISBN: 978-84-617-8694-7

and one as h_i for the thermal convection-diffusion equation. The collision step for these distributions can be expressed mathematically as

$$f_i^*(\mathbf{x},t) = f_i(\mathbf{x},t) - \tfrac{1}{\tau_f}\left[f_i(\mathbf{x},t) - f_i^{eq}(\mathbf{x},t)\right] + \delta f_i \quad , \tag{12}$$

$$h_i^*(\mathbf{x},t) = h_i(\mathbf{x},t) - \tfrac{1}{\tau_h}\left[h_i(\mathbf{x},t) - h_i^{eq}(\mathbf{x},t)\right] + \delta h_i \quad . \tag{13}$$

The relaxation parameters τ_f and τ_h are related to the respective transport coefficients and the additional terms δf_i and δh_i are used to recover the correct macroscopic equations. The fluid density ρ, equilibrium velocity $\mathbf{u}^{eq}$, and energy scalar A (could be the regular temperature T or its modified counterparts like $\widetilde{T}$, θ, or $\bar{\theta}$, depending on which energy equation to solve by distributions h_i) can be obtained from the density distributions f_i and h_i as

$$\rho = \sum_i f_i \ , \quad \mathbf{u}^{eq} = \sum_i f_i \mathbf{c}_i / \sum_i f_i \ , \quad A = \sum_i h_i \quad , \tag{14}$$

where $\mathbf{c}_i$ is the i-th lattice velocity. The equilibrium distributions f_i^{eq} and h_i^{eq} can then be calculated from these properties as [7, 4, 5]

$$f_i^{eq} = \omega_i \rho \left[1 + \frac{\mathbf{c}_i \cdot \mathbf{u}}{c_s^2} + \frac{(\mathbf{c}_i \cdot \mathbf{u})^2}{2c_s^4} - \frac{\mathbf{u}^2}{2c_s^2}\right] \quad , h_i^{eq} = A f_i^{eq}/\rho \quad . \tag{15}$$

The parameter ω_i is called the lattice weight factor and c_s is the lattice sound speed. After the collision, the post-collision distributions f_i^* and h_i^* will then move to the nearest neighboring lattice node at velocity $\mathbf{c}_i$ over a time step δt:

$$f_i(\mathbf{x} + \mathbf{c}_i \delta t, t + \delta t) = f_i^*(\mathbf{x},t) \quad ; h_i(\mathbf{x} + \mathbf{c}_i \delta t, t + \delta t) = h_i^*(\mathbf{x},t) \quad . \tag{16}$$

Appropriate mathematical analysis like the Chapman-Enskog expansion can be performed to the above distribution dynamics, and the following macroscopic equations can be derived [7, 16]:

$$\frac{\partial \rho}{\partial t} + \rho \nabla \cdot \mathbf{u} = 0 \quad , \tag{17}$$

$$\frac{\partial \mathbf{u}}{\partial t} + \nabla \cdot (\mathbf{u}\mathbf{u}) = -\frac{\nabla P}{\rho} + \nu \nabla^2 \mathbf{u} + \mathbf{F} \quad , \tag{18}$$

$$\frac{\partial A}{\partial t} + \mathbf{u} \cdot \nabla A = \alpha \nabla^2 A + S \quad . \tag{19}$$

The fluid properties $\mathbf{u}$, P, ν, and α are related to the LBM parameters by

$$\mathbf{u} = \mathbf{u}^{eq} + \frac{\mathbf{F}\delta t}{2\rho} \ , \quad P = c_s^2 \rho \ , \quad \nu = c_s^2 \left(\tau_f - \frac{1}{2}\right)\delta t \ , \quad \alpha = c_s^2 \left(\tau_h - \frac{1}{2}\right)\delta t \ . \tag{20}$$

The additional terms δf_i and δh_i in Eqs. (12) and (13) are related, respectively, to the forcing term $\mathbf{F}$ and source term S in the resulting macroscopic equations Eqs. (18) and (19) as

$$\delta f_i = \frac{\omega_i \mathbf{F} \cdot \mathbf{c}_i \delta t}{c_i^2} \ , \quad \delta h_i = \omega_i S \quad . \tag{21}$$

 ISBN: 978-84-617-8694-7

These δf_i and δh_i terms can be conveniently adjusted according to the forcing or source terms in the macroscopic equations to be solved. In our next validation and demonstration simulation examples, we use the simple D2Q9 (2D and $b = 9$) square lattice structure, for which the nine lattice velocities are

$$\mathbf{c}_0 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \quad \mathbf{c}_{1-4} = \begin{bmatrix} \cos(i-1)\pi/2 \\ \sin(i-1)\pi/2 \end{bmatrix} \frac{\delta x}{\delta t}, \quad \mathbf{c}_{5-8} = \sqrt{2} \begin{bmatrix} \cos(2i-9)\pi/4 \\ \sin(2i-9)\pi/4 \end{bmatrix} \frac{\delta x}{\delta t} \quad . \tag{22}$$

The lattice weight factors are $\omega_0 = 4/9$, $\omega_{1-4} = 1/9$, and $\omega_{5-8} = 1/36$; and the lattice sound speed $c_s = 1/\sqrt{3}\delta x/\delta t$. δx is the lattice grid resolution.

2.3 Simulating Periodic Thermal Flows with LBM

2.3.1 For Flow Field

Under a given pressure drop ΔP_L per periodic module, it is convenient to use the reduced pressure $\widetilde{P}$, and the LBM equations for f_i given in Section 2.2 can then be used $\mathbf{F} = (\Delta P_L/\rho L, 0)^T$. The classical periodic boundary condition [19] can then be applied at the periodic boundaries, meaning density distributions leaving the domain outlet will re-enter the domain at the inlet, or vice versa. This method has been widely used in LBM simulations, although more dedicate treatments are available to impose the pressure drop directly for some particular situations like multiphase or multicomponent flows [20]. To simulate a periodic flow with a specific flow rate, the pressure drop ΔP_L can be dynamically adjusted according to the simultaneous flow rate till the desirable value is established.

2.3.2 For Thermal Field: The Source Term (ST) Approach

Similarly, by tuning the source term δh_i according to Eq. (21), the LBM algorithm for h_i in Sect. 2.2 can be used to solve Eq. (7) for CWT systems

$$\delta h_i = \omega_i \left[(\alpha \lambda_L^2 + \lambda_L u_x)\bar{\theta} - 2\alpha\lambda_L \frac{\partial \bar{\theta}}{\partial x} \right] \tag{23}$$

for energy scalar $A = \bar{\theta}$; or to solve Eq. (10) for SHF cases with

$$\delta h_i = -\omega_i u_x \Delta T_L / L \tag{24}$$

for energy scalar $A = \widetilde{T}$. The temperature change ΔT_L for SHF cases can be readily calculated from the total heat flux over surface via Eq. (11); however, for CWT systems, the decaying rate λ_L is unknown before the simulation. In our practice, we start with an initial guess and run the simulation for some time (2000 time steps in our simulations) with that initial value. After that, a new λ_L value is calculated via Eq. (8) every certain time steps (we use 20 time steps), till the simulation becomes steady in flow and temperature fields. The differential term $\partial\bar{\theta}/\partial x$ in Eq. (7) can be estimated by a finite difference approximation.

 ISBN: 978-84-617-8694-7

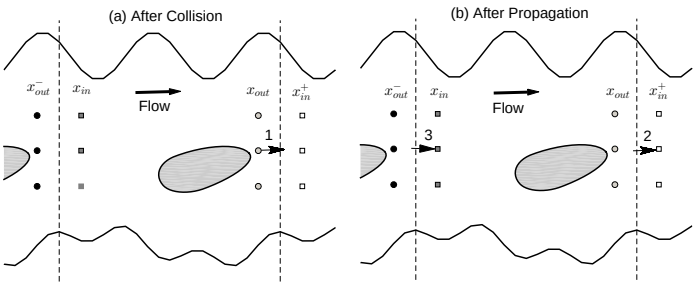


Figure 1: Schematic illustrations of the modified periodic boundary treatment for thermal field.

2.3.3 For Thermal Field: The Distribution Modification (DM) Approach

Another way to incorporate the periodic features of the temperature field described in Sect. 2.1 is to modify those density distributions h_i that cross the module inlet or outlet boundaries, as done in Ref. [20] for pressure periodic boundary conditions. We consider the CWT situation the periodic system shown in Fig. 1 with the first column of lattice nodes at x_{in} (dark gray squares) and the last column at x_{out} (light gray circles). Also displayed there are the outlet nodes of the upstream module at $x_{out}^- = x_{in} - \delta x = x_{out} - L$ (black circles) and the inlet nodes of the downstream module at $x_{in}^+ = x_{in} + L = x_{out} + \delta x$ as (white squares); although these nodes are actually not involved in the LBM calculation. To avoid the extra source term in Eq. (7), we will use LBM to solve the energy equation Eq. (5) for θ. Now let us take the post-collision distribution $h_1^*(x_{out})$ at x_{out} (Arrow 1 in Fig. 1a) as an example. In the propagation step, $h_1^*(x_{out})$ is supposed to move to the next node in velocity $\mathbf{c}_1 = (1, 0)^T$, and becomes the incoming distribution at the inlet node of the downstream module x_{in}^+ (Arrow 2 in Fig. 1b). However, now it is out of our simulation domain and therefore cannot participate in the LBM calculation anymore. On the other side, we need the incoming distribution $h_1(x_{in})$ at the domain inlet x_{in} (Arrow 3 in Fig. 1b), but it is not available since the nodes at the x_{out}^- are not in the simulation domain either. Based on the periodic relationship of temperature given in Eq. (6) and $\theta = \sum_i h_i$ (Eq. 14 with $A = \theta$), it is reasonable to assume the proportionality in θ can be extended to each distribution h_i, and therefore one has

$$h_1(x_{in}, t + \delta t) = e^{\lambda_L L} h_1(x_{in}^+, t + \delta t) = e^{\lambda_L L} h_1^*(x_{out}, t) \quad . \tag{25}$$

This analysis can be applied to other lattice distributions that cross the periodic boundaries during the propagation step, and a modified periodic boundary treatment for these distributions can be established as (for the D2Q9 lattice model used here)

$$h_i(x_{in}, y_{in}, t + \delta t) = e^{\lambda_L L} h_i^*(x_{out}, y_{out}, t) \quad , \quad y_{in} = y_{out} + c_{i,y}\delta t, \quad i = 1, 5, 8 \quad ; \tag{26}$$

 ISBN: 978-84-617-8694-7

$$h_i(x_{out}, y_{out}, t+\delta t) = e^{-\lambda_L L} h_i^*(x_{in}, y_{in}, t) \quad , \quad y_{out} = y_{in} + c_{i,y}\delta t, \quad i = 3, 6, 7 \quad . \tag{27}$$

Here $c_{i,y}$ is the $y-$component of the lattice velocity $\mathbf{c}_i$.

This distribution modification (DM) approach is also applicable to the SFH cases. Here we work with the original energy equation Eq. (3) and re-write the periodic relationship for the regular temperature T Eq. (9) to define a proportional factor β as

$$\beta = \frac{T(x_{in}^+)}{T(x_{in})} = 1 + \frac{\Delta T_L}{T(x_{in})} \quad ; \tag{28}$$

and, following the above discussion, the modified periodic boundary condition for h_i in SHF systems is

$$h_i(x_{in}, y_{in}, t+\delta t) = \beta^{-1} h_i^*(x_{out}, y_{out}, t) \quad , \quad y_{in} = y_{out} + c_{i,y}\delta t, \quad i = 1, 5, 8 \quad ; \tag{29}$$

$$h_i(x_{out}, y_{out}, t+\delta t) = \beta h_i^*(x_{in}, y_{in}, t) \quad , \quad y_{out} = y_{in} + c_{i,y}\delta t, \quad i = 3, 6, 7 \quad . \tag{30}$$

Note these modified periodic treatments revert back to the classical periodic boundary condition in LBM when the proportional factor $e^{-\lambda_L}$ or β is set to 1.

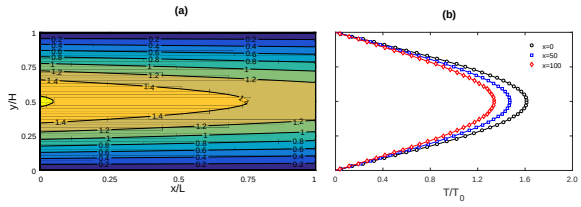


Figure 2: The simulated temperature field (a) and transverse profiles (b) for flow through the flat channel with CWT condition on the channel surfaces. In (b) the symbols are our LBM results and the underlying curves are from the analytical solution by Brown [3].

3 Validation and Demonstration Simulations

3.1 Heat Transfer of Laminar Flow through 2D Flat Channel

The fully developed flow in a uniform pipe or channel can be considered as an extreme example of periodic flows, for which the periodic module can be selected as a segment of the channel of any length. Here our simulation domain is a 2D rectangle of length $L = 200$

 ISBN: 978-84-617-8694-7

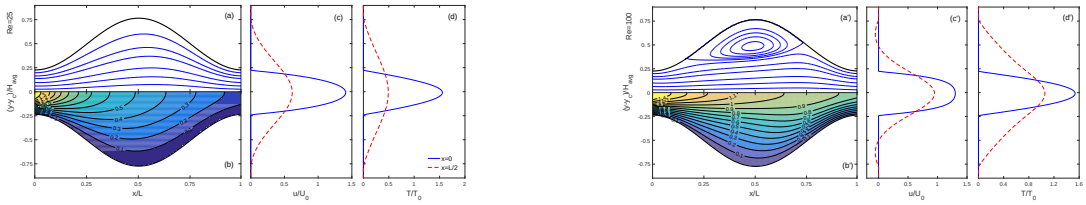


Figure 3: The simulated flow (a and a′) and temperature (b and b′) fields and transvers profiles of the streamwise velocity (c and c′) and temperature (d and d′) at two locations $x = 0$ and $x = L/2$ for the flows through a wavy channel with Reynolds number $Re = 25$ (a-d) and $Re = 100$ (a′-d′).

and height $H = 50$ (Both L and H, as well as other LBM parameters to be given in this section, are all non-dimensional values). The midpoint boundary method recently developed by Zhang and co-workers [18, 4, 5] is implemented at the solid-fluid boundaries for both flow and thermal LBM calculations for all simulations in this paper. The Reynolds number $Re = U_0 H/\nu$ is 40 (U_0 is the mean flow velocity) and the Prandlt number $Pr = \nu/\alpha$ is 0.7. The CWT situation is considered here and wall temperature T_w is set as 0; thus the regular temperature T and the reduced temperature θ are the same. During the simulation, the mean flow temperature at the domain inlet $T_0 = T_m(x = 0)$ is maintained at 1. This is the so-called Graetz problem, and the analytical solution can be expressed in series [3]. Fig. 2 shows the comparison between our LBM results using the DM periodic boundary method and those from Brown′s analytical solution [3] for the temperature field. No visible discrepancy can be observed. According to the analytical solution [3], the Nusselt number along the channel is constant at 3.7704; while our LBM yields an indeed constant value of 3.7757. The relative difference is only 0.14%, and we attribute it to the second-order finite difference approximation we use to evaluate the temperature gradient on surfaces.

3.2 Heat Transfer of Laminar Flow through 2D Wavy Channel

The flow and heat transfer through wavy channels have been extensively investigated for its practical applications [15, 17, 11, 2, 14]. As in the flat channel simulation, we have T_w=0, T_0=1, and Pr=0.7; and the DM method is used to incorporate the boundary periodicity. Two values, $Re = 25$ and 100, are tested in our simulations. Fig. 3 collects our LBM results of these two calculations, including the flow streamlines, the isotherms, and the spanwise profiles of streamwise velocity u_x and temperature T at the maximum and minimum width locations. The streamline and isotherm patterns are very similar to those reported in previous studies [2, 14]; however, a direct comparison is difficult due to the lack of original data for those publications. At $Re = 100$, a pair of circulation vortex have developed in the

 ISBN: 978-84-617-8694-7

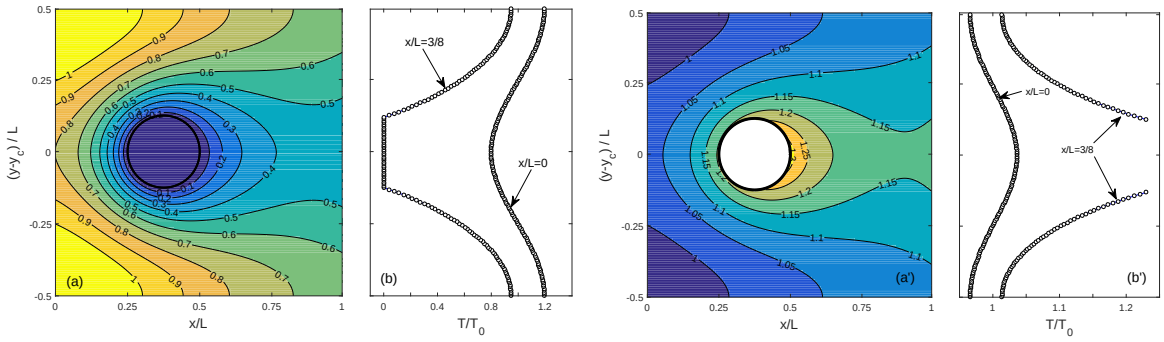


Figure 4: Result comparison of the simulations for the thermal flows around cylinder using the DM or ST approaches for the CWT (a and b) and SHF (a′ and b′) wall conditions. In the temperature fields (a and a′), the color patches are from the DM approach, and the isotherm lines are from the ST approach. The temperature profiles at $x/L = 0$ and $3/8$ (indicated by labels) are displayed in (b) and (b′), with the symbols from the DM approach and curves from the ST approach.

wide section, and the separation and reattachment locations are similar to those in Ref. [2].

3.3 Heat Transfer in Flow through Square Cylinder Array

The last system we simulate represents the heat transfer process associated to laminar flow through a square array of circular cylinders, which resembles the interior configuration of a cross-flow tube heat exchanger [9, 10]. The periodic module here is a square with $L = H = 160$, and the cylinder has a diameter of $D = 40$ and its center locates at $(3L/8, L/2)$. The flow direction is from the left to the right in the x-direction, and the regular periodic boundary condition is applied along the top and bottom edges. The Reynolds number $Re = U_0 H/\nu = 2.4$ and the Prandlt number $Pr = 1$. Four individual simulations are conducted: CWT+DM (the same one-module simulation above), CWT+ST, SHF+DM, and SHF+ST. To impose a desirable temperature gradient on the surface, the Neumann boundary method developed by Oulaid and Zhang [12] is employed. With the inlet mean temperature $T_0 = 1$ and wall temperature $T_w = 0$, the CWT case represents a cooling process. On the other hand, for SHF case, we use a uniform surface flux with $\partial T/\partial n = -0.01$ for the cylinder surface and it therefore is a heating process. Results from these simulations, including the temperature field in the domain and two representative transverse temperature profiles, are collected in Fig. 4. Figs. 4 a and a′ clearly show the cooling or heating effect from the cylinder on the fluid. It is interesting to see that, in Fig. 4a′ for the SHF boundary situation, the temperature increases near the outlet. This is understandable since the outlet

 ISBN: 978-84-617-8694-7

is close to the heating source (the cylinder) in the next module, and for this diffusion-dominant system (Peclet number $Pe = RePr = 2.4$), the heating flux from the cylinder can reach a relatively long distance even against the flow direction. As for the results from the DM or ST approaches for both CWT and SHF boundary conditions, one can see again excellent agreement exists in Fig. 4, and no apparent difference can be spotted. The decaying rate λ_L value is 3.88554×10^{-3} from the DM approach, and 3.88533×10^{-3} from the ST approach. Such a nearly perfect match indicates that both the DM and ST approaches can produce reliable results for simulating developed periodic thermal flows.

4 Summary and Concluding Remarks

We have examined the periodic relationships in flow and temperature fields for fully developed periodic thermal flows with CWT and SHF boundary conditions, and proposed two LBM implementations (the ST and DM approaches) for such flow situations. The methods have then been tested thoroughly in several simulations by comparing our LBM results to those from analytical solutions, previous publications, and our own LBM simulations using different numerical techniques. The good performance suggests that our methods could be useful for future LBM thermal simulations.

For the two numerical schemes to implement the periodic features of temperature in LBM, the ST approach has been typically used in traditional CFD studies, and certainly can also be adopted in LBM. On the other hand, the DM approach is unique for LBM with some computational advantages. In the DM method, extra calculations are only required for the thermal distributions crossing the periodic inlet/outlet boundaries; but in the ST method, an extra term has to be calculated for all distributions and at all lattice nodes. Furthermore, for systems with CWT boundaries, the ST approach also needs to calculate the streamwise derivative of temperature ($\partial\bar{\theta}/\partial x$ in Eq. 7), and this could further increase the computational demand. The method in Eq. (8) to calculate the decaying rate λ_L does not requires a volumetric integration of temperature over the entire computational domain, and thus it could also be useful for improving the computational efficiency of LBM as well as other CFD methods.

Acknowledgment

This work was supported by the Natural Science and Engineering Research Council of Canada (NSERC). The calculations have been enabled by the use of computing resources provided by WestGrid (westgrid.ca), SHARCNet (sharcnet.ca), and Compute/Calcul Canada (computecanada.org).

 ISBN: 978-84-617-8694-7

References

[1] T. Adachi and H. Uehara. *Int. J. Heat Mass Transfer*, 44:4333, 2001.

[2] H. M. S. Bahaidarah, N. K. Anand, and H. C. Chen. *Num. Heat Transfer A*, 47:417, 2005.

[3] G. M. Brown. *AIChE J.*, 6:179, 1960.

[4] Q. Chen, X. Zhang, and J. Zhang. *Phys. Rev. E*, 88:033304, 2013.

[5] Q. Chen, X. Zhang, and J. Zhang. *Commun. Comput. Phys.*, 17:937–959, 2015.

[6] M. Greiner, R. J. Faulkner, V. T. Van, H. M. Tufo, and P. F. Fischer. *ASME J. Heat Transfer*, 122:653, 2000.

[7] Z. Guo and C. Shu. *Lattice Boltzmann Method and Its Applications in Engineering.* World Scientific Publishing, Singapore, 2013.

[8] X. He, S. Chen, and G. D. Doolen. *J. Comput. Phys.*, 146:282, 1998.

[9] J. P. Holman. *Heat Transfer.* McGraw-Hill, New York, 1968.

[10] F. P. Incropera, D. P. DeWitt, T. L. Bergman, and A. S. Lavine. *Fundamentals of Heat and Mass Transfer.* John Wiley & Sons, New York, 6th ed. edition, 2006.

[11] B. Niceno and E. Nobile. *Int. J. Heat Fluid Flow*, 22:156, 2001.

[12] O. Oulaid, Q. Chen, and J. Zhang. *J. Phys. A*, 46:475501, 2013.

[13] S. V. Patankar, C. H. Liu, and E. M. Sparrow. *ASME J. Heat Transfer*, 99:180, 1977.

[14] A. G. Ramgadia and A. K. Saha. *Int. J. Therm. Sci.*, 67:152, 2013.

[15] E. Stalio and M. Piller. *ASME J Heat Transfer*, 129:769, 2007.

[16] S. Succi. *The Lattice Boltzmann Equation.* Oxford Univ. Press, Oxford, 2001.

[17] G. Wang and S. P. Vanka. *Int. J. Heat Mass Transfer*, 38:3219, 1995.

[18] X. Yin and J. Zhang. *J. Comput. Phys.*, 231:4295–4303, 2012.

[19] J. Zhang. *Microfluid. Nanofluid.*, 10:1–28, 2011.

[20] J. Zhang and D. Y. Kwok. *Phys. Rev. E*, 73:047702, 2006.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

Mucus Velocity in Human Lungs

Kanognudge Wuttanachamsri[1]

[1] *Department of Mathematics, King Mongkut's Institute of Technology Ladkrabang, Bangkok 10520, Thailand*

emails: whychamsri@hotmail.com

Abstract

With each breath, the human body breathes in air but also particles, bacteria and viruses. Particles penetrating into the system are trapped on a layer of mucus and propelled by tiny hairlike structure, cilia, that line the airways. In this paper, we calculate the average velocity of mucus residing above the cilia, which is mainly expelled by the cilia movement. For slow flow problem, we apply Stokes equation to find the mucus velocity where the bottom boundary condition of fluid velocity is obtained from the porous layer in which the fluid flows by the movement of cilia. A mixed finite element method is used to find the average mucus velocity in a three-dimensional domain. The numerical result is compared with that calculated by D. J. Smith et al. [Bull. Math. Biol., 69: 289-327, 2007] (38.3 $\mu m/s$) with a good agreement.

Key words: Mixed finite element method, Stokes equations, Cilia, Lung, Mucus, Three-dimensional domain

1 Introduction

Fluid transport is one of the most important phenomena in many biological problems. One example of the fluid transport problems is the transport of mucus in the human respiratory system. In pulmonary pathways, the epithelium surface is coated by a layer of mucus, which is considered to be either a highly viscous or viscoelastic fluid. The mucous layer traps, transports, or sometimes chemically disarms undesirable chemical or biological agents and natural debris, and so to avoid contamination or infection, mucus must be continuously expelled from the pulmonary pathway. Such an expelling force consists of two mechanisms: the coupling between liquid and airflow (e.g. breathing/coughing) and the force

 ISBN: 978-84-617-8694-7

from metachronal waves generated by the coherent beating of cilia - small hairlike structures. The primary transport mechanism responsible for the removal of mucus and cellular debris from the lungs is the second mechanism, the so-called muco-ciliary clearance. Due to the increasing patients with lung disease having problems with the transport of mucus in airways, researchers in biological fields are interested to find the mucus velocity in the bronchus and bronchioles. In this work, we focus on the primary mechanism and calculate the mucus velocity due to the movement of cilia.

Studying about mucus had been presented in various aspects in literatures. For example, Cees P van der Schans [10] experimentally studied mucus transport in airways. He concluded that the muco-ciliary transport was mostly at the peripheral airways and airflow transport was at the central airways. For airways disease, airflow transport was an important alternative to muco-ciliary clearance. The laboratory observations of Matsui et al. [8] indicated that the PCL and the mucous layers seemed to move in unison, while with the removal of the mucous layer, fluid transport in the PCL layer alone was drastically reduced. Hironori [6] provided that the average velocity above the cilia tips was a parabolic profile with the maximum velocity at the tips of cilia. H. Matsui et al. [7] studied on the pathogenesis of cystic fibrosis (CF) airways infection and suggested a treatment method by using salt and water instead of the modulation of ionic composition. The mucus velocities for normal and CF cases were also provided. The mucus velocity for normals was $26 \pm 5\mu m/s$ and that for CF was $1.2 \pm 0.2\mu m/s$ for 24 hr after addition of 50 μl of phosphate-buffered saline (PBS) containing tracers. The mucus velocity depended on the tracers and time. For different tracers and time, the mucus velocity values were various. S. K. Lai [4] reviewed the macro- and microrheology of human mucus and provided that for macroscale mucus behaved as a non-Newtonian fluid what it acted as a low viscosity liquid at the nanoscale.

Figure 1 shows the muco-ciliary system consisting of three layers. The middle layer, in which the cilia beat, is known as the periciliary layer (PCL) where the fluid in this layer is called PCL fluid, and the upper layer consists of mucus. Mucus is secreted by goblet cells in the lower layer that are interspersed among the ciliated cells to trap particles getting into the body. To model the slow flow problem, we employ Stokes equation coupling with the incompressible flow equation to calculate the mucus velocity using a mixed finite element method.

Literatures mathematically calculated the mucus velocity by using various models and methods. For instant, W. Hofmann and B. Asgharian [3] used the asymmetric multiple-path models of the bronchial tree to calculated the mucociliary clearance velocities in bronchial airway generations. They found that in human tracheal the mucus velocity was 5.5 mm/min and decreased in roughly exponential expression with increasing airway generation number. W. L. Lee et al. [5] simulated the muco-ciliary trannsport process using the projection method combined with the immersed boundary method (IBM) to describe the ciliary beating patterns. The mean mucus velocity is plotted with the viscosity of the mucus, the number

 ISBN: 978-84-617-8694-7

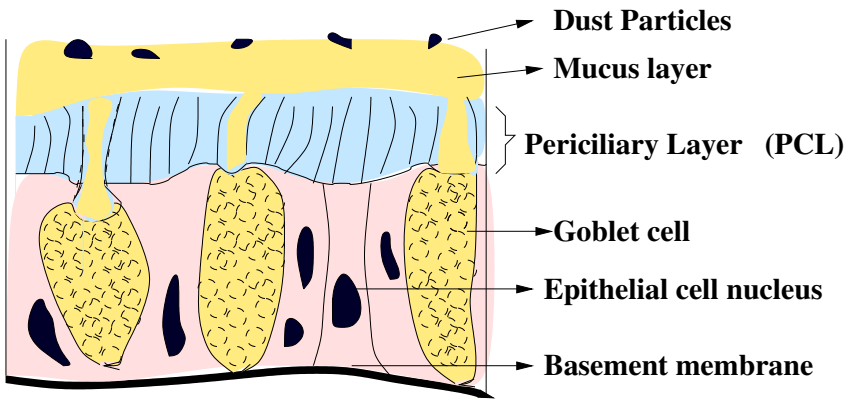


Figure 1: Schematic illustration the muco-ciliary system consisting of threee layers: epithelial layer, PCL and mucus layer.

of cilia, the depth of periciliary layer and the cilia beat frequency. D. J. Smith et al. [9] presented a new mathematical model, deriving from Navier-Stokes momentum equations, of the mucus transport in airways by cilia. They calculated the mean mucus speed, which is 38.3 $\mu m/s$. However, the literatures study only the case that the solid phases in PCL have no movement. In this article, we find the mucus velocity from the self-propelled of cilia in PCL. That is the boundary condition at the bottom of our numerical domain, mucus layer, is obtained from the PCL in which the PCL fluid is calculated from movement of cilia as provided in [2].

To calculate the numerical results, we employ the Stokes equations derived in [2] which will be briefly rewritten in Section 1 including the discretization of the model using a mixed finite element method written in [1] which will be rewritten here. In Section 3 we provide the boundary conditions obtained from [2] for coding where the numerical validation and results are presented in Section 4. The conclusion is drawn in Section 5.

2 Stokes Model and Its Discretization

In this section, we provide Stokes equation with the incompressible continuity equation and also its discretization using a mixed finite element method. The system of equation which is derived in [2] is as follows.

$$\nabla p - \mu \Delta \mathbf{v} = \rho \mathbf{g} \tag{1}$$

$$\nabla \cdot \mathbf{v} = 0, \tag{2}$$

 ISBN: 978-84-617-8694-7

where $\mathbf{v}$ is the velocity of the liquid phase; μ is the dynamic viscosity; p is pressure; $\mathbf{g}$ is gravity and ρ is fluid density.

Next, we transfer the Stokes model into discrete counterparts for making them suitable for numerical implementation. Although the model discretization has been provided in [1], they form an important part of our study. Then we briefly rewrite them here. Let Ω be our computational domain, T_h be a triangulation of domain Ω and

$$V_h = \{v \in H^1(\Omega) : v|_K \text{ is quadratic}, \forall K \in T_h\} \tag{3}$$

$$H_h = \{q \in L_0^2(\Omega) : q|_K \text{ is linear}, \forall K \in T_h\}, \tag{4}$$

be finite-dimensional subspaces of the Sobolev spaces $H^1(\Omega)$ and $L_0^2(\Omega)$, respectively, where

$$L_0^2(\Omega) = \{q \in L^2(\Omega) : \int_\Omega q d\Omega = 0\}. \tag{5}$$

By using a mixed finite element method provided in [1], the system of equations (1) - (2) can be written in the matrix form as

$$\begin{pmatrix} \tilde{\mathbf{B}} & 0 & 0 & -\tilde{\mathbf{Q}}_1^T \\ 0 & \tilde{\mathbf{B}} & 0 & -\tilde{\mathbf{Q}}_2^T \\ 0 & 0 & \tilde{\mathbf{B}} & -\tilde{\mathbf{Q}}_3^T \\ -\tilde{\mathbf{Q}}_1 & -\tilde{\mathbf{Q}}_2 & -\tilde{\mathbf{Q}}_2 & 0 \end{pmatrix} \begin{pmatrix} \mathbf{V}_1 \\ \mathbf{V}_2 \\ \mathbf{V}_3 \\ \mathbf{P} \end{pmatrix} = \begin{pmatrix} \tilde{\mathbf{B}}_1 \\ \tilde{\mathbf{B}}_2 \\ \tilde{\mathbf{B}}_3 \\ \mathbf{0} \end{pmatrix}, \tag{6}$$

where the approximate solutions $(v_i, p) \in V_h \times H_h$ are

$$v_i(\mathbf{x}) = \sum_{m=1}^{M} \psi_m(\mathbf{x}) v_i^m = \Psi^T \mathbf{V}_i, \tag{7}$$

$$p(\mathbf{x}) = \sum_{l=1}^{L} \phi_l(\mathbf{x}) p_l = \Phi^T \mathbf{P}. \tag{8}$$

where $\mathbf{V}_i$ and $\mathbf{P}$ are vectors of the velocities and pressure, respectively; ψ_m and ϕ_l are basis functions; Ψ and Φ are their vector forms and the integers M and L are determined by the interpolation function and

$$\tilde{\mathbf{B}} = (\mu/\epsilon^l)(\tilde{\mathbf{K}_{jj}}), \quad \tilde{\mathbf{K}_{ij}} = \int_{\Omega^e} \frac{\partial \Psi}{\partial x_i} \frac{\partial \Psi^T}{\partial x_j} d\Omega^e, \quad \tilde{\mathbf{Q}}_i = \int_{\Omega^e} \Phi \frac{\partial \Psi^T}{\partial x_i} d\Omega^e, \tag{9}$$

$$\tilde{\mathbf{B}}_i = \left(-\int_{\Gamma^e} \Psi \Phi^T n_i d\Gamma^e \right) \mathbf{P} + \frac{\mu}{\epsilon^l} \left(\int_{\Gamma^e} \Psi \frac{\partial \Psi^T}{\partial x_j} n_j d\Gamma^e \right) \mathbf{V}_i, i = 1, 2, 3. \tag{10}$$

The notation Ω^e indicates the element domain such that $\Omega = \bigcup_e \Omega^e$ and the repeated index j is the summation over $j, j = 1, 2, 3$.

 ISBN: 978-84-617-8694-7

3 Boundary conditions

In this section, we provide boundary conditions used to model the problem. First, we inspect the condition at the bottom of domain Ω. Since in this work we assume that the mucus flow depends on the cilia movement, we then begin by considering the PCL layer where the PCL fluid is moved by the ciliary beating provided in [2]. Therefore, the velocity of the PCL fluid at the tip of cilia becomes our boundary condition at the bottom of our numerical domain. Because K. Chamsri and L. Schreyer [2] assume that the cilia are an array of cylinders and the cilia move by making angles θ with the horizontal plane, to calculate the mucus velocity, they provide the average velocity over all angles at the top of PCL domain. Figure 2 shows the the boundary condition of the mean velocity at the bottom of our domain. The top one is the first component of the velocity and the middle and the bottom ones are the second and third components, respectively.

By employing the boundary conditions from [2], we now have the boundary condition at the bottom of our numerical domain. For other boundary conditions, they are written with the system of equations as follows.

$$\nabla p - \mu \Delta \mathbf{v} = \rho \mathbf{g} \qquad \text{in } \Omega \tag{11}$$

$$\nabla \cdot \mathbf{v} = 0, \qquad \text{in } \Omega \tag{12}$$

$$\mathbf{v} \text{ is employed from [2]} \qquad \text{at the bottom of } \Omega \tag{13}$$

$$\mathbf{v} \text{ and } p \text{ are periodic} \qquad \text{on the left and right sides.} \tag{14}$$

4 Numerical Validation and Results

The numerical solutions of the model and boundary conditions provided in Section 1 and 3, respectively, are presented in this section. By using the model discretization of Taylor-Hood type written in Section 1, for the case of average velocity of PCL at the tip of cilia or the bottom of mucus layer, the velocity of mucus presented in Figure 3 with 15,468 number of degree of freedom with CPU time about 1 month. The average mucous velocity is 31.67 $\mu m/s$. The result can be verified by comparing it with Matsui et al. [8] and Smith et al.[9]. Matsui et al. [8] experimentally shows a mean mucus transport of $39.8 \pm 4.2 \mu m/s$ and Smith et al. [9] presents a mean mucus velocity of 38.3 $\mu m/s$. The discrepancy between the solutions may be from the different assumptions and techniques used in the works. For instant, Smith et al. [9] considered the mucus layer and mucus-PCL interface as linear viscoelastic and completely flat, respectively by using a new mathematical model while Matsui et al. [8] assumed that the PCL is nearly stationary with human tracheobronchial cell cultures. Comparing our solution with them, it shows a reasonable agreement with our predicted value.

 ISBN: 978-84-617-8694-7

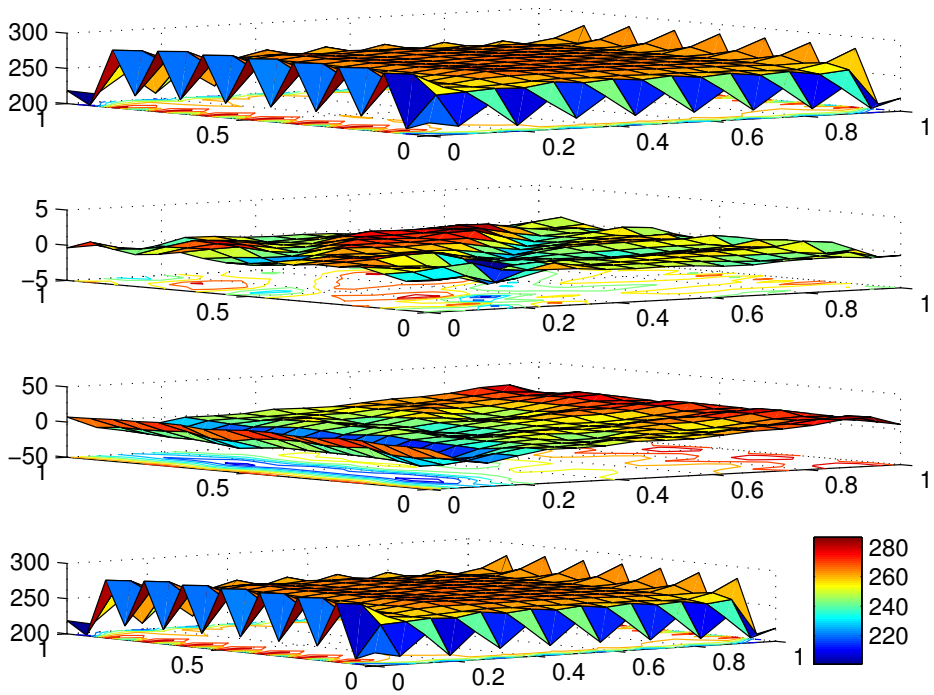


Figure 2: Boundary condition of the velocity at the bottom of our numerical domain in case of average velocity of PCL over all angles.

5 Conclusion

We calculate mucus velocity due to the movement of cilia consisting in the PCL. For the slow flow problem, we use Stokes equation on the domain of mucus layer. The boundary condition at the bottom of our numerical domain is employed from the velocity of PCL fluid at the top of PCL provided in [2]. The velocity is obtained from averaging the velocity of the PCL fluid over angles θ that cilia make with the horizontal plan. At the sides of our cube domain, periodic boundary conditions are applied and a mixed finite element method is exploited to determine the velocity profile. The average result over the z axis of the domain are compared with Matsui et al. [8] and Smith et al.[9]. The result is in good agreement.

 ISBN: 978-84-617-8694-7

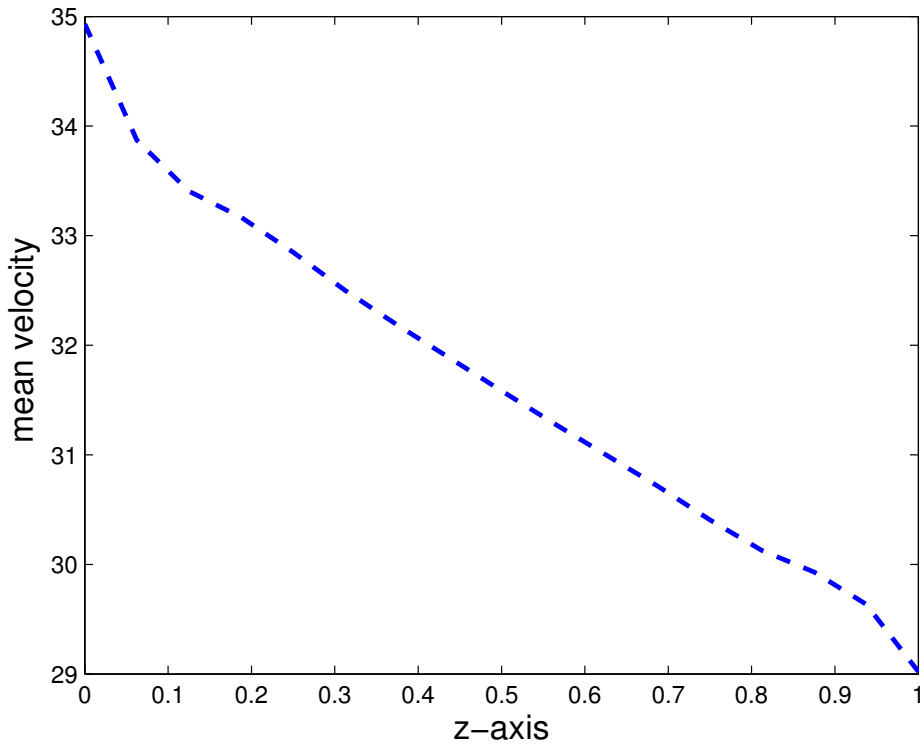


Figure 3: Mucus velocity when the bottom boundary condition is the average velocity of PCL

Acknowledgements

This work has been supported by a grant from KMITL Research Fund (KMITL Fund).

References

[1] K. Chamsri. N-dimensional stokes-brinkman equations using a mixed finite element method. *Australian Journal of Basic and Applied Sciences*, 8:30–36, Special 2014.

[2] K. Chamsri and L. Schreyer. Effect of the cilia movement on fluid. To be published, 2016.

[3] W. Hofmann and B. Asgharian. Comparison of Mucociliary Clearance Velocities in Human and Rat Lungs for Extrapolation Modeling. *Annals of Occupational Hygiene*, 46, Supplement 1:323–325, 2002.

[4] S. K. Lai, Ying-Ying Wang, D. Wirtz, and J. Hanes. Micro- and macrorheology of mucus. *Advanced Drug Delivery Reviews*, 61(2):86–100, 2010.

[5] W. L. Lee, P. G. Jayathilake, Z. Tan, Le D. V., H. P. Lee, and B. C. Khoo. Muco-Ciliary Transport: Effect of Mucus Viscosity, Cilia Beat Frequency and Cilia Density. *Computer & Fluids*, 49:214–221, 2011.

[6] R. Lima, Y. Imai, T. Ishikawa, and V.C. Cano. *Visualization and Simulation of Complex Flows in Biomedical Engineering*. Springer, 2014.

[7] H. Matsui, B.R. Grubb, R. Tarran, S.H. Randell, J.T. Gatzy, C.W. Davis, and R.C. Boucher. Evidence for Periciliary Liquid Layer Depletion, Not Abnormal Ion Composition, in the Pathogenesis of Cystic Fibrosis Airways Disease. *Cell*, 95:1005–1015, 1998.

[8] H. Matsui, S. H. Randell, S. W. Peretti, C. W. Davis, and R. C. Boucher. Coordinated Clearance of Periciliary Liquid and Mucus from Airway Surfaces. *The Journal of Clinical Investigation*, 102:1125–1131, 1998.

[9] D. J. Smith, E. A. Gaffney, and J. R. Blake. A Viscoelastic Traction Layer Model of Muco-Ciliary Transport. *Bulletin of Mathematical Biology*, 69:289–327, 2007.

[10] C. P. Van Der Schans. Bronchial Mucus Transport. *Respiratory Care*, pages 1150–1158, 2007.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

Variational Structure of a Class of Fractional Hamiltonian Systems and Its Applications

Yongzhen Yun[1,2], You-Hui Su[1] and Dongdong Wang[1]

[1] *School of Mathematics and Physics, Xuzhou University of Technology*

[2] *School of Mathematics and Statistic, Yili Normal University*

emails: yongzhen0614@163.com, suyh02@163.com, wdd@xzit.edu.cn

Abstract

In the paper, we are concerned with the existence criteria of the solution to a class of Riemann-Liouville fractional Hamiltonian system. The variational structure of this Hamiltonian system is demonstrated by using the Hamiltonian action and the Legendre transform. As an applications, some new results on the existence of at least one periodic solution are obtained by means of the dual last action principle. The distinction of this paper from others is that the variational methods and the critical point theory are used to study the fractional Hamiltonian systems.

Key words: Fractional Hamiltonian systems, periodic solutions, variational structure, the dual last action principle.
MSC 2000: 34B25; 34B18; 35G30

1 Introduction

Fractional calculus finds its wide applications in many different fields, such as mathematical modeling of control theory, fluid flow, bioengineering, biochemistry, electrical networks, astronomy, viscoelasticity and other fields of science, see [1, 4] and references therein. Therefore, the theory of fractional differential equation is an area intensively developed during the last decades [3, 5]. The study techniques employed frequently are fixed point theory (including the Banach contraction mapping principle and the Schaefer fixed point theorem [7, 8, 15]), topological degree theory (coincidence degree theory [6]), comparison method (including upper and lower solutions and monotone iterative method [2, 9]), and so on. It should be noted that the study techniques employed in this paper are quite from the above.

 ISBN: 978-84-617-8694-7

The results, which have been obtained in this paper about the existence and multiplicity of solutions to nonlinear fractional differential equation, are derived by using the variational methods and the critical point theory.

Critical point theory has been proved to be very effective tools in determining the existence of solution to integer order Hamiltonian systems, please refer to [16]. The idea behind it is that finding solutions to a given Hamiltonian systems (boundary value problem) is equal to looking for critical points of a suitable energy functional defined on an appropriate function space. But till now, there are few results obtained on the solution to fractional Hamiltonian systems (boundary value problem), by the critical point theory[10–12], since it is often difficult to establish a suitable space and variational functional of fractional Hamiltonian systems.

In [10], Jiao and Zhou have studied the following fractional boundary value problem

$$\begin{cases} {}_tD_T^{\alpha}\left({}_0D_t^{\alpha}u(t)\right) = \nabla F\left(t,u(t)\right), \quad t\in[0,T], \\ u(0)=u(T)=0, \end{cases} \tag{1}$$

where ${}_tD_T^{\alpha}$ and ${}_0D_t^{\alpha}$ are respectively the right and left Riemann-Liouville fractional derivative of order $0<\alpha\leq 1$, $F:[0,T]\times\mathbb{R}^N\to\mathbb{R}$ is a given function satisfying some assumptions and $\nabla F\left(t,u(t)\right)$ is gradient of F at u. The authors have proved the existence of solutions to boundary value problem (1) by means of critical point theory.

Motivated by the results[10–12], we make study on the existence criteria for the following fractional Hamiltonian system

$$\begin{cases} {}_0D_t^{\alpha}\left({}_0D_t^{\alpha}u(t)\right) + \nabla F\left(t,u(t)\right) = 0, \ a.e. \ \ t\in[0,T], \\ u(0)=u(T)=0, \end{cases} \tag{2}$$

where $0<\alpha\leq 1$, $T>0$, $F:(t,u)\to F(t,u)$ measurable in t for every $u\in\mathbb{R}^N$ and continuously differentiable and convex in u for a.e. $t\in[0,T]$, $\nabla F\left(t,u\right)$ is gradient of F at u, and ${}_0D_t^{\alpha}$ is the left Riemann-Liouville fractional derivative of order $0<\alpha\leq 1$. The variational structure of Hamiltonian systems mentioned above is obtained by using the Hamiltonian action and the Legendre transform. As an applications, the existence of at least one periodic solution is demonstrated by using the dual last action principle. The distinction of this paper from others is that the variational methods and the critical point theory are used to study the fractional Hamiltonian systems.

Moreover, consider the following second-order Hamiltonian systems

$$\begin{cases} u''(t) + \nabla F\left(t,u(t)\right) = 0, \quad t\in[0,T], \\ u(0)=u(T)=0. \end{cases} \tag{3}$$

Here it is obvious that the problem on the existence of solution for the system (3) is closely related to the Hamiltonian systems (2), and if the choice of $\alpha=1$, system (2) leads to the system (3).

 ISBN: 978-84-617-8694-7

2 Preliminary

In this section, in order to apply critical point theory to the study of the existence of solutions to the fractional Hamiltonian systems (2), we list some basic definitions and Lemmas which are to be used in the proof of our main results. Some basic definitions and properties of fractional calculus which are used in this paper, please refer to [10, 13, 14].

Let us recall respectively the spaces R, L^p, $L^p[0,t]$, and the norms are defined separately by

$$\| x \|_\infty = \max_{t\in[0,T]} | x(t) |, \quad \| x \|_{L^p} = \left(\int_0^T | x(t) |^p \, ds\right)^{1/p}$$

and

$$\| x \|_{L^p[0,t]} = \left(\int_0^t | x(t) |^p \, ds\right)^{1/p},$$

where $t \in [0,T]$ and $1 < p < \infty$.

Definition 1. [10] Let $0 < \alpha \leq 1$ and $1 < p < \infty$, the fractional derivative space $E_0^{\alpha,p}$ is defined by the closure of $C_0^\infty([0,T],\mathbb{R}^N)$ with respect to norm

$$\| x \|_{\alpha,p} = \left(\int_0^T | x(t) |^p \, dt + \int_0^T | {}_0D_t^\alpha x(t) |^p \, dt\right)^{1/p} \text{ for any } x \in E_0^{\alpha,p},$$

where $C_0^\infty([0,T],\mathbb{R}^N)$ denotes the set of all functions $x \in C^\infty\left([0,T],\mathbb{R}^N\right)$ with $x(0) = x(T) = 0$

We shall denote by $\Gamma_0\left((\mathbb{R}^N\right)$ the set of all convex lower semi-continuous(l.s.c.) functions $F : \mathbb{R}^N \to [-\infty, +\infty]$, whose effective domain $D(F) = \{x \in \mathbb{R}^N : F(x) < +\infty\}$ is non-empty.

Definition 2. [14] Let $H : [0,T]\times\mathbb{R}^{2N} \to \mathbb{R}, (t,u) \to H(t,u)$ be a smooth Hamiltonian such that for each $t \in [0,T]$, $H(t,\cdot) \in \Gamma_0(\mathbb{R}^N)$ is strictly convex and $\frac{H(t,u)}{|u|} \to +\infty$ if $| u | \to +\infty$. The Fenchal (or Legendre) transform is defined by $H^*(t,v) = \sup_{u\in\mathbb{R}^{2N}} (v,u) - H(t,u)$, or

$$\begin{aligned} H^*(t,v) &= (v,u) - H(t,u), \\ v = \nabla H(t,u), \quad u &= \nabla H^*(t,v), \end{aligned} \tag{4}$$

Finally, we present some Lemmas which are used to prove our main results.

Lemma 1. [10, Proposition 3.1] *Let $0 < \alpha \leq 1$ and $1 < p < \infty$, then the fractional derivative space $E_0^{\alpha,p}$ is a reflexive and separable Banach space.*

Lemma 2. [14, Proposition 2.4] If $F \in \Gamma_0(\mathbb{R}^N)$ is strictly convex and such that $F(u)/ | u | \to +\infty$ if $| u | \to \infty$, then $F^* \in C^1(R^N,\mathbb{R})$.

 ISBN: 978-84-617-8694-7

Lemma 3. [10, Proposition 3.2] *Let $0 < \alpha \leq 1$ and $1 < p < \infty$. For all $x \in E_0^{\alpha,p}$, if $1 - \alpha \geq 1/p$ or $\alpha > 1/p$, we have*

$$\| x \|_{L^P} \leq \frac{T^\alpha}{\Gamma(\alpha+1)} \|_0 D_t^\alpha x \|_{L^P} . \tag{5}$$

If $\alpha > 1/p$ and $\frac{1}{p} + \frac{1}{q} = 1$, then

$$\| x \|_\infty \leq \frac{T^{\alpha-1/p}}{\Gamma(\alpha)((\alpha-1)q+1)^{1/q}} \|_0 D_t^\alpha x \|_{L^P} . \tag{6}$$

According to (5), we can consider $E_0^{\alpha,p}$ with respect to the norm

$$\| x \|_{\alpha,p} = \|_0 D_t^\alpha x \|_{L^p} = \left(\int_0^T |_0 D_t^\alpha x |^p \, dt \right)^{1/p} \tag{7}$$

in the following analysis.

Lemma 4. [10, Proposition 3.4] *Let $0 < \alpha \leq 1$ and $1 < p < \infty$. Assume that $\alpha > 1/p$ and the sequence u_k converges weakly to u in $E_0^{\alpha,p}$ i.e. $u_k \rightharpoonup u$, then $u_k \to u$ in $C([0,T], \mathbb{R}^N)$, i.e. $\| u - u_k \|_\infty = 0$ as $k \to \infty$.*

Lemma 5. [14, Theorem 1.1] *If φ is w.l.s.c. on a reflexive Banach space X and has a bounded minimizing sequence, then φ has a minimum on X.*

3 Variational structure

In this section, the variational structure of the fractional Hamiltonian systems (2) is derived by means of the Hamiltonian action and the Legendre transform, which enables us to convert the existence of solutions to fractional Hamiltonian systems (2) into finding the critical points of a corresponding functional.

Let the space X be defined by

$$X = \left\{ v = (v_1, v_2) : v_1 \in E_0^{\alpha,q}\left(0, T; \mathbb{R}^N\right), v_2 \in E_0^{\alpha,p}\left(0, T; \mathbb{R}^N\right) \right\},$$

and the norm is equipped with $\| v \| = \| v_1 \|_{\alpha,q} + \| v_2 \|_{\alpha,p}$. It follows from Lemma 1 that X is a reflexive Banach space.

Assume that $\widetilde{E}_0^{\alpha,p} = \{x \in E_0^{\alpha,p} : \int_0^T x(s)ds = 0\}$, then

$$\widetilde{X} = \left\{ v = (v_1, v_2) : v_1 \in \widetilde{E}_0^{\alpha,q}\left(0, T; \mathbb{R}^N\right), v_2 \in \widetilde{E}_0^{\alpha,p}\left(0, T; \mathbb{R}^N\right) \right\}.$$

Let $(E_0^{\alpha,p})^*$ be standed by the conjugate space of $E_0^{\alpha,p}$, there is

$$X^* = \left\{ v = (v_1, v_2) : v_1 \in (E_0^{\alpha,q})^*, v_2 \in (E_0^{\alpha,p})^* \right\},$$

 ISBN: 978-84-617-8694-7

and we easily know that X^* is the conjugate space of X.

Moreover, let the space Y be defined by

$$Y = \left\{ u = \left(u_1, u_2\right) : u_1 \in E_0^{\alpha,p}\left((0,T;\mathbb{R}^N\right), u_2 \in E_0^{\alpha,q}\left((0,T;\mathbb{R}^N\right)\right\},$$

Let $h \in L^1(0,T;\mathbb{R}^N)$, the mean value is defined by $\bar{h} = \frac{1}{T}\int_0^t h(t)dt$. We denote by $J = \begin{pmatrix} 0_N & I_N \\ -I_N & 0_N \end{pmatrix}$ the symplectic matrix, then $(Ju, v) = -(u, Jv)$ for all $u, v \in \mathbb{R}^{2N}$.

Let $x = u_1$, ${}_0D_t^\alpha x = \lambda u_2$, then, the fractional Hamiltonian systems (2) reduces to

$${}_0D_t^\alpha u_1 + \lambda u_2 = 0 \text{ and } u_i(0) = u_i(T) = 0, \quad i = 1, 2,$$

that is, the fractional Hamiltonian systems (2) is equivalent to the following non-autonomous fractional Hamiltonian systems

$$\begin{cases} J {}_0D_t^\alpha u + \nabla H(t,u) = 0, \\ u(0) = u(T) = 0, \end{cases} \tag{8}$$

where ${}_0D_t^\alpha u = \begin{pmatrix} {}_0D_t^\alpha u_1 \\ {}_0D_t^\alpha u_2 \end{pmatrix}, \nabla H(t,u) = \begin{pmatrix} \frac{1}{\lambda}\nabla F(t,u_1) \\ -\lambda u_2 \end{pmatrix}$. Therefore, the Hamiltonian action ψ can be written

$$\psi(u) = -\int_0^T \left[\tfrac{1}{2}(J {}_0D_t^\alpha u(t), u(t)) + H(t,u(t))\right] dt. \tag{9}$$

Setting ${}_0D_t^\alpha v = -J {}_0D_t^\alpha u$, so that

$$u = Jv - c. \tag{10}$$

According to (8) and (11), we have ${}_0D_t^\alpha v = \nabla H(t,u)$, if the Legendre transform $H^*(t,.)$ of $H(t,.)$ existence, then

$$u = \nabla H^*\left(t, {}_0D_t^\alpha v\right). \tag{11}$$

Hence, in terms of (10), the Hamiltonian equations (11) reduces to $Jv - \nabla H^*\left(t, {}_0D_t^\alpha v\right) = c$.

In the following, we prove that the Legendre transform $H^*(t,.)$ of $H(t,.)$ existence. It follow from (9) and (11) that

$$\begin{aligned} &\psi(u) = -\int_0^T \left[\tfrac{1}{2}(J {}_0D_t^\alpha u(t), u(t)) + H(t,u(t))\right] dt = \int_0^T \left[\tfrac{1}{2}({}_0D_t^\alpha v(t), u(t)) - H(t,u(t))\right] dt \\ &= \int_0^T \left[-\tfrac{1}{2}({}_0D_t^\alpha v(t), u(t)) + ({}_0D_t^\alpha v(t), u(t)) - H(t,u(t))\right] dt \\ &= \int_0^T \left[\tfrac{1}{2}(J {}_0D_t^\alpha u(t), u(t)) + ({}_0D_t^\alpha v(t), u(t)) - H(t,u(t))\right] dt. \end{aligned}$$

Formula (4) suggests replacing $({}_0D_t^\alpha v, u) - H(t,u)$ by $H^*(t, {}_0D_t^\alpha v)$. The dual action is thus defined on a suitable space of T-periodic functions by

$$\chi(v) = \int_0^T \left[\tfrac{1}{2}({}_0D_t^\alpha v(t), v(t)) + H^*(t, {}_0D_t^\alpha v(t))\right] dt,$$

where $v = \left(v_1, v_2\right), v_1, v_2 \in \mathbb{R}^N$. In Section 4, we shall see that χ can be bounded below under reasonable assumptions upon H. So it suffices to find critical points of χ restricted to the space $\widetilde{X}$.

 ISBN: 978-84-617-8694-7

Lemma 6. *Assume that $H : [0,T] \times \mathbb{R}^{2N} \to \mathbb{R}, (t,u) \to H(t,u)$ be measurable in t for each $u \in \mathbb{R}^{2N}$ and strictly convex and continuously differentiable in u for almost every $t \in [0,T]$. If there exists $q \in [1,\infty], \lambda > 0, \delta > 0, \beta, \gamma \in L^p(0,T;\mathbb{R}^+)$ with $\frac{1}{p} + \frac{1}{q} = 1$, such that for all $u \in \mathbb{R}^{2N}$ and a.e. $t \in [0,T]$, one has*

$$\delta(\mid u \mid^q /q) - \beta(t) \leq H(t,u) \leq \lambda(\mid u \mid^q /q) + \gamma(t). \tag{12}$$

Then the dual action χ is a continuously differentiable on $\tilde{X}$ and, if $v \in \tilde{X}$ is a critical point of χ, the function u defined by $u(t) = \nabla H^*(t, {}_0D_t^\alpha v(t))$ satisfies fractional Hamiltonian systems (8).

Proof. It follows from Lemma 2 that $H^*(t,u)$ is continuously differentiable in u for a.e. $t \in [0,T]$. By assumption (12), for all $u \in \mathbb{R}^{2N}$ and a.e. $t \in [0,T]$, we obtain

$$\lambda^{-p/q}(\mid v \mid^p /p) - \gamma(t) \leq H^*(t,v) \leq \delta^{-p/q}(\mid v \mid^p /p) + \beta(t). \tag{13}$$

Proposition 2.2 in [14] implies that

$$\mid \nabla H^*(t,v) \mid \leq [(q/\delta)(\mid v \mid) + \beta(t) + \gamma(t) + 1]^{p-1} \leq c_1 \mid v \mid^{p-1} + c_2(\beta(t) + \gamma(t) + 1)^{p-1},$$

for some positive constants c_1 and c_2.

Let us note that $(\beta+\gamma+1)^{p-1} \in L^q$ since $(\beta+\gamma+1) \in L^p$. By (13) and (14), the function L defined by $L(t,x,y) = (1/2)(Jy,x) + H^*(t,y)$ satisfies the assumptions of Theorem 3.1 in [10]. Therefore, the dual action χ is continuously differentiable on X, and, hence, on $\tilde{X}$.

Finally, if $v \in \tilde{X}$ is a critical point of χ, Theorem 3.1 in [10] implies that, for all $h \in \tilde{X}$, one has

$$0 = \int_0^T \left[\tfrac{1}{2}(J_0D_t^\alpha v(t), h(t)) + (\nabla H^*(t, {}_0D_t^\alpha v(t)) - \tfrac{1}{2}Jv(t), {}_0D_t^\alpha h(t))\right] dt. \tag{14}$$

It is then easy to verify the preceding relation for all $h \in X$, and hence for all $h \in C_T^\infty$. By (14), we have

$$\nabla H^*(t, {}_0D_t^\alpha v(t)) - \tfrac{1}{2}Jv(t) = \int_0^t \tfrac{1}{2}J_0D_t^\alpha v(s)ds + c, \text{ a.e. on } [0,T],$$

i.e., $Jv(t) = \nabla H^*(t, {}_0D_t^\alpha v(t)) + \tilde{c}$, a.e. on $[0,T]$. Setting $u(t) = H^*(t, {}_0D_t^\alpha v(t)) = Jv(t) - \tilde{c}$, we obtain $u \in X$, ${}_0D_t^\alpha u = J_0D_t^\alpha v$ and by duality ${}_0D_t^\alpha v(t) = \nabla H(t,u(t))$, therefore ${}_0D_t^\alpha u = J_0D_t^\alpha v = J\nabla H(t,u(t))$, a.e. on $[0,T]$. Moreover, $u(0) = u(T)$ since $u \in X$. The proof is complete.

Lemma 7. *For every $v \in X$, there is*

$$\int_0^T (J_0D_t^\alpha v(t), v(t))\, dt \geq -\tfrac{T^\alpha}{\Gamma(\alpha+1)} \int_0^T \mid {}_0D_t^\alpha v(t) \mid^2 dt.$$

 ISBN: 978-84-617-8694-7

Proof. Let us write $v = \tilde{v} + \bar{v}$, where $\bar{v} = \frac{1}{T}\int_0^T v(s)ds$, then by Cauchy-Schwarz inequality and Lemma 3, we obtain

$$\int_0^T (J_0D_t^\alpha v(t), v(t))\,dt = \int_0^T (J_0D_t^\alpha v(t), \tilde{v}(t))\,dt \geq -\left(\int_0^T \mid J_0D_t^\alpha v(t) \mid^2 dt\right)^{1/2}\left(\int_0^T \mid \tilde{v}(t) \mid^2\right)^{1/2}$$
$$\geq -\frac{T^\alpha}{\Gamma(\alpha+1)}\left(\int_0^T \mid J_0D_t^\alpha v(t) \mid^2 dt\right)^{1/2}\left(\int_0^T \mid_0 D_t^\alpha v(t) \mid^2\right)^{1/2}$$
$$= -\frac{T^\alpha}{\Gamma(\alpha+1)}\int_0^T \mid_0 D_t^\alpha v(t) \mid^2 dt.$$

4 The existence of weak solutions

In this section, the existence theorem of at least one periodic solution to the fractional Hamiltonian systems (2) is proved.

Theorem 1. *Assume that the following conditions are satisfied*

(A1) *there exists* $l \in L^4(0,T;\mathbb{R}^N)$ *such that* $H(t,u) \geq (l(t),u)$ *for all* $u \in \mathbb{R}^{2N}$ *and a.e.* $t \in [0,T]$*;*

(A2) *there exists* $\lambda \in [0, 2\pi/T]$*, and* $\gamma \in L^2(0,T;\mathbb{R}^N)$*, such that for every* $y \in \mathbb{R}^{2N}$ *and a.e.* $t \in [0,T]$ *one has* $H(t,u) \leq \frac{\lambda}{2} \mid u \mid^2 + \gamma(t)$*;*

(A3) $\int_0^T H(t,u)dt \to +\infty$ *as* $\mid u \mid \to \infty, u \in \mathbb{R}^{2N}$*.*

Then the fractional Hamiltonian system (8) *has at least one solution* $u \in Y$ *such that* $v(t) = -J\left[u(t) - \frac{1}{T}\int_0^T u(s)ds\right]$ *minimizes the dual actian* $\chi : X \to [-\infty, +\infty], v \to \int_0^T \left[\frac{1}{2}(J_0D_t^\alpha v(t), v(t)) + H^*(t,{}_0D_t^\alpha v(t))\right]dt$*, that is to say, the fractional Hamiltonian system* (2) *has at least one solution* $x \in E_0^{\alpha,p}$*.*

Proof. We divide the proof into three steps.

Step 1. A perturbed problem is established and existence of a solution of it is proved.

Let $\epsilon_0 > 0$ be such that

$$0 < \epsilon_0 + \lambda < \frac{\Gamma(\alpha+1)}{T^\alpha} \tag{15}$$

and let

$$H_\epsilon : [0,T] \times \mathbb{R}^{2N} \to \mathbb{R}, (t,u) \to \frac{\epsilon \mid u \mid^2}{2} + H(t,u)$$

where $0 < \epsilon < \epsilon_0$. Since the function $F(t,x)$ is measurable on t for every $x \in \mathbb{R}^N$ and continuously differentiable and convex in x for a.e. $t \in [0,T]$, then $H(t,u)$ is measurable on t for all $u \in \mathbb{R}^{2N}$ and continuously differentiable and convex in u for a.e. $t \in [0,T]$. By the defined of H_ϵ, we know that $H_\epsilon(t,\cdot)$ is strictly convex and continuously differentiable for a.e. $t \in [0,T]$ and $H_\epsilon(\cdot,u)$ is measurable on $[0,T]$ for every $u \in \mathbb{R}^{2N}$. From (A1) and (A2), we have

$$-\mid l(u) \mid\mid u \mid + \frac{\epsilon \mid u \mid^2}{2} \leq H_\epsilon(t,u) \leq (\lambda+\epsilon_0)\frac{\mid u \mid^2}{2} + \gamma(t),$$

 ISBN: 978-84-617-8694-7

therefore

$$\frac{\epsilon \mid u \mid^2}{4} - \frac{\mid l(t) \mid^2}{2} \leq H_\epsilon(t,u) \leq (\lambda+\epsilon_0)\frac{\mid u \mid^2}{2} + \gamma(t), \tag{16}$$

So, by Lemma 6, the perturbed dual action

$$\chi_\epsilon(v) = \int_0^T \left[\frac{1}{2}(J_0 D_t^\alpha v(t), v(t)) + H_\epsilon^*(t, {}_0D_t^\alpha v(t))\right] dt$$

is continuously differentiable on $\tilde{X}$ and if $v_\epsilon \in \tilde{X}$ is a critical point of χ_ϵ, the function u_ϵ defined by

$$u_\epsilon(t) = \nabla H_\epsilon^*(t, {}_0D_t^\alpha v(t))$$

is a solution of

$$\begin{cases} J_0 D_t^\alpha u(t) + \epsilon u(t) + \nabla H(t,u) = 0, \\ u(0) = u(T), \end{cases} \tag{17}$$

and the relation $J_0 D_t^\alpha v_\epsilon =_0 D_t^\alpha u_\epsilon$ holds.

By (16) and Propositions 2.2 in [14], we have

$$H_\epsilon^*(t, {}_0D_t^\alpha v(t)) \geq \frac{1}{2}\left(\frac{1}{\lambda+\epsilon_0}\right) \mid_0 D_t^\alpha v(t) \mid^2 -\gamma(t),$$

which together with Lemma 7 implies that

$$\begin{aligned} \chi_\epsilon(v) &\geq \frac{1}{2}\left(\frac{1}{\lambda+\epsilon_0} - \frac{\Gamma(\alpha+1)}{T^\alpha}\right)\int_0^T \mid_0 D_t^\alpha v(t) \mid^2 dt - \int_0^T \gamma(t)dt \\ &= \delta_0 \parallel_0 D_t^\alpha v(t) \parallel_{L^2}^2 -\gamma_0, \end{aligned} \tag{18}$$

where $\delta_0 = \frac{1}{2}\left(\frac{1}{\lambda+\epsilon_0} - \frac{\Gamma(\alpha+1)}{T^\alpha}\right) > 0$ by (15) and $\gamma_0 = \int_0^T \gamma(t)dt$.

Let $\{v_k\}$ be a minimizing sequence for χ_ϵ. It follows from (18)that $\{\parallel_0 D_t^\alpha v_k(t) \parallel_{L^2}\}$ is bounded, then Lemma 3 implies that $\{v_k\}$ is bounded in $\tilde{X}$. Now we show χ_ϵ is weakly lower semi-continuous (w.l.s.c.) on $\tilde{X}$. From the continuity of H_ϵ and the definition of H_ϵ^*, it follows that $\chi_{\epsilon,1}(v) = \int_0^T H_\epsilon^*(t, {}_0D_t^\alpha v(t))dt$ is w.l.s.c. on $\tilde{X}$ and $\chi_{\epsilon,2}(v) = \frac{1}{2}\int_0^T (J_0 D_t^\alpha v(t), v(t))dt$ is w.l.s.c. on $\tilde{X}$. Thus $\chi_\epsilon = \chi_{\epsilon,1}(v) + \chi_{\epsilon,2}(v)$ is w.l.s.c. on $\tilde{X}$ and by Lemma 5, has a minimum at some point $v_\epsilon \in \tilde{X}$.

Step 2. A posteriori estimates on u_ϵ

By assumption (A1), (A2) and Proposition 2.2, we have

$$\mid \nabla H(t,u) \mid \leq 2(\lambda+1)\left[\mid u \mid + \frac{|l(t)|^2}{2} + \gamma(t)\right] + 1 + \mid u \mid .$$

It is then easy to verify that the function $\bar{H} : \mathbb{R}^{2N} \to \mathbb{R}, u \to \int_0^T H(t,u)dt$ is continuously differentiable. By assumption (A3), $\bar{H}$ has a minimum at some point $u \in \mathbb{R}^{2N}$ for which $\int_0^T \nabla H(t,\bar{u})dt = 0$, so that the problem

$${}_0D_t^\alpha v(t) = \nabla H(t,\bar{u}) \tag{19}$$

 ISBN: 978-84-617-8694-7

has a unique solution $\eta \in \tilde{X}$ such that $\int_0^T \eta(s)ds = 0$. By (19), $H^*(t,{}_0D_t^\alpha\eta(t)) = ({}_0D_t^\alpha\eta(t),\bar{u}) - H(t,\bar{u})$ so that $H^*(\cdot,{}_0D_t^\alpha\eta(\cdot)) \in L^1(0,T;\mathbb{R})$. From the inequality $H(t,u) \le H_\epsilon(t,u)$ we deduce $H_\epsilon^*(t,v) \le H^*(t,v)$ and from (18), we obtain

$$\begin{aligned}&\delta_0 \parallel {}_0D_t^\alpha v(t) \parallel_{L^2}^2 - \gamma_0 \le \chi_\epsilon(v_\epsilon) \le \chi_\epsilon(\eta)\\ &\le \int_0^T \left[\tfrac{1}{2}(J_0D_t^\alpha\eta(t),\eta(t)) + H_\epsilon^*(t,{}_0D_t^\alpha\eta(t))\right]dt = c_1 < \infty.\end{aligned}$$

Therefore $\parallel {}_0D_t^\alpha v_\epsilon \parallel_{L^2} \le c_2$. According to $J_0D_t^\alpha v_\epsilon = {}_0D_t^\alpha u_\epsilon$, we have $\parallel {}_0D_t^\alpha \tilde{u}_\epsilon \parallel_{L^2} = \parallel {}_0D_t^\alpha u_\epsilon \parallel_{L^2} \le c_2$, where $\tilde{u}_\epsilon = u_\epsilon - \bar{u}_\epsilon, \bar{u}_\epsilon = \frac{1}{T}\int_0^T u_\epsilon(t)dt$. Lemma 3 implies that $\parallel \tilde{u}_\epsilon \parallel \le c_3$. The convexity of $H(t,\cdot)$ and (17) reduce to

$$\begin{aligned}&H\left(t,\tfrac{\bar{u}_\epsilon(t)}{2}\right) = H\left(t,\tfrac{u_\epsilon(t)}{2} - \tfrac{\tilde{u}_\epsilon(t)}{2}\right) \le \tfrac{1}{2}H(t,u_\epsilon(t)) + \tfrac{1}{2}H(t,-\tilde{u}_\epsilon(t))\\ &\le \tfrac{1}{2}(\nabla H(t,u_\epsilon(t)),u_\epsilon(t)) + \tfrac{1}{2}H(t,0) + \tfrac{\lambda}{4} \mid \tilde{u}_\epsilon(t) \mid^2 + \tfrac{\gamma(t)}{2}\\ &\le \tfrac{1}{2}(-J_0D_t^\alpha u_\epsilon(t),u_\epsilon(t)) - \tfrac{\epsilon}{2} \mid u_\epsilon(t) \mid^2 + \tfrac{\lambda}{4} \mid \tilde{u}_\epsilon(t) \mid^2 + \tfrac{\gamma(t)}{2}.\end{aligned}$$

Using Lemma 7, we obtain

$$\begin{aligned}&\int_0^T H(t,\bar{u}_\epsilon(t))dt \le -\tfrac{1}{2}\int_0^T (J_0D_t^\alpha u_\epsilon(t),u_\epsilon(t)) + \tfrac{\lambda}{4} \parallel \tilde{u}_\epsilon(t) \parallel_{L^2}^2 + \gamma_0\\ &\le \tfrac{T^\alpha}{2\Gamma(\alpha+1)} \parallel {}_0D_t^\alpha\tilde{u}_\epsilon(t) \parallel_{L^2}^2 + \tfrac{\lambda}{4} \parallel \tilde{u}_\epsilon(t) \parallel_{L^2}^2 + \gamma_0\\ &\le \tfrac{T^\alpha}{2\Gamma(\alpha+1)}c_2^2 + \tfrac{\lambda}{4}c_3^2 + \gamma_0 = c_4.\end{aligned}$$

By assumption (A3), $\mid \bar{u}_\epsilon \mid \le c_5$, we have $\parallel u_\epsilon \parallel \le \parallel \tilde{u}_\epsilon \parallel + \parallel \bar{u}_\epsilon \parallel \le c_3 + \sqrt{T}c_5 = c_6$.

Step 3. Existence of a solution for the original problem.

From $\parallel u_\epsilon \parallel \le c_6$, there exists a sequence $\{\epsilon_n\}$ tending to 0 in $[0,\epsilon_0]$ and some $u \in Y$ such that u_{ϵ_n} converge weakly to u in Y. Moreover, as ${}_0D_t^\alpha v_\epsilon = -J_0D_t^\alpha u_\epsilon$, we have $v_\epsilon(t) = -J(u_\epsilon(t) - \bar{u}_\epsilon)$. so that v_{ϵ_n} converges weakly to

$$v(t) = -J(u - \bar{u}). \tag{20}$$

By Proposition 1.2 in [14], u_{ϵ_n} (resp. v_{ϵ_n}) converges uniformly to u (resp. v) on $[0,T]$. From (17) in integrated form

$$Ju_{\epsilon_n}(t) - Ju_{\epsilon_n}(0) + \int_0^T \left[\epsilon_n u_{\epsilon_n}(s) + \nabla H(s,u_{\epsilon_n}(s))\right]ds = 0,$$

it follows that

$$Ju(t) - Ju(0) + \int_0^T \nabla H(s,u(s))ds = 0,$$

therefore $u \in Y$ is a solution of (8).

Finally, we will show that $v = (v_1,v_2)$ minimizes the dual action χ. since

$$H_\epsilon^*(t,v) \le H^*(t,v) \text{ for all } h \in X,$$

 ISBN: 978-84-617-8694-7

we have

$$\chi_{\epsilon_n}(v_{\epsilon_n}) \leq \chi_{\epsilon}(h) \leq \chi(h).$$

Now by the definition $H^*(t,\cdot)$ and duality between u_{ϵ_n} and ${}_0D_t^{\alpha}v_{\epsilon_n}$, we obtain

$$\begin{aligned}\chi_{\epsilon_n}(v_{\epsilon_n}) &= \int_0^T \left[\tfrac{1}{2}(J_0D_t^{\alpha}v_{\epsilon_n}(t), v_{\epsilon_n}(t)) + (u_{\epsilon_n},{}_0D_t^{\alpha}v_{\epsilon_n}(t)) - H_{\epsilon_n}(t,u_{\epsilon_n}(t))\right]dt\\ &= \int_0^T \left[\tfrac{1}{2}(J_0D_t^{\alpha}v_{\epsilon_n}(t), v_{\epsilon_n}(t)) + (u_{\epsilon_n},{}_0D_t^{\alpha}v_{\epsilon_n}(t)) - H(t,u_{\epsilon_n}(t)) - \tfrac{\epsilon_n}{2} \mid u_{\epsilon_n}(t) \mid^2\right]dt.\end{aligned}$$

It follows from (8) and (20) that

$${}_0D_t^{\alpha}v(t) = \nabla H(t,u(t)) \qquad \text{a.e. on } [0,T]. \tag{21}$$

Letting $t \to \infty$ in (21), we have, by (21)

$$\begin{aligned}\lim_{n\to\infty}\chi_{\epsilon_n}(v_{\epsilon_n}) &= \int_0^T \left[\tfrac{1}{2}(J_0D_t^{\alpha}v(t), v(t)) + (u(t),{}_0D_t^{\alpha}v(t)) - H(t,u(t))\right]dt\\ &= \int_0^T \left[\tfrac{1}{2}(J_0D_t^{\alpha}v(t), v(t)) + H^*(t,{}_0D_t^{\alpha}v(t))\right]dt = \chi(v).\end{aligned}$$

Thus $\chi(v) \leq \chi(h)$ for all $h \in X$. The proof is completed.

When $H(t,\cdot)$ is strictly convex, we deduce from Theorem 1 that there exists a necessary and sufficient condition for the solvability of the fractional Hamiltonian systems (2)

Corollary 1. *Assume that $H(t,\cdot)$ is strictly convex for a.e. $t \in [0,T]$ and satisfies the conditions* (A1) *and* (A2) *of Theorem* 1. *Then the following conditions are equivalent.*

(A4) *Problem* (8) *is solvable, i.e. problem* (2) *is solvable.*

(A5) *There exists $\bar{x} \in \mathbb{R}^{2N}$ such that $\int_0^T \nabla H(t,\bar{x})dt = 0$.*

(A6) $\int_0^T H(t,x)dt \to +\infty$ *when* $\mid x \mid \to \infty$.

Acknowledgements

This work has been partially supported by NSF of China (11361047,11501560) and NSF of Jiangsu Province (BK20151160), the Six Talent Peaks Project of Jiangsu Province (2013-JY-003) and the 333 High-Level Talents Training Program of Jiangsu Province(BRA2016275).

References

[1] R. P. Agarwal, M. Benchohra, S. Hamani, A survey on existence results for boundary value problems of nonlinear fractional differential equations and inclusions, Acta Appl. Math. 109 (2010) 973-1033.

[2] Z. Bai, S. Zhang, S. Sun, C. Yin, Monotone iterative method for fractional differential equations, Electron. J. Diff. Equ. 2016 (2016) 1-8.

[3] Z. Bai, W. Sun, Existence and multiplicity of positive solutions for singular fractional boundary value problems, Comput. Math. Appl. 63 (2012) 1369-1381.

[4] A. Carpinteri, F. Mainardi, Fractals and Fractional Calculus in Continuum Mechanics, Springer, 1997.

[5] Y. Chang, J.J. Nieto, Some new existence results for fractional differential inclusions with boundary conditions, Math. Comput. Modelling 49 (2009). 605-609.

[6] T.Y. Chen, W. B. Liu, Z. G. Hu, A boundary value problem for fractional differential equation with p-Laplacian operator at resonance, Nonlinear Anal. 75 (2012) 3210-3217.

[7] C. Cheng, Z. Feng, Y. Su, Positive solutions of fractional differential equations with derivative terms, Electron. J. Differ. Equ. 2012, 215 (2012) 1-27.

[8] M. A. E. Herzallah, D. Baleanu, Existence of a periodic mild solution for a nonlinear fractional differential equation, Comput. Math. Appl. 64 (2012) 3059-3064.

[9] T. Jankowski, Boundary problems for fractional differential equations, Appl. Math. Lett. 28 (2014) 14-19.

[10] F. Jiao, Y. Zhou, Existence results for fractional boundary value problem via critical point theory, Internat. J. Bifur. Chaos, Sci. Engrg. 22 (2012), No. 4, Art. ID 1250086, 1-17.

[11] F. Jiao, Y. Zhou, Existence of solutions for a class of fractional boundary value problems via critical point theory, Comput. Math. Appl. 62 (2011) 1181-1199.

[12] Y. Zhou, L. Zhang, Existence and multiplicity results of homoclinic solutions for fractional Hamiltonian systems, Comput. Math. Appl. 73 (2017) 1325-1345.

[13] A. A. Kilbas, H. M. Srivastava, J. J. Trujillo, Theory and Applications of Fractional Differential Equations, vol. 204, Elsevier Science B.V. Amsterdam, 2006.

[14] J. Mawhin, M. Willem, Critical Point Theorey and Hamiltonian Systems, Springer, New York, 1989.

[15] Y.H. Su, Z. Feng, Existence theory for an arbitrary order fractional differential equation with deviating argument, Acta. Appl. Math. 118 (2012) 81-105.

[16] J. Zhou, Y. Li, Existence of solutions for a class of second-order Hamiltonian systems with impulsive effects, Nonlinear Anal. 72 (2010) 1594-1603.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

Developing a new method with memory based on Hermite's interpolation

Maryam Mohamadi Zade[1] **and Taher Lotfi**[1]

[1] *Department of Applied Mathematics, Hamedan Branch,,*
Islamic Azad University, Hamadan, Iran

emails: ma.mohamadizade@yahoo.com, Corresponding author:
lotfitaher@yahoo.com

Abstract

In this work, we concern to develop a new with memory method for solving nonlinear equations. To this end, we use three functional evaluation per iteration. In addition, our method is constructed by the information of Hermite's interpolation. It has convergence order 5.5. Some numerical examples are included to show the efficiency and its applicability.

Key words: With memory method, Hermit's interpolation, efficiency, accuracy, convergence order.

1 Introduction

There are so many problems in science and technology that lead to nonlinear of equations. Solving these equation is based on iterative methods. However, such methods should satisfy some standards, Kung and Traub [2], and Ostrowski [1] proposed two criteria for this purpose: optimality and efficiency.
Kung and Traub established that any two-step without memory method is optimal and uses three functional evaluations having convergence order four. Also, efficiency index is defined by $E(p,n) = p^{1/n}$, where p and n are convergence order and functional evaluations, respectively. Therefore any two- step without memory method has the same efficiency index $E(4,3) = 4^{1/3} \simeq 1.587$ [2]. Following Traub's idea of introducing with memory methods, many researcher [2] constructed many with memory methods [3]. Most of these methods are

 ISBN: 978-84-617-8694-7

free derivative [4], [5] and [6]. To the best of our knowledge, construction of with memory method in which they use derivative, are few.

It is well-known that iterative method for solving nonlinear equation is divided into two general categories with and without memory. There has been many attempts to construct optimal without memory methods of both kinds them [6]. Traub in his book shows that it is possible to increase the convergence order of Steffensen type method without any new functional evaluation. His idea has been followed by many authors. On the other hand, there are few methods of such kind [7] that use derivatives. Hence, in this work we focus on developing new kind of such methods, say, a new with memory method based on Hermite's interpolation.

The rest of this work is organised as follows: In Section 2, we design an optimal without memory method. Then we attempt to derive two kinds of with memory methods. In Section 3, some numerical examples are presented. Section 4 concludes this work.

2 The method and analysis of convergence

Consider the following two-step method

$$\begin{cases} y_n = x_n - \frac{f(x_n)}{f'(x_n)}, \\ x_{n+1} = y_n - \frac{f(y_n)}{f'(y_n)}, \qquad n = 0, 1, 2, \ldots. \end{cases} \tag{2.1}$$

when x_0 is given. This method is not optimal. We approximate $f'(y_n)$ in the second step by Hermite's interpolation.

Let

$$P(t) = f(y_n) + (t - y_n)f[y_n, x_n] + (t - y_n)(t - x_n)f[y_n, x_n, x_n]$$

where

$$f[x_n, x_n] = f'(x_n), f[x_n, y_n] = \frac{f(x_n) - f(y_n)}{x_n - y_n}$$

and

$$f[y_n, x_n, x_n] = \frac{f[y_n, x_n] - f[x_n, x_n]}{y_n - x_n}.$$

Therefore,

$$p'(t) = f[y_n, x_n] + (2t - y_n - x_n)f[y_n, x_n, x_n],$$

Hence

$$f'(y_n) \simeq p'(y_n) = f[y_n, x_n] + (y_n - x_n)f[y_n, x_n, x_n]$$

 ISBN: 978-84-617-8694-7

Consequently,

$$\begin{cases} y_n = x_n - \dfrac{f(x_n)}{f'(x_n)}, \\ x_{n+1} = y_n - \dfrac{f(y_n)}{f[y_n, x_n] + (y_n - x_n) f[y_n, x_n, x_n]}, \qquad n = 0, 1, 2, \ldots. \end{cases} \tag{2.2}$$

This method is optimal and has convergence order four.

Theorem 2.1. *Method (2.2) has convergence order four if x_0 is close enough to the sough zero. It's error equation is $e_{n+1} = (c_2^2 - c_2 c_3) e_n^4 + o(e_n^5)$*

Now, we modify Method (2.2) to as follows:

$$\begin{cases} y_n = x_n - \dfrac{f(x_n)}{f'(x_n) + \alpha_n f(x_n)}, \\ x_{n+1} = y_n - \dfrac{f(y_n)}{f[y_n, x_n] + (y_n - x_n) f[y_n, x_n, x_n]}, \qquad n = 0, 1, 2, \ldots. \end{cases} \tag{2.3}$$

Theorem 2.2. *The error equation of method (2.3)is given by $e_{n+1} = (\alpha_n + c_2)(c_2(\alpha_n + c_2) - c_3) e_n^4 + o(e_n^5)$*

To derive a with memory method, we use suppose $\alpha_n + c_n = 0$, or $\alpha_n = -c_2 = -\frac{f''(\alpha)}{2f'(\alpha)}$. Since α is unknown,we use $f''(\alpha) \simeq p''(y_n) = 2f[y_n, x_n, y_n]$ and $f'(\alpha) \simeq p'(y_n)$.
If in each iteration, we use the following accelerator $\alpha_n = -\dfrac{p''(y_n)}{2p(y_n)}$, then we have

Theorem 2.3. *If $\alpha_n = -\dfrac{p''(y_n)}{2p(y_n)}$, then convergence order of method (2.3) is 5.*

It is still possible to increase the convergence order. Consider the following modification of (2.3)

$$\begin{cases} y_n = x_n - \dfrac{f(x_n)}{f'(x_n) + \alpha_n f(x_n)}, \\ x_{n+1} = y_n - \dfrac{f(y_n)}{f[e, e_y] + (e_y - e) f[e_y, e, e] + \beta_n (e_y - e)^2}, \qquad n = 0, 1, 2, \ldots. \end{cases} \tag{2.4}$$

We have the following error equation

 ISBN: 978-84-617-8694-7

Theorem 2.4. *The error equation of method (2.4) is given by*

$$e_{n+1} = \frac{1}{f'(\alpha)}(c_2+\alpha_n)\Big(\beta_n + f'(\alpha)c_2(\alpha_n+c_2) - f'(\alpha)c_3\Big)e_n^4 + o(e_n^5)$$

If $\alpha + c_2 = 0$, then we can set $\beta_n = \dfrac{f'''(\alpha)}{6}$.
Here if is still possible to approximate the accelerator β_n by the interpolatory

$$N_3(t, y_n, y_{n-1}, x_{n-1}, x_{n-1}),$$

where

$$\begin{aligned} N_3(t) =& f(y_n) + (y_n - t)f[y_n, x_n] \\ &+(t-y_n)(t-x_n)f[y_n, x_{n-1}, x_{n-1}] \\ &+(t-y_n)(t-x_{n-1})f[y_n, x_{n-1}, x_{n-1}, y_{n-1}] \end{aligned}$$

Hence,

Theorem 2.5. *If* $\alpha_n = -\dfrac{p''(y_n)}{2p'(y_n)}$ *and* $\beta_n = \dfrac{N_3'''(y_n)}{6}$, *then method (2.4) has convergence order* 5.5.

Methods (2.3), (2.4) are called with memory method since they use information from the current and the previous iterations. To the best our knowledge, method (2.4) is the first method of its kind. Indeed previous works have increased the convergence order at most to five.

3 Numerical results and comparisons

In this section, the family of with memory methods

$$\begin{cases} y_n = x_n - \dfrac{f(x_n)}{\lambda_n f(x_n) + f'(x_n)}, \\ x_{k+1} = y_n - \dfrac{f(y_n)}{2\lambda_n f(x_n) + f'(x_n)}\Big(1 + 2\dfrac{f(y_n)}{f(x_n} + \tau(\dfrac{f(y_n)}{f(x_n)})^2\Big), \end{cases} \tag{3.1}$$

$$\begin{cases} y_n = x_n - \dfrac{f(x_n)}{\lambda_n f(x_n) + f'(x_n)}, \\ x_{n+1} = y_n - \dfrac{f(y_n)}{2\lambda_n f(x_n) + f'(x_n)}\Big(\dfrac{f(x_n) + (2+\beta)f(y_n)}{f(x_n) + \beta f(y_n)}\Big) \end{cases} \tag{3.2}$$

 ISBN: 978-84-617-8694-7

methods	$\lvert x_1-\alpha\rvert$	$\lvert x_2-\alpha\rvert$	$\lvert x_3-\alpha\rvert$	r_c
Method (3.1)	0.5505(-1)	0.7972(-5)	01081(-24)	4.99
Method (3.2)	0.5505(-1)	07972(-5)	0.4138(-24)	4.99
Our Method(2.3)	0.8524(-1)	0.8780(-4)	0.1952(-19)	5.0
Our Method(2.4)	0.9195(-1)	0.5435(-4)	0.3306(-23)	5.5

Table 1: $f(x)=\dfrac{1}{t^4}-t^2-\dfrac{1}{t}+1,\ x_0=1,\ \alpha=2,\ \lambda=-0.01$

where $\lambda_n=-c_2=-\dfrac{f''(\alpha)}{2f'(\alpha)}$

These tables include the values of the computational order of convergence r_c calculated by the formula [3]

$$r_c=\frac{\log\lvert f(x_n)/f(x_{n-1})\rvert}{\log\lvert f(x_{n-1})/f(x_{n-2})\rvert}. \tag{3.3}$$

Table 1 shows that the Method (2.4) competes the previous methods. In additional its efficiency index is better than the previous works. In other words, it has efficiency index $5.5^{\frac{1}{3}}\simeq 1.76$ greater than $5^{\frac{1}{3}}\simeq 1.70$.

4 Conclusion

In this work, we developed a new kind of with memory methods for solving nonlinear equations. To this end, based on Hermite's interpolation, two special kinds, of the methods were derived. One of them increases the convergence order from 4 to 5, and the other one from 4 to 5.5. The latter kind has not been studied yet in the literature and can compete with every current with memory in its class. This method is under developing for general case.

Acknowledgements

This work has been partially supported by Islamic Azad University- Hamedan Branch.

References

[1] A.M. Ostrowski, *Solution of equations and systems of equations*, , Prentice-Hall, Englewood Cliffs, NJ, USA, 1964.

[2] J.F. Traub, *Iterative Methods for the Solution of Equations*, Prentice Hall, New York, 1964.

[3] Petkovic, M.S., Neta, B., *Multipoint Methods for Solving Nonlinear Equations*, Elsevier, Amsterdam-Boston- Heidelberg, 2013.

[4] Cordero, A., Lotfi, T., Bakhtiari, P., Torregrosa J., *An efficient two parametric family with memory for nonlinear equations*,Numer Algor, **DOI 10.1007/s** (2014) 11075-014-9846-8.

[5] Cordero, A., Lotfi, T., Torregrosa, J.R., Assari, P., Mahdiani, K., *Some new bi-accelerator two-point methods for solving nonlinear equations*, Comp. Appl. Math, **DOI 10.1007/s** (2014) 40314-014-0192-1.

[6] Lotfi, T., Soleymani, G., Ghorbanzadeh, M., Assari, P., *On the construction of some tri-parametric iterative methods with memory*, Numer Algor, **DOI 10.1007/s** (2015) 11075-015-9976-7.

[7] Wang, X., Zhang, T., *A new family of Newton-type iterative methods with and without memory for solving nonlinear equations*, Calcolo, **DOI 10.1007/s** (2014) 10092-012-0072-2.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

Optimal Iterative Methods for Finding Multiple Roots of Nonlinear Equations using Free Parameters

Fiza Zafar[1], Alicia Cordero[2], Quratulain[1] and Juan R. Torregrosa[2]

[1] *Centre for Advanced Studies in Pure and Applied Mathematics, Bahauddin Zakariya University, Pakistan*

[2] *Instituto de Matemáticas Multidisciplinar, Universitat Politècnica de València, Spain*

emails: fizazafar@gmail.com, acordero@mat.upv.es, quratulainrana54@gmail.com, jrtorre@mat.upv.es

Abstract

In this paper, we propose a family of optimal eighth order convergent iterative methods for multiple roots with known multiplicity with the introduction of two free parameters and three univariate weight functions. Also numerical experiments have been applied to a number of test equations for a special scheme from this family that satisfies the conditions given in this paper.

Key words: Nonlinear equations, Optimal iterative methods, Multiple root
MSC 2000: 65H05, 65H10

1 Introduction

Newton's method for multiple roots is quadratically convergent for every root. It determines simple or multiple zeros of a non-linear equation. If the given function has only a simple zero, then Newton's method converges quadratically to the exact solution. In past, it was very difficult to construct a higher-order optimal multi-point scheme for multiple zeros of the given function with multiplicity $m \geq 1$. Nowadays, with the digital computer, advanced computer arithmetics, software and symbolic computation, the construction of higher-order optimal multi-point methods has become easier. Many researchers presented optimal fourth-order iterative methods for multiple zeroes like Li et al. [7] in 2009, Sharma and Sharma [9] in 2010, Sharifi et al.[8] in 2012, Liu and Zhou [6] and Zhou et al.[12] in 2013, Thukral [11] in 2014, Hueso et al.[5] in 2015 and Behl et al. [1] in 2016. In recent years, at most

 ISBN: 978-84-617-8694-7

sixth-order convergence method has been given for finding multiple zeros that can be found in the available literature. There are only three multi-point iterative schemes with sixth-order convergence for multiple zeros. First one was proposed by Thukral [11] and other two were presented by Geum et al.[3]. In 2013, Thukral [10] presented a multi-point iterative method with sixth-order convergence, which is given by

$$y_n = x_n - m\frac{f(x_n)}{f'(x_n)}, \tag{1}$$

$$z_n = x_n - m\frac{f(x_n)}{f'(x_n)}\sum_{i=1}^{3}\left(\frac{f(y_n)}{f(x_n)}\right)^{\frac{i}{m}}, \tag{2}$$

$$x_{n+1} = z_n - m\frac{f(x_n)}{f'(x_n)}\left(\frac{f(z_n)}{f(x_n)}\right)^{\frac{1}{m}}\left[\sum_{i=1}^{3}\left(\frac{f(y_n)}{f(x_n)}\right)^{\frac{i}{m}}\right]^2. \tag{3}$$

In 2015, Geum et al.[3], have given the following two-point sixth-order iterative scheme:

$$y_n = x_n - m\frac{f(x_n)}{f'(x_n)}, m > 1, \tag{4}$$

$$x_{n+1} = x_n - Q(p_n, s_n)\frac{f(x_n)}{f'(x_n)}, \tag{5}$$

where $p_n = \sqrt[m]{\frac{f(y_n)}{f(x_n)}}$, $s_n = \sqrt[m-1]{\frac{f'(y_n)}{f'(x_n)}}$ and $Q : C^2 \to C$ is holomorphic function in the neighborhood of origin (0,0). In 2016, Geum et al. [4], have again proposed a three-point iterative scheme with sixth-order convergence for multiple zeros. The proposed scheme was based on weight functions, which can be seen in the following expression:

$$y_n = x_n - m\frac{f(x_n)}{f'(x_n)}, m > 1, \tag{6}$$

$$w_n = x_n - mG(p_n)\frac{f(x_n)}{f'(x_n)}, \tag{7}$$

$$x_{n+1} = x_n - mK(p_n, t_n)\frac{f(x_n)}{f'(x_n)}, \tag{8}$$

where $p_n = \sqrt[m]{\frac{f(y_n)}{f(x_n)}}$, $t_n = \sqrt[m]{\frac{f(w_n)}{f(x_n)}}$ and $G : C \to C$ is analytic in a neighborhood of 0 and $K : C^2 \to C$ is holomorphic in the neighborhood of (0,0). All of the above three schemes (1)-(6) require four function evaluations in order to produce sixth-order convergence. The iterative method (4) has one drawback that it does not work for simple zeros (i.e. for $m = 1$). Moreover, there does not exist any optimal scheme greater than fourth-order convergence. So, we propose an optimal eighth-order convergent iterative method for multiple root of a nonlinear equation. The main reason of this proposed method is to present a new higher-order optimal scheme for finding simple as well as multiple zeros of nonlinear equations.

 ISBN: 978-84-617-8694-7

The rest of the paper is organized as follows: In Section 2, we propose a new family of optimal eighth-order iterative methods to find multiple roots of nonlinear equation and discuss its convergence analysis, defining an special case whose performance will be checked in Section 4, in comparison with some existing sixth-order ones. Concluding remarks are given in Section 5.

2 Construction of the Scheme

This section is devoted to the construction and convergence analysis of this proposed scheme with the main theorem. So, we propose a new eighth-order scheme for a known multiplicity $m \geq 1$ of the desired multiple zero as follows

$$\begin{aligned} y_n &= x_n - m\frac{f(x_n)}{f'(x_n)}, \\ z_n &= x_n - mu_nH(u_n)\frac{f(x_n)}{f'(x_n)}, \\ x_{n+1} &= x_n - u_nv_n(A_2 + A_3u_n)P(v_n)G(w_n)\frac{f(x_n)}{f'(x_n)}, \end{aligned} \tag{9}$$

where $A_2, A_3 \in R$ are three free parameters and the weight functions $H : \mathbb{C} \to \mathbb{C}$, $P : \mathbb{C} \to \mathbb{C}, /G : \mathbb{C} \to \mathbb{C}$ are analytic function in the neighborhood of 0 with

$$u_n = \left(\frac{f(y_n)}{f(x_n)}\right)^{\frac{1}{m}}, \quad v_n = \left(\frac{f(z_n)}{f(y_n)}\right)^{\frac{1}{m}}, \quad w_n = \left(\frac{f(z_n)}{f(x_n)}\right)^{\frac{1}{m}},$$

are the variables of the weight functions. It is worthy to note that we will obtain well known King's family of fourth-order iterative methods for $m = 1$ with the help of first two substeps. In the next Theorem 1, we demonstrate that the order of convergence of the proposed scheme will reach at optimal eight without using additional functional evaluations.

Theorem 1 *Let us consider $x = \alpha$(say) be a multiple zero with a multiplicity $m \geq 1$ of the involved function f. In addition we assume that $f : \mathbb{C} \to \mathbb{C}$ be an analytic function in the region enclosing a multiple zero α. The proposed scheme defined by (9) has an optimal eighth-order convergence, when it satisfies the following expressions:*

$$\begin{aligned} A_2 &= 1, A_3 = 2A_2, \\ H_0 &= H(0) = 1, H_1 = H'(0) = 2, H_2 = H''(0) = -2, H_3 = H'''(0) = 36, \\ P_0 &= P(0) = P'(0), G_0 = G(0) = \frac{m}{P_0A_2}, G_1 = G'(0) = \frac{2m}{P_0A_2} \end{aligned}$$

 ISBN: 978-84-617-8694-7

and the error equation is given as:

$$\begin{aligned} e_{n+1} &= \frac{1}{48m^7P_0}c_1c_1^2(11+m)-2mc_2)((14(59+12m+m^2)P_0-3\,(11+m)^2\,P_2)c_1^4 \\ &\quad -12m(4(7+m)P_0-(11+m))c_1^2c_2+12m^2\,(2P_0-P_2)\,c_2^2+24m^2P_0c_1c_3)e_n^8 \\ &\quad +O(e_n^9). \end{aligned}$$

Now, we define an special case of our proposed scheme that will be used in the numerical section. Let us describe the following polynomial weight function directly from the proposed Theorem 1:

$$H(u)=6u^3-u^2+2u+1,\ P(v)=v+1,\ G(w)=\frac{2mw}{A_2P_0}+\frac{m}{A_2P_0} \tag{10}$$

for

$$u_n=\left(\frac{f(y_n)}{f(x_n)}\right)^{\frac{1}{m}},\ v_n=\left(\frac{f(z_n)}{f(y_n)}\right)^{\frac{1}{m}},\ w_n=\left(\frac{f(z_n)}{f(x_n)}\right)^{\frac{1}{m}}.$$

Thus, the corresponding optimal eighth-order iterative scheme is given by

$$y_n = x_n-m\frac{f(x_n)}{f'(x_n)}, \tag{11}$$

$$z_n = y_n-mu(6u^3-u^2+2u+1)\frac{f(x_n)}{f'(x_n)}, \tag{12}$$

$$x_{n+1} = z_n-muv(1+2u)(v+1)(2w+1)\frac{f(x_n)}{f'(x_n)}. \tag{13}$$

In what follows, it will be denoted by M1.

3 Numerical experiments

This section is devoted to demonstrate the efficiency, effectiveness and convergence behavior of the presented schemes. In this regard, we consider the proposed scheme M1 and choose four test problems for comparison given in the Examples 1-4. Now, we want to compare our methods with other existing methods of same domain on the basis of error per iteration and computational order of convergence COC. We have chosen sixth-order iterative methods for the comparison which is the highest-order till date for multiple zeros. Therefore, we compare the proposed methods with the family of two-point sixth-order methods, which were presented by Geum et al. in [4], out of them we consider (4) and (6) denoted by GM1 and GM2 respectively for $Q(p_n,s_n)=m(1+2(m-1)(p_n-s_n)-4p_ns_n+s_n^2),\ G(p_n)=1+p_n+2p_n^2$, and $K\,(p_n,t_n)=1+p_n+2p_n^2+(1+2p_n)t_n$.

We did our calculations with several number of significant digits (minimum 1000 significant digits) to minimize the round off error. We calculate the values of all the constants

 ISBN: 978-84-617-8694-7

$f_1(x), x_0 = -0.92, m = 4$				
$\lvert x_n - \alpha\rvert$	$\lvert x_1 - \alpha\rvert$	$\lvert x_2 - \alpha\rvert$	$\lvert x_3 - \alpha\rvert$	p_n
$GM1$	2.595063462(-8)	8.177972395(-47)	8.009985192(-278)	5.999999996
$GM2$	7.630978643(-8)	1.380544256(-43)	4.840288751(-258)	5.999999989
$M1$	5.980551468(-10)	2.106002934(-74)	4.979673785(-590)	7.999999999

Table 1: Numerical results on $f_1(x)$

and functional residuals up to several number of significant digits but due to the limitations, we display the value of errors per iterations and absolute residual errors in the function up to 9 decimal digits with exponent power in Tables 1–4.

Also we calculate the computational order of convergence,

$$p_n \approx \frac{\log|(x_{k+1} - \alpha)/(x_k - \alpha)|}{\log|(x_k - \alpha)/(x_{k-1} - \alpha)|},$$

due to Weerakoon and Fernando, in order to check the theoretical order of convergence.

Example 1 *Let us consider the following standard nonlinear test function from Geum et al. [3]*

$$f_1(x) = [cos(\frac{\pi x}{2}) + e^{1-x^2} - x - 2]^4. \tag{14}$$

The above function has a multiple zero at $\alpha = -1$ of multiplicity 4 with initial guess -0.92, which has been used to test the methods. The obtained results have been presented in Table 1, where theoretical order of convergence is reached and the higher order of convergence is observed in better precision of the results per iteration.

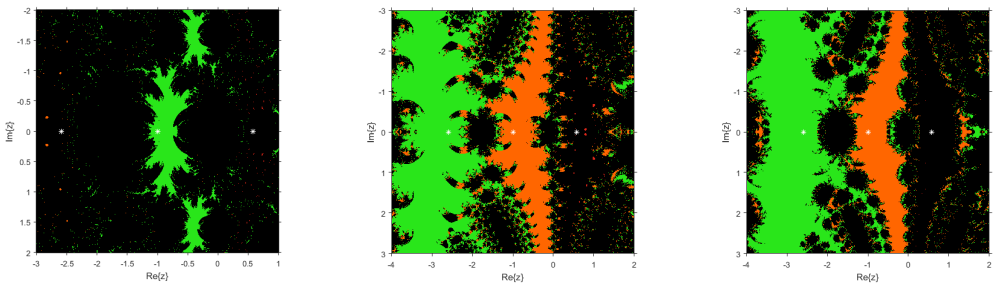
In order to check if this initial estimation is representative enough to show the performance of these methods, we present in Figure 4 the dynamical plane corresponding to each one of the methods applied on $f_1(x)$ with a mesh of 401×401 complex initial estimations. When the method is used with one of these initial guesses and it converges to one of the zeros of the nonlinear function in less than 80 iterations, it is plotted in the color assigned to the root. If it has not converged in 80 iterations, then this point is plotted in black. The programs used to make this figure appears in [2].

It is observed that the behavior is good in all the methods. The order makes the differences in two senses: the convergence, if exists, is quicker in M1 but, as it is usual, the set of convergent initial estimations decreases as the order increases.

Example 2 *We assume another standard test problem.*

$$f_2(x) = [cos(\frac{\pi x}{2}) + x^2 - \pi]^5. \tag{15}$$

 ISBN: 978-84-617-8694-7



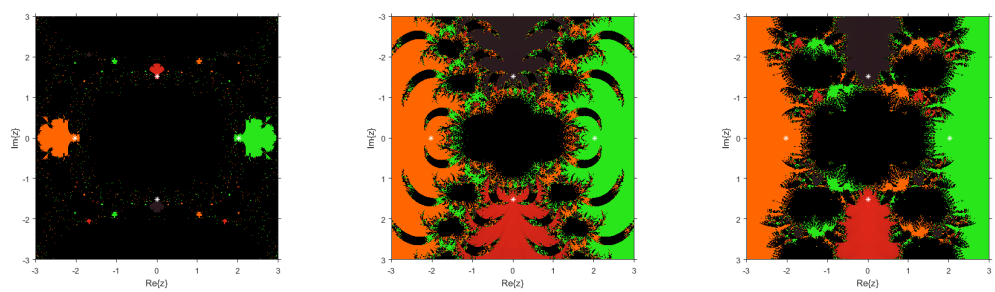
(a) M1 (b) GM1 (c) GM2

Figure 1: Dynamical planes for all the methods on $f_1(x)$

$f_2(x), x_0 = 2.5, m = 5$				
$\lvert x_n - \alpha \rvert$	$\lvert x_1 - \alpha \rvert$	$\lvert x_2 - \alpha \rvert$	$\lvert x_3 - \alpha \rvert$	p_n
$GM1$	0.522076246(-3)	8.422456030(-21)	1.488718830(-121)	5.999931812
$GM2$	0.111114189(-2)	2.533864417(-18)	3.588609342(-106)	5.999791032
$M1$	0.152779335(-3)	9.699141216(-31)	2.563327723(-240)	7.999972458

Table 2: Numerical results on $f_2(x)$

The above function has a multiple zero at $\alpha = 2.034724...$ *of multiplicity* 5 *with initial guess* 2.5. *Numerical tests on* $f_2(x)$ *appear in Table 2, showing a very good performance and the basins of attraction of the roots appear in Figure 2.*



(a) M1 (b) GM1 (c) GM2

Figure 2: Dynamical planes for all the methods on $f_2(x)$

 ISBN: 978-84-617-8694-7

$f_3(x), x_0 = 3.0, m = 2$				
$\lvert x_n - \alpha \rvert$	$\lvert x_1 - \alpha \rvert$	$\lvert x_2 - \alpha \rvert$	$\lvert x_3 - \alpha \rvert$	p_n
$GM1$	0.11849404(-5)	2.622019043(-37)	3.078026493(-221)	5.999999981
$GM2$	0.527453465(-5)	1.152582669(-32)	1.254891624(-192)	5.999999755
$M1$	1.401040040(-7)	1.304527036(-55)	7.370106153(-440)	7.999999992

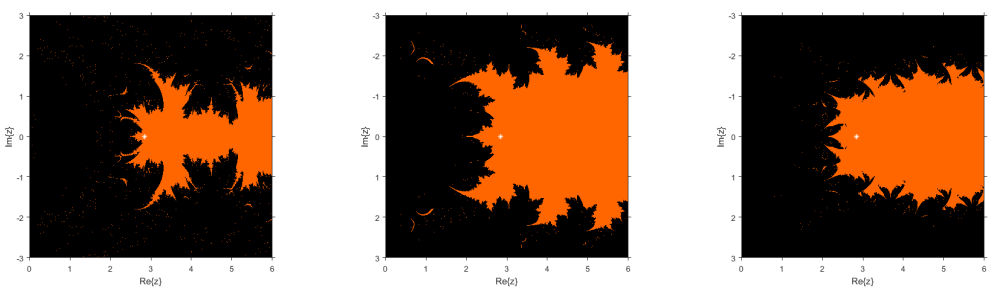
Table 3: Numerical results on $f_3(x)$

Example 3 *We assume another test problem involving exponential function as:*

$$f_3(x) = (e^x + x - 20)^2. \tag{16}$$

The above function has a multiple zero at $\alpha = 2.842438...$ *of multiplicity* 2 *with initial guess* 3.0.

In this case, the differences observed in the numerical tests correspond again to the optimality of the proposed method M1, meanwhile the basins of attraction are very good in all cases.



(a) M1 (b) GM1 (c) GM2

Figure 3: Dynamical planes for all the methods on $f_3(x)$

Example 4 *We assume another standard test problem involving trigonometric function as:*

$$f_4(x) = (cos(x) - x)^3. \tag{17}$$

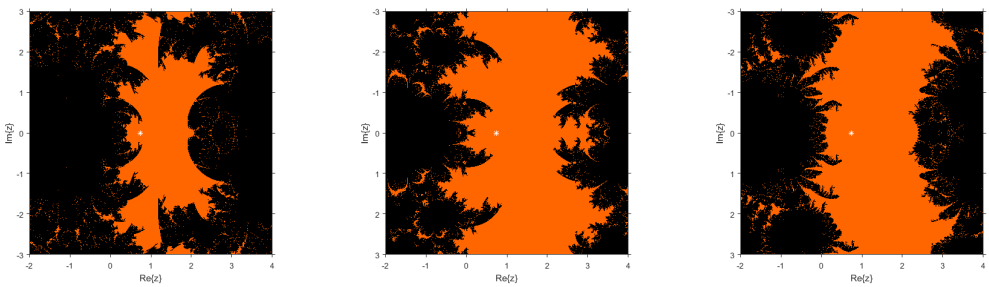
The above function has a multiple zero at $\alpha = 0.73908513...$ *of multiplicity* 3 *with initial guess* 1.0.

The numerical and dynamical behavior show good performance in all cases, both in Table 4 and in Figure 4.

 ISBN: 978-84-617-8694-7

$f_4(x), x_0 = 1.0, m = 3$				
$\lvert x_n - \alpha \rvert$	$\lvert x_1 - \alpha \rvert$	$\lvert x_2 - \alpha \rvert$	$\lvert x_3 - \alpha \rvert$	p_n
$GM1$	0.114347263(-5)	2.264182681(-38)	1.364669639(-228)	5.999999938
$GM2$	0.255308875(-5)	6.835881397(-36)	2.518668789(-213)	5.999999834
$M1$	4.905393922(-8)	4.062521585(-61)	8.990216944(-486)	7.999999997

Table 4: Numerical results on $f_4(x)$



(a) M1 (b) GM1 (c) GM2

Figure 4: Dynamical planes for all the methods on $f_4(x)$

4 Conclusion

We have proposed an optimal eighth-order family of iterative methods for solving nonlinear equations for multiple roots with known multiplicity. The family of methods include two free parameters and three weight functions involving function-to-function ratio. The methods involve only one derivative evaluation. Moreover, numeric and graphical tests show that the proposed method has good performance as compared with other similar methods.

References

[1] R. Behl, A. Cordero, S.S. Motsa, J.R. Torregrosa, V. Kanwar, *An optimal fourth-order family of methods for multiple roots and its dynamics*, Numer. Algor. **71**(4) (2016) 775–796.

[2] F. Chicharro, A. Cordero, J.R. Torregrosa, *Drawing dynamical and parameters planes of iterative families and methods*, The Scientific World Journal **2013** (2013) Article ID 780153.

 ISBN: 978-84-617-8694-7

[3] Y.H. Geum, Y.I. Kim, B. Neta, *A class of two-point sixth-order multiple-zero finders of modified double-Newton type and their dynamics*, Appl. Math. Comput. **270** (2015) 387–400.

[4] Y.H. Geum, Y.I. Kim, B. Neta, *A sixth-order family of three-point modified Newton-like multiple-root finders and the dynamics behind their extraneous fixed points*, Appl. Math. Comput. **283** (2016) 120–140.

[5] J.L. Hueso, E. Martinez, C. Teruel, *Determination of multiple roots of nonlinear equations and applications*, J. Math. Chem. **53** (2015) 880–892.

[6] B. Liu, X. Zhou, *A new family of fourth-order methods for multiple roots of nonlinear equations*, Non. Anal. Model. Cont. **18**(2) (2013) 143–152.

[7] S. Li, X. Liao, L. Cheng, *A new fourth-order iterative method for finding multiple roots of nonlinear equations*, Appl. Math. Comput. **215** (2009) 1288–1292.

[8] M. Sharifi, D.K.R. Babajee, F. Soleymani, *Finding the solution of nonlinear equations by a class of optimal methods*, Comput. Math. Appl. **63** (2012) 764–774.

[9] J.R. Sharma, R. Sharma, *Modified Jarratt method for computing multiple roots*, Appl. Math. Comput. **217** (2010) 878–881.

[10] R. Thukral,*Introduction to higher-order iterative methods for finding multiple roots of nonlinear equations*,J. Math. **2013** (2013) Article ID 404635, 3 pages http://dx.doi.org/10.1155/2013/404635.

[11] R. Thukral, *A new family of fourth-order iterative methods for solving nonlinear equations with multiple roots*, J. Numer. Math. Stoch. **6**(1) (2014) 37–44.

[12] X. Zhou, X. Chen, Y. Song, *Families of third and fourth order methods for multiple roots of nonlinear equations*, Appl. Math. Comput. **219** (2013) 6030–6038.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

Robust a posteriori error estimation for a weak Galerkin finite element discretization of Stokes equations

Xiaobo Zheng and Xiaoping Xie

School of Mathematics, Sichuan University, Chengdu 610064, China

emails: zhengxiaobosc@yahoo.com, xpxie@scu.edu.cn

Abstract

This paper proposes a robust residual-based a posteriori error estimator for a weak Galerkin finite element method for the Stokes equations in two and three dimensions. The estimator consists of two terms. The first term characterizes the difference between the L^2-projection of the velocity approximation on the element interfaces and the corresponding numerical trace, and the second term is related to the jump of the velocity approximation between the adjacent elements. We show that the estimator is reliable and efficient through two estimates of global upper and global lower bounds, up to two data oscillation terms caused by the source term and the nonhomogeneous Dirichlet boundary condition. The estimator is also robust in the sense that the constant factors in the upper and lower bounds are independent of the viscosity coefficient. Numerical results are provided to verify the theoretical results.

Key words: the Stokes equations, weak Galerkin method, a posteriori error estimator.

1 Introduction

Let $\Omega \subset \mathbb{R}^d$ $(d = 2, 3)$ be a bounded polygonal or polyhedral domain. We consider the following Stokes problem: find the velocity $\boldsymbol{u}$ and the pressure p such that

$$\begin{aligned} -\nu\Delta\boldsymbol{u} + \nabla p &= \boldsymbol{f}, && \text{in } \Omega, \\ \nabla \cdot \boldsymbol{u} &= 0, && \text{in } \Omega, \\ \boldsymbol{u} &= \boldsymbol{g}, && \text{on } \partial\Omega. \end{aligned} \tag{1.1}$$

ISBN: 978-84-617-8694-7

where $\nu > 0$ denotes the viscosity coefficient, $\boldsymbol{f} \in [L^2(\Omega)]^d$ is the body force, and $\boldsymbol{g}$ satisfies the compatibility condition $\int_{\partial\Omega} \boldsymbol{g} \cdot \boldsymbol{n} = 0$, with $\boldsymbol{n}$ being the unit outward vector normal to the boundary $\partial\Omega$.

The development of reliable and efficient a posteriori error estimators for finite element discretizations of the Stokes problem has become an active research area over the last several decades (see, e.g. [17, 1, 2, 18, 8, 4, 13, 10, 3, 15, 9, 11, 7], and the references therein). In [17], two a posteriori error estimators were presented for the mini-element based on the residual of the finite element solution and on the solution of local problems. Some related results can be found in [1, 2]. A posteriori error estimators were analyzed for non-conforming finite element approximations in [18, 8, 9], for discontinuous Galerkin methods in [13, 15], and for dual mixed finite element methods in [10, 3]. In [11] a unified framework for a posteriori error estimation for the Stokes problem was developed based on H^1-conforming velocity reconstruction and H(div)-conforming, locally conservative flux (stress) reconstruction. We also refer to [4, 7] for a posteriori error analysis for quasi-Newtonian fluid flows.

In [20] a weak Galerkin(WG) finite element method was presented for the Stokes equations (1.1) in the primary velocity-pressure formulation. The method uses P_k/P_{k-1} $(k \geq 1)$ discontinuous finite element combination for the velocity and pressure, with the velocity element being enhanced by polynomials of degree $k-1$ on the interface of the finite element partition. The usual gradient and divergence operators are implemented as distributions in properly-defined spaces. Optimal-order error estimates were established for the corresponding numerical approximation in various norms. We refer to [5, 21] for another two classes of WG methods for (1.1) and to [22] for a divergence-free WG method for the quasi-Newtonian Stokes flows. We note that in [6] the first a posteriori error analysis of WG methods was carried out for diffusion equations, where the residual type a posteriori error estimator is a combination of the standard conforming Galerkin finite elements and mixed finite elements.

In this paper, we develop a residual type a posteriori error estimator for the WG method in [20] for the Stokes problem (1.1) in two and three dimensions. The a posteriori error estimator for the velocity error plus the pressure error consists of two terms. The first term characterizes the difference between the L^2-projection of the velocity approximation on the element interfaces and the corresponding numerical trace, and the second term is related to the jump of the velocity approximation between the adjacent elements. We show that the estimator is reliable and efficient with two estimates of global upper and global lower bounds, up to two data oscillation terms caused by the source term and the nonhomogeneous Dirichlet boundary condition. We note that the a posteriori estimation is robust with respect to the viscosity coefficient. The main tool of our analysis is the Helmholtz decomposition for tensor fields.

The rest of the paper is organized as follows. Section 2 introduces notations and the WG scheme of [20]. Section 3 describes the a posteriori error estimator and proves its reliability and efficiency. Finally, Section 4 gives some numerical experiments.

 ISBN: 978-84-617-8694-7

Acknowledgements

This work has been partially supported by Major Research Plan of National Natural Science Foundation of China (91430105).

References

[1] Bank R E, Welfert B D. A posteriori error estimates for the Stokes equations: A comparison. Computer Methods in Applied Mechanics and Engineering, 1990, 82: 323-340.

[2] Bank R E, Welfert B D. A posteriori error estimates for the Stokes problem. SIAM Journal on Numerical Analysis, 1991, 28: 591-623.

[3] Carstensen C, Causin P, Sacco R. A posteriori dual-mixed adaptive finite element error control for Lamand Stokes equations[J]. Numerische Mathematik, 2005, 101: 309-332.

[4] Carstensen C, Funken S. A posteriori error control in low-order finite element discretisations of incompressible stationary flow problems[J]. Mathematics of Computation, 2001, 70(236): 1353-1381.

[5] Chen G, Feng M, Xie X. Robust globally divergence-free weak Galerkin methods for Stokes equations[J]. Journal of Computational Mathematics 2016, 34(5): 549-572.

[6] Chen L, Wang J, Ye X. A posteriori error estimates for weak Galerkin finite element methods for second order elliptic problems[J]. Journal of Scientific Computing, 2014, 59(2): 496-511.

[7] Congreve S, Houston P, Suli E, Thomas P. Wihler. Discontinuous Galerkin finite element approximation of quasilinear elliptic boundary value problems II: strongly monotone quasi-Newtonian flows[J]. IMA Journal of Numerical Analysis, 2013, 33: 1386-1415.

[8] Dari E, Durán R, Padra C. Error estimators for nonconforming finite element approximations of the Stokes problem[J]. Mathematics of Computation, 1995, 64(211): 1017-1033.

[9] Dörfler W, Ainsworth M. Reliable a posteriori error control for nonconforming finite element approximation of Stokes flow[J]. Mathematics of Computation, 2005, 74(252): 1599-1619.

 ISBN: 978-84-617-8694-7

[10] Farhloul M, Nicaise S, Paque L. A posteriori error estimation for the dual mixed finite element method of the Stokes problem. C. R. Acad. Sci. Paris, Ser. I, 2004, 339: 513-518.

[11] Hannukainen A, Stenberg R, Vohralk M. A unified framework for a posteriori error estimation for the Stokes problem[J]. Numerische Mathematik, 2012, 122: 725-769

[12] Kay D, Silvester D. A posteriori error estimation for stabilized mixed of the Stokes equations. SIAM Journal on Scientific Computing 1999, 21: 1321-1336

[13] Houston P, Schotzau D, and Wihler T P. hp-adaptive discontinuous Galerkin finite element methods for the Stokes problem[J]. European Congress on Computational Methods in Applied Sciences and Engineering, ECCOMAS 2004; Neittaanmaki P, Rossi P, Korotov S, Onate E, Periaux J, and Knorzer D (eds.) Jyvaskyla, 24-28 July 2004

[14] Mitchell W F. A comparison of adaptive refinement techniques for elliptic problems[J]. ACM Transactions on Mathematical Software (TOMS), 1989, 15(4): 326-347.

[15] Paul H, Schözau D, Wihler T P. Energy norm shape a posteriori error estimation for mixed discontinuous Galerkin approximations of the Stokes problem[J]. Journal of Scientific Computing, 2005, 22(1-3): 347-370.

[16] Shi Z, Wang M. Finite element methods[M]. Science Press, 2013.

[17] Verfürth R. A posteriori error estimators for the Stokes equations[J]. Numerische Mathematik, 1989, 55(3): 309-325.

[18] Verfürth R. A posteriori error estimators for the Stokes equations II non-conforming discretizations[J]. Numerische Mathematik, 1991, 60(1): 235-249.

[19] Verfürth R. A Review of A posteriori Error Estimation and Adaptive Mesh Refinement Techniques. Wiley Teubner, Chichester and Newyork (1996).

[20] Wang J, Ye X. A weak Galerkin finite element method for the Stokes equations[J]. Advances in Computational Mathematics, 2015: 1-20.

[21] Wang R, Wang X, Zhai Q, and Zhang R.. A Weak Galerkin Finite Element Scheme for solving the stationary Stokes Equations[J]. Journal of Computational and Applied Mathematics, 2016, 302: 171-185.

[22] Zheng X, Chen G, Xie X. A divergence-free weak Galerkin method for quasi-Newtonian Stokes flows [J]. Science China Mathematics, 2017, 60, doi: 10.1007/s11425-016-0354-8.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

An Approximation Algorithm for the BWC Problem

Shira Zucker[1]

[1] *Department of Computer Sciences, Sapir Academic College, Shaar Hanegev, Israel*

emails: zuckers@cs.bgu.ac.il

Abstract

Given a graph G and positive integers b and w, the black-and-white coloring problem asks about the existence of a partial vertex-coloring of G, with b vertices colored black and w white, such that there is no edge between a black and a white vertex. This problem is known to be NP-complete in general. We provide here a polynomial time approximation algorithm for the optimization version of the problem.

Key words: black-and-white coloring problem, BWC problem, approximation algorithm.

1 Introduction

The *Black-and-White Coloring (BWC) problem* is defined as follows. Given an undirected graph G and positive integers b, w, determine whether there exists a partial coloring of G such that b vertices are colored black and w vertices white (with all other vertices left uncolored), such that no black vertex and white vertex are adjacent. Such a partial coloring, if exists, is a *Black-and-White Coloring (BWC)* of the graph.

One application of the BWC problem comes from the chemical industry. A set of b samples of a product B and w samples of a product W has to be stored in $n \geq x + y$ different available places. For security reasons, due to the chemical nature of the samples and the configuration of the storing places, there are certain pairs of places that cannot contain two different types of products. The question is whether it is possible to store all samples by respecting these restrictions. By constructing a graph G that has a vertex for each storing place and an edge between each two places that are not allowed to contain two different types of products, this problem reduces to the BWC problem.

Another application is solved explicitly in [1]: Items of two data types, D_1 and D_2, are stored in a 2-dimensional table. A person would like to retrieve data of type D_1 or of type

 ISBN: 978-84-617-8694-7

D_2, but not of both. When retrieving data, we would like to allow a one-unit error in each of the table's indexes. In case of an error, we do not want a person trying to retrieve an element of type D_1 to extract an element of type D_2. An additional goal is to populate the elements of type D_1 in a way that will maximize the number of places left in the table for the elements of type D_2.

The problem was originated by Berge, who raised the following instance [4].

Problem 1.1 *Given positive integers n and $b \leq n^2$, place b black and w white queens on an $n \times n$ chessboard, so that no black queen and white queen attack each other, and with w as large as possible.*

The BWC problem has two natural corresponding optimization problems:

1. Given a graph G and a positive integer b, color b of the vertices black, so that there will remain as many vertices as possible which are non-adjacent to any of the b vertices. These latter vertices are to be colored white, and the resulting coloring is *optimal*. All other vertices are to be left uncolored.

2. Given a graph G and a positive integer b, color b of the vertices black, while all other vertices are to be colored white, in a way that there will be as few as possible edges connecting black and white vertices. In this optimization version, we should make the BWC as legal as possible, meaning, we would like to reduce the number of "erroneous" edges, i.e., edges connecting black and white vertices.

The BWC problem has been introduced in general, and proved to be NP-complete, by Hansen *et al.* [4]. In the same paper, an $O(n^3)$ algorithm for trees was given. In [2], an $O(n^2 \lg^3 n)$ algorithm for trees was given. Kobler *et al.* [5] gave a polynomial algorithm for partial k-trees with a fixed k. In [7], there is an algorithm for chordal graphs, which is $O(\chi n^3)$ time, where χ is the chromatic number of the graph.

Yahalom [6] investigated an analogous problem to that suggested by Berge, using rooks instead of queens, and gave a sub-linear algorithm to this problem. For special cases, in which the ratio between the sides of the board is an integer or close to an integer, she derived an explicit formula for the optimal solution. In [1], we investigated an analogous problem, using kings instead of queens, and provided explicit optimal solutions for the toroidal and the non-toroidal versions.

In [3] we examined several heuristic algorithms, in particular tabu search, for solving the problem in general.

The BWC problem admits a generalization for any number of colors. An *anticoloring* of a graph is a partial vertex coloring with two or more colors, in which no two adjacent vertices have distinct colors. In the general *anticoloring problem*, we are given an undirected graph G and positive integers $b_1, \ldots, b_k$, and have to determine whether there exists an anticoloring of G such that b_j vertices are colored in color j, $j = 1, \ldots, k$. We call such an anticoloring

 ISBN: 978-84-617-8694-7

a $(b_1, \ldots, b_k)$-anticoloring. Yahalom [6] noticed that it is easy to rewrite the anticoloring problem as an integer linear programming problem.

In this paper we focus on the second optimization version of the problem and present a polynomial-time approximation algorithm for it.

2 Results

Consider the following optimization version of the BWC problem,

Problem 2.1 *Given a graph G and a positive integer b, color b of the vertices black, while the rest of the vertices are to be colored white, in such a way that the number of "correct" edges, i.e, edges connecting vertices colored with the same color, is maximal.*

Note that, since all vertices are colored black or white, in case G is connected, there should be at least one "erroneous" edge.

Obviously, since the number of edges $|E|$ is given, the problem of finding a coloring with maximal correct edges is equivalent to the problem of finding a coloring with minimal erroneous edges. Thus, Problem 2.1 is equivalent to the definition we presented in Section 1.

Consider the following

Algorithm 2.1 *Given a graph $G = (V, E)$ and an integer $b \leq |V|$, choose randomly b vertices to be colored black. All other vertices are to be colored white.*

We will show

Theorem 2.2 *Algorithm 2.1 is a polynomial $\frac{1}{2}$-approximation algorithm for Problem 2.1.*

Proof:
Obviously, the algorithm has a polynomial time. Now, since the graph G contains n vertices, the probability for a vertex to be colored black is $\frac{b}{n}$. The probability for an edge to connect a black and a white vertices is

$$2 \cdot \frac{b}{n} \cdot (1 - \frac{b}{n}) = \frac{2b(n-b)}{n^2}.$$

A simple calculation shows that this function gets a maximum value where $b = \frac{n}{2}$. In this case, the maximum value is $\frac{1}{2}$. Therefore, the probability for an edge to connect a black and a white vertex (i.e., erroneous edge) is at most $\frac{1}{2}$. Surely, the probability for the opposite case, in which an edge is to be correct, is at least $\frac{1}{2}$. Therefore, the expected number of correct edges is at least $\frac{|E|}{2}$. The optimal solution contains at most $|E|$ correct edges. Thus, Algorithm 2.1 gives a $\frac{1}{2}$-approximation for Problem 2.1.

Note that this happens in the worst case. In most cases, where $b < |V|$ or $b > |V|$, the expected number of correct edges would be much greater than $\frac{|E|}{2}$, and therefore the approximation ratio would be much better.

 ISBN: 978-84-617-8694-7

3 Future work

We plan to find a better approximation algorithm for the problem, or prove that in the worst case there is no such. We also plan to find an algorithm for the online version of the problem, in which the graph is not given in advance, and we should color each vertex before the next one arrives.

References

[1] D. Berend, E. Korach and S. Zucker, Anticoloring of a family of grid graphs, *Discrete Optimization*, 5/3:647–662, 2008.

[2] D. Berend and S. Zucker, The Black-and-White coloring problem on trees, *Journal of Graph Algorithms and Applications*, 13/2:133–152, 2009.

[3] D. Berend, E. Korach and S. Zucker, Tabu Search for the BWC Problem, *Journal of Global Optimization*, 54/4:649–667, DOI: 10.1007/s10898-011-9783-1, 2012.

[4] P. Hansen, A. Hertz and N. Quinodoz, Splitting trees, *Disc. Math.*, 165/6:403–419, 1997.

[5] D. Kobler, E. Korach and A. Hertz, On black-and-white colorings, anticolorings and extensions, preprint.

[6] O. Yahalom, Anticoloring problems on graphs, M.Sc. Thesis, Ben-Gurion University, 2001.

[7] S. Zucker, The Black-and-White Coloring Problem on Chordal Graphs, *Journal of Graph Algorithms and Applications*, 16/2:261281, 2012.

 ISBN: 978-84-617-8694-7

Volume VI

Long Gas Pipeline Mathematical Modelling

Alaa Abdul-Ameer
Faculty of Engineering & IT, The British University in Dubai
PO Box 345015 Dubai - UAE
emails: alaa.ameer@buid.ac.ae

Abstract

Gas transportation via long pipelines is considered. Distributed parameter modelling with series and shunt energy dissipation and gas stream capacitance and inductance is incorporated. Hybrid analysis methods, wherein both the distributed and the lumped, concentrated elements of the pipeline system are included in the overall model, are advocated. An illustrative application study is outlined validating thereby the algebraic procedures employed.

Key words: gas, pipeline, transient, modelling, response

i. Introduction

The transportation of gas over long distances by pipelines will be considered in this contribution. This method of supply is a relatively safe, reliable and cost- effective form of conveying natural gas which is universally employed.
Constructing and the installation of gas pipelines is an expensive, labour intensive and politically sensitive operation. These networks often span remote regions, cross national boundaries and ecologically protected areas resulting in delicate, protracted negotiations.
Beyond this the running cost associated with gas pipelines is substantial. Owing to the frictional energy dissipation arising from the internal pipeline roughness, welds, joints, bends and discontinuities, there is a continuous reduction in the gas stream pressure an hence, volume gas steam flow rate. To counter this effect, centrifugal or axial flow compressors are installed at strategic locations along the pipeline restoring thereby the pressure loss and gas volume flow rate.

Due to the length of gas pipelines and the proportional pressure loss, the compressors employed operate continuously. Consequently, the running cost, maintenance and refit charges associated with this requirement, are substantial.

This problem is exacerbated by the remote locations, monitoring and operation of the compressor drive systems and gas coolers. These active devices must also respond to varying load demands with the requirement for constant delivery pressure and supply rates.

As with all continuously operated systems, any operational economy is translated into significant savings owing to the accumulating reduction in running, maintenance and delivery costs. However, to assess the energy efficiency of the system modelling procedures have to be derive enabling the dynamic system response to be investigated.

The classical theory for spatially dispersed pipeline systems results in irrational, multivariable, input- output models which are incomplete in the Laplace transform variable, see for example [1], [2]. Theoretically, it is possible to obtain the predicted, system responses from these models. However, the procedures involved do not provide simple, usable results which can be incorporated into design, analysis or optimisation investigations.

Finite element techniques may be used to assess the pipeline dynamics. With this procedure, large matrix models arise from the modelling procedure attracting thereby computational errors. Considerable speculation surrounding the computed pipeline performance may also be encountered in that the number and composition of the elements employed is unspecified.

In view of this, the focus of this contribution will be on deriving a pipeline modelling method which includes all of the salient features of classical analysis whilst avoiding the above complications.

ii. Long Pipeline Modelling Methods

In the transportation of gases via long pipeline and compressor networks, geographical dispersion is a significant feature. Owing to this, obtaining estimates for the volume flow of gas using conventional, lumped parameter theory is inappropriate.

It may be considered that using multiple lumping, with the application of finite element techniques, would be sufficient. However, this is not so, since the dynamics of spatially dispersed systems comprise a combination of travelling, stationary and reflected pressure and flow waveforms. Hence, there is no equivalent lumped, cascaded model counterpart since travelling and reflected transient components cannot materialise in lumped representations, in that all spatial dispersion effects are absent, from these models.

In any case when considering pipelines of 1-10 km in length, the matrix models, derived from fe methods, would be dimensionally very large [3]. Consequently, in

addition to the analytical disadvantages cited above, numerical computational errors which would further contaminate the results would be encountered [4].

Alternatively, with the employment of hybrid, distributed- lumped modelling an accurate modelling method is available [5], [6]. This procedure allows pipeline elements which are clearly distributed, to be modelled using distributed parameter methods. Otherwise, relatively pointwise components and sub - assemblies such as valves, compressors, bends and restrictions may be represented using lumped analysis methods without too much loss of accuracy. This allows engineering judgment to be exercised in selecting the appropriate modelling method, for each system element.

Importantly, the many boundary conditions and complexity arising from the use of distributed parameter methods universally are avoided with this approach, whilst including simple lumped parameter models for components which form the pipeline connecting elements and series- parallel branches.

Following the modelling of the individual elements of pipelines, distributed-lumped analysis allows the construction of an overall hybrid, matrix representation for the system. The final model provides a general, component identifiable, accurate, impedance- admittance realisation enabling dynamic simulation, analysis and regulator design.

iii. Series and Parallel Representations

Elements comprising an overall pipeline system may be assembled in series, parallel or in series parallel form where in each case, the steady state volume flow would be inversely proportional to the pipeline input impedance.

In the analysis herein, lumped parameter components are represented by simple frictional flow resistances R_1, R_2, R_3 etc. However, there is no need to adhere to this configuration. Any analogous two port network representation could be employed to model the lumped parameter units.

For the connecting elements, where there is energy storage, modelling using two port network analogues which contain inductance and capacitance elements, as shown in [7], could be employed.

iv. Distributed Parameter Modelling

Although the flow of gas, in pipeline systems is usually turbulent, three dimensional and non-linear, there are compelling reasons for the formulation of simple, usable models for perturbed flow changes, relative to given steady state conditions. Essentially, this type of model would enable the analysis of complex, inter connected applications to be analysed. This ideally would be via general solutions for the spatially distributed, pressure - flow relationships following input or disturbance changes. Moreover, simulation studies could be easily accommodated

using this form of representation with the advantage that regulator design exercises could now be embarked upon using existing theoretical techniques and algorithms.

In fact pioneering work, as detailed in [8], [9] and [10], showed that theoretically derived, first order, perturbed, acoustic, one dimensional approximations for the Navier-Stokes equations were available. These modelling restrictions included zero bulk modulus and radiant heat transfer effects. A continuous, homogenous medium was also assumed with no radial or axial heat transfer effects.

Further work within this framework was undertaken in [11] where a general discrete, distributed-lumped parameter representation, for linear systems was presented. Low temperature application studies using this approach were proposed in [12] where all of these representations related to the perturbed, pressure variation dynamics, relative to steady state, equilibrium conditions.

Extending this work, this contribution focuses on the distributed parameter system model shown in [1] and [2], where L_j, C_j, r_j and g_j are the pipeline system equivalent, distributed inductance, capacitance and series (longitudinal) and shunt (radial) flow resistance and conductance, per metre length of pipeline, respectively. The governing equations for this type of element, for the *j*th pipeline section are:

$$\frac{\partial p_j}{\partial x}(t,x) = -L_j \frac{\partial q_j}{\partial t} - r_j q_j(t,x) \tag{4.1}$$

and

$$\frac{\partial q_j}{\partial x}(t,x) = -C_j \frac{\partial p_j}{\partial t} - g_j p_j(t,x) \tag{4.2}$$

Following Laplace transformation, with zero initial conditions, equations 4.1 and 4.2 yield the solution, for the *j*th distributed parameter model of a system of *m* elements, of:

$$\begin{bmatrix} P_j(s,l_j) \\ P_{j+1}(s,l_{j+1}) \end{bmatrix} = \begin{bmatrix} \zeta_j^{-1}(s)(s)\mathrm{w}_j(s) & -\zeta_j^{-1}(s)\left(\mathrm{w}_j^2(s)-1\right)^{1/2} \\ \zeta_j^{-1}(s)\left(\mathrm{w}_j^2(s)-1\right)^{1/2} & -\zeta_j^{-1}(s)\mathrm{w}_j(s) \end{bmatrix} \begin{bmatrix} Q_j(s,l_j) \\ Q_{j+1}(s,l_{j+1}) \end{bmatrix} \tag{4.3}$$

where: $j = 2k+1,\ k = 0,1,\cdots m-1,$

($m =$ Number of distributed parameter elements),

$$\zeta_j(s) = \left[\left(L_j s + r_j\right)/\left(C_j s + g_j\right)\right]^{1/2}, \quad w_j(s) = \left(e^{2\Gamma_j(s)l_j}+1\right)/\left(e^{2\Gamma_j(s)l_j}-1\right)$$

and: $\Gamma_j(s) = \left[\left(L_j s + r_j\right)\left(C_j s + g_j\right)\right]^{1/2}$

Consequently, even with all of the constraints mentioned earlier the input-output relationship for a typical distributed parameter, gas pipeline network model is multivariable, irrational and is incomplete in the Laplace variable *s*. This difficulty effectively masks any correspondence between the actual system performance and

the governing equations so that extracting information from this representation is markedly impaired, see for example [13].

In this regard, of interest here, is the nature of the series impedance and shunt admittance of the infinitesimal airway element, shown in [1]. The series impedance frictional drag r_j, for example, represents the effect of the gas flow on the pressure gradient arising from shear action, at the pipeline wall, boundary layer. Contrasting this, the shunt g_j, admittance or conductance arises from compressibility effects, as shown in [14] where the frictional drag arises from varying gas path compliance, owing to cross-flow turbulence and molecular friction.
In pipeline systems, the flow impedance is principally due the entrance/ exit losses and to the r_j and g_j, distributed pipeline frictional resistance effects where:

$$\frac{1}{g_j} > r_j \tag{4.4}$$

In dimensionally "long" pipelines both the series r_j and the shunt g_j frictional factors, contribute to the overall pressure drop and diminishing volume flow characteristics. The analysis herein also confirms that the inclusion of both these dissipation mechanism is mandatory. Moreover, the per unit length energy storage parameters, as shown in [15] and [16], for a circular pipeline, diameter $2a_j$ are the gas path capacitance and inductance of:

$$C_j = \pi a_j^2 \Big/ \gamma R_g \theta_j \rho \tag{4.5}$$

and

$$L_j = 1 \Big/ \pi a_j^2 \rho, \text{ respectively, where:} \tag{4.6}$$

$$L_j >> C_j \tag{4.7}$$

for engineering applications. In view of the inequalities of equation 4.4 and 4.7 rationality may be recovered by equating:

$$\left(\frac{C_j s}{g_j}+1\right)\frac{\prod_{k=1}^{L}\left(T_{jk}s+1\right)^2}{\prod_{k=1}^{L}\left(\tau_{jk}s+1\right)^2} \cong \left(\frac{L_j s}{r_j}+1\right) \tag{4.8}$$

It should be noted that for the j^{th} section, with an appropriate choice of T_{jk} and τ_{jk} the approximation of equation 4.8, with $s = i\omega$, is accurate at:

low frequencies $\omega < r_j / L_j$, high frequencies $\omega > g_j / C_j$

and at $(2l-1)$ intermediate frequencies ω^* where:

$r_j \Big/ L_j < \omega^* < g_j \Big/ C_j$ Then since:

$$\Gamma_j(s)=\left[r_j\left(\frac{L_j s}{r_j}+1\right)\left(\frac{C_j s}{g_j}+1\right)g_j\right]^{1/2}, \tag{4.9}$$

equation 4.9 becomes, following the substitution shown in equation 4.8:

$$\Gamma_j(s)=\alpha_j\frac{\left(T_{j1}s+1\right)}{\left(\tau_{j1}s+1\right)}\frac{\left(T_{j2}s+1\right)}{\left(\tau_{j2}s+1\right)}\left(\frac{C_j s}{g_j}+1\right) \tag{4.10}$$

where: $\alpha_j=\sqrt{r_j g_j}$

Equally, since:

$$\zeta_j(s)=\sqrt{\frac{\left(L_j s+r_j\right)}{\left(C_j s+g_j\right)}} \tag{4.11}$$

then equation 4.11, with the substitution of equation 4.8, continuing with L=2 for illustration purposes, is:

$$\zeta_j(s)=\bar{\alpha}_j\frac{\left(T_{j1}s+1\right)}{\left(\tau_{j1}s+1\right)}\frac{\left(T_{j2}s+1\right)}{\left(\tau_{j2}s+1\right)},\quad \text{where } \bar{\alpha}_j=\sqrt{\frac{r_j}{g_j}} \tag{4.12}$$

The remaining important function of equation 4.2 is $w_j(s)$. Since;

$$w_j(s)=\left(e^{2l_j\Gamma_j(s)}+1\right)\Big/\left(e^{2l_j\Gamma_j(s)}-1\right) \tag{4.13}$$

then with $\Gamma_j(s)$ from equation 4.10:

$$w_j(s)=\left(e^{2l_j\sqrt{r_j g_j}\chi_j(s)}+1\right)\Big/\left(e^{2l_j\sqrt{r_j g_j}\chi_j(s)}-1\right) \tag{4.14}$$

Consequently, upon substituting for equation 4.14:

$$\left(w_j^2(s)-1\right)^{1/2}=2e^{l_j\sqrt{r_j g_j}\chi_j(s)}\Big/\left(e^{2l_j\sqrt{r_j g_j}\chi_j(s)}-1\right) \tag{4.15}$$

where in equation 4.15:

$$\chi_j(s)=\frac{\left(T_{j1}s+1\right)}{\left(\tau_{j1}s+1\right)}\frac{\left(T_{j2}s+1\right)}{\left(\tau_{j2}s+1\right)}\left(\frac{C_j s}{g_j}+1\right) \tag{4.16}$$

given by: $\chi_j\cong a_j s+b_j$ (4.17)

where from equation 4.17, expanding $\chi(s)$ for high frequencies gives:

$a_j=\left[\frac{C_j}{g_j}+\left(T_{j1}+T_{j2}\right)-\left(\tau_{j1}+\tau_{j2}\right)\right]$ and $b_j=1$

It is evident from equations 4.13, through to equation 4.17, that the distributed parameter model is now in an attractive form. The functions comprising equation 4.3 are free from origin branch point problems with each component $\zeta_j(s)$, $w_j(s)$ and $\left(w_j^2(s)-1\right)^{1/2}$ being single valued and complete, in the Laplace variable s with simple steady state values of:

$$\zeta_j(0)=\bar{\alpha}_j,\ w_j(0)=\frac{e^{2l_j\alpha_j\chi_j(0)}+1}{e^{2l_j\alpha_j\chi_j(0)}-1}\ \text{and}\ \left(w_j^2(0)-1\right)^{1/2}=\frac{2e^{l_j\alpha_j\chi_j(0)}}{e^{2l_j\alpha_j\chi_j(0)}-1}$$

Although values for the equivalent of the energy storage via the distributed capacitance and inductance per unit length in gas pipelines can be obtained with accuracy, from equations 4.5 and 4.6, the distributed, per unit length, series and shunt resistance values are more difficult to ascertain. From the theory of fluid dynamics the pressure drop due to friction is inversely proportioned to the Reynolds number, $\mathrm{Re}<2500,$ for a given gas flow velocity, gas density, ducting length and cross sectional area. For higher Reynolds numbers the empirical law of Blasius could be used to obtain estimates of frictional flow coefficients, as discussed in [17]. These estimates and those obtained from Moody diagrams [18] however, are based on steady flow conditions, in ducting of constant cross sectional area. Consequently, as in all engineering system problems, the frictional coefficient values are known with least confidence. Empirically based results, derived from direct measurement may be used if the system exists. Otherwise, upper and lower bounded values for r_j and g_j may be employed, with the system response characteristics reflecting these estimates.

v. An Overall Pipeline System Model

An overall model structure must now be derived enabling the assembly of the system matrix. For purposes of illustration, if a distributed- lumped configuration is assumed, then an appropriate system matrix can be constructed by adding consecutive distributed or lumped system descriptions and in so doing, eliminate all intermediate variables.

In this case the system model for a total of *m* distributed-lumped, interconnected sections for the system equation:

$$\left(P_1(s),0,0\cdots0\right)^T=\Omega(s)\left(Q_1(s),Q_2(s),\cdots,Q_m(s)\right) \tag{5.1}$$

where:

$$\Omega(s)=\begin{bmatrix}
\zeta_1(s)\mathrm{w}_1(s) & -\zeta_1(s)\left(\mathrm{w}_1^2(s)-1\right)^{1/2} & 0 & 0 & \cdots\\
\zeta_1(s)\left(\mathrm{w}_1^2(s)-1\right)^{1/2} & -\zeta_1(s)\mathrm{w}_1(s)-\bar{g}_{1,11}(s) & -g_{1,12}(s) & 0 & \cdots\\
0 & -g_{1,12}(s) & -\bar{g}_{1,22}(s)+\zeta_2(s)\mathrm{w}_2(s) & \zeta_2(s)\left(\mathrm{w}_2^2(s)-1\right)^{1/2} & \cdots\\
0 & 0 & \zeta_2(s)\left(\mathrm{w}_2^2(s)-1\right)^{1/2} & -\zeta_2(s)\mathrm{w}_2(s)-\bar{g}_{3,11}(s) & \cdots\\
\vdots & \vdots & 0 & 0 & 0\\
0 & 0 & 0 & 0 & 0
\end{bmatrix}$$

can be easily constructed. This impedance matrix is in respect of the distributed-lumped-distributed-lumped-system topology with $\bar{g}_{1,11}(s)$, $\bar{g}_{1,22}$ etc., being the diagonal elements of the termination matrix of equation A2.1. However, there is no

restriction on assembling alternative matrix descriptions for distributed-distributed-distributed realizations, for example, when modelling a series of purely distributed parameter pipeline elements, of varying dimensions.
As equation 5.1 shows, $\Omega(s)$ is a skew, symmetric tri-diagonal matrix enabling simple recursive procedures to be employed in the computational of the determinant, as shown in [21].

vi. Single Pipeline and Compressor

In this application a model representing a single, long pipeline and a compressor will be considered.
From the theory of section 5, the system equation 5.1 is relevant, since there is only a single distributed parameter section and a single termination, lumped resistance element. Hence:

$$\begin{bmatrix} P_1(s) \\ P_2(s) \end{bmatrix} = \begin{bmatrix} \zeta(s)\mathrm{w}(s) & -\zeta(s)\left(\mathrm{w}^2(s)-1\right)^{1/2} \\ \zeta(s)\left(\mathrm{w}^2(s)-1\right)^{1/2} & -\zeta(s)\mathrm{w}(s) \end{bmatrix} \begin{bmatrix} Q_1(s) \\ Q_2(s) \end{bmatrix} \tag{6.1}$$

where in equation 6.1, the termination relationship between the transformed pressure change $P_2(s)$ and the transformed airflow change $Q_2(s)$ is simply:

$$P_2(s) = RQ_2(s)$$

Consequently, following inversion equation 6.1 becomes:

$$\frac{\begin{bmatrix} \zeta(s)\mathrm{w}(s)+R & -\zeta(s)\left(\mathrm{w}^2(s)-1\right)^{1/2} \\ \zeta(s)\left(\mathrm{w}^2(s)-1\right)^{1/2} & -\zeta(s)\mathrm{w}(s) \end{bmatrix}}{\zeta(s)\left(\mathrm{w}(s)R+\zeta(s)\right)} \begin{bmatrix} P_1(s) \\ 0 \end{bmatrix} = \begin{bmatrix} Q_1(s) \\ Q_2(s) \end{bmatrix} \tag{6.2}$$

If the compressor unit is assumed to be relatively lumped, in comparison to the pipeline, comprising rotors and bearings, see for example [22] then:

$$P_1(s) = k_f U(s) \Big/ \left(\tau_f(s)+1\right)$$

where: k_f is the gain and τ_f is the fan time constant and $U(s)$ is the applied voltage for the electrical drive. For this particular application, the parameters are:

a = Pipeline raduis $= 0.5\mathrm{m}$, Θ = Absolute Gas Tempreture $= 313\mathrm{K}^{\circ}$, ρ = Gas density

R = Exit resistance $= 1.0\ \mathrm{Nsec/m^5}$, R_g = Characteristic gas constant $= 287\ \mathrm{J/kg\ K^{\circ}}$

l = Pipeline length $= 1000$, 5000 and 10,000 m, τ_c = Compressor time constant $= 5.0$ sec

k_c = Compressor gain $= 10\ \mathrm{kg\ sec^2/m^3v}$, γ = Adiabatic index $= 1.31$

g = Shunt conductance per meter $= 10^{-4}\ \mathrm{m^5/Nsec}$, r = Series resistance per meter $= 0.6\times10^{-4}\ \mathrm{Nsec/m^5}$

From equations 4.5 and 4.6 the gas capacitance and inductance per meter length of pipeline are:

$C = \pi a^2 / \gamma R_g \Theta \rho = 0.643 \times 10^{-4}$ m^{2}and: $L = 1/\pi a^2 \rho = 1.2835$ m^2, respectively.

From equations 4.10 and 4.12, and 4.11, respectively:

$$\bar{\alpha} = \sqrt{(r/g)} = 0.7745,\ \alpha = \sqrt{rg} = 0.7745 \times 10^{-4} \text{ and } \zeta(s) = \bar{\alpha}\left(\frac{T_1 s + 1}{\tau_1 s + 1}\right)\left(\frac{T_2 s + 1}{\tau_2 s + 1}\right)$$

Also equation 4.8 requires that for a single pipeline section where for illustration purposes, l = 2 and the section subscripts j have been dropped:

$$\left(\frac{T_1 s + 1}{\tau_1 s + 1}\right)\frac{(T_2 s + 1)^2}{(\tau_2 s + 1)^2} \cong \frac{L/r\, s + 1}{C/g\, s + 1} \tag{6.3}$$

where here: $L/r = 2.1392 \times 10^4$ sec and $C/g = 0.0643$ sec

From the Bode diagram shown in figure 3, where the frequency response curves for:

$$\left(L/r\, i\omega + 1\right)\Big/\left(C/g\, i\omega + 1\right) \text{ and } \frac{(T_1 i\omega + 1)(T_2 i\omega + 1)}{(\tau_1 i\omega + 1)(\tau_2 i\omega + 1)}$$

are shown in full and dotted lines, respectively, for the series resistance of: $r = 0.6 \times 10^{-4}$ Nsec/m^5 with a shunt admittance of $g = 10^{-4}$ m^5/Nsec.

The break frequencies selected for the Bode characteristics for $1/(L/r)$, $1/(C/g)$, $1/T_1$, $1/T_2$, $1/\tau_1$ and $1/\tau_2$ are presented in Table 1, below:

r Nsec/m^5	$1/(L/r)$ rad/sec	$1/(C/g)$ rad/sec	$1/T_1$ $1/T_2$ rad/sec	$1/\tau_1$ $1/\tau_2$ rad/sec
0.6x10^{-4}	0.4674x10^{-4}	15.552	6x10^{-4} 0.4	0.05 6.02

Table 1, Break frequencies $(g = 10^{-4}$ m^5/Nsec)

It is evident, from figure 1, that the approximations has the mid-range frequency intersection and exact low and high frequency correspondence, as stated earlier.

If now, as in equation 4.16,

$$\chi(s) = \frac{(T_1 s + 1)(T_2 s + 1)}{(\tau_1 s + 1)(\tau_2 s + 1)}\left(\frac{C}{g} s + 1\right) \tag{6.4}$$

is evaluated for the values in Table 1, then $w(s)$ and $\zeta(s)$, in equation 6.1, are fully defined and outputs may now be computed from this equation, where:

$$\frac{Q_1(s)}{P_1(s)} = \left(\zeta(s)w(s) + R\right)\Big/\left(\zeta(s)\left(w(s)R + \zeta(s)\right)\right) \tag{6.5}$$

$$\frac{Q_2(s)}{P_1(s)} = \left(w^2(s) - 1\right)^{1/2}\Big/\left(w(s)R + \zeta(s)\right) \tag{7.5}$$

where in delay form $w(s)=\left(1+e^{-2l\alpha\chi(s)}\right)\Big/\left(1-e^{-2l\alpha\chi(s)}\right)$, and $\hat{w}(s)=\left(w^2(s)-1\right)^{1/2}=\dfrac{2e^{-l\alpha x(s)}}{\left(1-e^{-2l\alpha x(s)}\right)}$

Alternatively, commensurate with the pipeline system topology, shown in figure 2, equations 7.4 and 7.5 may be written as:

$$\frac{Q_1(s)}{P_1(s)}=\frac{\left(\zeta(s)w(s)+R\right)}{\zeta(s)\left(w(s)R+\zeta(s)\right)} \tag{7.6}$$

and

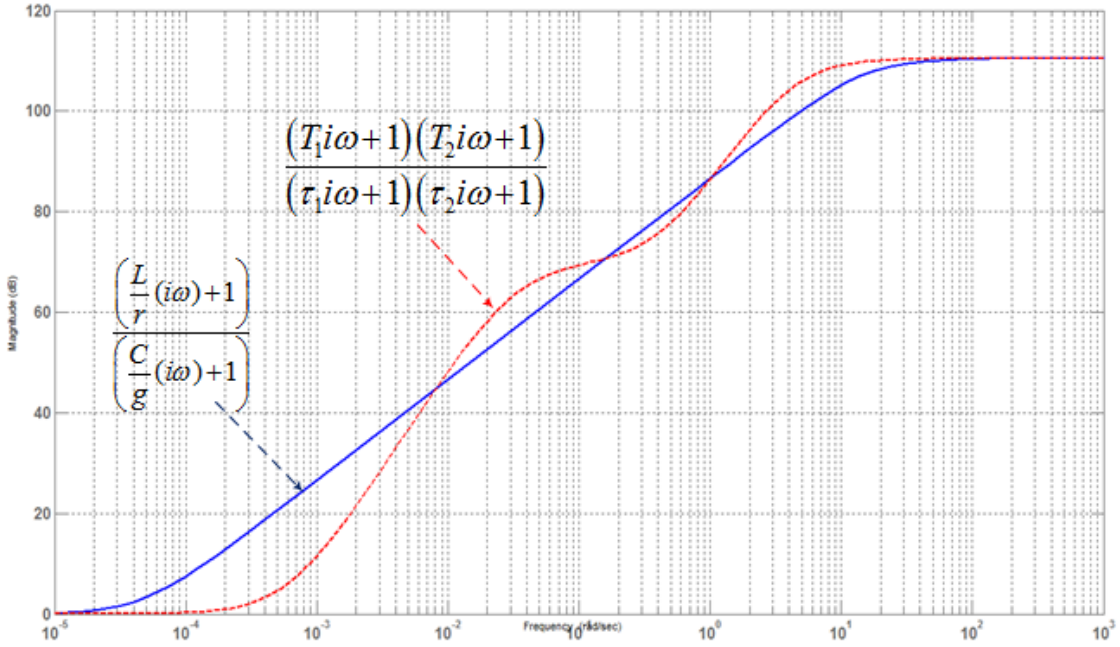


Figure 1, Bode Diagram Modulus for $\frac{\left(\frac{L}{r}(i\omega)+1\right)}{\left(\frac{C}{g}(i\omega)+1\right)}$ **and** $\frac{(T_1 i\omega+1)(T_2 i\omega+1)}{(\tau_1 i\omega+1)(\tau_2 i\omega+1)}$

$$\frac{Q_2(s)}{Q_1(s)}=\frac{\zeta(s)\left(w^2(s)-1\right)^{1/2}}{\left(\zeta(s)w(s)+R\right)} \tag{7.7}$$

The block diagram for the representation given by equations 7.6 and 7.7 is as shown in figure 2, which is in series form and whereas equations 7.4 and 7.5 provide the parallel equivalent realisation, for this system model.

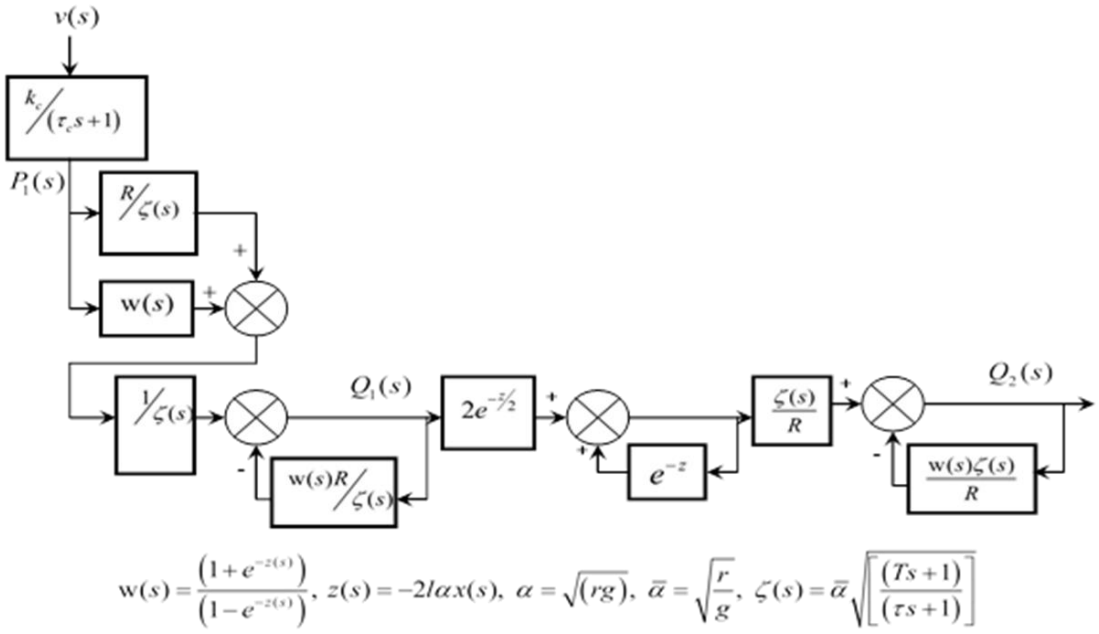


Figure 2, Distributed-Lumped Parameter Series Representation Block Diagram for a Pipeline and Compressor, *l* = 1000, 5000 and 10,000 m

To simplify the simulation process it would be prudent to construct sub-system, blocks for $w(s)$ and $\hat{w}(s)$. Since, from equation 4.17:

$\chi(s) = as + b$

where the low frequency approximation is:

$$a = \frac{C}{g} + \left(T_1 + T_2\right) - \left(\tau_1 + \tau_2\right) \text{ and } b = 1, \tag{7.8}$$

so that:

$$w(s) = \frac{\left(1 + e^{-2l\alpha(as+b)}\right)}{\left(1 - e^{-2l\alpha(as+b)}\right)} \tag{7.9}$$

Following a step input change on the $w(s)$ sub-system, it is easy to show that:

$$\frac{w(s)}{s} = \frac{1}{s}\left(1 + 2\sum_{n=1}^{\infty} e^{-2nl\alpha(as+b)}\right)$$

so that the output following any arbitrary finite input change would be stable since:

$b > 0$

From the geometry of the approximation given by equation 4.8, evidently:

$\left(T_1 + T_2\right) > \left(\tau_1 + \tau_2\right)$, so that, $a > 0$

resulting in the finite time delay $e^{-2l\alpha a}$ Also, for $\hat{w}(s)$, in delay form is:

$$\hat{w}(s) = \frac{2e^{-l\alpha(as+b)}}{\left(1 - e^{-2l\alpha(as+b)}\right)} \tag{7.10}$$

then following a unit step change on this sub-system:

$\frac{\hat{w}(s)}{s} = \frac{1}{s}\sum_{n=1}^{\infty} e^{-nl\alpha(as+b)}$ and this produces a stable output response since again:

$b > 0$ and $a > 0$ gives a finite time delay $e^{-nl\alpha a}$ and attenuation $e^{-nl\alpha b}$.

In this case, in accordance with Table 1, when: $r = 0.6\times10^{-4}$ Nsec/m^5

then following division the low frequency approximation from equation 7.8 is:

$\chi(s) = 1 + 1648.9s$ and $\zeta(s) = \bar{\alpha}\left(\frac{1666.6s+1}{20s+1}\right)\left(\frac{1.5s+1}{0.166s+1}\right)$

As the pipeline length varies, the characteristic impedance $\zeta(s)$, the exit resistance R and the compressor model remain invariant. Consequently, only $w(s)$ and $\hat{w}(s)$ need to be adjusted, in the simulation model, to obtain the gas flow characteristics for any length of pipeline with the same diameter and per unit length resistance values. In this regard, substituting for $\chi(s)$ in the equations for $w(s)$ and $\hat{w}(s)$ are given by equations 7.9 and 7.10.

Hence for the 1,000 m pipeline:

$w(s) = \frac{\left(1+0.8555e^{-77.0627s}\right)}{\left(1-0.8555e^{-77.0627s}\right)}$, $\hat{w}(s) = \frac{1.8508e^{-38.5813s}}{\left(1-0.8555e^{-77.0627s}\right)}$ for the 5,000 m pipeline:

$w(s) = \frac{\left(1+0.4609e^{-385.3s}\right)}{\left(1-0.4609e^{-385.3s}\right)}$, $\hat{w}(s) = \frac{1.3578e^{-192.5794s}}{\left(1-0.4609e^{-385.3s}\right)}$ and for the 10,000 m pipeline:

$w(s) = \frac{\left(1+0.2125e^{-770.62s}\right)}{\left(1-0.2125e^{-770.62s}\right)}$, $\hat{w}(s) = \frac{0.9218e^{-385.15s}}{\left(1-0.2125e^{-770.62s}\right)}$

The distributed – lumped parameter block representation for the series configuration including the compressor unit, given by: $\frac{P_1(s)}{v(s)} = \frac{k_c}{\tau_c s+1}$, is shown in figure 1. Following unit step changes on the compressor motor voltage input, the changes in the volume flow at $Q_1(t)$ and $Q_2(t)$ are shown in figure 3 in dotted and bold lines, respectively. This figure is initially for a 1000 m long, 1.0 m diameter, distributed parameter pipeline model, with $r = 0.6\times10^{-4}$ Nsec/m^5 and $g = 10^{-4}$ m^5 / Nsec, with the remaining parameters given earlier. Upon increasing the pipeline length to 5000 m and then 10,000 m the responses, following a 1% change in the compressor motor voltage, are as shown in this figure again with dotted and bold traces for the inlet and exit volume flow transients, respectively.

vii. Conclusion

In this contribution, the theory for modelling long pipelines was presented. It was shown that accurate, unambiguous, simple models could be easily constructed for pipeline – compressor configurations with the model-simulation block diagram

requiring with more than 4 basic sub-system models, for the compressor, $w(s), \left(w^2(s)-1\right)^{1/2}$ and $\zeta(s)$ for the complete system representation.

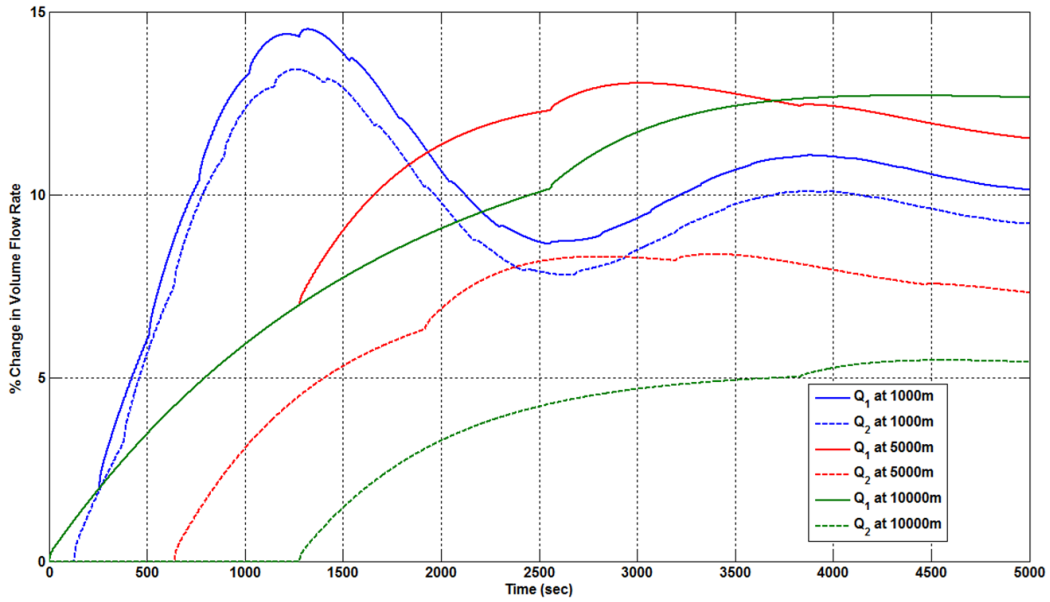


Figure 3, Percentage Changes in Inlet and Exit Pipeline Volume Flow Rates for 1m Diameter Gas Pipeline with an Exit Resistance of 1 N sec/m

The distributed parameter theory involved was also uncomplicated with the incorporation of the continuous series and shunt resistances and the gas stream capacitance and inductance, energy storage effects. Consequently, the perturbed transient volume flow responses following any input variation are ready available, from the theoretical development and simulation process, established herein.
The transient response representations for pipelines up to 10km were obtained from the application study. Owing to the generality of the theory pipelines with physical features could be accommodated.

viii. Acknowledgment

The author wishes to acknowledge the support and encouragement for this research provided by the Vice Chancellor, The British University in Dubai-UAE.

ix. References

[1] R. Schwartz and B. Freedland, *Linear Systems*, McGraw-Hill, NY,1965.
[2] Y. Takahashi, *Control and Dynamics Systems*, Addison-Wesley, NY, 1970.
[3] K. E. Rouch, and J. S. Kao, *Dynamic Reduction in Rotor Dynamic by FE Method*, Tras. ASME, Journal of Mech. Design, Vol. 102, 1980, pp. 360-367.

[4] B. Bradie, *Numerical Analysis*, Pearson Int. Inc., New Jersey, 2006.
[5] R. Whalley and A. Abdul-Ameer, *The Computation of Torsional, Dynamic Stress*, Proc IMechE, Journal of Mechanical Engineering Science, pt C, Vol. 223, 2009 pp1799-1814.
[6] R. Whalley, *The Response of Distributed-Lumped Parameter Systems,* Proc IMechE pt C, Vol. 202 (66) 1988 pp 421-429
[7] R. Whalley, *Interconnected Spatially Distributed Systems,* Trans Inst MC, Vol. 12, No 5, 1990 pp260-271
[8] A. S. Iberall, *Attenuation of Oscillatory Pressures in Instrument Lines*, Journal of Research, National Bureau of Standards, Vol. 4, 1960, R.P 2115
[9] N. B. Nichols, *The Linear Properties of Pneumatic Transmission Lines*, Inst. Soc. America Joint Automatic Control Conference, Boulder, 196.
[10] F. T. Brown, *The Transient Response of Fluid Lines*, ASME Journal of Basic Engineering, 1962, pp547-553.
[11] H. Bartlett and R. Whalley, *Analogue Solution to the Modelling and Simulation of Distributed-Lumped Parameter Systems*, Proc IMechE pt 1, Vol. 212, No 12, 1998, pp 99-114
[12] H. Bartlett and R. Whalley, *Gas Flow in Pipes and Tunnels,* Proc IMechE pt I, Vol. 209, No. 6, 1995 pp41-52.
[13] M. R. Spiegel, *Laplace Transforms*, Schaum Pub. Co., New York, 1965.
[14] J. A. Robertson and C. T. Crowe, *Engineering Fluid Mechanics*, Houghton Miffin, Boston, 1990.
[15] D. P. Eckman, *Automatic Process Control*, J. Wiley, London, 1958.
[16] W. J. Palm, *System Dynamics*, McGraw-Hill, New York, 2005.
[17] G. F. C. Rogers and Y. R. Mayhew, *Engineering Thermodynamics, Work and Heat Transfer*, Longmans Green, London, 1970.
[18] R. Gautam, *In Triccacies of Design of a Gas Pipeline*, 6th Prog. On Oil and Gas Transportation (PETROFED), New Delhi, 2009.
[19] A. W. Langill, *Automatic Control System Engineering*, Prentice-Hall, New Jersey, 1965.
[20] H. H. Rosenbrock, *Computer Aided Control System Design*, Academic Press, London, 1974
[21] S. Barnett, *Matrices, Methods and Applications*, Clarendon Press, Oxford, 1992.
[22] R. Hodder, *Screw Compressors*, Compressor and Optimisation Conference, paper 15, Aberdeen, 2008.
[23] B. Bunday, *Basic Optimisation Methods*, Edward Arnold, London, 1984.

Scrap Optimization in an Aluminium Extrusion Industry

M.F. de Almeida, Aldina Correia and Nuno Carvalho

ESTG - School of Management and Technology,
P.PORTO - Polytechnic of Porto,
CIICESI – Center for Research and Innovation in Business Sciences and Information

emails: 8150372@estg.ipp.pt, aic@estg.ipp.pt, mff@estg.ipp.pt

Abstract

The aluminium extrusion process is influenced by several parameters, such as the length and temperature of the billet, the extrusion speed, the extrusion ratio, the profiles temperature,. These parameters determine the quality of the product as well as the amount of scrap produced. In the bibliographic research, it was evident a strong relationship of dependency between the different variables being studied. This article's goal is to propose a set of improvements to reduce the percentage of scrap during the aluminium extrusion process and, consequently to increase the productivity and quality of the product commercialised by the company ADLA, S.A.

Thus, through this study, it is intended to obtain extrusion indicators that allow reduce the scrap percentage. So that in the future it is possible to settle which extrusion variables allow better results.

In a first stage, it is presented a brief introduction to the theme, the variables that influence the process are identified and the guidelines for the data collection are defined. After that, ANOVA and other statistical techniques are used to find out in which way the variables influence the scrap production.

Key words: Extrusion, Aluminium, Quality tools, Optimization, ANOVA
MSC2000: 62Pxx, 74Pxx

1. Introduction

According to Saha (2000), extrusion is a plastic deformation process in which a block of metal (billet) is forced to flow by compression through the die opening of a smaller cross-sectional area, as represented in Figure 1.

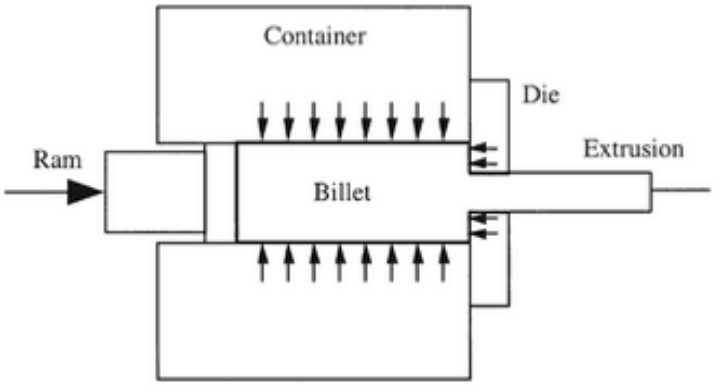


FIGURE 1 – Extrusion Process (Source: to Saha (2000))

The extrusion technology allows to produce varied geometry components, which can to be later applied in several areas (Shyamu *et al*, 2015, and Sreenivasulu and SrinivasaRao, 2016). This industrial process is used since the XIX century, but the high use of extrusion technique occurred in World War II, when different extruded aluminium profiles were used to produce aeronautical components (ABAL, 2007). Throughout the time, aluminium alloys have been widely studied and developed. Nowadays, the extruded aluminium profiles can be found in different markets such as construction, transports, motor sports, industry and structures, due to their excellent mechanical properties.

For the most demanding applications on the market is important predicting/controlling the properties of aluminum, if possible before, during and after extrusion. It allows collecting data, in order to improve the extrusion process, both in terms of optimization of equipment and inherent variables.

In order to face strong competition in all sectors and markets, it is necessary to move towards optimizing manufacturing processes, increasing competitiveness, reducing waste and maximizing profits. Thus, a key procedure is increasingly to attend the process quality improvement.

The main goal of this article is to obtain indicators that will allow us to optimise the fundamental parameters during the aluminium extrusion process, focus on the reducing of the scrap. Changing the experimental variables and checking the consequently response, is possible to understand their nature and the relationship between the variables. Our purpose is to produce improved products at the future in ADLA Aluminium Extrusion, SA company.

According to Bajimaya et. al. (2007) the properties of the extruded aluminium shapes are affected greatly by the way in which the metal flows during extrusion. Since, the metal flow is influenced by several factors, such as: temperature of billet, temperature of container, extrusion pressure, velocity of extrusion, size of billet (length and diameter) and extrusion ratio.
The length and temperature of the billet, the extrusion speed, the extrusion ratio and the profiles temperature are the variables process to have into account, in this study, because they are the actual available parameters in ADLA.
The application of statistical techniques to deal with the available variables are also an objective, in order to analyse their significance in the process.

2. Investigation hypothesis

In order to reach the proposed objective, a number of studies to examine the extrusion process and the variables that can influence it were carried out. In this section some research hypotheses has been formulated.

2.1. Billet length

Productivity can be define as the maximum amount of material, in good conditions, that can be produce per unit of time. Several studies have reported the importance of the billet length in the amount of scrap produced in the aluminium extrusion industry (Hajeeh, 2013; Tabucanon and Treewannakul, 1987).
Thus, the first hypothesis under investigation in this study is:

H1: The level of scrap produced is related to the length of the billet

2.2. Billet temperature

Several authors describe billet temperature as the key parameter in the aluminium extrusion industry, for example Flitta and Sheppard (2005) and Bajimaya, Park and Wang (2007). The temperature rise during extrusion depends on many parameters. Flitta and Sheppard (2005) affirms that the flow stress is reduced if the temperature is increased and deformation is therefore, easier, but at the same time, the maximum extrusion speed is reduced because localized temperature can lead to incipient melting temperature. Ab Rahim, Lajis and Ariffin (2015) affirms that the temperature of the billet is one of the variables to be taken into account during the extrusion process because it influences the quality of the extruded product and, therefore, it is connected to the amount of defective material. Given the below hypothesis (H2) is:

H2: The level of scrap produced is related to the temperature of the billet

2.3. Extrusion speed

The extrusion speed is a very important variable, because higher speed generates higher deformation/tension imposed into the billet inside the container. Consequently, increase the energy of the system, which will be converted to heat energy (Saha, 2000).

For a given profile extrusion, the speed of the press is considered to be one of the critical parameters. According to several authors, it is a variable that plays an important role, with respect to the surface quality of the profile and the productivity in the extrusion (Zhang et. al., 2012 and Flitta, 2004). According to Zhu, Couper and Dahle (2012) the extrusion temperature increases with increasing press speed, which directly influences the mechanical properties of the profile, as well as the surface quality of the product. However, the extrusion speed is limited by the appearance of tears on the surface of the extruded profile, which leads to an increase of non-conforming product (Arif et. al., 2002, and Li et. al., 2013). Thus, hypothesis 3 is:

H3: The level of scrap produced is related to the extrusion speed.

2.4. Extrusion ratio

The extrusion ratio (ER) of a profile is the clear indication of the amount of mechanical work that will occur during the extrusion of a given profile (Saha, 2000). This factor allows to calculate the effective deformation in the extrusion process (Extrusion Deformation = ln (ER)). It is expected that ER affects the flow stress of the metal and, consequently, also affects the maximum extrusion pressure (Karabay, Zeren and Yilmaz, 2003). It was not clear, in the literature, how the extrusion ratio influences the scrap production. However, according to some authors (Saha, 2000, and Karabay, Zeren and Yilmaz, 2003) it is well known that the increase in ER increases the maximum extrusion pressure.

Given the experience in ADLA, S.A. when the ER increased, the mechanical properties of the extruded profile are affected. The tensile strength increases as well as the hardness of the material increases, however a decrease in the percentage of elongation is observed. There is also an increase in the speed of the profile when the extrusion ratio is high, keeping the press speed constant during the extrusion process. Authors such as Abdul-Jawwad and Bashir (2011) indicate that ER directly influences the output temperature of the profile. If high temperatures influence the surface quality of the profile and its mechanical properties, we can consider that it has a relation with the levels of scrap produced. Thus, the 4 hypothesis is:

H4: The level of scrap produced is related to the extrusion ratio

2.5. Profile exit temperature

One of the main interests in the aluminium extrusion industry is to know the profile exit temperature, especially in alloys that are heat treated (6xxx series alloys). Such interest is due to the fact that the temperature of the profile at the exit of the press mouth influences in large part the obtaining of several characteristics and properties that determine the quality of the product, as the superficial quality of the product, hardness, resistance to traction and stretching (Abdul-Jawwad and Bashir, 2011).
The exit temperature is affected by several factors, namely the initial temperature of the billet, the extrusion speed, the heat generated during the extrusion due to the friction between the billet and the container (Flitta and Sheppard, 2005; Abdul-Jawwad and Bashir, 2011; Saha, 1998; and Farjad Bastani, Aukrust and Brandal, 2011). With regard to Saha (2000) the product quality is affected by the exit profile temperature, because influences the heat treatment processes and dimensional stability, causing defects during extrusion. These defects occur, according to Arif (2002), due to the fact that the high exit profile temperature leads to tears and a "*somewhat shaky*" appearance during extrusion of the profile.
In this way is predicted that the exit temperature of the profile influences the percentage of scrap produced. Then, hypothesis 5 is:

H5: The level of scrap produced is related to the profile exit temperature.

3. Results and Discussion

The experiments were carried out in an industrial extrusion press, having a maximal force of 25 MN and maximal ram speed of 20mm/s. Billet´s dimeter was Ø203,20 mm and was cuted behind the furnace.
These experiments were performed using five types of dies, with different dimensions and geometry (randomly chosen). During extrusion process several parameters were measured:

- billet length – L_{Bilet},
- billet temperature – T_{Bilet},
- extrusion speed – N,
- extrusion ratio – ER,
- profile exit temperature – T_{Exit}.

In order to verify the relationship between the extrusion process parameters similar to (Almeida, Correia and Costa, 2016) (considered as independent variables) and the percentage of scrap produced – %Scrap (dependent variable), a correlation matrix was calculated through the SPSS software.

From the obtained results, in Table 1, it was concluded that the variables *N* and T_{Exit} do not have a significant correlation with *%Scrap.* T_{Bilete} show a significant and positive correlation with the variable *%Scrap,* then we can consider that the billet temperature could be an explanatory element of a scrap increment during the aluminium extrusion process. On the other hand, the variables *ER* and L_{Bilet} has a significant and negative correlation with the *%Scrap,* then the billet length and extrusion ratio could be explanatories parameters of a scrap decrement during the aluminium extrusion process.

Table 1 – Pearson Correlations

		L_{Bilet}	T_{Bilet}	*N*	*ER*	T_{Exit}
%Scrap	Pearson Correlation	-0.398**	0.686**	0.096	-0.665**	0.043
	Sig. (2-tailed)	0.004	0.000	0.505	0.000	0.765
	n	50	50	50	50	50

*. Correlation is significant at the 0.05 level (2-tailed).

**. Correlation is significant at the 0.01 level (2-tailed).

According to this results we can consider verified the first, the second and the fourth investigation hypothesis, i.e., the level of scrap produced in this extrusion aluminium industrial process is related to the length of the billet, the temperature of the billet and the extrusion ratio.

In order to study if the relationships of the parameters under analysis affects the percentage of scrap, considering together, a multivariate linear regression model was tested in agreement with the hypothesis formulated. Then the dependent variable is the percentage of scrap and the extrusion parameters are the independent variables. In order to exclude non-significant parameters, the stepwise method is considered. The multivariate linear regression model summary is presented in Table 2.

Table 2 – Multivariate Linear Regression Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
4	0.844[d]	0.712	0.686	5.08573

[d] Predictors: *(Constant),* T_{Bilet}, *ER,* L_{Bilet}, *N*

The model has an Adjusted R Square 0.686, then approximately 68.6%. From this way we can say that the four independent variables explain 68,6% of the variance occurred on the dependent variable % Scrap.

The significance value of the ANOVA test is approximately zero, Table 3, then the estimated linear multivariate model is highly significant, i.e., the variables have a significant effect on the variation of the dependent variable (%Scrap).

Table 3 – Multivariate Linear Regression ANOVA test

Model		Sum of Squares	df	Mean Square	F	Sig.
4	Regression	2870.702	4	717.676	27.747	0.000[e]
	Residual	1163.908	45	25.865		
	Total	4034.610	49			

[e] Predictors: *(Constant)*, T_{Bilet}, *ER*, L_{Bilet}, *N*

Table 4 – Coeficientes

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
4	B	Std. Error	Beta				
(Constant)	-21.722	34.831		-0.624	0.536		
T_{Bilet}	0.166	0.071	0.259	2.346	0.023	0.526	1.903
ER	-0.131	0.024	-0.638	-5.383	0.000	0.456	2.193
L_{Bilet}	-0.022	0.006	-0.332	-3.843	0.000	0.860	1.162
N	-1.049	0.413	-0.238	-2.542	0.015	0.733	1.364

Finally, optimizing the extrusion process through better knowledge/control of the variables allows obtaining productivity gains, minimize the amount of waste and fundamentally improve the quality of the product. This is what we intend to achieve in the development of this work.

5. References

[1] Ab Rahim, S. N., Lajis, M. A., & Ariffin, S. (2015). A review on recycling aluminum chips by hot extrusion process. In *Procedia CIRP*. https://doi.org/10.1016/j.procir.2015.01.013

[2] Abdul-Jawwad, A. K., & Bashir, A. (2011). A Comprehensive Model for Predicting Profile Exit Temperature of Industrially Extruded 6063 Aluminum Alloy, 193–201. https://doi.org/10.1080/10426914.2010.505618

[3] Arif, A. F. M., Sheikh, A. K., Qamar, S. Z., Raza, M. K., & Al-Fuhaid, K. M. (2002). Product defects in aluminum extrusion and its impact on operational cost.

[4] Associação Brasileira do Alumínio - ABAL. (2007). *Fundamentos e Aplicações do Aluminio*. (Abal, Ed.), *Comissão Técnica - ABAL*. São Paulo.

[5] Bajimaya, S. M., Park, S., & Wang, G.-N. (2007). Predicting Extrusion Process Parameters Using Neural Networks. *Proceedings of World Academy of Science, Engineering & Technology*, *1*(11), 138–142.

[6] De Almeida, M. F., Correia, A., & Costa, E. (2016). Layered clays in PP polymer dispersion □ the effect of □ processing. *Journal of Applied Statistics*, 1-10.

[7] Farjad Bastani, A., Aukrust, T., & Brandal, S. (2011). Optimisation of flow balance and isothermal extrusion of aluminium using finite-element simulations. *Journal of Materials Processing Technology*. https://doi.org/10.1016/j.jmatprotec.2010.11.021

[8] Flitta, I. (2004). Simulation of aluminium extrusion process., (January). Retrieved from http://eprints.bournemouth.ac.uk/314/4/licence.txt

[9] Flitta, I., & Sheppard, T. (2005). Effect of pressure and temperature variations on FEM prediction of deformation during extrusion. *Materials Science and Technology*, *21*. https://doi.org/10.1179/174328405X29221

[10] Hajeeh, M. A. (2013). Optimizing an aluminum extrusion process. *Journal of Mathematics and Statistics Published Online*, *9*(92), 77–83. https://doi.org/10.3844/jmssp.2013.77.83

[11] Karabay, S., Zeren, M., & Yilmaz, M. (2003). Investigation extrusion ratio effect on mechanical behaviour of extruded alloy AA-6063, *135*.

[12] Li, T., Zhao, G., Zhang, C., Guan, Y., Sun, X., & Li, H. (2013). Effect of Process Parameters on Die Wear Behavior of Aluminum Alloy Rod Extrusion. *Materials and Manufacturing Processes*, *28*(3), 312–318.

https://doi.org/10.1080/10426914.2012.675536
[13] Saha, P. K. (1998). Thermodynamics and tribology in aluminum extrusion.
[14] Saha, P. K. (2000). *Aluminum Extrusion Technology*. (ASM International, Ed.). Materials Park, Ohio 44073-0002.
[15] Shyamu, V., Murali, T., Raju, M., Sapthagir, S., & Venkateswarlu, A. (2015). Analysis and Evaluation of Fractur Behaviour of Alluminium Alloy in Various Applications. *International Journal of Innovations in Engineering and Technology*.
[16] Sreenivasulu, R., & SrinivasaRao, C. (2016). Optimization of Surface Roughness and Circularity Deviation and Selection of Different Alluminium Alloys During Drilling for Automotive and Aerospace Industry. *Independent Journal of Management & Production*, *7*(2), 413–430. https://doi.org/10.14807/ijmp.v7i2.414
[17] Tabucanon, M. T., & Treewannakul, T. (1987). Scrap reduction in the extrusion process: the case of an aluminium production system. *Appl. Math. Modelling*, *11*.
[18] Zhang, C., Zhao, G., Chen, Z., Chen, H., & Kou, F. (2012). Effect of extrusion stem speed on extrusion process for a hollow aluminum profile. *Materials Science and Engineering B: Solid-State Materials for Advanced Technology*. https://doi.org/10.1016/j.mseb.2011.09.041
[19] Zhu, H., Couper, M. J., & Dahle, A. K. (2012). Effect of process variables on the formation of streak defects on anodized aluminum extrusions: An overview. *High Temperature Materials and Processes*. https://doi.org/10.1515/htmp-2012-0024

Efficient image based analysis of fruit surface: optimization of post-harvest costs.

J.A. Alvarez-Bermejo[1], D.P. Morales-Santos[2], E. Castillo-Morales[2], L. Parrilla[2] and J.A. Lopez-Ramos[3]

[1] *Department of Informatics, University of Almería*
[2] *Department of Electronics, University of Granada*
[3] *Department of Mathematics, University of Almeria*

emails: jaberme@ual.es, diegopm@ugr.es, ecastillo@ditec.ugr.es, lparrilla@ditec.ugr.es, jlopez@ual.es

Abstract

Southern Spain is a region with its economy mostly based in the agriculture sector. New techniques for optimizing crops and post harvest processes are on the cutting-edge research lines. Andalusia (south of spain) is advancing fast in greenhouse technologies, seed optimization, harvest and post-harvest procedures. The industry must, in addition, optimize their expenses during the post-harvest process to avoid food loss and to be able to compute expenses to the producers that do not fit standards. This paper develops an embedded technology to analyse post-harvest products in order to compute potential food losses.

Key words: post-harvest, food loss, embedded system, parallel image analysis, homography.
MSC2000:

1. Introduction

Andalusia (southern Spain) is an eminently agricultural region. Southern Spain's industry has to be efficient both in terms of production and economy. Harvest and post-harvest are, therefore, highly optimized processes. The costs associated with post-harvest depend on how the product comes from the field. The state in which it arrives must be attributed to the costs of the producers.

Processing companies have complex, synchronized product lines and the fact that a product line has to be pre-processed causes the line to slow down, affecting the whole process of processing, packaging and distribution.

Figure 1. Distribution lines in a processing company (part of the production pipeline).

In [1] there is an interesting definition of what is a post-harvest system, and it is defined as a system that should be thought of as encompassing the delivery of a crop from the time and place of harvest to the time and place of consumption. Paying attention and efforts in reducing the food loss, obtaining the maximum efficiency and return in all the steps involved. This implies being able to calculate costs and risks in every element of the system.

The key term *system* denotes a dynamic, complex aggregate of logically interconnected functions or operations within a particular sphere of activity. The term *chain or pipeline* highlights the functional succession of various operations but tends to ignore their complex interaction. Figure 1 describes precisely how complex these systems can be. Distribution lines in Figure 1 are carefully scheduled according to the predicted production of the next day. The production is related to a certain producer. A producer has a risk associated. This paper will develop a mean to add more information to the risk of the producer whose product is on the distribution line.

In considering the system or the agro-food chain as a whole, harvesting can be seen as the hinge, or as a ridge between the pre-harvest slope, corresponding to production activity and the post-harvest slope, extending from harvesting to consumption. These ideas are illustrated in Figure 2, which give Bourne's graphic representation of the food pipeline. It should be noted that the producer is directly related to the pre-processing stage.

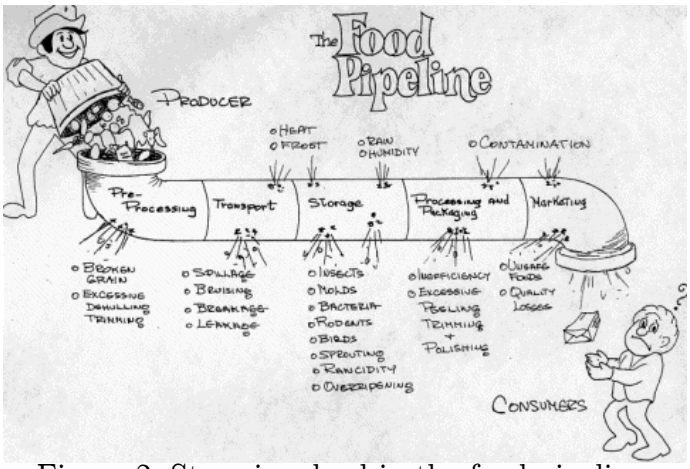


Figure 2. Steps involved in the food pipeline.

The post-harvest system encompasses a sequence of activities and operations that can be divided into two groups:

technical activities: harvesting, field drying, threshing, cleaning, additional drying, storage, processing;
economic activities: transporting, marketing, quality control, nutrition, extension, information and communication, administration and management.

Post-harvest losses

Losses are a measurable reduction in foodstuffs and may affect either quantity or quality. They arise from the fact that freshly harvested agricultural produce is a living thing that breathes and undergoes changes during post-harvest handling. Loss should not be confused with damage, which is the visible sign of deterioration, for example, chewed grain and can only be partial. Damage restricts the use of a product, whereas loss makes its use impossible.

The first distinction in agro-food losses is that between quantity and quality. Quantitative loss is a loss in terms of physical substance, meaning a reduction in weight and volume and can be assessed and measured. Qualitative loss, however, is concerned particularly with the food and reproductive value of products and requires a different kind of evaluation.
Damage is a clear deterioration in the product, e.g. broken or pitted grain, which affects more its quality than its quantity and can in the long-term result in a definite loss. Both damage and loss should be quantified in terms of weight and cost.

2. Materials and methods.

To deal with the food loss problems and to be able to calculate risks associated with each producer, it would be ideal to compute the damages or dirt that the fruit has when entering the line. This is an important point as if the product need pre-processing (such as cleaning) the production lines are slowed down. As this is the most worrying issue it is not the only problem that appears, also human intervention is planned to prepare the fruit.

The proposed solution is based in the analysis of a stream or sequence of images as captured from the CCD cameras incorporated in embedded devices. These embedded devices characterized by low powered processors, limited memory and connectivity are installed (as Figure 3 shows) in the head of production lines. A traditional issue regarding this approach is facing the computational concerns raised when processing the set of images. To this end advanced computational methods are needed.

When referring to set of images we have implemented and optimized efficient homography based analysis. As each tomato advances in the line it is rolled along its side. Cameras capture the computed unrolled surface and processes the damages or dirt on each piece. This is later sent to the cost system to advance risks in the production lines associated with each producer. The technique implemented to analyze the surface of each tomato is the image stitching - homography. In brief, an homography assume a camera takes two photographs--*a* and *b* of a set of points P_i. The camera's position is the same for the two photographs, but the orientation is different--in photograph *b* it has been rotated with the rotation matrix R. The image coordinates of the points P_i in the photographs will be related to each other with the affine transformation represented by a 3 by 3 homography matrix H

$$ {}^{a}\vec{p}_i = {}^{a}H_{b}\ {}^{b}\vec{p}_i $$

Figure 3. Production line transformation

If this matrix can be established, any point in *a* can be transformed to *b* and vice versa. That means that one of the photographs can be extended to contain everything that is seen by any of the cameras. More photographs can be added, and the extended photograph will grow to a bigger and more complete panorama image.

Image stitching has been solved in different ways and is widely commented in the literature [2]. The process of combining different images to one is used, for example to combine images taken by satellites for navigation, to generate panoramic views from impressive landscapes and huge objects [3], etc. Automatic image stitching can be automated through direct or feature-based methods. Direct methods use all image data and minimize the pixel-to-pixel dissimilarities. Feature-based methods match image features, which enables the automatic detection of the correlation of images with overlapping parts. In [4] we also find an interesting definition: image stitching is the task of combining images with overlapping parts to one big image. In [5] authors alert that this technique needs a sequence of complex computation steps, especially the execution on a mobile device can take long and consume a lot of energy. In this paper, contrary to the study exposed in [4] we have used techniques of advanced computation methods in order to accelerate the computational phase and to reduce the impact of the analysis on the energy consumption.

This procedure was applied to a set of views taken from a tomato. The results of the analysis throw a percentage of affected surface.

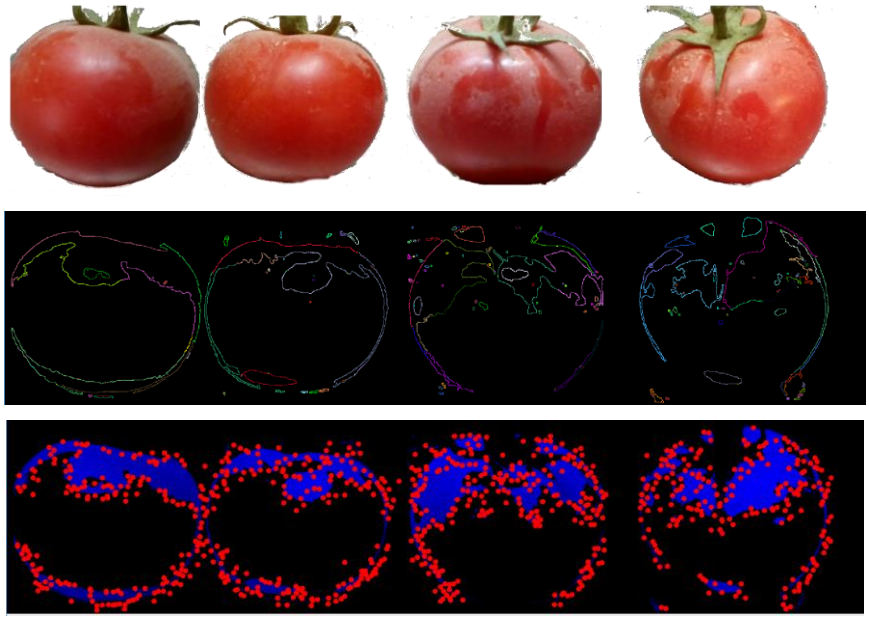


Figure 4. Image stitching and computation of the images.

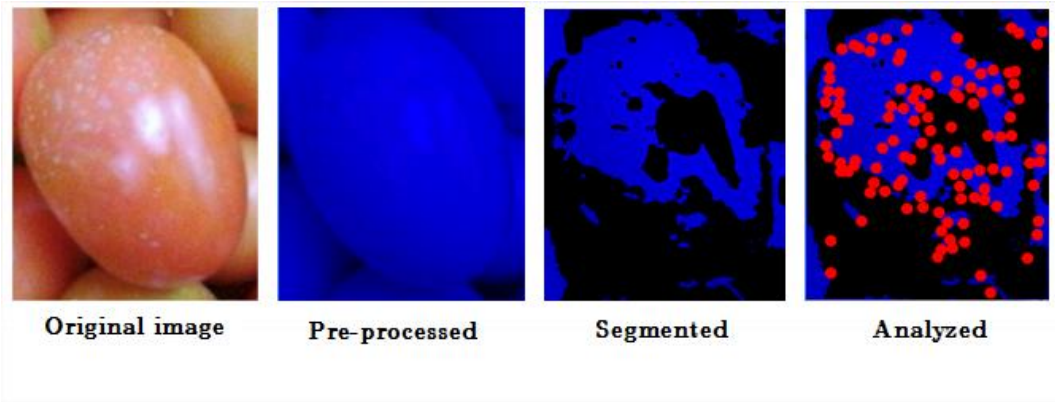
In Figure 4 different stages of the processing of the homography are shown. These stages are applied to get the percentage of the fruit with damages or dirt.

3. Implementation

This section is devoted to describe the implementation of the solution in an Android based device. We have chosen this sort of device as a sample of a embedded device that could be installed in the production lines.

To do this, let's first explain how the analysis is performed by explaining each step in the first approach, the generic algorithm. And a posteriori, we will see the variants of this generic algorithm applying parallelism to be able to improve its performance. In this case, 2 improvements have been made. The first, an improvement using Java threads (Android AsyncTasks); and the second is to use a kernel using the renderscript enhanced computing module for Android. These methods will be explained in detail later.

The stages involved in the analysis of the surface have been used in order to obtain features from the set of images; so that the percentage of damaged surface can be computed. Steps are processed sequentially.

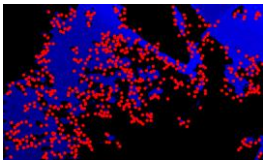
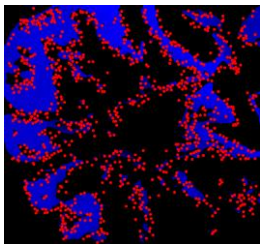
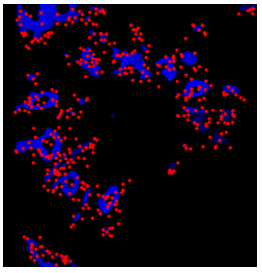
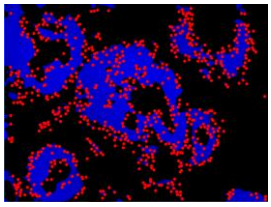
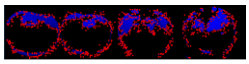


The algorithm has four elemental stages:

- Pre-processing: the space color is moved to HSV to make it light-bright independent.
- Segmentation: helps in identifying where the interesting areas are (dirt, damages,…). A threshold is added to the processing stage to select a subsect of HSV values that are of interest for the analysis. The parameters involved are H (hue) and S (saturation).

- Analysis: in the areas of interest detected in the previous stage, points of interest are calculated (points in surface where dirt is concentrated).
- Results: pixels that fall into the segmented and analyzed areas are computed in relation with the total surface and the ratio of damage is calculated as $\%\,damage = (\#\,points * \pi * \llbracket 5\text{^}2 * 100 \rrbracket\ \text{^}\,)/surface$

4. Results

Image	Analyzed image	% dirt	Lineal (ms)	Threaded (ms)	RS (ms)
		42,78	9789	7784	5376
		41,68	11976	8734	5526
		40,74	10905	8803	5317
		42,75	11224	8373	5348
		53,65	10194	8794	5252

The lineal version processes all the stages image after image. The threaded version uses Android AsyncTasks to concurrently compute the analysis steps in each image that is added to the set of the images.

The RS or RenderScript version was implemented using the C99 standard. RenderScript is a computation module operating at native level that allows to parallelize operations on the pixels of images. RS allows to use CPU and GPU as well as associate light computational operations on pixels instead of on images. In RS each pixel is not a process or a thread, operations on pixels are grouped in computational units to be processed concurrently.

As it can be seen, the RS version shows to be faster than the threaded version. As new data is added to the homography buffer, new computation is associated with the RS kernels and new operations are concurrently started. The fine grain that is achieve shows to perform better than the threaded version.

5. Conclusions

Production lines, in agricultural processing companies, are very sensible to the conditions of the products that each producer is serving. In order to calculate and to predict the risk of each producer to the production line, tools to measure and instrument these conditions are advised. We have proposed an algorithm to compute the conditions of the product (tomato in this case) that enters in the production lines.

The study that is realized is based on image analysis. Three versions of the algorithm were tested (with single images and with homography based images). The lineal version (which was not parallelized), the threaded based version which was based in assigning a thread to a new image that was added by the device and the renderscript based which uses a model where operations per pixel are grouped obtaining a fine grain and better results.

We can conclude that it is possible to conduct these tests using conventional embedded devices and commercial operating systems like Android. Also, a method to assign costs and risks to each producer can be developed as measures on the quality and damages of the product are possible and fast.

References

[1] U.N. FAO (2017, may 11) *Post-harvest system and food losses*. Retrieved from http://www.fao.org/docrep/004/ac301e/AC301e03.htm

[2] M. BROWN & D.G. LOWE (2007). *Automatic panoramic image stitching using invariant features.* International Journal of ComputerVision, **74**(1), 59–73.

[3] R. SZELISKI. *Computer vision: Algorithms and applications*. Texts in computer science. Springer, London and New York.
[4] Q. WANG, F. REIMEIER, & K. WOLTER. *Efficient Image Stitching through Mobile Offloading*. Electronic Notes in Theoretical Computer Science, **327**, 30, 125-146.
[5] C. HERBON, K. TONNIES, & B. STOCK. *Adaptive planar and rotational image stitching for mobile devices*. In Roger Zimmermann, editor, the 5th ACM Multimedia Systems Conference, 213–223.

The importance of robotics in early childhood education: first step of an intervention proposal using BeeBots

Arís, N.[1], Orcos, L.[1,3], and Magreñán, Á.A[2]

[1]*Facultad de Educación. Universidad Internacional de La Rioja*
[2] *Escuela superior de Ingeniería y Tecnología. Universidad Internacional de la Rioja*
[3]*Facultad de Educación. Universidad Nacional de Educación a distancia*

emails: nuria.aris@unir.net lara.orcos@unir.net, lorcos1@alumno.uned.es, alberto.magrenan@unir.net,

Abstract

Robotics has established itself as one of the emerging technologies that allow us to contribute an innovative and highly meaningful methodology in all stages of education. But its implementation in Early Childhood Education has a strong potential even in its incipient phase. In the first place it allows the formation of the first spatial and temporal concepts in a playful, interactive and constructive way. At the same time, another potential of Robotics in the first educational experiences is that it enhances learning according to the different types of intelligence.

This study, which has an introductory character and seeks an approximation to Robotics and its application in Early Childhood Education has been developed with a group of children of 3 years, where Robotics is implemented for the first time with the aim of establishing the impact in learning the first spatial concepts. With this, we intend to have the starting point, to be able to complete, in a future, the evolution of these students to the learning through Robotics in a longitudinal way.
Key words: Educational Robotics, Multiple Intelligences, Early Childhood Education, Meaningful Learning

1. Introduction

The globalized world in which we live requires that from early ages the children to be immersed in the advances of the Information and Communication

Technologies (ICT) since they are undoubtedly Digital Natives. These tools allow the creation of new spaces in the classrooms to be able to approach new learning processes allowing a more meaningful knowledge acquisition as long as there is a balance between the knowledge to be transferred, the methodologies used for it and the way in which technological are applied.

There is no doubt that as teachers we have to stand out that the challenges our students need require languages and dynamics designed for them in order to introduce critical and reflexive thinking and to enhance meaningful learning and; ultimately, to teach them based on the needs with which they will have to face in the future.

The stage of Early Childhood Education is ideal given the great cerebral plasticity of children in this period. All the learning that the child is acquiring being them of a cognitive, physical, motor or communicative character evolve in an integrated way. The child establishes the foundation of his identity, his emotions and the way of deciphering the reality in which he is interacting. By age 3, children are able to associate different situations with their own experiences and at the same time relate them with the emotion they can provoke. They are aware of the existence of different mental states and they have also internalized the rules of behaviour typical of their socio-cultural environment, [1]. In short, they are able to understand and attribute meaning to a given situation, all this based on previous knowledge, expectations and curiosity of their evolutionary moment. Moreover, an aspect to keep in mind in reference to the above statement is emotional education. Backett explains that school learning should be added to the purely cognitive learning emotional education in children, since both elements contribute to an integral development of the personality.

In relation to the insertion of ICT, the TPCK (Technological Pedagogical Content Knowledge) model proposed by Mishra and Koehler [5] establishes a clear connection between Content, Pedagogy and Technology allowing the integration of technology in three Theoretical, Pedagogical and Methodological levels. A special emphasis should be placed on the fact that what is important in technology is not to use it, but to know how to use it.

In the case of Early Childhood Education, free school trends are now being widely expanded, as they allow us to stage the methodologies that are key for educational systems to evolve towards open models [6]. In this sense, ICT represents a great opportunity to work towards these more open horizons [3]. Roig-Vila studies the role that ICT plays in the key aspects of Early Childhood Education. In his study, he notes that ICT represent a great opportunity to enhance aspects such as personal initiative, as they are consultation resources and through

which methodologies can be worked based on the needs of the students. Moreover, they also help students to create their own identity by making them inquiring about those aspects that appeal to them. This is very important to work the contents not only in a globalizing way, but also specifying in each of the daily situations that can arise. He also comments that ICTs can be a great help in promoting language development in these short ages due to the use of images, videos, etc. Robotics increases students' interest and participation in STEM subjects (Science, Technology, Engineering and Mathematics) and the fact of learning to program at an early age has a positive influence on improving decision-making and solving problems skills and therefore personal autonomy based on the constructivist approach in which the student is the centre of learning [4].

In Early Childhood Education, there are few publications on educational robotics and computational thinking, and especially in which the integration between the learning process and the use of robotics can be clearly seen, [2]. However, it is a fact that Robotics is increasingly introduced in the classrooms demonstrating that it fosters aspects as necessary as creativity, autonomy, learning based on trial error, and most importantly, the fact of developing the ability of being able to identify possible solutions using certain algorithms that can be transferred to different problems. In Childhood Education the most commercialized and used robots are the Bee-Bot® and Blue-Bot® and its use is mainly based on empowering students in the process of solving a series of challenges through a sequence of actions ordered through computational programming to foster logical-mathematical intelligence.

There are several aspects related with the fact that the introduction of educational robotics can be positive in this educational level, in addition to logical-mathematical intelligence. One of them is the improvement of spatial intelligence, since it promotes the development of space-time perception, having to order actions to sequence events and to get the robot to move in a desired direction. In addition, the collaborative work that implies the use of this tool is also positive to enhance respect, tolerance, socialization, interest, motivation and self-esteem because students see how the robot is managing to do the missions that they have programmed themselves. A meaningful learning based on trial-error is earned which takes the starting point in their motivation.

2. Methodology

The methodology will include two clearly differentiated and complementary parts. We will start with a study based on qualitative methodology which will be completed with quantitative cross-cut methodology. First we are oriented to the

understanding process to describe and to interpret what Robotics is and its application in the environment of the Early Childhood Education. The technique for the collection of information consists of a deep bibliographical revision that intents to understand the dimensions of the study object and its didactic application.

Secondly, data will be collected on the response of students to their participation in the session in which Robotics has been used. The technique for collecting information will be based on an evaluative questionnaire with 15 items with a response in the form of a liker scale with values in 1-5. With this, we will be able to objectively establish the learning and / or improvement obtained in the students, as a first diagnosis, in order to have a starting point for future longitudinal monitoring.

3. Experience Description

From a didactic application point of view, we will describe this experience of learning mathematics in Early Childhood Education with a Robotic educational resource. Two main learning objectives can be achieved: a) To promote the approximation and the initial learning of mathematical aspects such as sequencing, laterality and spatial notion: front, back, side and side, etc.; b) To develop multiple intelligences through the skills of teamwork, respect, curiosity and creativity. The resource used are the BeeBots which are an educational resource to work on computational thinking and favour the directional language, turns, laterality and basic spatial concepts by directing the movement of bees. They allow beginning to teach control, directional language and programming to children from the age of 3.

The experience is located at Escola Edumar in Barcelona, endowed with a Special Education Support Unit (SESU) and in which students with a great diversity of learning levels are enrolled. The first step is always based on the teacher's description of the situation to be solved. For example the first problem is to get the Robot to advance in a straight line, stop at the end and turn the head 90º to the right, 90º to the left and put it back straight. Students should take into account three basic aspects: a) Each time a button is pressed, Bee-Bot's eyes blink and a sound confirms the instruction; b) The Robot always go forward or backwards 15 cm and turn on itself 90 °. The sequence is done step by step, marking each action with light and sound. It has a memory of 40 movements; c) The X button erases the memory to start a new sequence, otherwise it will repeat the old sequence and then the new instructions.

Taking this information into account, the different groups of 3-year-old children perform the 3 phases of the project: a) They analyse the possible alternatives of

programming the robot through the technique of manipulation; b) they choose the solution they thought to be more appropriate; c) they implement it. All this generates great motivation and involvement in the children as they participate orally, manipulatively, playfully, etc. The teaching work in these activities is none other than to help the students to reflect, anticipate, rehearse and test and then rethink about the results obtained.

4. Conclusions

The development of the first stage of this study leads us to conclude that the true digital literacy must be consolidated from the earliest stages of life. This study is based only on a first experience as a starting point and we intend to follow the evolution of these students after the use of the tool.

On the other hand, it is necessary to highlight the lack of studies on the application of Robotics focused on the stage of Early Childhood Education and for this reason it is very useful to provide objective data about it. Teachers are very motivated by the potential of Robotics.

Classroom implementation of robotics-based technologies is an opportunity to innovate in the didactic applicability and the educational approach from a global point of view, due to the fact that it contemplates the didactic, methodological, and curricular dimensions.

5. References

[1] M. A. BACKETT, J. D. MAYER, R. M. WARNER.*"Emotional intelligence and its relation to everyday behavior"*. Personality and Individual Differences, *36, 1387-1402.2004.*

[2] F. BARRETO, F. BENITTI. Benitti, V. (2012). *Exploring the Educational potential of robotics in Schools: A systematic review*. Computers and Eduaction 58, 978-988. 2012.

[3] J. CUBERO. *Replanteando la tecnología educativa.* Revista comunicar, 21, 23-30. 2003.

[4] C. KIM, D. KIM, J.YUAN, R.B., HILL, P., DOCHI, C. N., THAI. (2015) *Robotics to promote elementary education pre-service teacher´s STEM engagement, learning and teaching.* Computers and Education 91, 14-31.2015.

[5] P., MYSHRA, M., KOEHLER. *Technological Pedagogical Content Knowledge: A Framework for teacher knowledge.* Teachers College Record, 108 (6), 1017-1054., 108 (6), 1017-1054. 2006.

[6] R. ROIG-VILA. *Tecnología, innovación e investigación en los procesos de enseñanza-aprendizaje.* Ediciones Octaedro, págs. 3002-3015. 2016.

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4-8 July, 2017

Solving Wave Equations on Fullerene Surfaces

James Avery[1]

[1] *Niels Bohr Institute, University of Copenhagen, Denmark*

email: avery@nbi.ku.dk

Abstract

Can we solve the electronic wave equations when there is no coordinate system?

The question arises from the wish to treat certain polyhedral carbon molecules, fullerenes and fulleroids[1,2], as two-dimensional closed surfaces. This would allow us to solve for their electronic structure on their intrinsic surface manifolds, which can be derived directly from the bond structure. The wave equation restricted to the (non-Euclidean) surface could then be solved without reference to any three-dimensional geometry of the molecule, and hence without the need for quantum chemical geometry optimization.

The resulting 2D system can potentially be solved several orders of magnitude faster than the full wave equation. But because it is a non-trivial task to find global coordinate systems for such curved surfaces, we must devise methods that can do without.

In this talk, I describe the mathematical challenges this poses, and my work in progress on solutions to overcome them.

Keywords: Fullerenes, electronic structure, discrete manifolds, non-Euclidean geometry

References

[1] P. SCHWERDTFEGER, WIRZ, LUKAS AND J.E. AVERY, *The Topology of Fullerenes*, Wiley Interdisciplinary Reviews: Computational Molecular Science, **5**, 1, 96-145, Wiley 2015. DOI: 10.1002/wcms.1207

[2] P. SCHWERDTFEGER, WIRZ, LUKAS AND J.E. AVERY, *Program Fullerene: A software package for constructing and analyzing structures of regular fullerenes*, J. Comp. Chem, **34**, 17, 1508-1526, Wiley, 2013. DOI: 10.1002/jcc.23278

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4-8 July, 2017

Topological Effects in 1-Pentagon Carbon Nanocones: Migrating Faces and Magic sizes

Adhemar Bultheel[1] and Ottorino Ori[2,3]

[1] *Department of Computer Science, KU Leuven, Belgium*
[2] *Actinium Chemical Research, Rome, Italy*
[3] *Laboratory of Renewable Energies-Photovoltaics, INCEMC, Timisoara, Romania*

emails: Adhemar.Bultheel@cs.kuleuven.be, ottorino.ori@gmail.com

Abstract

Changes in topology descriptors of one-pentagon carbon nanocones are investigated using both, direct F and dual graph F^D representations. *Topological compactness* and *topological roundness* - described by Wiener index W and topological efficiency ρ index respectively - dominate the growth of such a structure. Various chemical implications are forecasted on the basis of our purely topological model.

Key words: carbon nanocones, topological compactness, topological roundness
MSC2000: AMS Codes (none)

1. Introduction

Variations of the topological features of carbon nanocones with one pentagon at the apex (1-P NANOCONE) are investigated using both direct F and dual graph F^D representations. In this work we derived the asymptotic values for two important topological indices, the Wiener number W and the topological efficiency ρ index, improving previous numerical studies [1]. The two invariants are distance-based topological descriptors that, by taking into account the long-range structure of the graphs, describe important structural characteristics of the underlying nano-system, *i.e.* its *topological compactness* (W) and *topological roundness* (ρ).

Both indices drive the growth of such a structure when successive circles f of carbon atoms are added to it ($f \geq 0,1,2\ldots$, $f=0$ corresponding to the isolated pentagon case). When the 1-P NANOCONE is represented in the direct space (the N nodes of the graph correspond to the N carbon atoms of the chemical structure), the graph invariants follow the asymptotic curves reported in the box below.

F is made by 1 pivotal pentagon P surrounded by f concentric belts of hexagons $f = 0, 1, 2, 3, \ldots$.

Total number of hexagons is:
$n_6 = \frac{5}{2}(f^2 + f)$

Total number of faces $n_T = n_5 + n_6 = n_6 + 1$

for $f=0$ then $n_T = n_5 = 1$

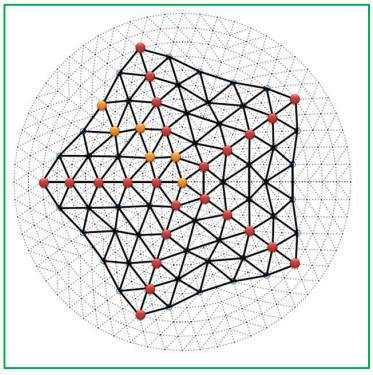
The number of nodes (carbon atoms) N obeys to the rule:
$N = 5 + 5(f^2 + 2f)$

Wiener index goes like:
$W(f) = (124f^5 + 620f^4 + 1205f^3 + 1135f^2 + 516f + 90)/6$

Other 1-P NANOCONE invariants have been derived in F and F^D.

2. Some Results

Present topological modelling is rich of original results on the chemical features of the nanocones, that are really very hard to achieve just by employing ab-initio investigations.



First of all, our topo-simulations show that the position of the most stable hexagonal face varies as a function of the nanocone size f, influencing in such a

way the stability of the whole structure (see the yellow path in the figure above). This migration follows a sequence of peculiar “jumps” in correspondence of specific *f* values.

Moreover, for certain numbers of hexagons (*magic sizes*), the conic cage results topologically more stable than graphenic fragments having similar number of atoms. Topological features make then 1-P NANOCONE suitable for generating fullerene-like structures whose presence in carbon black graphene sheets or in activated carbon has many experimental implications. The importance of fullerene-like structures also involves the interstellar carbon dust, playing in such a way a crucial role in catalyzing the formation of molecular hydrogen and other low-temperature chemical reactions.

3. References

[1] F. CATALDO, O. ORI AND S. IGLESIAS-GROTH, *Topological lattice descriptors of graphene sheets with fullerene-like nanostructures*, Mol. Sim., **36** (2010), 341.

An Ensemble Approach for *in silico* Prediction of Ames Mutagenicity

Gonzalo Cerruela García, Nicolás García-Pedrajas, Irene Luque Ruiz and Miguel Ángel Gómez-Nieto

Department for Computing and Numerical Analysis, University of Córdoba, Spain

emails: gcerruela@uco.es, npedrajas@uco.es, iluque@uco.es, mangel@uco.es

Abstract

In this paper, we propose and evaluate a method for molecular activity prediction. The proposal is based on the construction of ensembles of classifiers applying a supervised projection of the input space using Nonparametric Discriminant Analysis.

Testing these ensembles of classifiers for the *in silico* prediction of Ames mutagenicity, we have demonstrated the better behavior of our proposal with respect to classical methods.

Key words: QSAR, Ensembles of Classifiers, Mutagenicity prediction.

1 Introduction

The discovery and design of new drugs uses in vitro methods, in which laboratory experiments are performed to discover the activity of ligands with a target protein molecule. This approach is extremely expensive in terms of time and money when applied to determining activity on a large number of molecules.

Moreover, it is necessary for a drug that reaches market to possess perfect ADMET properties (absorption, distribution, metabolism, excretion, toxicity). The toxicology of drugs is a crucial research field in preclinical studies, constituting one of the leading causes of attrition in all stages of drug design.

Today, companies involved in the development of drugs lend a special interest in *in silico* prediction of toxicity to reduce time and production costs. In particular, there is great interest in the study of the mutagenic effects due to its close relationship with carcinogenicity. The most typical assay for mutagenicity is a

short-term bacterial reverse mutation assay proposed by Ames [1] that detects a large number of compounds that can induce genetic damage and frameshift mutations.

Quantitative studies of the structure-activity relationship (QSAR) using machine learning techniques is a useful *in silico* prediction method. QSAR is basically used to study biological activities with various properties associated with structures, which is useful in explaining how the structural features of a molecule influence biological activities.

Several studies have shown that it is not possible to apply a unique matching learning algorithm for QSAR applications [2]. On the contrary, in many cases, the use of combined prediction models is necessary to increase the applicability model beyond what is achievable with a single algorithm.

In the last decades, several algorithms have been used to construct QSAR models. For instance, in [3-6], the use of support vector machines (SVM) was proposed; in [7-10], models were based on decision trees, and in [11-13], the models were based on artificial neural networks.

Other authors [14, 15] proposed a combination of classifiers using an ensemble of different independent classifiers including support vector machines, decision trees, k-nearest neighbor (k-NN) and naive Bayes.

Ensemble methods are useful tools for modeling complex relationships between independent and dependent variables [16]. These methods are designed to overcome problems with single predictors and avoid over-fitting the training data [17]. Ensembles based on random subspaces of descriptors by averaging k-NN or SVM models have been constructed [18]. Though difficult to interpret, ensembles models can cope with highly nonlinear and difficult problems where simpler methods suffer from high bias due to their limited power.

Thus, in [2], different ensembles of decision trees such as boosting [19], bagging [20], and random forests [21] were compared with respect to standard classification methods. The authors concluded that a Random Forest (RF) model consistently improved classification accuracy with respect to a single tree over several QSAR applications.

The goal of this work is to extend a method to construct ensembles of classifiers [22] to the prediction of Ames mutagenicity. Our proposal is based on the use of Nonparametric Discriminant Analysis (NDA) as a linear supervised projection method used for constructing ensembles, comparing different classification methods such as C4.5, k-NN and SVM.

Also, we have devoted special attention to the chemical data used by the tested methods. Although QSAR statistical models have an acceptable predictive capacity, it is also necessary to pay attention to other necessary steps such as calculation of the molecular descriptors representing the essential information of a molecule in terms of their physicochemical properties such as constitutional,

electronic, geometric, hydrophobic, quantum, lipophilicity, solubility, and structural topology.

Several algorithms and software applications can be used to encode the chemical information in molecular descriptors, which will be used as independent variables to construct QSAR models [23-28]. In this work, we use the RDKit toolkit [28] that provides a wide range of molecular types of descriptors, including constitutional, topological, hybrid and van der Waals surface area.

The paper is organized as follows: Section 2 describes the characteristics of the benchmark dataset, and the method for constructing ensembles of classifiers; Sections 3 and 4 study the classification performance when applying ensembles with respect to standard methods and discusses the results; finally, Section 5 provides the conclusions of our work.

2 Materials and Methods

2.1 Data set and molecular representation

In order to obtain a technically feasible prediction model for mutagenicity, in [3], a clearly defined reference benchmark was proposed. This Ames mutagenicity benchmark dataset was derived from information contained in CCRIS [29], Helma et al. [30], Kazius et al. [31], Feng et al. [32] and Judson et al. [33].

The final dataset contains 6512 compounds as canonical smiles molecular representation with the corresponding Ames test results and references; this benchmark can be downloaded from [34]. The class's distribution behaves as follows: 3503 molecules correspond to the positive class and 3009 to the negative.

The predictive accuracy for the algorithms proposed in this paper was tested using the 177 descriptors generated by the RDKit toolkit software. The RDKit user manual contains a complete list of the descriptors and the corresponding bibliographic references [28].

2.2 Constructing ensembles of classifiers

In a previous work [22] the construction of ensembles projecting the input variables in a way that facilitated the accurate classification of misclassified instances without damaging the overall performance of the ensemble was proposed. This projection was implemented using the hidden layer of a multilayer perceptron [35].

That approach incorporated the advantages of boosting without its main drawbacks. The construction of the projection only accounted for instances that have been misclassified by a previous classifier, which permitted the new classifier to focus on difficult instances. However, as this classifier received a uniform distribution of the training instances, the sensitivity to noise and the effect of small datasets were greatly reduced. This method uses the whole input space reduced diversity and made the projection difficult to obtain, especially in the case of many inputs and few misclassified instances. To avoid this effect, a new method was

developed [36] in which the supervised projection uses random disjoint subspaces instead of all inputs and the neural network projection is substituted by a NDA linear supervised projection.

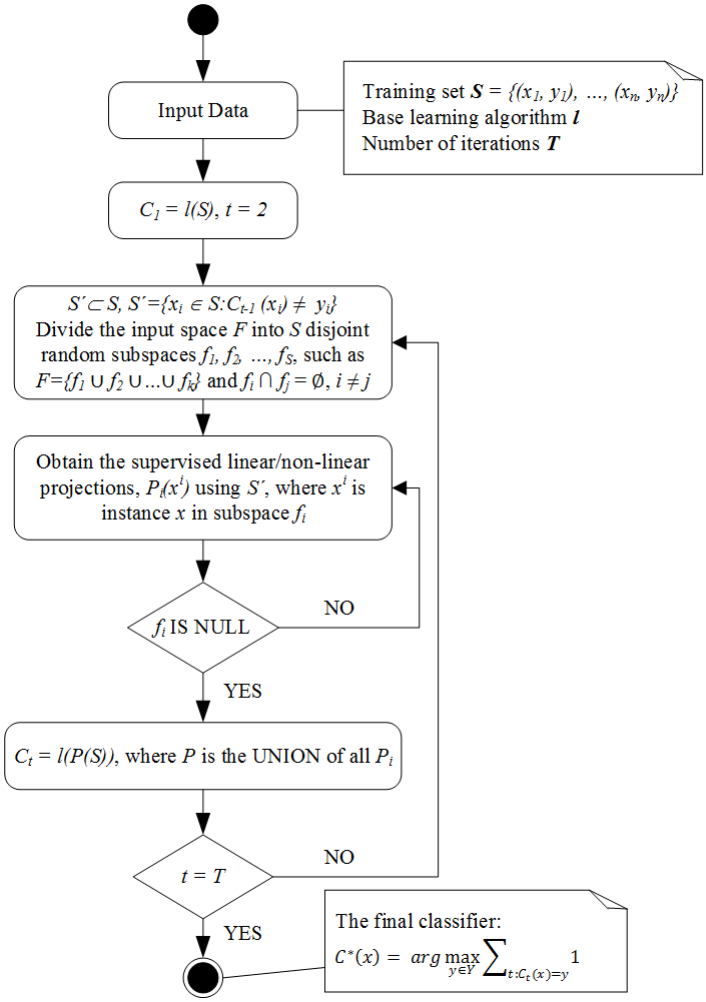


Figure 1: Algorithm for constructing the ensemble of classifiers using subspace linear/non-linear supervised projections

Figure 1 shows the proposed algorithm. At step *t*, the projection considers only the subset of instances $S' \subset S$ misclassified by the classifier added at step *t-1*. To train the classifier at step *t*, the input space is divided into several random disjoint subspaces. The instances in S' are used to obtain a supervised projection that focuses on misclassified instances. The original training set is subsequently projected using these transformations, and the next classifier is trained on these projected spaces using a uniform distribution of the instances. The projection is always performed using as inputs the original variables.

From the point of view of supervised projection algorithms, subspace projection has several advantages compared to projecting the whole input space. The algorithm is faster and more stable, as many inputs with few instances usually yield to ill-posed problems.

Our proposal is based on the use of Nonparametric Discriminant Analysis (NDA) as the linear supervised projection method. NDA is an alternative to the Linear Discriminant Analysis (LDA), which eliminates its two major problems: a) the Gaussian assumption over the class distribution in the dataset, b) limitation in dimension of the subspaces by the number of classes [37].

NDA can be formulated for the multiclass case as follows:

$$S_b^{\mathrm{NDA}} = \sum_{i=1}^{L} P_i \sum_{\substack{j=1 \\ j \neq i}}^{L} \sum_{l=1}^{n_i} \frac{w_l^{(i,j)}}{n_i} D_j\left(x_l^{(i)}\right) \cdot D_j\left(x_l^{(i)}\right)^T \tag{1}$$

$$S_w^{\mathrm{NDA}} = \sum_{i=1}^{L} P_i \sum_{l=1}^{n_i} \frac{w_l^{(i,j)}}{n_i} D_i\left(x_l^{(i)}\right) \cdot D_i\left(x_l^{(i)}\right)^T \tag{2}$$

$$w_l^{(i,j)} = \frac{min\{d\left(x_l^{(i)}, x_{\mathrm{kNN}}^{(i)}\right)^{\mathrm{p}}, d\left(x_l^{(i)}, x_{\mathrm{kNN}}^{(j)}\right)^{\mathrm{p}}\}}{d\left(x_l^{(i)}, x_{\mathrm{kNN}}^{(i)}\right)^{\mathrm{p}} + d\left(x_l^{(i)}, x_{\mathrm{kNN}}^{(j)}\right)^{\mathrm{p}}} \tag{3}$$

where p is a control parameter between zero and infinity, n_i is the number of instances in class i, p_i is the *a priori* probability of class i, $d\left(x_l^{(i)}, x_{\mathrm{kNN}}^{(j)}\right)$ is the distance from $x_l^{(i)}$ in class i to its k-th nearest neighbor in class j, and $D_j\left(x_l^{(i)}\right) = x_l^{(i)} - M_j^k\left(x_l^{(i)}\right)$ is the difference between the point $x_l^{(i)}$ and $M_j^k\left(x_l^{(i)}\right) = (1/k)\sum_{t=1}^{k} x_{\mathrm{kNN}}^{(j)}$, the mean vector of the k nearest neighbors of $x_l^{(i)}$ in class j, its "local k-NN mean".

3 Experimental setup

The experiments were conducted using a 10-fold cross-validation set-up [38]. The reported testing parameters are the average of the 10 experiments.

To evaluate the performance of the algorithms, we use accuracy (*ACC*) [39] considered as the number of successful hits compared to the total number of ratings. In the experimental results, we also included the values of sensitivity (*Sn*) defined as the fraction of positive patterns that are correctly classified and specificity (*Sp*) as the fraction of negative patterns that are correctly classified.

To the measurement of the balanced performance of a learning algorithm between these two classes, we use the *G-Mean* metric [40] as the square root of the

product of sensitivity and specificity. This measure tries to maximize accuracy in order to balance both classes at the same time.

Finally, to compare different classifiers, we use the area under the ROC curve (*AUC*) [41]. *AUC* has an important statistical property since it is equivalent to the probability that the classifier will rank a randomly chosen positive instance higher than a randomly chosen negative instance. In addition, *AUC* can be considered equivalent to the Wilcoxon test of ranks [42].

4 Experimental results

The ensembles were constructed applying the methodology described in section 2, using three different base classifiers (C4.5, k-NN, SVM). The ensembles were labeled according to the base classifier used, such as NDA-C4.5, NDA-kNN, and NDA-SVM.

All classifiers were tested for the prediction of Ames mutagenicity using the benchmark described in section 2. The 117 RDKit descriptors were used as independent variables. Moreover, the results were compared with those obtained by applying the classical methods.

Table 1: Experimental results

Method	Classical methods					Ensemble methods				
	Sp	Sn	ACC	GM	AUC	Sp	Sn	ACC	GM	AUC
C4.5	0.763	0.770	0.766	0.764	0.766	0.820	0.813	0.820	0.820	0.860
k-NN	0.784	0.773	0.779	0.778	0.834	0.823	0.773	0.800	0.800	0.864
SVM	0.770	0.824	0.795	0.796	0.865	0.832	0.810	0.820	0.814	0.873

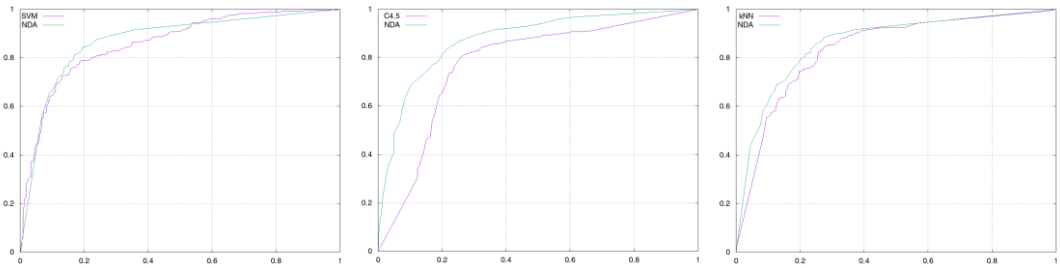
Table 1 shows the results for the standard methods C4.5, k-NN and SVM. The proportion of correctly classified instances is between 76.6% and 79.5%; the better performance was obtained when k-NN classifier was used. To better investigate the prediction ability of classifiers, Table 1 shows the values of sensitivity (*Sn*), specificity (*Sp*), G-mean (*GM*), and the area under the ROC curve (*AUC*). Regardless of the applied method, more than 76.3% of the negative patterns were correctly classified; for the case of positive patterns this value is greater than 77.0%.

As Table 1 shows the expected proportion of positives ranked before a uniformly drawn random negative (*AUC*) gets its maximum value (0.865) using the SVM classifier.

Table 1 also shows the results for the new ensembles proposed in this paper. The accuracy value obtained by the ensemble classifiers is between 2.7% to 7% higher than the standard classifiers (2.7% for k-NN, 3.14% for SVM and 7% for C4.5). In addition, the specificity is improved from 0.763 to 0.820 for C4.5, from 0.784 to 0.823 for k-NN and from 0.770 to 0.832 for SVM.

The *G-Mean* values obtained by the ensembles are around 2.83% to 7.33% higher than the standards k-NN and C4.5, respectively. A smaller increase (2.26%)

was obtained for SVM in this metric. The results for the *AUC* values showed an improvement in performance when the ensembles were used. This behavior is more marked for the ensemble based on C4.5 where an increase of the 12.2% was obtained.



(a) (b) (c)

Figure 2: ROC curves for the different models using: (a) SVM, (b) C4.5 and (c) k-NN as base classifiers.

Figure 2 shows the ROC curves for the classifiers. In all cases, the ensembles show better behavior than the standard methods. By setting an interesting area in a range of false positives between 0.2 to 0.4 (*x*-axis in the graphs), the ROC curves show good performance when SVM-NDA ensemble was used. The performance increases over the base classifier are significantly higher in this range for NDA-SVM (Fig. 2.a) and NDA-C4.5 (Fig. 2.b) with respect to the NDA-kNN (Fig. 2.c). According to these results, we concluded that the proposed ensemble is more accurate than the single classifier that makes them up.

5 Conclusions

This work evaluates new models for the prediction of mutagenicity using ensembles of classifiers constructed by applying a projection of the input variables using Nonparametric Discriminant Analysis.

The experimental results supported the advantage of the proposal compared to the application of classical methods. The greatest general improvement in the performance of the algorithm was obtained when using C4.5 and k-NN as base classifiers.

We can conclude that according to the results, the use of molecular descriptors yields good results for both the standard methods and the ensembles proposed in this paper.

In future work, we plan to extend the methods used here to evaluate the use of other linear and non-linear projections in order to further improve the accuracy in predicting QSAR methods.

Acknowledgments

This work was supported in part by Project TIN2015-66108-P of the Spanish Ministry of Science and Innovation.

References

[1] B. N. AMES, J. MCCANN, AND E. YAMASAKI, *Methods for detecting carcinogens and mutagens with the Salmonella/mammalian-microsome mutagenicity test*, Mutation Research/Environmental Mutagenesis and Related Subjects, **31** (1975) 347-363.

[2] C. L. BRUCE, J. L. MELVILLE, S. D. PICKETT, AND J. D. HIRST, *Contemporary QSAR Classifiers Compared*, Journal of Chemical Information and Modeling, **47** (2007) 219-227.

[3] K. HANSEN, S. MIKA, T. SCHROETER, A. SUTTER, A. TER LAAK, T. STEGER-HARTMANN, N. HEINRICH AND K. R. MÜLLER, *Benchmark Data Set for in Silico Prediction of Ames Mutagenicity*, Journal of Chemical Information and Modeling, **49** (2009) 2077-2081.

[4] F. R. BURDEN AND D. A. WINKLER, *Relevance Vector Machines: Sparse Classification Methods for QSAR*, Journal of Chemical Information and Modeling, **55** (2015) 1529-1534.

[5] G. HINSELMANN, L. ROSENBAUM, A. JAHN, N. FECHNER, C. OSTERMANN, AND A. ZELL, *Large-Scale Learning of Structure−Activity Relationships Using a Linear Support Vector Machine and Problem-Specific Metrics*, Journal of Chemical Information and Modeling, **51** (2011) 203-213.

[6] G. CERRUELA GARCÍA, I. LUQUE RUIZ, AND M. ÁNGEL GÓMEZ-NIETO, *Prediction of Drug Activity Using Molecular Fragments-Based Representation and RFE Support Vector Machine Algorithm*, in Modern Approaches in Applied Intelligence: 24th International Conference on Industrial Engineering and Other Applications of Applied Intelligent Systems, IEA/AIE 2011, Syracuse, NY, USA, June 28 – July 1, 2011, Proceedings, Part II, K. G. Mehrotra, C. K. Mohan, J. C. Oh, P. K. Varshney, and M. Ali, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 396-405.

[7] F. DA SILVA, J. DESAPHY, G. BRET, AND D. ROGNAN, *IChemPIC: A Random Forest Classifier of Biological and Crystallographic Protein–Protein Interfaces*, Journal of Chemical Information and Modeling, **55** (2015) 2005-2014.

[8] K. J. GRAHAM, *An Improved Decision Tree for Predicting a Major Product in Competing Reactions*, Journal of Chemical Education, **91** (2014) 1267-1268.

[9] W. TONG, H. HONG, H. FANG, Q. XIE, R. PERKINS, AND J. D. WALKER, *From Decision Tree to Heterogeneous Decision Forest: A Novel Chemometrics Approach for Structure-Activity Relationship Modeling*, Chemometrics and Chemoinformatics, **894** (2005) 173-185.

[10] F. V. BUONTEMPO, X. Z. WANG, M. MWENSE, N. HORAN, A. YOUNG, AND D. OSBORN, *Genetic Programming for the Induction of Decision Trees to Model Ecotoxicity Data*, Journal of Chemical Information and Modeling, **45** (2005) 904-912.

[11] J. MA, R. P. SHERIDAN, A. LIAW, G. E. DAHL, AND V. SVETNIK, Deep *Neural Nets as a Method for Quantitative Structure–Activity Relationships*, Journal of Chemical Information and Modeling, **55** (2015) 263-274.

[12] K. Z. MYINT, L. WANG, Q. TONG, AND X. Q. XIE, *Molecular Fingerprint-Based Artificial Neural Networks QSAR for Ligand Biological Activity Predictions*, Molecular Pharmaceutics, **9** (2012) 2912-2923.

[13] J. MEILER, R. MEUSINGER, AND M. WILL, *Fast Determination of 13C NMR Chemical Shifts Using Artificial Neural Networks*, Journal of Chemical Information and Computer Sciences, **40** (2000) 1169-1176.

[14] F. CHENG, Y. YU, J. SHEN, L. YANG, W. LI, G. LIU, P. W. LEE AND Y. TANG, *Classification of Cytochrome P450 Inhibitors and Noninhibitors Using Combined Classifiers*, Journal of Chemical Information and Modeling, **51** (2011) 996-1011.

[15] I. BONET, P. FRANCO-MONTERO, V. RIVERO, M. TEIJEIRA, F. BORGES, E. URIARTE, AND A. MORALES HELGUERA, *Classifier Ensemble Based on Feature Selection and Diversity Measures for Predicting the Affinity of A2B Adenosine Receptor Antagonists,* Journal of Chemical Information and Modeling, **53** (2013) 3140-3155.

[16] P. YANG, Y. HWA YANG, B. B ZHOU, AND A. Y ZOMAYA, *A review of ensemble methods in bioinformatics*, Current Bioinformatics, **5** (2010) 296-308.

[17] T. HANCOCK, R. PUT, D. COOMANS, Y. VANDER HEYDEN, AND Y. EVERINGHAM, *A performance comparison of modern statistical techniques for molecular descriptor selection and retention prediction in chromatographic QSRR studies,* Chemometrics and Intelligent Laboratory Systems, **76** (2005) 185-196.

[18] C. MERKWIRTH, H. MAUSER, T. SCHULZ-GASCH, O. ROCHE, M. STAHL, AND T. LENGAUER, *Ensemble methods for classification in cheminformatics*, Journal of Chemical Information and Computer Sciences, **44** (2004) 1971-1978.

[19] Y. FREUND AND R. E. SCHAPIRE, *A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting*, Journal of Computer and System Sciences, **55** (1997) 119-139.

[20] L. BREIMAN, *Bagging Predictors*, Machine Learning, **24** (1996) 123-140.

[21] L. BREIMAN, *Random forests*, Machine learning, **45** (2001) 5-32.

[22] N. GARCÍA-PEDRAJAS AND C. GARCÍA-OSORIO, *Constructing ensembles of classifiers using supervised projection methods based on misclassified instances*, Expert Systems with Applications, **38** (2011) 343-359.

[23] G. KARLSTRÖMA, R. LINDHA, P. MALMQVISTA, B. O ROOSA, U. RYDEA, V. VERYAZOVA, P. O. WIDMARKA, M. COSSIB, B. SCHIMMELPFENNIGC, P. NEOGRADYD, AND L. SEIJOE, *MOLCAS: a program package for computational chemistry*, Computational Materials Science, **28** (2003) 222-239.

[24] Y. SHAO ET AL., *Advances in methods and algorithms in a modern quantum chemistry program package*, Physical Chemistry Chemical Physics, **8** (2006) 3172-3191.

[25] A. MAURI, V. CONSONNI, M. PAVAN, AND R. TODESCHINI, *Dragon software: An easy approach to molecular descriptor calculations*, Match, **56** (2006) 237-248.

[26] A. R. KATRITZKY, R. PETRUKHIN, H. YANG, AND M. KARELSON, *CODESSA PRO, User's manual*, University of Florida, 2002.

[27] C. W. YAP, *PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints*, Journal of Computational Chemistry, **32** (2011) 1466-1474.

[28] RDKIT: OPEN-SOURCE CHEMINFORMATICS SOFTWARE. http://www.rdkit.org/ (accessed January 17, 2017).

[29] CHEMICAL CARCINOGENESIS RESEARCH INFORMATION SYSTEM; 2009. NCRI Informatics Initiative. http://www.cancerinformatics.org.uk/matrix/CCRIS.htm (accessed January 17, 2017).

[30] C. HELMA, T. CRAMER, S. KRAMER, AND L. DE RAEDT, *Data mining and machine learning techniques for the identification of mutagenicity inducing substructures and structure activity relationships of noncongeneric compounds*, Journal of Chemical Information and Computer Sciences, **44** (2004) 1402-1411.

[31] J. KAZIUS, R. MCGUIRE, AND R. BURSI, *Derivation and validation of toxicophores for mutagenicity prediction*, Journal of Medicinal Chemistry, **48** (2005) 312-320.

[32] J. FENG, L. LURATI, H. OUYANG, T. ROBINSON, Y. WANG, S. YUAN AND S. S. YOUNG, *Predictive Toxicology: Benchmarking Molecular Descriptors*

and Statistical Methods, Journal of Chemical Information and Computer Sciences, **43** (2003) 1463-1470.

[33] P. N. JUDSON, P. A. COOKE, N. G. DOERRER, N. GREENE, R. P. HANZLIK, C. HARDY, A. HARTMANN, D. HINCHLIFFE, J. HOLDER AND L. MÜLLER, *Towards the creation of an international toxicology information centre*, Toxicology, **213** (2005) 117-128.

[34] *Ames Mutagenecity Benchmark Dataset*. http://pubs.acs.org/doi/suppl/10.1021/ci900161g (accessed January 17, 2017).

[35] S. HAYKIN, *Neural Networks: A Comprehensive Foundation*. Prentice Hall PTR, 1998.

[36] N. GARCÍA-PEDRAJAS, J. MAUDES-RAEDO, C. GARCÍA-OSORIO, AND J. J. RODRÍGUEZ-DÍEZ, *Supervised subspace projections for constructing ensembles of classifiers*, Information Sciences, **193** (2012) 1-21.

[37] K. FUKUNAGA AND J. MANTOCK, *Nonparametric Discriminant Analysis*, IEEE Transactions on Pattern Analysis and Machine Intelligence, **6** (1983) 671-678.

[38] T. G. DIETTERICH, *Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms*, Neural Computation, **10** (1998) 1895-1923.

[39] I. H. WITTEN AND E. FRANK, *Data Mining: Practical Machine Learning Tools and Techniques*, Morgan Kaufmann, 2005.

[40] M. KUBAT, R. C. HOLTE, AND S. MATWIN, *Machine Learning for the Detection of Oil Spills in Satellite Radar Images*, Machine Learning, **30** (1998) 195-215.

[41] A. P. BRADLEY, *The Use of the Area under the ROC Curve in the Evaluation of Machine Learning Algorithms*, Pattern Recognition, **30** (1997) 1145-1159.

[42] J. A. HANLEY AND B. J. MCNEIL, *The Meaning and Use of the Area under a Receiver Operating Characteristic (ROC) Curve*, Radiology, **143** (1982) 29-36.

3D trajectory generation for rotating extensible manipulators using zenithal gnomic projection and polar piecewise interpolants

Mihai Dupac
Department of Design and Engineering, Bournemouth University

e-mail: mdupac@bournemouth.ac.uk

Abstract

In this paper the 3D trajectory planning of the end-effector of a rotating extensible manipulator arm is described using projective geometry and polar piecewise interpolants. Perspective projections of the 3D via points located of the northern hemisphere are seen from the manipulator base location which represents the centre of rotation of the extensible manipulator. Due to the geometry of the manipulator arm, polar zenithal gnomic projections are associated with the spherical coordinates named the radial distance related to the manipulator length along the polar and azimuthal angles.
Possible trajectories of the end effector are generated using a geometric transformation applied to the polar piecewise interpolants that approximate the gnomic projective trajectory of the 3D via points. Smoothness of the polygonal trajectory is obtained through the use of piecewise interpolants with continuous derivatives between the 3D via points/nodes. The novel approach generate fast trajectory interpolation of 3D via points, minimise execution time and allows easy calculation of kinematics variables. To verify the proposed approach and to validate the model, numerical simulations are conducted for two different configurations. In this context the reachable workspace (working volume) of the extensible manipulator is examined with respect to the base location and the end effector path.

Key words: Trajectory planning, Mechanisms, Robots, Interpolation
MSC2000: 65D05, 70B15

1. Introduction

Trajectory planning and control of robotic arms and manipulators represents critical key elements in industry especially when reduced production costs and improved productivity are a must [1, 2]. The ability to control the motion along a desired trajectory is imperative especially when minimal execution time [3, 4] and is considered. Moreover, optimal trajectory planning [5] require smoot trajectory generation (velocity, acceleration [11] and/or jerk) which can be achieved by the means of adequate interpolants.

Shape approximation methods and algorithms based on piecewise interpolating functions and splines [6, 7] are adequate tools in preserving positivity, monotonicity and convexity and generating smooth motion through continuous piecewise smooth functions [12] in order to reduce the induced vibrations and minimise resonant frequencies excitation. Since algebraic-trigonometric Hermite polynomials are easy to use in generating smooth and continuous motion while preserving the continuity [10], kinematic variables and joint-space trajectories can be efficiently calculated along a specified joints path. A path-planning associated methodology for a set of nodal points with kinematical requirements is presented in [9]. Quick 3D trajectory planning using cylindrical coordinates associated with the end-effector trajectory of motion is discussed in [15], an analytical approach for path planning interpolation for any industrial robot is considered in [8], and the general theory for actual kinematics and dynamics is presented in [17].

In this paper the 3D trajectory of the end-effector of a rotating extensible manipulator is generated by the use of perspective zenithal gnomic projections of the 3D via points and polar piecewise interpolants. The approach generate fast trajectory interpolation and continuous motion while minimise execution time. The proposed approach is verified by numerical simulations conducted for two different configurations. In this context the reachable workspace (working volume) of the extensible manipulator is examined with respect to the base location and the end effector path.

2. Mathematical Modelling

Manipulator Model

The rotating extensible manipulator arm is composed of a rotating *z*-Guide, a slider joint (Rigid Guide) and a sliding link as shown in Fig. 1. The z-guide of the manipulator can rotate at *O* - in the fixed Cartesian reference frame *Oxyz* – about the vertical axis *Oz*. The z-guide is connected with the slider joint guide by a rigid rod r_G which is rigidly attached to it. The slider joint can rotate relative to the z-guide about the rigid rod $r_{G..}$ The sliding link which is interconnected with the slider joint can slide in and out of the slider (Rigid Guide). The height (length) of the z-guide is l_Z, the length of the sliding joint/guide is l_G and the length of the sliding link id l_{SL}. The 3D coordinates of the end effector trajectory can be expressed using spherical coordinates (or cylindrical) coordinates by

$$\begin{cases} x = \rho \sin\varphi \cos\theta \\ y = \rho \sin\varphi \sin\theta \\ z = \rho \cos\varphi \end{cases} \overset{\Leftarrow \text{ spherical}}{\underset{\text{cylindrical} \Rightarrow}{\Leftrightarrow}} \begin{cases} x = r\cos\theta \\ y = r\sin\theta \\ z = z \end{cases} \overset{\Leftarrow \text{ cylindrtical}}{\underset{\text{Cartesian} \Rightarrow}{\Leftrightarrow}} \begin{cases} r = \rho \sin\varphi \\ \theta = \theta \\ z = \rho \cos\ \varphi \end{cases} \tag{1}$$

The end effector of the manipulator arm should move to the via-points $P_1, P_2, ..., P_k, ..., P_n$ set by the user (Fig.1).

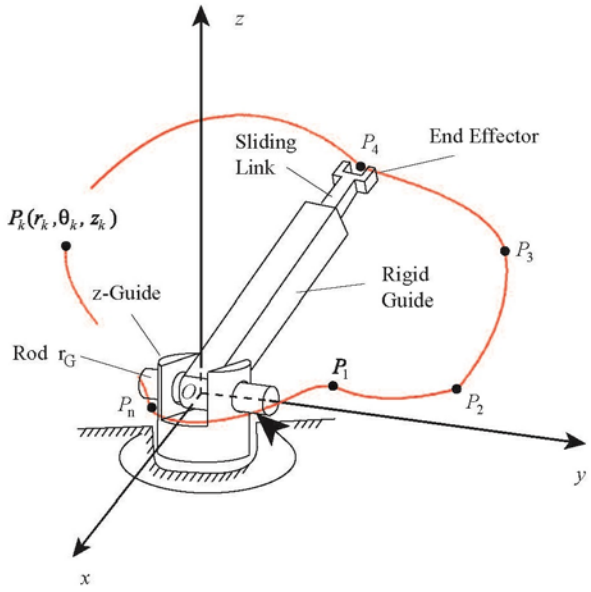


Figure 1. Rotating extensible manipulator model.

To interpolate between the data one can consider first the perspective projections of the via-points $P_i(r_i, \theta_i, \varphi_i)$ - viewed from the manipulator base location which represents the centre of projection - to a plane *N* perpendicular to the axis *Oz* and located at the distance $r_{Max} = \max\{r_i\}_{i=\overline{0,N_i}}$ from the origin *O* (tangent to the sphere of radius r_{Max} centred in origin) as shown in Fig. 2. Due to the geometry of the manipulator arm, polar zenithal gnomic projections are associated with the perspective projections denoted by $P_i'(R_i, \theta_i, \varphi_i) = P_i'(X_i, Y_i, r_{Max})$ calculated using

$$\begin{cases} \dfrac{X_i - x_O}{x_i - x_O} = \dfrac{Y_i - y_O}{y_i - y_O} = \dfrac{Z_i - z_O}{z_i - z_O} \\ Z_i = r_{Max} = \max\{r_i\}_{i=\overline{0,N_i}} \end{cases} \Leftrightarrow \begin{cases} X_i = x_O + \dfrac{x_i - x_O}{z_i - z_O}(r_{Max} - z_O) \\ Y_i = y_O + \dfrac{y_i - y_O}{z_i - z_O}(r_{Max} - z_O) \end{cases} \tag{1}$$

that is the intersection of the lines OP_i' (passing by O and $P_i, i=\overline{0,N_i}$) with the plane N. The associated radii $R_i = d(O,P'_i)$ of the points $P_i'(X_i, Y_i, r_{Max})$ - located in the projective plane N – are calculated using

$$R_i = d(O', P'_i) = \sqrt{(X_i)^2 + (Y_i)^2}$$

$$\overset{\text{Perspective Projection}}{=} \sqrt{\left(x_O + \frac{x_i - x_O}{z_i - z_O}(r_{Max} - z_O)\right)^2 + \left(y_O + \frac{y_i - y_O}{z_i - z_O}(r_{Max} - z_O)\right)^2}$$

$$\overset{\text{Gnomic Projection}}{=} \sqrt{\left(\frac{x_i}{z_i}(r_{Max})\right)^2 + \left(\frac{y_i}{z_i}(r_{Max})\right)^2} = \frac{r_i}{z_i} r_{Max} \tag{2}$$

Two projections P'_i and P'_{i+1} of the 3D via points P_i and P_{i+1} on the projective plane N are shown in Fig. 2 with the spherical coordinates named the radial distance related to the manipulator length along the polar and azimuthal angles.

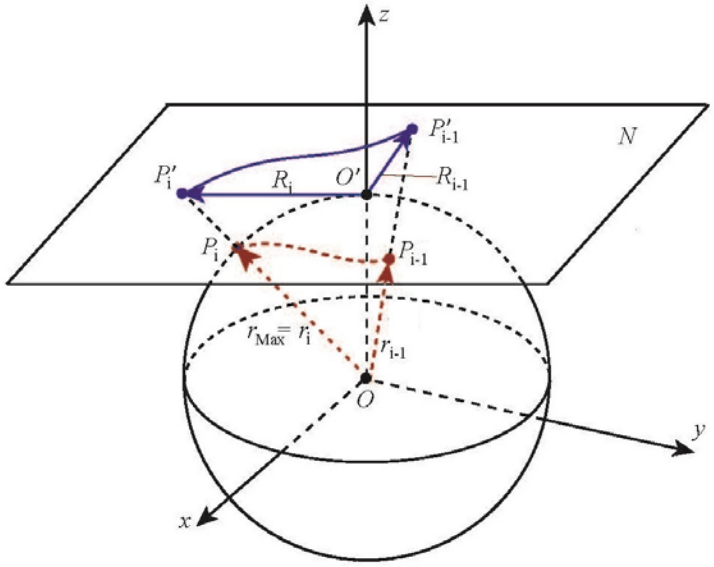


Figure 2. Perspective projections P_i' of the via-points P_i, and associated piecewise interpolation

To interpolate between the perspective projections P_i' of the via-points $P_i, i=\overline{0,N_i}$, located on the projective plane N, one can consider the angles $\theta_0 < \theta_1 < \theta_2 < ... < \theta_{N_k-1} < \theta_{N_i} = \theta_0 + 2\pi$, and the associated radii R_i. For each

interval $[\theta_i, \theta_{i+1}]_{\overline{i=0,N_i-1}}$ and associated radii R_i and R_{i+1} of two consecutive projections P'_i and P'_{i+1}, a piecewise polar interpolant that approximate the trajectory on the *projective* plane (Fig. 2) can be expressed as a Hermite-type function [13, 14] defined by

$$R(\theta) = \sum_{k=0}^{q} c_k^i (\theta - \theta_i)^k \tag{3}$$

where q is the order of the polynomial to be used (order 3 was used here), $c_0^i = R_i,\ c_1^i = \dot{R}_i, c_2^i = \frac{1}{h_i}\left[-\left(2\dot{R}_i + \dot{R}_{i+1}\right) + 3\Delta\dot{R}_i\right], c_3^i = \frac{1}{h_i^2}\left[\dot{R}_i + \dot{R}_{i+1} - 2\Delta\dot{R}_i\right],$ $R_i = R(\theta_i),\ R_{i+1} = R(\theta_{i+1}),\ h_i = \theta_{i+1} - \theta_i,\quad \Delta R_i = \frac{R_{i+1} - R_i}{h_i}$. The 3D trajectories of the end effector are calculated using the geometric transformation below

$$\begin{cases} x = \dfrac{R(\theta) z(\theta) \cos\theta}{r_M} \\ y = \dfrac{R(\theta) z(\theta) \sin\theta}{r_M} \\ z = z(\theta) \end{cases} \tag{4}$$

The trajectory height $z = z(\theta)$ for a step size S_i is approximated by

$$z(\theta) = z_i + kL_i, \qquad k = 1,2,..., N_i \tag{5}$$

where the height of a step is $L_i = \frac{|z_{i+1} - z_i|}{N_i}$ and the number of steps is $N_i = \frac{h_i}{S_i}$.

Existence of a Solution (End-Effector Trajectory)
Since the geometric path of the manipulator [7, 11] is generated using the via-points the end effector, manipulator base location and arm length, the trajectory can exist only inside the working envelope [16] defined by

$$\begin{cases} \max\limits_{P_{i_k} \in T_{P_i}} d(O, P_{i_k}) \le l_{LG} + l_{SL} \\ \max(l_{LG}, l_{SL}) \le \max\limits_{P_{i_k} \in T_{P_i}} d(O, P_{i_k}) \\ l_Z \le \sup\limits_{\overline{i=1,n;\ k=1,N_i}} (z_{i_k}) \end{cases} \tag{4}$$

where $P_{i_k}, k = \overline{1, N_i}$ represents interpolating points along the pricewise curve defined by the via-points P_i and P_{i+1}, $P_{i_1} = P_i, \forall i = \overline{1, n-1}$ and $P_{i_{N_i}} = P_{i+1}, \forall i = \overline{1, n-1}$.

1. Results

Two numerical examples are presented to illustrate trajectory generation including the existence of solutions for the extensible manipulator. The associated manipulator dimensions are: z-guide length l_z=0.5·10^{-1} m, slider length l_G = 7.5·10^{-1} m, sliding link length l_{SL} = 7.5·10^{-1} m. The z-guide is rotating with an angular velocity $\omega = 1$rad/s about the z-guide. The radial distance related to the manipulator length, and the polar and azimuthal angles of the manipulator via-points for the two chosen numerical examples are shown in Table 1 and Table 2.

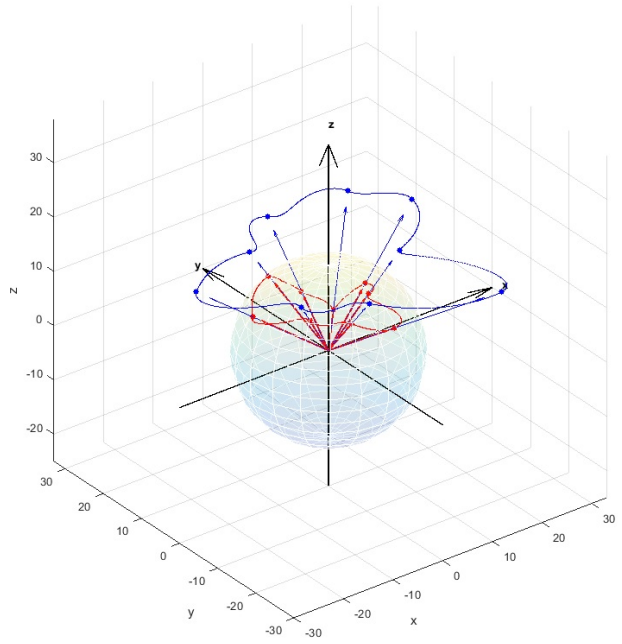
Variable Name	Value									
i	1	2	3	4	5	6	7	8	9	10=1
θ_i	20	45	85	125	165	215	260	305	345	360+20
r_i	13.0	7.2	10.6	15.6	15.8	14.9	12.0	12.5	11.4	13.0
φ_i	32	34	41	50	35	47	48	29	52	32

Table 1. Via-points for the 1st configuration of the manipulator (lengths in decimetres)

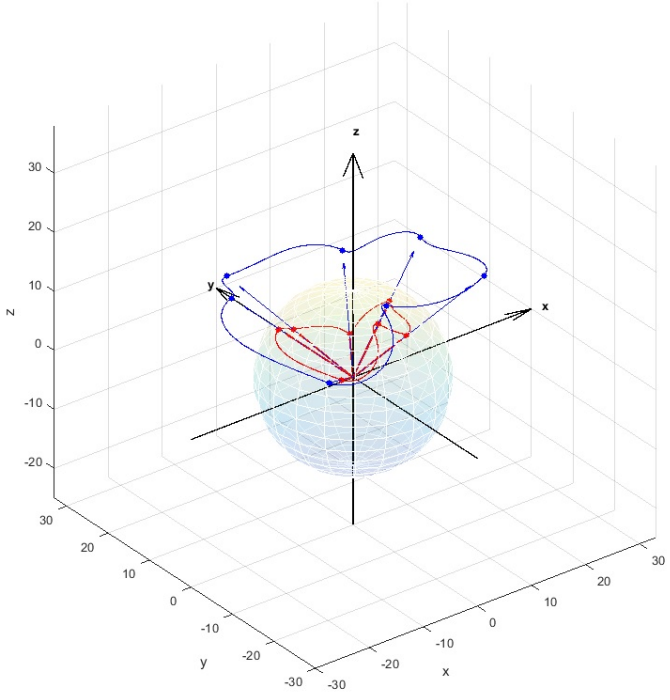
Variable Name	Value							
i	1	2	3	4	5	6	7	8=1
θ_i	20	60	130	150	225	280	335	360+20
r_i	12.8	7.8	11.4	14.2	14.8	12.0	10.0	12.8
φ_i	39	51	38	39	28	65	37	39

Table 2. Via-points for the 2nd configuration of the manipulator (lengths in decimetres)

The 3D end effector trajectory for the two manipulator configurations (Table 1 and Table 2) is shown in Fig. 3.a and Fig. 3.b. The projection of each trajectory on the *projective* plane shown in Fig. 3.a and Fig. 3.b using a blue colour – representing the radial trajectory of the end-effector in the projective plane – have been obtained using polar interpolation curves expressed as a Hermite-type functions. The red arrows pointing toward the via-points represents the radial distance related to the manipulator length in the spherical coordinates system.



(a)



(b)

Figure 3. Generated trajectory using gnomic projection and polar interpolation (a) manipulator arm goes outside the working envelope (b) end effector can reach all the via-points.

Possible trajectories of the end effector are generated using geometric transformations applied to the polar piecewise interpolants that approximate the gnomic projective trajectory of the 3D via points using the red interpolating curves shown in Fig. 3.a and Fig. 3.b. It can also be seen that the manipulator arm goes outside the working envelope/workspace [16], that is, the end effector of the manipulator arm cannot reach all the via-points in Table 1 along the planned trajectory shown in Fig. 3.a. However, for the 2nd simulation presented in Table 2, the end effector can reach all the via-points shown in Fig. 3.b.

2. Conclusions

In this paper the modelling and 3D trajectory planning of an extensible rotating manipulator end effector is presented. Polar zenithal gnomic perspective projections of the 3D via points and polar piecewise interpolants expressed as polar Hermite polynomials are considered for the end effector trajectory generations. Smoothness of the polygonal trajectory is obtained through the use of piecewise interpolants with continuous derivatives between the 3D via points/nodes. The existence of a solution in relation with the base of the manipulator arm is also addressed. Two numerical examples related to two different manipulator configurations are presented to illustrate trajectory planning and verify the proposed approach.

3. References

[1] A. GASPARETTO, P. BOSCARIOL, A. LANZUTTI AND R. VIDONI, *Trajectory planning in robotics*, Mathematics in Comp. Sci., **6** (2012) 269–279.

[2] B. LE BOUDEC, M. SAAD AND V. NERGUIZIAN, *Modeling and adaptive control of redundant robots*, Math. Comp. Sim., 19 (2006) 395-403.

[3] J.E. BOBROW, S. DUBOWSKY AND J.S. GIBSON, *Time-optimal control of robotic manipulators along specified paths*, International Journal of Robotics Research **4(3)** (1985) 3–17.

[4] D. CONSTANTINESCU AND E.A. CROFT, *Smooth and time-optimal trajectory planning for industrial manipulators along specified paths*, Journal of Robotic Systems 17(5) (2000) 233–249.

[5] L. ADHAMI AND E. COSTE, *Optimal planning for minimally invasive surgical robots*, IEEE Transactions on Robotics and Automation 19(5) (2003) 854–863.

[6] J. SANCHEZ-REYES, *Single-valued spline curves in polar coordinates*, Computer Aided Design 24 (1992) 307–315.

[7] L.M. KOCIC AND G.V. MILOVANOVIC, *Shape preserving approximations by polynomials and splines*, Computer & Mathematics with Applications 33(11) (1997) 59–97.

[8] I.G. KANG AND F.C. PARK, *Cubic spline algorithms for orientation interpolation*, International Journal for Numerical Methods in Engineering 46 (1999) 45-64.

[9] L.J. DU PLESSIS AND J.A. SNYMAN, *Trajectory-planning through interpolation by overlapping cubic arcs and cubic splines*, International Journal for Numerical Methods in Engineering 57 (2003) 1615–1641.
[10] B. SU AND L. ZOU, *Manipulator trajectory planning based on the algebraic trigonometric Hermite blended interpolation spline*, Procedia Engineering 29 (2012) 2093–2097.
[11] M. KALYONCU, *Mathematical modelling and dynamic response of a multistraight-line path tracing flexible robot manipulator with rotating-prismatic joint*, Applied Mathematical Modelling, 32 (2008) 1087–1098.
[12] S. AMAT, S. BUSQUIER, A. ESCUDERO AND L. CARLOS TRILLO, *Lagrange interpolation for continuous piecewise smooth functions*, Journal of Computational and Applied Mathematics, 221 (2008) 47–51.
[13] M. DUPAC, *Smooth trajectory generation for rotating extensible manipulators*, Mathematical Methods in Applied Sciences, 2016. (http://onlinelibrary.wiley.com/doi/10.1002/mma.4210/pdf).
[14] Y. IWASHITA, *Piecewise polynomial interpolation*, Open Gamma Quantitative Research, 15 (2014) 1-22.
[15] M. DUPAC AND P. SEWELL, *Quick 3D trajectory planning for rotating extensible manipulators using piecewise polynomial interpolation*, In Proceedings of the Congress on Numerical Methods in Engineering, Valencia, 3-5 July, 2017 (accepted).
[16] Y. CAO, K. LU, X. LI AND Y. ZANG, *Accurate numerical methods for computing 2D and 3D robot workspace*, International Journal of Advanced Robotic Systems 8(6) (2011) 1-13.
[17] D.B. MARGHITU AND M. DUPAC, *Advanced dynamics: analytical and numerical calculations with Matlab*, Springer, New York, 2012.

Influence of plotting positions on the Michael's acceptance regions in a Normal Q-Q Plot

María Dolores Estudillo-Martínez[1], Sonia Castillo-Gutiérrez[1] and Emilio Lozano-Aguilera[1]

[1] *Department of Statistics and Operation Research, University of Jaén*

emails: **mdestudi@ujaen.es, socasti@ujaen.es, elozano@ujaen.es**

Abstract

In 1983, Michael proposed acceptance regions for Normal Probability Plots to detect non-normality in a set of observations. In these acceptance regions, Michael used the plotting position proposed by Hazen in 1930. In this work we have studied the influence of plotting positions on the acceptance regions in a Normal Q-Q Plot to determinate if another plotting position is better than the used by Michael.

Key words: acceptance regions, plotting position, Normal Q-Q Plot, goodness-of-fit, graphical techniques

1. Introduction

Normal Q-Q Plots are used as graphical techniques of goodness-of-fit. Confidence bands are included in a normal probability plot to detect non-normality in a set of observations with the advantage that the conclusion is not influenced by the subjectivity of the observer.

In 1983, Michael [1] proposed acceptance regions for Probability Plots using the Hazen's plotting position (1930) [2]. In this contribution we have studied the influence of the choice of plotting positions on the acceptance regions in a Normal Q-Q Plot to determinate if another plotting position is better than the definition proposed by Hazen.

2. Acceptance regions of Michael in a Normal Q-Q Plot

In 1983, Michael proposed the acceptance regions for a Normal Q-Q Plot.

Given a set of ordered observations, $x_{(1)},\ldots,x_{(n)}$, the steps to construct a Normal Q-Q Plot with acceptance regions of Michael are as follows:

1. Obtain a Normal Q-Q Plot from the observations.
2. Calculate maximum-likelihood estimators for the mean and standard deviation of observations ($\hat{\mu}$ and $\hat{\sigma}$, respectively).
3. Get the value of d_{sp}, given a fixed level of significance α and the number of observations n, in the table of critical values of the statistic D_{sp} of Michael. The value of the test statistic is calculated in the following way:

$$D_{sp} = \max_i \left| r_i - s_i \right|, \; i = 1,...,n$$

where

$$r_i = \left(\frac{2}{\pi}\right)\arcsin\left(\sqrt{p_i}\right),$$

$p_i = \dfrac{i-0.5}{n}$ is the plotting position proposed by Hazen,

$$s_i = \left(\frac{2}{\pi}\right)\arcsin\left(\sqrt{u_i}\right), \;\; u_i = \Phi\left(\frac{x_{(i)} - \hat{\mu}}{\hat{\sigma}}\right)$$

and Φ is the distribution function of the standard normal.

4. Represent acceptance regions in the Normal Q-Q Plot defined by the following expression:

$$y = \hat{\mu} + \hat{\sigma}\Phi^{-1}\left(\sin^2\left[\arcsin\left(\Phi^{1/2}\left(x\right)\right) \pm 0.5\pi d_{sp}\right]\right)$$

5. Reject the hypothesis of normality if some observation falls outside the acceptance regions.

3. Other definitions of plotting positions

In the literature there are several definitions of plotting position formula for use on Normal Q-Q Plots.

In this contribution, we are going to specify briefly nine formulae for plotting positions.

The first definition what we're going to study it was proposed by Hazen in 1930:

$$p_i = \frac{i-0.5}{n} \qquad i = 1,\ldots,n$$

In 1939, Weibull [3] proposed the definition:

$$p_i = \frac{i}{n+1} \qquad i = 1, \ldots, n$$

The third formula was introduced by Benard and Bos-Levenbach [4] in 1953 with the following expression:

$$p_i = \frac{i-0.3}{n+0.4} \qquad i = 1, \ldots, n$$

The plotting position formula more commonly use was proposed by Blom [5] in 1958:

$$p_i = \frac{i-0.375}{n+0.25} \qquad i = 1, \ldots, n$$

Tukey [6], in 1962 introduced the following definition for plotting position:

$$p_i = \frac{i-(1/3)}{n+(1/3)} \qquad i = 1, \ldots, n$$

Gringorten [7] in 1963 proposed:

$$p_i = \frac{i-0.44}{n+0.12} \qquad i = 1, \ldots, n$$

In 1978, Cunnane [8] proposed the following definition:

$$p_i = \frac{i-0.4}{n+0.2} \qquad i = 1, \ldots, n$$

Yu and Huang (2001) [9] introduced the expression:

$$p_i = \frac{i-0.326}{n+0.348} \qquad i = 1, \ldots, n$$

Finally, Lozano-Aguilera et al. (2014) [10] proposed a plotting position as the median of the *ith* order statistic from a beta distribution Beta(i, n-i+1).

4. Simulation Study

In this section we have developed a simulation study to examine the power of the acceptance regions proposed by Michael using the previously specified definitions of plotting positions.

The power of a statistical test is the probability that the test will reject the null hypothesis when the null hypothesis is false, i.e. the probability of not committing a type II error, or making a false-negative decision. The power is calculated by

fixing the probability of the type I error, i.e. the significance level, α. We have fixed the significance level at 5%.

For this study, 10000 samples of size n=10(10)100 were generated from 19 alternative distributions. Here, we will only have to include tables concerning Chi-square(10) and Normal distribution.

Table 1: Empirical power (in %) of Michael's acceptance regions to detect non-normality in a Chi-square(10) distribution for different definitions of plotting positions at significance level 5%.

n	10	20	30	40	50	60	70	80	90	100
Hazen	8.63	18.25	28.23	39.69	49.99	60.57	67.86	75.94	80.94	86.80
Weibull	11.97	20.33	23.35	31.56	36.60	44.38	50.39	59.91	65.19	73.16
BBL	6.73	14.71	21.97	31.55	40.43	50.71	57.87	67.69	72.53	80.60
Blom	6.75	15.12	23.54	33.71	43.08	54.30	61.00	70.69	75.40	82.77
Tukey	7.01	15.00	22.76	32.33	41.51	52.23	59.43	68.95	73.67	81.57
Gring.	7.93	16.71	25.70	36.55	46.39	57.26	64.25	73.31	78.13	84.94
Cun.	7.53	15.92	24.35	34.85	44.21	55.33	62.24	71.78	76.45	83.68
YH	6.93	14.90	22.50	32.09	41.29	51.93	59.07	68.67	73.38	81.34
Lozano	6.89	14.77	22.03	31.47	40.57	51.06	58.19	67.94	72.77	80.83

Table 2: Empirical % of rejections under the null hypothesis of normality.

n	10	20	30	40	50	60	70	80	90	100
Hazen	5.16	4.56	4.79	5.21	4.94	5.13	4.93	5.36	4.66	5.49
Weibull	6.65	5.69	4.28	4.88	4.05	3.89	3.44	3.83	2.93	3.50
BBL	4.38	3.78	3.56	4.01	3.46	3.67	3.27	3.86	2.79	3.68
Blom	4.55	3.84	3.74	4.20	3.59	3.90	3.59	4.06	3.03	3.87
Tukey	4.47	3.74	3.65	4.10	3.55	3.84	3.39	4.06	2.92	3.77
Gring.	4.78	4.03	4.21	4.67	4.31	4.46	4.25	4.78	3.83	4.72
Cun.	4.68	3.85	3.88	4.43	3.94	4.15	3.86	4.39	3.55	4.35
YH	4.48	3.76	3.63	4.04	3.48	3.80	3.38	4.02	2.87	3.76
Lozano	4.45	3.73	3.57	4.00	3.40	3.68	3.28	3.91	2.78	3.66

In Table 1 we can observe that from n=30, the better power of the acceptance regions of Michael is reached using Hazen's plotting position. Similarly, Table 2 shows that in almost all sample sizes, with Hazen's plotting position will achieve results closer to the desired value (5%) than with the rest of definitions.

5. References

[1] J.R. MICHAEL, *The Stabilized Probability Plot*, Biometrika 70(1) (1983) 11-17.

[2] A. HAZEN, *Flood Flows. A Study of Frequencies and Magnitudes*, Wiley, New York, 1930.

[3] W. WEIBULL, *The Phenomenon of Rupture in Solids*, Ingeniors Vetenskaps Akademien Handlingar, 17 (1939).

[4] A. BENARD AND E.C. BOS-LEVENBACH, *The Plotting of Observations on Probability Paper*, Statistica 7 (1953).

[5] G. BLOM, *Statistical Estimates and Transformed Beta-Variables*, John Wiley & Sons, New York, 1958.

[6] J.W. TUKEY, *The future of data analysis*, Annals of Mathematical Statistics 33 (1962).

[7] I. GRINGORTEN, *A plotting rule for extreme probability paper*, J. Geophys. Res. 68(3) (1963), 813–814.

[8] C. CUNNANE, *Unbiased plotting positions. A review*, J. Hydrol. 37 (1978), 205–222.

[9] G.-H. YU AND C.-C. HUANG, *A distribution free plotting position*, Stoch. Environ. Res. Risk Assess. 15 (2001), 462-476.

[10] E.D. LOZANO-AGUILERA, M.D. ESTUDILLO-MARTÍNEZ AND S. CASTILLO-GUTIÉRREZ, *A proposal for plotting positions in probability plots.* J Appl Stat 41 (2014), 118–126.

Rate-Distortion/Complexity Analysis of Video Compression with Capability beyond HEVC

David Garcia-Lucas[1], Gabriel Cebrián-Márquez[1], Antonio J. Diaz-Honrubia[2] and Pedro Cuenca[1]

[1]Albacete Research Institute of Informatics (I3A)
[2]Technologies and Information Systems Department
University of Castilla-La Mancha, Spain

emails: David.Garcia72@alu.uclm.es, Gabriel.Cebrian@uclm.es, Antonio.DHonrubia@uclm.es, Pedro.Cuenca@uclm.es

Abstract

ITU-T Video Coding Expert Group (VCEG) and ISO/IEC Moving Picture Expert Group (MPEG) are studying the potential need for standardization of the future video coding technology with a compression capability that significantly exceeds that of the current High Efficiency Video Coding (HEVC) standard, including its current extensions. Both groups are working together on this exploration activity in a collaboration effort known as Joint Video Exploration Team (JVET) to evaluate compression technology designs proposed by their experts in this area. This paper describes the coding features that are under coordinated test model study by the JVET, and presents a rate-distortion/complexity analysis to study their real capabilities. Experimental results show that the new model achieves 25% bitrate reduction, but at a cost of extremely high computational complexity (11×) with respect to HEVC.

Key words: HEVC, JEM, Evaluation, Computational Cost

1. Introduction

High Efficiency Video Coding (HEVC) [1] was developed in 2013 by the *Joint Collaborative Team on Video Coding* (JCT-VC) to replace H.264/*Advanced Video Coding* (AVC) standard [2], which is the most used video coding standard and has dominated digital video services in the domestic and professional markets for over ten years. In terms of rate-distortion (RD) performance, HEVC roughly

doubles the compression performance of H.264/AVC, but at a cost of extremely high computational and storage complexities during encoding [3].

The amount of daily data generated in our current interconnected society is astounding. An important amount of this data is transmitted using different video formats. According to the latest Zettabyte Cisco report [4], in 2015, 70% of all IP traffic, both business and consumer, was IP video traffic. Moreover, the trend predicts that this ratio will keep increasing with the market penetration of novel services such as video-on-demand, internet video to TV, virtual reality, 360-degree video or internet-based video surveillance, and existing services such as cloud storage and video streaming services. Moreover, the report also details how all this information will, most likely, be consumed through portable devices, with limited computation and bandwidth capabilities. It is expected that smartphone traffic will double the traffic originated by computers by 2020. Along with the increase in the amount of video information exchanged on the Internet, and the bandwidth and computational limitations of portable devices, there is an increasing demand for higher bit rates, higher video resolutions and better video quality.

This issue results in the need of a new generation of video coding techniques to increase the quality and compression rates of previous standards. Since the release of HEVC, ITU-T *Video Coding Expert Group* (VCEG) and ISO/IEC *Moving Picture Expert Group* (MPEG) have been studying the potential need for the standardization of future video coding technologies with a compression capability that significantly exceeds that of the current HEVC standard. To better coordinate this study, VCEG and MPEG created the *Joint Video Exploration Team* (JVET) as a collaboration framework. The scope of the JVET activity includes the consideration of a variety of video sources and applications. For example, sources include camera-view content, screen content, consumer generated content, virtual reality/360-degree content, and high dynamic range content, while example applications include broadcasting (with live or pre-authored content), real-time video conferencing, video chat, on-demand viewing, storage-based media replay, and surveillance with fixed or moving cameras [5][6].

These new efforts of standardization and compression enhancements are being explored and implemented by the JVET in a software test model known under the name of *Joint Exploration Test Model* (JEM) [7]. JEM represents the software model which is being used as a starting point for the next generation of video coding standards following HEVC.

As it has been the case for all past ITU-T and ISO/IEC video coding standards, in HEVC and JEM only the bitstream structure, syntax, constraints and mapping for the generation of decoded pictures are standardized. Consequently, every decoder conforming to the standard will produce the same output for a given standard

conformant bitstream. This limitation provides encoder developers with maximum freedom to optimize their implementations and explore new techniques to improve the standard. Moreover, to assist the industry community in learning how to use the standard, the standardization effort not only includes the development of a text specification document, but also the reference software source code of the encoder and decoder standard implementation. This reference software is usually used as research tool to improve the standard and explore new techniques. For the development of the rate-distortion/complexity analysis detailed in this paper, the reference software for JEM version 3.0 will be used as base encoder. This paper focuses on a comparative evaluation of the quality/computational cost of the HEVC and JEM codecs using objective measures of assessment to analyse their real capabilities. The comparison was performed using the JVET *Common Test Conditions* presented in [8].

The remainder of this paper is organized as follows. Section 2 includes some technical background of the HEVC standard. Section 3 presents the algorithm description of JEM under study, followed by the experimental results in Section 4. Finally, Section 5 concludes the paper.

2. High Efficiency Video Coding (HEVC) Standard

HEVC can be considered an evolution of the current H.264/AVC, since it maintains the same block-based hybrid approach used in all previous video compression standards. In addition, new tools have been introduced in HEVC that increase its coding efficiency compared with H.264/AVC [9].

One of the most important changes affects the picture partitioning [10]. HEVC defines a new flexible *Coding Tree Unit* (CTU) structure which is a replacement of the *Macroblocks* (MBs), 16×16 pixel blocks, used in the previous standards. With the aim of achieving an optimal adaptation to the content details, CTUs can vary from a size of 64×64 pixels, to something much smaller, as it can be subsequently partitioned into four square sub-blocks of half resolution, named *Coding Units* (CUs), with a minimum allowable size of 8×8 pixels. Therefore, a CTU can be further partitioned into four depth levels, from *d=0* for 64×64 CU to *d=3* for 8×8 CUs, having 4^d CUs in each depth level. Thus, a CU in depth level *d* can be denoted as $CU_{d,k}$ (k=0,1,.., 4^d-1), and the four sub-CUs pending on $CU_{d,k}$ are denoted as $CU_{d+1,4k+i}$ (i=0,1,.., 3). In a CTU of 64×64, it can be observed that the maximum number of available CUs is $\sum_{d=0}^{d=3} 4^d$.

HEVC increases even more the flexibility of the CTU by defining two tree structures containing new unit types: *Prediction Units* (PUs), and *Transform Units* (TUs). For intra-picture prediction, a PU uses the same 2N×2N size as for the $CU_{d,k}$ to which it belongs, allowing it to be split into four N×N PUs only for CUs at the minimum depth level. Therefore, the PU size can range from 64×64 to

4×4 pixels. For inter-picture prediction, several non-square rectangular block shapes are available in addition to the square ones, providing a total of eight different PU sizes (2N×2N, 2N×N, N×2N, N×N, 2N×nU, 2N×nD, nL×2N, nR×2N). The prediction residue obtained in each of the PUs is transformed using various TU sizes from 32×32 to 4×4. In Figure 1, an example of the partitioning is shown, depicting how a CTU is structured in a hierarchical tree where each CU branch ends in a leaf ($CU_{d,k}$) that is the root for the prediction and transform trees. Figure 2 shows the partitioning of CTU into CUs (white), PUs (green) and TUs (black) applied to the *Basketball Pass* sequence.

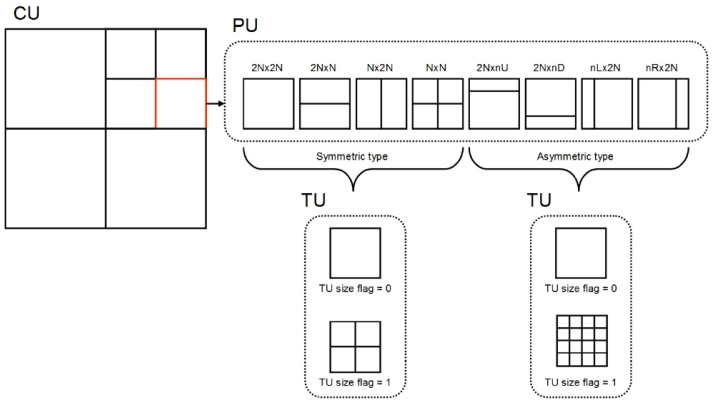


Figure 1. Partitioning of CTU into CUs, PUs and TUs

Figure 2. Partitioning of CTU into CUs (white), PUs (green) and TUs (black) applied to *Basketball Pass* sequence

It should be noted that a CTU can be split into 85 different CUs ($\sum_{d=0}^{d=3} 4^d$), which can be split into 1702 PUs, and each of these PUs has to be evaluated for all intra or inter prediction modes available, and each of the obtained residual blocks can be transformed into up to three TU sizes. HEVC checks most of the PUs (inter and intra modes) to decide whether it should split a CU or not by choosing the best

RD case. Furthermore, in the case of inter prediction, for each of these PU partitions a motion estimation algorithm is called. This wide range of possibilities makes HEVC much more computationally expensive than its predecessor, H.264/AVC. HEVC introduces changes in other modules too, such as intra prediction (where a total of 35 different coding modes can be selected), new image filters or new transform sizes [9], among others.

The above analysis evidences the need to reduce the *Rate-Distortion Optimization* (RDO) complexity for the HEVC intra/inter prediction, to make real-time HEVC video codecs with the best possible performance. With the aim of reducing this huge RDO complexity, several suboptimal fast HEVC video codecs have been developed by the industry using a reduced set of prediction modes that are previously selected as candidates, in a low complexity evaluation process.

3. Algorithm Description of Joint Exploration Test Model

This section describes the coding features that are under coordinated study by the JVET. They have been included in the JEM test model as potential enhanced video coding technology beyond the capabilities of HEVC.

JEM is being developed on top of HEVC test model (HM version 16.6) [11]. The basic encoding and decoding flowchart of HEVC is kept unchanged in JEM. However, the design elements of the most important modules, including the modules of block structure, intra and inter prediction, residue transform, loop filter and entropy coding, have been modified, introducing new tools. The following new coding features are included in JEM [7]:

- Block structure
  - Quadtree plus binary tree (QTBT) block structure
- Intra prediction improvements
  - 65 intra prediction directions
  - 4-tap interpolation filter for intra prediction
  - Boundary filter applied to other directions in addition to horizontal/vertical
  - Cross-component linear model (CCLM) prediction
  - Position dependent intra prediction combination (PDPC)
  - Adaptive reference sample smoothing
- Inter prediction improvements
  - Sub-PU level motion vector prediction
  - Locally adaptive motion vector resolution (AMVR)
  - 1/16 pel motion vector storage accuracy
  - Overlapped block motion compensation (OBMC)
  - Local illumination compensation (LIC)
  - Affine motion prediction
  - Pattern matched motion vector derivation
  - Bi-directional optical flow (BIO)
- Transform
  - Explicit multiple core transform

  - Mode dependent non-separable secondary transforms
  - Signal dependent transform (SDT)
- Adaptive loop filter (ALF)
- Enhanced CABAC design
  - Context model selection for transform coefficient levels
  - Multi-hypothesis probability estimation
  - Initialization for context models

All the methods listed above have been integrated into the main software branch of JEM [12], and in particular, in the software implementation of JEM 3.0.

3.1. Block Structure

In JEM, some additional concepts are included to this picture partitioning stage compared to HEVC. JEM incorporates a *QuadTree plus Binary Tree* (QTBT) structure for blocks. By the use of this structure, the separation in CU, PU and TU is no longer needed. This gives more flexibility for CU partition shapes to better match the local characteristics of the video sequence. Therefore, in JEM, through QTBT, CUs, TUs and PUs have the same block size. In QTBT, CUs can have either square or rectangle shapes like PUs in HEVC. Each CTU (up to 256×256 pixels) is first partitioned by a quadtree structure in squared CUs. Then, leaf nodes can be further partitioned by a binary tree structure. By the use of this tree, each CU can be split in horizontal and vertical CUs, defining the final structure of the CTU and, in consequence, its specific subdivision in CUs. An example of a CTU block structure in JEM is depicted in Figure 3. We may see how the binary tree determines the type of CU splitting. A value of 1 on the binary structure of the QTBT determines a symmetric vertical split for a CU, while a 0 value specifies a horizontal split. For the quadtree splitting, there is no need to indicate the splitting type since a block is always split horizontally and vertically into 4 sub-blocks with an equal size.

The following parameters are defined to have efficient signalling of a QTBT:

- *CTUSize*: the root node size of a quadtree, same concept as in HEVC.
- *MinQTSize*: the minimum allowed quadtree leaf node size.
- *MaxBTSize*: the maximum allowed binary tree root node size.
- *MaxBTDepth*: the maximum allowed binary tree depth.
- *MinBTSize*: the minimum allowed binary tree leaf node size.

In one example of the QTBT partitioning structure, the CTU size is set as 128×128 pixels (for the luma, and in the case of YUV420 subsampling format 64×64 pixels for the chroma), the *MinQTSize* is set to 16×16, the *MaxBTSize* is set to 64×64, the *MinBTSize* (for both width and height) is set to 4, and the *MaxBTDepth* is set to 4. The quadtree partitioning is applied to the CTU first to generate quadtree leaf nodes. The quadtree leaf nodes may have a size from 16×16 (i.e. the *MinQTSize*) to 128×128 (i.e. the *CTUSize*). If the leaf quadtree

node is 128×128, it will not be further split by the binary tree since the size exceeds the *MaxBTSize* (i.e. 64×64). Otherwise, the leaf quadtree node could be further partitioned by the binary tree. In this way, the quadtree leaf node is also the root node for the binary tree with a binary tree depth of 0. When the binary tree depth reaches *MaxBTDepth* (i.e. 4), it implies no further splitting. When the width of the binary tree node equals to *MinBTSize* (i.e. 4), it implies no further horizontal splitting. Similarly, when the height of the binary tree node equals to *MinBTSize*, it implies no further vertical splitting. The leaf nodes of the binary tree correspond to CUs that are further processed by the prediction and transform modules without any further partitioning.

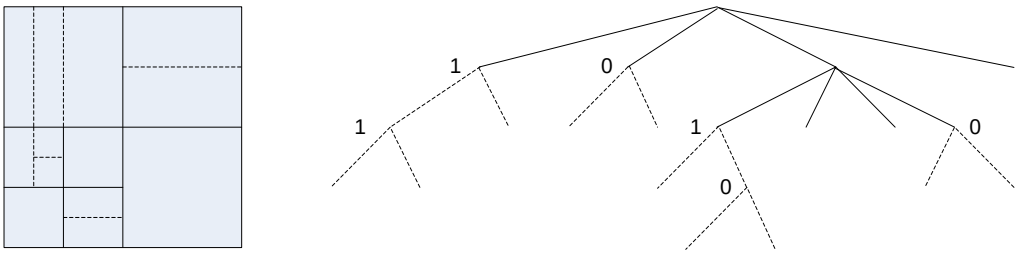


Figure 3: Illustration of a QTBT structure [7]

3.2. Intra and Inter prediction improvements

For intra prediction, to capture finer edge directions presented in natural videos, the directional intra modes are extended from 33, as defined in HEVC, to 65. The new directional modes are depicted as red dotted arrows in Figure 4. The Planar and DC modes remain the same. These denser directional intra prediction modes apply for all block sizes and both luma and chroma intra predictions. To accommodate the increased number of directional Intra modes, an Intra mode coding method with 6 *Most Probable Modes* (MPMs) is used. Two major technical aspects are involved: 1) the derivation of 6 MPMs, and 2) entropy coding of 6 MPMs and non-MPM modes.

Regarding inter prediction, with QTBT, each CU can have at most one set of motion information for each prediction direction. Two sub-CU level motion vector prediction methods are studied by splitting a large CU into sub-CUs and deriving motion information for all the sub-CUs of the large CU. *Advanced Temporal Motion Vector Prediction* (ATMVP) method allows each CU to fetch multiple sets of motion information from multiple blocks smaller than the current CU in the collocated reference picture. In the *Spatial-Temporal Motion Vector Prediction* (STMVP) method, motion vectors of the sub-CUs are derived

recursively by using the temporal motion vector predictor and spatial neighbouring motion vector. As in HEVC, JEM defines the merge mode to copy the motion information from neighbour blocks. In HEVC, motion vector accuracy is one-quarter pel (one-eighth in the case of chroma components for YUV420 video). In JEM, the accuracy for the internal motion vector storage and the merge candidate increases to one-sixteenth pel. The highest motion vector accuracy (1/16 pel) is used in motion compensation inter prediction for the CU coded with Skip/Merge mode

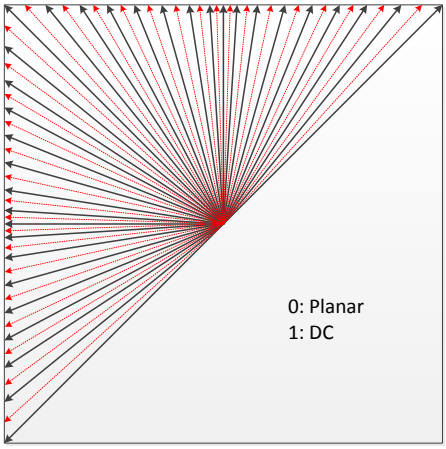


Figure 4: Proposed 67 intra prediction modes [7]

3.3. Other improvements

In JEM, in addition to the *Discrete Cosine Transform* (DCT)-II and the 4×4 *Discrete Sine Transform* (DST)-VII which have been employed in HEVC for transform coding, an *Adaptive Multiple Transform* (AMT) scheme has been chosen to encode the inter and intra residual CUs. It uses different transforms from the DCT and DST families than the ones that are used in HEVC. The newly introduced transform matrices are DST-VII, DCT-VIII, DST-I and DCT-V.

In JEM, an *Adaptive Loop Filter* (ALF) with block based filter adaption is applied. For the luma component, according to the direction and activity of local textures, one among 25 filters is selected for each 2×2 block. ALF aims to reduce visible artifacts such as ringing and blurring by reducing the mean absolute error between the original image and the reconstructed image.

In HEVC, the entropy coder used is *Context-based Adaptive Binary Arithmetic Coding* (CABAC). JEM uses an enhanced version of CABAC with a modified context modelling for transform coefficients, a multihypothesis probability estimation with context-dependent updating speed, and an adaptive initialization of models.

4. Rate-Distortion/Complexity Analysis

4.1. Encoding Parameters

This section aims to evaluate the rate-distortion/complexity capabilities of JEM version 3.0 with respect to HEVC. Our experiments rely on the default configuration of HM 16.6, which is used as an anchor for the obtained results. For a fair comparison, the experiments are conducted with a single-threaded implementation of the encoders. The hardware platform used in the experiments is composed of an Intel® Xeon® E5-2630L v3 CPU running at 1.80 GHz and 16 GB of main memory. To ensure a common framework for the simulations, the experiments were conducted under the *Common Test Conditions and Software Reference Configurations* recommended by the JVET [8] for the *All-Intra* (AI), *Low-Delay* (LD) and *Random-Access* (RA) mode configurations. That recommendation specifies the use of four *Quantization Parameter* (QPs) (22, 27, 32 and 37) and a set of 24 test sequences classified in six classes, A1, A2, A, B, C and D, which cover a wide range of resolutions from the largest (4096×2160 pixels) to the smallest (416×240 pixels), and frame rates from 100 fps to 24 fps. All the sequences use YUV420 chroma subsampling and a bit-depth of 10 (A1, A2 and A classes) and 8 bits (B to D classes).

4.2. Metrics

The rate-distortion/complexity analysis was evaluated in terms of *Encoder Time Ratio* (ETR) and R-D performance. The ETR measure was computed following the Equation (1). Regarding the R-D performance, the *Bjøntegaard Delta Rate* (BD-Rate) metric defined by ITU [14] is used. The BD-Rate provides the average difference between the R-D curves measured as a percentage of bit rate that is necessary to increase or decrease to achieve the same *Peak Signal to Noise Ratio* (PSNR) quality in both curves. In our simulation, a negative BD-Rate means the encoded bit rate using the JEM codec under study is lower than the bit rate obtained with HM, thus it is denoted as the gain in terms of bit rate saving.

$$Encoder\_Time\_Ratio = \frac{Enc.Time(JEM)}{Enc.Time(HM)} \qquad (1)$$

4.3. Simulation Results

Table 1 shows the performance progress of JEM compared with the HM 16.6 reference software for the RA configuration in terms of BD-rate gain and ETR. It can be observed that JEM codec obtains up to 26% bitrate reduction while increasing the computational complexity by around 11×. The performance of JEM 3.0 compared to the HM reference software is also summarized in terms of encoder time in Table 2 for the AI, LP-P, LD-B and RA mode configurations. A significant increase of computational complexity can be observed for the AI configuration (up to 60×).

Table 1: Comparison of different versions of JEM with HM 16.6 [13]

JEM version	Random Access (RA)	
	BD-Rate	ETR
JEM 1	-20.84 %	5.35×
JEM 2	-22.93 %	5.32×
JEM 3	-26.62 %	11.32×

Table 2: JEM3.0 compared to HEVC coding performance summary with different configurations [13]

Test configuration	BD-rate			ETR	
	Y	U	V	Enc.	Dec.
All Intra	-18%	-21%	-21%	60×	2×
Random Access	-26%	-30%	-29%	11×	10×
Low Delay-B	-21%	-25%	-26%	7×	7×
Low Delay-P	-24%	-28%	-29%	6×	4×

Figure 5 shows the profiling results of JEM 3.0 obtained for the RA configuration. These timing values have been extracted from the average of all the tested sequences and the four QP values. As can be seen, more than 70% of the encoding time is devoted to the Inter prediction, becoming the most computationally expensive operation in the encoder, whereas 23% of the total execution time corresponds to the Intra prediction. The remaining encoder modules (labelled as "Others"), including transform, quantization, entropy coding, and in-loop filtering represents only around 6% of the total encoding time. The complexity of the Inter module can be justified by the large amount of repetitive operations that the encoder has to perform on the same picture samples but with different block partitions. Finally, Table 3 shows in detail the JEM profiling results obtained for the RA configuration for all the tested sequences and for different stages of the encoder. It can be observed that the new features of "Merge" operation in JEM represent an expensive operation in the encoder.

5. Conclusions

This paper presents the coding features that are under coordinated test model study by the JVET, and presents a rate-distortion/complexity analysis. Experimental results show that the new model (JEM 3.0) achieves 25% bitrate reduction, but at a cost of an extremely high computational complexity (11×) with respect to HEVC. Future versions of the model should consider encoder complexity as one of the criteria when evaluating the new tools to be included, encouraging further encoder complexity reduction. Moreover, different techniques to accelerate the encoder should be proposed before the new standard is ready to be used in real applications. These techniques might make use of CPU and/or GPU parallelism, as well as other soft computing algorithms.

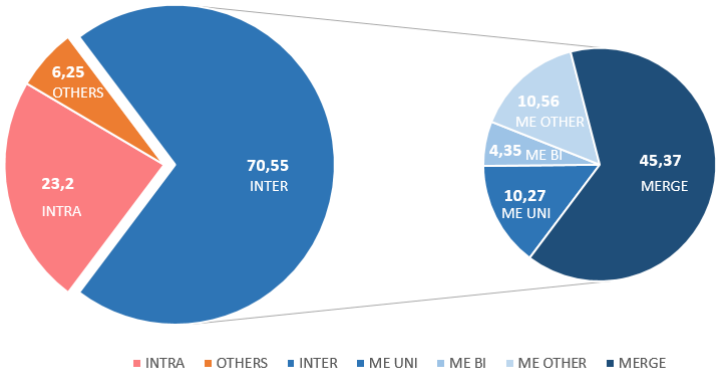


Figure 5: Distribution of encoding time (RA configuration)

Table 3: Profiling results for RA configuration (JEM 3.0)

Class	Sequence	ME_UNI_IME	ME_UNI_FME	ME_UNI_OTH	ME_BI_IME	ME_BI_FME	ME_BI_OTH	ME_OTHER	INTRA_LUMA	INTRA_CHRO	AffineMerge	Merge2Nx2N	MergeFRUC	OTHER
A1	Tango	1.61	5.23	4.89	2.17	3.02	0.32	11.62	21.13	2.17	0.26	18.76	22.68	6.16
	Drums100	1.33	3.89	4.00	0.77	2.23	0.26	12.37	16.71	1.62	0.56	27.45	22.72	6.11
	CampfireParty	1.32	4.00	5.57	1.94	2.26	0.29	5.98	35.35	5.27	0.09	12.10	20.04	5.79
	ToddlerFountain	1.36	2.09	4.22	0.56	1.17	0.20	5.52	41.44	4.33	0.05	17.19	17.17	4.72
A2	CatRobot	1.15	4.60	4.30	1.50	2.70	0.30	13.46	16.19	1.57	1.12	23.72	23.01	6.40
	TrafficFlow	0.55	4.57	2.95	0.65	2.64	0.21	14.92	23.42	1.04	0.92	25.28	16.65	6.21
	DaylightRoad	1.18	4.39	4.26	1.48	2.56	0.29	12.17	20.03	1.77	1.05	22.78	21.96	6.08
	Rollercoaster	1.75	6.10	5.83	2.68	3.46	0.36	11.73	16.26	1.85	1.11	17.83	24.84	6.21
A	Traffic	0.55	3.85	3.31	0.69	2.23	0.24	12.59	21.75	0.78	0.46	26.77	21.23	5.58
	PeopleOnStreet	1.43	2.88	4.48	0.81	1.62	0.28	9.02	24.25	1.86	0.33	24.43	22.94	5.69
	Nebuta	0.60	2.79	2.85	0.50	1.60	0.19	8.53	23.22	2.25	3.49	24.97	22.56	6.47
	SteamLocomotive	1.54	5.62	4.63	2.33	3.23	0.31	11.56	22.71	2.14	0.20	17.96	21.83	5.96
B	Kimono	1.17	3.93	4.00	1.07	2.27	0.27	11.98	19.22	1.49	0.41	25.21	23.23	5.77
	ParkScene	0.80	3.68	3.70	0.96	2.14	0.27	11.96	22.38	1.14	0.31	25.35	21.59	5.73
	Cactus	1.10	4.53	4.76	1.68	2.60	0.31	10.67	21.45	1.82	1.01	20.18	23.85	6.09
	BasketballDrive	1.43	4.08	5.04	1.62	2.32	0.31	9.64	25.43	2.58	0.51	18.53	22.44	6.08
	BQTerrace	0.75	4.95	4.57	1.58	2.85	0.34	13.22	13.80	0.70	0.33	25.63	24.22	7.07
C	BasketballDrill	1.37	5.16	6.23	2.40	2.89	0.41	9.44	19.58	1.97	0.38	16.40	27.32	6.48
	BQMall	1.22	5.00	5.77	2.29	2.85	0.40	10.69	16.44	1.38	0.40	18.98	28.08	6.51
	PartyScene	0.92	3.86	4.98	1.57	2.18	0.36	10.02	18.74	1.36	0.79	21.52	27.46	6.25
	RaceHorsesC	1.60	4.36	6.44	2.27	2.45	0.43	7.83	23.49	2.27	0.59	15.45	26.16	6.66
D	BasketballPass	1.51	4.40	6.75	2.37	2.46	0.45	8.05	21.80	2.44	0.27	15.53	27.07	6.91
	BQSquare	0.59	4.05	4.72	1.35	2.36	0.37	12.95	12.39	0.36	0.88	26.36	26.28	7.37
	BlowingBubbles	0.87	4.39	5.59	2.14	2.52	0.42	10.10	15.66	0.92	0.60	19.92	30.02	6.89
	RaceHorses	1.55	4.40	6.90	2.50	2.48	0.49	7.99	20.35	1.92	0.54	15.52	28.27	7.10
	Class A1	1.40	3.80	4.67	1.36	2.17	0.27	8.87	28.66	3.35	0.24	18.87	20.65	5.70
	Class A2	1.16	4.92	4.33	1.58	2.84	0.29	13.07	18.97	1.56	1.05	22.40	21.61	6.22
	Class A	1.03	3.78	3.81	1.08	2.17	0.26	10.43	22.98	1.76	1.12	23.53	22.14	5.92
	Class B	1.05	4.23	4.41	1.38	2.43	0.30	11.49	20.45	1.54	0.51	22.98	23.06	6.15
	Class C	1.28	4.59	5.86	2.13	2.59	0.40	9.50	19.56	1.74	0.54	18.09	27.26	6.47
	Class D	1.13	4.31	5.99	2.09	2.45	0.43	9.77	17.55	1.41	0.57	19.33	27.91	7.06
AVERAGE		**1.17**	**4.27**	**4.83**	**1.59**	**2.44**	**0.32**	**10.56**	**21.32**	**1.88**	**0.66**	**20.95**	**23.74**	**6.25**

Acknowledgements

This work was jointly supported by the Spanish Ministry of Economy and Competitiveness and the European Commission (FEDER funds) under the project TIN2015-66972-C5-2-R, and by the Spanish Ministry of Education, Culture and Sports under the grant FPU13/04601.

References

[1] ISO/IEC AND ITU-T, *High Efficiency Video Coding (HEVC). ITU-T Recommendation H.265 and ISO/IEC 23008-2 (version 3)*, April 2015.

[2] ISO/IEC AND ITU-T, *Advanced Video Coding for Generic Audiovisual Services, ITU-T Recommendation H.264 and ISO/IEC 14496-10 (AVC)*, February 2012.

[3] J.-R. OHM, G. J. SULLIVAN, H. SCHWARZ, T. K. TAN, AND T. WIEGAND, *Comparison of the Coding Efficiency of Video Coding Standards - Including High Efficiency Video Coding (HEVC)*, IEEE Trans. Circuits Syst. Video Technol., vol. 22, no. 12, pp. 1669–1684, Dec. 2012.

[4] Cisco Visual Networking Index (VNI) and VNI Service Adoption. Global Forecast Update, 2015–2020. June 2016.

[5] Requirements for a Future Video Coding Standard v4. ISO/IEC WG11 MPEG, 115th Meeting, Geneva, June 2016, Doc. N16359.

[6] Requirements for Future Video Coding (H.FVC). Annex Q6.B of report of Q6/16 and TD 8R1/WP3, ITU-T SG 16, Geneva, 16-27 January 2017.

[7] J. CHEN, E. ALSHINA, G. J. SULLIVAN, J.-R. OHM, AND J. BOYCE. *Algorithm Description of Joint Exploration Test Model 3 (JEM 3)*, JVET-C1001, 3rd Meeting, Geneva, June 2016.

[8] K. SUEHRING, AND X. LI. *JVET Common Test Conditions and Software Reference Configurations*. JVET-B1010, 2nd Meeting, San Diego, February 2016.

[9] G. J. SULLIVAN, J. R. OHM, W. J. HAN, AND T. WIEGAND, *Overview of the High Efficiency Video Coding (HEVC) Standard*, IEEE Trans. Circuits Syst. Video Technol., vol. 22, no. 12, pp. 1649-1668, December 2012.

[10] I.-K. KIM, J. MIN, T. LEE, W.-J. HAN, AND J.-H. PARK, *Block Partitioning Structure in the HEVC Standard*, IEEE Trans. Circuits Syst. Video Technol., vol. 22, no. 12, pp. 1697-1706, December 2012.

[11] Joint Collaborative Team on Video Coding Reference Software, ver. HM 16.6. [Online]. Available: https://hevc.hhi.fraunhofer.de/svn/svn_HEVCSoftware/

[12] JEM reference software, https://jvet.hhi.fraunhofer.de/svn/svn_HMJEMSoftware/.

[13] M. KARCZEWICZ, AND E. ALSHINA. *JVET AHG Report: Tool Evaluation (AHG1)*, JVET-D1001, 4th Meeting, Chengdu, October 2016.

[14] G. BJØNTEGAARD, *Improvements of the BD-PSNR Model*, ITU-T SG16 Q6 Document VCEG-AI11, 35th VCEG Meeting, July 2008.

Texture orientation detection over parallel architectures: a qualitative overview

Elena Georgiana Paraschiv[1], Damián Ruiz-Coll[2], Maria Pantoja[3] and Gerardo Fernández-Escribano[1]

[1] *Computer System Department, Instituto de Investigación en Informática de Albacete, University of Castilla-La Mancha, Spain.*
[2] *Signal Theory and Communications Department, University of Rey Juan Carlos, Spain.*
[3] *Computer Science & Software Engineering Department, California Polytechnic State University, USA.*

emails: elenag.paraschiv@uclm.es, damian.ruiz.coll@gmail.com, mpanto01@calpoly.edu, gerardo.fernandez@uclm.es

Abstract

In this paper, we propose a high parallel implementation model of an optimized version of the Mean Directional Variance algorithm using Sliding Window (MDV-SW). On the one hand, our approach allows to speed-up the HEVC (High Efficiency Video Coding) Intra-prediction in two different ways: first, reducing the total number of operations (Intra modes tested) by exploiting the directional locality of the image and, second, exploiting the parallelization opportunities to reduce the time needed for processing a block (operations involved). On the other hand, we can obtain the same perceptual information (texture orientation) from a raw video file as from a file obtained by decoding a HEVC stream. Therefore, we have focused on the real-time applications with an improved parallel version of the MDV-SW algorithm, covering most of the 33 angular directions proposed in the HEVC standard. The experimental results show that our improved version of the original MDV-SW algorithm and its parallel version can achieve a high computational complexity reduction without perceptive quality losses.

Key words: HEVC, Intra-prediction, parallel programming, OpenMP, CUDA

1. Introduction

High Efficiency Video Coding (HEVC) [1] is a new video coding standard approved by the Joint Collaborative Team Video Coding (JCT-VC) working group, from ITU (International Telecommunication Union) and ISO (International Organization for Standardization) international organizations. It is called to replace the H.264/AVC standard [2] for high resolution formats beyond High Definition (HD), such as the emerging Ultra High Definition format (UHD), known as 4K.

The high HEVC efficiency is reached through the high flexibility to adapt the three new units called Coded Units (CU), Prediction Units (PU) and Transform units (TU) to the local image features. Those units can adopt a wide range of block sizes from 4x4 to 64x64. The high computational complexity of HEVC is due to fact that the encoder must select the optimal size of the different units and also the optimal Intra-prediction mode. This process is implemented by the Rate Distortion Optimization (RDO) stage, which carries out an exhaustive evaluation of all available combinations of unit sizes and prediction modes, demanding a high computational burden.

HEVC has proved to be the state-of-the-art video coding solution for a wide range of applications, services, bit rates and formats, from the very low resolution to beyond 4K, but with an extremely high computational cost. For example, the high HEVC Intra-prediction performance is achieved by using of the brute force in the computation of the optimal angular predictor. However, that mass computing approach demands a huge amount of computation resources.

In this paper, we propose a parallel version of the MDV-SW algorithm to speed-up the HEVC Intra-prediction in two different ways: first, reducing the number of operations of the optimal gradient detection by exploiting the directional locality of the image and, second, using different parallel schemes to reduce the time needed for processing a block, trying different compilers and programming languages to fully exploit this. Hence, the MDV-SW algorithm is improved from [3]:

1. By increasing the number of texture orientations from 12 to 24 rational slopes, which improves performance and reduces the quality loss detected in the previous version. We use the Mean Directional Variance (MDV) to measure the local directionality of the image.
2. By designing a novel parallel implementation for the computation of the Mean Directional Variance.

Our approach can be used not only for HEVC encoding scenarios but also with low-power resources cameras where the video information is only available in

raw format. The explanation of the algorithm, the simulations and the performance evaluation was done with 8x8 pixels blocks.

The rest of this paper is organized as follows. Section 2 and 3 briefly introduces the scheme of the HEVC Intra-prediction, as well as the improved version of the MDV-SW algorithm, respectively. Section 4 presents the details of our parallel approach using OpenMP and CUDA schemes. Finally, Section 5 will show the experimental results. Conclusions are detailed in Section 6.

2. High Efficiency Video Coding

The HEVC standard use the well-known architecture of hybrid encoding model, where the spatial decorrelation is carried out by the Intra-prediction and the temporal decorrelation by the Inter-prediction based on the Motion Estimation and Motion Compensation (ME-MC) scheme. The residual generated by both predictions are transformed to frequential domain with the popular DCT (Discrete Cosine Transform).

HEVC has expanded the 9 Intra-prediction modes of H.264/AVC to 35 modes, covering an extremely high number of angular orientations. These modes include 33 directional modes and two non-angular modes (DC and Planar). The 33 angular modes can be classified into two categories: the first is the Integer Position Modes (IPM) composed by five modes, whose orientations match with the exact location of the reference pixels, and the second category, named Fractional Position Modes (FPM), includes the rest of the modes whose orientation is between two reference samples. Therefore, the predictor needs to be computed by the interpolation of the two nearest samples, increasing the computational cost of the process [3].

Intra-prediction follows the TUs partition tree and therefore prediction modes are applied to 4×4, 8×8, 16×16 and 32×32 size blocks. This makes the calculation of Intra-prediction stage very computationally intensive, because all of them have to be computed in order to estimate the best one between the full ranges of block size. These operations do not present any dependency among block sizes, for example 4x4 and 8x8 blocks can be calculated at the same time, but most encoders still compute the Intra-prediction using a sequential code, as the HM reference model do. The combination of data independence and computational complexity makes Intra-prediction a perfect candidate for parallelization techniques. In this article, we implemented parallelism at the block level and we evaluated the implementation on different parallel architectures to determine which one can achieve the highest Intra-prediction performance that means the minimum distortion with the minimum bit rate.

The Intra-prediction are built by using as reference samples the neighboring pixels of the left and top blocks, and these predictors have to be built in both the encoder and the decoder. However, the decoder has only the decoded pixels of the neighbor blocks that have been quantizated, and consequently their values can be far away from the original ones, depending of the quantization step used. This is the reason why the encoder is forced to use also as reference pixels the decoded pixels of the neighbor blocks, which prevents to build the predictors of one block without having been previously encoded and decoded its neighboring blocks.

3. The MDV-SW algorithm

Continuing with the work presented in [3], this work has a double aim: to improve the MDV-SW algorithm performance by doubling the number of angular prediction modes (covered in this section), and to exploit the scalability property of MDV-SW algorithm, which favors the implementation in parallel architectures.

It was proved in [3] that, when HEVC reference model is compared to the MDV-SW algorithm, the computational complexity is reduced around 30% on average, with a rate penalty of only 0.4%. This is achieved by detecting the dominant texture orientation of a block with a low complexity, which allows the selection of a reduced number of angular modes as candidates for the RDO stage.

In MDV-SW, the texture or gradient orientation of a block is given by the lowest mean directional variance along certain spatial directions (or co-lines) with rational slopes $r = r_x/r_y$, where r_x and r_y are integer positions of lattice $\Lambda \in \mathbb{Z}^2$, which is free of any pixel interpolation processing. Each pair of rational slopes (r_1 and r_2, for example) is considered to form a sublattice $\Lambda \subset \mathbb{Z}^2$, which allows the computation of the pixels belonging to the co-lines with slope r_1 and r_2. That is to say, pixels (x,y) of a co-line can be obtained from the lineal combination of r_1 and r_2, eq. (1):

$$\begin{bmatrix} x \\ y \end{bmatrix} = c_1 \begin{bmatrix} r_{x1} \\ r_{y1} \end{bmatrix} + c_2 \begin{bmatrix} r_{x2} \\ r_{y2} \end{bmatrix} + \begin{bmatrix} s_{k_x} \\ s_{k_y} \end{bmatrix} \Rightarrow \begin{matrix} x = c_1 r_{x_1} + c_2 r_{x_2} + s_{k_x} \\ y = c_1 r_{y_1} + c_2 r_{y_2} + s_{k_y} \end{matrix} \quad \forall\, c_1, c_2 \in \mathbb{Z}, \qquad (1)$$

where $s_k = [s_{kx}, s_{ky}]^T \ \forall k = 0, 1, \ldots, \det(M_\Lambda) - 1$ is a shifting vector of the sublattice Λ, which can be represented by the generator matrix $M_\Lambda = [r_1, r_2]$, as is described in [3].

As can be seen in Table I, the number of rational slopes used in this paper is 24, which is twice the number of directions proposed in [3], so we are increasing a 50% the angular accuracy of the algorithm. Figure 1 depicts the 33 angular Intra-prediction modes defined in HEVC (orange solid lines), as well as the 24 rational slopes proposed in this work (dashed blue lines for the ones used in [3], and dashed green lines for the new ones).

Table I. Rational slopes used for the texture orientation

Initial rational slopes [REF]	r_0	r_1	r_2	r_3	r_4	r_5	r_6	r_7	r_8	r_9	r_{10}	r_{11}
	-1/1	-1/2	-1/4	0	1/4	1/2	1/1	2/1	4/1	∞	-4/1	-2/1
New rational slopes	r_{12}	r_{13}	r_{14}	r_{15}	r_{16}	r_{17}	r_{18}	r_{19}	r_{20}	r_{21}	r_{22}	r_{23}
	-3/4	-2/3	-1/3	1/3	2/3	3/4	4/3	3/2	3/1	-3/1	-3/2	-4/3

By using 24 directions instead of 12, the texture orientation detection is significantly improved, since it achieves a gradient direction closer to the real dominant texture orientation of the block under evaluation. As seen in Figure 1, most of the rational slopes are either overlapping the Intra-prediction angular modes or they are placed next to them. Therefore, the MDV-SW algorithm is improved and it can be considered as good a predictor of the optimal Intra-prediction mode as HEVC.

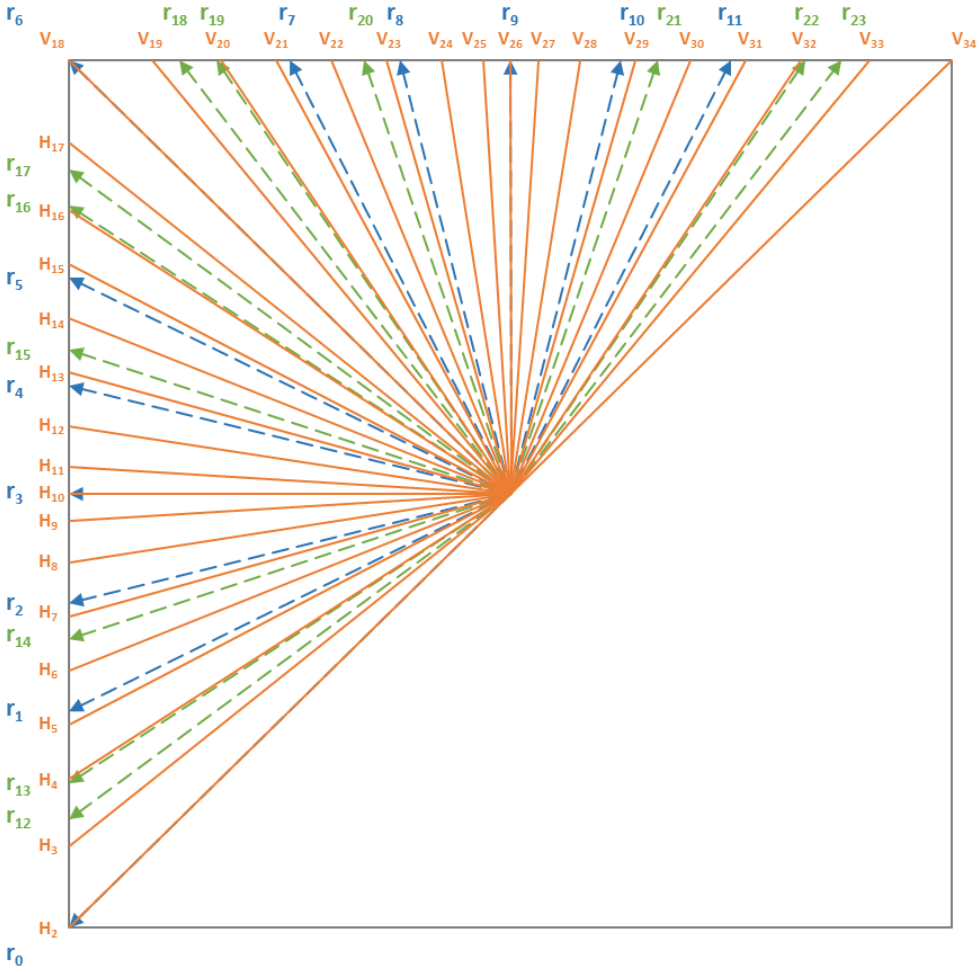


Figure 1. The 33 angular Intra-prediction modes of HEVC (solid orange lines), rational slopes used in [3] (dashed blue lines), and rational slopes added to the previous ones (dashed green lines)

In this paper, we use a block size of 8x8 pixel, which means that the MDV-SW algorithm uses 9x9 blocks, as depicted in Figure 2, where the pixels of the left-column and top-row are the reference pixels from adjacent blocks that have already been coded and decoded. In [3] it was revealed that the use of the original pixel, that means, the non-distorted pixels by the quantization stage, causes a poor orientation detection accuracy.

For each rational slope, there is a set of co-lines, as it can be seen in Table II, which are used to compute the directional variance by applying eq. (2), where $p_j(r_i,n)$ are pixels belonging to the co-line $CL(r_i,n)$, and N is the number of pixels of the n co-line with slope r_i:

$$\sigma^2[CL(r_i,n)] = \frac{1}{N}\left(\sum_{j=0}^{N-1} P_j^2(r_i,n) - \frac{1}{N}\left(\sum_{j=0}^{N-1} P_j(r_i,n)\right)^2\right) \quad \forall n \epsilon \mathbb{Z} \qquad (2)$$

The directional variance values obtained in eq. (2) are then used in eq. (3), which computes the mean directional variance for each set of co-lines with the same rational slope.

$$MDV(PU, r_i) = \frac{1}{L}\sum_{n=1}^{L} \sigma^2[CL(r_i,n)] \qquad (3)$$

The result obtained when the MDV-SW algorithm is used can be seen in Figure 3, which shows the dominant direction of the blocks selected in the synthetic image analyzed.

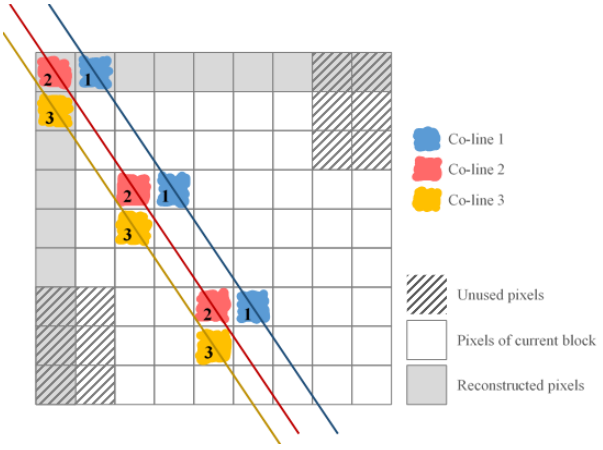


Figure 2. Example of co-lines with slope r_{19} in a 9x9 pixel block

Table II. Co-lines of a 9x9 block

Rational slopes	Number of pixel of each co-line (N)	Number of co-lines (L)
r_0, r_6	2,3,4,5,6,7,8,9,8,7,6,5,4,3,2	15
r_1, r_5, r_7, r_{11}	2,2,3,3,4,4,5,4,5,4,5,4,5,4,5,4,4,3,3,2,2	21
r_2, r_4, r_8, r_{10}	2,2,2,2,3,2,2,2,3,2,2,2,3,2,2,2,3,2,2,2,3,2,2,2,3,2,2,2,3,2,2,2,2	33
r_3, r_9	9,9,9,9,9,9,9,9,9	9
r_{12}, r_{17}, r_{18}, r_{23}	2,2,2,2,2,2,2,2,2,3,2,2,2,3,2,2,2,3,2,2,2,2,2,2,2,2,2	27
r_{13}, r_{16}, r_{19}, r_{22}	2,2,2,2,2,3,2,3,3,3,3,3,3,3,3,3,3,3,3,2,3,2,2,2,2,2	27
r_{14}, r_{15}, r_{20}, r_{21}	2,2,2,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,2,2,2	27

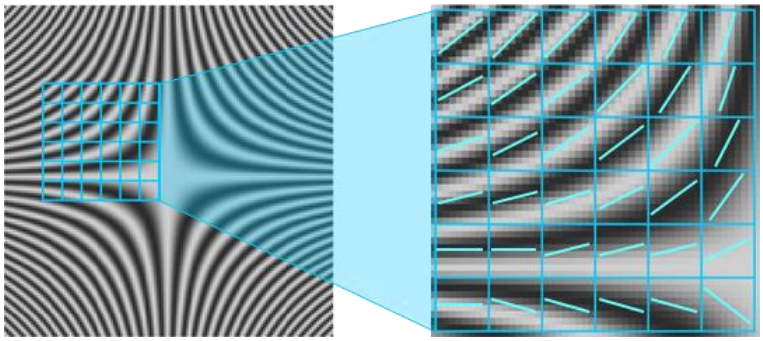


Figure 3. Texture orientation of a synthetic image

4. Parallel implementation

In this section, we are going to detail how the MDV-SW algorithm is implemented to be executed over different parallel architectures.

Block level parallelism

For every block of size NxN in the frame, we need to calculate 24 different gradient directions. So, the algorithm can be summarized as detailed in Figure 4.

```
forEach direction[1-24]
//The number of diagonals for each direction varies with the angle
//The next forEach example is for the rational slopes r1, r5, r7 and r11
   forEach diagonal[1-21]
      //calculate diagonal lines
      find all pixels that belong to the diagonal (Figure 2)
      add all the pixels inside the diagonal ->diagN
      calculate the mean for the pixels in the diagonal mean->diagN/npixels
      calculate the standard deviation of pixels in the diagonal (Equation 1)
    //average per gradient direction
    average[direction]=sum[diagonals]/num_diagonals (Equation 2)
 //calculate the minimum of the 21 directions
 IntraPred[BlockN]= min(average[direction])
```

Figure 4. Pseudocode

The MDV-SW direction mode for the entire block is therefore the direction that gives the minimal average. For a fixed 8x8 block size, this results in an average of 24x24=576 gradient direction calculations (24.5 is a mean value, because not all the rational slopes have the same number of diagonals). Since we have four different block sizes (4x4, 8x8, 16x16 and 32x32) in total, if we assume that the average number of diagonals is 24 for all of them, we will need 4 x 576 = 2304 gradient directions calculation (this is an estimated value, as the number of diagonals depend on the size of the blocks); the directions do not present any dependency among each other and can be calculated in parallel. In addition, we are decoding HD video where every frame contains 14399 blocks of size 8x8 and there are 30 frames per second in video, which means that for a 1 minute of encoded video we need to calculate the Intra-prediction for around 30 million blocks. Any improvement in execution time of the encoder is clearly hugely beneficial.

We chose to assign one thread per gradient line calculation which results in 24x24, 576 threads for an 8x8 blocks and a total of 2304 threads to calculate all blocks sizes. Every thread performs around 17 operations (multiplications and additions); this will lead to low computations done per thread. Part of the article purpose was to determine if the use of a multicore architecture, like Intel's Xeon [4], would be better than a many-core architecture, like the Nvidia GPUs [5]. We decided to use OpenMP [6] to program a multicore Intel's Xeon E5 architecture, using a standard open source compiler like gcc [7], and also an Intel's proprietary compiler, the Intel Parallel Studio icc [8]. We also used two different compilers to evaluate the many-core architecture: Nvidia's nvcc compiler, which requires the algorithm to be implemented using the low-level programming language/API called CUDA C [9], and the OpenACC [10] compiler, which is not proprietary, supports multiple GPU platforms and is pragma based high-level programming language. In the next section, we explain the different implementations over these architectures.

CPU multicore acceleration using OpenMP

Intel's Xeon E5 family is a multicore CPU architecture with 12 cores and HyperThreading, allowing a maximum of 24 threads of simultaneous execution. Intel's recommend the use of their proprietary C compiler, icc, to optimize the hardware specifically for fully exploiting vectorization and data alignment. The primary advantages of the Xeon CPUs are their multi-socket capabilities, higher core counts, larger cache memory, and support for ECC memory. A Xeon Phi co-processor, or co-processors, can be used alongside an existing Intel Xeon processor to provide increased/improved computing power. The architecture also allows the use of common parallel programming languages OpenMP.

Since most multicore architectures commercially available have around 4-12 cores per thread, it does not make any sense to run 504 threads on them because it will only slow the execution time. So, we decided to run 24 threads and let the compiler choose dynamically (by using the pragma schedule) how many operations we want to perform per thread. The main areas accelerated by OpenMP pragmas are:

1. `#pragma omp task`, to calculate the different blocks directions
2. `#pragma omp`, for on the one outside loop in *Figure C*.

a.1 Using gcc compiler
Compiler command and flags: **gcc -g -Wall -03 sourcefilename -fopenmp**
a.2 Using icc compiler
Compiler command and flags: **icc -g -Wall -03 -vec-report=3 -align sourcefile –qopenmp**

GPU Many-core implementation using CUDA C and OpenACC

Nvidia GPUs is a many-core architecture used in modern computers as a coprocessor to specifically accelerate graphics (video and image). Programming for these heterogeneous computer systems has been an area of research for many years to accelerate scientific computations. Since the threads used by these architectures are mostly hardware managed, the creation of the thread is basically free and their parallelism model is based on the assumption that you can run hundreds of thousands of threads in parallel. The performing improvements provided by this architecture on many different scientific computations resulted in the development of several low-level APIs (CUDA, OPENCL, etc.). Programming these APIs can take time and usually requires some level of expertise to implement them. To solve this problem, recent approaches employ directive-based high-level programming like OpenACC. OpenACC is a pragma directive based programming language designed to allow easy development for a variety of hardware accelerators, including GPUs from different vendors, multicore architectures and FPGAs. The goal of the OpenACC compiler is to improve the execution time of existing code written in Fortran, C or C++, by adding different pragma directives to the code that will allow it to run on the available accelerator. OpenACC allows programmers to quickly develop new architectures without the need to understand much of the hardware or the need to learn new vendor specific programming languages. However, since the compiler will take most of the decisions, the performance speed is usually lower than the one that can be by using from hardware specific programming languages and compilers. OpenACC support is provided by a number of vendors and is defined by an open standard.

Since we can run “many” threads in a GPU, initially we chose to calculate a diagonal per thread. However, each line is at maximum nine additions (r_3 and r_9), so the work done per thread was very low and we decided to perform two diagonals from same direction per thread.

CUDA C

Each CUDA block calculates one direction. This way, we can copy the block into shared memory reducing the global memory bandwidth bottleneck. Each thread inside the CUDA block calculates two of the diagonals. We cannot optimize for divergence, since each direction requires different pixels in the block and the calculations are slightly different - each thread performs two line and lines differ in size. The average of the standard deviations is performed as a reduction to avoid atomic adds and obtain maximum thread parallelism.

Compiler command and flags: **nvcc -arch=sm_55 sourcefile.cu**

OpenACC

The implementation in OpenACC is based on the C implementation and mainly consists in substituting the OpenMP pragmas described in a.1 by openACC pragmas. The for loops in the Intra-prediction are perfect to be accelerated by OpenACC directives, !$acc for or !$acc kernels. We added !$acc data specific pragmas to specifically express if used data should be copy on the GPU and remain there; this way we improve the communication bottleneck.

Compiler command and flags: **pgcc -acc -ta:nvidia:managed -Minfo=accel sourcefile.c**

Analysing the compiler output we can notice that the number of threads created by the GPU is below 400, which confirms that is beneficial to perform more than one diagonal calculation per thread.

5. Simulation results

If we compare the timings presented in Figure 5 for a block size of 8x8, the CUDA implementation is clearly the best one. We expect the CUDA implementation to perform even better if the video encoded is 4K instead of HD; and even more if instead of doing block level parallelism we do parallelism at the Image Level calculating all blocks Intra directions at the same time.

We find also very interesting the developing time taken to implement each solution. The summary of the lines of codes written for each implementation are presented in Table III. The solution using CUDA is the most complex one to develop. The solution in OpenACC just took around 1 hour to develop, since the code was based in a C implementation. We do expect to have a performance overhead when going from a low-level API to a high-level pragma based API, but the results are competitive and if development time or lack of a GPU

programming experts is a constrain it is a very viable option for scientists to try use GPU with OpenACC.

The multicore implementation also shows very good improvements over the sequential implementation, which means that even if accelerators like GPUs are not available, video encoding can easily be speed up without much rewriting of code, just by adding very common pragmas directives.

Table III. Lines of code needed to program for the specific compiler/ Programming language

Compiler + Library	Number of lines of code added
C (original code)	1000
C + OpenMP	4 extra line more than the original code
C + OpenACC	2 extra lines more than original code
CUDA	Complete rewrite of original code 1000. Uses Shared memory and reduction

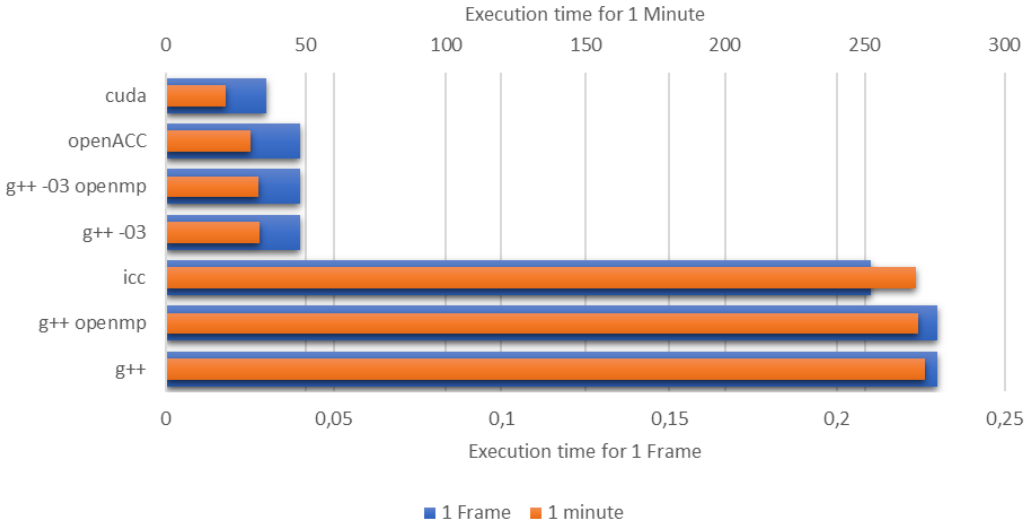


Figure 5. Execution times with parallel implementation, using an 8x8 block size

6. Conclusions

In this paper, we have presented an improved and parallel version of the MDV-SW algorithm, which is based on the dominant gradient detection. The experimental results proved that, even if the MDV-SW algorithm is not implemented within the HEVC reference software, it can be easily implemented to process raw video files over parallel architectures and obtain the texture orientation. Our approach obtained a very significant computational complexity reduction with no quality losses from a perceptual point-of-view. The MDV-SW

algorithm could also be used in a HEVC encoder scenario, since it reduces the number of Intra-prediction modes to be tested by a HEVC encoder.

Block level parallelism achieves a significant improvement over sequential execution. We did evaluate different architectures and programming paradigms to provide the optimal solution for different situations. Therefore, it will allow us to develop hardware/software systems to implement the MDV-SW algorithm focused in low-power and consumption devices.

We leave for future work how to solve the dependencies between prediction blocks in order to make the parallel implementation at Image Level possible.

7. Acknowledges

The MINECO and European Commission (FEDER funds) supported this work under the project TIN2015-66972-C5-2-R.

8. References

[1] High Efficiency Video Coding, Rec. ITU-T H.265 and ISO/IEC 23008-2, Jan. 2013.

[2] Advanced Video Coding for Generic Audiovisual Services,Rec.ITU-T H.264 and ISO/IEC 14496-10 (MPEG-4 AVC), 2012.

[3] DAMIÁN RUIZ, GERARDO FERNÁNDEZ-ESCRIBANO, JOSÉ LUIS MARTÍNEZ, PEDRO CUENCA, *Fast Intra mode decision algorithm based on texture orientation detection in HEVC*, Signal Processing: Image Communication 44 (2016) 12-28.

[4] Intel's Xeon E5 family. http://www.intel.com/content/www/us/en/products/processors/xeon/e5-processors.html

[5] Nvidia Maxwell GPU Architecture. https://developer.nvidia.com/maxwell-compute-architecture.

[6] OpenMP Specification for Parallel Programming, http://www.openmp.org/.

[7] Using GNU Compiler Collection. https://gcc.gnu.org/onlinedocs/gcc/

[8] Intel Parallel Studio XE with Intel C++ compiler. https://software.intel.com/en-us/c-compilers.

[9] CUDA C Programming Guide v 8.0. http://docs.nvidia.com/cuda/cuda-c-programming-guide/#axzz4eY3Ta3OR.

[10] The OpenACC Application Programming Interface, http://www.openacc-standard.org, November 2011-2017.

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4-8 July, 2017

Augmenting Complex Networked Systems under Improved Pinning Controllability Condition

Mahdi Jalili
School of Engineering, RMIT University, Melbourne, Australia

email: mahdi.jalili@rmit.edu.au

Abstract

In this paper we introduce a method for efficient augmentation of networked systems provided that the pinning controllability of the final network is improved. The problem is how to connect a sub-network to an already existing network such that the controllability is maximised. We consider the eigenratio of the augmented Laplacian matrix as a pinning controllability metric, and use graph perturbation theory to approximate influence of edge addition on the eigenratio. The resulting metric can be efficiently used to find the links connecting two disjoint networks. We also provide numerical simulations on synthetic networks and show that the proposed method is more effective than heuristics such as connecting nodes with high degrees or betweenness centrality values.

Key words: complex networks, pinning controllability, network augmentation, scale-free networks, small-world networks

1. Introduction

Many natural and man-made systems can be modelled as networks where a number of individual entities, known as agents, nodes or vertices, are coupled through connecting links. Real networks have been shown to share some common structural properties such as community structure, densification, power-law degree distribution and small-world property [1]. There is overwhelming interest to study various dynamical phenomena on complex networks. Examples include synchronisation [2], consensus [3], information propagation [4] and cascaded failures [5]. Synchronisation is the most well-known collective behaviour studied in networked systems, where a number of individual dynamical systems coordinate their motion through a connection graph. It has been shown that

structure of the connection graph has a major role in determining its synchronisation properties [6].

In some applications one needs to control dynamics of a networked system to a specific set point. Although dynamical networks can synchronise without presence of an external input, one can speed up the synchronisation by controlling some of the nodes. This is referred to as pinning control (or pinning synchronisation) in the literature, where external input is fed into (often) small fraction of nodes, known as driver nodes, to direct the dynamics of the whole network into a reference state [7, 8]. A number of approaches have been proposed to measure pinning controllability of networked systems. Eigenratio of the augmented Laplacian matrix of the connection graph in as effective metric measuring pinning controllability of networks [8-10]. This method is based on the assumption of having the same dynamics for the reference state and individual nodes.

Engineering network structure to have optimal levels of pinning controllability is important in many applications for which the connections graph is required to have high controllability. A number of methods have been proposed to improve pinning controllability of networks. For example, one can employ link weighting [11] or efficient rewirings [12] to enhance the pinning controllability. Another important issue in this field is to determine the best drivers. Controlling a node involves significant cost, and one is often required to achieve the control task with minimum number of drivers, or the best drivers if their number is fixed. The pinning control with best drivers has been modelled as an optimisation problem and evolutionary optimisation methods have been used to solve the problem to obtain the list of best drivers [9, 13]. Moradi Amani et al. proposed a simple metric based on the eigenvectors of the Laplacian matrix to find the most influential nodes for pinning control [14].

In this paper, we propose an approach for effective augmentation of complex networks under improved pinning controllability conditions. Network augmentation is a phenomenon with significant engineering applications [15]. For example, when a new suburb is added to the energy grid, the question is how to create the connections between the two networks provided that a certain utility function is maximised. Here we consider the case when a new sub-network is added to an already existing networked with known driver nodes. The question is how to create inter-network links between these two networks such that high level of pinning controllability is obtained in the final network.

2. Pinning Controllability of Complex Dynamical Networked Systems

Here we consider coupled identical dynamical systems where all individual nodes follow the same dynamics. Let's consider a network with N nodes, all having the

same dynamics. When there is no pinning control, the dynamics of the motion of coupled systems reads

$$\frac{d\boldsymbol{x}_i}{dt} = F(\boldsymbol{x}_i) + \sigma \sum_{j=1}^{N} a_{ij} H\left(\boldsymbol{x}_i - \boldsymbol{x}_j\right) \quad ; \quad i = 1, 2, ..., N \,, \tag{1}$$

where $\boldsymbol{x}_i \in R^d$ are d-dimensional state vectors, $F: R^d \to R^d$ defines the (uniforms) dynamics of each node, σ is a coupling strength and H is a projection matrix indicating from which dimensions the individual dynamical systems are coupled to one another. $A = (a_{ij})$ is the adjacency matrix determining the connection graph; $a_{ij} = a_{ji} = 1$ when there is a link between nodes i and j, and $a_{ij} = a_{ji} = 0$ otherwise. We suppose that there are no self-loops in the network, i.e., $a_{ii} = 0$. One can rewrite the above equation as

$$\frac{d\boldsymbol{x}_i}{dt} = F(\boldsymbol{x}_i) - \sigma \sum_{j=1}^{N} l_{ij} H \boldsymbol{x}_j \quad ; \quad i = 1, 2, ..., N \,, \tag{2}$$

where $L = (l_{ij})$ is Laplacian matrix of the connection graph; $l_{ij} = -a_{ij}$ for $i \neq j$, and $l_{ii} = k_i$, where k_i is degree of node i, which is calculated by summing all connections pointing to this node.

In pinning control (or synchronisation), often small fraction of nodes are considered as drivers to which the external control input is fed. The final aim of pinning control is to force the individual dynamical units to follow a reference state. Let's suppose that the reference state $s(t)$ has the same dynamics as the individual dynamical nodes

$$\frac{d\boldsymbol{s}(t)}{dt} = F\left(\boldsymbol{s}(t)\right) \,. \tag{3}$$

If the difference between the trajectories of the individual unites and the above reference state goes to zero as time goes to infinity, one can argue that global pinning synchronisation is obtained. To achieve this, one needs to apply the following control input to driver i

$$u_i = \sigma \beta_i k_i \left(\boldsymbol{x}_i - \boldsymbol{s}\right) \,. \tag{4}$$

where k_i is the feedback control gain, and β_i indicates the driver nodes such that $\beta_i = 1$, if node i is a driver node, and $\beta_i = 0$ otherwise. Therefore, equations of motion of the controlled networked system read

$$\frac{d\boldsymbol{x}_i}{dt} = F(\boldsymbol{x}_i) - \sigma \sum_{j=1}^{N} l_{ij} H \boldsymbol{x}_j - \sigma \beta_i k_i (\boldsymbol{s} - \boldsymbol{x}_i) \quad ; \quad i = 1, 2, ..., N \ , \qquad (5)$$

An important problem in designing pinning control strategy is to determine the parameters that guarantee stability of the synchronised solution $\boldsymbol{x}_1(t) = \boldsymbol{x}_2(t) = \ldots = \boldsymbol{x}_N(t) = \boldsymbol{s}(t)$. A number of approaches have been proposed to obtain sufficient or necessary conditions for the stability of synchronised solution in pinning control. The approach based on master stability function is a prime choice in this field [8, 12, 13, 16], which is based on studying linear stability of variational equations. It leads to a criterion for linear stability of the synchronised solution, i.e., a necessary stability condition. This solution is based on the eigen-decomposition of the augmented Laplacian matrix, defined as

$$L_a = \begin{pmatrix} l_{11} + k_1\beta_1 & l_{12} & \ldots & l_{1N} \\ l_{21} & l_{11} + k_1\beta_2 & \ldots & l_{2N} \\ \vdots & \vdots & \ldots & \vdots \\ l_{N1} & l_{N2} & \ldots & l_{NN} + k_N\beta_N \end{pmatrix}. \qquad (6)$$

The augmented Laplacian matrix is indeed the original Laplacian matrix (for the case without any pinning control input) with added diagonal terms associated with the pinning control. It has been shown that the eigenvalues of the augmented Laplacian matrix are non-zero and positive as $0 < \lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_N$ for connected networks. The eigenratio $R = \lambda_N / \lambda_1$ is considered as a metric measuring the pinning controllability such that the smaller the eigenratio R is, the better the pinning controllability is. Indeed this metric is based on the interpretation that "network N_1 is more controllable than network N_2, if for a larger range of coupling parameter σ it is possible to control N_1 as compared to N_2" [12]. An interesting thing about the eigenratio is that is depends solely on the structure of the connection graph and is independent of dynamics of individual dynamical systems.

3. Effective Network Augmentation under Optimal Controllability Condition

Our problem here is how to efficiently augment networks improving their pinning controllability. We assume that there is an existing network with fixed topology and predetermined driver nodes; let's denote this network by N_A. We then consider a new sub-network, denoted by N_B, which is required to be connected to

N_A. There will be I inter-network connections to be made between the two networks. The problem is to find the tipping nodes from N_A and N_B such as that creating I inter-network links between them results in high levels of pinning controllability as measured by the eigenratio of the augmented Laplacian of the whole network. Here we use graph perturbation theory [12, 17] in order to find an approximate solution to this problem, and show that it is more effective than heuristics such as connecting high-degree nodes from both sides.

Let's denote the eigenvalues and eigenvectors of the augmented Laplacian L_A by λ_i and $\boldsymbol{u}_i$, respectively. Let's suppose that adding an edge results in the changes in L_A, λ and $\boldsymbol{u}$ as ΔL_A, $\Delta\lambda$ and $\Delta\boldsymbol{u}$, respectively. Considering the basic definition of eigen-decomposition problem and some simple operations, one obtains

$$\Delta\lambda_i = \frac{\boldsymbol{u}_i^T \Delta L \boldsymbol{u}_i}{\boldsymbol{u}_i^T \boldsymbol{u}_i} \quad , \quad i = 1, 2, ..., N \, . \tag{7}$$

The pinning controllability index depends on λ_1 and λ_N. If a link is created between two non-connected nodes i and j, it is simple to show that the change in these eigenvalues due to this link addition can be approximated as

$$\begin{cases} \Delta\lambda_1 \approx (\boldsymbol{u}_{1i} - \boldsymbol{u}_{1j})^2 \\ \Delta\lambda_N \approx (\boldsymbol{u}_{Ni} - \boldsymbol{u}_{Nj})^2 \end{cases}, \tag{8}$$

where $\boldsymbol{u}_{1i}$ and $\boldsymbol{u}_{Ni}$ are the i-th entry of the eigenvectors corresponding to λ_2 and λ_N, respectively. Using equation (8) and assuming $\Delta\lambda_1 << \lambda_1$, the change in the eigenratio R by adding a link between nodes i and j can be approximated as

$$\Delta R \approx \frac{\lambda_1 (\boldsymbol{u}_{Ni} - \boldsymbol{u}_{Nj})^2 - \lambda_N (\boldsymbol{u}_{1i} - \boldsymbol{u}_{1j})^2}{\lambda_1^2} \, . \tag{9}$$

The above equation is the key equation to use here. One can first create a random link between networks N_A and N_B to make them a connected network, provided that networks N_A and N_B are connected networks themselves. Then, one should make an eigen-decomposition and obtain the above value for any inter-network links. Finally, the top-I values are selected and inter-networks are created based on them. Note that one has to remove the link initially created between the two networks.

4. Numerical Simulations

In order to show the effectiveness of the proposed method, we include numerical simulations on synthetic networks. As network structures for both networks N_A and N_B, we considered scale-free networks constructed as follow. First, an all-to-all connected network with m nodes is considered. Then, at each step a new node is added to the network and creates m connections. The probability that an old node is connected to the newly added node is proportional to its degree; the higher the degree of an old node is, the higher the probability of being connected to the new node. The networks evolved using this algorithm will have scale-free degree distribution, meaning that many nodes will have small degrees, while a few nodes will have large degree values. Furthermore, the average degree of the network will be almost $2m$.

We construct the existing network N_A with size $N = 1000$ and $m = 5$. The network N_B that is augmented in the existing network is considered with different size and average degree. $I = 10$ inter-network connections are created. We compare the performance of the proposed method with heuristic approaches including connecting the nodes with the highest degree or betweenness centrality from each side. Figures 1 and 2 shows the results, where outperformance of the proposed method is clear. Note that as the eigenratio R get lower values, the pinning controllability is better.

5. Conclusions

Control and synchronisation of complex networks has many potential applications in science and engineering. In this paper we considered the problem of network augmentation subject to maximising pinning controllability of the final network. A new sub-network with a specific structure is connected to an existing network with fixed topology and driver nodes. The problem is to find nodes from these two networks to create a number of inter-network connections. We used the eigenratio of the augmented Laplacian matrix, the largest eigenvalue divided by the smallest one, as a metric measuring pinning controllability of networks. Networks with lower eigenratio have better pinning controllability. We proposed a method based on graph perturbation theory and introduced a metric to find the best inter-network links. Our numerical simulations on synthetic networks showed that our proposed method outperforms heuristic methods including connecting the nodes with the highest degree or betweenness centrality from each side.

Acknowledgments

This research is supported by Australian Research Council through project No DP170102303.

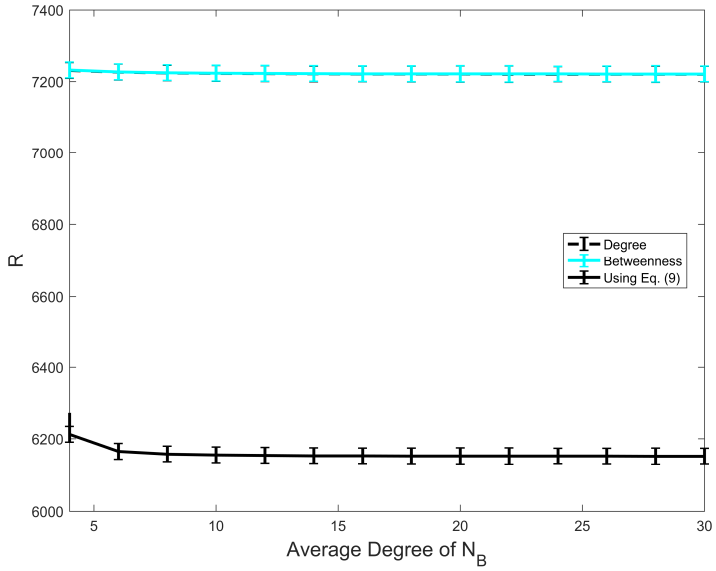


Figure1: The eigenratio R as a function of average degree (m/2) of network N_B. The size of network N_B is set to 200. I = 10 inter-networks are created based on the proposed method (black line), by connecting high degree nodes from each side (dashed black line), or by connecting those with high betweenness centrality (cyan line). Data show mean values with errorbars corresponding to the standard deviation over 50 realizations.

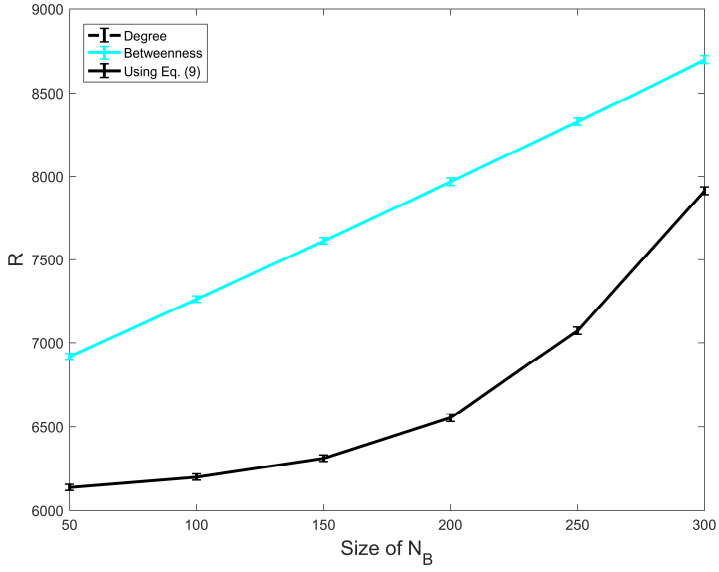


Figure 2: The eigenratio R as a function of size of network N_B. We fix m = 5. Other designations are as Figure 1.

References

[1] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, and D. U. Hwang, "Complex networks: structure and dynamics," *Physics Reports,* vol. 424, pp. 175-308, 2006.

[2] A. Arenas, A. Díaz-Guilera, J. Kurths, Y. Moreno, and C. Zhou, "Synchronization in complex networks," *Physics Reports,* vol. 469, no. 3, pp. 93-153, 2008.

[3] R. Olfati-Saber, J. A. Fax, and R. M. Murray, "Consensus and cooperation in networked multi-agent systems," *Proceedings of the IEEE,* vol. 95, no. 1, pp. 215 - 233, 2007.

[4] D. Easley, and J. Kleinberg, *Networks, crowds, and markets*: Cambridge University Press, 2010.

[5] B. Mirzasoleiman, M. Babaei, M. Jalili, and M. A. Safari, "Cascaded failures in weighted networks," *Physical Review E,* vol. 84, pp. 046114, 2011.

[6] M. Jalili, "Synchronization in dynamical networks: synchronizability, neural network models and EEG analysis," EPFL, Lausanne, 2008.

[7] J. Zhou, J.-a. Lu, and J. Lü, "Pinning adaptive synchronization of a general complex dynamical network," *Automatica,* vol. 44, no. 4, pp. 996-1003, 2008.

[8] F. Sorrentino, M. di Bernardo, F. Garofalo, and G. Chen, "Controllability of complex networks via pinning," *Physical Review E,* vol. 75, pp. 046103, 2007.

[9] M. Jalili, O. Askari Sichani, and X. Yu, "Optimal pinning controllability of complex networks: Dependence on network structure," *Physical Review E,* vol. 91, no. 1, pp. 012803, 2015.

[10] L. Gao, X. Liao, and H. Li, "Pinning controllability analysis of complex networks with a distributed event-tiggered mechanism," *IEEE Transactions on Circuits and Systems II: Express Briefs,* vol. 61, no. 7, pp. 541-545, 2014.

[11] Y. Orouskhani, M. Jalili, and X. Yu, "Optimizing dynamical network structure for pinning control," *Scientific Reports,* vol. 6, pp. 24252, 2016.

[12] M. Jalili, and X. Yu, "Enhancing pinning controllability of complex networks through link rewiring," *IEEE Transactions on Circuits and Systems II: Express Briefs*, 2016.

[13] Y. Tang, H. Gao, J. Kurths, and J.-a. Fang, "Evolutionary pinning control and its application in UAV coordination," *IEEE Transactions on Industrial Infromatics,* vol. 8, no. 4, pp. 828-838, 2012.

[14] A. Moradi Amani, M. Jalili, L. Stone, and X. Yu, "Finding the most influential nodes in pinning controllability of complex networks," *IEEE Transactions on Circuits and Systems II: Express Briefs*, 2016.

[15] J. Wang, X. Yu, and L. Stone, "Effective augmentation of complex networks," *Scientific Reports*, vol. 6, pp. 25627, 2016.
[16] F. Sorrentino, "Effects of the network structural properties on its controllability," *Chaos*, vol. 17, pp. 033101, 2007.
[17] M. Jalili, and X. Yu, "Enhancement of synchronizability in networks with community structure through adding efficient inter-community links," *IEEE Transactions on Network Science and Engineering*, vol. 3, no. 2, pp. 106-116, 2016.

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4-8 July, 2017

High Performance Parallel Implementation of the Jaya Optimization Algorithm: a Manycore GPU Approach

Antonio Jimeno-Morenilla[1], José-Luis Sánchez-Romero[1], Héctor Migallón[2] and Higinio Mora-Mora[1]

[1] *Department of Computer Technology, University of Alicante*
[2] *Department of Physics and Computer Architecture, Miguel Hernandez University*

emails: jimeno@dtic.ua.es, sanchez@dtic.ua.es, hmigallon@umh.es, hmora@dtic.ua.es

Abstract

The aim of optimization methods consists in searching for an optimal value for a specific function within a constrained or unconstrained domain. These methods are useful for a wide range of scientific and engineering applications. Recently, a new optimization method called Jaya has generated a growing interest due to its simplicity and efficiency. In this paper, a parallel version of Jaya is developed and implemented for manycore platforms. The performance of the parallel implementation is compared with that of the sequential version of the algorithm when optimizing a function benchmark. The results indicate that the parallel Jaya algorithm achieves a significant speed-up for each of the functions, up to 190x.

Key words: Jaya, optimization, parallelism, GPU, CUDA

1. Introduction

Optimization methods are devoted to find an optimal value for a given function, generally a minimum. Each function to be optimized has its specific domain, behaviour and number of variables involved. Indeed, some of these functions have local minima, so the find of the absolute optimum can become very difficult.

Optimization methods can be mainly divided into deterministic and heuristic approaches. Deterministic approaches take advantage of the analytical properties of the function (see [1]). When coping with non-convex or large-scale optimization problems, determining the global optimum may become a very

complex task. In this case, heuristic methods should be used since they are usually more flexible and efficient than deterministic ones, and the computational time required to find the optimum can be highly reduced.

Heuristic optimization methods are classified into Evolutionary Algorithms (EA) and Swarm Intelligence (SI) algorithms. Among the EA methods, it is worthwhile mentioning Genetic Algorithm (GA), Differential Evolution (DE), Evolutionary Strategy, and some others. Among the SI methods (see [2]), it is worthwhile mentioning Particle Swarm Optimization (PSO), Artificial Bee Colony (ABC), and some others. Other methods based on nature phenomena have been developed, such as Harmony Search (HS), Biogeography-Based Optimization (BBO), and some others (see [3, 4]). The success of most of the mentioned algorithms is greatly conditioned by their specific parameters. The proper tuning of these parameters is a crucial factor for an efficient find of the global optimum.

Recently, two optimization methods have been proposed, namely TLBO (Teacher-Learner Based Optimization) [5] and Jaya [6]. Both optimizations algorithms have the advantage of not needing specific parameter tuning. They only require general parameters such as the number of iterations and the population dimension. Although they are very similar, TLBO uses two phases each one of iterations, whereas Jaya only performs one phase each one of the iterations. The Jaya algorithm has generated a growing interest in many scientific and engineering areas due to its simplicity and efficiency, see for example [7-15].

Optimization algorithms have been usually implemented on computer systems following a traditional, sequential approach. However, most of these algorithms are feasible to be decomposed into independent tasks and executed in parallel. In the last years, the performance of parallel hardware architectures has greatly increased, while their cost has been highly reduced. Nevertheless, parallelizing an algorithm is not a simple task since it requires a reformulation and adequacy to the specific architecture to be used. Among these architectures, it is worthwhile mentioning the multiprocessors and the manycore Graphics Processing Units (GPU). GPUs are originally dedicated to graphics processing but, since they have become massively parallel resources, they are suitable to be applied to other high performance processing tasks. Some research works can be found in the literature which demonstrate the advantages of executing parallel implementations of optimization algorithms on multiprocessors (see for example [16-20]) and on manycore GPUs (see for example [21-26]).

2. The Jaya Algorithm

As mentioned earlier the Jaya algorithm has the advantage of not requiring specific tuning parameter: only population size (number of different individuals) and generations (number of iterations) should be configured. This algorithm is based on the fact that the optimal solution for a given problem can be obtained moving towards the best partial solution and, at the same time, avoiding the worst

solution. Compared with other optimization methods such as GA, ABC, DDE, PSO, and TLBO, Jaya obtained better results in terms of best, mean, and worst values of different unconstrained benchmark functions [27].

The description of the Jaya algorithm is as follows. Let $f(x)$ be the objective function to be minimized (or maximized). At any iteration i, assume that there are n design variables (i.e. $j = 1, 2, \ldots, n$) and p candidate solutions (i.e. population size, $k = 1, 2, \ldots, p$). The *best* candidate obtains the best value of $f(x)$ (i.e. $f(x)_{best}$) in the whole candidate solutions, and the *worst* candidate obtains the worst value of $f(x)$ (i.e. $f(x)_{worst}$) in the whole candidate solutions. If $X_{j,k,i}$ is the value of the jth variable for the kth candidate during the ith iteration, then this value is modified by means of the following equation:

$$X'_{j,k,i} = X_{j,k,i} + r_{1,j,i} \left(X_{j,best,i} - \left| X_{j,k,i} \right|\right) - r_{2,j,i} \left(X_{j,worst,i} - \left| X_{j,k,i} \right|\right) \quad (1)$$

where $X_{j,best,i}$ is the value of the variable j for the *best* candidate, and $X_{j,worst,i}$ is the value of the variable j for the *worst* candidate.

In Equation (1) $X'_{j,k,i}$ is the updated value of $X_{j,k,i}$, and $r_{1,j,i}$ and $r_{2,j,i}$ are two random numbers in the range [0, 1], for the jth variable computed in the ith iteration. The term $r_{1,j,i} \left(X_{j,best,i} - \left| X_{j,k,i} \right|\right)$ indicates the tendency of the algorithm to move closer to the best solution, whereas the term $- r_{2,j,i} \left(X_{j,worst,i} - \left| X_{j,k,i} \right|\right)$ indicates the tendency of the algorithm to avoid the worst solution. Obviously, the new candidate ($X'_{j,k,i}$) is accepted only if it gives a better function evaluation. All the accepted function values at the end of each one of the iterations are maintained, so these values become the input to the next iteration.

3. Parallelization of Jaya

The Jaya algorithm has inherent parallel features which can be exploited. On the one hand, each candidate solution (individual $k = 1, 2, \ldots, p$ in the algorithm) into the population can independently perform the function evaluation. Moreover, each design variable ($j = 1, 2, \ldots, n$ in the algorithm) can update its value, taking into account the current best and worst values.

On the other hand, the Jaya algorithm performs some independent executions (*Runs*) of the algorithm, in our proposal we try to execute all these executions simultaneously. Considering all the computed solutions, the statistical data about the results (best, worst and mean solution, and also standard deviation) are the algorithm output.

Our parallel Jaya algorithm was designed using CUDA 7.5. Two different Nvidia Maxwell GPUs were used to evaluate the parallel performance respect to the sequential implementation, evaluated on two general purpose Intel processors. The GPUs used were the Nvidia GTX950 (768 CUDA cores, 1.025 GHz, 2 GB memory) and the GTX970 (1,664 CUDA cores, 1.05 GHz, 4 GB memory). Intel processors used were the i7-4790 (4 cores, 3.6 GHz) and the i7-6700 (4 cores, 3.4 GHz).

Both GPUs used are Maxwell Nvidia GPUs, which therefore are composed by Maxwell Streaming Multiprocessors (SMM), each one with 128 CUDA cores. Regarding the Jaya algorithm, at each iteration the updated whole population have to be shared for all threads involved in the computing, thus, in order to obtain good parallel behaviour, the population data should be stored in the shared memory. Given that the shared memory is owned by each SMM, each independent execution of the algorithm is mapped on one SMM, thus the number of thread blocks in the grid of the kernel launched is equal to the desired number of independent executions (*Runs*). On the other hand, the number of threads per block in the grid were configured in a 2D array, being the row size equal to the population size and the column size is equal to the number of design variables, which depends on the function to be optimized. The shared memory is used to store: the whole population, the partial values of the function evaluation, the new candidates, and other data as for example the indices of the current best and worst solution. Each thread performs the updating of one design variable and computes the part of the function evaluation corresponding to the assigned design variable. Only one thread have to perform the reduction process of computing the final value of the global function evaluation for each specific individual. Note that the partial values of the function evaluation have to be stored in the shared memory, increasing the shared memory requirements, to be able to efficiently carry out the last reduction process. Figure 1 depicts the parallel computing of the partial values of the function evaluation for whole population respect to each thread block, where P is the population size and N is the number of the design variables of the function to be optimized. N that the value of P is assigned as tuning parameter while the value of N cannot be modified since it is inherent to the chosen function.

Worthy to note that the shared memory performance and the available amount is a key to obtain good speed-ups. On the one hand the shared memory performance allows share efficiently the data involved in the computing, and on the other hand the amount of shared allows to increase the population size.

4. Experimentation

The comparison between the sequential and the parallel implementations of the algorithm was made taking into account 30 unconstrained functions frequently used as a well-known benchmark in several works about optimization. First, the Rosenbrock function was chosen so as to provide a complete set of statistical results. This function is usually defined with 30 design variables.

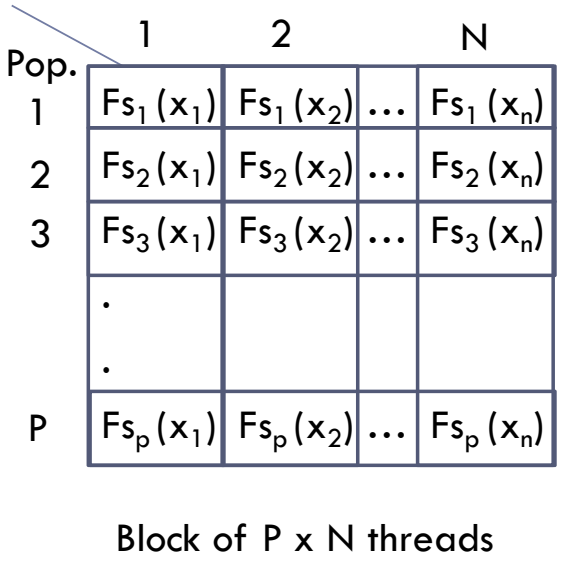


Figure 1. Data computing distribution for one independent run (block).

The number of iterations was fixed to 30,000. Two parameters were modified so as to evaluate the speed-up when compared with the sequential implementation and the error in the function optimization: number of *Runs* (which value sets the number of thread blocks) was varied in the set of values {2, 4, 8, 16, 32, 64, 128, 256}; with regard to population size (number of individuals), its value was varied within the set {8, 16, 32}. Figure 2 shows the speed-up achieved when comparing the parallel implementation on the Nvidia GTX970 GPU with the sequential execution on the Intel i7-4790 processor. The X-axis corresponds to the number of independent executions (*Runs*), whereas each bar from left to right within a specific value of *Runs* corresponds to a population size of 8, 16, and 32 individuals respectively. Obviously, when the value of *Runs* is lower than the number of MSS (13 for the Nvidia GTX970) the GPU cannot be fully occupied obtaining low speed-up values. It can be observed that a maximum speed-up higher than 50x was obtained with 64 *Runs* and a population size of 8. In this case, the error was in the order of 10^{-3} for both the sequential and the parallel implementations. Worthy to note that the parallel performance improves as the population size decreases, this is due to the reduction processes become more significant respect to the total computational cost. Obviously the optimal population size depends on the computational cost of the function evaluation to be optimized.

Second, the full set of unconstrained functions was also optimized by means of the sequential and the parallel version of the Jaya algorithm. Speed-up was calculated by following the same criteria as with the Rosenbrock function with regard of fixed number of iterations, number of independent executions, and population size. Remark that, depending on the function to be optimized, the

value of *Runs* were increased up to 1024, whereas in other cases this parameter had to be decreased to 128 or even 64 due to the features of the GTX950 GPU. Figure 3 shows the maximum speed-up obtained when optimizing the whole benchmark functions. It can be observed that, in some cases, the speed-up is higher than 100x. Indeed, in case of F05 (the Matyas function), the speed-up is near 190x with a population size of 64 and 1024 *Runs*. Table 1 shows the value of *Runs* and the population size related to the maximum speed-up obtained for each one of the benchmark functions.

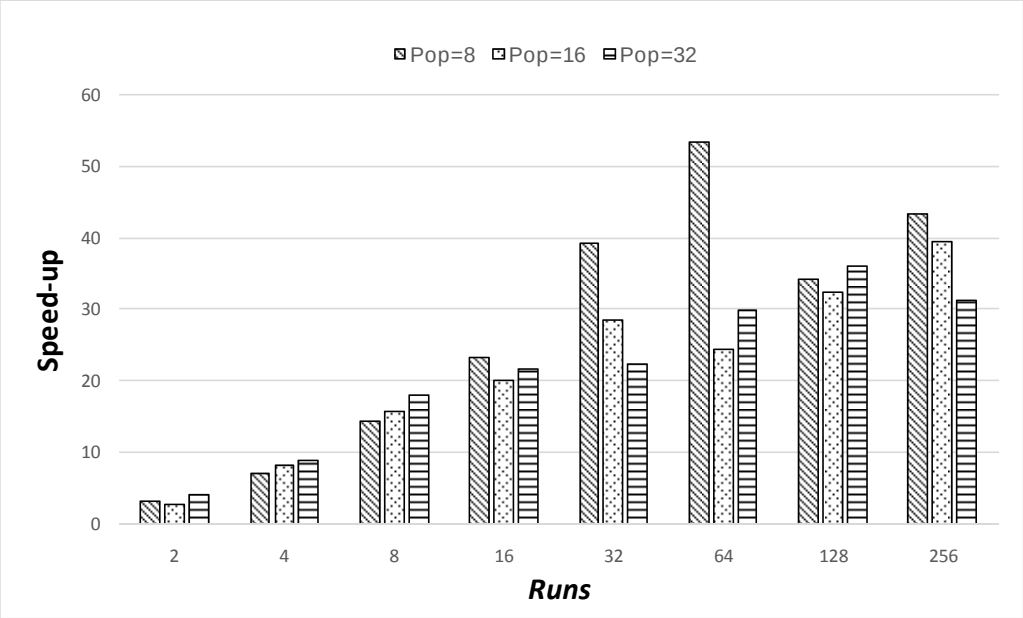


Figure 2. Speed-up obtained for the Rosenbrock function with different independent executions (*Runs*) and population sizes.

5. Conclusions

In this work, a parallelization of Jaya, a recent optimization algorithm, is proposed. The algorithm is implemented in a manycore GPU platform using CUDA, and its performance is compared with that of a conventional sequential version of the algorithm implemented in general purpose processors. Results obtained using two different GPUs are tested by means of a benchmark of 30 unconstrained functions. The results of the experimentation demonstrate that the parallel implementation of Jaya on GPUs provides a significant speed-up when compared with the sequential execution. In the best case, the speed-up rises to 190x, being the mean speed-up for the 30 benchmark functions equal to 53x, whereas the median is 41x.

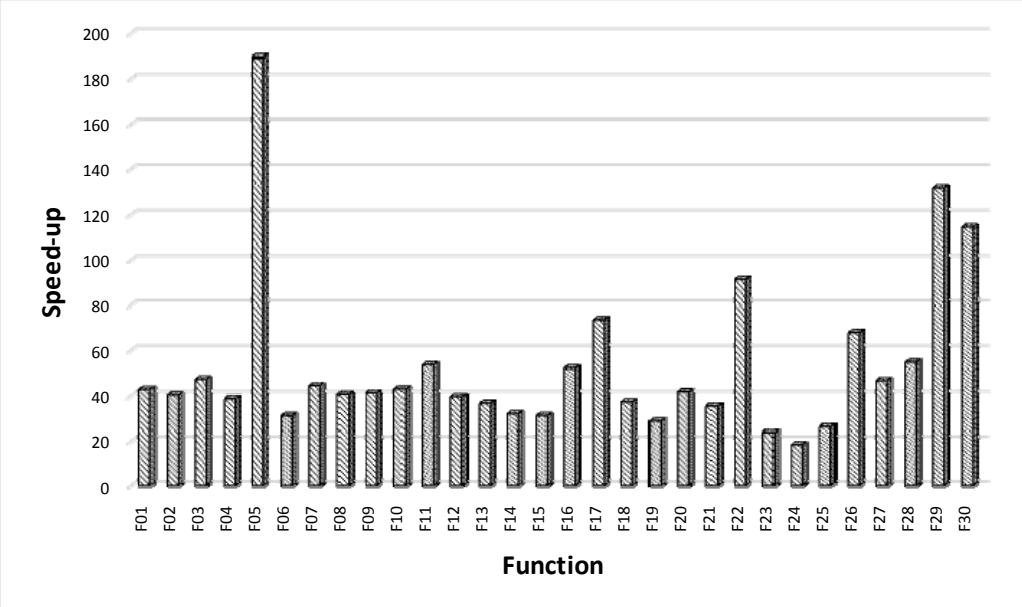


Figure 3. Maximum speed-up for the 30 benchmark functions.

Function	Name of function	Popul.	Runs	Function	Name of function	Popul.	Runs
F01	Sphere	8	256	F16	Booth	256	256
F02	Sum of squares	16	128	F17	Michalewicz_2	64	256
F03	Beale	64	128	F18	Michalewicz_5	128	64
F04	Easom	64	512	F19	Bohachevsky_2	64	256
F05	Matyas	64	1024	F20	Bohachevsky_3	64	512
F06	Colville	32	128	F21	Goldstein-Price	64	128
F07	Trid_6	32	512	F22	Perm	32	512
F08	Trid_10	8	512	F23	Hartmann_3	64	64
F09	Zakharov	32	64	F24	Ackley	8	128
F10	Schweffel_1.2	16	32	F25	Penalized_2	8	128
F11	Rosenbrock	8	64	F26	Langerman_2	64	512
F12	Dixon-Price	8	256	F27	Langerman_5	32	256
F13	Foxholes	64	256	F28	Langermann_10	32	128
F14	Branin	128	256	F29	Fletcher-Powell_5	32	128
F15	Bohachevsky_1	128	512	F30	Fletcher-Powell_1	16	256

Table 1. Population size and *Runs* at the maximum speed-up for the 30 benchmark functions.

As future work, new benchmark functions should be analysed so as to obtain additional performance results of the parallel Jaya algorithm, paying special attention to constrained functions. Furthermore, other optimization algorithms should be parallelized and implemented on manycore platforms and compared to the Jaya algorithm.

6. References

[1] M.-H. LIN, J.-F. TSAI AND C.-S. YU, *A Review of Deterministic Optimization Methods in Engineering and Management*. Mathematical Problems in Engineering, vol. 2012 (2012).

[2] E. BONABEAU, M. DORIGO AND G. THERAULAZ, *Swarm Intelligence: From Natural to Artificial Systems*, Oxford University Press, NY, 1999.

[3] R.V. RAO AND V. PATEL, *An elitist teaching-learning-based optimization algorithm for solving complex constrained optimization problems*. International Journal of Industrial Engineering Computations, 3 (2012) 535–560.

[4] R.V. RAO AND V. PATEL, *Comparative performance of an elitist teaching-learning-based optimization algorithm for solving unconstrained optimization problems*. International Journal of Industrial Engineering Computations, 4 (2013) 29–50.

[5] R.V. RAO, V.J. SAVSANI AND D.P. VAKHARIA, *Teaching–learning-based optimization: A novel method for constrained mechanical design optimization problems*. Computer-Aided Design, 43 (2011), 303–315.

[6] R.V. RAO, *Jaya: A simple and new optimization algorithm for solving constrained and unconstrained optimization problems*. International Journal of Industrial Engineering Computations, vol. 7 (2016) 19–34.

[7] S.P. SINGH, T.PRAKASH, V. SINGH AND M.G. BABU, *Analytic hierarchy process based automatic generation control of multi-area interconnected power system using Jaya algorithm*. Engineering Applications of Artificial Intelligence, vol. 60 (2017) 35–44.

[8] K. GAO, A. SADOLLAH, Y. ZHANG, R. SU AND K.G.J. LI, *Discrete Jaya algorithm for flexible job shop scheduling problem with new job insertion*, International Conference on Control, Automation, Robotics & Vision (ICARCV). IEEE (2016) 1–5.

[9] H. SHAYEGHI, H. SHAYANFAR, S. ASEFI AND A. YOUNESI, *Optimal tuning and comparison of different power system stabilizers using different performance indices via Jaya algorithm*. Proceedings of the International Conference on Scientific Computing (CSC). The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp) (2016) 34.

[10] K. GAO, Y. ZHANG, A. SADOLLAH AND R. SU, *Jaya algorithm for solving urban traffic signal control problem*, International Conference on Control, Automation, Robotics & Vision (ICARCV). IEEE (2016) 1–6.

[11] R. AZIZIPANAH-ABARGHOOEE, M.MALEKPOUR, M. ZARE AND V. TERZIJA, *A new inertia emulator and fuzzy-based lfc to support inertial and governor responses using Jaya algorithm*. Power and Energy Society General Meeting (PESGM) IEEE (2016) 1–5.

[12] K. ABHISHEK, V.R. KUMAR, S. DATTA AND S.S. MAHAPATRA, *Application of Jaya algorithm for the optimization of machining performance characteristics during the turning of cfrp (epoxy) composites: comparison with TLBO, GA, and ICA*. Engineering with Computers (2016) 1–19.

[13]M. BHOYE, M. PANDYA, S. VALVI, I. N. TRIVEDI, P. JANGIR AND S.A. PARMAR, *An emission constraint economic load dispatch problem solution with microgrid using Jaya algorithm*. International Conference on Energy Efficient Technologies for Sustainability (ICEETS), 2016. IEEE (2016) 497–502.

[14]I.N. TRIVEDI, S.N. PUROHIT, P. JANGIR AND M.T. BHOYE, *Environment dispatch of distributed energy resources in a microgrid using Jaya algorithm*. 2016 2nd International Conference on Advances in Electrical, Electronics, Information, Communication and Bio-Informatics (AEEICB). IEEE (2016) 224–228.

[15]S. MISHRA AND P.K. RAY, *Power quality improvement using photovoltaic fed dstatcom based on Jaya optimization*. IEEE Transactions on Sustainable Energy, vol. 7, no. 4 (2016) 1672–1680.

[16]A. J. UMBARKAR, N. M. ROTHE AND A.S. SATHE, *OpenMP Teaching-Learning Based Optimization Algorithm over Multi-Core System*. International Journal of Intelligent Systems and Applications, vol. 7 (2015) 57-65.

[17]A.J. UMBARKAR, M.S. JOSHI AND P.D. SHETH, *OpenMP Dual Population Genetic Algorithm for Solving Constrained Optimization Problems*. International Journal of Information Engineering and Electronic Business, vol. 1 (2015) 59-65.

[18]R. BANOS, J. ORTEGA AND C. GIL, *Comparing multicore implementations of evolutionary meta-heuristics for transportation problems*. Annals of Multicore and GPU Programming, vol. 1, no.1 (2014).

[19]R. BANOS, J. ORTEGA AND C. GIL, *Hybrid MPI/OpenMP Parallel Evolutionary Algorithms for Vehicle Routing Problems*. Applications of Evolutionary Computation: 17th European Conference, EvoApplications 2014, Granada, Spain (2014)

[20]P. DELISLE, M. KRAJECKI, M. GRAVEL AND C. GAGNÉ, *Parallel implementation of an ant colony optimization metaheuristic with OpenMP*. Proceedings of the 3rd European Workshop on OpenMP (EWOMP'01), Barcelona, Spain (2001).

[21]Y. TAN AND K. DING, *A Survey on GPU-Based Implementation of Swarm Intelligence Algorithms*, IEEE Transactions on Cybernetics, vol. 46, no. 9 (2016), 2028-2041.

[22]G.-H. LUO, S.-K. HUANG, Y.-S. CHANG, S.-M. YUAN, *A parallel Bees Algorithm implementation on GPU*, Journal of Systems Architecture 60 (2014) 271–279.

[23]A. DELÉVACQ, P. DELISLE, M. GRAVEL AND M. KRAJECKI, *Parallel Ant Colony Optimization on Graphics Processing Units*, Journal of Parallel and Distributed Computing, 73 (2013) 52–61.

[24]L. MUSSI, F. DAOLIO AND S. CAGNONI, *Evaluation of parallel particle swarm optimization algorithms within the CUDA architecture*, Information Sciences, 181 (2011) 4642–4657.

[25]L.P. VERONESE AND R.A. KROHLING, *Differential Evolution Algorithm on the GPU with C-CUDA*, 2010 IEEE Congress on Evolutionary Computation –CEC (2010) 1-7.

[26]Y. ZHOU AND Y. TAN, *GPU-based Parallel Particle Swarm Optimization*, 2009 IEEE Congress on Evolutionary Computation - CEC 2009 (2009) 1493-1500.

[27]R.V. RAO AND G.G. WAGHMARE, *A new optimization algorithm for solving complex constrained design optimization problems*. Engineering optimization, vol. 49, no. 1 (2017) 60–83.

Helical Gold Nanorod and Chiral Gold Nanocage Structures

Xiaojing Liu [1] and Ian Hamilton[2]

[1] *Department of Chemistry, Wilfrid Laurier University, Waterloo, ON, Canada N2L 3C5*

emails: xxliu@wlu.ca, ihamilton@wlu.ca

Abstract

We present intrinsically chiral gold Au_{9n+6} structures consist of an Au_{9n} tube which is capped with Au_3 units at each end. The intrinsic chirality of these structures results from the helicity of the gold strands which form the tube and not because an individual Au atom is a chiral center. The symmetry of these structures is C_3 and substructures of gold hexagons with a gold atom in the middle are particularly prominent.

Key words: gold nanostructures, relativistic density functional theory

Introduction

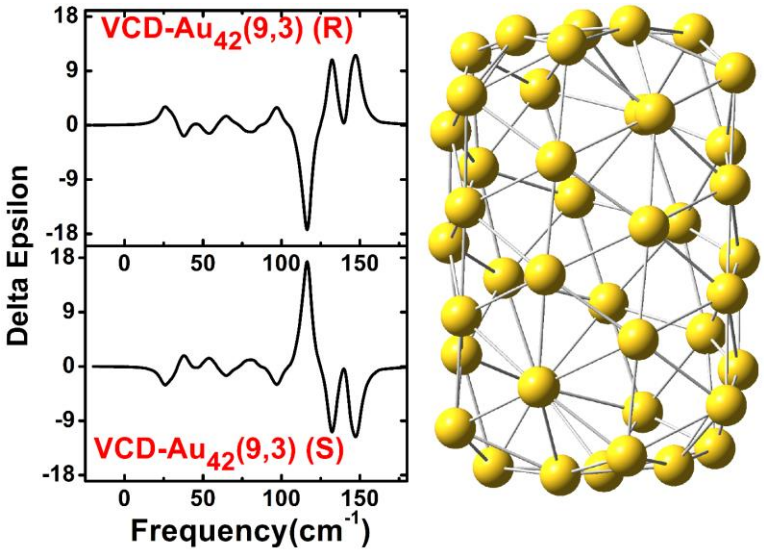
Gold nanoparticles have received a great deal of attention due to their size-related properties which can differ significantly from those of bulk gold. Thus, whereas bulk gold is inert, it was shown that gold nanoparticles are effective catalysts for the oxidation of carbon monoxide and a number of other chemical reactions.

For gold clusters, compact, tube, and cage structures have all been shown to be stable. Cage structures are of particular interest because it may be possible to encapsulate (and stabilize) species inside the cage.

There are a number of instances in which chirality is induced in a gold cluster by chiral ligands or a chiral arrangement of achiral ligands and this chirality may be retained if the ligands are subsequently removed. However, there are only a few instances in which a bare gold cluster is intrinsically chiral.

Results

We have established the stability of helical gold nanorods with 24, 32, and 40 gold atoms.[1] These structures have a core comprised of a linear strand of 3, 4, and 5 gold atoms and a shell comprised of seven helical strands of 3, 4, and 5 gold atoms. We have studied the IR spectra of these species and also the corresponding vibrational circular dichroism (VCD) spectra. A detailed examination of the VCD spectra with adsorbed CClHDT allowed us to establish helical gold nanorods as chiral recognition nanostructures.[2] Starting from the helical gold nanorods, removing the central linear strand, increasing the number of helical strands, and capping the resulting helical gold nanotube at each end, we obtain a series of chiral gold nanocage structures. These gold nanorod and nanocage structures, which exhibit unique electronic and magnetic properties, have potential use in chiral catalysis and as components in nanostructured devices.



References

[1] X.J. LIU, I.P. HAMILTON, R.P. KRAWCZYK AND P. SCHWERDTFEGER, J. Comp. Chem. **33** (2012) 311-318.

[2] X.J. LIU AND I.P. HAMILTON, J. Am. Chem. Soc. **136** (2014) 17757-17761.

Proceedings of the 17th International Conference on Computational and Mathematical Methods in Science and Engineering, CMMSE 2017
4-8 July, 2017

Hopf and homoclinic bifurcation of a new SEIRS epidemic model

M.P. Markakis[1] and P.S. Douris[1]

[1] *Department of Electrical & Computer Engineering, University of Patras, GR 26504, Greece*

e-mails: markakis@upatras.gr, pandouris@ece.upatras.gr

Abstract

A new SEIRS epidemic model with nonlinear incidence rate and nonpermanent immunity is presented in the present paper. Initially, a stability analysis is performed and we proceed to the determination of the necessary and sufficient conditions for the occurence of Hopf bifurcation of codimension 1 or higher. The fact that the incidence rate per infective individual is given by a nonlinear function, product of rational powers of two state variables leads to a more realistic modeling of the physical problem itself. A stability analysis is performed and the features of Hopf bifurcation are investigated. Both the corresponding critical regions in the parameter space and their stability characteristics are presented. Furthermore, by using algorithms based on a new symbolic form as regards the restriction of an n-dimensional nonlinear parametric system to the center manifold and the normal forms of the corresponding Hopf bifurcation, as well, the associated bifurcation diagram is derived. Also, various emerged limit cycles, as well as a homoclinic point-to-point (P2P) connecting orbit are numerically obtained by appropriate implemented methods.

Keywords: Epidemic models; Hopf bifurcations; Limit cycle, Homoclinic P2P connection

1. Introduction

In this paper we present a new epidemic SEIRS model with nonlinear incidence rate and nonpermanent immunity. We deal with the stability analysis and we proceed to the determination of the necessary and sufficient conditions for the occurrence of Hopf bifurcation of codimension 1 and 2. The problem is governed by the following 4D differential dynamical system:

$$\begin{aligned}
&\dot{S} = rN - dS + \varepsilon R - h(S,I)I \\
&\dot{E} = -(d+\sigma)E + h(S,I)I \\
&\dot{I} = -(\gamma + d + a)I + \sigma E \\
&\dot{R} = \gamma I - (d+\varepsilon)R \\
&and \\
&\dot{N} = rN - dN - aI
\end{aligned} \tag{1.1}$$

where S, E, I, R, N denote the number of susceptible, exposed (incubating), infective, recovered individuals and the total population, respectively, while $r, d, \varepsilon, \sigma, \alpha, \gamma$ stand for the system parameters and $h(S,I)$ is the incidence rate per infective individual. Regarding their physical meaning, r denotes the birth rate, d denotes the physical death rate, ε denotes the rate of loss of immunity, σ denotes the rate of incubation, α is the additional death rate due to the epidemic, and γ is the recovery rate. By setting

$$h(S,I) = \beta S^m I^{s-1} / N^{m+s-1} \tag{1.2}$$

with β, m, s positive constants ($s > 1$), and normalizing with respect to the total population N, which is considered constant, taking into account the latter of (1.1), the above system is reduced to a 3D one, namely

$$\begin{aligned}
&\dot{x} = r + \varepsilon - (r+\varepsilon)x - \varepsilon w - \varepsilon y + axy - \beta x^m y^s \\
&\dot{w} = -(r+\sigma)w + awy + \beta x^m y^s \\
&\dot{y} = \sigma w - (r + \gamma + \alpha)y + ay^2
\end{aligned} \tag{1.3}$$

where $x = S/N,\ w = E/N,\ y = I/N$ and $x + w + y + z = 1$. The analysis is multi-parametric, meaning that the parameter space of the system is structured by three varying parameters. In Section 2 a stability analysis of the system is performed, where the active parameters are determined and various graphical representations are obtained concerning the critical (with respect to Hopf bifurcations) values of the varying parameters, as well as the critical, noncritical and stability regions in the parameter space considered. Via effective algorithms based on a new proper symbolic form regarding the center manifold (see [4]), and carried out using the symbolic computation software Mathematica, the corresponding bifurcation portraits are plotted throughout the regions of the parameter space under consideration. Using a custom multiple shooting method, as well as a modified orthogonal collocation method on finite elements, limit cycles corresponding to the cases resulted from the analysis and a homoclinic P2P connecting orbit are presented together with the associated technical details.

2. Stability analysis – Hopf bifurcation

The final reduced system (1.3) possesses two types of equilibria; a disease-free one, namely

$$\Sigma_0 = (1,0,0) \tag{1.4}$$

and an endemic one of the form $\Sigma_1(x^0, y^0, z^0)$ with $y^0 \neq 0$ obtained after some tedious algebraic manipulations, as

$$x^0 = \frac{1}{\sigma\left(\alpha y^0 - \kappa_0\right)}\left[\alpha^2\left(y^0\right)^3 - \alpha\kappa_4\left(y^0\right)^2 + \left(\kappa_3 + \sigma\varepsilon\right)y^0 - \sigma\kappa_0\right] \tag{1.5}$$

$$\beta\sigma\left(x^0\right)^m\left(y^0\right)^{s-1} = \alpha^2\left(y^0\right)^2 - \alpha\left(\kappa_4 - \varepsilon\right)y^0 + \kappa_3 - \varepsilon\kappa_2 \tag{1.6}$$

$$w^0 = \frac{1}{\sigma}\left[-\alpha\left(y^0\right)^2 + \kappa_2 y^0\right] \tag{1.7}$$

where $\kappa_0 = r + \varepsilon,\ \kappa_1 = r + \varepsilon + \sigma,\ \kappa_2 = r + \gamma + \alpha,\ \kappa_3 = \kappa_1\kappa_2,\ \kappa_4 = \kappa_1 + \kappa_2$. We focus on the endemic equilibrium $\Sigma_1 = \left(x^0, y^0, w^0\right)$ with x^0, y^0, w^0 given in (1.5) - (1.7), since it corresponds to persistence of the disease.

Regarding the local stability of Σ_1, taking into account Eqns. (1.5), (1.6) and (1.7), the Jacobian matrix evaluated at Σ_1 becomes

$$J_0 = \begin{pmatrix} \alpha y^0 - \kappa_0 - \dfrac{m}{\sigma}A\dfrac{y^0}{x^0} & -\varepsilon & -\varepsilon + \alpha x^0 - \dfrac{s}{\sigma}A \\ \dfrac{m}{\sigma}A\dfrac{y^0}{x^0} & \alpha y^0 - \kappa_1 + \varepsilon & \dfrac{\alpha}{\sigma}\left[-\alpha\left(y^0\right)^2 + \kappa_2 y^0\right] + \dfrac{s}{\sigma}A \\ 0 & \sigma & 2\alpha y^0 - \kappa_2 \end{pmatrix} \tag{1.8}$$

where

$$\begin{aligned} &y^0 = y^0\left(\gamma, r, \alpha, \varepsilon, \sigma, \beta, m, s\right), \quad x^0 = x^0\left(y^0; \gamma, r, \alpha, \varepsilon, \sigma\right), \\ &A = A\left(y^0; \gamma, r, \alpha, \varepsilon, \sigma\right) = \alpha^2\left(y^0\right)^2 - \alpha\left(\kappa_4 - \varepsilon\right)y^0 + \kappa_3 - \varepsilon\kappa_2 \end{aligned} \tag{1.9}$$

The associated characteristic equation is

$$\lambda^3 + B_2\lambda^2 + B_1\lambda + B_0 = 0 \tag{1.10}$$

By considering the well-known Routh-Hurwitz necessary and sufficient stability conditions, namely

$$B_0 > 0, \quad B_1 > 0, \quad B_1B_2 - B_0 > 0, \tag{1.11}$$

related to the equilibrium Σ_1, we conclude with the formulas

$$\begin{aligned} &B_0 = \frac{1}{\sigma x^0}\left(P_{00} + P_{01}x^0\right), \quad B_1 = \frac{1}{\sigma x^0}\left(P_{10} + P_{11}x^0\right), \\ &B_3 = B_1B_2 - B_0 = \frac{1}{\sigma^2\left(x^0\right)^2}\left(P_{30} + P_{31}x^0 + P_{32}\left(x^0\right)^2\right), \end{aligned} \tag{1.12}$$

where $P_{ij} = f_{ij}(y^0; \gamma, r, \alpha, \varepsilon, \sigma, m, s),\ i = 0,1,3,\ j = 0,1,2$ are polynomials with respect to y^0. By solving the equation $B_3 = 0$ (resulting from (1.10) for $\lambda = i\omega$) with respect to y^0 (after substitution of the right-hand side of (1.5) for x^0, we evaluate the real

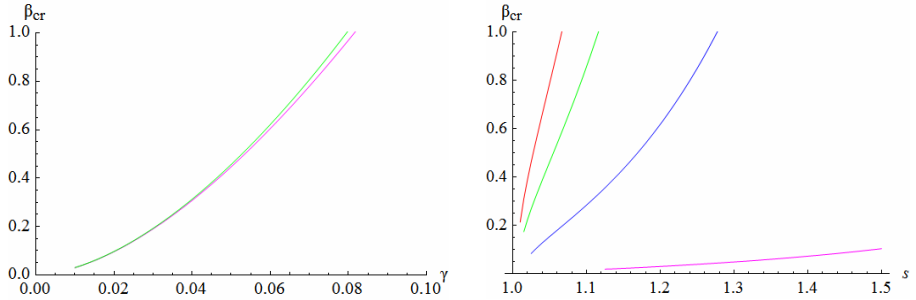
roots of a 9th degree polynomial numerically), we further evaluate x^0 and w^0 by substituting the obtained root of B_3 into (1.5) and (1.7), respectively. Then, taking into account that $\kappa_2 > \alpha$, we arrive at $w^0 > 0$ by means of (1.7) in the case where $0 < y^0 < 1$. Thus if

$$0 < x^0 < 1, \quad 0 < y^0 < 1, \quad w^0 < 1, \quad B_0 > 0, \quad B_1 > 0 \tag{1.13}$$

then (x^0, y^0, w^0) represent the critical values $(x^0_{cr}, y^0_{cr}, w^0_{cr})$. Then, by solving (1.6) with respect to β and considering (γ, σ, β) varying parameters of the problem we obtain the critical value

$$\beta_{cr} = \frac{\alpha^2 \left(y^0_{cr}\right)^2 - \alpha\left(\kappa_4 - \varepsilon\right) y^0_{cr} + \kappa_3 - \varepsilon\kappa_2}{\sigma \left(x^0_{cr}\right)^m \left(y^0_{cr}\right)^{s-1}} \tag{1.14}$$

provided that $0 < \beta_{cr} < 1$, where x^0_{cr}, y^0_{cr} the aforementioned critical equilibrium of the system (1.1). Thus a *critical surface* $\beta_{cr} = \beta_{cr}(\gamma, \sigma)$ is generated in the parameter space (γ, σ, β), (we have $y^0_{cr} = y^0_{cr}(\gamma, \sigma)$ and due to (1.5) we also have $x^0_{cr} = x^0_{cr}(\gamma, \sigma, y^0_{cr}(\gamma, \sigma))$, with $r, \alpha, \varepsilon, m, s$ fixed), defined over the area of the parameter plane (γ, σ), where the critical values of β are obtained via (1.13) and (1.14); we call this area *critical region*. Moreover, by evaluating the appropriate derivatives with respect to the active parameters (γ, σ, β), according to Liu criterion [5], the *transversality condition* holds. Hence, a Hopf bifurcation occurs at the critical equilibrium. Graphical representations of β_{cr} (evaluated by using (1.14)) versus γ (for different values of σ, with $r, \alpha, \varepsilon, m, s$ fixed) and versus s (for different values of γ, with $r, \alpha, \varepsilon, \sigma, m$ fixed) are presented in Figures 1a, 1b and 1c, respectively, while *critical regions* are obtained in the parameter plane (γ, σ) for fixed values of $r, \alpha, \varepsilon, m, s$ in Figures 2a and 2b.

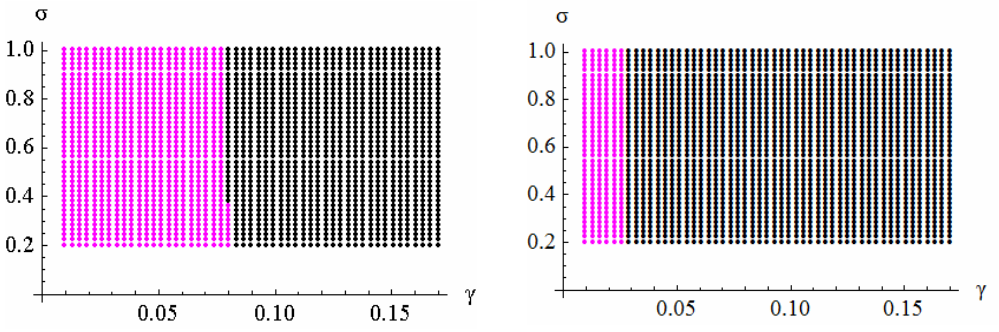


(a) **(b)**

Fig.1 The critical value, β_{cr}, as a function of the parameters: **(a)** γ for $\sigma = 0.25$ (magenta), $\sigma = 0.75$ (green), **(b)** s for $\sigma = 0.5$ and $\gamma = 0.01$ (magenta), $\gamma = 0.06$ (blue), $\gamma = 0.12$ (green), $\gamma = 0.17$ (red).

We note that variation of the values of fixed parameters does not affect the

number of critical values as regards β. Thus in any case, the expression (1.13), under restrictions (1.14), yields zero or at most one critical value ($0<\beta_{cr}<1$). Additionally, we note that an increase in s or m gives rise to an expansion of the *zero region*, namely the area of the parameter plane (γ,σ) where no critical values of β exist. Furthermore, the status of the equilibrium points corresponding to the values of β in the range $(0,1)$, is shown in Figures 3a and 3b. More precisely, after the right-hand side of (1.5) has been substituted for x^0, by using standard numerical computation routines, Eqn. (1.6) is solved with respect to y^0 for any given value of β and fixed values for the other parameters.

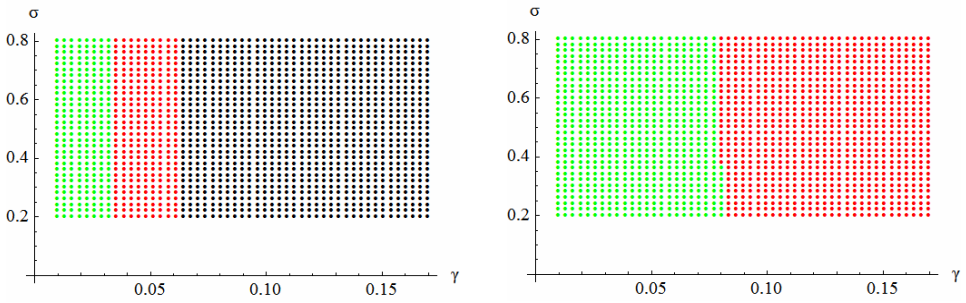


(a) **(b)**

Fig.2 *Critical regions* (magenta – one critical value for β) and *zero regions* (black – no critical value for β), for $r=10^{-4}$, $\alpha=10^{-4}$, $\varepsilon=10^{-3}$, $m=0.5$ and **(a)** $s=1.2$, **(b)** $s=1.6$.

Then x^0 and w^0 can be obtained by means of (1.5) and (1.7), respectively, and finally the coefficients B_i, $i=0,1,2$ of the characteristic equation (1.10) are determined. Thus, concerning the *critical* pairs (γ,σ) (the points of the *critical region*), we focus on the slope of the real part of the complex conjugate eigenvalues of the Jacobian as a function of β, inside an interval $(\beta_{cr}-d_1,\beta_{cr}+d_2)$, $d_1,d_2>0$ (where a pair of complex eigenvalues exist), in order to determine the direction and stability of the occurring codimension 1 Hopf bifurcation. As regards the *zero regions* (no β_{cr}), one investigates the existence of equilibrium points, as well as their stability, as β varies within the range $(0,1)$. As a result, we conclude that regardless of the parameter values, the real part of the complex eigenvalues transverses the β- axis (at β_{cr}) with negative slope, all over the *critical surface* (for β lying into the aforementioned "complex" interval, the sign of the real eigenvalue is always negative). Moreover, we encounter an unstable endemic equilibrium depending on β in the *zero regions*, for this active parameter lying inside an interval $(\delta,1)$, $\delta=\delta(\gamma,r,\alpha,\varepsilon,\sigma,m,s)>0$ (we have no equilibria at $(0,\delta)$), with δ notably sensitive to γ- variation, in the sense that as

γ increases, δ shifts rapidly towards unity, shrinking the (unstable) "equilibrium" β - interval. By increasing s or m, a "0 - equilibrium" region is emerged inside the *zero region* (there exist no equilibria in the whole range $(0,1)$), getting larger as these two rational exponents (especially s) increase.



(a) **(b)**

Fig.3 Stability features regarding the *critical* (green) and the *zero* (black and red) *regions*. Green denotes the negative slope of the real part of the complex eigenvalues versus β, around β_{cr}, red indicates the existence of unstable endemic equilibria for $\beta \in (\delta,1)$ and black outlines the pairs (γ,σ) for which no endemic equilibria arise in the whole range $(0,1)$ for β. The parameters are defined as: $r=10^{-4}$, $\alpha=10^{-4}$, $\varepsilon=10^{-3}$, $m=0.5$ and **(a)** $s=1.2$, **(b)** $s=1.6$.

3. Continuation of limit cycles via multiple shooting - Point-to-point homoclinic orbit

The stable limit cycles emerging from the supercritical Hopf bifurcation of the endemic equilibrium Σ_1 are numerically continued with respect to the fundamental period with one active parameter, the transmission coefficient β, via a custom multiple shooting algorithm with an integral phase condition. For the numerical continuation of the limit cycles of interest the original system is transformed to:

$$\dot{u} = T_p f(u,\alpha) \tag{3.1}$$

by use of the time scaling $\tau = t/T_p$, which maps the independent variable $[0,T_p]$ to $[0,1]$, so that the period appears explicitly as a system parameter. Then, the time variable lies in $[0,1]$ and this interval is subdivided into a total number of $N+1$ mesh points and consider a number of initial conditions, s_i, $i=1,2,..,N$ along an initial approximation of the limit cycle of interest. The problem to be solved numerically is:

$$\dot{u} = T_p f(u,a),\ u(t_i) = s_i,\ u,s \in \mathbb{R}^n,\ \alpha \in \mathbb{R} \tag{3.2}$$

where $t_i \le t \le t_{i+1}$ for $i=1,2,..,N$ within $[0,1]$. Last, a phase condition is needed in order to fix the time-shift. An integral phase condition has been used:

$$\int_0^1 \dot{\hat{u}}(t)\left[u(t)-\hat{u}(t)\right]dt=0 \tag{3.3}$$

where $\hat{u}(\tau)$ represents an initial guess for the solution, typically obtained from the previous continuation step. As soon as the active parameter remains practically unchanged and the step of continuation greatly increases, a good enough initial approximation of the homoclinic orbit has become available. Now, consider the parameterized differential system of interest. Assuming that this system possesses one hyperbolic equilibrium, $z_{\pm}$. we call a solution $z(t)$, for $-\infty<t<+\infty$, of the system at $a=a_0$ a homoclinic connecting orbit (to this equilibrium) if the limit:

$$z_{\pm}=\lim_{t\to\pm\infty} z(t) \tag{3.4}$$

exists. As long as f is continuous $z_{\pm}$ is a stationary point, so:

$$f(z_{\pm},a_0)=0 \tag{3.5}$$

Regarding the numerical computation of a homoclinic orbit, the infinite time horizon $(-\infty,+\infty)$ is truncated to a finite one of the form $[T_-,T_+]$, so that the system is transformed to:

$$\dot{u}=2Tf(u,\alpha) \tag{3.6}$$

accompanied by the integral phase condition (3.3).

In order to determine the number of parameters, p, for which the connecting orbits are isolated and structurally stable phenomena in the system and the orbits of the system appear in a generic manner, we assume that (2.16) possesses the stationary point of interest, $z_{\pm}$. These stationary points are in fact two invariant sets of the system, $M_-,M_+\subset\mathbb{R}^{n+p}$, respectively and the orbit $\gamma=\{(z(t),\mu(t)):\ t\in\mathbb{R}\}$ where $q=(z,\mu)\in\mathbb{R}^{n+p}$ is the connecting orbit pair with:

$$\dot{z}=g(z)=(f(z,\mu),0) \tag{3.7}$$

is called a connecting orbit from M_- to M_+ if:

$$dist(z(t),M_{\pm})\to 0 \text{ as } t\to\pm\infty \tag{3.8}$$

If $M_{\pm}(\mu)$ have stable manifolds $M_{\pm s}(x_-)$ and unstable manifolds $M_{\pm u}(x_+)$ and the dimensions of these manifolds are $n_{\pm c}+n_{\pm s}$, $n_{\pm c}+n_{\pm u}$, respectively, so that $n=n_{\pm c}+n_{\pm s}+n_{\pm u}$. For the steady case, however, $n_{-c}=n_{+c}=0$, so $n_{\pm u}=n-n_{\pm s}$, since $M_-(\mu)$ is a hyperbolic equilibrium $x_-(\mu)$ of (3.7) and $M_+(\mu)$ is also a hyperbolic equilibrium $x_+(\mu)$ of (3.7). We expect the orbit γ to be an isolated connecting orbit if [8]:

$$T_{z(t)}M_{-u}\cap T_{z(t)}M_{+s}=T_{z(t)}\gamma=span\{\dot{z}(t)\} \text{ for all } t\in\mathbb{R} \tag{3.9}$$

for the tangent spaces at $z(t)$. Furthermore, the connecting orbit is persistent in $p-$parametric systems as long as the intersection is transversal, that is if

$$T_{z(t)}M_{-u}+T_{z(t)}M_{+s}=\mathbb{R}^{n+p} \tag{3.10}$$

holds. So, based on (3.9), (3.10) we get:

$$p = m_{+u} - m_{-u} + 1 \tag{3.11}$$

Since the orbit sought is a homoclinic point-to-point one, then $m_{+u} = m_{-u}$, so according to (3.11) $p = 1$, so β has been chosen as the active parameter. The numerically continued cycles are presented in the following Figure:

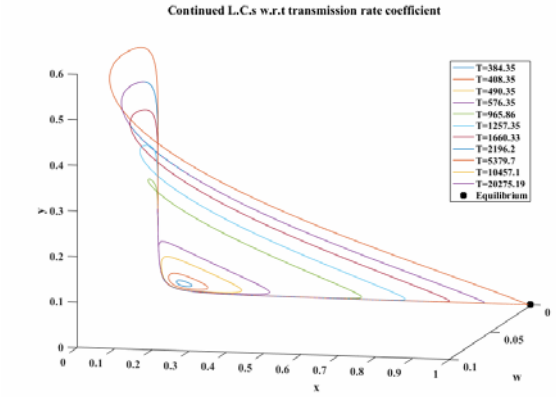


Fig. 4.1 Numerically continued limit cycles emerging from a supercritical Hopf bifurcation

4. High order accurate boundary conditions

An appropriate combination of multiple scales approximation method and that of successive approximations leads to a technique for the determination of high order accurate boundary conditions for the numerical location of homoclinic P2P orbits. Both the widely used techniques of projection boundary conditions and the method of eigenvectors are first order accurate thus requiring quite good initial data for the successful computation of the orbits of interest and this is becoming harder and harder to achieve as the number of state variables together with the number of active parameters increases. A quick description of the technique follows. Consider the dynamical differential system (3.7) and let this system possess a number of fixed points, from which $E_1(z_1^0, z_2^0, .., z_n^0)$ is the fixed point of interest (i.e. the one associated with the homoclinic P2P orbit). Then E_1 is translated to $O(0,0,..,0)$ via the transformation:

$$\xi = z - z^0,\ z \in \mathbb{R}^n \tag{3.12}$$

and z^0 the coordinates of E_1, that is $z^0 = \left(z_1^0, z_2^0, .., z_n^0\right)^T$, so that the initial differential system becomes:

$$\dot{\xi} = f\left(\xi, a\right),\ \xi \in \mathbb{R}^n,\ a \in \mathbb{R}^1 \tag{3.13}$$

and assuming the solution of interest can be approximated up to the order of interest in positive integer powers of a small amplitude orbital parameter, ε, as:

$$\xi(t) = \sum_{i=1}^{k} \varepsilon^i \varphi_i(t) \tag{3.14}$$

where k is the desired order of approximation (Deprit and Henrard, 1965) we insert $\xi(t)$ of (3.13) in the translated system due to (3.12), that is in (3.14) and we equate the terms of the same order, thus getting k first order differential systems, one for each order of approximation. Then, further assuming the fixed point of interest is hyperbolic (i.e. no eigenvalue associated with it is trivial), we set the initial conditions of the first order approximation corresponding to the eigenspace of interest equal to zero. Thus, the first order systems of the desired order of approximation (scale order) are extracted. Let us describe the definition and application of high order B.C.'s. Assuming the solutions of the dynamical differential system of interest can be approximated by:

$$x_{i,j}(t) \approx \sum_{j=1}^{j_{\max}} \varepsilon^j x_{i,j}(t) \tag{3.15}$$

where x_i, $i=1,2,3$ are the state variables (i.e. $x_1(t)=x(t)$, $x_2(t)=w(t)$, $x_3(t)=y(t)$), ε denotes the orbital parameter and $j_{\max}$ is the maximal scale order (i.e. power of ε) or equivalently the highest order of approximation. Then, by substituting the expressions of (3.15) into (1.1) and equating the terms of the same order, we get the respective linear systems of ordinary differential equations (ODEs for short). Thus, the state variables axes of (1.1) are translated by setting:

$$\Delta x_i = x_i - x_i^*,\ i=1,2,3 \tag{3.16}$$

where x_i^*, $i=1,2,3$ are the values of the equilibrium of interest, so that the new equilibrium point of the translated system now becomes $O(0,0,0)$ and the right-hand sides of (1.1) are generally expanded in a Taylor series and terms up to the fourth order are kept, so that 4th order accurate boundary conditions are extracted. Thus, the linearized systems of each order of approximation for the determination of both the outgoing and the incoming solution vector can be obtained. By successively solving the linearized systems of ordinary differential equations and choosing the appropriate initial conditions for each solution, the boundary conditions up to the desired order become available by use of (3.15). Regarding the approximation of the outgoing (incoming) solution vector, the solution of the corresponding system is obtained for every order of approximation and the integration constants corresponding to the eigenvalues [or the eigenvalues with negative (positive) real part more generally] are set equal to zero. Then, the total approximation up to the order of interest is of the form:

$$\Delta x_i(t) \approx \varepsilon \Delta x_{i,1}(t) + \varepsilon^2 \Delta x_{i,2}(t) + \varepsilon^3 \Delta x_{i,3}(t) + \varepsilon^4 \Delta x_{i,4}(t) \tag{3.17}$$

The solutions of the linearized systems of successive (order) approximations for the outgoing solution vector (i.e. associated with the unstable manifold of the equilibrium, to which the orbit under consideration is homoclinic) are presented:

k order approximation for outgoing (unstable) solution vector:

$$\Delta x_{i,k}(t) = q_{xi,k} c_1^{\,k} e^{k\lambda_u t} \tag{3.18}$$

for $i=1,2,3$, $k=1,2,3,4$, where c_1 denotes the integration constant corresponding to the locally asymptotically unstable solution mode (i.e. the positive eigenvalue

of the Jacobian matrix linearized around the equilibrium of interest) in (3.18) and q_{xi}, q_{wi}, q_{yi}, for $i = 1,2,3$ are parameters associated with the numerical values of the chosen system parameters. The systems are solved sequentially and all the rest integration constants (i.e. except for c_1) are set equal to zero. The effectiveness of the higher order approximations is presented in Figure 4.2

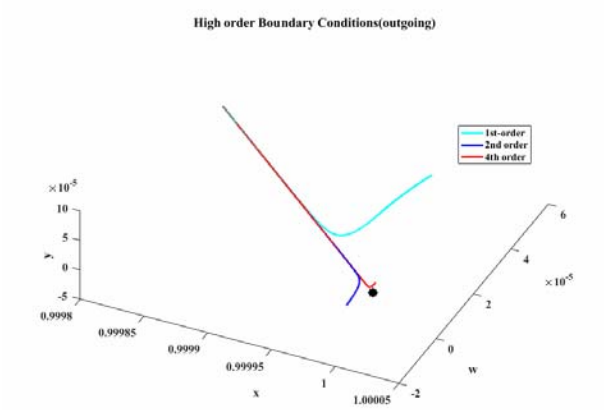


Fig. 4.2 Effectiveness of high order boundary conditions

Thus, the high order B.C.'s for the computation of the homoclinic orbit are (See (3.6)):

$$g(s_1, s_N) = c_{w,j1}\left(x_{j0} - \Delta x_{j_{out}}\right)^2 + c_{w,j2}\left[x_j(\tau = 1) - \Delta x_{j_{in}}\right]^2 \tag{3.19}$$

for $j = 1,2,3$, where $s_1 = [x(\tau = 0; s_1), w(\tau = 0; s_1), y(\tau = 0; s_1)]^T = [x_0, w_0, y_0]^T$ and $c_{w,xjm}$, m=1,2 for are some weighting (scaling) coefficients (all of which have been set equal to 1 in the present case). By employing the aforementioned high order boundary conditions a faster, more efficient computation of the homoclinic orbit of interest is possible.

The transversality of the homoclinic orbit can be verified by use of the Mel'nikov integral. If that integral is not equal to zero, then the homoclinic orbit is transversal, meaning that the stable and the unstable manifold associated with the equilibrium point of the homoclinic orbit intersect transversally. So, consider the extended system:

$$\dot{z} = f(z, a),\ z \in \mathbb{R}^n,\ a \in \mathbb{R}^1,\ \dot{\alpha} = 0 \tag{3.20}$$

and the extended vector of phase variables is $(z, \alpha)^T \in \mathbb{R}^{n+1}$. Also, let (3.20) possess a P2P homoclinic orbit for $a = a_c$ denoted H_c around the saddle equilibrium $z_0(a_c)$, where $z_0(a)$ represents a one-parameter family (curve) of saddle equilibria. This curve has an unstable manifold W^u and a stable manifold W^s. For a specific value of the system parameter a, the corresponding slices of these manifolds coincide with the unstable and stable (respectively) manifolds of

the saddle $z_0(a)$. Consider the linearization of (3.20) around the homoclinic orbit $z_h(t)$ at $\alpha = \alpha_c$:

$$\dot{\gamma} = f_z\left(z_h(t), a_c\right)\gamma + f_a\left(z_h(t), a_c\right)\mu,\ \dot{\mu} = 0 \tag{3.21}$$

Then $\dot{z}_h(t)$ (and every scalar multiple of it) is the unique solution to the variational equation around. The adjoint variational equation around H_c is:

$$\dot{\theta} = -A^T(t)\theta,\ \theta \in \mathbb{R}^n \tag{3.22}$$

where $A(t) = f_z(z_h(t), \alpha_c)$ and it has a unique bounded solution denoted by $\theta(t) = \delta(t)$. Thus, the nondegeneracy (or regularity) of the homoclinic orbit is equivalent to the transversality of the intersection of W^u and W^s along in the extended $n+1$-dimensional phase space of (3.20) and it translates to the condition

$$M_a(a_c) = \int_{-\infty}^{+\infty} \left\langle \delta(t), f_a\left(z_h(t), a_c\right)\right\rangle dt \neq 0 \tag{3.23}$$

which is the Mel'nikov integral. If a scalar function, $M(a)$, measuring the displacement of the invariant manifolds is defined, then:

$$M(a) = M_a(a_c)(a - a_c) + O\left(a^2\right) \tag{3.24}$$

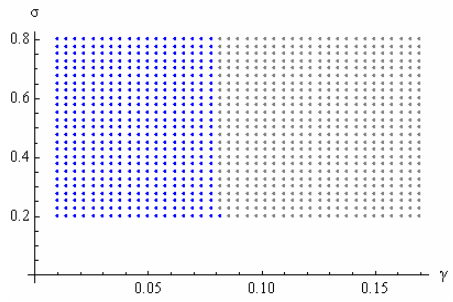
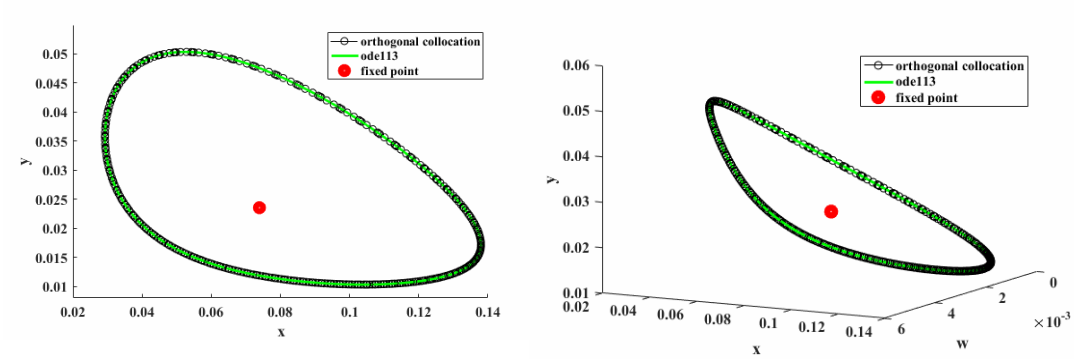


Fig.4.3 Bifurcation portrait of the supercritical Hopf bifurcation for $r = 10^{-4},\ \alpha = 10^{-4},\ \varepsilon = 10^{-3},\ m = 0.5,\ s = 1.2$. Positive sign (blue region) of the first Lyapunov coefficient at the *critical surface* $\beta_{cr} = \beta_{cr}(\gamma, \sigma)$, in the parameter plane (γ, σ) (no β_{cr} in the gray region).

5. Bifurcation Results-Discussion

After setting $r = 10^{-4},\ \alpha = 10^{-4},\ \varepsilon = 10^{-3},\ m = 0.5,\ s = 1.2$ regarding the values of the fixed parameters, by following the procedure developed in [4] (briefly discussed in Section 3 of the present paper), we arrive at the bifurcation portrait of the system with respect to the Σ_1 equilibrium path, presented in Fig.4.3. Regarding the sign of the Lyapunov coefficient, it remains strictly negative on the whole *critical surface* $\beta_{cr} = \beta_{cr}(\gamma, \sigma)$. Thus stable limit cycles are generated through supercritical Hopf bifurcations, arising for (γ, σ) taking values in the *critical region* and $\beta < \beta_{cr}(\gamma, \sigma)$ (the real part of the complex conjugate eigenvalues

is a decreasing function of β around β_{cr} - see Section 2). Since the critical Lyapunov coefficient l_1 never becomes zero, the system undergoes solely a ***codimension 1*** Hopf bifurcation, for any critical triplet of the active parameters.

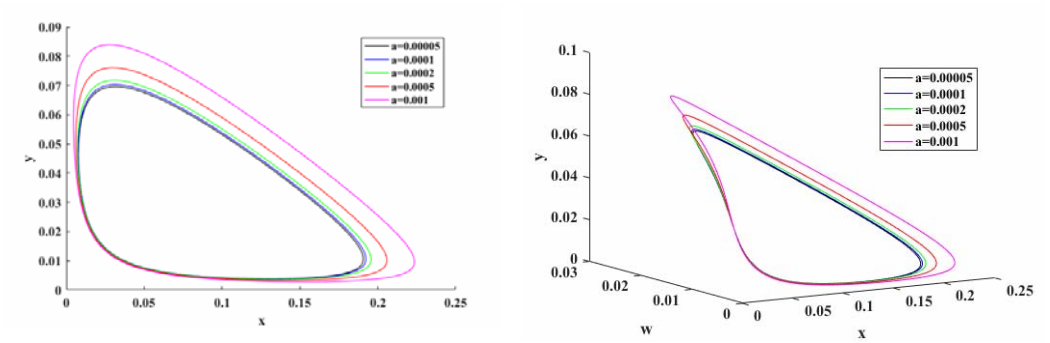


(a) **(b)**

Fig.5 (a,b) Stable cycle generated by supercritical Hopf bifurcation in xy (SI), plane and in $3D$ (SEI), respectively, for:
$(r,\alpha,\varepsilon,m,s,\gamma,\sigma,\beta)=(10^{-4},10^{-4},10^{-3},0.5,1.2,0.042,0.5,0.328525)$ ($\beta_{cr}=0.338525$), determined by a custom orthogonal collocation on finite elements algorithm. Unstable endemic equilibrium (red marker): $(x^0,w^0,y^0)=(0.073948,0.001986,0.023535)$.
Period: $T=410.572533$ days.

The fact that the limit cycles bifurcated are stable means that the phenomenon is persistent, that is the flows in a nearby neighbourhood of the limit cycle are attracted by the cycle itself, leading to the corresponding disequilibrium fluctuations defined by the periodic trajectory, as expected in a considerable number of epidemics. The bifurcation results are verified by the computation and presentation of one cycle for specific values of the parameters by use of a custom algorithm of orthogonal collocation on finite elements, shown in Figures 5 (a, b) and a family of limit cycles obtained for different values of the epidemic-induced parameter α, shown in Figures 6 (a, b), where the corresponding β_{cr} and the period T of the periodic orbits increase with α. The stability of the obtained cycles is additionally verified by numerical computation of the respective Floquet-multipliers and exponents. For the first limit cycle presented in Figures 5, the Floquet-multipliers, $\mu_i=e^{\lambda_i\cdot T}$, i=1,2,3, with $\lambda_i=(1/T)\cdot\ln(\mu_i)$ the respective exponents and T the fundamental period of the cycle, computed are: $\mu_i=\{1,0.82,1.1\times10^{-17}\}$ and $\lambda_i=\{0,-4.7\times10^{-4},-0.095\}$, respectively. Also, we note that the same cycles are generated by the variable-step, variable-order Adams-Bashforth-Moulton predictor-corrector method of orders 1 to 12, which is the standard integrated Matlab routine used to solve non-stiff ODEs, noted as "ode113" in the graphs illustrated in Figures 5 and 6.

Regarding the significance of the introduction of the additional epidemic-induced death rate, a, and the corresponding new nonlinear temrs involved in all equations of (1.3), as shown in Figures 6, it has an important effect on the numeric value of the fundamental period of the bifurcating cycles, which would have been overseen, otherwise. The introduction of "a" could also contribute in estimating the additional resources needed, based on the changes in the duration of critical phases and stages during an epidemic cycle and the maximum number of infected individuals (i.e. estimating the additional cost and health resources needed), thus making it possible to manage and control the epidemics more effectively, efficiently and with a better allocation of available resources.



(a) **(b)**

Fig.6 (a,b) Stable cycles generated by supercritical Hopf bifurcation in xy (SI), plane and in $3D$ (SEI), respectively, for:
$(r,\varepsilon,m,s,\gamma,\sigma,\beta)=(10^{-4},10^{-3},0.5,1.2,0.0548,0.24,0.46825)$, $\alpha\in[5\cdot10^{-5},10^{-3}]$, located by a custom orthogonal collocation on finite elements algorithm.
$\min\beta_{cr}=0.51825$ ($\alpha=5\cdot10^{-5}$). Periods: $T_{\min}=410.414990$ days ($\alpha=5\cdot10^{-5}$), $T_{\max}=441.550487$ days ($\alpha=10^{-3}$).

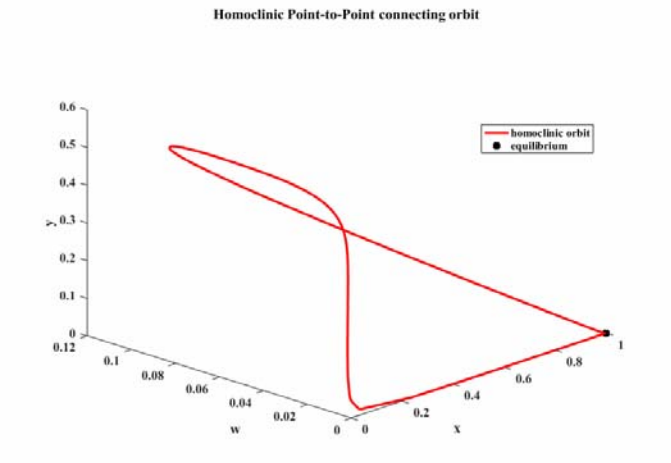


Fig. 7 Point-to-Point homoclinic orbit

Moreover, a homoclinic orbit presented in Figure 7 has been located starting from the continuation of limit cycles emerging from a Hopf bifurcation. This connecting orbit is homoclinic to a saddle endemic equilibrium located close to the disease-free equilibrium Σ_0. The significance of such an orbit is that it acts as a separatrix, as it splits the phase space in regions of qualitatively different motions; the region of oscillatory motions and the region of non-oscillatory ones. We further note that variation of the fixed parameters changes the shape and the size of the *critical* and *zero* regions, without affecting the qualitative profile of the results. In particular, increase in the rational exponents involved in the incidence rate, especially in that concerning the infective, results in a shrinkage of the *critical region*, while the *zero region*, as well as the non-equilibrium area inside that region (as regards the endemic one), expands towards lower values of the recovery rate. Last, the algorithm of orthogonal collocation on finite elements with Legendre orthogonal polynomials is proved excellent in the fast and precise numerical computation of the bifurcated limit cycles.

5. References

[1] H. M. YANG AND A. S. BARREIROS SILVEIRA, The Loss Of Immunity In Directly Transmitted Infections Modeling: Effects On The Epidemiological Parameters, *Bulletin Of Mathematical Biology*, 60 (1998) 355-372.

[2] W.M. LIU, S.A. LEVIN, Y. IWASA, Influence Of Nonlinear Incidence Rates Upon The Behavior Of SIRS Epidemiological Models, *Journal Of Mathematical Biology*, 23 (1986) 187–204.

[3] WEI-MIN LIU, HERBERT W. HETHCOTE, AND SIMON A. LEVIN, Dynamical Behavior Of Epidemiological Models With Nonlinear Incidence Rates, *Mathematical And Computer Modelling,* 51 (2010) 810-822.

[4] MICHAIL P. MARKAKIS AND PANAGIOTIS S. DOURIS, On The Computation Of Degenerate Hopf Bifurcations For N-Dimensional Multiparameter Vector Fields, *International Journal Of Mathematics And Mathematical Sciences*, Vol. 2016, Article ID 7658364, (2016) 12 Pages, Doi:10.1155/2016/7658364.

[5] WEI-MIN LIU, Criterion Of Hopf Bifurcations Without Using Eigenvalues, *Journal Of Mathematical Analysis And Applications*, 182 (1994) 250-256.

[6] Y.A. KUZNETSOV, *Elements Of Applied Bifurcation Theory*, 3rd Ed., Springer, New York, 2004.

[7] H.W. HETHCOTE, Qualitative Analyses Of Communicable Disease Models, *Mathematical Biosciences,* 28 (1976) 335-356.

[8] W.-J. BEYN, On well-posed problems for connecting orbits in dynamical systems, *Contemporary Mathematics* Volume **172** (1994).

High-Performance Paradigm for Digital Transform Processing

H. Mora[1], M.T. Signes-Pont[1], A. Jimeno-Morenilla[1] and J.L. Sánchez-Romero[1]

[1] Department of Computer Science Technology and Computation, University of Alicante

emails: hmora@ua.es, teresa@dtic.ua.es, jimeno@dtic.ua.es, sanchez@dtic.ua.es

Abstract

The digital transforms are intensive in multiplication and accumulation operations which have a high computational cost. Advances in computer arithmetic and digital technologies allow simplifying the processing of complex algorithms when they are implemented in modern circuits. New computation techniques can be explored to provide efficient operational methods for implementing algorithms that avoid much of the complex and costly mathematical operations. This work aims to design a high-performance paradigm for computing some common digital transforms. The proposed architecture has been implemented in a reconfigurable platform to evaluate their performance when compared to other methods. The transform used as example in this work is the Discrete Cosine Transform. The results show that the proposal offers high performance results comparable or better than best-known methods.

Key words: Digital Transform Implementation, Computational Techniques Design, Computer Arithmetic
MSC2000: 65Yxx

1. Introduction

The new capacities in circuit manufacture provided by VLSI and ULSI high-density integration methods, open up new possibilities in the design of digital transform operators. Regular circuit modelling allows compact structures and provide efficiency advantages over other methods which require complex control stages. Therefore, in this work, better calculation techniques are proposed to leverage the new technology advances. The main idea consists in designing a

direct calculation of the arithmetic expressions by means stored logic to facilitate and simplify its implementation in signal processors.

This approach could lay the basis of designing a processor core for a set of transforms. Among the wide range of signal transforms, the digital image transforms are widely used in many consumer applications. These operations have an image matrix as input to produce a result in time and/or frequency for analysis. The separable transforms adopt a simpler formulation that reduces complexity and allows the calculation of this matrix separately for rows and columns. For example, belong to this group are the known Fast Fourier Transform (FFT), Discrete Cosine Transform (DCT), Discrete Sine Transform (DST) or Discrete Hartley Transform (DHT). Their general expression is shown in (1).

$$T(u) = \sum_{x=0}^{N-1} f(x)g(x,u) \tag{1}$$

where, g(x, u) is the kernel of the transform and u ∈ 0..N-1.

2. Problem Definition

The efficient calculation of the transforms is made using the separability property. That is, the 2-D transform is organized in two consecutive 1-D transforms. This section describes the mathematical expressions from which the calculation method is proposed and defines de computation problem.

In this work, the DCT is used as example to describe how the method works due to several reasons: (a) this transform has a simple mathematical formulation but needs significant amount of computing resources; (b) it is intensely studied and a huge volume of proposals exists to compare with them; and (c) it is being part of the most common functions required by new multimedia applications [1,2].

$$F(u) = \frac{C(u)}{2} \sum_{x=0}^{7} f(x) cos\left[\frac{(2x+1)u\pi}{16}\right] \tag{2}$$

where the kernel of the transform and C(u) are:

$$g(x,u) = cos\left[\frac{(2x+1)u\pi}{16}\right] \quad C(u) = \begin{cases} \frac{1}{\sqrt{2}}, if\ u = 0 \\ 1, otherwise \end{cases}$$

Other mentioned transforms have similar computing needs.

3. Proposed Computing Architecture

The methodology uses the basic mathematical formulation of the transform described by (2). According to this expression, the calculation of each F(u) contains several multiplications and additions. The set of constants are depending of the kernel g(x, u) of the transform.

The key design features of the proposed architecture are two: (a) how multiplication and addition operations required for each F(u) are composed, and (b) how the constants C_{uj} are including in the processing of each multiplication. With regard to the first issue, (a), the standard method of calculating each multiplication operation consists of the well-known following stages [3]: partial product generation, partial product reduction and final addition. The proposed idea consists in computing the eight multiplications and seven additions of each F(u) in a combined way, so that, the partial product reductions are integrated with the additions required in each calculation. This approach allows to accelerate a variety of arithmetic operators where additions and multiplications are present. Thus, it should improve overall computation of the transform, where former operations are very frequent, instead of performing complicated designs for reducing the complexity. Regarding the constants involved (b), the proposed design avoids reiterated calculations by the direct generation of only non-zero digits of each constant. To improve the generation of partial products, we use signed-digit format to represent those constants and so minimize the number of nonzero digits while keeping the error within the *p* bits length. The next figure depicts the overall scheme of the transform calculation and the detailed generation of each F(u).

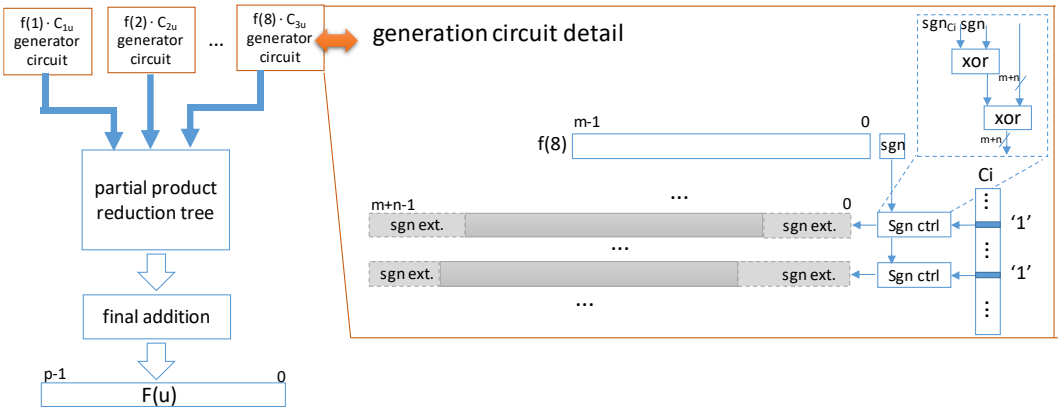


Fig. 1. Overall block diagram with direct partial product generation detail for n=8

This proposal doesn't have to compute symmetries neither decomposition operations for transform calculation to reduce the processing delay because all addition and multiplication operations involved in each F(u) are considered in an integrated way.

4. Evaluation the Performance of the Architecture

The proposal has been evaluated in term of area, delay and power consumption. A normalized balance in equality terms by using homogeneous costs of the components of the architecture is conducted to compare with other designs or proposals [4, 5]. The calculated delays correspond to the critical path of the designs, and they are set by the delay of the slowest path through the circuits.

The results of the estimations and simulations shown demonstrate that the proposal described in this research significantly reduces the execution time of other designs. Additionally, our proposal, allows a more regular structure when calculating all components of the transform in a uniform way. It therefore allows implementation of segmented or pipelined processing strategies that could reduce combined computation.

5. Conclusions

The architecture provides a compact structure for performing operations that does not require ROMs to perform multiplications. The study focus on Discrete Cosine Transform as example of the architecture for being very useful in design of embedded devices for multimedia applications. This approach doesn't need to insert precompute stages to reduce the operations involved. It has proven that a simple approach can provide good performance results.

In comparison to the existing designs, the approach offers some advantages that can be explored for high-speed calculators. We have shown that VLSI implementation through FPGA synthesis can easily meet challenging requirements for high performance applications.

6. References

[1] Q. SUN, L. WANG, Y. SHAO, J. ZUO, Watermarking technique based on three-coefficient comparison in DCT domain, *The Journal of Supercomputing*, **72** (2016), 2594–2608.

[2] J. MORA-PASCUAL, H. MORA-MORA, A. FUSTER-GUILLÓ, J. AZORÍN LÓPEZ, Adjustable compression method for still JPEG images, *Signal Processing: Image Communication*, **32** (2015) 16–32.

[3] H MORA-MORA, J MORA-PASCUAL, JL SÁNCHEZ-ROMERO, JM GARCÍA-CHAMIZO, Partial product reduction by using look-up tables for M × N multiplier, *Integration, the VLSI journal*, **41** (2008) 557-571.

[4] S. S. MISHRA, A. K. AGRAWAL, R.K. NAGARIA, A comparative performance analysis of various CMOS design techniques for XOR and XNOR circuits, *Int. Journal on Emerging Technologies*, **1** (2010).

[5] R. RAJSUMAN, Design and test of large embedded memories: an overview, *IEEE Design and Test of Computers*, **18** (2001) 16-27.

A proposal for computing congestion from trajectories

Francisco Javier Moreno Arboleda[1], Simón Zea Gallego[1] and Jaime Guzmán Luna[1]

[1] *Department of Computer Science, Universidad Nacional de Colombia*

emails: fjmoreno@unal.edu.co, szeag@unal.edu.co, jaguzman@unal.edu.co

Abstract

Vehicle congestion is a problem that overwhelms the big cities. A slow segment is the part of a trajectory (i.e., a subtrajectory) where a vehicle (e.g., a taxi) is considered to have a slow movement according to an established speed limit. In this work, we propose an algorithm that identifies slow segments from the historical data of trajectories of moving objects (e.g., vehicles) and based on a formal definition of congestion that we also propose in this work.

.

Key words: congestion, trajectories, slow segments

1. Introduction

Vehicle congestion [1] is a problem that overwhelms the big cities. The negative impacts of congestion include more time for commuting (to work) and for the delivery of goods and services, more fuel consumption, more air pollution, more mortality (diseases associated with pollution), and loss of money for individuals and companies.

For example, in 2015 in Belgium the average time spent by people in traffic jams was 44 hours, and according to Martin Powell [2], a Siemens mobility expert, 1% of Europe's GDP is wasted in traffic jams, i.e., around 300 million euros.

The reduction of congestion is part of the strategic vision of the development of a city, which articulates mobility, competitiveness, human and urban growth with the sustainability of the city and its quality of life. The reduction of congestion requires multidisciplinary efforts and appropriate policies and measures.

A first step in solving the problem of congestion is to determine when a region is congested. A second step is to identify *where* (regions) and *when recurrent congestion* occurs. For example, a segment of a road where there is always congestion on business days between 7 am and 8 am. In this work, we propose an algorithm that identifies these segments (for space limitations it is not presented here but in the full paper) from the historical data of trajectories of moving objects (e.g., vehicles) and based on a formal definition of congestion that we propose in this work.

2. A Simple example

Suppose we have the annual history of the daily trajectories [3] followed by vehicles in a city. Each trajectory is made up of a set of coordinate points (latitude, longitude, and time). For simplicity, the region of analysis will be rectangular.

From each trajectory we will extract its *slow segments*. A slow segment is the part of a trajectory (i.e., a subtrajectory) where the vehicle is considered to have a slow movement according to an established speed limit (defined by the analysts). For instance, a segment of a trajectory between 7 am and 7:08 am during which the vehicle maintained a speed below 10 km/h. The slow segment must also be *maximum in duration*, so in the previous example, the speed of the vehicle must be greater than or equal to 10 km/h *immediately before* 7am or *immediately after* 7:08 am.

On the other hand, in order to analyze the congestion, the region of interest is segmented into *squares* (*a square tesselation*) with side length l (the analyst defines this value). For the identification of periods of congestion, time is also segmented. For temporal segmentation, the day is divided into n time segments each of size dt (the analyst defines this value), e.g., dt = 15 minutes. The representation of the region of interest squared in a specific time segment is called a *scenario*. Thus, for a squared region, there are n scenarios in one day.
In each scenario, the slow segments are identified. Note that a slow segment can be spread across multiple scenarios. For example, a slow segment of a trajectory between 7:12am and 7:43am is included in two scenarios (one from (7:00am to 7:15am] and the other from 7:15 am to 7:30am]). In Figure 1 we show a scenario and the slow segments in one of its squares (the square is zoomed).

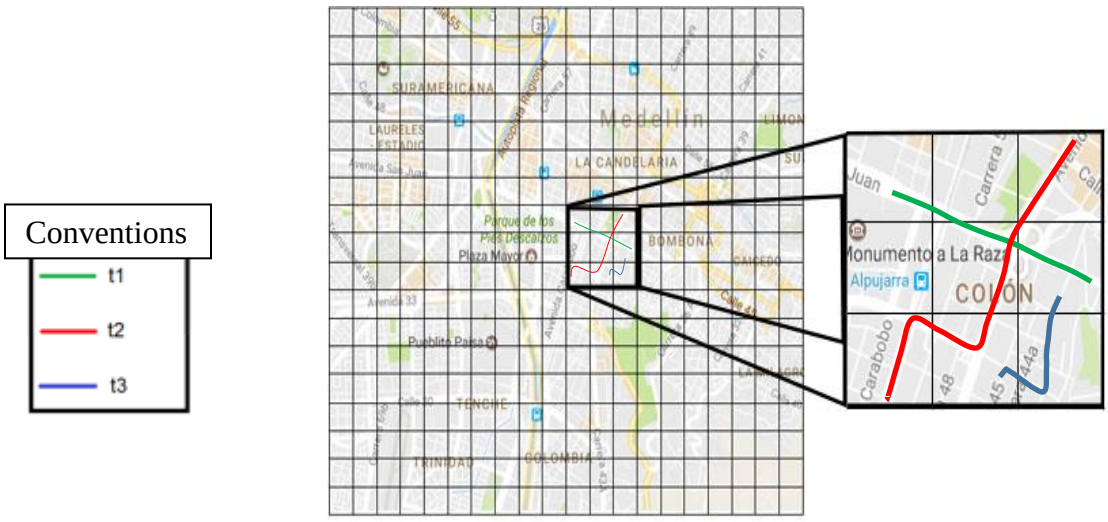


Figure 1. Scenario and square with slow segments of three trajectories.

After having identified the slow segments in a scenario, we can calculate the congestion in a square. This measure is a number between 0 and 1 and corresponds to the ratio between the total distance of slow segments in the square over the total distance of the trajectory in the square, where 0 means no congestion and 1 is the maximum congestion.

For example, consider the square of a scenario shown in Figure 2. We show in brackets the slow segments (with their corresponding distance) of two trajectories that pass through this square. There, trajectory 1 had a slow segment whose distance was 5 and trajectory 2 had a slow segment whose distance was 4. Therefore, the congestion of this square is (5 + 4) / ((8 + 5 + 10) + (10 + 4 + 8)) = 0.2.

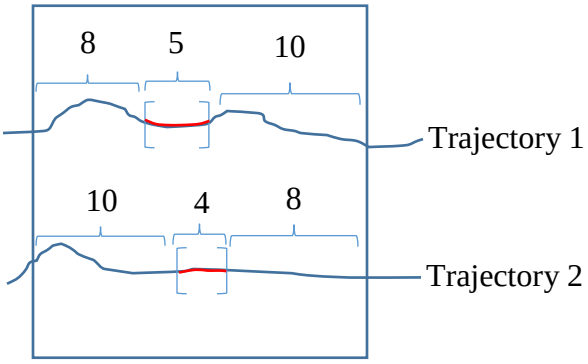


Figure 2. Square with segments of two trajectories.

After obtaining the congestion of each square, we analyze in each scenario how congestion is spread across the squares. For example, we can analyze congestion between two adjacent squares, i.e., those having one side or one point (for the case of diagonal squares) in common, during all temporal segments of the same day or every morning of the same weekday (e.g., every Monday morning or every Saturday morning) for a year.

Definition: Congestion in a square of a scenario: Let a square $R[v, w]$ of a scenario E associated with a region R of order p X q (i.e., p X q squares), $0 < v \leq p$, $0 < w \leq q$, $SL = \{sl_1, sl_2, \ldots, sl_p\}$ the set of all the slow segments of all trajectories that passed through $R[v, w]$ and $SM = \{sm_1, sm_2, \ldots, sm_k\}$ the set of all segments (whether it be slow or not) of all trajectories that passed through $R[v, w]$.The congestion of $R[v, w]$ is:
$(\sum \text{lenght}(sl_i), i = 1,\ldots,p) / (\sum \text{lenght}(sm_i), i = 1, \ldots, k)$.
Where the length() function calculates the total distance of a segment (whether it be slow or not).

3. References

[1] M. D. AFTABUZZAMAN, Measuring Traffic Congestion A Critical Review. 30th Australasian Transport Research Forum, 2007.

[2] M. POWELL, *Helping Cities Get Smarter*, https://www.siemens.com/customer-magazine/en/home/cities/helping-cities-get-smarter.html

[3] S. SPACCAPIETRA, C. PARENT, M.L. DAMIANI, J.A MACEDO, F. PORTO, C. VANGENOT, *A Conceptual View on Trajectories*. Data and Knoledge Engineering 65(1):126-146.

Holographic tools for cell division contents learning

Orcos, L.[1,3], Arís, N.[1] and Magreñán, Á. A.[2]

[1]*Facultad de Educación. Universidad Internacional de La Rioja*
[2] *Escuela superior de Ingeniería y Tecnología. Universidad Internacional de la Rioja*
[3]*Facultad de Educación. Universidad Nacional de Educación a distancia*

emails: lara.orcos@unir.net, lorcos1@alumno.uned.es, nuria.aris@unir.net, alberto.magrenan@unir.net.

Abstract

The purpose of this work is to develop a methodology based on the use of holographic tools to learn Biology and Geology contents related with cellular division. The work is thought to be carried out in the 4th course of high school education with two groups of students, one them will work in a traditional way and the other with the use of this methodology. To test is there is a significant difference between the results obtained by the two groups after having answered a questionnaire related to cell division concepts, a mean comparison test is going to be carried out. Moreover, it also intended to assess the level of motivation of student when using the methodology to evaluate its use and improvement in the future.

Keywords: Meaningful learning, Cellular division, Holographic tool, High School level

1. Introduction

The importance of science in society requires that teaching methods used in classrooms encourage meaningful learning based on competency achievement. The STEM (Science, Technology, Engineering and Mathematics) pedagogical methodology in the classroom is crucial as it aims to approach interdisciplinary learning based on project resolution.

In science learning, a high number of researches have been carried out to learn about difficulties in students [3, 4] as they negatively affect the acquisition of meaningful learning. In the case of cellular biology, the concepts of photosynthesis, respiration, genetics and cell division are the ones that present the most difficulties.

A study based on a methodology to favour the conceptual change of students, concludes that it is necessary to use three-dimensional models and referents as close as possible to the student to enhance their meaningful learning [6]. In this context, the potential of the hologram in training programs is undeniable especially in the sciences. The hologram was invented in 1947 by Dennis Gabor, who won the Nobel Prize for Physics in 1971. There have been advances in the holographic projection in the visual improvement of the technique "Ghost Pepper" that was applied in the mid-nineteenth century.

The prisms are also based on the same "Ghost Pepper" technique, with the difference that only a reflective surface is placed on top of a monitor or screen so that the image is seen on the surface. A study establishes a pedagogical foundation of how the hologram is a teaching medium considering that it is supported by the principles of general pedagogy [10]. The student has the feeling that this object is present and feels more predisposed towards learning. There are three factors that allow us to argue this fact: a) The possibility of such observation facilitates mental representation and the formation of concepts, laws, etc.; b) It allows to obtain the representations based on the relations between the form and the content; c) It strengthens the students relationships as they originate conceptual relations of an individual character and also those of the whole in the class group. Likewise, their motivational potential is projected in the possibility of generating learning contexts among equals that encourage the creation of a shared work environment [9].

Balogh *et al.* made a project in which they used different optical modules that sent light to a holographic display to show a hologram without the need for additional use of lenses [2]. Agócs *et al.* collaborator of the previous one, used diverse optical modules besides mirrors to obtain certain interactivity [1]. Jones *et al.* developed a device composed of a field light visualizer that allows human eyesight of binocular type to be able to see an image formed in 360 degrees, this is possible thanks to a high-speed projector, which transmits images to a mirror with holographic diffuser and electronic circuitry to decode digital video signals, this shows as a result a projection of the object that can be observed without the need to wear special lenses and also avoid the restriction of seeing yourself only from a reference point [7]. A study carried out with 400 teachers on the effectiveness of holograms in education. The results showed that teachers consider this technique potentially effective in achieving meaningful learning [5].

It is worth noting that the research works that refer to the use of holograms for educational purposes, derived from the collections of Serra *et al.* [10], are based on analogic or transmission holograms which are in static plates and are not in motion. It deals with interactive holographic applications through posterior projection or mobile prisms whose objective is to create interactive contents for people of both commercial and institutional applications. Another study comments about how the proper implementation of holograms in classrooms makes students see themselves submerged in a striking environment that makes them to concentrate and build their own learning from their own previous experiences [8].

2. Objectives

The general objective of the present work is to design a holographic technological tool that contributes to the meaningful learning of the students of 4th year of High School Education and to evaluate the degree of satisfaction of the same ones after the use of the tool.

3. Methodology

This study is thought to be carried out with two groups, control and experimental, in the 4th year of High School Education in the subject Biology and Geology. Firstly it will be required that students answer a pre-test to know their previous concepts and them a post-test to assess if there are significant differences between the results obtained after the application or not of the tool. Finally, a questionnaire related to the use of the tool to assess the degree of satisfaction was answered. For the work with the experimental group, after having answered the pre-test, the students have to build the holographic pyramidal prisms with polyethylene music and plastic CD cases and to make videos, both for mitosis and meiosis, taken from Youtube® and edited with the Camtasia® Video Editor. For the analysis of post-test results, a comparison of means in independent groups will be carried out so as to test if there are significant differences between them.

4. Conclusions

A methodology has been stablish for the understanding process of cell division concepts using a hologram due to the fat that this tool if thought to be better for science learning than the typical videos and animations is because it increases the degree of motivation of the students when working with it. The next step is the apply it on the classroom, firstly as a pilot study and then with a greater sample, so as to verify whether the assumptions have been achieved or not. Moreover, it is

also though to make an improvement of this tool to work with interactive holograms through the phenomena "Ghost Pepper" and "Rear Projection" in which semitransparent sheets are used to work the contents through an interface of interaction using a motion sensor as Kinect.

5. References

[1] T. AGOCS, T. BALOGH, T. FORGACS, F. BETTIO, E. GOBBETTI, G. ZANETTI, AND E. BOUVIER. *A large scale interactive holographic display. In VR '06:* Proceedings of the IEEE Virtual Reality Conference, IEEE Computer Society, Washington, DC, USA, 57. 2006.

[2] T. BALOGH, Z. DOBRANYI, T. FORGACS, A. MOLNAR, A. L. SZLOBOD, E. GOBBETTI, F. MARTON, F. BETTIO, G. PINTORE, G. ZANETTI, E. BOUVIER, AND R. KLEIN. *An interactive multi-user holographic environment. In SIGGRAPH '06:* ACM SIGGRAPH 2006 Emerging technologies, ACM Press, New York, NY, USA, 18.2006.

[3] A. CIMER. *What makes biology learning difficult and effective; student´s view*. Educational Research and Reviews, 7 (3), 61-71. 2012.

[4] M. GARCÍA, Y. SEGOVIA, M.J. GÓMEZ, J.M. SEMPERE, P. PEINADO, A. ROMERO. *Dificultades en el aprendizaje de la Biología Celular según la opinión del alumnado. XIII Jornadas de Redes de Investigación en Docencia Universitaria [Recurso electrónico]: Nuevas estrategias organizativas y metodológicas en la formación universitaria para responder a la necesidad de adaptación y cambio.* Universidad de Alicante, 2015. ISBN 978-84-606-8636-1, pp. 2585-2596. 2015.

[5] H. GHULOUM. *3D Hologram Technology in Learning Environment.* Informing Science & IT Education Conference (pp. 693–704). 2010.

[6] F. J. ÍÑIGUEZ, M. PUIGCERVER,. *Una propuesta didáctica para la enseñanza de la genética en la educación secundaria.* Revista Eureka sobre Enseñanza y Divulgación de las Ciencias 10(3), 307-327. 2013.

[7] A. JONES, I. MACDOWALL, H. YAMADA, M. BOLAS, P. DEBEVEC. *Rendering for an Interactive 360◦ Light Field Display.* ACM Transactions on Graphics (TOG), vol 26. 2007.

[8] H. LEE. *3D Holographic Tecnology and Its Educational Potential.* TeachTrends Vol.57 N. 4. Pags. 34-39. 2013.

[9] J. I. POZO, AND C. MONEREO. *Introducción: la nueva cultura del aprendizaje universitario o por qué cambiar nuestras formas de enseñar y aprender.* Psicología del aprendizaje universitario: la formación en competencias. Madrid: Morata: 9-28. 2009.

[10] R. SERRA, G. VEGA. Á. FERRAT, J.J. LUNAZZI, Y S. F. D. MAGALHÃES. *El holograma y su utilización como un medio de enseñanza de la física en Ingeniería.* Revista Brasileira de Ensino de Física, v. 31, n. 1, 1401. 2009.

Multivariate conditional quantile dependence between energy prices and clean energy stock returns

Juan C. Reboredo[1] and Andrea Ugolini[2]

[1] *Department of Economics, Universidade de Santiago de Compostela, Spain*

[2] *Post Graduate program in management – PPGA, Unifacs, Brazil*

emails: juancarlos.reboredo@usc.es, andrea.ugolini85@gmail.com

Abstract

We assessed the impact of quantile energy price movements on the quantiles of clean energy stock price returns using a multivariate vine-copula dependence setup. For the period 2009-2016, our evidence shows that oil and electricity prices were major contributors to the dynamics of clean energy stock returns in the USA and the EU, respectively, whereas coal prices played a minor role in shaping clean energy stock price returns. Furthermore, we found evidence of a symmetric energy price impact, so extreme upward or downward energy price movements had a similar impact on clean energy stock returns. This evidence has potential implications for risk management decision making by energy investors and for policy maker decision making regarding support for clean energy deployment.

Key words: Energy prices; clean energy stock price returns; copulas

1. Introduction

The dynamics of energy prices is one of the main energy-related risk factors that can impact on the financial performance of clean energy investments projects, rendering the substitution of exhaustible for sustainable energy resources more or less viable on economic grounds (see, e.g., Kumar et al., 2012; Reboredo, 2015). Therefore, identifying how the dynamics of different energy prices impacts on the value of renewable energy companies is of particular interest to investors who need to assess the sensitivity of their renewable energy investments to energy prices, in particular, when energy prices are especially low or high. Policy makers are also interested in how fluctuations in energy prices shape renewable energy

stock prices as market forces driving energy prices may provide investors with market-based incentives to invest in the green energy industry, so public investment efforts could be optimally managed.

We modelled the impact of energy prices on new energy stock prices by considering the impact of different energy prices on renewable energy stock prices in a multivariate setup, in which dependence between different energy prices and renewable stock prices was measured, while taking into account direct and indirect price transmission channels. Specifically, we characterized the multivariate dependence structure between oil, gas, coal and electricity energy prices and clean energy stock prices using vine-copula models (Joe, 1996), which characterize high-dimensional joint distributions using a hierarchical structure comprised of a set of bivariate copulas that capture dependence between two variables. This empirical approach offers modelling flexibility, as marginal models and multivariate dependence structures are modelled independently; in particular, it enables the conditional quantile dependence between renewable energy price changes on energy prices — and vice versa — to be assessed, considering both direct and indirect channels of influence. Furthermore, this conditional dependence information can be used to compute the contribution of each energy price change to clean energy stock price movements.

Our empirical analysis characterized the multivariate dependence structure between clean energy stock price indices and oil, gas, coal and electricity prices in the EU and the USA for the period January 2009-September 2016. Our results indicate that the multivariate dependence structure is given by a C-vine hierarchical structure, in which electricity and oil prices play a central role in determining conditional dependence in the EU and USA, respectively. Dependence was mostly time-varying and the analysis of the quantile impact of different energy prices revealed that movements in energy prices played an important role in renewable energy price dynamics, in particular, in extreme downward or upward energy price fluctuations in the EU and the USA. Furthermore, our evidence also indicates that oil prices were the main contributor to new energy stock price movements in the USA, with coal prices playing a minor role. Gas prices played a minor role in shaping renewable energy stock prices in the EU and a more significant role in the USA. Finally, our empirical evidence reveals that the impact of energy prices on renewable energy prices was symmetric, so extreme upward or downward energy price movements had a similar impact on stock prices.

2. Methods

Let o_t, g_t, c_t and e_t be the (log) change in oil, gas, coal and electricity prices, respectively, and let r_t be the (log) change in the renewable energy stock

price. The impact of fluctuating energy prices (for oil, say) of a size given by its β-quantile on the α-quantile of the stock price return distribution for renewable energy prices, given the prices of the other energies, can be measured as:

$$P(r_t \le q^{r_t|o_t}_{\alpha,\beta,t} \mid o_t \le q^{o_t}_{\beta,t}, g_t, c_t, e_t) = \alpha, \quad (1)$$

where $q^{r_t|o_t}_{\alpha,\beta,t}$ is the conditional α-quantile for renewable energy returns at time t and where $q^{o_t}_{\beta,t}$ is the unconditional β-quantile for oil prices, which can be obtained from the inverse of their distribution functions F as:

$$q^{r_t|o_t}_{\alpha,\beta,t} = F^{-1}_{r_t|o_t \le q^{o_t}_{\beta,t},\, g_t, c_t, e_t}(\alpha), \quad (2)$$

$$q^{o_t}_{\beta,t} = F^{-1}_{o_t}(\beta). \quad (3)$$

We can thus measure the impact of oil price fluctuations of different sizes on renewable energy stock prices under different market circumstances, as given by the stock price quantiles. We can also assess the contribution of oil price movements to renewable energy prices at the α-quantile by considering the difference between the conditional and unconditional values:

$$\gamma_{o_t} = q^{r_t|o_t}_{\alpha,\beta,t} - q^{r_t}_{\alpha,t}. \quad (4)$$

where $q^{r_t}_{\alpha,t} = F^{-1}_{r_t}(\alpha)$ is the unconditional α-quantile for the return distribution. Note that when $\gamma_{o_t} = 0$, oil price fluctuations have a negligible impact on stock returns, and when $\gamma_{o_t} < 0 (> 0)$ oil price movements move stock returns in the same (opposite) direction.

Similarly, we can consider the quantile impact arising from other energy price fluctuations, namely, g_t, c_t and e_t, and compute γ_{g_t}, γ_{c_t} and γ_{e_t}. As the sum of γs is not equal to 1, we can normalize the contribution of energy price fluctuations to stock returns as:

$$\hat{\gamma}_{o_t} = \frac{|\gamma_{o_t}|}{|\gamma_{o_t}| + |\gamma_{g_t}| + |\gamma_{c_t}| + |\gamma_{e_t}|}. \quad (5)$$

Eq. (1) can now be expressed in terms of the copula function as:

$$C_{r_t,o_t|g_t,c_t,e_t}\left(F_{r_t|g_t,c_t,e_t}\left(q^{r_t|o_t}_{\alpha,\beta,t}\right), F_{o_{2t}|g_t,c_t,e_t}\left(q^{o_t}_{\beta,t}\right)\right) = \alpha\beta. \quad (6)$$

where $C_{r_t,o_t|g_t,c_t,e_t}(\cdot)$, the conditional bivariate copula between oil and renewable energy returns, can be obtained by partially deriving the copula function in Eq. (7). Next, given the values for α and β and given that $F_{o_{2t}|g_t,c_t,e_t}\left(q^{o_t}_{\beta,t}\right) = \beta$, we can solve from the copula specification in Eq. (9) to obtain $F_{r_t|g_t,c_t,e_t}\left(q^{r_t|o_t}_{\alpha,\beta,t}\right)$. We obtain $q^{r_t|o_t}_{\alpha,\beta,t}$ by inverting the conditional distribution function of r_t, which can be

obtained from the conditional copula. This procedure is also applied to the other energy prices in order to obtain their conditional quantiles.

We characterized multivariate dependence using a vine-copula, which factorizes multivariate copula density in terms of a successive mixing of 5(5-1)/2 bivariate linking copulas with a hierarchical structure (see Joe, 1997; Bedford and Cooke, 2001, 2002; Kurowicka and Cooke, 2006; Aas et al., 2009). Specifically, we considered the C-vine, D-vine and R-vine copulas given by:

$$f\left(x_1,x_2,x_3,x_4,x_5\right)=\prod_{k=1}^{5}f_k\left(x_k\right)\prod_{h=2}^{5}c_{1,h}\left(F_1\left(x_1\right),F_h\left(x_h\right)\right)\prod_{j=2}^{5-1}\prod_{i=1}^{5-j}c_{j,j+i|1,\ldots,j-1}\left(F\left(x_j\mid x_1,\ldots,x_{j-1}\right),F\left(x_{j+i}\mid x_1,\ldots,x_{j-1}\right)\right), \quad (7)$$

$$f\left(x_1,x_2,x_3,x_4,x_5\right)=\prod_{k=1}^{5}f_k\left(x_k\right)\prod_{h=1}^{5-1}c_{h,h+1}\left(F_h\left(x_h\right),F_{h+1}\left(x_{h+1}\right)\right)\prod_{j=2}^{5-1}\prod_{i=1}^{5-j}c_{i,i+j|i+1,\ldots,i+j-1}\left(F\left(x_i\mid x_{i+1},\ldots,x_{i+j-1}\right),F\left(x_{i+j}\mid x_{i+1},\ldots,x_{i+j-1}\right)\right), \quad (8)$$

$$f\left(x_1,x_2,x_3,x_4,x_5\right)=\prod_{k=1}^{5}f_k\left(x_k\right)\prod_{i=1}^{5-1}\prod_{e\in E_i}c_{j(e),k(e)|D(e)}\left(F\left(x_{j(e)}\mid x_{D(e)}\right),F\left(x_{k(e)}\mid x_{D(e)}\right)\right), \quad (9)$$

where $f(\cdot)$ denotes the density function. Marginal models in Eq. (6) are given by an average ARMA (p,q) with TGARCH components.

3. Results

We used daily data for energy prices for the USA and the EU, expressed in USD and EUR, respectively, for the period 2 January 2009-1 September 2016. The data were sourced from Bloomberg as follows: (a) for crude oil, the WTI and Brent benchmark prices for the USA and the EU, respectively; (b) for gas, natural gas futures (NYMEX) for the USA and the UK natural gas futures for the EU; (c) for coal, the Nymex Clearport Central Appalachian Coal Futures for the USA and the ARA (Argus/McCloskey) for Europe; and (d) for electricity, the NYMEX PJM Electricity futures for the USA and the Phelix index for the EU. Finally, for renewable energy prices, we used the ECO Clean Energy Index (ECO) for the USA and the European Renewable Energy Index (ERIX) for the EU.

Figures 1 and 2 depict the estimated multivariate dependence structures for the EU and the USA, respectively. In both cases, the best fit was offered by the C-vine copula structure. Electricity prices played a pivotal role in the EU, whereas

oil prices were central in the USA. Our estimates also show that the Student-t copula was the best fitting bivariate copula in most cases, indicating that symmetric dependence characterized multivariate dependence between energy price movements and clean energy stock returns.

Figure 1. Vine-copula structure for EU.

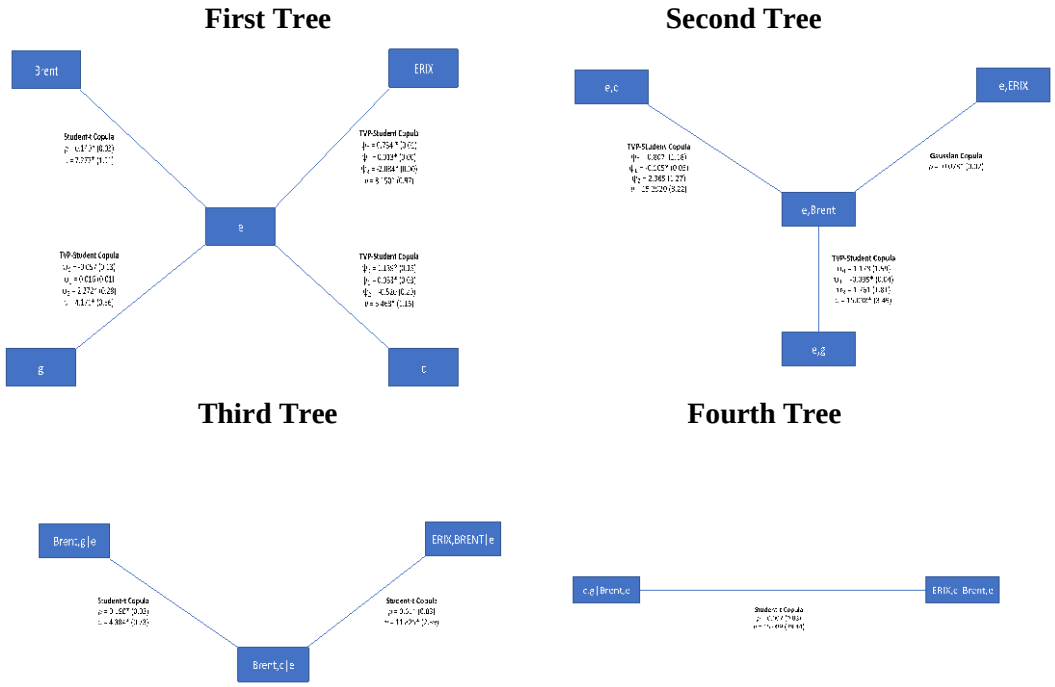
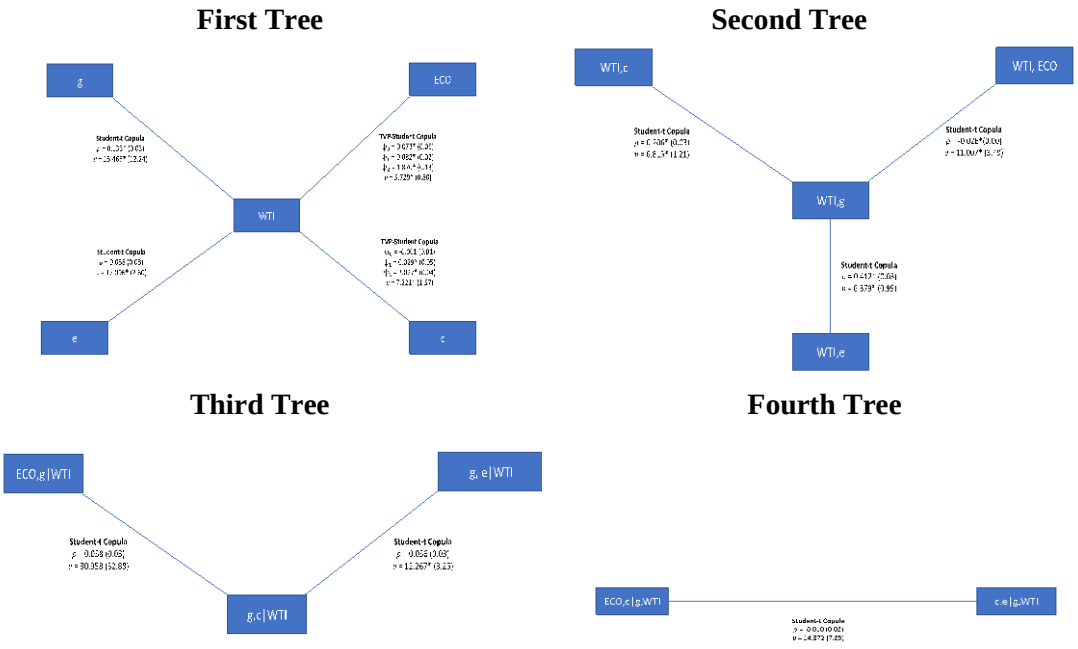


Figure 2. Vine-copula structure for the USA.



Tables 1 and 2 present empirical evidence for the contribution of energy prices to clean energy stock returns at different quantiles. This quantitative evidence confirms that: (a) the contribution of energy prices is symmetric and (b) the main contributor in the EU is electricity, whereas in the USA this role is played by oil.

Table 1. Contribution of energy prices to changes in the ECO index.

Quantile		WTI	Gas	Coal	Electricity
5%	Mean	0.58	0.14	0.13	0.14
	SD.	(0.17)	(0.06)	(0.06)	(0.06)
25%	Mean	0.44	0.18	0.19	0.19
	SD	(0.21)	(0.08)	(0.08)	(0.07)
75%	Mean	0.41	0.20	0.20	0.20
	SD	(0.21)	(0.08)	(0.08)	(0.07)
95%	Mean	0.56	0.15	0.14	0.15
	SD	(0.17)	(0.06)	(0.06)	(0.06)

Table 2. Contribution of energy prices to changes in the ERIX index.

Quantile		Electricity	BRENT	Coal	Gas
5%	Mean	0.59	0.12	0.18	0.11
	SD.	(0.14)	(0.06)	(0.11)	(0.09)
25%	Mean	0.48	0.20	0.17)	0.15
	SD	(0.13)	(0.10)	(0.11)	(0.09)
75%	Mean	0.47	0.22	0.16	0.15
	SD	(0.12)	(0.11)	(0.11)	(0.09)
95%	Mean	0.58	0.13	0.18	0.11
	SD	(0.13)	(0.06)	(0.11)	(0.09)

4. References

[1] REFERENCESAAS, K., CZADO, C., FRIGESSI, A., & BAKKEN, H. (2009). PAIR-COPULA CONSTRUCTIONS OF MULTIPLE DEPENDENCE. INSURANCE: MATHEMATICSAND ECONOMICS, 44, 182–198.

[2] BEDFORD, T., & COOKE, R. M. (2001). PROBABILITY DENSITY DECOMPOSITION FOR CONDITIONALLY DEPENDENT RANDOM VARIABLES MODELEDBY VINES. ANNALS OF MATHEMATICS AND ARTIFICIAL INTELLIGENCE, 32, 245–268.

[3] BEDFORD, T., & COOKE, R. M. (2002). VINES – A NEW GRAPHICAL MODEL FOR DEPENDENT RANDOM VARIABLES. ANNALS OF STATISTICS, 30,1031–1068.

[4] JOE, H. (1996). FAMILIES OF M-VARIATE DISTRIBUTIONS WITH GIVEN MARGINS AND M(M − 1)/2 BI-VARIATE DEPENDENCE PARAMETERS. IN L.RÜSCHENDORF, B. SCHWEIZER, & M. D. TAYLOR (EDS.), DISTRIBUTIONS WITH FIXED MARGINALS AND RELATED TOPICS. HAYWARD: INSTITUTE OFMATHEMATICAL STATISTICS.

[5] JOE, H. (1997). MULTIVARIATE MODELS AND DEPENDENCE CONCEPTS. MONOGRAPHS IN STATISTICS AND PROBABILITY (VOL. 73) LONDON: CHAPMANAND HALL.

[6] KUMAR, S., MANAGI, S., MATSUDA, A., 2012. STOCK PRICES OF CLEAN ENERGY FIRMS, OIL AND CARBON MARKETS: A VECTOR AUTOREGRESSIVE ANALYSIS. ENERGY ECONOMICS 34, 215-226.

[7] KUROWICKA, D., & COOKE, R. M. (2006). UNCERTAINTY ANALYSIS WITH HIGH DIMENSIONAL DEPENDENCE MODELLING. CHICHESTER: JOHN WILEY.

[8] REBOREDO, J.C., 2015. IS THERE DEPENDENCE AND SYSTEMIC RISK BETWEEN OIL AND RENEWABLE ENERGY STOCK PRICES? ENERGY ECONOMICS. 48, 32–45.

Estimating hospital production functions through flexible regression models

Reyes-Santías, F. [1], Cadarso-Suarez, C.[2] and Espasandin, J.[2]

[1] *GEN, Universidad de Vigo,*
[2]*Universidad de Santiago de Compostela (USC)*

emails: francisco.reyes@uvigo.es carmen.cadarso@usc.es
jenifer.espasandin@usc.es

Abstract

1. Background

Two models are commonly used in the estimation of hospital production function: the Cobb-Douglas and the transcendental logarithmic (Translog model). Using these models for estimation and prediction, the functional shape of continuous inputs is "forced" to follow a linear parametric form, which frequently does not fit the data closely. The relative lack of flexibility of parametric models has led to the development of non-parametric regression techniques based on the broad family of generalized additive models (GAMs).

This paper studies the use of Additive Models (AMs) to calculate hospitals production functions. The results of this new approach are compared with the Cobb-Douglas and the Translog models.

2. Data description

The variables we use consist of inputs to hospital production in the form of capital and labour, and outputs from production.

We have chosen the output of Inpatient care measured as number of admissions standardized by means of complexity, obtaining homogeneous units of production (UPHs),calculated by multiplying the number of admissions by the complexity (weight) obtained from the Diagnostic Related Groups (DRGs).

Following Ferrier and Valmanis hospital inputs are measured as follows: in terms of capital we use the average number of beds (Beds) in each hospital.

Labour inputs are measured by the number of consultant full-time equivalents (FTEs) employed in each hospital.

Workload statistics were collected by hospital, as panel data, from the Regional Ministry's Information System, for the period 2010-2016. Hospitals have been classified within three clusters by Reyes following their size: Cluster 1 (small; < 200 beds), Cluster 2 (medium; 200 − 650 beds), and Cluster 3 (large; > 650 bedts).

3. The statistical models

The Cobb-Douglas production function proposed by Charles W. Cobb and Paul H. Douglas takes the following form for our model:

$$\ln UPHs = \beta_0 + \beta_1 \ln(FTEs) + \beta_2 \ln(Beds) + \beta_4 Year + \sum_{h=1}^{nH-1} \alpha_h Hospital_h + \varepsilon. (1)$$

A standard procedure for introducing the possibility of technical change is to include a time trend (Year). This captures observed changes in the technology.

An alternative to the Cobb-Douglas production function is the translog production function. The form of translog production function used is as follows:

$$\ln UPHs = \beta_0 + \beta_1 \ln(FTEs) + \beta_2 \ln(Beds) + \beta_3 \ln(FTEs)\ln(Beds) + \beta_4 Year + \sum_{h=1}^{nH-1} \alpha_h Hospital_h + \varepsilon. (2)$$

The flexible model considered was the following AM including a Beds-by-FTEs interaction:

$$\ln UPHs = \beta_0 + \beta_1 Year + f_1(\ln FTEs) + f_2(\ln Beds) + f_3(\ln FTEs, \ln Beds) + \sum_{h=1}^{nH-1} \alpha_h Hospital_h + \varepsilon, (5)$$

where *f1* and f2 are unknown smooth functions of the number of beds (log scale) and the number of physicians (log scale) respectively, and f3 is a unknown smooth function representing the possible interaction between the number of beds and the number of physicians (both in log scale).

It should be noted that the categorical covariate 'Hospital' was also included in the previous models. In these models, *nH* denotes the number of hospitals (in our study 10), and *Hospitalh* is a dummy variable taking the value 1 for the *hth* hospital and 0 otherwise.

With regard to the estimation of the model, penalized thin plate splines were used to represent the smooth functions *f1*, *f2* and *f3*, and the optimal smoothing parameters were estimated via Restricted (or Residual) Maximum Likelihood (REML).

All the statistical analysis was performed using R software, version 2.9.1. AMs were fitted using *mgcv* package.

4. Results

In this section, we describe the results of each estimated model, for the Regional Health Service hospitals as an overall and every hospital Cluster. We evaluated the models based on the AIC (Akaike Information Criterion) and the economic interpretation for an output change due to changes in input factors.

First of all, the models have been estimated for the Regional Health Service hospitals as an overall. Related to goodness of fit for the models, both the R2 and the AIC indicate that the AM provides a better fit in comparison with the two classic models, CD and Translog.

Following the results for Cluster 1, AM regression model is the only one able to detect a significant interaction between Beds and FTEs inputs ($p = 0.041$). Paying attention to AIC (CD = 1073.084, Translog = 1072.302, AM = 875.337) and R2 values (CD = 63.41, Translog = 63.54, AM = 78.50) we could observe a higher explanation power from the AM rather than for classic ones.

Paying attention to the results for Cluster 2, the effect of interaction between input factors is captured by models translog and flexible one ($p < 0.003$). Even more, the AM model, unlike the classic ones, is able to show that changes in production technology, captured by time trends, would affect the output ($p = 0.022$). The significant hospital effect for the three models ($p < 0.001$, in all cases) seems to reflect some variability related to the size of the hospitals included in Cluster 2. As in the previous results, not only the AIC estimates but also the R2 (CD=77.68, Translog=78.32, AM=85.30) seem to show the advantage of AM behind the classic ones.

The estimation results for Cluster 3 show significant effects of Beds for the three models ($p < 0.001$) while at the same time there are not any significant effects related to FTEs, hospital and technical change (Years) variables for the CD and Translog, but the AM detects effect for the FTEs ($p = 0.037$). However, there is a significant interaction between Capital and Labour factors for the AM ($p = 0.022$) whereas there is not for the Translog ($p =0.772$). The goodness-of-fit of the AM measured by the R2 (CD=78.01, Translog=77.96, AM=86.50) as well as the AIC

(CD=775.441, Translog=777.311, AM=619.218) have been more satisfactory for the AM compared with Cobb-Douglas and Translog.

Graphical output of the results for the AMs, for the overall sample and for each hospital cluster separately, are depicted in Figure 1.

5. Discussion and Conclusions

The decision to measure production of hospitals by the AM made an attempt to improve flexibility for the functional form. The model proposed is certainly a simplified version of the complete econometric model specification (some other variables, in fact, can affect the analyzed phenomenon) but, also at this preliminary stage, the obtained results are really closed to the desirable hypotheses.

A selected set of simple indicators of production has been analyzed. These indicators have been compared across different hospital typologies. This comparative analysis gives important insights to the different variations among hospitals.

As reported in Figure 1, while medium size and small basic-care hospitals are almost homogenous in terms of bed productivity, large size hospitals presents a more complex bed productivity trend. Among hospital typologies, the AM presents a large variability for consultants' productivity. The interpretation of these results is surely an interesting instrument for decision makers in order to analyze the productive conditions of each hospital and the health care sector as an overall. Moreover, AMs may also be applied to check the classical models performance.

Results in this study suggest that AM is a promising technique for the research and application areas on health economics. Moreover, results allow to characterize the domains in which our approach may be effective like those related to demand, costs and utility functions in health care.

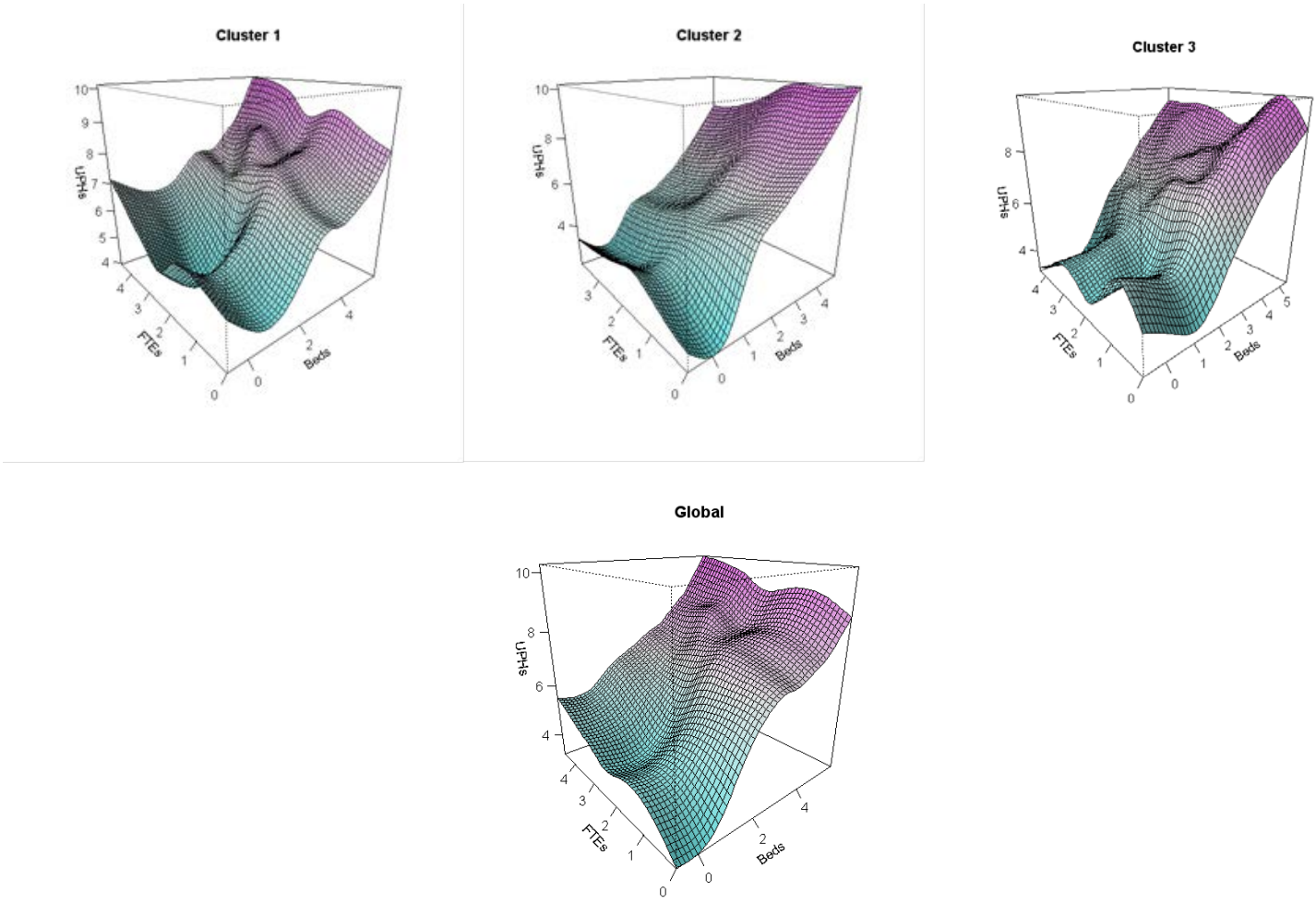


Figure 1
Productivity growth Global, Cluster 1, Cluster 2, Cluster 3, Flexible model. Variables (UPHs, FTEs, Beds) are expressed in Logarithm scale

6. Results

[1] Hastie T J, Tibshirani R J. Generalized Additive Models. London, Chapman and Hall, 1990.
[2] Wood S N. Generalized Additive Models: An Introduction with R. Chapman and
Hall/CRC Press, 2006.
[3] López Rois F J, Mateo R, Gómez J R, Ramón C, Pereiras M. Methodological criteria for drawing up a contract-programme or singular sector-based agreement of specialized care using HPUs. Secretara Xeral SERGAS. Consellera de Sanidade e Servicios Sociais. Xunta de Galicia. Santiago de Compostela, 1999.

[4] Ferrier G, Valmanis V. Do mergers improve hospital productivity? Journal of the
Operational Research Society, 55: 1071–1080, 2004.
[5] Reyes F. Adopción, difusión y utilización de la Alta Tecnologa Médica en Galicia. Tomografía Computerizada y Resonancia Magnética. Universidade de A Coruña, Servizo de Publicacions. A Coruña, 2009.
[6] Cobb C W, Douglas P H. A Theory of Production The American Economic Review, 18 (Supplement): 139–165, 1928.
[7] Christensen L R, Jorgenson D W, Lau L J. Transcendental Logarithmic Production
Frontiers The Review of Economics and Statistics, 55: 28–45, 1973.
[8] Ruppert D,Wand M P, Carroll R J. Semiparametric Regression. Cambridge University Press, 2003.
[9] R Development Core Team. R: A language and environment for statistical computing (2009). R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org
[10] Akaike H. A new look at the statistical model identification IEEE Transactions on Automatic Control, 19 (6): 716–723, 1974.

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4-8 July, 2017

Determining the Best Routes on Dual Gauge Railway Networks using Graphs

Eugenio Roanes-Lozano[1], Alberto Almech[2], Carmen Solano-Macías[3] and Antonio Hernando[4]

[1] *Instituto de Matemática Interdisciplinar (IMI) & Departamento de Álgebra, Universidad Complutense de Madrid*
[2] *Universidad Complutense de Madrid*
[1] *Departamento de Información y Comunicación, Universidad de Extremadura*
[1] *Departamento de Sistemas Informáticos, Universidad Politécnica de Madrid*

emails: eroanes@mat.ucm.es, albermech@gmail.com, csolano@unex.es, ahernando@etsisi.upm.es

Key words: railway networks, routes, graphs, computer algebra systems

1. Extended Abstract

The origin of this work is a conversation with a politician of Extremadura region (Spain) about the railway network related to the region and its possible improvements. We found that there was a lack of simple tools that could help in showing different alternatives in a railway network with two different track gauges.

Two of the authors developed some years ago a complex piece of software (denoted *RutasOptiRed*) for the Spanish Railway Foundation. It evaluated best routes, precise timings, CO_2 emissions and costs for the different pieces of rolling stock of Renfe (Spanish Railways operator) running on Adif (Spanish infrastructure administrator) railway network [1,2,3]. The Spanish case is complex. The network has tracks of two gauges (standard: *1,435 mm*; and Iberian, a broad gauge: *1,668 mm*) and some double-gauge sections (with three rails). There are gauge changeovers installed at some stations, connecting both networks. Moreover, five completely different signalling systems are used in the network (ASFA, ASFA Digital, LZB, EBICAB, ETCS). Finally, some classes of

the rolling stock are dual gauge and one of them is hybrid. The performances (top speed, acceleration) of the pieces of flexible rolling stock depend on how they are working. Therefore, the problem is not trivial at all.

RutasOptiRed required complete details of all sections in the railway network as well as of all the different pieces of rolling stock. Each edge of the graph had a list of characteristics associated (like gauge, maximum speed, etc.). The existence of dual gauge sections (without a gauge changeover at their endpoints), as on the Madrid-Canfranc route, required of the development of a complex adaptation of Dijkstra algorithm [4].

That piece of software was used to study alternatives to the connection Madrid-Badajoz [5]. Soon after, the proposed connection, partially using the Seville high-sped line, was adopted by Renfe (without installing a gauge changeover, but with a guaranteed transfer), and is now still available.

It was also used to study the different alternatives to access Galicia with the opening of some parts of the new NW high-speed line and the possibilities of the new hybrid trains (Class 730) [6]. The present route is the one considered optimal by *RutasOptiRed*.

The key idea in the new approach presented here is to consider a single graph without gauge details associated to its edges. This way a standard Dijkstra algorithm can be used. Let us detail it afterwards.

Let S be the set of stations. Three undirected graphs are considered as input:

- G_1: graph corresponding to the Iberian gauge railway network (the vertices are the stations and the sections of the network are the edges)
- G_2: graph corresponding to the standard gauge railway network
- G_3: graph corresponding to the dual gauge railway network.

If a station can be reached by trains of the two gauges (either through normal tracks or double gauge tracks), it will be described by two vertices. For instance, we shall have a vertex $Zaragoza_i$ and another vertex $Zaragoza_s$ (standing for Iberian / standard, respectively). Let S_i and S_s be the sets of stations with both Iberian and standard gauges. Let $C_i \subseteq S_i$ and $C_s \subseteq S_s$ be the sets of stations where a gauge changeover is installed.

Step 1: Two new undirected graphs are considered, each one corresponding to the accessibility in each gauge:

- $G_1^* = G_1|S_i \cup G_3|S_i$ (corresponding to the sections of the network that can be traversed by an Iberian gauge train)

- $G_2^* = G_2|S_s \cup G_3|S_s$ (corresponding to the sections of the network that can be traversed by a standard gauge train)

where $G_x|S_y$ represents exchanging in graph G_x the name of all vertices, v, such that $v_y \in S_y$, by v_y.

Step 2: The two previous graphs are merged and the connectivity between the two graphs is given by the gauge changeovers positioning (Figure 1):

$$G^* = G_1^* \cup G_2^* \cup \{\{v_i, v_s\} : v_i \in C_i,\ v_s \in C_s\}$$

Observe that the elements of the last set in the formula above are of the form $\{Zaragoza_i , Zaragoza_s\}$.

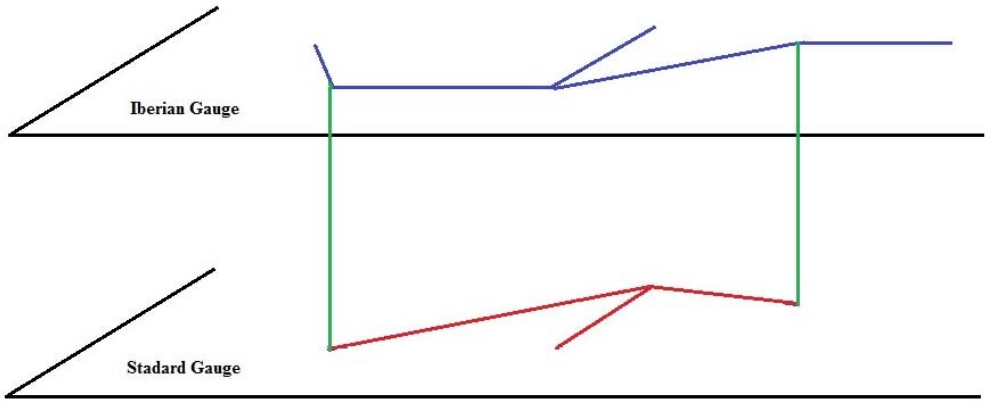


Figure 1: Graph G^* representing the global accessibility.

This way, only the real or theoretical timings have to be assigned to each edge of G^*. Edges correspond to sections of railway line or to gauge changeovers, and endpoint to endpoint timings can be easily computed. Let us note that the purpose of this new piece of software is far simpler than that of *RutasOptiRed*, although the approach for merging the subnetworks of different gauges is much simpler.

The new package is implemented in Maple, using its graphs package, and will be exemplified by analysing the Madrid–Badajoz and Badajoz–Lisbon routes, based on present and historically available lines [7,8].

2. Acknowledgements

This work was partially supported by the research project TIN2015-66471-P (Government of Spain) and grant CASI-CAM S2013/ICE-2845 (Comunidad de Madrid)..

3. References

[1] A. HERNANDO, E. ROANES-LOZANO, A. GARCÍA-ÁLVAREZ, L. MESA, I. GONZÁLEZ-FRANCO, *Optimal Route Finding and Rolling Stock Selection in a Dual Gauge Multi-Voltage Railway Network.* Comput. Sci. Eng. **14/4** (2012) 82-–89. DOI: 10.1109/MCSE.2012.80

[2] E. ROANES-LOZANO, A. HERNANDO, A. GARCÍA-ÁLVAREZ, L. MESA, I. GONZÁLEZ-FRANCO, *Calculating the Exploitation Costs of Trains in the Spanish Railways,* Comput. Sci. Eng. **15/3** (2013) 89-95. DOI: 10.1109/MCSE.2013.54

[3] E. ROANES-LOZANO, A. HERNANDO, A. GARCÍA-ÁLVAREZ, L. MESA, I. GONZÁLEZ-FRANCO, *Possibilities of RutasOptiRed package*. In: A. Ibeas, J. L. Moura, L. dell'Olio, B. Alonso (editors): XI Congreso de Ingeniería del Transporte (CIT 2014), Elsevier Procedia-Social and Behavioral Sciences **160**, Amsterdam (2014) 102-111. DOI: 10.1016/j.sbspro.2014.12.121

[4] A. HERNANDO, E. ROANES-LOZANO, A. GARCÍA-ÁLVAREZ, *A recommender system for train routing: when concatenating two minimum length paths is not the minimum length path,* Appl. Math. Comput. (To appear).

[5] E. ROANES LOZANO, A. HERNANDO, A. GARCÍA ÁLVAREZ, L. MESA, I. GONZÁLEZ FRANCO, *Comparación de posibles rutas y material móvil para el itinerario Madrid-Badajoz con el paquete RutasOptiRed*. In: VII Congreso de Innovación Ferroviaria. Ponencias, UNED, Calatayud, Spain (2012) 607-618.

[6] E. ROANES LOZANO, A. HERNANDO, A. GARCÍA ÁLVAREZ, L. MESA, I. GONZÁLEZ FRANCO, *Comparación por tipo de tren de Renfe Operadora de posibles enrutamientos alternativos en la red de Adif usando el paquete RutasOptiRed*. In: Actas del X Congreso de Ingeniería del Transporte CIT 2012 (CD-ROM), Universidad de Granada, Granada, Spain (2012).

[7] E. ROANES-LOZANO, L.M. LAITA, E. ROANES-MACÍAS, M.J. WESTER, J.L. RUIZ, C. RONCERO, *Evolution of Railway Network Flexibility: the Spanish Broad Gauge Case,* Math. Comput. Simul. **79/8** (2009) 2317-2332. DOI: 10.1016/j.matcom.2008.11.007

[8] E. ROANES-LOZANO, J. GÓMEZ-CASTAÑO, C. SOLANO-MACÍAS, *Evolution of the connectivity of the Portuguese broad gauge railway network (1948-2012),* In: C. Reinhardt, K. Shroeder (editors): Railways: Types, Design and Safety Issues, Nova Science Publishers, Inc., New York (2013) 149-164.

Linguistic Description of Behaviours based on Fuzzy Deformable Prototypes. A Study Case using Time Tracking logs.

Francisco P. Romero[1], José A. Olivas [1], Juan García[1] and Jesús Serrano-Guerrero

[1] *Department of Information Technologies and Systems, University of Castilla La Mancha*

emails: franciscop.romero@uclm.es, Juan.GAguilar @uclm.es, JoseAngel.Olivas@uclm.es, Jesus.Serrano@uclm.es

Abstract

This paper presents an approach for the linguistic description of time series based related to human behaviour based on Fuzzy Deformable Prototypes. The main purpose is to create a complex behaviour model that allows to offer recommendations regarding the current situation and predictions. It is also presented a case of study based on data captured by a time-tracking tool.

Key words: Fuzzy Deformable Prototypes, Time Tracking Logs

1. Introduction

Nowadays, there is a proliferation of tools, devices and applications that allow to capture data of people activity, but, merely offer the possibility of its visualization. The integration of the available information and its subsequent analysis using intelligent algorithms would give the possibility of offering personalized services to the user. The main problem to be addressed when trying to analyse these data sources lies in the inherent inaccuracy in the recording process. With the aim of handling data of these characteristics, Fuzzy Logic [1] offers mechanisms that allow to introduce expert knowledge and formalize the uncertainty that exists on the subject. For this purpose, our proposal is based on the use of fuzzy logic techniques such as Fuzzy Deformable Prototypes [2] that allow not only the description of the phenomena but inferences to be made about the presented situations.

2. Linguistic description of human behaviours using fuzzy deformable prototypes

The use of "ideal descriptions" for human behaviours has several shortcomings. Ideal descriptions prototypically describe a fact, but in fact only approximate the ideal to a degree. Common description sentences show uncertainty or vagueness associated to the fulfilment of this ideal. Then, in the context of description, a fact or a set of facts is associated with a paradigm so that the paradigm interprets the behaviour. Thus, it is possible to describe the behaviour according to this interpretation. To generalize, many of the descriptions depend on the way the most similar paradigm or prototype for the circumstances of the problem is found.

Fuzzy Deformable Prototypes (from now on FDPs) can provide an adequate formal framework for working with this idea. FDPs come from the confluence of two approaches to the concept of prototype: the "deformable prototype" of Bremermann [3] ("*a real element is classified according to the minimum energies required for physically deforming the closest prototype"*) and the fuzzy prototypes of Zadeh [4] ("*a fuzzy prototype is a reunion of good, bad and borderline elements of a category*"). The definition of FDPs inherits some features of Zadeh's fuzzy prototype approach but includes some extensions to manage the complexity of real-world problems. For example, in FDPs there are many prototypes of a behaviour depending on its relevant features. This work uses the concept of fuzzy deformable prototypes (FDPs) in order to model different phenomena related to human behaviour.

The principle to obtain a fuzzy deformable prototype of a population of elements (for example, time segments) is to stratify it by grouping them sharing the same membership degree. For each level of stratification, the fuzzy prototype is obtained using an iterative process of clustering. During the iterative process, an object maximally summarized from each level of clustering is obtained which can be viewed as a fuzzy prototype. Where the number of prototypes for a behaviour is given, it may be meaningful to compute the collective properties of the prototypes and consider them as the reference for the corresponding descriptions and recommendations.

In addition, each fuzzy prototype can also be represented as fuzzy sets. It means that it is possible to calculate a membership degree between an element and the fuzzy set. The use of FDP's allows to evaluate new situations from these patterns, to deform [4] the most similar prototypes to this new behaviour and describe it using a combination of prototypes with the membership degrees as coefficients.
As as result of this process, we can obtain some sentences to describe the behaviour with different degrees of typicality. For example, we can use the prototypical way to describe the situation, or show some uncertainty or vagueness or show higher and lower relationships with the prototypical behaviour.

3. Case of Study: Personal Time Logs

Nowadays, tracking time is a relevant topic because it is directly related to efficiency and productivity. For this purpose, it is important to find a way to realistically and practically analyse the use of time.

Following the theory of prototypes of Zadeh [2], the concept of a time segment with respect to its productivity can include a set of prototypes, which represent the good, low or medium compatibility of the instances with the concept.

The first step of our analytics process to discover these prototypes is to calculate the behavioural features from the data provided by Time Tracking Tool. For this purpose, we analyse all relevant historic records and perform the aggregations that feature definition demands—for example, aggregating the amount of interruptions this user spent during that time segment.

Then, the following step is detecting the relationships between the different time segments and thus to obtain the fuzzy deformable prototypes based definition of them. With this aim, first a clustering process is performed based on the features previously defined. Subsequently, aggregation functions are applied on each type of time segments to calculate the features values that will define each of the prototypes. Finally, each of these prototypes is represented using fuzzy numbers.

From these definitions, we try to simulate the capacity of interpretation of the situation, that is, to find the model of evolution of productivity more adapted to the real circumstances. To this end, the user is first asked the real value of the factors that a priori determine the progression of a time segment. Their combination will give a degree of compatibility with prototypes, those prototypes with a degree of non-zero compatibility will be modified to give rise to the characterization of the time segment by a combination whose coefficients are the degrees of compatibility previously obtained.

Specifically, the system describes the situation based on the level of stress of the user. To do this, it calculates the current stress state of the user by dividing for each prototype the stress curve [5] in five stages: start, up, top, down, end. Identifying within the prototype the phase in which the user is, through a finite state machine, can be generated the linguistic description of the situation and therefore recommendations from it. In addition, the linguistic description is enriched by the characteristics of the use of time (“maker”, “manager”, “mixture”). The description provided to the current time segment would be valid if there are no influence of *local factors*. In this case, it is mandatory to modify any description or recommendation.

4. Conclusions and Future Work

Continuous analysis of time series captured from human activity allows us to describe changes in the people behaviour and offer recommendations. Fuzzy Deformable Prototypes of time segments show the different behavioural possibilities. Subsequently, each time segment type goes through a "deformation process" until a prototype match is achieved. The better we can describe the "time segment" during any window time, the easier it becomes to offer specific recommendations and predictions.

It would be interesting in future work to devise a version of the analysis in which reinforcement learning could play an important role. For example, the user's opinion about the descriptions and recommendations provided by the system can be fed back the application with this knowledge.

5. References

[1] ZADEH, L. A. "FUZZY SETS." INFORMATION AND CONTROL 8.3 (1965): 338-353.

[2] OLIVAS, J. A. CONTRIBUCIÓN AL ESTUDIO EXPERIMENTAL DE LA PREDICCIÓN BASADA EN CATEGORÍAS DEFORMABLES BORROSAS. PHD THESIS. UNIVERSITY OF CASTILLA LA MANCHA, 2000

[3] H. BREMERMANN, 'PATTERN RECOGNITION'. H. BOSSEL: SYSTEMS THEORY IN THE SOCIAL SCIENCES. BIRKHÄUSER VERLAG, 1976, PAGES 116-159.

[4] ZADEH, L. A. "A NOTE ON PROTOTYPE THEORY AND FUZZY SETS." COGNITION 12.3 (1982): 291-297.

[5] TEIGEN, K. H. "YERKES-DODSON: A LAW FOR ALL SEASONS." THEORY & PSYCHOLOGY 4.4 (1994): 525-547.

Proceedings of the 17th International Conference on Computational and Mathematical Methods in Science and Engineering, CMMSE 2017
4-8 July, 2017

On a sufficient condition for commutative orthogonal block structure

Carla Santos[1], Célia Nunes[2], Cristina Dias[3] and João Tiago Mexia[4]

[1] *Department of Mathematics and Physical Sciences, Polytechnical Institute of Beja, and CMA -Center of Mathematics and its Applications, New University of Lisbon, Portugal*
[2] *Department of Mathematics and Center of Mathematics and Applications, University of Beira Interior, Portugal*
[3] *College of Technology and Management, Polytechnical Institute of Portalegre and CMA - Center of Mathematics and its Applications, New University of Lisbon, Portugal*
[4] *Department of Mathematics and CMA- Center of Mathematics and its Applications, Faculty of Science and Technology, New University of Lisbon, Portugal*

emails: carla.santos@ipbeja.pt, celian@ubi.pt, cpsilvadias@gmail.com, jtm@fct.unl.pt

Abstract

A model has orthogonal block structure if it has variance-covariance matrix that is a linear combination of known pairwise orthogonal orthogonal projection matrices that add to the identity matrix. When the orthogonal projection matrix on the space spanned by the mean vector commutes with the orthogonal projection matrices, in the expression of the variance-covariance matrix, the model has commutative orthogonal block structure. Resorting to B-matrices we present a general condition for this commutativity.

Key words: B-matrices, mixed models, models with commutative orthogonal block structure
MSC2000: AMS Codes (optional)

1. Introduction

Linear mixed models are a powerful tool for analysing experimental data from several areas, such as agriculture, biology, medicine or industry. In the framework of the design of experiments in agricultural trials, in 1965, a special class of linear mixed models as emerged, called models with orthogonal block structure, OBS, based on the structure of the variance-covariance matrix [6,7]. Later on, in order to obtain optimal estimation for variance components of blocks and contrasts of treatments, using the algebraic structure of OBS, arose a particular case of these models, those of models with commutative orthogonal block structure, COBS [4].

2. Models with commutative orthogonal block structure

Let us consider a mixed model

$$\boldsymbol{Y} = \sum_{i=0}^{w} \boldsymbol{X}_i \boldsymbol{\beta}_i$$

where $\boldsymbol{\beta}_0$ is fixed and $\boldsymbol{\beta}_1$, ... , $\boldsymbol{\beta}_w$ are independent random vectors with null mean vectors, variance-covariance matrices $\sigma_1^2 \boldsymbol{I}_{c_1}$ ... $\sigma_w^2 \boldsymbol{I}_{c_w}$, where $c_i = rank(\boldsymbol{X_i}), i = 1, \dots, w$ and null cross-covariance matrices.

$\boldsymbol{Y}$ has mean vector

$$\boldsymbol{\mu} = \boldsymbol{X}_0 \boldsymbol{\beta}_0$$

and variance-covariance matrix

$$\boldsymbol{V}(\boldsymbol{\theta}) = \sum_{i=1}^{w} \sigma_i^2 \boldsymbol{M}_i$$

where $\boldsymbol{M}_i = \boldsymbol{X}_i \boldsymbol{X}_i^T, i = 1, \dots, w$.

Since the space spanned by the mean vector is $\Omega = R(\boldsymbol{X}_0)$, the orthogonal projection matrix, OPM, on Ω, is $\boldsymbol{T} = \boldsymbol{X}_0 (\boldsymbol{X}_0^T \boldsymbol{X}_0)^+ \boldsymbol{X}_0^T = \boldsymbol{X}_0 \boldsymbol{X}_0^+$, where $+$ indicates Moore–Penrose inverse.

When the matrices $\boldsymbol{M}_1, \dots, \boldsymbol{M}_w$ commute, they generate a commutative Jordan algebra of symmetric matrices, CJAS,A. This is a linear space constituted by symmetric matrices that commute and containing the squares of their matrices [5]. The CJAS, A, as one unique basis, called the principal basis, Q, that is constituted by known pairwise orthogonal orthogonal projection matrices, POOPM, [8] thus the matrices $\boldsymbol{M}_i, i = 1, \dots, w$ are linear combinations of the matrices of the CJAS principal basis

$$\boldsymbol{M}_i = \sum_{j=1}^{m} b_{i,j} \boldsymbol{Q}_j$$

With $\gamma_j = \sum_{i=1}^{w} b_{i,j}\, \sigma_i^2,\ j = 1, \ldots, m,$ the canonical variance components, the variance-covariance matrix of $\boldsymbol{Y}$ will take the form

$$\boldsymbol{V} = \sum_{j=1}^{m} \gamma_j \ \boldsymbol{Q}_j$$

Since $\sum_{i=1}^{w} \boldsymbol{M}_j \in$ A is invertible, A is a complete CJAS and $\sum_{j=1}^{m} \boldsymbol{Q}_j = \boldsymbol{I}_n$.

So

$$\boldsymbol{Y} = \sum_{i=0}^{w} \boldsymbol{X}_i \boldsymbol{\beta}_i$$

is a model with orthogonal block structure, OBS. These models, introduced in [6,7], have been intensively studied and play a central part in the theory of randomized block designs, see, e.g., [1,2].

An important class of OBS, models with commutative orthogonal block structure, COBS, arises when T, the OPM on the space spanned by the mean vector, commutes with the POOPM $\boldsymbol{Q}_j$, $j = 1, \ldots, m$. [4]. So $\boldsymbol{T}$ and $\boldsymbol{V}$ commute and the least square estimators, LSE, for estimable vectors, will give BLUE (best linear unbiased estimators) whatever the variance components [10].

Assuming the rows of $\boldsymbol{X}_0$ to correspond to the sets of levels of the fixed effects factors, the mean values of the observations will be determined by those sets.

Let us consider that there will be $\dot{n}$ sets of the levels associated to $r_1, \ \ldots, \ r_{\dot{n}}$, contiguous rows of $\boldsymbol{X}_0$. If the components of $\boldsymbol{\beta}_0$, $\beta_{0,1}, \ \ldots, \beta_{0,\dot{n}}$, are the corresponding mean values, we can reorder the observations to have the block diagonal matrix

$$\boldsymbol{X}_0 = D\left(1_{0,1}, \ldots, 1_{0,\dot{n}}\right)$$

So the orthogonal projection matrix on , the space spanned by the mean vector, is given by

$$\boldsymbol{T} = D\left(\frac{1}{r_1} J_{r_1}, \ldots, \frac{1}{r_{\dot{n}}} J_{r_{\dot{n}}}\right)$$

where $J_r = 1_r\ 1_r{}^T$

The fundamental partition of $\boldsymbol{Y}$ will be constituted by the sub-vectors $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_{\dot{n}}$, corresponding to the $\dot{n}$ sets of the levels of the fixed effects factors [3]. Then the variance covariance matrix can be defined by

$$\boldsymbol{V} = \begin{bmatrix} \boldsymbol{V}_{1,1} & \cdots & \boldsymbol{V}_{1,\dot{n}} \\ \vdots & & \vdots \\ \boldsymbol{V}_{\dot{n},1} & \cdots & \boldsymbol{V}_{\dot{n},\dot{n}} \end{bmatrix}$$

with $\boldsymbol{V}_{l,l}$ the variance-covariance matrix of $\boldsymbol{Y}_l$, $l = 1, \ldots, \dot{n}$, and $\boldsymbol{V}_{l,h}$ the cross-covariance matrix of $\boldsymbol{Y}_l$ and $\boldsymbol{Y}_h$, $l \neq h$.

Since

$$TV = \begin{bmatrix} \frac{1}{r_1} J_{r_1} \boldsymbol{V}_{1,1} & \cdots & \frac{1}{r_1} J_{r_1} \boldsymbol{V}_{1,\dot{n}} \\ \vdots & & \vdots \\ \frac{1}{r_{\dot{n}}} J_{r_{\dot{n}}} \boldsymbol{V}_{\dot{n},1} & \cdots & \frac{1}{r_{\dot{n}}} J_{r_{\dot{n}}} \boldsymbol{V}_{\dot{n},\dot{n}} \end{bmatrix} \text{ and } VT = \begin{bmatrix} \boldsymbol{V}_{1,1} \frac{1}{r_1} J_{r_1} & \cdots & \boldsymbol{V}_{1,\dot{n}} \frac{1}{r_1} J_{r_1} \\ \vdots & & \vdots \\ \boldsymbol{V}_{\dot{n},1} \frac{1}{r_{\dot{n}}} J_{r_{\dot{n}}} & \cdots & \boldsymbol{V}_{\dot{n},\dot{n}} \frac{1}{r_{\dot{n}}} J_{r_{\dot{n}}} \end{bmatrix}$$

the matrices T and V commute if and only if

$$\begin{cases} \frac{1}{r_1} J_{r_1} \boldsymbol{V}_{1,1} = \boldsymbol{V}_{1,1} \frac{1}{r_1} J_{r_1} & \cdots & \frac{1}{r_1} J_{r_1} \boldsymbol{V}_{1,\dot{n}} = \boldsymbol{V}_{1,\dot{n}} \frac{1}{r_1} J_{r_1} \\ \vdots & & \vdots \\ \frac{1}{r_{\dot{n}}} J_{r_{\dot{n}}} \boldsymbol{V}_{\dot{n},1} = \boldsymbol{V}_{\dot{n},1} \frac{1}{r_{\dot{n}}} J_{r_{\dot{n}}} & \cdots & \frac{1}{r_{\dot{n}}} J_{r_{\dot{n}}} \boldsymbol{V}_{\dot{n},\dot{n}} = \boldsymbol{V}_{\dot{n},\dot{n}} \frac{1}{r_{\dot{n}}} J_{r_{\dot{n}}} \end{cases}$$

Which occurs when the matrices $\boldsymbol{V}_{l,h}$, $l = 1, \ldots, \dot{n}$, h= $1, \ldots, \dot{n}$ are B-matrices, see, e.g. [9], this is, when

$$\frac{1}{\dot{n}} \sum_{l=1}^{\dot{n}} \boldsymbol{V}_{l,h} = \frac{1}{\dot{n}} \sum_{h=1}^{\dot{n}} \boldsymbol{V}_{l,h} = \frac{1}{\dot{n}\dot{n}} \sum_{l=1}^{\dot{n}} \sum_{h=1}^{\dot{n}} \boldsymbol{V}_{l,h}$$

With

$$V = D\left(\sigma_1{}^2\ I_{r_1}, \ldots, \sigma_{\dot{n}}{}^2\ I_{r_{\dot{n}}}\right)$$

matrices T and V commute.

With this commutativity condition $\boldsymbol{Y}$ will be a model with commutative orthogonal block structure (COBS) and, according to the version of the Gauss-Markov theorem in [10], the LSE for estimable vectors will be BLUE.

3. References

[1] T. CALINSKI, S. KAGEYAMA, *Block Designs: A Randomization Approach, vol. I*: Analysis, Lecture Note in Statistics, 150, Springer-Verlag, New York (2000).

[2] T. CALINSKI, S. KAGEYAMA, *Block Designs: A Randomization Approach, vol. II*: Design, Lecture Note in Statistics, 170, Springer-Verlag, New York (2003).

[3] F. CARVALHO, J.T. MEXIA, R. COVAS, C. FERNANDES, *A fundamental partition in models with commutative orthogonal block structure*. AIP Conf. Proc. 1389 (2011) 1615–1618

[4] M. FONSECA, J.T. MEXIA, R. ZMYSLONY, *Inference in normal models with commutative orthogonal block structure*, Acta Comment. Univ. Tartu. Math. 12 (2008) 3–16.

[5] P. JORDAN, J. VON NEUMANN, E. WIGNER, *On the algebraic generalization of the quantum mechanical formalism*, Ann. Math. 36 (1934) 26–64

[6] J.A. NELDER, *The analysis of randomized experiments with orthogonal block structure I. Block structure and the null analysis of variance*, Proc. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci. 283 (1965) 147–162.

[7] J.A. NELDER, *The analysis of randomized experiments with orthogonal block structure II. Treatment structure and the general analysis of variance*, Proc. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci. 283 (1965) 163–178.

[8] J. SEELY, *Quadratic subspaces and completeness*, Ann. Math. Statist. 42 (1971) 710–721.

[9] C. SANTOS, *Error orthogonal models: structure, operations and inference*, PhD thesis. University of Beira Interior.

[10] R. ZMYŚLONY, *A characterization of best linear unbiased estimators in the general linear model*, Mathematical Statistics and Probability Theory, 2 (1978) 365–373.

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4-8 July, 2017

From Graphene to Graphyne, Fullerenes, Fulleroids, Gaudienes and their Golden Duals

Peter Schwerdtfeger[1]

[1] *Centre for Theoretical Chemistry and Physics, The New Zealand Institute for Advanced Study,*
Massey University (Albany Campus), Auckland, New Zealand

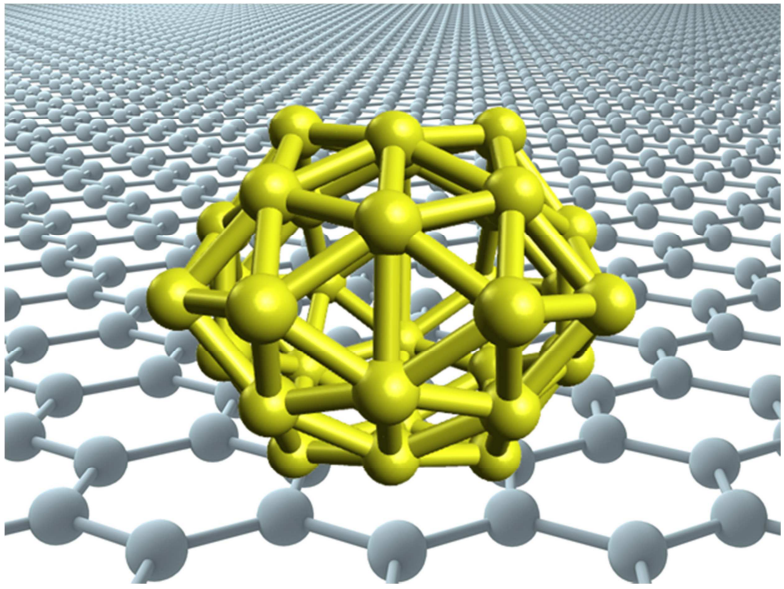
email: peter.schwerdtfeger@gmail.com

Abstract

Graphene is a material with many potential applications. For example, we can introduce nano-holes into graphene membranes for the design of efficient molecular sieves. Our research group has recently succeeded to efficiently separate He-3 from He-4 by quantum tunneling.

Graphene can also be wrapped around a sphere to form fullerene structures. 12 pentagons have to be however introduced to satisfy Euler's polyhedral formula [1,2]. Graphene can also be modified to graphyne by vertex insertions. From this one obtains new fullerene structures called gaudienes. Such structures can be explained using topology and graph theory which form the basis of our general-purpose computer program Fullerene that creates accurate 3D structures for any fullerene isomer. It also creates dual structures corresponding to a triangulation of a sphere. Such a dual structure has recently been found experimentally in Lai-Shang Wang's group at Brown University, and consists of 16 gold atoms on a sphere. These unique triangulations of a sphere related to fullerene duals have exactly 12 vertices of degree five. The icosahedral hollow gold cages previously postulated are related to halma transforms of C_{20}. This dual mapping also relates the (chiral) gold nanowires observed to the (chiral) carbon nanotubes. In fact, the Mackay icosahedra well known in gold cluster chemistry are related to the dual halma transforms of the smallest possible fullerene C_{20}.

Keywords: Fullerenes, Fulleroids, Golden Duals



References

[1] P. SCHWERDTFEGER, WIRZ, LUKAS AND J.E. AVERY, *The Topology of Fullerenes*, Wiley Interdisciplinary Reviews: Computational Molecular Science, **5**, 1, 96-145, Wiley 2015. DOI: 10.1002/wcms.1207

[2] P. SCHWERDTFEGER, WIRZ, LUKAS AND J.E. AVERY, *Program Fullerene: A software package for constructing and analyzing structures of regular fullerenes*, J. Comp. Chem, **34**, 17, 1508-1526, Wiley, 2013. DOI: 10.1002/jcc.23278

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4-8 July, 2017

Feature selection by means of genetic algorithms

Antonio J. Tallón-Ballesteros[1], Beatriz Ruiz-Reina[2], and Luis Rus-Pegalajar[2]

[1] *Department of Languages and Computer Systems, University of Seville, 41012 Spain*
[2] *Higher Technical School of Computer Science Engineering, University of Seville, 41012 Spain*

emails: atallon@lsi.us.es, bearuirei2@alum.us.es, luiruspeg@alum.us.es

Abstract

This paper explores the efficacy of a population-based approach such as genetic algorithms in the field of feature selection, although these algorithms were created as an alternative to optimization methods like EVOP or simplex. A framework called Correlation-based Feature Selection guided by a Genetic Search algorithm (CFS-GeSe) was introduced. The results are promising especially for C4.5 classifier, bearing in mind the accuracy and Roc measures.

Key words: supervised machine learning, feature selection, genetic algorithms, data mining, dimensionality reduction.

1. Introduction

A possible taxonomy of classifiers could distinguish between different kinds of algorithms [1]. The most well-known methods fall into the category of decision trees, rule-based classifiers and black-box methods. Some pros and cons characterise every type of approach. Nowadays the available information is continuously making some progress and every second more data to be processed increases. We speak of attribute [2] or feature as the description of some existing measure in the universe that takes values in a certain domain. On the other hand, we talk about domain as the set of values of the same type. That features are taken from the data mining. The data mining [3] contains a lot of data that are treated to

be converted in knowledge. This knowledge is taken when the features are treated with or without the genetic algorithms.

Therefore, machine learning-techniques and data mining were developed to discover knowledge and recognize patterns from these data. What is more feature selection is almost a mandatory task in most of application domains in order to minimize redundancy and maximize relevance to the target, as the data collected is usually associated with a high level of noise.

2. Feature selection and genetic algorithms

Feature selection could be defined as the way to get the most relevant attributes, deleting those who are redundant or not relevant, for to obtain the most optimal result. When feature selection is applied, we get less data, greater accuracy, simpler results and less attributes; for this reason, the efficiency is increased and the accuracy and the understanding of the results [4].

The only way to guarantee the optimal selection of this final subset is to do an exhaustive search of all the possible subsets. However, this could take a long time as for N features the number of possible subsets is 2^N. The purpose of feature selection is to select a subset of the most relevant features out of the larger set, maximizing the classification performance. In other words, those features which are highly correlated with the class [5].

Genetic algorithms are based on natural evolution, identifying the attribute as chromosome, and the parts of the attribute as genes. This evolution is based on the idea that the chromosomes with "good structures" survive more likely.

There are several ways to obtain these attributes to make the selection (randomly, by tournament, according to their evaluation,...), but the genetic algorithms [6] make it favouring the best while allowing the diversity. This diversity is obtained by giving all the genes a probability of being chosen, so that although the best is favoured, one with less probability can be part of the gene that allows us to obtain the most optimal result. The methods used [7] in genetic algorithms are based on mutations or crossings of genes through which the best attributes are obtained. While permutations are crosses based in cycles, crossings exchange chromosome segments.

The most commonly used technique in feature selection is the stepwise approach [8], which can run forward or backward. In the forward version, the variables are selected and added to the model one at time, until no improvement in the results is obtained. In the backward version, the first model to be computed is the one with all the variables; the variables are then selected and removed one at time. A mix of these two approaches is also used, in which each addition of a variable by the forward selection is followed by a backward elimination. The stepwise techniques have two main disadvantages: first, each choice heavily affects the following choices; and second, the final result is expressed by a single combination, and then no choice is given to the user.

Genetic algorithms, in their favour, always allow the exploration of the whole experimental space: due to the occurrence of the mutations, each possible combination can occur at any moment. The result obtained with genetic algorithms is a whole population of solutions: the user can then choose the one he prefers, taking into account at the same time the response and the variables are used. From this it is also evident that a relevant feature of genetic algorithm is the ability to detect several local maxima.

In short, data mining contains those data that which we can do a feature selection without genetic algorithms, but of course, the use to this type of algorithms give us the better features with a minimal time.

Our proposal considers a genetic search (GS) in conjunction with an attribute evaluator based on correlation known as CFS (Correlation-based Feature Selection). In a first step CFS [9] is applied to get the assessment of the features according to a correlation measure; after that a genetic search algorithm guides the search following a metaheuristic approach. The framework has been called CFS-GeSe that stands for Correlation-based Feature Selection guided by a Genetic Search algorithm.

3. Experimentation

Table 1 summarises the main parameters of the CFS-GeSe framework along with their numerical values. It is very outstanding to stress that a preliminary experimental design was carried and after that we chose the values described due to a good performance; cross validation procedure using the train set to create preliminary train and preliminary test sets. It is straightforward that these values may not be the optimal ones but were the result of a previous experimentation.

Parameter	Value
Population size	200
Number of generations	20
Probability of crossover	0.6
Probability of mutation	0.033

Table 1: Parameter values for CFS guided by a Genetic Search algorithm (CFS-GeSe).

Table 2 represents the data set used throughout the experimentation. They come mostly from binary and multi-class supervised machine learning real-world problems available in the public repository from the University of California at Irvine (UCI). We have also included the number of selected features for the proposal coined as CFS-GeSe with the training set. Last row shows the dimensionality reduction on average over the original data sets.

Data set	#Samples	#Train	#Test	#Features	#Classes	#Selected(CFS-GeSe)
Hepatitis	155	117	38	19	2	10
Ionosphere	351	263	88	33	2	12
Labor	57	43	14	29	2	9
Promoter	106	80	26	58	2	20
Waveform	5000	3750	1250	40	3	12
Average	1133.80	850.60	283.20	35.80	2.20	12.60
Dim. Reduction						64.80

Table 2: Summary of the test-bed and features selected for CFS-GeSe proposal.

Concerning the classifiers, C4.5 and BayesNet were applied to every problem a) with the original data set and b) also after the data preparation over the train set took place and the list of selected attributes were projected onto the initial sets to get the reduced train and test subsets, respectively. As evaluation measures we have included in the forthcoming tables Accuracy and Roc in the test set.

Table 3 reports the test results obtained with C4.5 classification algorithm. By rows, we have represented a concrete result in boldface meaning that the current proposal overcame the situation will all the features. If no value is in bold, no improvements happened and the single advantage of the new proposal for this problem is that a reduction in the number of features took place, arising therefore a trade-off between the computational time and the performance. From the point of view of accuracy, CFS-GeSe gets better results in two out of the five problems tested. Moving to Roc measure, three data sets take advantage of the new proposal. In both cases, the averages are also higher which accompanied to the improvements in individual results indicate that CFS-GeSe is very appropriate for C4.5 algorithm.

Data set	Accuracy		Roc	
	Original	CFS-GeSe	Original	CFS-GeSe
Hepatitis	84.21	84.21	0.7760	**0.8890**
Ionosphere	92.05	92.05	0.8970	0.8920
Labor	85.71	85.71	0.8000	0.8000
Promoter	69.23	**76.92**	0.5860	**0.6570**
Waveform	74.80	**76.00**	0.8050	**0.8500**
Average	81.20	82.98	0.7728	0.8176

Table 3: Test results with C4.5 classifier.

Table 4 shows the accuracy and Roc results via the BayesNet supervised machine learning approach. We have followed the same conventions aforesaid for

the Table 3. Now, sometimes punctual enhancements happen. In other cases worst results are obtained which is enough to be very careful with the application of CFS-GeSe in combination with BayesNet. These experiments encourage to test with different feature selection approaches or even measures or kinds of searches to guide the exploration and exploitation with the findings of the first stage of the feature selection procedure.

Data set	Accuracy		Roc	
	Original	CFS-GeSe	Original	CFS-GeSe
Hepatitis	94.73	94.73	0.9740	0.9740
Ionosphere	92.05	92.05	0.9520	**0.9680**
Labor	85.71	85.71	0.9560	0.8670
Promoter	76.92	**80.76**	0.9020	0.8820
Waveform	78.08	77.84	0.9330	0.9310
Average	85.50	86.22	0.9434	0.9244

Table 4: Test results with BayesNet classifier.

4. Conclusions

This paper introduced CFS-GeSe which is a Correlation based Feature Selection guided by a Genetic Search algorithm. Experiments revealed that CFS-GeSe is very suitable for C4.5 classifier letting to get better test results with Accuracy and Roc measures. On the other hand, for BayesNet is not very clear that the combination with CFS-GeSe is a handy symbiosis and other ways of feature selection should be explored or even to analyse the measures, if any, that could be convenient to this very powerful supervised machine learning algorithm.

5. References

[1] S. FINLAY, *Predictive analytics, data mining and big data: Myths, misconceptions and methods*. Springer, 2014.

[2] R. OSSWALD, *A Logic of Classification*, PhD Dissertation, University of Düsseldorf (Germany), 2002.

[3] C. WESTPHAL, *Data Mining for Intelligence, Fraud & Criminal Detection: Advanced Analytics & Information Sharing Technologies*. CRC Press, 2008.

[4] J. HAN, J. PEI, AND M. KAMBER, *Data mining: concepts and techniques*, Elsevier, 2011.

[5] J. TANG, S. ALELYANI, AND H. LIU, Feature selection for classification: A review. *Data Classification: Algorithms and Applications*, 37, 2014.

[6] Z. MICHAELWICZ, *Genetic Algorithms + Data Structures = Evolution Programs* Springer, 1992.
[7] J. DEVILLERS, *Genetic algorithms in molecular modeling,* Academic Press, 1996.
[8] D. DELEN, *Real-world data mining: applied business analytics and decision making*. FT Press, 2014.
[9] M. A. HALL, *Correlation-based feature selection for machine learning,* PhD Dissertation, The University of Waikato (New Zealand), 1999.

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4-8 July, 2017

A class of two-step Steffensen type with memory methods with efficiency 2

Vali Torkashvand [1] and Taher Lotfi [1]

[1] *Department of Mathematics, Hamedan Branch, Islamic Azad University, Hamedan, Iran*

emails: vali_torkashvand@yahoo.com, lotfitaher@yahoo.com

Abstract

In this work,we develop a new two-step method given by Ren et al.(2009).It includes three parameters.We approximate these parameters and increase the convergence order from 4 to 8. convergence analysis along with numerical illustrations are provided to justify and accomplish theoretical and practical aspects of the proposed methods.

Key words: Nonlinear equation, Newton's interpolatory polynomial, adaptive with memory method,R-order convergence,self accelerating parameter.

1. Introduction

Finding the root of a nonlinear equations or systems is an interesting task in numerical analysis and applied mathematics, which has attracted so much attention recently. In the last years, iterative techniques have been applied in many diverse fields as economics, engineering, physics, dynamical models,and so on.Newton's method is the most well-known method for solving nonlinear equations[1].However,the existence of first derivative is compulsory for the convergence of Newton's method, which bounds its applications in practice. To overcome on this difficulty,Steffensen replaced the first derivative of the function in the Newton's iterate by forward finite difference approximation[2].Both the methods possess the quadratic convergence and the same efficiency but second one is derivative free. Multipoint iterative methods for solving nonlinear equations are of great practical importance since they overcome theoretical limits of one-point methods concerning the convergence order and computational efficiency [2-11]. In the case of multipoint without memory methods,this requirement is closely connected with results of Kung and Traub [3],who conjectured that the order of convergence of any multipoint method without memory, consuming n + 1 function evaluations per iteration, cannot with this property are usually called optimal order.The exceed the bound 2^n.Multipoint methods recent past,researchers have focused on optimizy the existing non-optimal iterative methods without any additional evaluation of functions and derivatives.Let m_k represent the r+1 quantities x_k, $t_1(x_k)$, $t_2(x_k)$, ..., $t_r(x_k)$ and define an iterative process by $x_{k+1} = F(m_k; m_{k-1}, m_{k-2}, ...,$

m_{k-r}).Following Traub's terminology [4], F is called a multipoint iterative function with memory. To compare iterative methods theoretically, Ostrowski [5] introduced the idea of efficiency index and given by $d^{1/n}$, where d is the order of convergence and n is the number of function evaluations per iteration. In this paper we present an improvement of the existing with memory family Ren et al [6], constructed by introducing two more iterative parameters which are calculated with help of Newton's interpolatory polynomial of different degrees. In Section 2, a family of two-point methods without memory with improved order of convergence from 4 to 8 without adding more evaluations is presented.The comparisons of absolute errors and computational efficiencies are given in Section 3 to illustrate convergence behavior.Finally,we give the concluding remarks.

2. New modified three- parametric methods

In this section, we deal with modifying two-point without memory method by Ren et al [6].Its error equation has three accelerator elements.Ren et al method has the iterative expression

$$y_k = x_k - \frac{f\ x_k}{f\ x_k, y_k}\ , w_k = x_k + f\ x_k\ \ , \beta \in \mathrm{R}$$

$$x_{k+1} = y_k - \frac{f\ y_k}{f\ x_k, y_k\ + f\ y_k, w_k\ - f\ x_k, w_k\ + \beta\ y_k - x_k\ \ y_k - w_k} \quad (2.1)$$

To transform (2.1) in a method with memory, with three accelerators, we consider the following modification of (2.1) [11].

$$y_k = x_k - \frac{f\ x_k}{f\ x_k, y_k\ + pf[w_k]}\ , w_k = x_k + \gamma f\ x_k\ \ , \qquad (2.2)$$

$$x_{k+1} = y_k - \frac{f\ y_k}{f\ x_k, y_k\ + f\ y_k, w_k\ - f\ x_k, w_k\ + \beta(y_k - x_k)(y_k - w_k)}$$

where $\gamma\ \neq 0, \beta$ and p are arbitrary parameters.In what follows, we present the error equation of (2.2).

Theorem 2.1 Let $I\ \subseteq\ R$ be an open interval, $f: I \to R$ our times continuously differentiable and let αϵI be a simple zero of f. If the initial point x_0 is sufficiently close to α, then the method defined by (2.2) converges to α with order four. Moreover, the error equation is

$$e_{k+1} = \frac{1 + \gamma f l a\ ^2 (p + c_2)(\beta + f l a c_2\ p + c_2\ - f l a c_3)}{f l a} e_k^4 + O\ e_k^5 \qquad (2.3)$$

where $c_k = \frac{f^{\,k}(\alpha)}{k! f'(\alpha)}\ \ for\ k = 2,3, \ldots.$

It must be noted that schemes (2.1) and (2.2) are still methods without memory.In the next section, we discuss how they can be transformed in methods with memory with higher order of convergence without increasing the number of functional evaluations per step.

3. The development of a new method with memory

This section concerns with extracting a new method with memory from (2.2) by using three self-accelerating parameters. Theorem 2.1 states that modified method (2.2) has convergence order four if $\gamma \neq \frac{-1}{f'\ \alpha}$, $p \neq -c_2$ and $\beta \neq$ fla c_3 Now, we pose some questions: Is it possible to increase the convergence order of this method? If so, how can it be done and what is the new convergence order? For answering these questions, we look at the error equation (2.2).
It can be seen that if we set $\gamma = \frac{-1}{f'\ \alpha}$, $p = -c_2 = -\frac{f''\ \alpha}{2f'\ \alpha}$ and $\beta =$ fla $c_3 = \frac{f'''\ \alpha}{6}$, then at least the coefficient of e_k^4 disappears. However, we do not know α ,and consequently, $f'\ \alpha$, $f''\ \alpha$ and $f'''\ \alpha$ cannot be data comuted. On the other hand, we can approximate α using available data design of methods with memory consists on the calculation of the parameters $\gamma = \gamma_k$, $p = p_k$ and $\beta = \beta_k$as the iteration proceeds by the formulas $\gamma_k = \frac{-1}{f'\ \alpha}$, $p_k = -c_2 = -\frac{f''\ \alpha}{2f'\ \alpha}$ and $\beta_k =$ fla $c_3 = \frac{f'''\ \alpha}{6}$, $for\ k = 1,2, \ldots\ \ where\ \gamma_k, c_2$ and c_3 are approximations toγ_k, c_2 and c_3, respectively. To this end, the following approximates are applied

$$\gamma_k = \frac{-1}{N'_3(x_k)} \approx \frac{-1}{f'\ \alpha}, p_k = -\frac{N''_4\ w_k}{2N'_4\ w_k} \text{ and } \beta_k = \frac{N'''_5\ y_k}{6}\ , k = 1,2, \ldots. \quad (3.1)$$

where $N_3\ x_k = N_3\ t; x_k, x_{k-1}, y_{k-1}, w_{k-1}$, $N_4\ w_k = N_4\ t; w_k, x_k, x_{k-1}, y_{k-1}, w_{k-1}$, $N_5(y_k) = N_5(t; y_k, w_k, x_k, x_{k-1}, y_{k-1}, w_{k-1})$,

are Newton's interpolating polynomials of third ,fourth and fivth degree, set through four,five and six best available approximations (nodes) $x_k, x_{k-1}, y_{k-1}, w_{k-1}$, $w_k, x_k, x_{k-1}, y_{k-1}, w_{k-1}$ and $y_k, w_k, x_k, x_{k-1}, y_{k-1}, w_{k-1}$, respectively. It should be noted that if one uses lower Newton's interpolation, lower accelerators are obtained. Replacing the fixed parameters γ , p and β in the iterative formula (2.1) by varying γ_k, p_k and β_k calculated by (3.1), we propose the following new method with memory, denoted by NM7,

$$x_0, \gamma_0, p_0, \beta_0\ are\ given, then\ w_0 = x_0 + \gamma_0 f\ x_0\ ,$$

$$\gamma_k = -\frac{1}{N'_3\ x_k}\ , p_k = -\frac{N''_4\ w_k}{N'_4\ w_k}\ , \beta_k = -\frac{N'''_5\ y_k}{6},$$

$$y_k = x_k - \frac{f\ x_k}{f\ x_k, y_k\ + p_k f\ w_k}\ , w_k = x_k + \gamma_k f\ x_k\ ,$$

$$x_{k+1} = y_k - \frac{f\ y_k}{f\ x_k, y_k\ + f\ y_k, w_k\ - f\ x_k, w_k\ + \beta_k\ y_k - x_k\ \ y_k - w_k} \quad (3.2)$$

Here, we concerns the second question regarding convergence order of the method with memory (3.2). We need the next technical result (see[4]).

Lemma 3.1 If $\gamma_k = -\frac{1}{N'_3\ x_k}$, $p_k = -\frac{N''_4\ w_k}{N'_4\ w_k}$, $\beta_k = -\frac{N'''_5\ y_k}{6}$, k = 1,2, ...

then the estimates $(1+\gamma_k fla) \sim C_1 e_{k-1} e_{k-1,w} e_{k-1,y}, p + c_k \sim C_2 e_{k-1} e_{k-1,w} e_{k-1,y},$

$$\beta_k + flac_2\ p + c_2\ - flac_3 \sim C_3 e_{k-1} e_{k-1,w} e_{k-1,y} \quad (3.3)$$

hold, where $C_1, C_2\ and\ C_3$ are some asymptotic constants. The following result determines the convergence order of the two-point iterative method with memory (3.2).

Theorem 3.2 If an initial estimation x_0 is close enough to a simple root α of $f\ x\ = 0$ and $\gamma_0, p_0\ and\ \beta_0$ must be uniformly bounded above, being f a real sufficiently differentiable function, then the R-order of convergence of the two-point method with memory (3.2) is arbitrary close to 7.53.

Theorem 3.3 If an initial estimation x_0 is close enough to a simple root α of $f\ x\ = 0$ and $\gamma_0, p_0\ and\ \beta_0$ must be uniformly bounded above, being f a real sufficiently differentiable function, then the R-order of convergence of the two-point method adaptive with memory (3.2) is arbitrary close to 8 (NM8).

4. Numerical Examples and Conclusion

The errors $|x_k - \alpha|$ of approximations to the sought zeros, produced by the different methods at the first three iterations, are given in Tables 1, 2 where m(−n) stands for $m \times 10^{-n}$. These tables also include, for each test function, the initial estimation values and the last value of the computational order of convergence coc [9] computed by the expression

(if it is stable) $$coc = \frac{\mathrm{Log}(\frac{f(x_k)}{f(x_{k-1})})}{\mathrm{Log}(\frac{f(x_{k-1})}{f(x_{k-2})})} \ . \quad (4.1)$$

The software Mathematica 9, with 2000 arbitrary precision arithmetics, has been used in our computations.Several iterative methods with and without memery, for comparing with our Derivative-free Zheng et al two-step order 4 proposed methods, have been chosen as comes next (ZM)[7]. $y_k = x_k - \frac{f\ x_k}{f\ x_k, y_k}$, $w_k = x_k + \gamma f\ x_k$,

$$x_{k+1} = y_k - \frac{f\ y_k}{f\ x_k, y_k\ + f\ y_k, x_k, w_k\ \ y_k - x_k} \quad (4.2)$$

Derivative-free Petkovic et al two-step with memory with order $2 + \sqrt{6} \cong 4.44$ (PM)[8].

$$y_k = x_k - \frac{f\ x_k}{\emptyset_k}\ , w_k = x_k - \gamma f\ x_k\ \ , u_k = \frac{f\ y_k}{f\ x_k}, v_k = \frac{f\ y_k}{f\ w_k},$$

$$h\ u_k, v_k\ = \frac{1-u_k}{1+v_k}\ , \emptyset_k = \frac{f\ x_k\ - f\ w_k}{x_k - w_k}\ x_{k+1} = y_k - h\ u_k, v_k\ * \frac{f\ y_k}{\emptyset_k} \qquad (4.3)$$

Derivative-free Lotfi et al two-step with memory order 6 (LM)[10].

$$y_k = x_k - \frac{f\ x_k}{f\ x_k, y_k}\ , w_k = x_k + \gamma f\ x_k\ , x_{k+1} = y_k - \frac{f\ y_k}{p'_2\ y_k} \qquad (4.4)$$

$$p'_2\ y_k\ = \frac{2y_k - y_k - x_k}{x_k - w_k\ \ x_k - y_k} f\ x_k\ + \frac{2y_k - x_k - y_k}{w_k - x_k\ \ w_k - y_k} f\ w_k\ + \frac{2y_k - w_k - x_k}{y_k - w_k\ \ y_k - x_k} f\ y_k\ f\ w_k$$

Derivative-free Cordero et al two-step with memory order 7 (CM)[11].

$$\gamma_k = -\frac{1}{N'_3\ x_k}\ , p_k = -\frac{N''_4\ w_k}{N'_4\ w_k}\ ,$$

$$y_k = x_k - \frac{f\ x_k}{f\ x_k, y_k\ + p_k f\ w_k}\ , w_k = x_k + \gamma_k f\ x_k\ ,$$

$$x_{k+1} = y_k - \frac{f\ y_k}{f\ x_k, y_k\ + f\ y_k, w_k\ - f\ x_k, w_k\ + \beta\ y_k - x_k\ \ y_k - w_k} \qquad 4.5$$

Table 1 $\mathrm{f}_1\ \mathrm{t}\ = 1 + \frac{1}{\mathrm{t}^4} - \frac{1}{\mathrm{t}} - \mathrm{t}^2, \alpha = 1, \mathrm{x}_0 = 1.4, \mathrm{p}_0 = 0.1, \beta_0 = 0.01, \gamma_0 = 0.1$

Methods	$\lvert x_1 - \alpha \rvert$	$\lvert x_2 - \alpha \rvert$	$\lvert x_3 - \alpha \rvert$	*coc*	*EI*
ZM (4.3)	0.879 (–2)	0.305 (–8)	0.430 (–34)	4.00	1.587
LM (4.4)	0.177 (–1)	0.197 (–10)	0.267 (–63)	6.00	1.817
PM (4.3)	0.559 (–1)	0.114 (–5)	0.364 (–9)	4.46	1.646
RM (2.1), $\beta = 1$	0.494 (–1)	0.163 (–3)	0.259 (–13)	4.00	1.587
CM (4.5), $\beta = 0$	1.000 (1)	0.431 (–1)	0.200 (–8)	7.00	1.913
*NM*7 (3.2)	1.000 (1)	0.433 (–1)	0.186 (–10)	7.53	1.96
*NM*8 (3.3)	0.433 (–1)	0.186 (–10)	0.189 (–82)	8.00	2.00

Table 2 $\mathrm{f}_2\ \mathrm{t}\ =\ \mathrm{t} - 2\ \ \mathrm{t}^{10} + \mathrm{t} + 2\ \mathrm{e}^{-5\mathrm{t}}, \alpha = 2, \mathrm{x}_0 = 2.2, \mathrm{p}_0 = 0.1, \beta_0 = 0.01, \gamma_0 = 0.1$

Methods	$\lvert x_1-\alpha\rvert$	$\lvert x_2-\alpha\rvert$	$\lvert x_3-\alpha\rvert$	coc	EI
ZM (4.2)	0.957 (−3)	0.164 (−13)	0.162 (−56)	3.72	1.549
LM (4.4)	0.261 (0)	0.236(−1)	0.478(−10)	6.00	1.817
PM (4.3)	0.134 (−2)	0.166(−15)	0.877(−74)	4.44	1.644
RM (2.1), $\beta=1$	0.116 (−1)	0.460 (−8)	0.205 (−33)	3.96	1.582
CM (4.5), $\beta=0$	0.200 (0)	0.704 (−3)	0.368(−22)	7.00	1.913
NM7 (3.2)	0.200 (0)	0.846(−3)	0.434 (−25)	7.54	1.961
NM8 (3.3)	0.846 (−3)	0.434(−25)	0.439(−206)	8.00	2.00

In Tables 1–2 we have examined some methods with different kinds of convergence order. It is observed that these methods support their theoretical aspects.The last columns of Tables show computational efficiency index defined by IE = $\mathrm{COC}^{1/3}$. It is asymptotically 2. In other words, our three parametric with memory method (3.3), shows a behavior as optimal n-point without memory methods.Therefore, we can conclude the following:
We have developed a new three parametric with memory method which has efficiency index 2. However, such task is only of theoretical importance, since developing methods with efficiency index more than 2 is impossible.Moreover,our developed method (3.3) does not need any derivatives and can be used even for non-smooth functions. Studying basins of attraction of the methods in this work may be considered for future research.

Acknowledgement

This work was supported by Hamedan Branch, Islamic Azad University.

Referenes

[1] Ortega, J.M., Rheinboldt, W.G. (ed.).: Iterative Solutions of Nonlinear Equations in Several Variables,Ed. Academic Press, New York (1970).
[2] J. F. Steffensen, Remarks on iteration, Scandinavian Aktuarietidskr, vol.16, pp. 64–72,1933.
[3] H.T. Kung, J.F. Traub, Optimal order of one-point and multipoint iteration, J. Assoc. Comput. Math. 21(1974) 634–651.
[4] J. F. Traub, Iterative Methods for the Solution of Equations, Prentice Hall, New York, NY, USA,1964.
[5] Ostrowski, A.M.: Solution of Equations and System of Equations. Prentice-Hall, Englewood Cliffs, NJ (1964).
[6]Hongmin Ren ,Qingbiao Wu ,Weihong Bi, A class of two-step Steffensen type methods with fourth-order convergence, Applied Mathematics and Computation 209 (2009) 206–210.
[7] Q. Zheng, J. Li, and F. Huang, An optimal Steffensen-type family for solving nonlinear equations, Applied Mathematics and Computation, vol. 217, no. 23, pp. 2011, 9592–9597.
[8] M.S. Petkovic, S. Ilic, J. Dzunic, Derivative free two-point methods with and without memory for solving nonlinear equations, Applied Mathematics and Computation 217 (2010) 1887–1895.
[9] Jay, I.O.: A note on Q-order of convergence. BIT Numer. Math. 41, (2001) 422–429.
[10] T. Lotfi, F. Soleymani, Z. Noori, A. Kilicman, F. Khaksar Haghani, Efficient Iterative Methods with and without Memory Possessing High Efficiency Indices, Hindawi Publishing

CorporationDiscrete Dynamics in Nature and Society Volume 2014, Article ID 912796, 9 pageshttp://dx.doi.org/10.1155/2014/912796.
[11]A.Cordero,T.Lotfi,J.Torregrosa,P.Assari,K.Mahdiani.,Some new bi-accelarator two-point methods for solving nonlinear equations.,Comp.Appl.Math(2016)251-267.Doi:10.1007/s40314-014-0912-1.

Conservative finite-difference scheme for the 2D problem of femtosecond laser pulse interaction with kink structure of high absorption in semiconductor.

Vyacheslav A. Trofimov[1], Maria M. Loginova[1] and Vladimir A. Egorenkov[1]

[1] *Department of Computational Mathematics, Lomonosov Moscow State University, Russian Federation*

emails: vatro@cs.msu.ru, mloginova@cs.msu.ru, egorenkov-v-a@mail.ru

Abstract

The problem of high-intensive laser pulse influence on semiconductor with nonlinear absorption is considered. The conservative finite-difference scheme efficiency for complicated nonlinear processes computation is proposed. Computer simulation results are presented.

Key words: conservative finite-difference scheme, optical bistability, femtosecond pulse

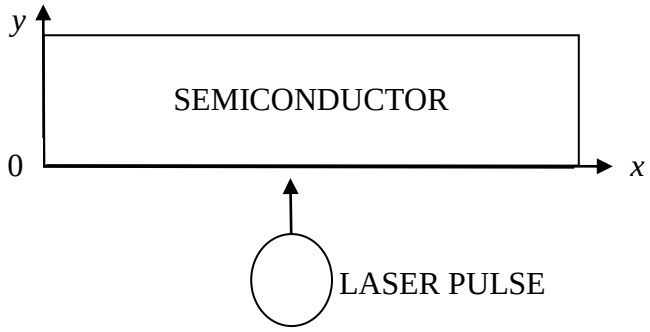
1. Introduction

We consider a 2D problem of femtosecond laser pulse interaction with semiconductor under the condition of field optical bistability (OB) occurrence. This is very promising phenomenon for the creation and developing of all-optical data processing. In the case of OB existence, the hysteresis dependence of semiconductor characteristics on laser pulse intensity and also kink structure of a high absorption appears. Existence of this phenomena is based on the laser radiation non-linear absorption of a semiconductor.

The laser pulse interaction with a semiconductor is described by the set of nonlinear partial differential equations. The laser pulse propagation in a semiconductor is described by a nonlinear Schrödinger equation with respect to the envelope (complex amplitude) of wave pocket. We developed a conservative finite-difference scheme (FDS) for this problem. It is a nonlinear one, so to realize it we proposed an original two-step iteration process. The FDS is also conservative one on the each of iterations. Computer simulation results confirmed

that the FDS possesses an asymptotic stability property. This property is very important because we should provide calculation during long time interval. One of the features of proposed approach for the nonlinear FDS realization is an opportunity of its generalizing for multidimensional problem.

2. Problem Statement and FDS



The process of laser pulse interaction with a semiconductor is described by the following set of 2D dimensionless differential equations concerning semiconductor characteristics - free electron concenration *n(x,y,t)*, ionized donors concentration *N(x,y,t)*, electric field potential $\varphi(x,y,t)$ [1, 2] and complex amplitude *A(x,y,t)*:

$$\frac{\partial^2\varphi}{\partial x^2}+\frac{\partial^2\varphi}{\partial y^2}=\gamma(n-N),\ \frac{\partial N}{\partial t}=G(n,N,\varphi)-R(n,N),\ 0<x<L_x,\ 0<y<L_y,\ t>0.$$

$$\frac{\partial n}{\partial t}=D_x\frac{\partial}{\partial x}\left(\frac{\partial n}{\partial x}-\mu_x n\frac{\partial\varphi}{\partial x}\right)+D_y\frac{\partial}{\partial y}\left(\frac{\partial n}{\partial y}-\mu_y n\frac{\partial\varphi}{\partial y}\right)++G(n,N,\varphi)-R(n,N),$$

$$\frac{\partial A}{\partial t}+iD_{Ax}\frac{\partial^2 A}{\partial x^2}+iD_{Ay}\frac{\partial^2 A}{\partial y^2}+\frac{\beta_A\delta_0}{2}\delta(n,N,\varphi)A=0\,,0<x<L_x,\ -L_y<y<L_y,\ t>0.$$

Boundary conditions (BC) and initial condition are written below:

$$\left(\frac{\partial n}{\partial x}-\mu_x n\frac{\partial\varphi}{\partial x}\right)\Bigg|_{x=0,L_x}=\left(\frac{\partial n}{\partial y}-\mu_y n\frac{\partial\varphi}{\partial y}\right)\Bigg|_{y=0,L_y}=0,\ \frac{\partial\varphi}{\partial x}\Bigg|_{x=0,L_x}=\frac{\partial\varphi}{\partial y}\Bigg|_{y=0,L_y}=0.$$

$$A\big|_{x=0,L_x}=A\big|_{y=0,L_y}=0,\ A\big|_{t=0}=e^{-\frac{(x-L_{xc})^2}{a^2}-(y-L_{yc})^{10}-2i\pi\chi(y-L_{yc})}.$$

This homogenous BC correspond to both the electric current absence through semiconductor faces and external electric field absence. Below the following functions are introduced: generation *G* and recombination *R* of free electrons

$$G(n,N,\varphi,A)=q_0|A|^2\,\delta(n,N,\varphi),\ R(n,N)=\frac{nN-n_0^2}{\tau_R}.$$

Absorption coefficient $\delta(n,N,\varphi)$ could be approximated by different ways, in our paper we consider its folowing approximation:

$$\delta_n(n,N,\varphi)=(1-N)e^{-\psi(1-\xi n)}.$$

It should be stressed, that we take into account the diffraction phenomena to describe the laser pulse reflection from the inhomogeneities induced by laser radiation in semiconductor. This is a very important feature of such problem statement and it has not been described in literature earlier.

The law of charge preservation takes place for this problem, therefore the FDS conservatism consists in validity of this invariant difference analog.

To solve the differential initial-boundary problem numerically we approximate it by the set of finite-difference equations. We use uniform space and time grids in rectangular domain and standard notations of the first and the second difference derivatives for this purpose. As a result, we propose the following FDS in the inner nodes of the grids:

$$\varphi_{\bar{x}x}+\varphi_{\bar{y}y}=\gamma(\hat{n}-\hat{N}),\quad \frac{\hat{N}-N}{\tau}=0.5\left(\hat{G}+G\right)-0.5\left(\hat{R}+R\right),$$

$$\frac{\hat{n}-n}{\tau}=0.5D_x\left(\hat{n}_{\bar{x}x}+n_{\bar{x}x}\right)+0.5D_y\left(\hat{n}_{\bar{y}y}+n_{\bar{y}y}\right)+0.5\left(\hat{G}+G\right)-0.5\left(\hat{R}+R\right)-$$

$$-0.5\frac{D_x\mu_x}{h_x}\left(\hat{n}_{k+0.5}\hat{\varphi}_x-\hat{n}_{k-0.5}\hat{\varphi}_{\bar{x}}+n_{k+0.5}\varphi_x-n_{k-0.5}\varphi_{\bar{x}}\right)-$$

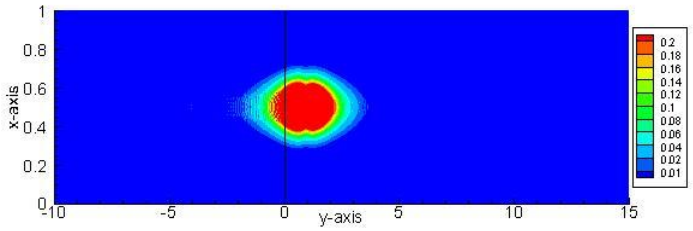
$$-0.5\frac{D_y\mu_y}{h_y}\left(\hat{n}_{j+0.5}\hat{\varphi}_y-\hat{n}_{j-0.5}\hat{\varphi}_{\bar{y}}+n_{j+0.5}\varphi_y-n_{j-0.5}\varphi_{\bar{y}}\right),$$

$$\frac{\hat{A}-A}{\tau}+iD_{A_x}\overset{0.5}{A}_{\bar{x}x}+iD_{A_y}\overset{0.5}{A}_{\bar{y}y}+\frac{\beta_A\delta_0}{2}\overset{0.5}{\delta}\overset{0.5}{A}=0.$$

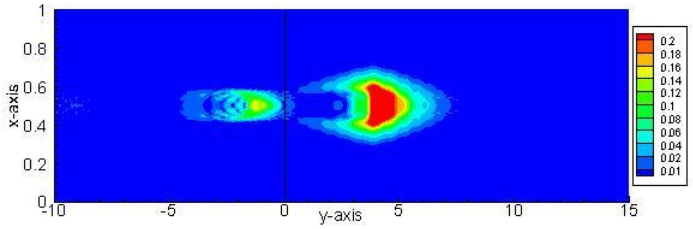
This FDS has the second order of approximation on time and space coordinates in inner grid nodes. BC for electric field potential and free electron concentration are approximated with the first order because of the scheme conservative property requirement.

To solve the obtained system of 2D nonlinear difference equations we use two-stage iteration process [3]. We constructed it in such a way, that the FDS becomes a conservative one on each of iterations. It is an important feature of our approach. Such approach allows to avoid disadvantages which arise from split-step method using, because this method accumulates computing mistakes at calculating non-stationary problem on a big time interval and, therefore, asymptotic stability property violation takes place.

3. Computer simulation results



(a)



(b)

Figure 1: Distribution of square root from laser pulse intensity $|A|^2$ at time moment t=3 (a), 6 (b) for computation with parameters $a = 0.1, \delta_0 = 0.25$, $q_0 = 4, n_0 = 0.01, \xi = 3, \psi = 2.553$ $D_x = D_y = 10^{-5}$, $\mu_x = \mu_y = 10^{-5}$, $D_{Ax} = 10^{-5}, D_{Ay} = \frac{1}{4\pi\chi}, \chi = 5$.

In Fig.1 an example of the laser pulse interaction with a high absorption domain in semiconductor is shown in two time moments. First time moment (Fig.1a) corresponds to laser pulse transmission the semiconductor face. We see a reflected pulse formation. In the second time moment (Fig.1b) a motion of reflected sub-pulse occures.

4. Acknowledgments

This work supported by the Russian Science Foundation (Grant №14-21-00081).

5. References

[1] R. SMITH, *Semiconductors*, Cambridge University Press, 1978.

[2] H.M. GIBBS, *Optical Bistability: Controlling Light with Light*, Academic Press, NY, 1985.

[3] V.A. TROFIMOV, M.M. LOGINOVA, V.A. EGORENKOV, *New two-step iteration process for solution of semiconductor plasma generation problem with arbitrary Boundary Conditions in 2D case*, WIT transactions on modelling and simulation, **59** (2015), 85-96.

Addendum

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

Exponentially fitted symmetric and symplectic diagonally implicit Runge-Kutta methods for Hamiltonian oscillators

Julius Osato Ehigie[1], Dongxu Diao[2], Ruqiang Zhang[2], Yonglei Fang[3], Xilin Hou[1] and Xiong You[2]

[1] *College of Horticulture, Nanjing Agricultural University, Nanjing 210095, China*

[2] *College of Sciences, Nanjing Agricultural University, Nanjing 210095, China*

[3] *School of Mathematics and Statistics, Zaozhuang University, Zaozhuang 277160, China*

emails: jehigie@unilag.edu.ng, 2016111003@njau.edu.cn, 2016204021@njau.edu.cn, ylfangmath@163.com, hxl@njau.edu.cn, youx@njau.edu.cn

Abstract

The symmetric and symplectic conditions and exponential fitting conditions for modified Runge-Kutta methods are presented. Two new exponentially fitted symmetric and symplectic diagonally implicit Runge-Kutta (EFSSDIRK) methods of respective second order and fourth order are derived. Implementation on the Kepler problem shows the high effectiveness and competence of the new methods compared to three four order DIRK methods in the recent literature.

Keywords: Exponential fitting, symmetry, symplecticity, Hamiltonian problem
MSC 2000: 65L04, 65L05, 65L06

1 Introduction

In this paper, we focus on the effective integration of the initial value problem of the first order differential system in the form

$$\dot{y} = f(t, y), \quad y(x_0) = y_0, \tag{1}$$

which arise in sciences such as celestial mechanics, astrophysics, chemistry biology and engineering. In many applications, the problem (1) takes the form of a Hamiltonian system,

that is the function $f(t,y) = J^{-1}\nabla H(p,q)$ with $J = \begin{pmatrix} 0 & I_d \\ -I_d & 0 \end{pmatrix}$, where $H = H(p,q) = H(p_1,\ldots,p_d,q_1,\ldots,q_d)$ is the *Hamiltonian energy*. A special class of numerical integrators, called *symplectic integrators*, have been proved to be very effective for Hamiltonian systems due to the preservation of the symplectic structure of the solution.

2 Exponentially fitted symmetric and symplectic diagonally implicit Runge-Kutta methods

Suppose the solution to the problem (1) is oscillatory with frequency ω, we consider the modified implicitly diagonal Runge-Kutta methods of the form

$$\begin{aligned} Y_i &= \eta_i y_n + h\sum_{j=1}^{i} a_{ij} f(x_n + c_j h, Y_j), \quad i = 1,\ldots,s, \\ y_{n+1} &= y_n + h\sum_{i=1}^{s} b_i f(x_n + c_i h, Y_i), \end{aligned} \tag{2}$$

where h is the stepsize, η_i, b_i, a_{ij}, c_i, $1 \le j \le i \le s$ are assumed to be even functions of $z = \mathrm{i}\omega h$ ($\mathrm{i}^2 = -1$). It is usually assumed that $\lim_{z\to 0}\eta_i(z) = 1$ so that as $z \to 0$, the scheme (2) reduces to a traditional Runge-Kutta method.

According the Tsitouras et al. [6], the local truncation error of the modified RK method (2) can be expressed as

$$LTE = \sum_{\tau\in\mathcal{T}} \frac{h^{\rho(\tau)}}{\rho(\tau)!}(\gamma(\tau)b(z)^T\Phi(\tau) - 1)\alpha(\tau)\mathcal{F}(\tau)(y_0) + \sum_{\tau\in\widetilde{\mathcal{T}}} \frac{h^{\rho(\tau)}}{\rho(\tau)!}(\gamma(\tau)b(z)^T\widetilde{\Phi}(\tau) - 1)\alpha(\tau)\widetilde{\mathcal{F}}(\tau)(y_0), \tag{3}$$

where $\mathcal{T}$ ($\widetilde{\mathcal{T}}$) is the set of (modified) rooted trees, $\mathcal{F}(\tau)(y_0)$ ($\widetilde{\mathcal{F}}(\tau)(y_0)$) is the elementary differential of f associated to $\tau \in \mathcal{T}$ ($\widetilde{\mathcal{T}}$) at y_0, $\gamma(\tau)$ is the density of the tree τ, $\alpha(\tau)$ is the number of monotonic labellings of τ and $\Phi(\tau)$ ($\widetilde{\Phi}(\tau)$) is the elementary weight coefficient which depends on $A(z)$ and $\eta(z)$.

The modified Runge-Kutta method of the form (2) has order p if for any sufficiently smooth problem (1), under the assumption that $y(x_0) = y_0$, the local truncation error of the solution satisfies

$$LTE = y(x_0 + h) - y_0 = \mathcal{O}(h^{p+1}) \quad \text{as} \quad h \to 0. \tag{4}$$

 ISBN: 978-84-608-6082-2

For example, the conditions for the method (2) to be of order four are listed as follows:

$$\begin{array}{lll}
\eta_i = 1 + \mathcal{O}(z^2), & \sum\limits_i b_i = 1 + \mathcal{O}(z^4), & \sum\limits_i b_i\eta_i = 1 + \mathcal{O}(z^4), \\
\sum\limits_{i,j} b_i a_{ij}\eta_j = \frac{1}{2} + \mathcal{O}(z^3), & \sum\limits_{i,j} b_i\eta_i a_{ij} = \frac{1}{2} + \mathcal{O}(z^3), & \sum\limits_{i,j} b_i a_{ij} = \frac{1}{2} + \mathcal{O}(z^3), \\
\sum\limits_{i,j,k} b_i a_{ij} a_{jk} = \frac{1}{6} + \mathcal{O}(z^2), & \sum\limits_{i,j} b_i c_i^2 = \frac{1}{3} + \mathcal{O}(z^2), & \sum\limits_{i,j} b_i c_i^3 = \frac{1}{4} + \mathcal{O}(z), \\
\sum\limits_{i,j} b_i a_{ij} c_j^2 = \frac{1}{12} + \mathcal{O}(z), & \sum\limits_{i,j,k} b_i c_i a_{ij} a_{jk} = \frac{1}{8} + \mathcal{O}(z), & \sum\limits_{i,j,k,l} b_i a_{ij} a_{jk} a_{kl} = \frac{1}{24} + \mathcal{O}(z).
\end{array} \tag{5}$$

The method (2) is symmetric if

$$\begin{array}{l}
c_i = 1 - c_{s+1-i}, \quad \eta_i = \eta_{s+1-i}, \quad b_i(z) = b_{s+1-i}(z), \quad i = 1, \ldots, s, \\
a_{ij} = \eta_i(z) b_j(z), \quad 1 \le j < i \le s, \\
a_{ii} + a_{s+1-i,s+1-i} = \eta_i(z) b_i(z), \quad i = 1, \ldots, s.
\end{array} \tag{6}$$

The method (2) is symplectic if its coefficients satisfy the following conditions

$$a_{ij} = \eta_i b_j \quad \text{for} \quad 1 \le j < i \le s, \quad a_{ii} = \frac{1}{2}\eta_i b_i \quad \text{for} \quad 1 \le i \le s. \qquad i, j = 1, \ldots, s. \tag{7}$$

The method (2) is exponentially fitted if its internal stages and the update are exact for the functions $\{\exp(\pm\omega x)\}$ leading to:

$$\begin{array}{l}
\sum\limits_{j=1}^{i} a_{ij} \sinh(c_i z) = \frac{\cos(c_i z) - \eta_i}{z}, \quad \sum\limits_{j=1}^{i} a_{ij} \cosh(c_i z) = \frac{\sin(c_i(z) z)}{z}, \quad i = 1, \ldots, s, \\
\sum\limits_{i=1}^{s} b_i \sinh(c_i z) = \frac{\cos(z) - 1}{z}, \quad \sum\limits_{i=1}^{s} b_i \cosh(c_i z) = \frac{\sin(z)}{z}.
\end{array} \tag{8}$$

A 2-stage symmetric and symplectic exponentially fitted DIRK method of order two is given by

$$\begin{array}{l}
c_1 = 1/4, \quad c_2 = 3/4, \quad \eta_1 = \frac{1}{\cosh(c_1 z)}, \quad \eta_2 = \eta_1, \quad a_{11} = \frac{\tanh(c_1 z)}{z}, \quad a_{21} = 2a_{11}, \\
a_{22} = \frac{\sin(c_2 z) - 2\sinh(c_1 z)}{z\cosh(c_2 z)}, \quad b_1 = \frac{\cosh((c_2 - 1)z) - \cosh(c_2 z)}{z\sinh((c_1 - c_2)z)}, \quad b_2 = b_1.
\end{array} \tag{9}$$

The method is denoted as EFSSDIRK2.

A 3-stage symmetric and symplectic exponentially fitted DIRK method of order four is

 ISBN: 978-84-608-6082-2

obtained as follows

$$
\begin{aligned}
&c_1 = \left(\tfrac{1}{3} + \tfrac{2^{1/3}}{6} + \tfrac{2^{2/3}}{12}\right) - \tfrac{1}{864}\left(1 + 2^{2/3} + 2^{4/3}\right) z^2 - \tfrac{1}{414720}\left(23 \cdot 2^{1/3} - 22 \cdot 2^{2/3} - 36\right) z^4 + \ldots,\\
&c_2 = 1/2, \quad c_3 = 1 - c_1, \quad \eta_1 = \tfrac{1}{\cosh((1-c_3)z)}, \quad \eta_3 = \eta_1\\
&\eta_2 = \tfrac{4(3c_3^2-3c_3+1) - \cosh\left(\frac{z}{2}(1-2c_3)\right)}{4\left(3c_3^2-3c_3+1\right)\cosh\left(\frac{z}{2}\right) + z\sinh\left(\frac{z}{2}(1-2c_3)\right) - \cosh(z-c_3 z)},\\
&a_{11} = a_{33} = \tfrac{1}{2}\eta_3 b_3, \quad a_{21} = \eta_2 b_3, \quad a_{22} = \tfrac{1}{2}\eta_2 b_2, \quad a_{31} = 2a_{11}, \quad a_{32} = \eta_3 b_2.\\
&b_1 = \tfrac{-2z\sinh\left(\frac{z}{2}(1-2c_3)\right) + \cosh((1-c_3)z) - \cosh(3z)}{z\left(8c_3^3\sinh\left(\left(c_3-\frac{1}{2}\right)z\right) + 2\sinh\left(\frac{z}{2}(1-2c_3)\right)\left(\cosh\left(\frac{z}{2}(1-2c_3)\right) + 4(c_3-1)^3\right)\right)},\\
&b_2 = \tfrac{4\sinh\left(\frac{z}{2}(1-2c_3)\right)\left(z\cosh\left(\frac{z}{2}(1-2c_3)\right) - 4\left(3c_3^2-3c_3+1\right)\sinh\left(\frac{z}{2}\right)\right)}{z\left(8c_3^3\sinh\left(\left(c_3-\frac{1}{2}\right)z\right) + 2\sinh\left(\frac{z}{2}(1-2c_3)\right)\left(\cosh\left(\frac{z}{2}(1-2c_3)\right) + 4(c_3-1)^3\right)\right)},\\
&b_3(z) = b_1(z).
\end{aligned}
\tag{10}
$$

The method is denoted as EFSSDIRK4. As $z \to 0$, the method reduces to a classical four-stage DIRK proposed by Sans-Serna and Abia [5] and Feng and Qin [3], which is denoted as SSDIRK4.

3 Numerical experiments on the Kepler problem

To examine the numerical effectiveness of the newly constructed SSEFDIRK4, we apply it to the Kepler problem

$$
p_1' = -\tfrac{q_1}{\sqrt{(q_1^2+q_2^2)^3}}, \quad p_2' = -\tfrac{q_2}{\sqrt{(q_1^2+q_2^2)^3}}, \quad q_1' = p_1, \quad q_2' = p_2, . \tag{11}
$$

The Hamiltonian of this problem is given by $H(p_1, p_2, q_1, q_2) = \frac{1}{2}(p_1^2 + p_2^2) - \left(q_1^2 + q_2^2\right)^{-1/2}$. For initial values $q_1(0) = 1,\ p_1(0) = 0,\ q_2(0) = 0,\ p_2(0) = 1$, the exact solution of (11) is $q_1(x) = \cos(x)$, $q_2(x) = \sin(x)$.

Four highly efficient DIRK methods of order four are selected from the literature for comparison listed as follows: DIRK4A (Alexander [1]), DIRK4C (Cash [2]), SSDIRK4SA (Sanz-Serna and Abia [5]) and TFSDIRK4K (Kalogiratou [4]).

The effectiveness of the previous methods will be compared in terms of accuracy and efficiency. We test the accuracy of the methods in the terms (P, Q), where P represents the decimal logarithm of the maximum global error $\log_{10}(\text{MGE})$ and Q represents the computational effort required by each method which is measured by the decimal logarithm of the number of function evaluations$\log_{10}(\text{NFE})$. We also check the efficiency in terms of $R = \text{MGE} \times \text{NFE}$.

We take the fitting frequency as $\omega = 1$ and integrate the problem on the interval $[0, 100\pi]$ with different stepsizes $h = \frac{\pi}{16}, \frac{\pi}{32}, \frac{\pi}{64}, \frac{\pi}{128}$. The numerical results are presented in Table 1 and Table 2 .

 ISBN: 978-84-608-6082-2

Table 1: Maximum global error vs number of function evaluations

(P,Q)	$\frac{\pi}{16}$	$\frac{\pi}{32}$	$\frac{\pi}{64}$	$\frac{\pi}{128}$
EFSSDIRK2	$(-11.84, 4.57)$	$(-11.41, 4.74)$	$(-11.68, 4.95)$	$(-11.66, 5.19)$
EFSSDIRK4	$(-8.00, 4.96)$	$(-10.62, 5.09)$	$(-11.41, 5.25)$	$(-11.77, 5.45)$
DIRK4N	$(0.34, 4.85)$	$(0.22, 4.97)$	$(-1.24, 5.11)$	$(-2.76, 5.25)$
SSDIRK4SA	$(-0.35, 4.78)$	$(-1.61, 4.90)$	$(-2.83, 5.06)$	$(-4.03, 5.22)$
TFSDIRK4K	$(-6.42, 4.76)$	$(-5.67, 4.87)$	$(-6.46, 5.04)$	$(-6.08, 5.19)$

Table 2: Comparison of computational efficiency

R	$\frac{\pi}{16}$	$\frac{\pi}{32}$	$\frac{\pi}{64}$	$\frac{\pi}{128}$
EFSSDIRK2	$5.26E-8$	$2.12E-7$	$1.87E-7$	$3.33E-7$
EFSSDIRK4	$9.04E-4$	$2.96E-6$	$6.99E-7$	$4.82E-7$
DIRK4N	$1.51E-0$	$1.55E-0$	$7.37E-2$	$3.15E-3$
SSDIRK4SA	$2.69E-0$	$1.97E-1$	$1.72E-2$	$1.55E-3$
TFSDIRK4K	$2.17E-2$	$1.56E-1$	$3.75E-2$	$1.26E-1$

From Tables 1 and 1 we can see that the new methods EFSSDIRK2 and EFSSDIRK3s4 outperform the other methods we choose for comparison when applied to the Kepler problem. Among all the five methods, the second order EFSSDIRK2 is the most accurate and most efficient although the other methods are fourth order.

Acknowledgements

This research was partially supported by the National Natural Science Foundation of China (NSFC) under Grant No. 11171155 and No. 11571302.

References

[1] R. Alexander, Diagonally implicit Runge-Kutta methods for stiff ODEs, *SIAM J. Numer. Anal.*. 14(6) (1997) 1006–1021.

[2] J.R. Cash, Diagonally implicit Runge-Kutta formulae for the numerical integration of nonlinear two-point boundary value problems, *Comput. Math. Appl.* 10 (1984) 123–137.

[3] K. Feng, M. Qin, *Symplectic Geometric Algorithms for Halmitonian systems*, Zhejiang Publishing United Group and Springer-Verlag, 2010.

[4] Z. Kalogiratou. Diagonally implicit trigonometrically fitted symplectic Runge-Kutta methods, *Appl. Math. Comput.* 219 (2013), 7406–7412.

 ISBN: 978-84-608-6082-2

[5] J.M. Sanz-Serna, L. Abia, Order conditions for canonical Runge-Kutta schemes, *SIAM J. Numer. Anal.* 28 (1991) 1081–1096.

[6] Ch. Tsitouras, I.Th. Famelis, T.E. Simos, On modified RungeKutta trees and methods, *Comput. Math. Appl.* 62 (2011) 21012111.

 ISBN: 978-84-608-6082-2

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

Multidimensional adapted RKN methods for multi-frequency oscillatory systems

Yonglei Fang[1]

[1] *School of Mathematics and Statistics, Zaozhuang University*

emails: ylfangmath@163.com

Abstract

Multidimensional adapted Runge-Kutta-Nyström (MARKN) methods with optimized phase properties for multi-frequency oscillatory systems are obtained in this paper. The new methods are of order five. We analyze the stability and the phase properties of the higher order methods. Numerical results are carried out to show the efficiency of our new methods for the numerical integration of the multi-frequency oscillatory systems.

Key words: Multidimensional ARKN pair, Multi-frequency oscillatory system, Variable step-size, Phase-lag.
MSC 2000: 65L05

1 Introduction

In this paper we are concerned with the numerical integration of the system of the form

$$\begin{cases} y''(t) + My(t) = f(y(t)), \quad t \in [t_0, t_{\text{end}}], \\ y(t_0) = y_0, \quad y'(t_0) = y_0', \end{cases} \tag{1}$$

in which $M \in \mathbb{R}^{m\times m}$ is a symmetric positive semi-definite matrix (stiffness matrix) and $f : \mathbb{R}^m \to \mathbb{R}^m$, $y_0 \in \mathbb{R}^m, y_0' \in \mathbb{R}^m$. We will consider the construction of multidimensional ARKN methods with optimized phase properties for the numerical integration of multi-frequency oscillatory systems.

 ISBN: 978-84-617-8694-7

2 Multidimensional ARKN methods

For any matrix $M \in \mathbb{R}^{m\times m}$, we introduce the following matrix-valued functions [1]:

$$\phi_j(M) := \sum_{k=0}^{\infty} \frac{(-1)^k M^k}{(2k+j)!}, \quad j = 0, 1, 2, \ldots. \tag{2}$$

For the numerical integration of problem (1), a multidimensional ARKN method has the form

$$\begin{cases} Y_i = y_n + hc_i y_n' + h^2 \sum\limits_{j=1}^{s} \bar{a}_{ij}\big(f(Y_j) - MY_j\big), \qquad i = 1, \cdots, s, \\ y_{n+1} = \phi_0(V) y_n + h\phi_1(V) y_n' + h^2 \sum\limits_{i=1}^{s} \bar{b}_i(V) f(Y_i), \\ y_{n+1}' = \phi_0(V) y_n' - hM\phi_1(V) y_n + h \sum\limits_{i=1}^{s} b_i(V) f(Y_i), \end{cases} \tag{3}$$

in which the weight functions $b_i(V) \in R^{m\times m}$ and $\bar{b}_i(V) \in R^{m\times m}, i = 1, \cdots, s$ in the final stages are functions of V with $V = h^2M$ and the scheme (3) can be expressed in the Butcher tableau as

$$\begin{array}{c|c} c & \bar{A} \\ \hline & \bar{b}^T(V) \\ \hline & \bar{b}^T(V) \end{array} = \begin{array}{c|ccc} c_1 & \bar{a}_{11} & \ldots & \bar{a}_{1s} \\ \vdots & \vdots & \ddots & \vdots \\ c_s & \bar{a}_{s1} & \cdots & \bar{a}_{ss} \\ \hline & \bar{b}_1(V) & \cdots & \bar{b}_s(V) \\ \hline & b_1(V) & \ldots & b_s(V) \end{array}$$

or equivalently by the quadruple $(c, \bar{A}, \bar{b}(V), b(V))$. The order conditions for the multidimensional ARKN method are given by Wu et. al in [2].

The stability of ARKN methods can be analyzed by the test equation of the form (see [3])

$$y''(t) + \omega^2 y(t) = -\epsilon y(t), \tag{4}$$

with $\omega^2 + \epsilon > 0$. Applying the ARKN methods to the test problem (4) yields

$$\begin{pmatrix} y_{n+1} \\ hy_{n+1}' \end{pmatrix} = M(V, z) \begin{pmatrix} y_n \\ hy_n' \end{pmatrix},$$

where

$$M(V, z) = \begin{pmatrix} \phi_0(V) - z\bar{b}^T(V)N^{-1}e & \phi_1(v) - z\bar{b}^T(V)N^{-1}c \\ -V\phi_1(V) - zb^T(V)N^{-1}e & \phi_0(V) - zb^T(V)N^{-1}c \end{pmatrix},$$

with $N = I + (V + z)\bar{A}$, $e = (1, 1, \cdots, 1)^T$ and I the identity matrix.

For the numerical integration of oscillatory problems, it is very important to consider the phase properties of ARKN methods. So, we refer to the following definition.

 ISBN: 978-84-617-8694-7

Definition 2.1 *The quantities [1, 3, 4, 5]*

$$\phi(H) = H - \arccos\left(\frac{\mathrm{tr(M)}}{2\sqrt{\det(M)}}\right), \quad d(H) = 1 - \sqrt{\det(M)},$$

are called the dispersion error and the dissipation error, respectively, where $H = \sqrt{V+z}$ *. Therefore, an ARKN method is said to be dispersive of order* q *and dissipative of order* p *if*

$$\phi(H) = \mathcal{O}(H^{q+1}), \quad d(H) = \mathcal{O}(H^{p+1}).$$

If $\phi(H) = 0$ *and* $d(H) = 0$*, then the method is said to be zero dispersive and zero dissipative.*

3 Construction of new multidimensional ARKN methods

We consider the fourth-stage explicit multidimensional ARKN methods which are expressed by the following Butcher-tableau:

$$\begin{array}{c|c} c & \bar{A} \\ \hline & \bar{b}^T(V) \\ \hline & b^T(V) \end{array} = \begin{array}{c|cccc} 0 & 0 & 0 & 0 & 0 \\ c_2 & \bar{a}_{21} & 0 & 0 & 0 \\ c_3 & \bar{a}_{31} & \bar{a}_{31} & 0 & 0 \\ c_4 & \bar{a}_{41} & \bar{a}_{42} & \bar{a}_{43} & 0 \\ \hline & \bar{b}_1(V) & \bar{b}_2(V) & \bar{b}_3(V) & \bar{b}_4(V) \\ \hline & b_1(V) & b_2(V) & b_s(V) & b_4(V) \end{array}$$

With the choice of $\bar{A}e = \frac{c^2}{2}$ and we solve the order condition equations up to order five and obtain

$$\bar{b}_1(V) = \bar{b}_1(c_2, c_3, c_4, V), \bar{b}_2(V) = \bar{b}_2(c_2, c_3, c_4, V), \bar{b}_3(V) = \bar{b}_2(c_2, c_3, c_4, V),$$

$$\bar{b}_4(V) = \bar{b}_2(c_2, c_3, c_4, V), b_1(V) = b_1(c_2, c_3, c_4, V), b_2(V) = b_2(c_2, c_3, c_4, V),$$

$$b_3(V) = b_3(c_2, c_3, c_4, V), b_4(V) = b_4(c_2, c_3, c_4, V), \bar{a}_{31} = \frac{c_3^2}{2} - a_{32}, \ \bar{a}_{41} = \frac{c_4^2}{2} - a_{42} - a_{43},$$

$$\bar{a}_{43} = (c_4(5c_2 - 2)(c_2 - c_4)(c_3 - c_4))/(10c_3(c_3(4c_3 - 3) + c_2^2(6c_3 - 4) + c_2(3 - 6c_3^2))),$$

$$\bar{a}_{32} = (c_2 - c_3)c_3(5c_4 - 4)/(10c_2(3 - 4c_4 + c_2(6c_4 - 4))),$$

$$\bar{a}_{42} = (-2a_{43}c_3^2(c_3(4c_3 - 3) + c_2^2(6c_3 - 4) + c_2(3 - 6c_3)) + (c_2 - c_4)c_4((c_2 - c_3)c_3(c_3 - c_4)$$

$$+2a_{32}c_2(3 - 4c_2 - 4c_4 + 6c_2c_4)))/(2c_2(c_2 - c_3)c_3(3 - 4c_3 + c_2(6c_3 - 4))),$$

$$c_3 = (12 - 15c_2 - 15c_4 + 20c_2c_4)/(15 - 20c_2 - 20c_4 + 30c_2c_4).$$

 ISBN: 978-84-617-8694-7

As an example, choosing $c_2 = \frac{3}{10}$, we give the dispersion and the dissipation of the methods

$$\phi(H) = -\frac{\epsilon\big(45(164 - 433c_4 + 280c_4^2)\epsilon + (4752 - 15643c_4 + 11830c_4^2)\omega^2\big)H^7}{2016000((11c_4 - 9)\big(\epsilon + \omega^2)^2\big)} + \mathcal{O}(H^9),$$

$$d(H) = -\frac{\epsilon\big((96 - 254c_4 + 165c_4^2)\epsilon + (69 - 221c_4 + 165c_4^2)\omega^2\big)H^6}{14400\big((11c_4 - 9)(\epsilon + \omega^2)^2\big)} + \mathcal{O}(H^8).$$

Assuming that $\omega \gg \epsilon$, we select the free parameter c_4 so that the dispersion or the dissipation are optimized.

Case(a): Firstly, we choose the free parameter so that the error constant of the dispersion is minimal in the sense that this constant is of order $\mathcal{O}(\epsilon)$, obtaining $c_4 = (15643 + \sqrt{19838809})/23660$. We denote this method as ARKN4S5I.

Case(b): Secondly, we choose the free parameter so that the error constant of the dissipation is minimal and we obtain $c_4 = (221 + \sqrt{3301})/330$. We denote this method as ARKN4S5II.

4 Numerical experiments

In this section, we will compare the numerical performance of the new methods with some existing codes proposed in the scientific literature. The criterion used in the numerical comparisons is the decimal logarithm of the maximum global error (LOG10 (ERROR)) versus the computational effort measured by the number of function evaluations (NUMBER OF FUNCTION EVALUATIONS) required by each method. We select two multidimensional multi-frequency oscillatory systems as test problems and the methods for comparison are listed below:

- ARKN3S4F: the three-stage ARKN method of order four with minimal error constant of dispersion error given in [3].
- ARKN3S4W: the three-stage ARKN method of order four with minimal error constant of dissipation error given in [1].
- W1ERKN3s3, W2ERKN3s3: the two three-stage multidimensional ERKN methods derived in [5].
- SSMERKN3S4: the three-stage symplectic and symmetric Multidimensional ERKN given in [6].
- ARKN4S5W: the four-stage multidimensional ARKN method given in [1].

 ISBN: 978-84-617-8694-7

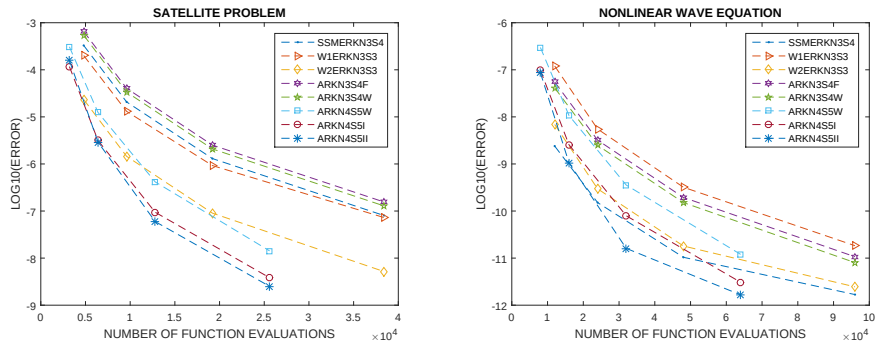


Figure 1: Efficiency curves in Problem 1 and Problem 2

- ARKN4S5I, ARKN4S5II: the two four-stage multidimensional ARKN methods derived in this paper.

Problem 1. We consider the equation governing the motion of an artificial satellite [7]

$$\vec{u}'' + \frac{\hbar}{2}\vec{u} = -\frac{1}{4}\frac{\partial(V|\vec{u}|^2)}{\partial\vec{u}},$$

with $\hbar = \dfrac{K^2}{r_0} - \dfrac{1}{2}|\dot{u}_0|^2 - V_0$, where $\hbar$ is the total energy of the elliptic motion, V is the perturbing potential of the earth. The initial conditions are considered on an elliptic equatorial orbit with $e = 0.1$, $K^2 = 3.98601 \times 10^5$, $r = 6.8 \times 10^3$, $\vartheta = 3.0 \times 10$, $\varrho = 3.844 \times 10^5$, $\lambda = 4.90266 \times 10^3$. The problem is solved on the interval $[0, 100]$ and the numerical results are reported in Fig. 1 (on the left).

Problem 2. Consider the initial-boundary value problem of the nonlinear wave equation

$$\begin{cases} \dfrac{\partial^2 u}{\partial t^2} - \dfrac{\partial^2 u}{\partial x^2} = u^5 - u^3 - \dfrac{1}{2}u, \quad 0 < x < 1, t > 0, \\ u(0,t) = 0, \quad u(1,t) = 0, \quad u(x,0) = \dfrac{x(1-x)}{100}, \quad u_t(x,0) = 0. \end{cases}$$

The method of lines is a technique for solving partial differential equation (PDEs). By using second order symmetric differences, the problem is converted to the initial value problem of an oscillatory system of the form

$$\begin{cases} \ddot{U} + MU = F(U), \\ U(0) = \left(\dfrac{x_1(1-x_1)}{100}, \cdots, \dfrac{x_{N-1}(1-x_{N-1})}{100}\right)^T, \ U'(0) = \mathbf{0}, \end{cases}$$

 ISBN: 978-84-617-8694-7

in which $U(t) = (u_1(t), \cdots, u_{N-1}(t))^T, u_i(t) \approx u(x_i, t), x_i = i\Delta x,\ i = 1, \cdots, N-1, \Delta x = 1/N$,

$$M = \frac{1}{\Delta x^2}\begin{pmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ & & & -1 & 2 \end{pmatrix},$$

$$F(U) = \left(u_1^5 - u_1^3 - \frac{1}{2}u_1, \cdots, u_{N-1}^5 - u_{N-1}^3 - \frac{1}{2}u_{N-1}\right)^T.$$

In this test, we choose $N = 20$. The system is integrated on the interval [0,100] and the numerical results are reported in Fig. 1 (on the right).

Acknowledgement

This research was partially supported by the National Natural Science Foundation of China (NSFC) under Grant No. 11571302.

References

[1] X. Wu, B. Wang, *Multidimensional adapted Runge-Kutta-Nyström methods for oscillatory systems*, Comput. Phys. Commun. **181** (2010) 1955–1962.

[2] X. Wu, X. You, J. Xia, *Order conditions for ARKN methods solving oscillatory systems*, Comput. Phys. Commun. **180** (2009) 2250–2257.

[3] J.M. Franco, *New methods for oscillatory systems based on ARKN methods*,Appl. Numer. Math. **56** (2006) 1040–1053.

[4] X. Wu, B. Wang, K. Liu, H. Zhao, *ERKN methods for long-term integration of multidimensional orbital problems*, Appl. Math. Model. **37** (2013) 2327–2336.

[5] X. Wu, X. You, W. Shi, B. Wang, *ERKN integrators for systems of oscillatory second-order differential equations*, Comput. Phys. Commun. **181** (2010) 1873–1887.

[6] X. Wu, B. Wang, J. Xia, *Explicit symplectic multidimensional exponential fitting modified Runge-Kutta-Nystrom methods*, BIT Numer. Math. **52** (2012) 773-795.

[7] E.L. Stiefel, G. Scheifele, Linear and Regular Celestial Mechanics, Springer-Verlag, New York, 1971.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

An adapted four-step method for the numerical solution of perturbed oscillators

Shiwei Liu[1], Juan Zheng[1], Xiong You[2] and Yonglei Fang[1]

[1] *School of Mathematics and Statistics, Zaozhuang University, Zaozhuang 277160, China*

[2] *College of Sciences, Nanjing Agricultural University, Nanjing 210095, China*

emails: sw.goof@gmail.com, llzhengjuan@sina.com, youx@njau.edu.cn, ylfangmath@163.com

Abstract

An adapted explicit four-step method for the numerical integration of perturbed oscillators is obtained by refining the classical method of power series. The local truncation error, phase properties and linear stability of the new method are analyzed. Numerical experiments are reported to show the accuracy and efficiency of the new method when it is compared with some high-quality methods recently proposed in the literature.

Keywords: Variation-of-constant formula; adapted method; perturbed oscillator; efficiency

1 Introduction

We are concerned the numerical integration of perturbed oscillators which are second-order initial value problems of the form

$$y'' + w^2 y = g(t, y), \quad y(t_0) = y_0, \quad y'(t_0) = y'_0, \tag{1}$$

where the magnitude of the perturbing force satisfies $\|g(t, y)\| \leq \|y\|$. This type of problems arise in different scientific areas such as celestial mechanics, nuclear physics, quantum chemistry, electronics, and so on. In this talk, we investigate the construction of an adapted explicit four-step method adapted to the numerical integration of perturbed oscillators. The new derived method has simple structure, has algebraic order six, and can integrate the harmonic oscillator without truncation error.

 ISBN: 978-84-608-6082-2

2 Construction of an adapted explicit four-step method

The variation-of-constant formula for (1) reads

$$y(t_n+h)=y(t_n)\cos(v)+hy'(t_n)\frac{\sin(v)}{v}+h^2\int_0^1\hat{g}(t_n+hz)\frac{\sin(v(1-z))}{v}\mathrm{d}z, \tag{2}$$

where $v=wh$ and $\hat{g}(t)=g(t,y(t))$. Suppose that $\hat{g}(t)$ is smooth enough we can write

$$\hat{g}(t_n+hz)=\sum_{j=0}^{\infty}h^j\hat{g}^{(j)}(t_n)\frac{z^j}{j!}. \tag{3}$$

Substituting the equation in (2), the exact solution of (1) can be expressed in terms of the ϕ-functions

$$y(t_n+h)=y(t_n)\phi_0(v)+hy'(t_n)\phi_1(v)+\sum_{j=0}^{\infty}h^{j+2}\hat{g}^{(j)}(t_n)\phi_{j+2}(v). \tag{4}$$

where

$$\phi_0(v)=\cos(v),\phi_1(v)=\frac{\sin(v)}{v},\phi_{j+2}(v)=\int_0^1\frac{\sin(v(1-z))}{v}\frac{z^j}{j!}\mathrm{d}x,j\geqslant 0.$$

For interesting properties of the ϕ-functions, the reader is referred to You et al. [12] and Gonzlez et al. [5] .

Keeping in mind that $\phi_j(v)$ is an even function [10], we substitute the step size h with $2h$ and $-2h$ respectively in (4) and get

$$\begin{aligned}y(t_n+2h)&=y(t_n)\phi_0(2v)+2hy'(t_n)\phi_1(2v)+\sum_{j=0}^{\infty}(2h)^{j+2}\hat{g}^{(j)}(t_n)\phi_{j+2}(2v),\\ y(t_n-2h)&=y(t_n)\phi_0(2v)-2hy'(t_n)\phi_1(2v)+\sum_{j=0}^{\infty}(-2h)^{j+2}\hat{g}^{(j)}(t_n)\phi_{j+2}(2v).\end{aligned} \tag{5}$$

Adding the two equations in (5), we obtain

$$y(t_n+2h)+y(t_n-2h)=2y(t_n)\phi_0(2v)+2\sum_{j=0}^{\infty}(2h)^{2j+2}\hat{g}^{(2j)}(t_n)\phi_{2j+2}(2v). \tag{6}$$

The derivatives $\hat{g}^{(2)}(t_n),\hat{g}^{(4)}(t_n)$ in (6) can be approximated by the finite difference formulae

$$\begin{aligned}\hat{g}^{(2)}(t_n)&=\frac{-\hat{g}(t_n-2h)+16\hat{g}(t_n-h)+16\hat{g}(t_n+h)-\hat{g}(t_n+2h)-30\hat{g}(t_n)}{12h^2}+\mathcal{O}(h^4),\\ \hat{g}^{(4)}(t_n)&=\frac{\hat{g}(t_n-2h)-4\hat{g}(t_n-h)-4\hat{g}(h+t_n)+\hat{g}(2h+t_n)+6\hat{g}(t_n)}{h^4}+\mathcal{O}(h^2).\end{aligned}$$

 ISBN: 978-84-608-6082-2

Thus we have the following express

$$\begin{aligned} & y\left(t_n+2h\right)+y\left(t_n-2h\right)-2y(t_n)\phi_0(2v) \\ & = h^2\left(\left(128\phi_6(2v)-\tfrac{8}{3}\phi_4(2v)\right)\left(\hat{g}(t_n-h)+\hat{g}(t_n+h)\right)\right. \\ & \quad +\left(\tfrac{128}{3}\phi_4(2v)-512\phi_6(2v)\right)\left(\hat{g}(t_n-2h)+\hat{g}(t_n+2h)\right) \\ & \quad \left.+\left(768\phi_6(2v)+8\phi_2(2v)-80\phi_4(2v)\right)\hat{g}(t_n)\right)+\mathcal{O}(h^8). \end{aligned} \tag{7}$$

Based on the equation (7), we propose a symmetric four-step sixth-order method for the perturbed oscillator (1) in the following form

$$y_{n+2}+a_1y_n+y_{n-2}=h^2\left(b_1\left(g_{n+2}+g_{n-2}\right)+b_2\left(g_{n+1}+g_{n-1}\right)+b_3g_n\right) \tag{8}$$

where

$$\begin{aligned} a_1 &= -2\phi_0(2v), & b_1 &= 128\phi_6(2v)-\tfrac{8}{3}\phi_4(2v), \\ b_2 &= \tfrac{128}{3}\phi_4(2v)-512\phi_6(2v), & b_3 &= 768\phi_6(2v)+8\phi_2(2v)-80\phi_4(2v). \end{aligned}$$

For small values of $|v|$ the above formulae are subjected to heavy cancellations, then the following Taylor series expansions should be used:

$$\begin{aligned} a_1 &= -2+4v^2-\frac{4v^4}{3}+\frac{8v^6}{45}-\frac{4v^8}{315}+\frac{8v^{10}}{14175}-\frac{8v^{12}}{467775}+\cdots, \\ b_1 &= \frac{1}{15}+\frac{2v^2}{945}-\frac{v^4}{2025}+\frac{2v^6}{66825}-\frac{134v^8}{127702575}+\frac{16v^{10}}{638512875}+\cdots, \\ b_2 &= \frac{16}{15}-\frac{176v^2}{945}+\frac{208v^4}{14175}-\frac{64v^6}{93555}+\frac{544v^8}{25540515}-\frac{304v^{10}}{638512875}+\cdots, \\ b_3 &= \frac{26}{15}-\frac{304v^2}{315}+\frac{706v^4}{4725}-\frac{592v^6}{51975}+\frac{892v^8}{1702701}-\frac{3448v^{10}}{212837625}+\cdots. \end{aligned}$$

It is observed that when $v\to 0$, this adapted method reduces to the classical Störmer-Cowell implicit four-step method [6].

The scheme (8) is implicit. At each step a nonlinear algebraic system must be solved by, for example, the Newton-Raphson iteration in order to find y_{n+2} so that the computational cost is expensive. As a recipe, we use a explicit symmetric four-step Störmer-Cowell method [6] as a predictor for an approximation of y_{n+2} and then correct the result with (8). Thus an adapted explicit four-step method is obtained in the following form

$$\begin{aligned} \bar{y}_{n+2} &= h^2\left(\frac{7}{6}(g_{n+1}-w^2y_{n+1})-\frac{1}{3}(g_n-w^2y_n)+\frac{7}{6}(g_{n-1}-w^2y_{n-1})\right), \\ y_{n+2}+a_1y_n+y_{n-2} &= h^2\left(b_1\left(\bar{g}_{n+2}+g_{n-2}\right)+b_2\left(g_{n+1}+g_{n-1}\right)+b_3g_n\right), \end{aligned} \tag{9}$$

where $\bar{g}_{n+2}=g(t_{n+2},\bar{y}_{n+2})$. The algebraic order of the adapted explicit four-step method (9) is of sixth and the principal term of the local truncation error is given by

$$LTE=\frac{1}{37800}h^8\left(y^{(6)}\left(189g_y-80w^2\right)-80y^{(8)}\right)$$

 ISBN: 978-84-608-6082-2

The new explicit method defined by (9) is denoted as AFS. It is observable that the adapted explicit method is capable of integrating exactly the harmonic oscillator. i.e. the equation(1) with $g(t, y) = 0$).

3 Numerical illustrations

In order to show the accuracy and efficiency of the new adapted method derived above, two model problems are considered. Five two-step or four-step methods are selected from the literature for comparison. The efficiency is measured by the decimal logarithm of the maximum global error ($\text{Log}_{10}|\text{ERR}|$) vesus the computational effort measured by the number of function evaluations (NEF) required by each method. The methods used in the numerical comparison are listed as follows:

- FS: Classical explicit four-step method with the free parameter $c_1 = -1/10$ based on [1].
- SFS: Explicit four-step method with the phase-lag and its first and second derivatives vanished developed in [8].
- SFSH: Explicit hybrid four-step method with the phase-lag and its first, second and third derivatives vanished developed in [9].
- TS: Classical explicit two-step method proposed in [2].
- ATS: Adapted explicit two-step method developed in [10].
- AFS: The new adapted explicit four-step method (9) derived in this paper.

Problem 1. We consider "almost" periodic problem

$$y''(x) + 169y(x) = (480 - 160i)e^{3ix}, \quad y(0) = 4 + i, \quad y'(0) = -23 + 22i,$$

or equivalently by

$$\begin{aligned} u''(x) + 169u(x) &= 480\cos(3x) + 160\sin(3x), \\ v''(x) + 169v(x) &= 480\sin(3x) - 160\cos(3x), \\ u(0) = 4, u'(0) &= -23, v(0) = 1, u'(0) = 22. \end{aligned}$$

The analytic solution is

$$y(x) = u(x) + \mathrm{i}v(x),$$

with

$$\begin{aligned} u(x) &= 3\cos(3x) + \cos(13) + \sin(3x) - 2\sin(13x), \\ u(x) &= -\cos(3x) + 2\cos(13) + 3\sin(3x) + \sin(13x). \end{aligned}$$

 ISBN: 978-84-608-6082-2

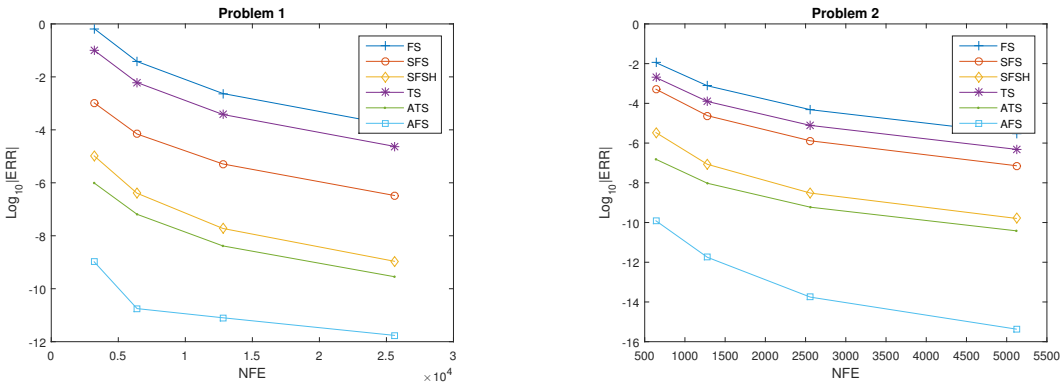


Figure 1: Efficiency comparison

In our test we choose $w = 13$, and the equation has been integrated in the interval $[0, 100]$.
Problem 2. A linear test problem was studied by Franco [3]

$$y'' + w^2 y = (w^2 - 4t^2)\cos(t^2) - 2\sin(t^2), \quad y(0) = 1, \quad y'(0) = w,$$

the analytic solution is given by

$$y(t) = \sin(wt) + \cos(t^2).$$

In our test we choose the integration interval $[0, 5]$, and the parameter (also is the estimated frequency) $w = 50$.

Figure 1 depicts the efficiency curves for the above two problems. We see that the new adapted method AFS outperforms the other five methods.

4 Conclusions

In this paper, an adapted explicit hybrid four-step method based on the algorithm of Scheifele which is derived by refining the classical Taylor expansion. The good property of Scheifele's based methods is that they can integrate exactly unperturbed oscillator, we give the principal local truncation error (PLTE) of the new adapted method which also confirms this fact. The analysis of phase-lag and linear stability property is made and discussed. The numerical experiments show that the new adapted method is much more accurate and efficient than some well-known methods proposed in the scientific literature. Finally, we should note that the new adapted method depend on the fitted frequency, a good estimate of the dominant frequency is needed for the adapted method to be applied effectively. For the choice of the fitting frequency, the reader is referred to the papers [11, 7].

Acknowledgements

This research was partially supported by the National Natural Science Foundation of China (NSFC) under Grant No. 11571302 and No. 11171155.

References

[1] Z.A. Anastassi, T. E. Simos, A parametric symmetric linear four-step method for the efficient integration of the Schrdinger equation and related oscillatory problems, J. Comput. Appl. Math. 236 (2012) 3880-3889.

[2] M.M. Chawla Numerov made explicit has better stability, BIT Numer. Math. 24 (1984) 117-118.

[3] J.M. Franco, Runge-Kutta methods adapted to the numerical integration of oscillatory problems, Appl. Numer. Math. 50 (2004) 427-443.

[4] A.B. Gonzlez, P. Martn, J.M. Farto, A new family of RungeKutta type methods for the numerical integration of perturbed oscillators, Numer. Math. 82 (1999) 635-646.

[5] J.D. Lambert, Computational methods in ordinary differential equations, John Wiley & Sons Inc, New York, 1973.

[6] H. Ramos, J. Vigo-Aguiar, On the frequency choice in trigonometrically fitted methods. Appl. Math. Lett. 23(11) (2010) 1378-1381.

[7] T.E. Simos, An explicit four-step method with vanished phase-lag and its first and second derivatives, J. Math. Chem. 52 (2014) 833-855.

[8] T.E. Simos, A new explicit hybrid four-step method with vanished phase-lag and its derivatives, J. Math. Chem. 52 (52) 1690-1716.

[9] H. Van de Vyver, An adapted explicit hybrid method of Numerov-type for the numerical integration of perturbed oscillators, Appl. Math. Comput. 186 (2007) 1385-1394.

[10] J. Vigo-Aguiar, T.E. Simos, JM. Ferrandiz, Controlling the error growth in long-term numerical integration of perturbed oscillations in one or several frequencies, Proc. R. Soc. Lond. Seri. A-Math. Phys. Eng. Sci. 460 (2004) 561-567.

[11] X. You, Y. Fang, J. Zhao, Special extended Nystrm tree theory for ERKN methods, J. Comput. Appl. Math. 263 (2014) 478499

 ISBN: 978-84-608-6082-2

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

Two-Derivative Runge-Kutta Method with increased phase-lag and dissipation order for the Schödinger equation

Ping Wang[1] **and Yonglei Fang**[1]

[1] *School of Mathematics and Statistics, Zaozhuang University*

emails: happywpnn@sohu.com, ylfangmath@163.com

Abstract

Optimized explicit two-derivative Runge-Kutta (TDRK) method with increased phase-lag and dissipation order for the numerical integration of the Schrödinger equation is constructed in this paper. The methods is of order five with increased order of phase and dissipation. Numerical results are reported to show the efficiency of the new method.

Key words: TDRK method, Schrödinger equation, Phase-lag, Dissipation.
MSC 2000: 65L05

1 Introduction

In this paper, we are concerned with the numerical integration of the Schrödinger equation of the form

$$y''(x) = \big(V(x) - E\big)y(x), \tag{1}$$

where the real number E is the *energy* and the function $V(x)$ is the *effective potential* satisfying $V(x) \to 0$ as $x \to \infty$. Regarding the oscillatory feature of the problem (1), researchers have proposed many numerical methods with frequency-dependent coefficients by some techniques such as trigonometrical/exponential fitting.

In [1], Kosti et al. constructed optimized constant coefficients Runge-Kutta method with increased phase-lag order for the numerical integration of the Schrödinger equation.

In this paper we construct an optimized explicit TDRK method with optimized phase-lag and dissipation for the numerical integration of Schrödinger equation based on the TDRK methods in [2].

 ISBN: 978-84-617-8694-7

2 Basic Theory

2.1 Two-derivative Runge-Kutta methods

For the numerical integration of (1), we consider the following explicit TDRK methods in [2] of the form,

$$\begin{cases} Y_1 = y_n, \\ Y_k = y_n + c_k h f(x_n, y_n) + h^2 \sum_{j=1}^{k-1} a_{kj} g(x_n + c_j h, Y_j), \ k = 2, \ldots, s, \\ y_{n+1} = y_n + h f(x_n, y_n) + h^2 \sum_{k=1}^{s} b_k g(x_n + c_k h, Y_k), \end{cases} \tag{2}$$

in which $g(x, y) := y''(x) = \frac{\partial}{\partial x} f(x, y) + \frac{\partial}{\partial y} f(x, y) \cdot f(x, y)$. The method (2) can be expressed briefly by the Butcher tableau

$$\begin{array}{c|c} c & A \\ \hline & b^T \end{array} = \begin{array}{c|cccc} 0 & & & & \\ c_2 & a_{21} & & & \\ \vdots & \vdots & \ddots & & \\ c_s & a_{s1} & \ldots & a_{ss-1} & \\ \hline & b_1 & \cdots & b_{s-1} & b_s \end{array}$$

or simply by (c, A, b). As the scheme (2) indicates, at each step, this special explicit TDRK method involves only one evaluation of the function f and s evaluations of the function g. The order conditions for the general-purpose TDRK method (2) are given in Chan et. al in [2].

2.2 Phase-lag property of the TDRK methods

For the purpose of phase analysis, we consider the following test equation

$$y' = \mathrm{i}\omega y, \quad \mathrm{i}^2 = -1, \tag{3}$$

in which $\omega > 0$ is the frequency of the problem. Applying the TDRK method (2) to (3) yields

$$y_{n+1} = M(\nu) y_n, \quad \nu = \omega h, \tag{4}$$

where $M(\nu)$ is called the stability function.

Definition 2.1 *(see [3]) For the TDRK method (2) with stability function $M(\nu)$, the quantities*

$$PL(\nu) = \nu - \arg(M(\nu)), \quad DIS(\nu) = 1 - |M(\nu)| \tag{5}$$

 ISBN: 978-84-617-8694-7

are called phase-lag and dissipation. If

$$PL(\nu) = c_\phi \nu^{q+1} + \mathcal{O}(\nu^{q+3}), DIS(\nu) = c_d \nu^{p+1} + \mathcal{O}(\nu^{p+3}),$$

the method is said to have phase-lag order q and dissipation order p, respectively.

Denoting $M(\nu) = U(\nu) + \mathrm{i}V(\nu)$ with $U(\nu),\ V(\nu)$ the real and imaginary parts of $M(\nu)$, we have

$$U(\nu) = 1 - \nu^2 b^T (I + \nu^2 A)^{-1} e, \quad V(\nu) = \nu\big(1 - \nu^2 b^T (I + \nu^2 A)^{-1} c\big).$$

The phase-lag and the dissipation become

$$PL(\nu) = \nu - \arctan\left(\frac{V(\nu)}{U(\nu)}\right), \quad DIS(\nu) = 1 - \sqrt{V^2(\nu) + U^2(\nu)} \tag{6}$$

Theorem 2.1 *([4]) For the TDRK method given by (2), we have the following formula for the direct calculation of the phase-lag order p and the phase-lag constant c_ϕ*

$$\tan(\nu) - \frac{V(\nu)}{U(\nu)}) = c_\phi \nu^{p+1}.$$

3 Construction of the new method

In this section, we shall give a new kind of TDRK method with increased phase-lag order and dissipation order. We consider the four stages explicit TDRK method which is simplify denoted by the Butcher tableau

$$\begin{array}{c|cccc} 0 & & & & \\ c_2 & a_{21} & & & \\ c_3 & a_{31} & a_{32} & & \\ c_4 & a_{41} & a_{42} & a_{43} & \\ \hline & b_1 & b_2 & b_3 & b_4 \end{array} \tag{7}$$

Together with the simplified assumptions $Ae = c^2/2$. Applying this method to the test equation (3), we have

$$y_{n+1} = (U(\nu) + \mathrm{i}V(\nu)) y_n, \quad \nu = \omega h, \tag{8}$$

in which

$$U(\nu) = 1 - \nu^2 s_1 + \nu^4 s_2 - \nu^6 s_3 + \nu^8 s_4,$$

$$V(\nu) = \nu - \nu^3 u_1 + \nu^5 u_2 - \nu^7 u_3,$$

 ISBN: 978-84-617-8694-7

with $s_k = b^T.A^{k-1}.e(k = 1, 2, 3, 4), u_j = b^T.A^{j-1}.c(j = 1, 2, 3)$. We have the following Taylor series

$$\begin{aligned} \frac{V(\nu)}{U(\nu)} &= \phi_3\nu^3 + \phi_5\nu^5 + \phi_7\nu^7 + \phi_9\nu^9 + \phi_{11}\nu^{11} + \cdots, \\ DIS(\nu) &= d_2\nu^2 + d_4\nu^4 + d_6\nu^6 + d_8\nu^8 + \cdots, \end{aligned} \tag{9}$$

in which

$$\phi_3 = (u_1 - s_1), \phi_5 = -s_1^2 + s_2 + s_1u_1 + u_2,$$

$$\phi_7 = -s_1^2(s_1 + u_1) - s_3 - s_2u_1 - s_1(u_2 - 2s_2) + u_3,$$

$$\phi_9 = s_1^2(3s_2 - u_2 + u_1s_1 - s_1^2) + s_2(u_2 - s_2) + s_4 + s_3u_1 + s_1(u_3 - 2s_3 - 2s_2u_1),$$

$$\phi_{11} = s_1^3(4s_2 - u_2 + s_1u_1 - s_1^2) + 2s_2s_3 + s_2u_1 - s_4u_1 - s_3u_2 - s_2u_3$$
$$+s_1(2s_2u_2 + 2(s_4s_3u_1) - 3s_2^2) + s_1^2(u_3 - 3s_3 - 3s_2u_1),$$

$$d_2 = (2s_1 - 1)/2, d_4 = (1 - 4s_1 - 8s_2 + 8u_1)/8,$$

$$d_6 = (8(s_2 - s_1 - u_1 + 2s_3 + 2s_1u_1 - 2u_2) - 1 + 6s_1)/16,$$

$$d_8 = (16(3u_1 - s_2(3 + 8u_1)) - 40s_1 - 64(s_1^3 + s_3 + 3s_1u_1 + s_1u_1^2 + u_2)$$
$$+32(3u_1^2 + 3s_1^2) + 128(s_1^2u_1 - s_4 + s_1s_2 - s_1u_2 + u_1u_2 + u_3) - 5)/128.$$

We solve the fifth order equations and get

$$a_{32} = (2a_{43}c_3(3 - 5(c_2 + c_3 - 2c_2c_3)) + 2a_{42}c_2(3 - 5(c_3 - c_2(2c_3 - 1)))$$
$$+c_4(c_2-_4)(c_4 - c_3))(c_2 - c_3)c_3/(2c_2(c_2 - c_4)c_4(3 - 5c_4 + 5c_2(2c_4 - 1))),$$

$$b_4 = \frac{3 - 5c_3 + 5c_2(2c_3 - 1)}{60c_4(c_4 - c_2)(c_4 - c_3)}, \quad b_3 = \frac{-3 + 5c_4 + c_2(5 - 10c_4)}{60(c_2 - c_3)c_3(c_3 - c_4)}, a_{21} = c_2^2/2,$$

$$b_2 = \frac{3 - 5c_4 + 5c_3(2c_4 - 1)}{60c_2(c_2 - c_3)(c_2 - c_4)}, b_1 = \frac{1}{2} - b_2 - b_3 - b_4,$$

$$a_{31} = c_3^2/2 - a_{32}, a_{41} = c_4^2/2 - a_{42} - a_{43}.$$

There are five parameters in the new five order TDRK method, the choice of the free paremeters is very essential for the oscillatory systems. We select optimized parameters so that the dispersion and the dissipation are minimal.

 ISBN: 978-84-617-8694-7

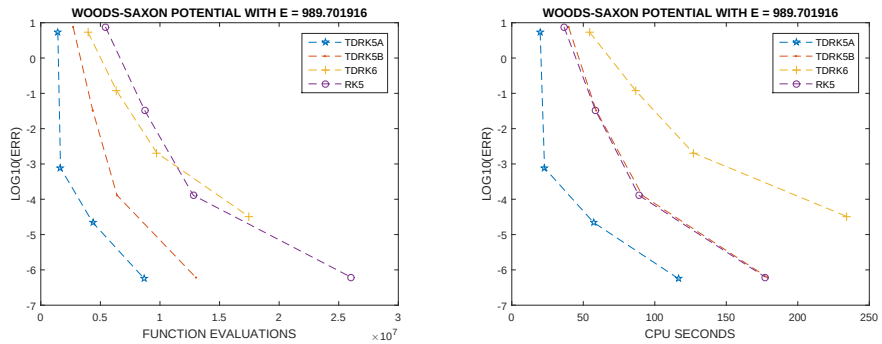


Figure 1: Efficiency curves for $E = 989.701916$

Case(I): Selecting the parameters such that the dispersion to be order twelve. Solving $\phi_7 = \frac{17}{315}, \phi_9 = \frac{62}{2835}, \phi_{11} = \frac{1382}{155925}$ yields $c_2 = \frac{1}{7}$ and

$$\begin{aligned} a_{43} &= \frac{127(c_3 - c_4)c_4(7c_4 - 1)}{198c_3(16 - 137c_3 + 175c_3^2)}, \\ a_{42} &= \frac{7c_4(7c_4 - 1)(6336 + 88011c_3^2 + 18802c_4 - c_3(73054 + 18711c_4))}{25146(7c_3 - 1)(25c_3 - 16)}. \end{aligned} \tag{10}$$

The proposed TDRK method with free parameter $c_3,\ c_4$. We try to get an optimal method minimizing the error constant

$$C^{(6)} = \left((\tau_1^{(6)})^2 + (\tau_2^{(6)})^2 + (\tau_3^{(6)})^2\right)^{1/2} \tag{11}$$

in which $\tau_i^{(6)}$ are the error constants of a five-order TDRK method. We minimize the errr constant (11) and obtain $c_3 = \frac{101}{254}, c_4 = \frac{62}{81}$. The method is of phase-lag order twelve and dissipation order five. We denote this method as TDRK5-12-5 and the method is list as

$$\begin{array}{c|cccc} 0 & & & & \\ \frac{1}{7} & \frac{1}{98} & & & \\ \frac{101}{254} & \frac{14443}{16387064} & \frac{320271}{4096766} & & \\ \frac{62}{81} & \frac{827410679635}{8178059102301} & \frac{316504531259}{36679809637053} & \frac{678470997055292}{3704660773342353} & \\ \hline & \frac{6187}{125240} & \frac{1611071}{9594540} & \frac{1040578564}{5193194265} & \frac{272629233}{3312227240} \end{array} \tag{12}$$

4 Numerical Illustrations

In this section, we shall examine the numerical performance of the new method. The methods we choose for comparison are as follows:

- TDRK5A: the optimized sixth order TDRK method derived in this paper. .
- TDRK6: the sixth order TDRK method with $c_4 = 1$ given in [2].
- TDRK5B: the fifth order TDRK method derived in [5].
- RK5: the optimized fifth order RK method derived by Kosti in [1].

We consider the numerical integration of the Schrödinger equation (1) with the well-known Woods-Saxon potential of the form in [3]

$$V(x) = c_0 z\big(1 - a(1-z)\big),$$

where $z = \Big(\exp\big(a(x-b)+1\big)\Big)^{-1}$, $c_0 = -50,\ a = 5/3,\ b = 7$. We solve the problem on the interval $[0, 15]$.

In the numerical experiments, we consider the so-called resonant-state problem $E > 0$, that is to find the energies (or resonances) $E \in [0, 1000]$ for which the phase shift is equal to $\frac{\pi}{2}$. The boundary conditions for this problem are

$$y(0) = 0 \quad \text{and} \quad y(x) = \cos(\sqrt{E}x) \quad \text{for large } x.$$

In Fig. 1, we plot the logarithm of error $|E_{\text{analytical}} - E_{\text{calculated}}|$ (LOG(ERR)) versus the computational effort by the number of function evaluations (FUNCTION EVALUATIONS) and the cpu times (CPU SECONDS) required by each methods for $E_{\text{analytical}} = 989.701916$. In view of Fig.1, we observe that the new method TDRKA shows more advantage over the selected methods.

Acknowledgement

This research was partially supported by the National Natural Science Foundation of China (NSFC) under Grant No. 11571302.

References

[1] A. A. Kosti, Z. A. Anastassi,T. E. Simos,*An optimized explicit Runge-Kutta method with increased phase-lag order for the numerical solution of the Schrödinger equation and related problems*,J. Math. Chem, **47** (2010) 315–330.

[2] R.P.K. Chan, A.Y.J. Tasi, *On explicit two-derivative Runge-Kutta methods*, Numer. Algor. **53** (2010) 171–194.

 ISBN: 978-84-617-8694-7

[3] H. Van de Vyver, *Stability and phase-lag analysis of explicit Runge-Kutta methods with variable coefficients for oscillatory problems*, Comput. Phys. Comm. **173** (2005) 115–130 .

[4] T.E. Simos, *RungeCKutta interpolants with minimal phase-lag*, Comput. Math. Appl. **26** (1993) 43–49

[5] Y.P. Yang, P. Wang, Y.L. Fang,*A two derivative Runge-Kutta method with increased phase-lag order for the Schrödinger equation*,ICIC Exp. Lette. B, **7** (2016) 1835–1841.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

Modified Two-Derivative Runge-Kutta Methods for the Schrödinger Equation

Yanping Yang[1], Qinghe Ming[1] and Yonglei Fang[1]

[1] *School of Mathematics and Statistics, Zaozhuang University*

emails: ypyang0703@163.com, mqh9015@163.com, ylfangmath@163.com

Abstract

A family of modified two-derivative Runge-Kutta (MTDRK) methods for the integration of Schrödinger equation are obtained. These methods are of algebraically order five. The numerical results in the integration of the radial Schrödinger equation with the Woods-Saxon potential are reported to show the high efficiency of our new methods. The result of error analysis is confirmed by the resonance problem.

Key words: Modified TDRK methods, Dispersion, Error analysis
MSC 2000: AMS codes 65L05

1 Introduction

In this paper we are focused on the numerical integration of one-dimensional Schrödinger equation of the form

$$y''(x) = \big(W(x) - E\big)y(x), \tag{1}$$

where the real number E is the *energy* and the function $W(x)$ is the *effective potential* satisfying $W(x) \to 0$ as $x \to \infty$. Two boundary conditions are associated with this equation: one is $y(0) = 0$ and the other imposed at large x is determined by physical considerations. The form of this second boundary condition depends crucially on the sign of the energy E.

Inspired by the ideas of Van de Vyver [1, 2], the purpose of this paper is to construct practical optimized modified TDRK methods for the numerical integration of the radial Schrödinger equation (1).

 ISBN: 978-84-617-8694-7

2 Modified TDRK methods and order conditions

We begin with consideration of an initial value problem of systems of first-order differential equations in the form

$$y'(x) = f(x, y), \quad y(x_0) = y_0, \tag{2}$$

whose solution has an oscillatory character, where $y \in \mathbb{R}^d$, $f : \mathbb{R} \times \mathbb{R}^d \to \mathbb{R}^d$ is a smooth function. Regarding the oscillatory property of the solution of the problem (2), we consider the following special form of explicit *modified TDRK method* [3]

$$\begin{cases} Y_1 = y_n, \\ Y_i = y_n + c_i h f(x_n, y_n) + h^2 \sum\limits_{j=1}^{i-1} a_{ij} g(x_n + c_j h, Y_j), \quad i = 2, \cdots, s, \\ y_{n+1} = y_n + h\beta(\nu) f(x_n, y_n) + h^2 \sum\limits_{i=1}^{s} b_i(\nu) g(x_n + c_i h, Y_i), \end{cases} \tag{3}$$

where c_i, a_{ij} $(1 \le j < i \le s)$ are real constants, $\beta(\nu)$ and $b_i(\nu)$ $(1 \le i \le s)$ are real even functions of $\nu = h\omega$ with ω an estimate of the principal frequency of the problem.

The scheme (3) can also be expressed compactly by the Butcher tableau

$$\begin{array}{c|c} c & A \\ \hline \beta(\nu) & b^T(\nu) \end{array} = \begin{array}{c|cccc} 0 & & & & \\ c_2 & a_{21} & & & \\ \vdots & \vdots & \ddots & & \\ c_s & a_{s1} & \cdots & a_{ss-1} & \\ \hline \beta(\nu) & b_1(\nu) & \cdots & b_{s-1}(\nu) & b_s(\nu) \end{array}$$

or simply by $(c, A, \beta(\nu), b(\nu))$. We assume that as $\beta(\nu) \to 1$ as $\omega \to 0$ so that when $\omega \to 0$ the modified RK method (3) reduces to a traditional two-derivative RK method with constant coefficients (see [3]).

3 Construction of new modified TDRK methods

Our aim is to determine the b-values and β-value in the scheme (3). Following the approach in [1, 2], we apply the method to the linear scalar equation

$$y' = \mathrm{i}\omega y, \quad y(x_0) = y_0, \quad \omega \in \mathbb{R}^+, \quad \mathrm{i}^2 = -1 \tag{4}$$

and obtain the recursive relation

$$y_1 = R(\nu) y_0, \tag{5}$$

where

$$R(\nu) = 1 + \mathrm{i}\nu\beta(\nu) + b(\nu)^T (I_s + \nu^2 A)^{-1} (e + \mathrm{i}\nu c)$$

 ISBN: 978-84-617-8694-7

is called the *stability function*. For the explicit scheme (3)

$$R(\nu) = U(\nu) + \mathrm{i}V(\nu)$$

where

$$\begin{aligned} U(\nu) &= 1 + b(\nu)^T(I_s - \nu^2 A + \nu^4 A^2 - \ldots + (-1)^{s-1}\nu^{2(s-1)}A^{s-1})e, \\ V(\nu) &= \nu\beta(\nu) + \nu b(\nu)^T(I_s - \nu^2 A + \nu^4 A^2 - \ldots + (-1)^{s-1}\nu^{2(s-1)}A^{s-1})c. \end{aligned} \tag{6}$$

In the sequel we consider a three-stage explicit modified TDRK given by

$$c_2 = \frac{3}{10}, a_{21} = \frac{9}{200}, c_3 = \frac{3}{4}, a_{31} = 0, a_{42} = \frac{9}{32}.$$

For this method,

$$R(\nu) = 1 + \sum_{j=1}^{6} r_j(\mathrm{i}\nu)^j, \tag{7}$$

where

$$r_1 = \beta(\nu), \quad r_2 = \sum_i b_i, \quad r_3 = \sum_i b_i c_i, \quad r_4 = \sum_{i,j} b_i a_{ij}, \quad r_5 = b_3 a_{32} c_2, \quad r_6 = b_3 a_{32} a_{21}.$$

Then the fitting condition becomes

$$U(\nu) = 1 - r_2\nu^2 + r_4\nu^4 - r^6\nu^6 = \cos(\nu), \quad V(\nu) = r_1\nu - r_3\nu^3 + r_5\nu^5 = \sin(\nu). \tag{8}$$

We choose the values A and c from the modified TDRK (3) and we get two linear equations in four unknowns $b_i(\nu),\ i = 1, 2, 3$ and $\beta(\nu)$. For the determination of these unknown parameters, we consider two different choices.

3.1 The first optimized modified TDRK method

Together with the order conditions

$$b_1(\nu) + b_2(\nu) + b_3(\nu) = \frac{1}{2}, \quad b_2(\nu)c_2 + b_3(\nu)c_3 = \frac{1}{6}, \tag{9}$$

we solve equations (8) and derive the following solution

$$\begin{aligned} \beta(\nu) &= \tfrac{10\nu^3 - 60\nu(1-\cos(\nu)) + \sin(\nu)(9\nu^2-120)}{9\nu^3 - 120\nu}, \\ b_1(\nu) &= \tfrac{6400 - 3200\nu^2 + 200\nu^4 - 3\nu^6 - 6400\cos(\nu)}{54\nu^6 - 720\nu^4}, \\ b_2(\nu) &= \tfrac{5(1600\nu^2 - 3200(1-\cos(\nu)) - 200\nu^4 + 9\nu^6)}{27\nu^4(3\nu^2-40)}, \\ b_3(\nu) &= \tfrac{160(40 - 20\nu^2 + \nu^4 - 40\cos(\nu))}{27\nu^4(3\nu^2-40)}. \end{aligned}$$

 ISBN: 978-84-617-8694-7

It is easy to check that

$$\beta(\nu) = 1 - \frac{\nu^6}{100800} + \mathcal{O}(\nu^8),\ b(\nu)^T c^2 = \frac{1}{12} - \frac{\nu^2}{3600} + \mathcal{O}(\nu^4),$$
$$b(\nu)^T c^3 = \frac{1}{20} - \frac{7\nu^2}{24000} + \mathcal{O}(\nu^4),\ b(\nu)^T Ac = \frac{1}{120} - \frac{\nu^2}{14400} + \mathcal{O}(\nu^4).$$

Therefore the new method is of order five and we denote this method as MTDRK5I.

3.2 The second optimized modified TDRK method

Another approach is to consider two linear equations of the form (4) with frequencies $\omega = \omega_1$ and $\omega = \omega_2$, respectively. Then substituting $\nu = \nu_1 = \omega_1 h$ and $\nu = \nu_2 = \omega_2 h$ in the equations (8), we obtain a system of four linear equations in the four parameters $b_i(\nu)$, $i = 1, 2, 3$ and $\beta(\nu)$. The solutions b_i, $i = 1, 2, 3$ and β contain two parameters ν_1 and ν_2. Taking the limits $\nu_1 \to \nu$ and $\nu_2 \to \nu$, we get

$$\beta(\nu) = \frac{40\nu - 80\nu\cos(\nu) + 3\nu^3\cos(\nu) + 120\sin(\nu) - 23\nu^2\sin(\nu)}{80\nu},$$
$$b_3(\nu) = \frac{4(40 + (3\nu^2 - 40)\cos(\nu) - 23\nu\sin(\nu))}{27\nu^4},$$
$$b_2(\nu) = \frac{40(9\nu^2 - 40) + \cos(\nu)(1600 - 660\nu^2 + 27\nu^4) + (1100\nu - 207\nu^3)\sin(\nu)}{108\nu^4},$$
$$b_1(\nu) = (40(3200 - 480\nu^2 + 27\nu^4) + (52800\nu^2 - 128000 - 4680\nu^4 + 81\nu^6)\cos(\nu)$$
$$-\nu(97600 - 20400\nu^2 + 621\nu^4))/(14400\nu^4).$$

Again it is easy to check that

$$\beta(\nu) = 1 + \frac{13\nu^6}{50400} + \mathcal{O}(\nu^8),\ b(\nu)^T e = \frac{1}{2} - \frac{\nu^4}{720} + \mathcal{O}(\nu^6),$$
$$b(\nu)^T c = \frac{1}{6} + \frac{\nu^4}{3150} + \mathcal{O}(\nu^6),\ b(\nu)^T c^2 = \frac{1}{12} - \frac{\nu^2}{1800} + \mathcal{O}(\nu^4),$$
$$b(\nu)^T c^3 = \frac{1}{20} - \frac{7\nu^2}{12000} + \mathcal{O}(\nu^4),\ b(\nu)^T Ac = \frac{1}{120} - \frac{\nu^2}{7200} + \mathcal{O}(\nu^4).$$

Therefore the new method is of order five and we denote this method as MTDRK5II.

4 Numerical results

In this section, we test the numerical performance of the new fifth-order modified TDRK methods in the integration of the radial Schrödinger equation with the Woods-Saxon potential. The methods we choose for comparison are as follows:

- EFTDRK4I: optimized TDRK method with order four derived by Fang et al. in [5]
- EFTDRK4II: optimized TDRK method with order four derived by Fang et al. in [5]
- RK5V: optimized RK method with order five derived by Van de Vyver in [6]

 ISBN: 978-84-617-8694-7

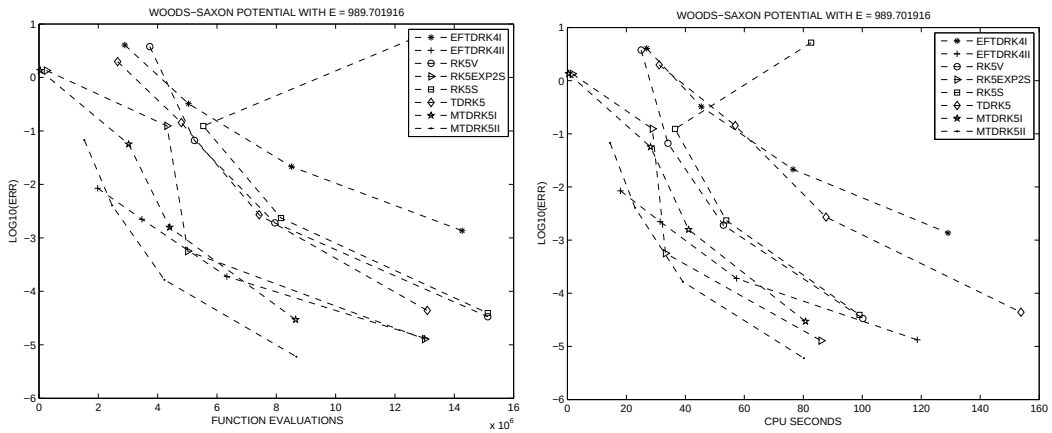


Figure 1: Efficiency curves for $E = 989.701916$

- RK5S: phase fitted RK method with order five presented by Simos in [7]
- RK5EXP2S: the exponentially fitted fifth order RK method with exponential order two presented by Simos in [4].
- TDRK5: the classical TDRK with order five in [3].
- MTDRK5I: modified optimized two-derivative fifth order RK method derived in Section 3.1.
- MTDRK5II: modified optimized two-derivative fifth order RK method derived in Section 3.2.

We consider the numerical integration of the Schrödinger equation (1) with the well-known Woods-Saxon potential

$$V(x) = c_0 z\big(1 - a(1-z)\big),$$

where $z = \Big(\exp\big(a(x-b)+1\big)\Big)^{-1}$, $c_0 = -50$, $a = 5/3$, $b = 7$. The problem is solved on the interval $[0, 15]$. In the numerical experiments, we consider the so-called resonant-state problem $E > 0$, that is, to find the energies (or resonances) $E \in [0, 1000]$ for which the phase shift is equal to $\frac{\pi}{2}$. The boundary conditions for this problem are

$$y(0) = 0 \quad \text{and} \quad y(x) = \cos(\sqrt{E}x) \quad \text{for large } x.$$

We follow the lines of [6, 8] and choose the fitted frequency

$$\omega = \begin{cases} \sqrt{50+E}, & x \in [0, 6.5], \\ \sqrt{E}, & x \in [6.5, 15]. \end{cases}$$

 ISBN: 978-84-617-8694-7

The numerical results $E_{\text{calculated}}$ are compared with the analytical solution $E_{\text{analytical}}$ of the Woods-Saxon potential, rounded to six decimal places. In Fig. 1, we plot the error $\log_{10}|E_{\text{analytical}} - E_{\text{calculated}}|$ versus the computational effort measured by the number of function evaluations (FUNCTION EVALUATIONS) and the CPU time (CPU SECONDS) required by each method for $E_{\text{analytical}}$=989.701916.

Acknowledgement

This research was partially supported by the National Natural Science Foundation of China (NSFC) under Grant No. 11571302.

References

[1] H. Van de Vyver, *Comparison of some special optimized fourth-order Runge-Kutta methods for the numerical solution of the Schrödinger equation*, Comput. Phys. Commun. **166** (2005) 109–122 .

[2] H. Van de Vyver, *Modified explicit Runge-Kutta methods for the numerical solution of the Schrödinger equation,* Comput. Phys. Commun. **171** (2005) 1025–1036 .

[3] R.P.K. Chan, A.Y.J. Tsai, *On explicit two-derivative Runge-Kutta methods*, Numer. Algor. **53** (2010) 171–194 .

[4] Z.A. Anastassi, T.E. Simos,*Trigonometrically fitted Runge-Kutta methods for the numerical solution of the Schrödinger equation,* J. Math. Chem. **37** (2005) 281–293 .

[5] Y.L. Fang, X. You, Q.H. Ming, *Exponentially fitted two derivative Runge-Kutta methods for the Schrödinger equation,* Int. J. Mod. Phys. C **24**, (2013) 1350073, 9 pages.

[6] H. Van de Vyver,*An embedded phase-fitted modified Runge-Kutta method for the numerical integration of the radial Schrödinger equation*, Phys. Lett. A **352** (2006) 278–285.

[7] T.E. Simos, J.V. Aguiar, *A modified phase-fitted Runge-Kutta method for the numerical solution of the Schrödinger equation,* J. Math. Chem. **30** (2001) 121–131 .

[8] L.Gr. Ixaru, M. Rizea, *A Numerov-like scheme for the numerical solution of the Schrödinger equation in the deep continuum spectrum of energies*, Comput. Phys. Comm. **19** (1980) 23–27.

 ISBN: 978-84-617-8694-7

Proceedings of the 17th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2017
4–8 July, 2017.

Trigonometrically-fitted multi-derivative methods for the Schrödinger equation

Yanwei Zhang[1] **and Yonglei Fang**[1]

[1] *School of Mathematics and Statistics, Zaozhuang University*

emails: zywmuke@163.com, ylfangmath@163.com

Abstract

A family of trigonometrically-fitted multi-derivative linear methods involving several derivatives for the numerical integration of the Schrödinger equation are obtained in this paper. Numerical results are reported to show the efficiency and robustness of the new method IV specially adapted to the integration of the radial time-independent Schrödinger equation for large energies.

Key words: Schrödinger equation, Obrechkoff method,Trigonometrically fitted.
MSC 2000: 65L05

1 Introduction

In this paper, we are interested in the numerical integration of one-dimensional Schrödinger equation of the form

$$y''(x) = \big(W(x) - E\big)y(x), \quad x \in [x_0, X], \tag{1}$$

where E is a real number representing the *energy*, the function $W(x)$ represents the *potential* with the property $W(x) \to 0$ as $x \to \infty$. In [1], by using trigonometric fitting, Wang derived a trigonometrically fitted Obrechkoff one-step method which is more accuracy and effective than the original method for the numerical solution of one-dimensional Schrödinger equation. This new method differs from the original Obrechkoff one-step method only in one simple coefficient. In this paper, we present a family of trigonometrically fitted one-step methods with multi-derivative of higher algebraic order than method in [1]. The numerical tests are carried out to show the efficiency of our new methods.

 ISBN: 978-84-617-8694-7

2 Multi-derivative linear methods with variable coefficients

Now we consider the one-step four-derivative linear methods of the Obrechkoff type in the form

$$\begin{aligned} y(x+h)-y(x) = hc_0\big(y'(x+h)+y'(x)\big) + h^2c_1\big(y''(x+h)-y''(x)\big) \\ +h^3c_2\big(y^{(3)}(x+h)+y^{(3)}(x)\big) + h^4c_3\big(y^{(4)}(x+h)+y^{(4)}(x)\big). \end{aligned} \tag{2}$$

If we choose $\{c_0,\ c_1,\ c_2,\ c_3\} = \{\frac{1}{2},\ -\frac{3}{28},\ \frac{1}{84},\ -\frac{1}{1680}\}$,we can derive Obrechkoff one step method of order seven which can be found in [2].In this section we will apply trigonometical fitting to the original one step method to construct more accuracy and efficient one step methods for the numerical solution of one-dimensional Schrodinger equation.

2.1 First new method

If we require the method (2) to integrate exactly the function $\exp(\mathrm{i}\omega x)$ ($\mathrm{i}^2=-1$), then we obtain

$$\cos(\frac{\nu}{2})(\nu^3c_2-\nu c_0)+\sin(\frac{\nu}{2})(1+c_1\nu^2-c_3\nu^4)=0, \quad \nu=\omega h. \tag{3}$$

With the choice $c_1=-\dfrac{3}{28}$ $c_2=\dfrac{1}{84}$ and $c_3=-\dfrac{1}{1680}$, solving the equation (3) yields

$$c_0=\frac{\nu^2}{84}+\left(\frac{1}{\nu}-\frac{3}{28}\nu^2+\frac{1}{1680}\nu^3\right)\tan(\frac{\nu}{2}). \tag{4}$$

We denote this method as method I. Its local truncation error is

$$LTE=-\frac{h^9}{25401600}\big(y^{(9)}(x)-\omega^8y'(x)\big)+\mathcal{O}(h^{10}).$$

2.2 Second new method

Now we set free the two coefficients $\{c_0,c_1\}$ and require that the method (2) to integrate exactly the functions

$$\{\exp(\mathrm{i}\omega x),\ x\exp(\mathrm{i}\omega x)\}, \tag{5}$$

leading to the following system of equations

$$\begin{cases} \cos(\frac{\nu}{2})(\nu^3c_2-\nu c_0)+\sin(\frac{\nu}{2})(1+c_1\nu^2-c_3\nu^4)=0, \\ \cos(\nu)(1-c_0+(c_1+3c_2)\nu^2-c_3\nu^4)+\sin(\nu)((c_0+2c_1)\nu-(c_2+4c_3)\nu^3)=-3c_2\nu^2+c_0. \end{cases} \tag{6}$$

 ISBN: 978-84-617-8694-7

Substituting$c_2 = \dfrac{1}{84}$ and $c_3 = -\dfrac{1}{1680}$ in (6) ,the following result emerges

$$\begin{aligned} c_0 &= \frac{1680 + 9\nu^4 - (1680 - \nu^4)\cos(\nu) - 10\nu^3\sin(\nu)}{840\nu(\nu + \sin(\nu))}, \\ c_1 &= -\frac{1680\nu + 40\nu^3 + \nu^5 + 40\nu^3\cos(\nu) - (1680 - 3\nu^4)\sin(\nu)}{1680\nu^2(\nu + \sin(\nu))}. \end{aligned} \tag{7}$$

We denote this method as method II. Its local truncation error is

$$LTE = -\frac{1}{25401600}\big(y^{(9)}(x) + 3\omega^8 y'(x) + 4\omega^6 y^{(3)}(x)\big)h^9 + \mathcal{O}(h^{10}).$$

2.3 Third new method

We free the three coefficients $\{c_0, c_1, c_2\}$ and keep the rest of c_3, then we demand the one-step method (2) to integrate exactly the functions

$$\{\exp(\mathrm{i}\omega x),\ x\exp(\mathrm{i}\omega x),\ x^2\exp(\mathrm{i}\omega x)\}.$$

We obtain the following equations

$$\begin{cases} \cos(\frac{\nu}{2})(\nu^3 c_2 - \nu c_0) + \sin(\frac{\nu}{2})(1 + c_1\nu^2 - c_3\nu^4) = 0, \\ \cos(\nu)(1 - c_0 + (c_1 + 3c_2)\nu^2 - c_3\nu^4) + 3c_2\nu^2 - c_0 \\ \quad + \sin(\nu)((c_0 + 2c_1)\nu - (c_2 + 4c_3)\nu^3) = 0, \\ \cos(\nu)(1 - 2c_0 - 2c_1 + (c_1 + 6c_2 + 12c_3)\nu^2 - c_3\nu^4) + 2c_1 - 12c_3\nu^2 \\ \quad + \sin(\nu)((c_0 + 4c_+6c_2)\nu - (c_2 + 8c_3)\nu^3) = 0. \end{cases} \tag{8}$$

Solving (8), we have

$$\begin{aligned} c_0 &= \frac{(-5040 - 3360\nu^2 - \nu^4 + 2\nu^6)\cos(\frac{\nu}{2}) + (5040 + \nu^4)\cos(\frac{3\nu}{2}) + \nu(16800\nu - 2\nu^5)\sin(\frac{\nu}{2})}{3360\nu(\nu\cos(\frac{\nu}{2}) - (1 - \nu^2 + \cos(\nu))\sin(\frac{\nu}{2}))}, \\ c_1 &= -\frac{(-5040\nu + 5\nu^5)\cos(\frac{\nu}{2}) + (1680 + \nu^4)(3 + \nu^2 + 3\cos(\nu))\sin(\frac{\nu}{2})}{1680\nu^2(\nu\cos(\frac{\nu}{2}) - (1 - \nu^2 + \cos(\nu))\sin(\frac{\nu}{2}))}, \\ c_2 &= \frac{(1680 - 3360\nu^2 + 3\nu^4 + 2\nu^6)\cos(\frac{\nu}{2}) - (1680 + 3\nu^4)(\cos(\frac{3\nu}{2}) - 2\nu\sin(\frac{\nu}{2}))}{3360\nu^3(\nu\cos(\frac{\nu}{2}) - (1 - \nu^2 + \cos(\nu))\sin(\frac{\nu}{2}))}. \end{aligned}$$

 ISBN: 978-84-617-8694-7

We denote this method as method III. Its local truncation error is

$$LTE = -\frac{1}{25401600}(y^{(9)}(x) - 8\omega^6 y^{(3)}(x) - 6\omega^4 y^{(5)}(x) - 3\omega^8 y'(x))h^9 + \mathcal{O}(h^{10}).$$

2.4 Fourth new method

Let $\{c_0, c_1, c_2, c_3\}$ are variable coefficients. We try to find the values of $\{c_0, c_1, c_2, c_3\}$ in such way that the one-step method (2) integrate exactly the functions

$$\{\exp(\mathrm{i}\omega x),\ x\exp(\mathrm{i}\omega x),\ x^2\exp(\mathrm{i}\omega x),\ x^3\exp(\mathrm{i}\omega x)\}.$$

We obtain the following equations

$$\begin{cases} \cos(\frac{\nu}{2})(\nu^3 c_2 - \nu c_0) + \sin(\frac{\nu}{2})(1 + c_1\nu^2 - c_3\nu^4) = 0, \\ \cos(\nu)(1 - c_0 + (c_1 + 3c_2)\nu^2 - c_3\nu^4) + \sin(\nu)((c_0 + 2c_1)\nu - (c_2 + 4c_3)\nu^3) + 3c_2\nu^2 - c_0 = 0, \\ \cos(x)(1 - 2c_0 - 2c_1 + (c_1 + 6c_2 + 12c_3)\nu^2 - c_3\nu^4) + \sin(\nu)((c_0 + 4c_+6c_2)\nu - (c_2 + 8c_3)\nu^3) \\ +2c_1 - 12c_3\nu^2 = 0, \\ \cos(\nu)(1 - 3c_0 - 6c_1 - 6c_2 + (c_1 + 9c_2 + 36c_3)\nu^2 - c_3\nu^4) + \sin(\nu)((c_0 + 6c_1 + 18c_2 + 24c_3)\nu \\ -(c_2 + 12c_3)\nu^3) - 6c_2 = 0. \end{cases} \tag{9}$$

Solving (9), we have

$$\begin{aligned} c_0 &= \frac{8(\nu(3 + 2\nu^2) + \cos(\nu)(-3\nu + \nu^3 - 3\sin(\nu)) + (3 - 6\nu^2)\sin(\nu))}{\nu(3 - 6\nu^2 + 2\nu^4 - 12\nu^2\cos(\nu) - 3\cos(2\nu) + 12\nu\sin(\nu) - 4\nu^3\sin(\nu))}, \\ c_1 &= -\frac{2(-9 + 18\nu^2 + 2\nu^4 + 24\nu^2\cos(\nu) + 9\cos(2\nu) - 24\nu\sin(\nu))}{\nu^2(3 - 6\nu^2 + 2\nu^4 - 12\nu^2\cos(\nu) - 3\cos(2\nu) + 12\nu\sin(\nu) - 4\nu^3\sin(\nu))}, \\ c_2 &= \frac{8(-3\nu + 2\nu^3 - 3\sin\nu + \cos(\nu)(3\nu + \nu^3 + 3\sin(\nu)))}{\nu^3(3 - 6\nu^2 + 2\nu^4 - 12\nu^2\cos(\nu) - 3\cos(2\nu) + 12\nu\sin(\nu) - 4\nu^3\sin(\nu))}, \\ c_3 &= \frac{-3 + 6\nu^2 - 2\nu^4 - 12\nu^2\cos(\nu) + 3\cos(2\nu) + 12\nu\sin(\nu) - 4\nu^3\sin(\nu)}{\nu^4(3 - 6\nu^2 + 2\nu^4 - 12\nu^2\cos(\nu) - 3\cos(2\nu) + 12\nu\sin(\nu) - 4\nu^3\sin(\nu))}. \end{aligned}$$

We denote this method as method IV. Its local truncation error is

$$LTE = -\frac{1}{25401600}(y^{(9)}(x) + 4\omega^2 y^{(7)}(x) + 6\omega^4 y^{(5)}(x) + 4\omega^6 y^{(3)}(x) + \omega^8 y'(x))h^9 + \mathcal{O}(h^{10}).$$

 ISBN: 978-84-617-8694-7

2.5 Formula for the first derivative

Because the Schrödinger equation (1) is linear, so all the high derivative $y^{(m)}$ at x can be expressed explicitly in terms of $y(x)$ and $y'(x)$ such as

$$y''(x) = f(x)y(x), \quad y^{(3)}(x) = f'(x)y(x) + f(x)y'(x),$$
$$y^{(4)}(x) = \big(f''(x) + f^2(x)\big)y(x) + 2f'(x)y'(x), \ldots$$

It is not sufficient for the problem (1) to have only formula (2). Therefore, we consider the following formula, which is obtain by differentiating (2) to x

$$\begin{aligned} y'(x+h) - y'(x) = hc_0\big(y''(x+h) + y''(x)\big) + h^2c_1\big(y^{(3)}(x+h) - y^{(3)}(x)\big) \\ +h^3c_2\big(y^{(4)}(x+h) + y^{(4)}(x)\big) + h^4c_3(y^{(5)}(x+h) - y^{(5)}(x)\big), \end{aligned} \tag{10}$$

where $c_0,\ c_1,\ c_2$ and c_3 are derived in this paper. Therefore, with the same ideas as in [1], we can compute the Schrödinger equation by combining (2) with (10).

3 Numerical experiment

In this section we carry out some numerical experiments to illustrate the performance of the new methods constructed in Section 2. We consider the numerical integration of the one-dimensional Schrödinger equation (1) with the Woods-Saxon potential

$$W(x) = c_0 z\big(1 - a(1-z)\big) \tag{11}$$

where $z = (\exp(a(x-b)+1))^{-1},\ c_0 = -50,\ a = 5/3,\ b = 7.$

For the test potential we shall consider the *resonance* problem consists in finding those eigenvalues (or energies) E in the range $1 \le E \le 1000$, at which the phase shift is equal $\pi/2$ with the domain $0 \le x \le 15$, and the initial phase is $\delta(E) = -\pi/2$. Basing on Quantum Mechanics [3], when $x \to \infty,\ v(x) \to 0$, the wave function has the following asymptotic expression

$$y(x) \to A\cos\big(\sqrt{E}x + \delta(E)\big). \tag{12}$$

In this test we choose $\omega = \sqrt{|W(x) - E|}$ and the exact eigenvalues are $E_{s8} = 53.588872, E_{s9} = 163.215341, E_{s10} = 341.495874, E_{s11} = 989.701916$.

In Figs. 1, we plot the decimal logarithm of the absolute error in 10^6 units versus the step-size with four eigenvalues among four methods I, II, III, IV.

From Figs.1, we can see that method IV are more superior to methods I , II and III. Among all the methods I, II, III and IV, method IV is the most efficient.

 ISBN: 978-84-617-8694-7

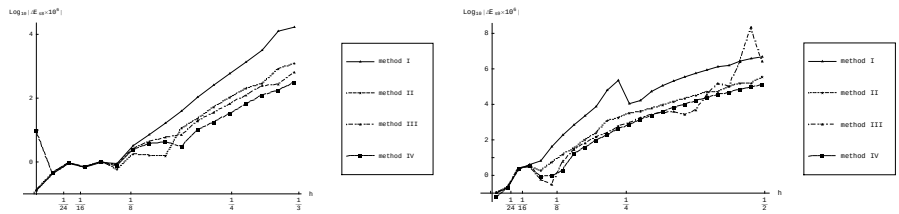


Figure 1: Efficiency curves for $E_{s8} = 53.588872$ and $E_{s9} = 163.215341$.

Acknowledgement

This research was partially supported by the National Natural Science Foundation of China (NSFC) under Grant No. 11571302.

References

[1] Z. Wang, Q. Chen, *A trigonometrically-fitted one-step method with multi-derivative for the numerical solution to the one-dimensional Schrödinger equation*, Comput. Phys. Commun. **170** (2005) 49–64.

[2] J.D. Lambert, A.R. Mitchell, *On the solution of $y' = f(x, y)$ by a class of high accuracy difference formula of low order*, Z. Angew. Math. Phys. **13** (1962) 223–232.

[3] L. Landau, E. Lifshitz, *Quantum Mechanics*, Oxford Univ. Press 1974.

 ISBN: 978-84-617-8694-7