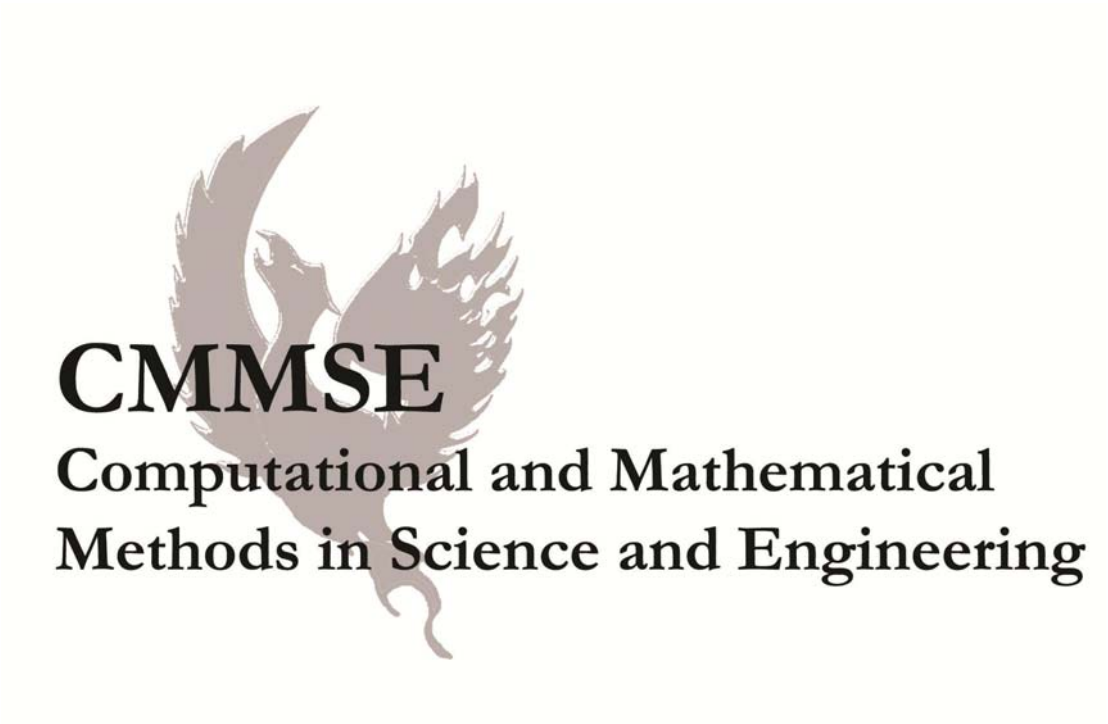
Proceedings of the 2011 International Conference on Computational and Mathematical Methods in Science and Engineering

Benidorm, Spain

June 26-30 2011



Editor:
J. Vigo-Aguiar

Associate Editors:
R. Cortina, S. Gray, J.M. Ferrándiz, A. Fernández, I. Hamilton, J.A. López Ramos, P. de Oliveira, R. Steinwandt, E. Venturino, J. Whiteman, B. Wade

ISBN 978-84-614-6167-7

Printed on acid-free paper

Preface:

We are pleased to bring the reader these proceedings containing the articles and extended abstracts presented at the ***"11th International Conference on Computational and Mathematical Methods in Science and Engineering"*** (CMMSE 2011), held in Alicante, Spain, from 26 to June 30, 2011.

The essence of science lies in the human motivation to understand the real world. This understanding has helped us to predict or even control some natural and physical phenomena. This goal would have been impossible without the use of technology, whose development is, in turn, undoubtedly linked to new scientific achievements. In this setting, Mathematics is the branch of Science that is most important to the development of other scientific disciplines and certainly contributes to important advances in technology that helps humans in their interaction with the real world.

Throughout history, Mathematics has become an essential tool to strengthen the relation between other fields of Science and Engineering, especially during the last decades of the preceding century and the first years of the present one due to the increased use of computational and mathematical methods. Only a few years ago some challenges that Science and Engineering faced were out of reach without a technology that allowed that goal. Today, interdisciplinary groups of scientists and engineers are able to create models and solve problems of a complex nature that entail huge computations using methods developed during these last years. In this way, computation has joined the two traditional components—experimentation and theory—of the scientific method. CMMSE aims to show how computational methods influence new mathematical achievements and how Mathematics plays a fundamental role in the development of new technology that allows scientists and engineers to speed up and carry out complex computations.

CMMSE is a forum where experts in many different scientific fields present their latest advances and share ideas and experiences in order to explore new directions in Science and Engineering. The CMMSE 2011 special sessions represent some of these emerging disciplines: from differential equations to Cryptology, from Biomathematics to mathematical models in Celestial Mechanics or Chemistry.

Three of the mini-symposia to be held at the conference will deal with security issues from different points of view. First, Mathematics in Cryptology, will explore cryptographic challenges, solutions and techniques of mathematical interest, including the design and analysis of cryptographic protocols. Second, we have a session dedicated to the mathematical modelling of the future internet and developing the corresponding security technology. The third mini-symposium, Computational Methods for Privacy Protection, will offer the latest advances in methods of computational intelligence, such as evolutionary algorithms and fuzzy logic that offer new and not yet fully exploited opportunities to discover threats and protect data.

Two other mini-symposia will be devoted to computational methods applied to Chemistry and Physics. Computational Nanoscience is becoming a growing and important discipline. The purpose of the Computational Nanoscience mini-symposium is to present new results of phenomena at the nanoscale, whose study generates a great need for modelling and computer simulations with applications in many fields but mainly focused on Chemistry and Physics. Mathematics in Optometry and Vision Sciences is

another mini-symposium where computational mathematical aspects apply to other sciences.

Partial differential equations will be the topic of three other mini-symposia. The first mini-symposium will be on the use of meshfree methods combined with sampling theory with applications in signal and image processing. Partial differential equations also apply in the development of other sciences as will be evidenced in the mini-symposium on Celestial Mechanics and Rotation of Earth. Lastly, from both analytical and numerical points of view, the mini-symposium on Mathematics and Numerics of Mechanics and Diffusion will pay special attention to viscoelasticity and non-Fickian diffusion with particular regard to drug delivery from biodegradable and non-biodegradable platforms.

Sophisticated numerical methods as well as efficient use of current high-performance computers employing, for instance, parallel architectures to address complex problems with high computational requirements are important for many different scientific and engineering applications. Again, and for the third consecutive year, we have the pleasure of working with the Spanish Network CAPAP-H "High Performance Computing on Heterogeneous Parallel Architectures." We would like to give a special mention to José Ranilla Enrique S. Quintana-Ortí and Diego Llanos for their very organized efforts.

We would like to thank the plenary speakers for their excellent contributions in research and leadership in their respective fields. We express our gratitude to the special session organizers and to all members of the Scientific Committee, who have been a very important part of the conference, and, of course, to all participants.

These four volumes contain all the proceedings of the conference. As a matter of style, volumes I, II and III contain the articles written in LaTeX and volume IV contains the articles written in Word and short-abstracts.

We cordially welcome all participants to CMMSE 2011. We hope you enjoy this conference.

Alicante, Spain, June 26, 2011

J. Vigo-Aguiar, R. Cortina, S. Gray, J.M. Ferrándiz, A. Fernández,
I. Hamilton, J.A. López Ramos, P. de Oliveira, R. Steinwandt, E. Venturino,
J. Whiteman, B. Wade

Acknowledgements:

We would like to express our gratitude to our sponsors:

- Generalitat Valenciana,
- Ministerio de Ciencia e Innovación,
- Universidad de Alicante.

Finally, we would like to thank all of the local organizers:

S. Díaz, J.F. Navarro, A. Pérez Carrión, V. Requena (Universidad de Alicante), P. Alonso, R. Cortina, A. Rodríguez Campa (Universidad de Oviedo), J.A. López Ramos (Universidad de Almería), F.J. Martínez Zaldívar (Universidad de Valencia), B. Martín García, M. T. Bustos, A. Fernández (Universidad de Salamanca) for his support to make possible the conference.

CMMSE 2010 Plenary Speakers :

- Stephen Gray, Center for Nanoscale Materials, Argonne National Labs, USA.
- Rober D. Iskander, Institute of Biomedical Engineering and Instrumentation, Wroclaw University of Technology, Poland.
- Bruno Raffin, INRIA, France.
- Rainer Steinwandt, Florida Atlantic University, USA.
- Ezio Venturino, University of Torino, Italy
- John Whiteman, Brunel University, UK,

The organization of the “Minisymposium on High Performance Computing was part of the activities of the Spanish Network CAPAP-H2 "High Performance Computing on Heterogeneous Parallel Architectures", supported by the Spanish Ministry of Science and Innovation (project TIN2009-08058-E).

Scientific Committee

H. M. Amman - Utrecht University, Utrecht, The Netherlands
H.T. Banks - North Carolina State University, USA
E. Boros - Rutgers University, USA
A. Brandt - UCLA & Weizmann Instit, USA & Israel
R. Criado - Universidad Rey Juan Carlos, Spain
M.I. García Fernández - Universidad de Málaga, Spain
S. Gray - Center for Nanoscale Materials, Argonne National Labs, USA
M. Greferath - University College, Dublin, Ireland
D. Estep - Colorado State University, USA
G. Fasshauer - Illinois Institute of Technology, USA
P. Forsyth - University of Waterloo, Canada
W. Han - University of Iowa, USA
F. Hickernell - Illinois Institute of Technology, USA
D. Higham - University of Strathclyde, Scotland
R. D. Iskander - Institute of Biomedical Engineering and Instrumentation Worclaw University of Technology, Poland
Y. Jiang - Los Alamos National Laboratory, USA
S. Jin - University of Madission Wisconsin, USA
A. Q.M. Khaliq - Middle Tennessee University, USA
J. Li - Pennsylvania State University, USA
J. A. López Ramos - UAL, Spain
P. Mezey - University of Newfoundland, Canada
J. J. Moreno Balcázar - University of Almería, Spain
S. Oharu - Chuo University, Japan
G. Papanicolaou - Stanford University, USA
J. Pasciak - Texas A & M University, USA
L. Petzold - Univ. of California, St Barbara, USA
G. Pinter - Univ. of Wisconsin-Milwaukee, USA
E.S. Quintana Ortí - Universidad Jaime I, Spain
B. Raffin - INRIA, France
J. Ranilla Pastor - Universidad de Oviedo, Spain
V. Rayward-Smith - University of East Anglia, UK
J. Rosenthal - University of Zurich, Switzerland
Q. Sheng - Baylor University, USA
R. Steinwandt - Florida Atlantic University, USA
E.H. Twizell - Brunel University, UK
G. Vanden Berghe - University of Ghent, Belgium
E. Venturino - University of Torino, Italy
A.M. Vidal Maciá - Universidad Politécnica Valencia, Spain
J. Vigo Aguiar - Universidad de Salamanca, Spain
B. A. Wade - Univ. of Wisconsin-Milwaukee, USA
J. Whiteman - Brunel University, U.K.
J. Xu - Pennsylvania State University, USA
D. Xie - Univ. of Wisconsin-Milwaukee, USA
Z. Zhang - Wayne State University, USA
E. Zuazua - BCAM - Basque Center for Applied Math

Volume I

 ISBN: 978-84-614-6167-7

 ISBN: 978-84-614-6167-7

Contents:

Volume I

 ISBN: 978-84-614-6167-7

 ISBN: 978-84-614-6167-7

 ISBN: 978-84-614-6167-7

Contents:

Volume II

 ISBN: 978-84-614-6167-7

 ISBN: 978-84-614-6167-7

 ISBN: 978-84-614-6167-7

Contents:

Volume III

 ISBN: 978-84-614-6167-7

 ISBN: 978-84-614-6167-7

Contents:
Volume IV

 ISBN: 978-84-614-6167-7

 ISBN: 978-84-614-6167-7

Contents:
Abstracts & Late Papers

 ISBN: 978-84-614-6167-7

 ISBN: 978-84-614-6167-7

Proceedings of the 11th International Conference on Computational and Mathematical Methods in Science and Engineering, CMMSE 2011 26–30 June 2011.

Inversion of general tridiagonal matrices: Preserving the numerical approach.

J. Abderramán Marrero[1], M. Rachidi[2] and V. Tomeo[3]

[1] *Department of Mathematics Applied to Information Technologies, Telecommunication Engineering School, U.P.M. Tech. University of Madrid, Spain*

[2] *Département de Mathématiques et Informatique, Faculté des Sciences, Université Moulay Ismail, Méknés, Morocco*

[3] *Department of Algebra, School of Statistics, U.C.M. University Complutense, Spain*

emails: jc.abderraman@upm.es, mu.rachidi@hotmail.fr, tomeo@estad.ucm.es

Abstract

After a brief introduction and justification of the method, a simple numerical algorithm for the inversion of tridiagonal nonsingular matrices (unreduced as well as reduced) is introduced. We maintain here the numerical approach, without invoking the symbolic computation, which has produced recent advances in the inversion of such matrices.

Key words: Computational complexity, difference equation, inverse matrix, numerical algorithm, tridiagonal matrix.

MSC 2000: 15A09, 15A29, 39A06, 65F05, 65Y20.

1 Inversion of general tridiagonal matrices

Algorithms for the inversion of tridiagonal nonsingular matrices are frequently used in applied sciences. The tridiagonal matrix is usually denoted as $T = \{a_i, b_i, c_i\}$ $(1 \leq i \leq n)$, with $a_1 = c_n = 0$. The coefficients $\{b_i\}$ correspond to the principal diagonal and $\{a_i\}$, $\{c_i\}$ correspond to the lower and upper subdiagonal, respectively. Commercial packages are based in gaussian inversion algorithms with pivoting strategies. These packages are efficient, specially when the tridiagonal matrix is reduced, but a great amount of memory is required. Their run times are greater than other specialized and simpler algorithms. Concerning the literature about such simple algorithms, we can cite e.g., [3, 8, 9]. In general, they are applicable only to unreduced matrices.

An analysis of the inversion of tridiagonal nonsingular matrices, without imposing any condition on the coefficients, was introduced in [5]. Nevertheless, the resulting

 ISBN: 978-84-614-6167-7

numerical algorithm breaks down when some principal submatrices are singular. Recently, some advances has been introduced with the symbolic computation [4, 6]. The computational complexity of the algorithms given in [4, 5, 6] is $O(n^2)$. As a continuation on the numerical line from [5], we introduce here an algorithm to obtain the entries of the inverse of any tridiagonal nonsingular matrix. It is based on the determinants of its principal submatrices, which verify second order linear recurrences. These determinants appear in a compact form for the entries of the inverse matrix, see [7],

$$(T^{-1})_{ij} = \begin{cases} (-1)^{i+j} \left(\prod_{k=i}^{j-1} c_k \right) \frac{\det T_{i-1} \cdot \det T_{n-j}^{(j)}}{\det T} & \text{if} \quad i \leq j, \\ \\ (-1)^{i+j} \left(\prod_{k=j+1}^{i} a_k \right) \frac{\det T_{j-1} \cdot \det T_{n-i}^{(i)}}{\det T} & \text{if} \quad i > j. \end{cases} \tag{1}$$

Matrix T_{i-1} is the left principal submatrix of order $i-1$. Matrix $T_{n-j}^{(j)}$ is the right principal submatrix of order $n-j$, which begins from the $(j+1)$-th row and column. We define here $\det T_0 = \det T_0^{(n)} = 1$. This representation for the entries of the inverse of tridiagonal matrices is a particular case of the closed representation for inverses of regular Hessenberg matrices, see e.g., [2]. It can also be obtained easily using the companion decomposition introduced recently in [1].

The complexity for the inversion of tridiagonal nonsingular matrices, is related to the obtainment of the determinants of all their principal (left and right) submatrices. Fortunately, such determinants have a fast computation when using second order linear difference equations, with complexity $O(n)$. The linear recurrence for determinants of left principal submatrices is, with initial conditions $\det T_1 = b_1$, $\det T_2 = b_2 b_1 - a_2 c_1$,

$$\det T_{k+2} = b_{k+2} \det T_{k+1} - a_{k+2} c_{k+1} \det T_k, \qquad (1 \leq k \leq n-2)\,. \tag{2}$$

For determinants of right principal submatrices, with initial conditions $\det T_1^{(n-1)} = b_n$, $\det T_2^{(n-2)} = b_{n-1} b_n - c_{n-1} a_n$, we have,

$$\det T_{k+2}^{(n-k-2)} = b_{n-k-1} \det T_{k+1}^{(n-k-1)} - c_{n-k-1} a_{n-k} \det T_k^{(n-k)}, (1 \leq k \leq n-2)\,. \tag{3}$$

Overflow and underflow can appear in the computation of such recurrences. Then, our algorithm works into the usage range. For example, the solutions for the recurrences of some diagonally dominant matrices grow quickly in magnitude. Then we must introduce other methods, as scaling transformations on the recurrences.

The inverse entries are computed with the expression from (1). The introduction of both vectors, prod$a$ for the product of the a_i and prod$c$ for the c_i, is important for the efficacy of the algorithm, in special for matrices with large order n. For the computations of $(T^{-1})_{i,j}$ in (1), we can made the following substitutions,

$$(-1)^{i+j} \left(\prod_{k=j+1}^{i} a_k \right) = \frac{\text{prod}a(i)}{\text{prod}a(j)}, (i > j) \quad ; \quad (-1)^{i+j} \left(\prod_{k=i}^{j-1} c_k \right) = \frac{\text{prod}c(j)}{\text{prod}c(i)}, (i < j)\,.$$

The determinants are evaluated with the two vector solutions of the recurrences (2)-(3).

 ISBN: 978-84-614-6167-7

2 An algorithm of inversion

Input:

- Order n and the components $\{a_i, b_i, c_i\}$ of the tridiagonal matrix T..
- A vector with the positions of rows which have a zero in the lower subdiagonal, (asign$a$).
- The total number of zeros in the lower subdiagonal, (*numbera*).
- A vector with the positions of columns which have a zero in the upper subdiagonal, (asign$c$).
- The total number of zeros in the upper subdiagonal, (*numberc*).

(For unreduced matrices we take number = 1 and asign = 1 for both subdiagonals)

Output: T^{-1} the inverse of the tridiagonal matrix T.

1. Initialize T^{-1} as the null matrix of size n.
2. Set the initial conditions. For $k = 3 : n + 1$; built the two vectors of principal determinants.
3. For $i = 1 : n$; evaluate $(T^{-1})_{ii}$, the entries of the main diagonal.
4. Evaluate the entries of the lower triangle of T^{-1}.
   (a) prod$a$(1)= 1. For $k = 2$:asign$a$(1)$-$1; prod$a$(k)= $-a_k$*prod$a$(k-1).
   For $j = 1 : k - 1$; evaluate $(T^{-1})_{k,j}$.
   (b) For $m = 2 :$ *numbera*; prod$a$(asign$a$(m-1))= 1.
   For $k =$ asign$a$(m-1)+1:asign$a$(m)$-$1: prod$a$(k)= $-a_k$*prod$a$(k-1);
   For $j =$ asign$a$(m-1):$k - 1$; evaluate $(T^{-1})_{k,j}$.
   (c) prod$a$(asign$a$(*numbera*))= 1;
   For $k =$ asign$a$(*numbera*)+1:n; prod$a$k) = $-a_k$*prod$a$(k-1);
   For $j =$ asign$a$(*numbera*):$k - 1$; evaluate $(T^{-1})_{k,j}$.
5. Evaluate the entries of the upper triangle of T^{-1}.
   (a) prod$c$(1)= 1. For $k = 2$:asign$c$(1)$-$1; prod$c$(k)= $-c_{k-1}$*prod$c$(k-1).
   For $i = 1 : k - 1$; evaluate $(T^{-1})_{i,k}$.
   (b) For $m = 2 :$ *numberc*; prod$c$(asign$c$(m-1))= 1.
   For $k =$ asign$c$(m-1)+1:asign$c$(m)$-$1: prod$c$(k)= $-c_{k-1}$*prod$c$(k-1);
   For $i =$ asign$c$(m-1) to $k - 1$; evaluate $(T^{-1})_{i,k}$.
   (c) prod$c$(asign$c$(*numberc*))= 1;
   For $k =$ asign$c$(*numberc*)+1:n; prod$c$(k) = $-c_{k-1}$*prod$c$(k-1);
   For $i =$ asign$c$(*numberc*):$k - 1$; evaluate $(T^{-1})_{i,k}$.

The algorithm does not break down when some principal submatrices are singular. It is especially useful in the reduced case. The computational complexity is $O(n^2)$ for unreduced tridiagonal matrices. This complexity diminishes when the number of zeros in the matrix subdiagonals increases. In this situation the inverse is not a full matrix. The unnecessary computation of almost all null entries is here avoided. This is illustrated in Figure 1. The algorithm is valid in the limit case of two-band triangular matrices. For unreduced matrices, the n^2 entries of the inverse matrix are evaluated.

In Figure 1 we introduce graphics for the general as well as unreduced case. We can observe the advantages of our algorithm with respect to the built-in function $inv()$ of the *Matlab* package. The tridiagonal matrices, with order $75 \leq n \leq 500$ in steps of 25 units, take random values from $[-5, 5]$. Note as its complexity diminishes in the general case (reduced as well as unreduced) with respect to the unreduced case.

 ISBN: 978-84-614-6167-7

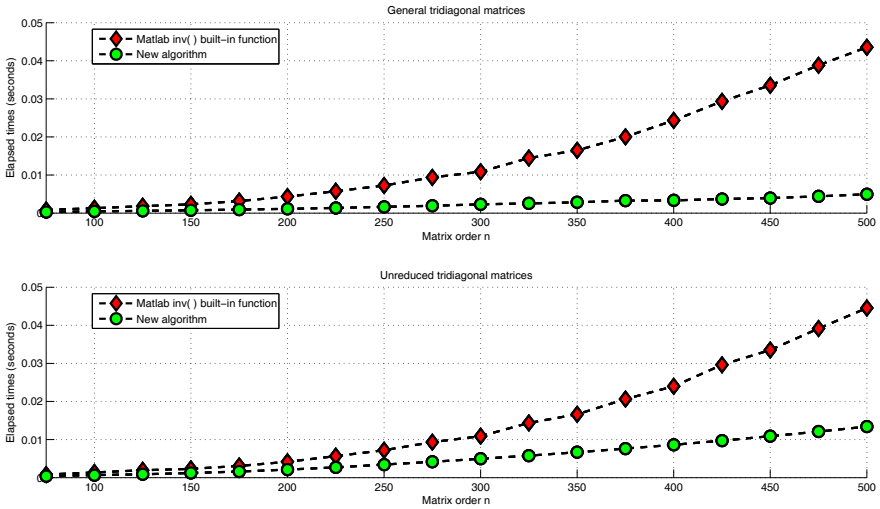


Figure 1: Mean value of the elapsed time, 150 trials, in the computations of the inverses.

References

[1] J. Abderramán Marrero and M. Rachidi, *Companion factorization in the general group $GL(n;\mathbb{C})$ and applications*, Linear Alg. Appl. 434 (2011), p.1261-1271.

[2] J. Abderramán Marrero and V. Tomeo, *On the closed representation for the inverses of Hessenberg matrices*, J. Comp. App. Math., (2010), submitted.

[3] B. Bukhberger and G.A. Emel'yanenko, *Methods of inverting tridiagonal matrices*, USSR Computational Math. and Math. Phys., **13** (1973), 10–20.

[4] M. El-Mikkawy and A. Karawia, *Inversion of general tridiagonal matrices*, Appl. Math. Lett., **19** (2006), 712–720.

[5] Y. Huang and W.F. McColl, *Analytical inversion of general tridiagonal matrices*, J. Phys. A: Math. Gen., **30** (1997), 7919–7933.

[6] H. Li, T. Huang, X. Liu, and H. Li, *On the inverses of general tridiagonal matrices*, Linear Alg. Appl. **433** (2010), 965-983.

[7] P. N. Shivakumar, and C. Ji, *Upper and Lower Bounds for Inverse Elelements of Finite and Infinite Tridiagonal Matrices*, Linear Alg. Appl. **247** (1996), 297-316.

[8] S. R. Vatsya and H. O. Pritchard, *An explicit inverse of a tridiagonal matrix*, Intern. J. Computer Math., **14:3** (1983), 295-304.

[9] T. Yamamoto and Y. Ikebe, *Inversion of band matrices*, Linear Alg. Appl. **24** (1979) 105-111.

 ISBN: 978-84-614-6167-7

Proceedings of the 11th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2011
26–30 June 2011.

From latex specifications to parallel codes

Alejandro Acosta[1], Francisco Almeida[1] and Ignacio Peláez[1]

[1] *Dpt. Statistics and Computer Science, La Laguna University, Spain*

emails: aacostad@ull.com, falmeida@ull.com, ignacio.pelaez@gmail.com

Abstract

The advent of multicore systems, joined to the potential acceleration of the graphics processing units, has given us a low cost computation capability unprecedented. The new systems alleviate some well known important architectural problems at the expense of a considerable increment of the programmability wall. The heterogeneity, both at architectural and programming level at the same time, raises the programming difficulties. As a contribution in this context, we propose a development methodology for the automatic source-to-source transformation on specific domains. This methodology is successfully instantiated as a framework to solve Dynamic Programming problems. As a result of applying our framework, the end user (a physicist, a mathematician or a biologist) can express her problem through a latex equation and automatically derive efficient parallel codes for current homogeneous or heterogeneous architectures. The approach allows an easy portability to new potential emergent architectures.

Key words: Translators, Dynamic Programming, Portability

1 Introduction

Current generation of computers is based on architectures based on multiple identical processing units composed of several cores (multicores) and it is expected that the number of cores per processor be incremented every year. It is also a well know fact that the current generation of compilers is not being able to transfer automatically the capacity of the new processing units to the applications. The situation is further complicated given that current architectures are of heterogeneous nature, where this multicore systems can be combined, for example, with the capabilities of using GPU system as general purpose processing architectures. This fact constitutes a severe difficulty that is appearing in the form of a barrier to the programmability.

 ISBN: 978-84-614-6167-7

Many are the proposals to tackle with this problem. Leaving aside the proposals based on the development of new programming languages, due to inconvenience caused to the user (new learning effort and code reusability), many of the approaches are based in the source-to-source transformation of sequential code into parallel code or in the transformation of parallel code designed for one architecture into parallel code for a different one [25, 9, 26, 16]. Another different approach is based in the use of skeletons. The programmer is provided with a set of patterns already parallelized that constitute a frame to develop parallel code, just by supplying sequential code [4, 13]. It worth also to mention the reasearch on frameworks devoted to build the former source-to-source transformers [23, 20, 3, 7].

Although technologically impressive, none ofthe projects based in skeletal parallel programming have achieved significant popularity in the wider parallel programming community. However, we claim that many of the developments made in the context of skeletal programming may play an important role in the automatic code generation based in source-to-source transformations. An important difficulty in the source-to-source transformation process is to transform sequential code sections into their parallel equivalent sections. That implies that the transformer must know in advance the sections to be parallelized, and how they should be translated, typically the user annotates the sections to be transformed.

An interesting feature of parallel skeletons is that the parallelism is hidden to the end user and is encapsulated into parallel patterns. Usually, the user fills gaps in the skeleton by providing sequential codes. New parallelizations (for new architectures for example) can be developed without any modification of the sequential code supplied by the user.

We have developed a source-to-source translator based in skeletons that generates code for many parallel architectures. The main goal is that the end user may obtain parallel code, without any knowledge in programming, just by defining her problem using a more natural language as the mathematics.

An advantage of our approach is that, in general, a source-to-souce transformation from sequential code to sequential code is semantically easier to develop that a transformation from sequential to parallel code. That is one of the fundamentals of our project, we automatically fill the sequential gaps in a parallel skeleton starting from a very user friendly specification. The parallelism is automatically provided by the skeleton and can be very easily extended. Since many parallel skeletons have been already developed and they work efficiently in current architectures, once the transformers have been developed, the level of productivity in terms of parallel code generated is highly increased.

As a proof of concept we apply the methodology to the dynamic programming technique, this technique is frequently applied to many research areas such as Control Theory, Operations Research, Biology, etc., [19, 11, 14]. As a result of this research it raises an specification language for Dynamic Programming problems that also constitutes a contribution of this work.

This remaining of the paper has been structured as follows: in section 2 we present the

 ISBN: 978-84-614-6167-7

methodology that we propose to broach the problem, in section 3 we raise the framework developed in the context of Dynamic Programming problems and in section 4 we include some computational results obtained from our tool and point out the high productivity achieved by the approach while keeping the efficiency at the same time. Finally we end the paper with some concluding remarks and future lines of work.

2 The methodology

Usually, source-to-source translators are used to make easier the work of developers. The source language use to have a higher abstraction level than the target language. Many translator have been developed and they typically follow the common structure shown at Figure 1 that operates in two different phases, the Front-end and the Back-end. The Front-end analyzes the input code and is responsible for the correctness of the code. For that purpose it performs the Lexical Analysis, the Syntax Analysis and the Semantical Analysis. The output of this phase is an intermediate code that will be the input of the next phase. The Back-end generates the output on the target language. In this phase optimization techniques can be applied to generate code so efficient as possible in the target architecture.

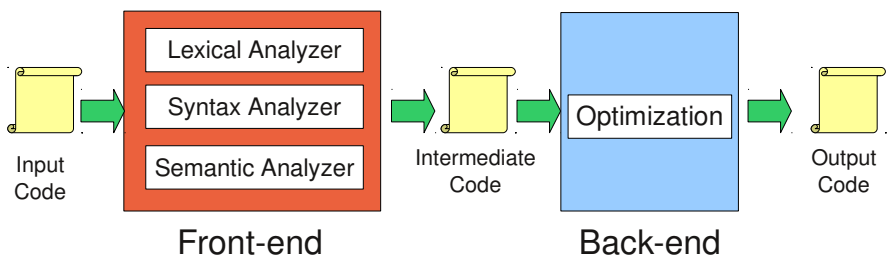


Figure 1: Basic model in source-to-source transformation

This division provides a high flexibility to the translator since the Front-end depends on the input language and is independent on the architecture. The Back-end on the other hand depends on the platform, and is independent from the input language. That allows to reuse the same Back-end to generate output code starting from several input languages, if different Front-ends are used. At the same time, the same Front-end can be used to generate output codes for different architectures, if Back-ends adapted to the target platforms are used.

In the skeletal based translation we propose to follow the same structure presented in

 ISBN: 978-84-614-6167-7

Figure 1 but introducing new layers and providing an increased general abstraction view (Figure 2). The proposal is close to that presented in [3]. In this case, the code generated by the Back-ends is the input code for a parallel skeleton. The input code in a parallel skeleton use to be a sequential code that is guarantied to be executed in parallel through the parallel patterns encapsulated into the skeleton. This patterns are adapted to several platforms. New parallel patterns can be developed for new architectures and also de skeletons are suitable for static and dynamic optimizations. Note that we separate the source to source transformation from the parallelization. The transformations are only of sequential codes to sequential codes, while the parallelizations are abstracted into the skeletons. The parallel code generated by our Back-end is suitable to be executed in many parallel architectures in terms of the parallel skeleton to be combined for the execution.

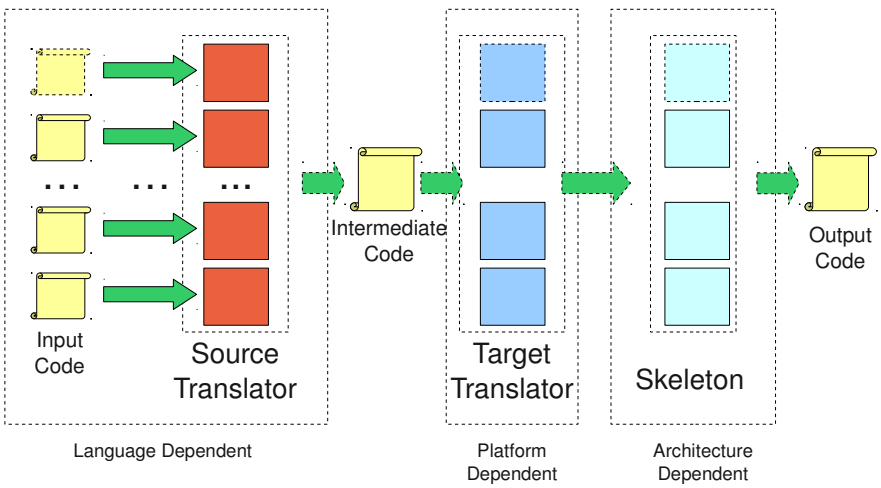


Figure 2: Model for the proposed architecture

Using this model, we propose a design structure where each of the phases can be overviewed as source-to-source translators (Figure 2). The Front-end may be seen as a source-to-source translator that generates an intermediate code, and the Back-end receives as input this source intermediate code and generate the output. By adding the skeletons the model allows to develop a *source translator*, that generates intermediate code independent from the architecture, and also a second *target translator* independent from the input intermediate language that generates an output code adapted and optimized for different architectures.

This model is quite flexible since allows the use the same *target translator*, and different *source translators*, to generate parallel code starting from various input languages. Or, at the same time, the use or the same *source translator*, and different *target translators*, to

 ISBN: 978-84-614-6167-7

produce output code for different skeletal software platforms. The target code generated by the *target translator* can be used in many different parallel architectures just by combining the skeleton. New adapted skeletons can be developed for new emergent architectures without any change nor new developments in the whole translation process.

As a proof of concept we have implemented a source-to-source translator that follows this model (Figure 3). The translator is directed to solve Dynamic Programming problems on parallel architectures. Of course, although the specific development of this paper is oriented to the Dynamic Programming technique, the same development model can be applied to many other contexts.

3 The dynamic programming technique: a proof of concept

Dynamic Programming (DP) is an important problem-solving technique that has been widely used in various fields such as control theory, operations research, economy, biology and computer science [19, 11, 14]. In DP an optimal sequence of decisions is arrived at by making explicit appeal to the principle of optimality.

For example [15, 8], however most of the parallelizations presented are for specific DP problems (see [1, 2]) or are restricted to limited classes of recurrences. A unified parallel general approach was presented in [6] as an extension to the work of [15] but the strong theoretical effort introduced in some cases dissuades us from using it as a model for developing parallel tools.

In conclusion, generic parallel approaches for DP are limited to classes of problems or they are not suitable to be assumed by a software component. It is worth mentioning that another source of difficulties is the fact that the notation used changes substantially from one formalization to the other. In most of the cases, how to obtain the optimal policy providing the optimal solution is left outside of the formalizations, and usually remains expressed as a non-formalized, sometimes intuitive, procedure.

Analyzing the software approaches for DP, we found a group of general libraries for combinatorial optimization problems such as [10, 5]. They are used to supply interfaces for sequential and parallel executions but in most of the cases DP is not considered at all. Next, we can find specific DP sequential libraries such as [18], and interesting software approaches derived from laboratories that apply solvers such as LINGO [17] to DP problems, following particular methodologies. In [22] we contributed with DPSKEL, a parallel skeleton where many efficient parallelizations for DP on different architectures are offered to the end user. The end user fills gaps on a C++ sequential code and the parallelism is automatically provided. In [21] we presented DPSPEC, a XML specification for DP problems that could be used as an alternative instead of the C++ interface for the DP parallel skeletons. Although for a scientist (a biologist, a physician, or an economist) XML is easier to manage, it still remains as a non natural approach. By other side, the problem of finding the optimal policy

 ISBN: 978-84-614-6167-7

$$InputData \equiv \begin{cases} n \in N & \#\ The\ number\ of\ objects \\ C \in N & \#\ The\ capacity\ of\ the\ Knapsack \\ p_k \in N;\ k \in \{0 \dots n-1\} & \#\ The\ profit\ of\ object\ k \\ w_k \in N;\ k \in \{0 \dots n-1\} & \#\ The\ weight\ of\ object\ k \end{cases}$$

$$OutputData \equiv \begin{cases} x_k \in \{0,1\};\ k \in \{0 \cdots n-1\} & \#\ The\ solution\ vector \\ n-1, C & \#\ The\ index\ solution \end{cases}$$

$$DecisionDef \equiv \{\ d_{k,c} \in \{0,1\};\ k \in \{0 \cdots n-1\};\ c \in \{0 \cdots C\} \quad \#\ The\ decisions$$

$$DPRecurrence \equiv$$
$$f_{k,c} = \begin{cases} 0 \ \Rightarrow\ d_{k,c} = 0 & if\ c < w_k \\ p_k \ \Rightarrow\ d_{k,c} = 1 & if\ k = 0\ and\ c \geq w_k \\ \max\{f_{k-1,c} \Rightarrow d_{k,c} = 0, f_{k-1,c-w_k} + p_k \Rightarrow d_{k,c} = 1\} & if\ k \neq 0\ and\ c \geq w_k \end{cases}$$

$$FormerDecision \equiv \begin{cases} & \#\ Assign\ solution\ k \\ x_k = d_{k,c} & if\ k \geq 0\ and\ c \geq 0 \\ & \#\ Next\ decision\ to\ assign \\ k-1; c-(w_k * x_k); & if\ k > 0\ and\ c \geq (w_k * x_k) \end{cases}$$

Table 1: Latex specification for the Knapsack Problem

after the optimal value is computed remained unsolved at that moment.

As a contribution of this paper, we propose a new specification language for DP problems that integrates all the elements of the DP technique, including how to compute the optimal policy. DP problems using this specification can be transformed automatically into our parallel skeletons through our intermediate language DPSPEC. To achieve it, DPSPEC and the parallel skeletons have been conveniently extended.

The input defines a structure where the user can define the DP problem without any knowledge of programming, using a more natural language as the mathematics. The code for this structure is defined using the LaTeX processor, widely used by the scientific community. We define a template where the user can define the problem to be solved and its parameters. We illustrate this specification using the well known DP approach for the Knapsack Problem(Table 1). As we can see the input data and solution or a problem are described at the *InputData* and *OutputData* sections respectively, the DP recurrence in the section *DPRecurrence*. Note that when this section is defined, the decisions $d_{k,c}$ are included so that the optimal policy can be obtained from the specification presented in section *FormerDecision*.

This LaTeX code will be transformed using the transformer Tex2DPS to DPSPEC.

 ISBN: 978-84-614-6167-7

The use of LaTeX is just a proof of concept, however from the methodological point of view, any other transformer producing the intermediate DPSPEC code is valid. DPSPEC is the input of a second translator (XML2Cpp), which is the responsible for translating the XML specification to C++ code adapted to the parallel platform. The C++ code generated corresponds with the sequential sections (Required sections in Figure 3) of the parallel skeleton (Provided sections in Figure 3). The flexibility in our approach allows to develop new translators from XML to another library of skeletons providing new or different functionalities.

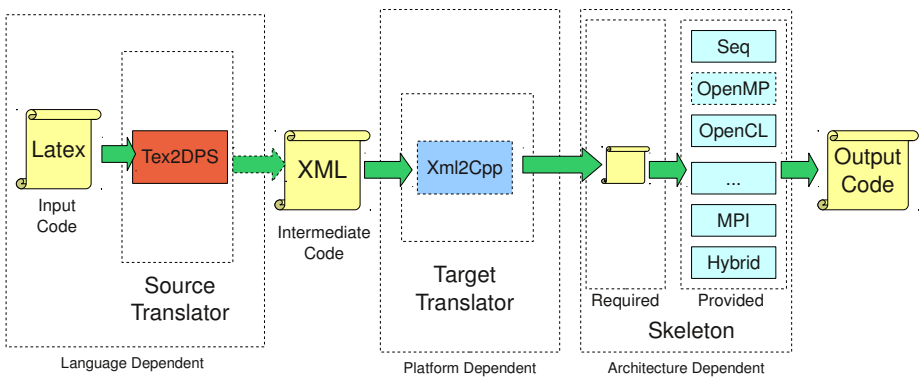


Figure 3: The proposed software architecture

Since skeletons are defined for different architectures, shared memory, message passing, hybrid shared memory/message passing (see Figure 3), the methodology provides a huge portability, specially if one consider that extending the approach to a new emergent platform, just means to include a new parallel skeleton.

4 Computational Results

To validate our methodology and the framework that we have developed, we tested with several Dynamic Programming problems (Table 2). Five different DP recurrences have been considered that are quite representative of wide class of problems. Note that the data dependences are different in most of the formula considered. That means that different parallel traverses of the DP table can be required. We just represented the DP problems using our LaTeX specification language and automatically generated the parallel codes. Four parallel skeletons have been considered, the sequential one and parallel versions on OpenMP, MPI and MPI/ULL_CALIBRATE. This last version is a distributed memory MPI version combined with the ULL_CALIBRATE library [12] to optimize in run time through dynamic

 ISBN: 978-84-614-6167-7

Problem	Recurrence
0/1 Knapsack KP	$f_{i,j} = max\{f_{i-1,j}, f_{i-1,j-w_i} + p_i\}$
Resource Allocation RAP	$f_{i,j} = p_{1,j}\ if\ i = 1\ and\ j > 0$ $f_{i,j} = \max_{0 \le k < j}\{f_{i-1,j-k} + p_{i,k}\}\ if\ (i > 1)\ and\ (j > 0)$
Matrix Parentization MPP	$f_{i,j} = \min_{i \le k < j}\{f_{i,k} + f_{k+1,j} + (dim_i * dim_{k+1} * dim_{j+1})$
Triangulation Convex Polygons TCP	$f_{i,j} = cost_i * cost_{i+1} * cost_{i+2}\ if\ (i = (j-2))$ $f_{i,j} = \min_{i<k<j}\{f_{i,k} + f_{k,j} + (cost_i * cost_k * cost_j)\}\ if\ (i \neq j-1)$
Guillotine Cut GCP	$f_{i,j} = \max \begin{cases} \max_{0 \le k < object}\{profit_k\} \\ \max_{0 \le z \le i/2}\{f_{z,j} + f_{i-z,j}\} \\ \max_{0 \le y \le j/2}\{f_{i,y} + f_{i,j-y}\}\} \end{cases}$

Table 2: Dynamic Programming test problems

load balancing.

The parallel platform used to execute our experiments is an AMD Opteron 6128 node (4 processors, each processor composed of 8 cores), with 32 cores sharing the memory. We have used only 20 of them in our tests. To simplify the experience, the tests have been developed using squared matrices of sizes 1000, 2000 and 5000. Note that according to the dependences of problems on Table 2, that are graphically represented in Figure 4, several traversing parallel approaches can be used to obtain the solution. The parallel skeletons used compute rows in parallel in the case of the RAP and KP, the MPP and TCP are processed by computing the diagonals down-top in parallel and in the case of GCP the diagonals are computed in parallel top-down.

Table 3 shows the running times of the sequential executions for all the proposed problems. This table provides a general view on the granularity of each problem. All the times are expressed in seconds. We can see as the KP is the problem with the finest granularity.

Size	1000	2000	5000
KP	0.4	1.63	10.2
RAP	20	162	2545
TCP	30	275	5128
MPP	31	282	5240
GCP	84	773	14033

Table 3: Sequential execution for test problems

 ISBN: 978-84-614-6167-7

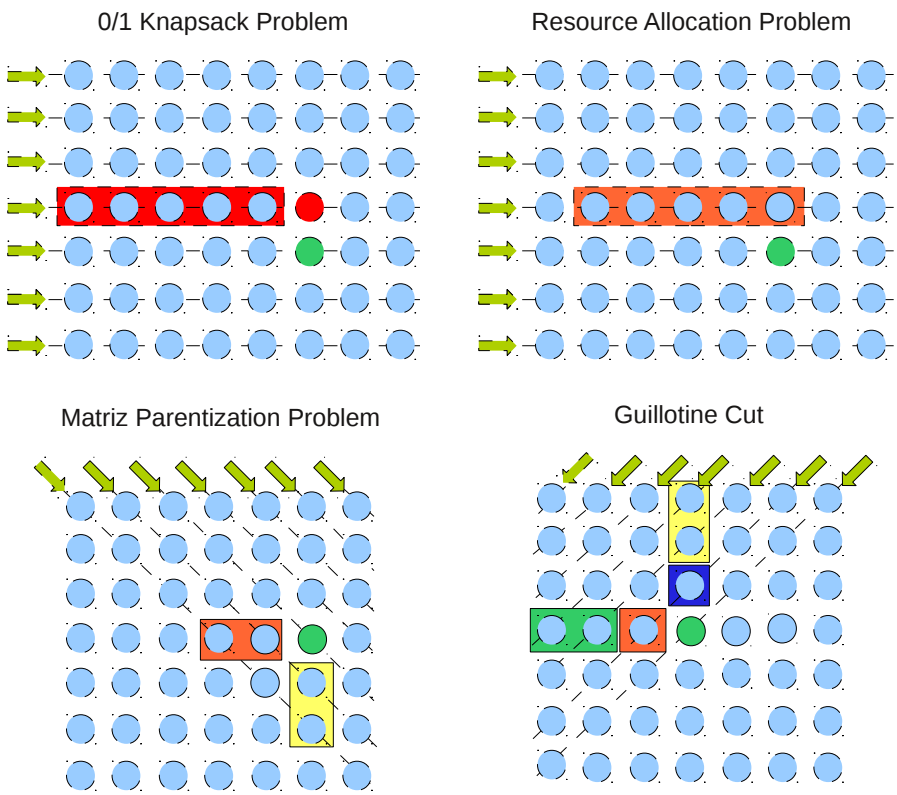


Figure 4: Dependences on the Dynamic Programming test problems

 ISBN: 978-84-614-6167-7

Skeleton	OpenMP					MPI				
Size	# cores					# cores				
	2	4	8	16	20	2	4	8	16	20
1000	15	9	4	2.4	2	14	9	4	2.4	2
2000	126	73	39	20	16	119	70	37	19	16
5000	1993	1153	616	311	257	1882	1093	591	311	255

Skeleton	MPI+ULL_CALIBRATE				
Size	# cores				
	2	4	8	16	20
1000	11	5	3	1.7	1.5
2000	81	44	23	12	10
5000	1304	700	399	180	141

Table 4: Running times of the RAP for several Skeletons

In Table 4 we show the running times obtained of executing all the skeletons over the RAP. As expected in this machine, we observe no significative differences between the the OpenMP and the MPI skeletons. When using mixed MPI/ULL_CALIBRATE skeleton the improvement is quite significative. This skeletons take advantage of the heterogeneous nature of the RAP recurrence to develop a dynamic load balancing of the rows.

In Table 5 we show the running times obtained of the execution with the remaining problems. We use the OpenMP skeleton for the KP, and the MPI skeletons for the MPP, TCP and GCP. We can see the benefits obtained from the parallelizations with the little effort of development imposed by our tool.

Figure 5 shows the speedup obtained for each one of the problems and for sizes 1000 and 5000. We can see as the KP decreases the speedup after a few number of processors, this is due to the low granularity, and how the RAP keeps an increasing speedup as a consequence of using the dynamic load balancer.

5 Conclusion and Future Work

We propose a source-to-source transformation methodology based in skeletal programming. The model is quite flexible and allows high levels of code reusability, portability and productivity. Since the parallel structure is decoupled from the translation model, no loss of efficiency is introduced by the approach. As a case of use, we applied the technique to Dynamic Programming problems. Several different problems expressed in LaTeX are automatically transformed into parallel programs that follow different parallelization patterns

 ISBN: 978-84-614-6167-7

Problem	MPP					TCP				
Size	# cores					# cores				
	2	4	8	16	20	2	4	8	16	20
1000	17	8	3.9	1.9	1.6	18	9	3.7	1.9	1.5
2000	146	73	41	22	21	143	72	36	22	20
5000	2659	1358	669	349	386	2596	1344	669	349	373
Problem	KP					GCP				
Size	# cores					# cores				
	2	4	8	16	20	2	4	8	16	20
1000	0.22	0.11	0.06	0.08	0.07	57	32	17	7.3	7.5
2000	0.86	0.44	0.23	0.24	0.24	480	271	193	74	69
5000	5.7	2.7	1.4	1.3	1.1	8772	4858	2559	1309	1517

Table 5: Running times for MPP, TCP, KP and GCP

implemented in various parallel libraries. The efficiency of the parallel code generated has been proved. For the near future we will be involved in two research directions. At the level of the skeletons we aim to extend DPSKEL with an OpenCL new skeleton, that would allow the portability of our methodology to GPUs. At the level of the Back-end translator, we propose to generate code for a new output language from the intermediate language. This output combined with the OpenCF [24] framework will provide web services interfaces so that the translators, and the parallel platforms, can be used transparently through web interfaces.

Acknowledgment

This work has been supported by the EC (FEDER) and the Spanish MEC with the I+D+I contract number TIN2008-06570-C04-03 and by the Canary Government project SolSubC200801000307.

References

[1] R. Andonov, S. Balev, S. Rajopadhye, and N. Yanev. Otimal semi-oblique tiling and its application to sequence comparison. In *13th ACM Symposium on Parallel Algorithms and Architectures (SPAA)*, 2001.

[2] R. Andonov and S. Rajopadhye. Optimal Orthogonal Tiling of 2-D Iterations. *Journal of Parallel and Distributed Computing*, 45:159–165, September 1997.

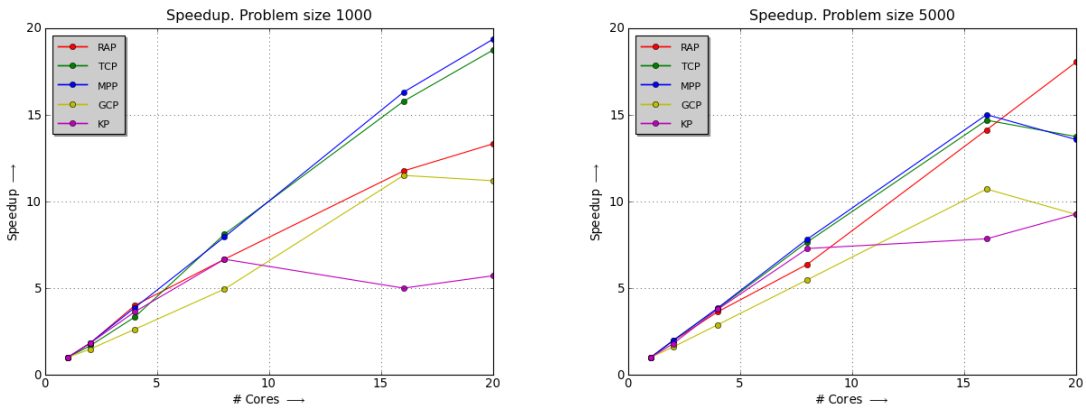


Figure 5: Speedups for the problems of sizes 1000 and 5000

[3] S. Benkner, E. Mehofer, and S. Pllana. Towards an intelligent environment for programming multi-core computing systems. In *Proceedings of the 2nd Workshop on Highly Parallel Processing on a Chip (HPPC 2008), in conjunction with Euro-Par 2008*, August 2008.

[4] A. Benoit and M. Cole. Two fundamental concepts in skeletal parallel programming. In *The International Conference on Computational Science (ICCS 2005) , Part II, LNCS 3515*, pages 764–771. Springer Verlag, 2005.

[5] B. L. Cun. Bob++ library illustrated by VRP. In *European Operational Research Conference (EURO'2001)*, page 157, Rotterdam, 2001.

[6] M. D., A. F., R. C., R. J., and D. A. Coloma I. Parallel dynamic programming and automata theory. *Parallel Computing*, 2000.

[7] C. Dave, H. Bae, S.-J. Min, S. Lee, R. Eigenmann, and S. P. Midkiff. Cetus: A source-to-source compiler infrastructure for multicores. *IEEE Computer*, 42(11):36–42, 2009.

[8] O. de Moor. Dynamic programming as a software component. In N. Mastorakis, editor, *Proc. 3rd WSEAS Int. Conf. Circuits, Systems, Communications and Computers*, 1999.

[9] I. Dooley. Automated source-to-source translations to assist parallel programmers. Master's thesis, Dept. of Computer Science, University of Illinois, 2006. http://charm.cs.uiuc.edu/papers/DooleyMSThesis06.shtml.

[10] J. Eckstein, C. A. Phillips, and W. E. Hart. PICO: An object-oriented framework for parallel branch and bound. Technical report, RUTCOR, 2000.

 ISBN: 978-84-614-6167-7

[11] A. Erdelyi and H. Topaloglu. A dynamic programming decomposition method for making overbooking decisions over an airline network. *INFORMS J. on Computing*, 22:443–456, July 2010.

[12] I. Galindo, F. Almeida, V. Blanco, and J. Badía. Dynamic load balancing on dedicated heterogeneous systems. In A. Lastovetsky, T. Kechadi, and J. Dongarra, editors, *Recent Advances in Parallel Virtual Machine and Message Passing Interface. EuroPVM/MPI 2009*, volume 5205 of *Lecture Notes in Computer Science*, pages 64–74, Dublin, Ireland, Sept. 2008. Springer Verlag.

[13] H. González-Vélez and M. Leyton. A survey of algorithmic skeleton frameworks: high-level structured parallel programming enablers. *Softw. Pract. Exper.*, 40:1135–1160, November 2010.

[14] K. Huang and Y.-T. Liang. A dynamic programming algorithm based on expected revenue approximation for the network revenue management problem. *Transportation Research Part E: Logistics and Transportation Review*, In Press, Corrected Proof:–, 2010.

[15] T. Ibaraki. *Enumerative Approaches to Combinatorial Optimization, Part II.* Annals of Operations Research. Volume 11, 1-4, 1988.

[16] F. V. Lionetti, A. D. McCulloch, and S. B. Baden. Source-to-source optimization of cuda c for gpu accelerated cardiac cell modeling. In *Proceedings of the 16th international Euro-Par conference on Parallel processing: Part I*, EuroPar'10, pages 38–49, Berlin, Heidelberg, 2010. Springer-Verlag.

[17] P. Lohmander. Deterministic and stochastic dynamic programming. www.sekon.slu.se/ PLO/diskreto/dynp.htm.

[18] B. C. Lubow. SDP: Generalized software for solving stochastic dynamic optimization problems. *Wildlife Society Bulletin*, 23:738–742, September 1997.

[19] J. Nascimento and W. Powell. Dynamic programming models and algorithms for the mutual fund cash balance problem. *Manage. Sci.*, 56:801–815, May 2010.

[20] S. Pai, R. Govindarajan, and M. J. Thazhuthaveetil. Plasma: Portable programming for simd heterogeneous accelerators.

[21] I. Peláez, F. Almeida, and D. González. An xml specification for automatic parallel dynamic programming. In *International Conference on Computational Science (1)*, pages 872–875, 2006.

 ISBN: 978-84-614-6167-7

[22] I. Peláez, F. Almeida, and F. Suárez. Dpskel: A skeleton based tool for parallel dynamic programming. In *Seventh International Conference on Parallel Processing and Applied Mathematics, PPAM2007*, 2007.

[23] ROSE. www.rosecompiler.org.

[24] A. Santos, F. Almeida, V. Blanco, D. Díez, J. Regueira, and E. Sicilia. Towards automatic service generation and scheduling in the OpenCF project. *International Journal of Web and Grid Services*, 4(4):367–378, Dec. 2008.

[25] M. Schordan and D. J. Quinlan. A source-to-source architecture for user-defined optimizations. In *JMLC*, pages 214–223, 2003.

[26] S. zee Ueng, M. Lathara, S. S. Baghsorkhi, and W. mei W. Hwu. W.m.w.: Cuda-lite: Reducing gpu programming complexity. In *In: LCPC08. Volume 5335 of LNCS*, pages 1–15. Springer, 2008.

 ISBN: 978-84-614-6167-7

Proceedings of the 11th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2011
26–30 June 2011.

A new watermarking algorithm based on multichannel wavelet functions

Santa Agreste[1] **and Luigia Puccio**[1]

[1] *Department of Mathematics, University of Messina*

emails: sagreste@unime.it, gina@unime.it

Abstract

Copyright is a form of protection provided by laws to the authors of "original works of authorship". In spite of several international laws the digital piracy is growing. Therefore it becomes crucial to develop numerical algorithms in order to hinder such phenomenon. In this paper we focus the attention on a new watermarking technique applied to digital color images. Our aim is to describe the realized watermarking algorithm based on multichannel wavelet functions with multiplicity $r = 3$, called MCWM 1.0. At the end we describe some experimental tests executed to analyze the properties of our algorithm. In particular we report some important numerical results in order to show its robustness to geometrical attacks. *Key words: Copyright protection; watermarking techniques; digital color image processing; multichannel wavelet functions*
MSC 2000: 65T60, 68U10, 94A08

1 Introduction

Copyright is an automatic author's rights protection under international laws relative to "original work of authorship". The digital piracy involving images, music, movies, books, and so on, is a legal problem that has not found a solution. Therefore it becomes crucial to create and to develop methods and numerical algorithms in order to solve the copyright problems. One of the most popular approaches considered as a tool for providing copyright protection is *digital watermarking*. It is a method or technique that hides information into digital objects. The insertion takes place by manipulating the content of digital data. The watermarking techniques alone do not protect from illegal copying. Indeed they are used to discourage a user from illegally redistributing copies of the media. In this paper we focus the attention on watermarking techniques applied to digital color images.
Generally watermarking algorithms are based on the *embedding* and the *detection* processes. These two phases are temporarily independent, but without the embedding

 ISBN: 978-84-614-6167-7

phase the detection phase would not be effective. An important property of the watermarking algorithms is the robustness. In this context a *robust* algorithm is an algorithm that is resistant to digital alterations, called *attacks*. These operations have the aim to modify, delete, or substitute the watermark [1, 2]. A classification of watermarking algorithm is based on type of different techniques for the embedding and detection phases. Concerning the first phase the watermarking algorithms are classified into *spatial domain* and *frequencies domain* algorithms according to the domain where the watermarking signal is embedded. In the spatial domain techniques watermarking scheme works directly on the pixel bit [3]. In the frequency domain techniques the watermark is embedded by altering some frequency coefficients of the image transformed by Discrete Cosine Transform (DCT) [4, 5], Discrete Wavelet Transform (DWT) [6, 7, 8, 9] or any other transform. The spatial domain techniques are less robust than frequency domain ones.In this article we present a novel approach for the watermarking process based on multichannel wavelets for digital color images. In the literature [10, 11] *multichannel wavelet* functions have been introduced to process *vector-valued* signals that must be processed as multichannel signals. A digital color image in RGB (or HSV) color model is a vector-valued two-dimensional signal because each pixel is a 3-vector. Therefore we have used multichannel wavelet functions to realize a frequency domain watermarking algorithm. We have applied our algorithm to images belonging to web virtual images galleries. The numerical results show an improvement of robustness property compared with our previous developed algorithms based on wavelet functions [6, 7]. The work is organized into three parts. In the first part we introduce the multichannel wavelet theory. In the second part we explain the realized watermarking algorithm for digital color images. In the third part we show the results relative to a large experimentation to verify the properties of imperceptibility and robustness to different attacks.

2 Multichannel wavelet theory

In this section we introduce multichannel wavelets from the point of view of *multichannel MRA* [11]. Let $L_2(\mathbb{R})^{\mathbb{Z}_r}$ be the space

$$L_2(\mathbb{R})^{\mathbb{Z}_r} = \left\{ \boldsymbol{f} : \mathbb{R} \to \mathbb{R}^r \ : \ \|\boldsymbol{f}\|_2 = \left(\sum_{j \in \mathbb{Z}_r} \int_{\mathbb{R}} \|f_j(x)\|^2 dx \right)^{1/2} < \infty \right\}$$

of square integrable vector fields. From now on we use the abbreviation "MRA" to consider the *multichannel MRA*.

A MRA is defined by a nested sequence of closed subspaces $V_0 \subset V_1 \subset ... \subset L_2(\mathbb{R})^{\mathbb{Z}_r}$, with the following properties:

1. they are *shift invariant*: $\boldsymbol{f} \in V_k \Leftrightarrow \boldsymbol{f}(\cdot + j) \in V_k$, $j \in \mathbb{Z}$
2. they are *scaled* versions of each other: $\boldsymbol{f} \in V_0 \Leftrightarrow \boldsymbol{f}(2^k \cdot) \in V_k$

 ISBN: 978-84-614-6167-7

3. they are generated by *stable* integer translates of certain vector fields. Let

$$V_0 = span\left\{\boldsymbol{f}_j(\cdot - k) : j \in \mathbb{Z}_r, k \in \mathbb{Z}\right\}$$

then the translates of the vector fields $\boldsymbol{f}_j$, $j \in \mathbb{Z}_r$ form a stable Riesz basis in $L_2(\mathbb{R})^{\mathbb{Z}_r}$, that is,

$$\left\| \sum_{j\in\mathbb{Z}_r} \sum_{k\in\mathbb{Z}} c_{jk} \boldsymbol{f}_j(\cdot - k) \right\|_2 \sim \left(\sum_{j\in\mathbb{Z}_r} \sum_{k\in\mathbb{Z}} |c_{jk}|^2 \right)^{1/2} \tag{1}$$

Any MRA is generated by a *matrix refinable function* $\boldsymbol{F} \in L_2(\mathbb{R})^{\mathbb{Z}_r\times\mathbb{Z}_r}$, that is, a function for which there exists a finitely supported mask

$$\boldsymbol{A} = (\boldsymbol{A}(k) \in \mathbb{R}^{\mathbb{Z}_r\times\mathbb{Z}_r} : k \in \mathbb{Z})$$

such that:

$$\boldsymbol{F} = (\boldsymbol{F} * \boldsymbol{A})(2\cdot) := \sum_{k\in\mathbb{Z}} \boldsymbol{F}(2\cdot - k)\boldsymbol{A}(k)$$

where $*$ symbol represents the *convolution operator* between matrix valued function $\boldsymbol{F}$ and matrix sequences $\boldsymbol{A}$.
The matrix valued function $\boldsymbol{F} \in L_2(\mathbb{R})^{\mathbb{Z}_r}$ which generates the MRA and whose columns are the vector-fields f_j, $j \in \mathbb{Z}_r$, is also called the *scaling function*. A MRA is generated by an *orthogonal* matrix function $\boldsymbol{F} \in L_2(\mathbb{Z})^{\mathbb{Z}_r\times\mathbb{Z}_r}$ if

$$\langle \boldsymbol{F}, \boldsymbol{F}(\cdot - j)\rangle = \delta_{0j}\boldsymbol{I}$$

where $\langle\cdot,\cdot\rangle$ represents the skew symmetric bilinear form

$$\langle\cdot,\cdot\rangle : L_2(\mathbb{R})^{\mathbb{Z}_m\times\mathbb{Z}_k} \times L_2(\mathbb{R})^{\mathbb{Z}_m\times\mathbb{Z}_l} \longrightarrow \mathbb{R}^{\mathbb{Z}_k\times\mathbb{Z}_l}$$

defined as

$$\langle F, G\rangle := \int_{\mathbb{R}} \boldsymbol{F}^T(x)\boldsymbol{G}(x)dx.$$

Since an MRA consists of a nested sequence of spaces

$$... \subset V_{j-1} \subset V_j \subset V_{j+1} \subset ... \qquad , j \in \mathbb{N}$$

we can define the relative orthogonal complements $W_j := V_{j-1} \oplus V_j$. If $\boldsymbol{F} \in V_j$ and $\boldsymbol{G} \in W_j$ the orthogonality condition can be rewritten as:

$$\langle \boldsymbol{F}, \boldsymbol{G}\rangle = \boldsymbol{0}.$$

If $\boldsymbol{F}$ is an orthonormal $\boldsymbol{A}$-refinable matrix function, i.e., $\boldsymbol{F} = \boldsymbol{F} * \boldsymbol{A}(2\cdot)$ then the bi-infinite block matrix

$$A = [\boldsymbol{A}(j - 2k) : j,\, k \in \mathbb{Z}]$$

 ISBN: 978-84-614-6167-7

satisfies:

$$\boldsymbol{A}^T\boldsymbol{A} = 2\boldsymbol{I} \tag{2}$$

A function $\boldsymbol{G} \in V_1$ is called *wavelet* for MRA if

$$W_j = \left\{\boldsymbol{G} * \boldsymbol{c}(2^j\cdot) : \boldsymbol{c} \in \ell_2(\mathbb{Z})^{\mathbb{Z}_r}\right\},\ j \in \mathbb{N}.$$

and it is called an *orthonormal wavelet* if $\boldsymbol{G}$ is moreover orthonormal. Suppose that $\boldsymbol{F}$ is an orthonormal $\boldsymbol{A}$-refinable matrix function, i.e. $\boldsymbol{A}^T\boldsymbol{A} = 2\boldsymbol{I}$, then there exists a bi-infinite block matrix

$$B = [\boldsymbol{B}(j-2k) : j,\ k \in \mathbb{Z}]$$

where

$$\mathbf{B} = \left(\mathbf{B}(k) \in \mathbb{R}^{\mathbb{Z}_r\times\mathbb{Z}_r} : k \in \mathbb{Z}\right) \in \ell_{00}^{\mathbb{Z}_r\times\mathbb{Z}_r}$$

such that $B^TA = 0$ and $B^TB = 2I$. Moreover this matrix B satisfies $AA^T + BB^T = 2I$. The matrix B is the key to the wavelet construction, in fact

$$\boldsymbol{G} := \boldsymbol{F} * \boldsymbol{B}(2\cdot) \in V_1.$$

Let ℓ_{00} be the vector space of all *finitely supported* matrix-valued and

$$[\,\cdot\,]:\quad \ell(\mathbb{Z})^{\mathbb{Z}_r} \quad\longrightarrow\quad \ell(\mathbb{Z})$$

be a function mapping a vector-valued sequence $\boldsymbol{c} \in \ell(\mathbb{Z})^{\mathbb{Z}_r}$ to $c = [\boldsymbol{c}]$ via

$$c(rj+k) = \boldsymbol{c}(j)_k,\ j \in \mathbb{Z},\ k \in \mathbb{Z}_r.$$

We define the *subdivision operator* $S_{\boldsymbol{A}}$ and the *decimation operator* $S^*_{\boldsymbol{A}}$, for any $\boldsymbol{A} \in \ell_{00}^{\mathbb{Z}\times\mathbb{Z}_r}$, to introduce a *fast wavelet transform*. Let $A \in \mathbb{R}^{\mathbb{Z}\times\mathbb{Z}}$ be a bi-infinite matrix with the block representation $A = [\boldsymbol{A}(j-2k) : j,k \in \mathbb{Z}]$ and let $\boldsymbol{c} \in \ell_2(\mathbb{Z})^{\mathbb{Z}_r}$ be the vector-valued input signal:

***Subdivision operator*:**

$$S_{\boldsymbol{A}}\boldsymbol{c} = \sum_{k\in\mathbb{Z}} \boldsymbol{A}(\cdot - 2k)\boldsymbol{c}(k)$$

such that $[S_{\boldsymbol{A}}\boldsymbol{c}] = A[\boldsymbol{c}]$.

***Decimation operator*:**

$$S^*_{\boldsymbol{A}}\boldsymbol{c} = \sum_{j\in\mathbb{Z}} \boldsymbol{A}^T(j - 2\cdot)\boldsymbol{c}(j)$$

which has the property that $[S^*_{\boldsymbol{A}}\boldsymbol{c}] = A^T[\boldsymbol{c}]$.

We can define the decomposition part of the pyramid scheme by setting

$$\boldsymbol{c}_{j-1} = \frac{1}{2}S^*_{\boldsymbol{A}}\boldsymbol{c}_j \quad \text{and} \quad \boldsymbol{d}_{j-1} = \frac{1}{2}S^*_{\boldsymbol{B}}\boldsymbol{c}_j \tag{3}$$

 ISBN: 978-84-614-6167-7

$C^L =$

$C^2_{1,1}$	$C^2_{2,1}$	$C^1_{2,1}$	$C^2_{1,2}$	$C^2_{2,2}$	$C^1_{2,2}$	$C^2_{1,3}$	$C^2_{2,3}$	$C^1_{2,3}$
$C^2_{3,1}$	$C^2_{4,1}$		$C^2_{3,2}$	$C^2_{4,2}$		$C^2_{3,3}$	$C^2_{4,3}$	
$C^1_{3,1}$		$C^1_{4,1}$	$C^1_{3,2}$		$C^1_{4,2}$	$C^1_{3,3}$		$C^1_{4,3}$

Figure 1: C^L, $L = 1, l$, with $l = 2$ represents 2-d DMCWT using filter with multiplicity $r = 3$

and the reconstruction part as

$$\boldsymbol{c}_j = S_{\boldsymbol{A}}\boldsymbol{c}_{j-1} + S_{\boldsymbol{B}}\boldsymbol{d}_{j-1} \tag{4}$$

where A and B are the bi-infinite matrices related to scaling and wavelet functions, respectively. The formula 3 is called *Discrete Multichannel Wavelet Transform* (DMCWT). The formula 4 is called *Inverse Discrete Multichannel Wavelet Transform* (IDMCWT).

The multichannel wavelet decomposition can be generalized to the 2D case and interpreted as a signal decomposition in a set of independent, spatially oriented frequency channels. For the sake of simplicity let $C^0 \in \mathbb{R}^{n\times m\times 3}$ be a three-dimensional matrix with n rows and m columns representing the color image. The matrix $C^1 \in \mathbb{R}^{n'\times m'\times 3}$, with $n' = n/2$ and $m' = m/2$, is obtained applying level $l = 1$ of 2-d DMCWT using filter with multiplicity $r = 3$. For any matrix plane, $j = 1, 2, 3$, four coefficient sub-bands form C^1:

$$C^1 = \begin{bmatrix} C^1_{1,j} & C^1_{2,j} \\ C^1_{3,j} & C^1_{4,j} \end{bmatrix} \tag{5}$$

where:

- the three-dimensional matrix $C^1_{1,j}$ represents the lowest frequencies;
- the three-dimensional matrix $C^1_{2,j}$ represents the vertical high frequencies and horizontal low frequencies (horizontal edges);
- the three-dimensional matrix $C^1_{3,j}$ the horizontal high frequencies and vertical low frequencies (vertical edges);
- the three-dimensional matrix $C^1_{4,j}$ represents high frequencies in both horizontal and vertical directions (the corners).

In Figure 1 three matrix planes C^k, for two levels of decomposition, $k = 1, 2$ are shown. $C^k_{\theta,j}$, $\theta = 1, 2, 3, 4$ and $j = 1, 2, 3$ represent the sub-bands the decomposition at level k.

 ISBN: 978-84-614-6167-7

3 The watermarking algorithm MCWM 1.0

In this section we explain MCWM 1.0, our watermarking algorithm based on multichannel wavelets with multiplicity $r = 3$ for digital color images belonging to web virtual image galleries. It hides an *invisible* mark into an image and detects it by comparing the original and the watermarked images. Therefore MCWM 1.0 is a *private* algorithm. The color images are modeled in Hue Saturation Value (HSV) color model in according to the Human Visual System study, which indicates that the human eye is less sensitive to noise in those areas of the image where the brightness is high or low. The watermarking signal is inserted in the coefficients of high frequency sub-bands of the Discrete MultiChannel Wavelet Transform (DMCWT) in the Value. In the *embedding phase* our efficient watermarking algorithm balances the properties of imperceptibly and robustness of the mark in an accurate way. In the *detection process* the input watermarked image could be different in pixel and dimension from the watermarked image output of the *embedding process*. In fact it could have been modified by *attacks* changing its dimension as *cut* or *resize*. Therefore we have introduced a new step called *alignment* step in which the dimension of original and watermarked image are to each other. In *detection process* the *Neyman-Pearson statistic criterion* is used to compute the correlation between the DMCWT coefficients of the watermarked image and the watermark [12]. This criterion allows to determine a *detection threshold* minimizing the probability of missing detection respect to the given probability of *false alarm*.
We have introduced two thresholds to compare this correlation minimizing the false positive case and obtaining a robust algorithm. The experimentation has been accomplished on images, in high and low resolution, building a real and commercial database including CEI database. The images have been subjected to several attacks.

3.1 Embedding process

The embedding process is composed of the following steps:

Step E1. Let $I \in \mathbb{R}^{m' \times n' \times 3}$ be the matrix corresponding to the original color image in HSV color model representation. It is necessary that m' and n' are divisible by 2^l to compute the l levels of the 2D DMCWT. If these conditions are not verified we compute a new matrix I from I'. The dimension of I are $m \times n \times 3$, such that:

$$m = \begin{cases} m' & \text{if } mod(m', 2^l) = 0 \\ 2^l m_l & \text{otherwise} \end{cases} \qquad n = \begin{cases} n' & \text{if } mod(n', 2^l) = 0 \\ 2^l n_l & \text{otherwise} \end{cases}$$

where $m_l = \left\lfloor \frac{m'}{2^l} \right\rfloor$ and $n_l = \left\lfloor \frac{n'}{2^l} \right\rfloor$ represent the integer part of $\frac{m'}{2^l}$ and $\frac{n'}{2^l}$. The new matrix C is obtained cutting the first and the last $\frac{(m'-m)}{2}$ rows and the first and the last $\frac{(n'-n)}{2}$ columns of each plane of I.

Step E2. C is decomposed l-times by means DMCWT. We use the new class of full rank filter built in [13]. In particular Cotronei-Puccio-Vocaturo multichannel wavelet with multiplicity $r = 3$ has been applied [14]. The matrix $C^l \in \mathbb{R}^{n \times m \times 3}$

 ISBN: 978-84-614-6167-7

is the decomposition of the image I. We extract the decomposed matrix C_3^l that represents the Value component of the image. In C_3^l we consider the three sub-matrices of high frequencies relative to the last level l of the decomposition: $C_{\theta,3}^l = \left\{c_{i,j,3}^{l,\theta}\right\}_{i=1,..,m_l,\, j=1,..,n_l}$ with $\theta = 2,3,4$.

Step E3. The watermark is embedded by means of the following casting formula:

$$\tilde{C}_{\theta,3}^l = C_{\theta,3}^l + \omega \cdot S_\theta \ , \quad \theta = 2,3,4 \tag{6}$$

where:

- the parameter $\omega \in \mathbb{R}$ represents the watermark strength:

$$\omega = \underset{j=1,..,n_l}{\texttt{std}} \ \underset{\theta=2,3,4}{\texttt{max}} \left\{ a_{\theta,j} \in \mathbb{R} : a_{\theta,j} = \underset{i=1,..,m_l}{\texttt{mean}} \, c_{i,j,3}^{\theta,l} \right\} \tag{7}$$

  where `std` and `mean` respectively the standard deviation and the mean.

- $S_\theta \in \mathbb{Z}^{m_l \times n_l}$, with $\theta = 2,3,4$ are three weight matrices whose value of elements $s_{i,j}^\theta$, depend on the respective coefficients of the sub-matrices $C_{\theta,3}^l$:

$$s_{i,j}^\theta = \begin{cases} 0 & \text{if} \quad |c_{i,j,3}^{l,\theta}| < \omega \\ 1 & \text{if} \quad c_{i,j,3}^{l,\theta} > T_S \\ -1 & \text{otherwise} \end{cases} \qquad i = 1,..,m_l, \ j = 1,..,n_l \tag{8}$$

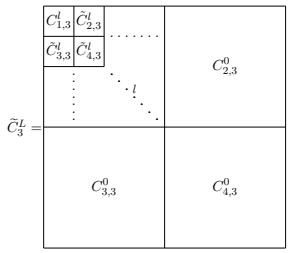
- the threshold T_S is computed such that the mean of not null entries of the block matrix $[S_2 \ S_3 \ S_4]$ is zero:

$$T_S = \texttt{median sort} \left\{ c_{i,j,3}^{l,\theta} \in C_\theta^l \, | c_{i,j,3}^{l,\theta} \notin [-\omega, \omega], \ \ i = 1,..,m_l, j = 1,..,n_l, \theta = 2,3,4 \right\} \tag{9}$$

  where `median` and `sort` are the operators that respectively, find the median value and sort the vector elements.

The parameters ω and T_S and the matrices S_2, S_3, S_4 have been computed to have a good arrangement between *robustness* property of algorithm and *imperceptibility* property of watermark. It is important to emphasize that ω guarantees the robustness property while S_2, S_3, S_4 guarantee the imperceptibility property.

Step E4. We substitute the sub-matrices $C_{\theta,3}^l$ with $\tilde{C}_{\theta,3}^l$, $\theta = 2,3,4$ as in 6 obtaining $\tilde{C}^L$, whose value plane is $\tilde{C}_3^L$:



 ISBN: 978-84-614-6167-7

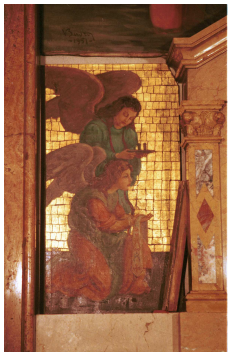


Figure 2: Original image

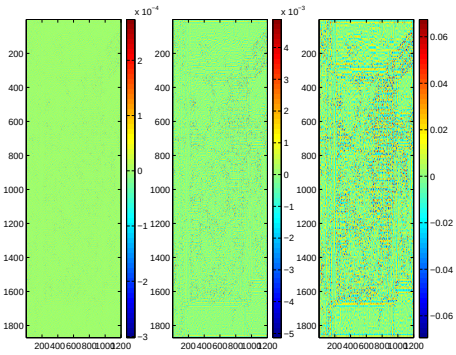


Figure 3: Difference between the original image I and the reconstructed one $\tilde{I}$

We apply l-times the Inverse of DMCWT (IDMCWT) to $\tilde{C}^L$ for reconstructing image $\tilde{I} \in \mathbb{R}^{n\times m\times 3}$. Note that, the watermark signal appears well-distributed in all color planes after this step. For sake of simplicity we explain this feature of our algorithm by an example. The original image is represented in Figure 2. In Figure 3 the differences between $\tilde{I}$ and I are presented. The watermarking signal has been distributed also Hue and Saturation planes by means IDCMT2. We have achieved this important result using multichannel wavelets that are vectorial functions.

Step E5. The inverse process of **Step E1.** is applied to adjust the dimensions and obtain the watermarked image $\tilde{I} \in \mathbb{R}^{n\times m\times 3}$ from $\tilde{I} \in \mathbb{R}^{n\times m\times 3}$.

3.2 Detection process

The detection process is composed of these steps:

Step D1. The presented algorithm is private, so the original image is necessary in this process. Let I' be the original image and $\tilde{I}^\star$ be the original and watermarked image possibly modified by geometrical attacks. Their dimensions are respectively

 ISBN: 978-84-614-6167-7

$m' \times n' \times 3$ and $\tilde{m}^\star \times \tilde{n}^\star \times 3$. A check procedure on matrices dimensions is applied to verify if $m' = \tilde{m}^\star$ and $n' = \tilde{n}^\star$. One of these conditions can be false because the watermarked image could have been modified by geometric attacks, such as resizing or cutting. Then the original and watermarked images are aligned by executing the following operations:

- take out the central block B of dimension 8×8 from I';
- compare B with some blocks $\tilde{B}^\star_{ij}$ extracted from $\tilde{I}^\star$ by means the *Mean Square Error* function.
- resize the matrix I' from its original dimension to $\tilde{m}^\star \times \tilde{n}^\star \times 3$ according to the position of the block $\tilde{B}^\star_{ij}$ congruent to B.

Step D2. It is equivalent to **Step E1**. I' and $\tilde{I}^\star$ are processed in order to obtain I and $\tilde{I}$ of dimension $m' \times n' \times 3$.

Step D3. I and $\tilde{I}$ are decomposed by means of the multichannel wavelet. As in the previous process the following matrices are computed: $C^l_\theta = \left\{c^{l,\theta}_{i,j}\right\}_{i=1,..,m_l,\, j=1,..,n_l}$ relative to original image and $\tilde{C}^l_\theta = \left\{\tilde{c}^{l,\theta}_{i,j}\right\}_{i=1,..,m_l,\, j=1,..,n_l}$ relative to the watermarked image, with $\theta = 2, 3, 4$.

Step D4. To detect the watermarking signal the correlation between the watermarked/original coefficients and watermarked signal are computed:

$$\rho = \frac{1}{z_l} \sum_{\theta=2,3,4} \left[\sum_{i=1}^{m_l} \sum_{j=1}^{n_l} S^\theta_{i,j} (\tilde{c}^{l,\theta}_{i,j,3} - c^{l,\theta}_{i,j,3}) \right] \tag{10}$$

where $z_l \in \mathbb{N}$ is the total number of the non-zero elements of the matrices S_θ with $\theta = 2, 3, 4$.
The ρ value is compared with two thresholds: T_1 and T_ρ. This latter is estimated minimizing the probability of false alarm, according to the Neyman-Pearson statistic criterion [15]:

$$P_f \leq \frac{1}{2} \texttt{erfc} \left(\frac{T_\rho}{\sqrt{2\sigma^2}} \right) \tag{11}$$

where $\texttt{erfc}(x)$ is the complementary error function and σ represents the standard deviation.
In order to refine the detect interval of the watermark, we have introduced a novel threshold T_1. To compute it, the C^θ_l-entries are sorted and inserted into a vector v:

$$T_1 = \underset{i=1,..,z_l}{\texttt{mean}}\, v_i + \omega \tag{12}$$

Then the detection of watermarking signal is the following:

- if $|\rho| \leq T_1$ and $|\rho| \geq T_\rho$ then the watermark exists;
- otherwise the watermark does not exist.

 ISBN: 978-84-614-6167-7

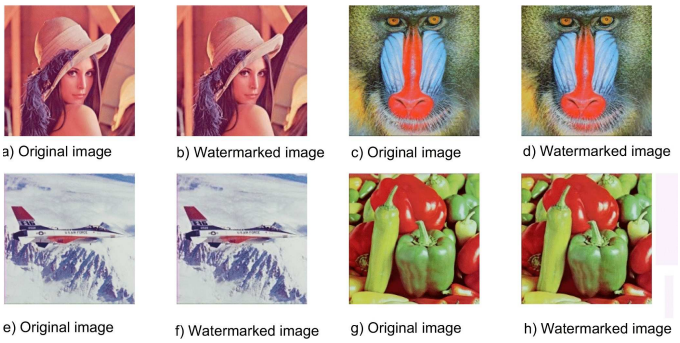


Figure 4: Example of original and watermarked images

4 Numerical results

In this section we describe the experimental tests executed to analyze the properties of our watermarking algorithm. The invisibility property has been observed by finding out the Peak Signal to Noise Ratio (PSNR) and the Weight Peak Signal to Noise Ratio (WPSNR). In Table 1 we have inserted the results of these measures obtaining on four classic test images Lena, Baboon, Airplane and Peppers. Figure 4 provides a visual quality comparison between the original and the watermarked images.

The second experimental test on the algorithm has the aim to verify the robustness of

Table 1: PSNR and WPSNR of the watermarked image

	Lena	*Baboon*	*Airplane*	*Peppers*
PSNR	+35.74	+32.00	+33.77	+34.90
WPSNR	+52.61	+49.34	+49.70	+53.65

watermarking algorithm. To estimate it we have applied several geometric attacks (i.e. resize, distortion, StirMark) to the watermarked image before executing the detection process. This important property has been tested on images (79), in high (44) and low resolution (35), belonging the CEI's database (http://www.chiesacattolica.it/beweb). The typology of the images belonging to this inventory includes a wide range of historical artistic assets that can be ordered according to objects and scenes: images with typical objects of the sacerdotal equipment, vestments, altar, cloths and holy vessels, equipment of churches (i.e. crosses, confessional, icons); images with scenes from frescoes and paintings; images of statues. The differences of historical-artistic images considered during the experimental tests have given a diversified and complex survey. We have applied the following attacks: better blurring; distortion (deform, zigzag); rotation; adding noise; cutting of image part; resize (+1%); resize (-1%); StirMark. The results are showed in Table 2. In the fist row we have inserted the applied attacks. In the second, third and fourth rows there are showed the number of low and high resolution images and the total number of images which inserted watermark has been detected after the attack. The last row represents the percentage of *success* obtained

 ISBN: 978-84-614-6167-7

by attacks our watermarked images.

Image	Blur	Distortion		Rotation	Noise	Cut	Resize -1%	Resize +1%	StirMark
		Deform	Zigzag						
Low res.	43	43	42	44	42	42	43	42	42
High res.	31	35	29	34	29	35	35	32	35
Tot	74	78	77	78	71	77	78	74	79
Perc.	94%	99%	97%	99%	90%	97%	99%	94%	100%

Table 2: Results of the attacks applied on the watermarked image before detecting the watermark by our algorithm. The used images to test the robustness belonging to CEI's database

This experimentation has shown that our algorithm is robust against geometrical attacks and filtering with a ratio more than 90% on 79 images. Furthermore it is very important to enphatize that the percentage of success respect to StirMark attack is 100% against the results relative to digital watermarking commercial software in [16]. We have developed a watermarking algorithm MCWM 1.0 based on multichannel wavelet functions that is private, invisible and robust. We have obtaned a improvement of the results respect to our previous algorithm [6].

4.1 Acknowledgements

This work was supported by: PRA Projects of University of Messina and GNCS of INdAM.

References

[1] S. Voloshynovskiy, S. Pereira, T. Pun, J. Eggers, J. Su, Attacks on digital watermarks: Classification, estimation-based attacks and benchmarks, IEEE Communications Magazine 39 (2001) 118–126.

[2] F. A. P. Peticolas, R. J. Anderson, M. G. Kuhn, Attacks on copyright marking system, Second workshop on information hiding, vol 1525 pp. 218-238 of Lecture Note in Computer Science,Portland, Oregon, USA.

[3] D.P.Mukheriee, S. Maitra, S. Acton, Spatial domain digital watermarking of multimedia objects for buyer authentication, IEEE Transactions on Multimedia 6 (1).

[4] M. Barni, F. Bartolini, V. Cappellini, A. Piva, A dct-domain system for robust image watermarking, Signal Processing 66 (3) (1998) 357–372.

[5] X. Li, X. Xue, Improved robust watermarking in dct domain for color images, in: AINA '04: Proceedings of the 18th International Conference on Advanced Information Networking and Applications, IEEE Computer Society, Washington, DC, USA, 2004, p. 53.

[6] S. Agreste, G. Andaloro, D. Prestipino, L. Puccio, An image adaptive, wavelet-based watermarking of digital images, J. Comput. Appl. Math. 210 (1-2) (2007) 13–21. doi:http://dx.doi.org/10.1016/j.cam.2006.10.087.

[7] S. Agreste, G. Andaloro, A new approach to pre-processing digital image for wavelet-based watermark, J. Comput. Appl. Math. 221 (2) (2008) 274–283. doi:http://dx.doi.org/10.1016/j.cam.2007.10.057.

[8] Z. Dawei, C. Guanrong, L. Wenbo, A chaos-based robust wavelet-domain watermarking algorithm, Elvier, Chaos, Solitons and Fractals 22 22 (2004) 47–54.

[9] M. Barni, F. Bartolini, A. Piva, Improved wavelet-based watermarking through pixel wisemasking, IEEE Transactions, Image Processing 10 (5) (2001) 783–791.

[10] S. Bacchelli, M. Cotronei, T.Sauer, Multifilters with and without prefilters, BIT 42 2 (1998) 231–261.

[11] S. Bacchelli, M. Cotronei, T. Sauer, Wavelets for multichannel signals, Adv. Appl. Math. 29 (4) (2002) 581–598. doi:http://dx.doi.org/10.1016/S0196-8858(02)00033-7.

[12] C. Scott, The neyman-pearson criterion, connexions Web site. http://cnx.org/content/m11548/1.2/ (2004).

[13] S. Agreste, A. Vocaturo, A new class of full rank filters in the context of digital color image processing., Vol. Stereology and Image Analysis. Ecs10 - Proceedings of the 10th European Congress of ISS of MIRIAM Project Series, ESCULAPIO Pub. Co, 2009, pp. 193–198.

[14] M. Cotronei, L. Puccio, A. Vocaturo, On the application of full rank filters to color image processing, in: F. Pistella, R. M. Spitaleri (Eds.), Proceeding of MASCOT 05, Vol. 10 of IMACS, 2006, pp. 125–129.

[15] M. Barni, F. Bartolini, V. Cappellini, A. Lippi, A. Piva, A DWT-based technique for spatio-frequency masking of digital signatures, in: Proceedings of the SPIE/IS&T International 20 Conference on Security and Watermarking of Multimedia Contents, Vol. 3657, 1999, pp. 31–39.
URL `citeseer.ist.psu.edu/barni99dwtbased.html`

[16] M.Kutter, F. Petitcolas, Fair evaluation methods for image watermarking systems, Electronic Imaging 9 (4) (2000) 445–455.

 ISBN: 978-84-614-6167-7

Proceedings of the 11th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2011
26–30 June 2011.

Improving Newton's Method for Nonlinear Optimization Problems in Several Variables

Kamel Al-Khaled[1], Ameen Alawneh[1,2] and Nadia Al-Rashaideh[1]

[1] *Department of Mathematics and Statistics , Jordan University of Science and Technology, IRBID 22110, JORDAN*

[2] *Qatar University, Department of Mathematics, Statistics and Physics, On Sabbatical Leave from Jordan University of Science and Technology*

emails: kamel@just.edu.jo, ameen@just.edu.jo,

Abstract

In this paper, Adomian decomposition method is used to improve Newton's method for minimizing functions of several variables. Numerical solutions are calculated in the form of convergent power series with easily computable components. The significant of this work is that the improvement of Newton's method reduces computations, improves the accuracy, and yields fast convergence.

Key words: Adomian decomposition method, Newton's method, Nonlinear Optimization.
MSC 2000: AMS codes (90C05, 65K05, 34K28)

1 Introduction

One of the most important problems in Mathematics is solving systems of nonlinear equations. These systems arise often from the numerical solutions of mathematical models in real life; especially in science and engineering [1]. These systems also arise in discretization of boundary value problems by finite difference equations, or finite element methods. The problem of solving systems of nonlinear equations can be seen in finding the solutions of optimization problems. It is not an easy task to locate all critical points of a real valued function $f(\mathbf{x})$ on R^n by attempting to find exact solutions of $\nabla f(\mathbf{x}) = \mathbf{0}$. Instead, iterative methods are used to search out the extremum by means of an approximating sequence whose points are generated in some computationally acceptable way from $f(\mathbf{x})$. More efficient methods for solving systems of nonlinear equations are continuously being sought. Some of these methods depend on variations of Newton's approach [2], and spectral method [3]. Adomian decomposition method (ADM) was first introduced by Adomian [4, 5], since the beginning of the 1980's, and

 ISBN: 978-84-614-6167-7

its used for solving a wide range of problems. This new iteration method has proven rather successful in dealing with both linear as well as nonlinear problems, as it yields analytical solutions and offer certain advantages over standard numerical methods [5]. ADM was used to solve a wide range of physical problems. An advantage of this method is: it can provide analytical approximation or an approximated solution to a rather wide class of nonlinear (and stochastic) equation without linearization, perturbation, closure approximation, or discretization methods [1]. Unlike the common method, i.e., weak nonlinearity and small perturbation which change the physics of the problem, ADM gives the approximated solution of the problem without any simplification. Thus, its results are more realistic [6]. In [7], the author has modified ADM to approximate the root of a nonlinear equation in one variable. Abbasbandy [8] introduced a powerful improvement of Newton-Raphson method by ADM for solving nonlinear equations. Recently, ADM is applied to solve systems of nonlinear equations [9, 10]. Authors in [12] proposed a modifications of ADM and used to obtain solutions of systems of nonlinear equations. In [13] the author introduced an efficient extension of Newton's method by modified ADM. The results suggest that the use of extension techniques introduced a promising tool for solving system of nonlinear equations. Abbaoui and Cherrualt [14] applied this techniques to solve the equation $f(\mathbf{x}) = \mathbf{0}$ and proved the convergence of the series solution.

In this paper, we will focus on using the same technique mentioned in [13], and setup an algorithm using Newton's method to solve unconstrained optimization problems in three variables by ADM. The proposed technique, as mentioned in [11] will be used to solve two types of functions: exponential and logarithmic functions.

2 Preliminary Results

In this section, we introduce some basic materials that will be useful for this paper.

2.1 Nonlinear Programming

Optimization in mathematics referes to the study of problems which one seeks to minimize or maximize a real-valued function by systematically choosing the values of real or integer variables from within the feasible region. Here, we bring out on some topics in nonlinear programming that will be used in solving optimization problems in section 4.

Let $f : R^n \to R$ be an objective function, and X is the feasible region, then the general form of nonlinear optimization problem is:

$$\min_{x \in X} f(\mathbf{x}), \;\; g_i(\mathbf{x}) \leq 0, \; i = 1, 2, ..., m, \;\; h_i(\mathbf{x}) = 0, \; i = 1, 2, ..., \ell. \tag{1}$$

The constraints g_i and h_i are real-valued functions. The feasible solution is a solution $(x_1, x_2, ..., x_n)$ which satisfies all components. The feasible region is the set of all possible feasible solutions.

 ISBN: 978-84-614-6167-7

2.2 Newton's Method

Newton's method is an efficient algorithm for finding approximations to zeros of a real-valued function. Consider the equation $f(x) = 0$, suppose that α is a root of f, and f is a continuously differentiable function on an interval containing α. Newton-Raphson method is given by

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)},\ n = 0, 1, ... \tag{2}$$

which started with the initial guess x_0. It is well known that Newton's method is quadratically convergent. Newton's method is used to find solution for systems of nonlinear equations, as in the following algorithm:

Algorithm: Suppose that g is a differentiable function of n variables with values in R^n, and that $x^{(0)} \in R^n$. Then the sequence $\{x^{(k)}\}$ generated by Newton's method for solving $g(x) = 0$ is defined by

$$x^{(k+1)} = x^{(k)} - \left[\nabla g(x^{(k)}\right]^{-1} g(x^{(k)}),\ k \geq 0 \tag{3}$$

Newton's method can also be used to find a minimum, or maximum value for a real-valued function in one or more variables, as in the following algorithm:

Algorithm: Suppose that $f(x)$ is twice continuously differentiable real-valued function of n variables, and suppose that $x^{(0)} \in R^n$. Then the sequence $\{x^{(k)}\}$ generated by Newton's method for minimizing $f(x)$ is defined by

$$x^{(k+1)} = x^{(k)} - \left[Hf(x^{(k)})\right]^{-1} \nabla f(x^{(k)}),\ k \geq 0 \tag{4}$$

where Hf is the Hessian of f.

2.3 Adomian Decomposition Method

In this subsection, following [9] we present an algorithm based on ADM, which is used to find an approximate solution for systems of nonlinear algebraic equations of the form

$$f(\mathbf{x}) = 0 \tag{5}$$

where $f(\mathbf{x}) = (f_1(x), f_2(x), ..., f_m(x))^T$, $\mathbf{x} = (x_1, x_2, ..., x_m)^T$, and $f : R^m \to R^m$ is a nonlinear mapping with the following properties:

- There exists an $\mathbf{x} \in R^m$ with $f(\mathbf{x}) = 0$.
- f is continuously differentiable in a neighborhood of $\mathbf{x}$.
- $f'(\mathbf{x})$ (Jacobian of f) is non-singular.

The ith component in (5) $f_i(x_1, ..., x_m)$ can be written in the form

$$x_i = \alpha + g_i(x_1, ..., x_m) \tag{6}$$

 ISBN: 978-84-614-6167-7

where $g_i,\ i = 1, 2, ..., m$ are nonlinear functions. The ADM consists of calculating the solution in the series form

$$x_i = \sum_{j=0}^{\infty} x_{i,j},\ i = 1, ..., m, \tag{7}$$

where the components $x_{1,0}, x_{2,0}, ...$ are usually determined recursively. From (6) we get the following recursive relation:

$$\begin{aligned} x_{i,0} &= \alpha, \\ x_{i,k+1} &= A_{i,k},\ k \geq 0. \end{aligned} \tag{8}$$

where $A_{i,k}$ are Adomian polynomials depending on $x_{1,0}, ..., x_{1,n}; x_{2,0}, ..., x_{2,n}; x_{m,1}, ..., x_{m,n}$. Moreover, the nonlinear terms $g_i(x_1, ..., x_m)$, in equation (6) can be expressed in terms of Adomian polynomial $A_{i,j}$ as follows:

$$g_i(x_1, ..., x_m) = \sum_{j=0}^{\infty} \lambda^j A_{i,j},\ \ i = 1, ..., m \tag{9}$$

Adomian polynomials $A_{i,j}$ can be formally obtained from the relation

$$A_{i,n}(x_{1,0}, ..., x_{1,n}; x_{m,1}, ..., x_{m,n}) = \frac{1}{n!}\frac{d^n}{d\lambda^n} g_i\Big[\sum_{j=0}^{\infty} \lambda^j x_{1,j}, ..., \sum_{j=0}^{\infty} \lambda^j x_{n,j}\Big]_{\lambda=0}, i = 1, 2, ..., m. \tag{10}$$

Upon substituting (7) and (9) into (6) yields

$$\sum_{j=0}^{\infty} x_{i,j} = \alpha + \sum_{j=0}^{\infty} \lambda^j A_{i,j},\ \ i = 1, 2, ..., m. \tag{11}$$

In practice it is difficult to compute the terms in the series (7). However, the $kth-$ term approximate solution $x_i,\ i = 1, ..., m$, will be determined by only evaluating finite number of terms as

$$S_{i,k} = \sum_{m=0}^{k-1} x_{i,m},\ k \geq 1,\ i = 1, ..., m \tag{12}$$

and the approximate solution $x_i,\ i = 1, ..., m$ of the system (5) is given by

$$x_i = \lim_{k \to \infty} S_{i,k}, \tag{13}$$

which is usually converges to an accurate solution [9].

3 Extended Newton's Method

Newton's method is a well-known algorithm for finding roots of equations in one or more dimensions. Here, we use the same procedure mentioned in [13], where we go one dimension further than [13], also we set up our algorithm to find local minima of

 ISBN: 978-84-614-6167-7

functions in three variables. It can also be used to find local maxima of functions. Consider a real-valued function of three-variables $f : R^3 \to R$, where (α, β, γ) be the minimizing of f, then (α, β, γ) must be a critical point of f. For solving this system of nonlinear equations, suppose f_x, f_y and f_z are continuously differentiable in the open convex set $D \subset R^3$ including (α, β, γ). Suppose the inverse of the Hessian matrix of f at (α, β, γ) exists and bounded, i.e., $H(f(x, y, z))^{-1}$ exist with $\| H(f(x, y, z))^{-1} \| \leq \eta$, for $\eta > 0$. Using Taylor's expansion of f_x, f_y, f_z near (x, y, z), say at $(x-h, y-k, z-w)$ for some small values h, k and w, we obtain

$$\begin{pmatrix} h \\ k \\ w \end{pmatrix} = H(f(x, y, z))^{-1} \begin{pmatrix} f_x(x, y, z) \\ f_y(x, y, z) \\ f_z(x, y, z) \end{pmatrix} \tag{14}$$

Adding the column vector (x, y, z) to both sides of the above equation, Newton's method for function minimization is then given by the following iterative formula

$$\begin{pmatrix} x_{n+1} \\ y_{n+1} \\ z_{n+1} \end{pmatrix} = \begin{pmatrix} x_n \\ y_n \\ z_n \end{pmatrix} - H(f(x, y, z))^{-1} \begin{pmatrix} f_x(x_n, y_n, z_n) \\ f_y(x_n, y_n, z_n) \\ f_z(x_n, y_n, z_n) \end{pmatrix} \tag{15}$$

with initial guess (x_0, y_0, z_0). It is possible to go one step further, and write Taylor's expansion of f_x, f_y and f_z to a higher order, we get

$$\begin{pmatrix} h \\ k \\ w \end{pmatrix} = \begin{pmatrix} c_1 \\ c_2 \\ c_3 \end{pmatrix} + N \begin{pmatrix} h \\ k \\ w \end{pmatrix} = \begin{pmatrix} c_1 \\ c_2 \\ c_3 \end{pmatrix} + \begin{pmatrix} N_1(h, k, w) \\ N_2(h, k, w) \\ N_3(h, k, w) \end{pmatrix} \tag{16}$$

where $c_1 = \frac{f_x}{f_{xx}}(x, y, z), c_2 = \frac{f_y}{f_{yy}}(x, y, z)$ and $c_3 = \frac{f_z}{f_{zz}}(x, y, z)$ are constants, and N is a vector quadratic polynomial. For approximating h, k and w, we can apply the multi-variable Adomain decomposition [8], which consists of representing

$$h = \sum_{n=0}^{\infty} h_n, \ k = \sum_{n=0}^{\infty} k_n, \ w = \sum_{n=0}^{\infty} w_n \tag{17}$$

and the nonlinear functions are decomposed as

$$N_i(h, k, w) = \sum_{n=0}^{\infty} A_{i,n}(h_0, h_1, ..., h_n; k_0, k_1, ..., k_n; w_0, w_1, ..., w_n) \tag{18}$$

where the $A'_{i,n}s$ are Adomian polynomials given by

$$A_{i,n} = \frac{1}{n!} \frac{d^n}{d\lambda^n} \Big[N_i \Big(\sum_{j=0}^{\infty} \lambda^j h_j, \sum_{j=0}^{\infty} \lambda^j k_j; \sum_{j=0}^{\infty} \lambda^j w_j \Big) \Big]_{\lambda=0}, i = 1, 2, 3; \quad n = 0, 1, 2, ...$$

Substituting equations (17), (18) into (16) yields

 ISBN: 978-84-614-6167-7

$h_0 = c_1, h_{n+1} = A_{1,n}, k_0 = c_2, k_{n+1} = A_{2,n}$ and $w_0 = c_3, w_{n+1} = A_{3,n}, n = 0, 1, 2, ...$ For $i = 1, 2, 3$, we have $A_{i,0} = N_i(h_0, k_0, w_0)$ and

$$A_{i,n} = \sum_{\Omega} \frac{h_1^{p_1}}{p_1!} \cdots \frac{h_n^{p_n}}{p_n!} \frac{k_1^{q_1}}{q_1!} \cdots \frac{k_n^{p_1}}{q_n!} \frac{w_1^{r_1}}{p_n!} \cdots \frac{w_n^{r_n}}{r_n!} \frac{\partial^{\Omega_1+\Omega_2+\Omega_3}}{\partial h^{\Omega_1} \partial k^{\Omega_2} \partial w^{\Omega_3}} N_i(h_0, k_0, w_0), n \neq 0$$

where Ω stands for

$$(p_1 + 2p_2 + ... + np_n) + (q_1 + 2q_2 + ... + nq_n) + (r_1 + 2r_2 + ... + nr_n) = n$$

and, $\Omega_1 = p_1 + p_2 + ... + p_n, \Omega_2 = q_1 + q_2 + ... + q_n$ and $\Omega_3 = r_1 + r_2 + ... + r_n$. However, in practice it is difficult to compute all terms of the series (16), instead the $(M+1)-$ term approximate solutions h, k and w, will be determined by only evaluating finite number of terms as:

$$H_M = h_0 + h_1 + ... + h_M = h_0 + A_{1,0} + A_{1,1} + ... + A_{1,M-1} \tag{19}$$

$$K_M = k_0 + k_1 + ... + k_M = k_0 + A_{2,0} + A_{2,1} + ... + A_{2,M-1} \tag{20}$$

and

$$W_M = w_0 + w_1 + ... + w_M = w_0 + A_{3,0} + A_{3,1} + ... + A_{3,M-1} \tag{21}$$

Since the series converges very rapidly, then the above equations can serves as a practical solution in each iteration. Now, we will show that the number of terms required to obtain an accurate computable solution is very small. Choosing different values for M in equations (19)-(21) we discuss the following three cases:

1. $M = 0$:

   Equations (19),(20) and (21) yields

   $$h \approx H_0 = c_1 - \frac{f_x}{f_{xx}}, \;\; k \approx K_0 = c_2 - \frac{f_y}{f_{yy}}, \;\; w \approx W_0 = c_3 - \frac{f_z}{f_{zz}}$$

   The approximate solution is

   $$\alpha = x - h \approx x - H_0 = x - \frac{f_x}{f_{xx}}, \;\; \beta = y - k \approx y - K_0 = y - \frac{f_y}{f_{yy}}$$

   and

   $$\gamma = z - w \approx z - W_0 = z - \frac{f_z}{f_{zz}}$$

   The generalized Newton's method for minimizing functions is given by

   $$x_{n+1} = x_n - \frac{f_x(x_n, y_n, z_n)}{f_{xx}(x_n, y_n, z_n)}, \;\; y_{n+1} = y_n - \frac{f_x(x_n, y_n, z_n)}{f_{xx}(x_n, y_n, z_n)}, \;\; z_{n+1} = z_n - \frac{f_x(x_n, y_n, z_n)}{f_{xx}(x_n, y_n, z_n)}$$

 ISBN: 978-84-614-6167-7

2. $M = 1$: Equations (19),(20) and (21) yields

$$h_1 = A_{1,0} = N_1(h_0, k_0, w_0) =$$

$$\Big[-\frac{f_y f_{xy}}{f_{yy} f_{xx}} - \frac{f_x f_{xz}}{f_{zz} f_{xx}} + \frac{1}{2}\Big(\frac{f_x^2 f_{xxx}}{f_{xx}^3} + \frac{f_y^2 f_{xyy}}{f_{yy}^2 f_{xx}} + \frac{f_z^2 f_{xzz}}{f_{zz}^2 f_{xx}} + \frac{f_x f_y f_{xxy}}{f_{xx}^2 f_{yy}} + \frac{f_x f_z f_{xxz}}{f_{xx}^2 f_{zz}} \frac{f_y f_z f_{xyz}}{f_{yy} f_{zz} f_{xx}}\Big)\Big]$$

while

$$k_1 = A_{2,0} = N_2(h_0, k_0, w_0) =$$

$$\Big[-\frac{f_x f_{yx}}{f_{yy} f_{xx}} - \frac{f_z f_{yz}}{f_{zz} f_{yy}} + \frac{1}{2}\Big(\frac{f_x^2 f_{yxx}}{f_{xx}^3} + \frac{f_y^2 f_{xyy}}{f_{yy}^2 f_{xx}} + \frac{f_z^2 f_{xzz}}{f_{zz}^2 f_{xx}} + \frac{f_x f_y f_{xxy}}{f_{xx}^2 f_{yy}} + \frac{f_x f_z f_{xxz}}{f_{xx}^2 f_{zz}} \frac{f_y f_z f_{xyz}}{f_{yy}^2 f_{zz} f_{xx}}\Big)\Big]$$

and,

$$w_1 = A_{3,0} = N_3(h_0, k_0, w_0) =$$

$$\Big[-\frac{f_x f_{yx}}{f_{yy} f_{xx}} - \frac{f_x f_{xz}}{f_{zz} f_{xx}} + \frac{1}{2}\Big(\frac{f_x^2 f_{xxx}}{f_{xx}^3} + \frac{f_y^2 f_{xyy}}{f_{yy}^2 f_{xx}} + \frac{f_z^2 f_{xzz}}{f_{zz}^2 f_{xx}} + \frac{f_x f_y f_{xxy}}{f_{xx}^2 f_{yy}} + \frac{f_x f_z f_{xxz}}{f_{xx}^2 f_{zz}} \frac{f_y f_z f_{xyz}}{f_{yy} f_{zz} f_{xx}}\Big)\Big]$$

The approximate solution is given by

$$\alpha = x - h \approx x - H_1 = x - h_0 - A_{1,0}, \quad \beta = y - k \approx y - K_1 = y - k_0 - A_{2,0}$$

and

$$\gamma = z - w \approx z - W_1 = z - w_0 - A_{3,0}$$

and hence we have the following iterations

$$x_{n+1} = x_n - \frac{f_x(x_n, y_n, z_n)}{f_{xx}(x_n, y_n, z_n)} - A_{1,0}(x_n, y_n, z_n)$$

$$y_{n+1} = y_n - \frac{f_y(x_n, y_n, z_n)}{f_{yy}(x_n, y_n, z_n)} - A_{2,0}(x_n, y_n, z_n)$$

and,

$$z_{n+1} = z_n - \frac{f_z(x_n, y_n, z_n)}{f_{zz}(x_n, y_n, z_n)} - A_{3,0}(x_n, y_n, z_n)$$

3. $M = 2$: Equations (19),(20) and (21) yields

$$\alpha = x - h \approx x - H_2 = x - h_0 - h_1 - h_2 = x - h_0 - A_{1,0} - A_{1,1}$$

$$\beta = y - k \approx y - K_2 = y - k_0 - k_1 - k_2 = y - k_0 - A_{2,0} - A_{2,1}$$

and,

$$\gamma = z - w \approx z - W_2 = z - w_0 - w_1 - w_2 = z - w_0 - A_{3,0} - A_{3,1}$$

 ISBN: 978-84-614-6167-7

The small values of h_2, k_2, w_2 can be found as follows:

$$h_2 = A_{1,1} = h_1\frac{\partial N_1}{\partial h}(h_0,k_0,w_0)+k_1\frac{\partial N_1}{\partial k}(h_0,k_0,w_0)+w_1\frac{\partial N_1}{\partial w}(h_0,k_0,w_0) = (h_1T_1+k_1T_2+w_1T_3)$$

$$k_2 = A_{2,1} = h_1\frac{\partial N_2}{\partial h}(h_0,k_0,w_0)+k_1\frac{\partial N_2}{\partial k}(h_0,k_0,w_0)+w_1\frac{\partial N_2}{\partial w}(h_0,k_0,w_0) = (h_1T_4+k_1T_5+w_1T_6)$$

and,

$$w_2 = A_{3,1} = h_1\frac{\partial N_3}{\partial h}(h_0,k_0,w_0)+k_1\frac{\partial N_3}{\partial k}(h_0,k_0,w_0)+w_1\frac{\partial N_3}{\partial w}(h_0,k_0,w_0) = (h_1T_7+k_1T_8+w_1T_9)$$

where

$$T_1 = h_0\frac{f_{xxx}}{f_{xx}} + \frac{k_0}{2}\frac{f_{xxy}}{f_{xx}} + \frac{w_0}{2}\frac{f_{xxz}}{f_{xx}},\ T_2 = -\frac{f_{xy}}{f_{xx}} + k_0\frac{f_{xyy}}{f_{xx}} + \frac{h_0}{2}\frac{f_{xxy}}{f_{xx}} + \frac{w_0}{2}\frac{f_{xyz}}{f_{xx}}$$

$$T_3 = -\frac{f_{xz}}{f_{xx}} + k_0\frac{f_{xzz}}{f_{xx}} + \frac{h_0}{2}\frac{f_{xxz}}{f_{xx}} + \frac{k_0}{2}\frac{f_{xyz}}{f_{xx}},\ \ T_4 = -\frac{f_{yx}}{f_{yy}} + h_0\frac{f_{yxx}}{f_{yy}} + \frac{k_0}{2}\frac{f_{yxy}}{f_{yy}} + \frac{w_0}{2}\frac{f_{yxz}}{f_{yy}},$$

$$T_5 = k_0\frac{f_{yyy}}{f_{yy}} + \frac{h_0}{2}\frac{f_{yxy}}{f_{yy}} + \frac{w_0}{2}\frac{f_{yyz}}{f_{yy}},\ \ T_6 = -\frac{f_{yz}}{f_{yy}} + w_0\frac{f_{yzz}}{f_{yy}} + \frac{h_0}{2}\frac{f_{yxz}}{f_{yy}} + \frac{k_0}{2}\frac{f_{yyz}}{f_{yy}},$$

$$T_7 = -\frac{f_{zx}}{f_{zz}} + h_0\frac{f_{zxx}}{f_{zz}} + \frac{k_0}{2}\frac{f_{zxy}}{f_{zz}} + \frac{w_0}{2}\frac{f_{zyz}}{f_{zz}},\ \ T_8 = -\frac{f_{zy}}{f_{zz}} + k_0\frac{f_{zyy}}{f_{zz}} + \frac{h_0}{2}\frac{f_{zxy}}{f_{zz}} + \frac{w_0}{2}\frac{f_{zyz}}{f_{zz}}$$

Finally,

$$T_9 = -\frac{f_{zzz}}{f_{zz}} + \frac{h_0}{2}\frac{f_{zxz}}{f_{xx}} + \frac{k_0}{2}\frac{f_{zyz}}{f_{zz}},$$

and hence

$$x_{n+1} = x_n - h_0(x_n,y_n,z_n) - h_1(x_n,y_n,z_n) - h_2(x_n,y_n,z_n)$$

$$y_{n+1} = y_n - k_0(x_n,y_n,z_n) - k_1(x_n,y_n,z_n) - k_2(x_n,y_n,z_n)$$

and,

$$z_{n+1} = z_n - w_0(x_n,y_n,z_n) - w_1(x_n,y_n,z_n) - w_2(x_n,y_n,z_n)$$

 ISBN: 978-84-614-6167-7

i	Newton	$M = 1$	$M = 2$
1	(1.92, 1.92, 1.92)	(1.984, 1.984, 1.984)	(2.0096, 2.0096, 1.99936)
2	(1.9968, 1.9968, 1.9968)	(2.0, 2.0, 2.0)	(2.0, 2.0, 2.0)
3	(1.99999, 1.99999, 1.99999)		
4	(2.0, 2.0, 2.0)		

Table 1: Results for Example 4.1.

i	Newton	$M = 1$	$M = 2$
1	1.0×10^{-3}	5.00025×10^{-3}	4.99988×10^{-3}
2	5.00029×10^{-3}	2.50016×10^{-3}	1.25020×10^{-3}
3	4.99995×10^{-3}	1.25008×10^{-3}	6.24984×10^{-4}
4	2.50018×10^{-3}	6.25042×10^{-4}	1.56275×10^{-4}
5	2.49975×10^{-3}	3.12521×10^{-4}	7.81231×10^{-5}
6	1.25010×10^{-3}	1.56260×10^{-4}	1.95344×10^{-5}
7	1.24988×10^{-3}	7.81302×10^{-5}	9.76538×10^{-6}

Table 2: Results for Example 4.2.

4 Applications to Optimization Problems

In this section, two optimization problems are given to illustrate the efficiency of using the extensions of Newton's method by ADM.

Example 4.1**: Logarithmic function** Consider the following nonlinear function in R^3

$$f(x, y, z) = \ln(x^2y^2z^2) - x - y - z.$$

The function f has a maximum point at $(2, 2, 2)$. Changing the sign of the function, then the problem become of finding a minimizer for the new function $-f$. With initial guess $(x_0, y_0, z_0) = (1.6, 1.6, 1.6)$, Table 1, shows some numerical results using Newtons method and the extension of Newtons method for $M = 1$ and $M = 2$. The numerical results obtained by the proposed method justify the advantage of using ADM to extend Newtons method.

Example 4.2**: Exponential function** Consider the following nonlinear function in R^3,

$$f(x, y, z) = e^{x-y} + e^{y-x} + e^{x^2} + z^2.$$

The Hessian $H(f(x, y, z))$ is positive definite, and so the function f has a strictly global minimizer at any critical point of f, namely at $(0, 0, 0)$. With initial guess $(x_0, y_0, z_0) = (0.1, 0.1, 0.1)$, Table 2 shows the supremum norm error between the exact solution $(0, 0, 0)$ and the approximate solution using Newton's method and the extended Newton's method, and reported as $\| (x_i, y_i, z_i) - (0, 0, 0) \|_\infty$.

 ISBN: 978-84-614-6167-7

Acknowledgements

This work has been partially supported by Jordan University of Science and Technology.

References

[1] J. M. Ortega, W.C. Rheinboldt; Iterative Solution of Nonlinear Equations in Several Variables, Academic Press, New York, 1970.

[2] A. L. Perssini; F. E. Sullivan; J. J. Uhl, Jr., The Mathematics of Nonlinear Programming. Springer-Verlag, New York 1991. *J. Comput. Physics*, Volume 166, pp 383-99 (2001).

[3] A. L. Perssini; F. E. Sullivan; J. J. Uhl, Jr., The Mathematics of Nonlinear Programming. *Comput. Math. Appl.*, Volume 40(2000), pp. 1015-1025.

[4] G. Adomian; R. Rach, On the solution of algebraic equations by the decomposition method. *Math. Anal. Appl.*, Volume 105(1985), pp. 141-166.

[5] G. Adomian; Solving Frontier Problems of Physics: The Decomposition Method. Kluwer Academic Publishers, Dordrecht, 1994.

[6] G. Adomian, A review of the decomposition method in applied mathematics. *Math. Anal. Appl.*, Volume 135(1988), pp. 501-544.

[7] S. M. El-Sayed, The modified decomposition method for solving nonlinear algbraic equations. *Appl. Math. Comput.*, Volume 132(2002), pp. 589-597.

[8] S. Abbasbandy, Improving Newton-Raphson method for nonlinear equations by modified Adomian decomposition method. *Appl. Math. Comput.*, Volume 145(2003), pp. 887893.

[9] Dogan Kaya; Salah M. El-Sayed, Adomian's decomposition method applied to systems of nonlinear algebraic equations. *Appl. Math. Comput.*, Volume 154(2004), pp. 487-493.

[10] E. Babolian; J. Biazar; A. R. Vahidi, Solution of a system of nonlinear equations by Adomian decomposition method. *Appl. Math. Comput.*, Volume 150(2004), pp. 847-854.

[11] Nadia Al-Rashaideh; Improving Newton's Method by Adomian Decomposition for Nonlinear Optimization Problems, Master thesis, Jordan University of Science and Technology, December 2008.

[12] Hossein Jafari; Varsha Daftardar-Gejji, Revised Adomian decomposition method for solving a system of nonlinear equations. *Appl. Math. Comput.*, Volume 175(2006), pp. 1-7.

 ISBN: 978-84-614-6167-7

[13] S. Abbasbandy, Extended Newton's method for a system of nonlinear equations by modified Adomian decomposition method. *Appl. Math. Comput.*, Volume 170(2005), pp. 648-656.

[14] J. Bizara, E. Babolian, R. Islam, Solution of systems of . *Appl. Math. Comput.*, Volume 170(2005), pp. 648-656.

 ISBN: 978-84-614-6167-7

Proceedings of the 11th International Conference on Computational and Mathematical Methods in Science and Engineering, CMMSE 2011 26–30 June 2011.

Efficient tools for detecting point sources in Cosmic Microwave Background maps

Pedro Alonso[1], Francisco Argüeso[1], Raquel Cortina[2], José Ranilla[2] and Antonio M. Vidal[3]

[1] *Department of Mathematics, Universidad de Oviedo, Spain*

[2] *Department of Computer Science, Universidad de Oviedo, Spain*

[3] *Department of Computer Systems and Computation, Universidad Politécnica de Valencia, Spain*

emails: palonso@uniovi.es, argueso@uniovi.es, raquel@uniovi.es, ranilla@uniovi.es, avidal@dsic.upv.es

Abstract

The Cosmic Microwave Background is a diffuse radiation which is contaminated by the radiation emitted by point sources. In this work, we present an efficient algorithm, with a high degree of parallelism, which can replace advantageously the classical approaches for detecting point sources in Cosmic Microwave Background maps. High performance computing libraries and parallel computing techniques have allowed to construct a portable, fast and numerically stable algorithm.

Key words: Efficiency, Cosmic Microwave Background
MSC 2000: 65F05, 65Y20, 68W10

1 Introduction

The cosmic microwave background (CMB) is a radiation which comes from the beginning of the Universe, carrying relevant information about its origin, evolution and structure. Since its discovery in 1964 by Penzias and Wilson (see [6]), the CMB has been measured by instruments aboard balloons and satellites such as the NASA COBE satellite (1992) ([7]), whose principal investigators, Mather and Smoot, detected the CMB fluctuations for the first time. In 2003, another Nasa satellite, WMAP, used these fluctuations to determine the cosmological parameters with unprecedented accuracy (see [8]). In 2009 the ESA Planck satellite was launched and nowadays it is gathering CMB data in order to improve the knowledge about our Universe (see [9]).

 ISBN: 978-84-614-6167-7

The CMB is a diffuse radiation which is contaminated by the radiation emitted by point sources. The detection of these point sources is vital for cleaning the radiation maps and also from the astrophysical point of view ([5]).

In [1] we presented the Neville Elimination as an efficient tool for detecting point sources in CMB maps. In this work, we present other efficient method for the same goal. The detection of these point sources is vital for cleaning the radiation maps and also from the astrophysical point of view.

2 Problem description

In this section we will present a typical method of point source detection in CMB maps.

In a region of the celestial sphere, we suppose to have a certain number n of radio sources that can be considered as point-like objects if compared to the angular resolution of our instruments. This means that their actual size is smaller than our smallest resolution cell. The emission of these sources is superimposed to the radiation $f(x, y)$. In our particular case this radiation is the CMB. A model for the emission as a function of the position (x, y) is:

$$\tilde{d}(x, y) = f(x, y) + \sum_{\alpha=1}^{n} a_\alpha \ \delta(x - x_\alpha, y - y_\alpha)$$

where $\delta(x, y)$ is the 2D Dirac delta function, the pairs are the locations of the point sources in our region of the celestial sphere, and a_α are their intensities. We observe this radiation through an instrument, with beam pattern $b(x, y)$, and a sensor that adds a random noise $n(x, y)$ to the signal measured. Again, as a function of the position, the output of our instrument is:

$$d(x, y) = \sum_{\alpha=1}^{n} a_\alpha \ b(x - x_\alpha, y - y_\alpha) + (f * b)(x, y) + n(x, y) \tag{1}$$

where the point sources and the diffuse radiation have been convolved with the beam. In our application, we are interested in extracting the locations and the intensities of the point sources. We thus assume that the intensities of the point sources are sufficiently above the level of the rest of the signal, and consider the latter as just a disturbance superimposed to the useful signal. If $c(x, y)$ is the signal which does not come from the point sources, model (1) becomes

$$d(x, y) = \sum_{\alpha=1}^{n} a_\alpha \ b(x - x_\alpha, y - y_\alpha) + c(x, y). \tag{2}$$

If our data set is a discrete map of N pixels, the above equation can easily be rewritten in vector form, by letting d be the lexicographically ordered version of the discrete map $d(x, y)$, a be the n-vector containing the positive source intensities a_α, c the lexicographically ordered version of the discrete map, $c(x, y)$, and ϕ be an $N \times n$

 ISBN: 978-84-614-6167-7

matrix whose columns are the lexicographically ordered versions of n replicas of the map $b(x, y)$, each shifted on one of the source locations. Equation (2) thus becomes

$$d = \phi\ a + c. \tag{3}$$

Looking at Eqs. (2) and (3), we see that, if the goal is to find locations and intensities of the point sources, our unknowns are the number n, the list of locations (x_α, y_α), with $\alpha = 1, ...n$ and the vector a. It is apparent that, once n and (x_α, y_α) are known, matrix ϕ is perfectly determined. Let us then denote the list of source locations by the $n \times 2$ matrix R, containing all their coordinates. For the CMB we can assume that c is a Gaussian random field with zero mean and known covariance ξ. Thus the likelihood function is

$$p(d|n, R, a) = \exp\left(-(d - \phi\ a)^t \xi^{-1} (d - \phi\ a)/2\right). \tag{4}$$

If we define $M = \phi^t \xi^{-1} \phi$, $e = \phi^t \xi^{-1}\ d$, the maximization of (4) leads us to the linear system

$$Ma = e \tag{5}$$

where M is a $n \times n$ matrix. The solution of this system will yield the maximum likelihood estimator of the source intensities.

One problem with this approach is that, in principle, we know neither the number n of point sources nor their positions. One standard way of dealing with this difficulty is considering (5) the local maxima of e and selecting as source positions these local maxima above a certain threshold. A 5σ threshold, with σ the standard deviation of e, is typically used, since such high fluctuations rarely have their origin in the CMB or the noise.

Finally, we have to find suitable methods to solve the system shown in (5), taking into account that the number of sources can range from tens to thousands depending on the size of the region studied and also on the frequency analyzed.

The problem statement, as described above, involves the processing of large matrices, if we want to cover a significant region of space. This can lead to excessive computation times as well as loss of precision. In this work, we provide an efficient solution with both aspects in mind: to get a reasonable run time and a numerically stable algorithm. To achieve both objectives, we resorted to the use of parallel computing and high performance numerical libraries.

3 Efficient implementations

First, it should be noted that when we are building the system $Ma = e$ we consider that the matrices M, ϕ and ξ are of order N, while the vectors e and d have N rows. Once vector e has been filtered by using the fixed threshold, matrix ϕ is set to $N \times n$ by choosing the adequate rows.

To solve the system (5) is necessary to calculate the matrix $M = \phi^t \xi^{-1} \phi$, and the vector $e = \phi^t \xi^{-1}\ d$, which should do so as efficiently as possible. A classical approach

 ISBN: 978-84-614-6167-7

could start by computing the inverse of ξ as a mean for calculating vector e and matrix M. Then, after applying a threshold process, the linear system $Ma = e$ can be solved. The computational cost of the classical approach implies $2N^3 + 2N^2 + 6Nn$ Flops.

However, from the numerical point of view, it should be desirable to obtain M and e without calculating the inverse of ξ. Some other ideas should be used in order to obtain an efficient algorithm. We have applied the following ones:

- Avoid unstable operations like computing inverses or multiplying large matrices. Instead use orthogonal transformations if possible.
- Try to solve large scale problems, having in mind this:
  - Use moderately the memory, avoiding unnecessary storage of data.
  - Get a moderately execution time.
- Organize the algorithms in such a way that high performance sequential or parallel libraries can be used.

With these ideas an efficient algorithm can be derived. As $\xi \in R^{N\times N}$ is a symmetric positive definite matrix, Cholesky decomposition can be used to obtain a lower triangular matrix such that: $\xi = LL^t$ ([4]). Hence, vector e can be expressed as:

$$e = \phi^t \xi^{-1} d = \phi^t L^{-t} L^{-1} d = \phi^t L^{-t} c_1 = \phi^t c_2$$

with $c_1 = L^{-1}d$ and $c_2 = L^{-t}c_1$.

Thus, vector e can be computed by performing a matrix-vector product, where matrix $\phi \in R^{N\times N}$. Observe that these operations involve a cost of $(N^3)/6 + 4(N^2)$ Flops.

As explained in the previous section, thresholding can be applied now to vector e, obtaining those positions with a value higher than 5σ. This is equivalent to obtain a selection matrix $P \in R^{N\times n}$, which consists of those columns of the identity matrix with a '1' in the position determined by the thresholding of vector e, and obtain $\tilde{e} = P^t e = (\phi P)^t c_2$.

Now, in order to construct the part of matrix M which is involved in the threshold linear system, we construct

$$\tilde{M} = P^t M P = (P\phi)^t \xi^{-1} (\phi P) = \tilde{\phi} L^{-t} L^{-1} \tilde{\phi},$$

with $\tilde{\phi} = \phi P \in R^{N\times n}$.

Thus, $\tilde{M} = (L^{-1}\tilde{\phi})^t (L^{-1}\tilde{\phi}) = Z^t Z$, with $Z = L^{-1}\tilde{\phi}$. If we compute the QR decomposition of $Z = QR$, with $Q \in R^{N\times N}$, orthogonal, and $R \in R^{N\times n}$, upper triangular, $\tilde{M} = Z^t Z = (QR)^t (QR) = R^t R$ and linear system $Ma = e$ can be expressed as $(R^t R)a = \tilde{\phi} c_2$. Thus, vector a can be computed by solving the triangular linear systems $R^t y = \tilde{\phi} c_2$ and $Ra = y$.

The construction of $\tilde{M}$ involves $(N^2)n + 2n^2(N - n/3)$ Flops and the solution of the final linear systems involves $2n^2$ Flops.

 ISBN: 978-84-614-6167-7

These ideas can be summarized in the following algorithm:

Algorithm CMB

Input ϕ', ξ, d, with $\phi, \xi \in R^{N \times N}$, $d \in R^{N \times 1}$

Step 1. Compute $\xi = LL^t$ (Cholesky factorization)

Step 2. Obtain e:

Solve $L * c_1 = d$ and $L^t c_2 = c_1$

Compute $e = \phi^t c_2$

Step 3. Calculate the positions of e that are above the threshold ($e(i) \geq 5\sigma$, $e \rightarrow \tilde{e}$)

Step 4. Get the columns of ϕ associated with the indices from Step 3: $\phi \rightarrow \tilde{\phi}$

Step 5. Solve $LZ = \tilde{\phi}$

Step 6. Compute the QR factorization of Z ($Z = QR$)

Step 7. Solve the triangular systems: $R^t y = \tilde{e}$, $Ra = y$

Output a

Finally, it should be noted that considering the high number of data and operations involved in the resolution of the problem, the use of parallel strategies is particularly suitable.

4 Parallelization strategy and experimental results

We have implemented the two algorithms described in the previous section. We called Classical algorithm to which constructs the matrix M and the vector e, starting from the inverse of ξ. In turn, the CMB algorithm avoids the inverse computation and uses matrix decomposition techniques.

Experiments reported in this section employ IEEE 754 double precision arithmetic. The hardware platform has one Intel Xeon E5530 Quad-Core processors (that is, 4 cores), running at 2.40 GHz, and it is equipped with 48 GB RAM. The operating system utilized is Ubuntu Linux distro (10.04.2 LTS).

High performance implementations of the BLAS were provided by Intel MKL (*Mathematical Kernel Library*, version 10.3). Arithmetic intensive operations of algorithms described in Section 3 have been addressed through calls to the appropriate subroutines of MKL, i.e. Cholesky factorization with *DPOTRF*, QR factorization with *DGEQRF*, and so on.

Figure 1 shows the execution times obtained for the algorithms considered. It shows that the calculated times are much higher for the classical algorithm. For example, if N is between 2^{13} (8192) and 2^{15} (32768), the time of classical algorithm is more than 13 times that of the CMB algorithm. These results can also be seen in Table 1.

The time of CMB algorithm is reasonably small when the problem size grows. This suggests the possibility of studying wider regions of space, which involve the processing of larger matrices at an affordable execution time, by using a larger number of cores. In addition, the technique used in the CMB algorithm is an efficient alternative that can be applied also in more complex computational methods such as Bayesian methods proposed in [3].

 ISBN: 978-84-614-6167-7

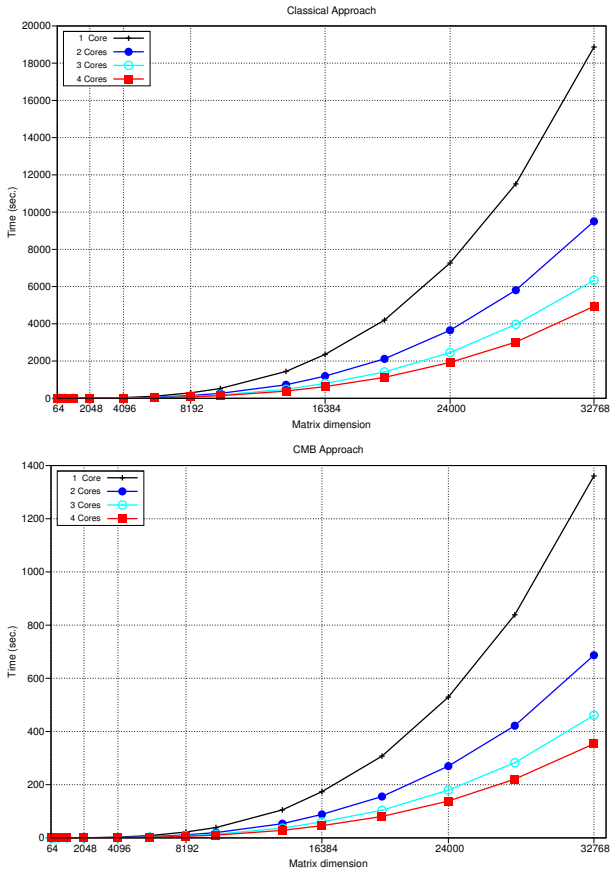


Figure 1: Execution time for the Classical and CMB algorithms.

Algorithm/Cores	1	2	3	4
Classical	18870.30	9503.11	6336.63	4939.95
CMB	1360.64	686.77	460.99	353.76

Table 1: Time (sec.) for $N = 2^{15}$.

5 Concluding Remarks

We have proposed an efficient algorithm, with a high degree of parallelism that can replace advantageously the classical approaches.

As main advantages of algorithm CMB we can cite:

- It allows to confront large scale problems with a reasonable execution time, optimizing the memory usage.
- Its parallelization is very efficient; near-optimal speedups have been obtained in

 ISBN: 978-84-614-6167-7

many cases.

- The constructed algorithm is scalable in the sense that execution time can be maintained, by increasing the problem size and the number of cores at the same rate.
- The use of high-performance libraries and the organization of the algorithm guarantees numerical stability and portability.
- Techniques developed can also be applied to the resolution of Bayesian methods described in [3], thus completing an important analysis tool.

Acknowledgements

This work was financially supported by the Spanish Ministerio de Ciencia e Innovación and by FEDER (Projects TIN2010-14971, TIN2008-06570-C04-02, TEC2009-13741 and CAPAP-H3 TIN2010-12011-E), Universitat Politècnica de València through Programa de Apoyo a la Investigación y Desarrollo (PAID-05-10) and Generalitat Valenciana through project PROMETEO/2009/013.

References

[1] P. Alonso, R. Cortina, J. Ranilla, A.M. Vidal, An efficient and scalable block parallel algorithm of Neville Elimination as a tool for the CMB maps problem. J. Math. Chem. DOI: 10.1007/s10910-010-9769-0

[2] E. Anderson *et al.*, LAPACK Users' Guide, Society for Industrial and Applied Mathematics, Philadelphia, PA, 1999.

[3] F. Argüeso *et al.*, A Bayesian technique for the detection of points sorces in cosmic microware backround maps, Mon. Not. Roy. Astron. Soc., 2011, in press.

[4] G. H. Golub, C. F. Van Loan, Matrix Computations, Johns Hopkins University Press, Baltimore, 1996.

[5] M. Lopez-Caniego *et al.*, Comparison of filters for the detection of point sources in Planck simulations. Mon. Not. Roy. Astron. Soc. **370**, 2047 (2006)

[6] A.A. Penzias, R.W. Wilson, A Measurement of Excess Antenna Temperature at 4080 Mc/s. Astrophys. J. **142**, 419 (1965)

[7] G. Smoot *et al.*, Structure in the COBE Differential Microwave Radiometer First-Year Maps. Astrophys.J. **396**, L1 (1992)

[8] D.N. Spergel *et al.*, First-Year Wilkinson Microwave Anisotropy Probe (WMAP) Observations: Determination of Cosmological Parameters. Astrophys. J. Suppl. **148**, 175 (2003)

[9] J.A. Tauber, The Planck Mission, in: A. Lasenby, A. Wilkinson (Eds.), New Cosmological Data and the Values of the Fundamental Parameters, Proceedings of IAU Symposium 201, 2005, 86.

[10] L. Toffolatti *et al.*, Extragalactic source counts and contributions to the anisotropies of the cosmic microwave background: predictions for the Planck Surveyor mission. Mon. Not. Roy. Astron. Soc. **297**, 117 (1998)

 ISBN: 978-84-614-6167-7

Proceedings of the 11th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2011
26–30 June 2011.

Computations with Pascal matrices

Pedro Alonso[1], Jorge Delgado[2], Rafael Gallego[1] and Juan Manuel Peña[2]

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

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

emails: palonso@uniovi.es, jorgedel@unizar.es, rgallego@uniovi.es, jmpena@unizar.es

Abstract

Bidiagonal factorizations of Pascal matrices are considered and they can be applied to obtain algorithms with high relative accuracy. A result on the conditioning of Pascal matrices is announced.

Key words: Pascal matrices, Bidiagonal factorizations, Accurate computations, Conditioning, Vandermonde matrices,
MSC 2000: 65F05,65F15, 65F35

1 Introduction

Pascal matrices have a long history (cf. [1], [5], [8]) and present important applications in filter design and image and signal processing ([5], [11]), as well as in probability, combinatorics, numerical analysis and electrical engineering ([4]), among other fields. Recently, several papers have presented fast algorithms for solving linear systems whose coefficient matrices are Pascal matrices ([9], [12]) and fast eigenvalue algorithms ([13]).

In Section 3 of this paper we announce a result on the ill-conditioning of Pascal matrices, showing that they are always worse conditioned than the Vandermonde matrices of the same order. In spite of this result, we announce that one can obtain algorithms with high relative accuracy (HRA) for the computation of eigenvalues and inverses of Pascal matrices, as well as for solving certain linear systems whose coefficient matrices are Pascal matrices. HRA means that the computations are such that subtractions (except for initial data) are not required so that the accuracy is preserved. In order to construct HRA algorithms, two tools are used. On the one hand, a result on bidiagonal factorizations of Pascal matrices

 ISBN: 978-84-614-6167-7

mentioned in Section 2. On the other hand, the algorithms with HRA for totally nonnegative matrices presented in [7]. Recall that a *totally nonnegative* matrix is a matrix whose minors are all nonnegative and that Pascal matrices are totally nonnegative (see Section 2).

2 Factorization of Pascal matrices

A Pascal matrix of order n is the symmetric matrix

$$P = (p_{ij})_{1\leq i,j\leq n}; \quad p_{ij} := \binom{i+j-2}{j-1}. \tag{1}$$

A lower triangular Pascal matrix of order n is the lower triangular matrix

$$P_L = (q_{ij})_{1\leq i,j\leq n}; \quad q_{ij} := \binom{i-1}{j-1}, \tag{2}$$

where $q_{ij} := 0$ if $j > i$.

It is well known (cf. [10]) that the lower triangular Pascal matrix P_L is the factor of the Cholesky factorization of the Pascal matrix P:

$$P = P_L P_L^T. \tag{3}$$

The following well known result provides a bidiagonal decomposition of a lower triangular Pascal matrix and will be a key tool in the accurate algorithms announced in this paper.

Lemma 1 *The lower triangular Pascal matrix of order n P_L given by (2) satisfies*

$$P_L = \begin{pmatrix} 1 & & & & \\ 0 & 1 & & & \\ & \ddots & \ddots & & \\ & & 0 & 1 & \\ & & & 1 & 1 \end{pmatrix} \begin{pmatrix} 1 & & & & \\ 0 & 1 & & & \\ & \ddots & \ddots & & \\ & & 1 & 1 & \\ & & & 1 & 1 \end{pmatrix} \cdots \begin{pmatrix} 1 & & & & \\ 1 & 1 & & & \\ & \ddots & \ddots & & \\ & & 1 & 1 & \\ & & & 1 & 1 \end{pmatrix}. \tag{4}$$

The joint use of the factorizations (3) and (4) gives a bidiagonal decomposition of a matrix P that will be denoted by $\mathcal{BD}(P)$. This factorization can also be used as a proof of the total nonnegativity of a Pascal matrix because all factors are obviously totally nonnegative and it is well known (cf. Theorem 3.1 of [3]) that the product of totally nonnegative matrices is also totally nonnegative.

It should be noted that the factorization (4) also appeared in a previous work by the authors (see proof of Lemma 1 of [2]). In this case, and considering the LU factorization obtained by Neville elimination with a pivoting strategy, it was shown that $|L| \leq P_L$. Given a matrix $A = (a_{ij})_{1\leq i,j\leq n}$, we denote by $|A|$ the matrix whose (i,j)-entry is $|a_{ij}|$.

 ISBN: 978-84-614-6167-7

3 Conditioning and accurate algorithms for Pascal matrices

Let us start this section by recalling the definition of a Vandermonde matrix. A Vandermonde matrix of order n is the matrix

$$V = (v_{ij})_{1\leq i,j\leq n} := (j^{i-1})_{1\leq i,j\leq n}. \tag{5}$$

It is well known that Vandermonde matrices are ill-conditioned.

Given a nonsingular matrix A, let us consider the traditional and the Skeel condition numbers of A, denoted by $\kappa_\infty(A)$ and $\mathrm{Cond}(A)$, respectively, and given by

$$\kappa_\infty(A) := \|A\|_\infty \, \|A^{-1}\|_\infty, \quad \mathrm{Cond}(A) := \| \, |A^{-1}| \, |A| \, \|_\infty.$$

Let us recall the following two properties:

- $\mathrm{Cond}(A) \leq \kappa_\infty(A)$ and it can be much smaller, and,
- in contrast to $\kappa_\infty(A)$, $\mathrm{Cond}(A)$ is invariant under row scaling: if D is a nonsingular diagonal matrix then
$$\mathrm{Cond}(DA) = \mathrm{Cond}(A).$$

These properties provide some of the reasons that explain why the Skeel condition number $\mathrm{Cond}(A)$ is more satisfying than the traditional condition number $\kappa_\infty(A)$ (see also Section 7.2 of [6]).

Let us announce that Pascal matrices always present a worse conditioning than the Vandermonde matrices of the same order.

Theorem 1 *Let P and V be the Pascal and Vandermonde matrices of order n given by (1) and (5), respectively. Then*

$$\mathit{Cond}(P) \geq \mathit{Cond}(V). \tag{6}$$

However, in spite of the ill-conditioning of a Pascal matrix P, one can use its bidiagonal decomposition $\mathcal{BD}(P)$ to provide accurate algorithms. In fact, as shown in [7], if the diagonal entries of the diagonal matrix of the bidiagonal decomposition $\mathcal{BD}(A)$ of a totally nonnegative matrix A and the off-diagonal entries of the remaining factors of $\mathcal{BD}(A)$ are known with HRA, then we can find algorithms with HRA to perform some computations with these matrices, such as the computation of their singular values, the computation of their eigenvalues, the computation of their inverses or solving certain linear systems $Ax = b$ (those where b has a chessboard pattern of alternating signs). So, we can construct such accurate algorithms for Pascal matrices because all mentioned entries are in this case 1's.

 ISBN: 978-84-614-6167-7

Acknowledgements

This work has been partially supported by the Spanish Research Grant MTM2009-07315 and under MEC and FEDER Grant TIN20010-14971.

References

[1] L. Aceto, D. Trigiante, *The matrices of Pascal and other greats*, Amer. Math. Monthly **108** (2001) 232–245.

[2] P. Alonso, J. Delgado, R. Gallego and J. M. Peña, *Growth Factors of Pivoting Strategies Associated to Neville Elimination*, J. Comp. Appl. Math. **235** (2011) 1775–1762.

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

[4] V. Biolkova, D. Biolek, *Generalized pascal matrix of first-order S-Z transforms*, in: ICECS99 Pafos, Cyprus 1999, pp. 929–931.

[5] A. Edelman, G. Strang, *Pascal Matrices*, MIT, Cambridge, MA, 2003.

[6] N. J. Higham, *Accuracy and stability of numerical algorithms*, second ed., Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2002.

[7] P. Koev, *Accurate computations with totally nonnegative matrices*, SIAM J. Matrix Anal. Appl. **29** (2007) 731–751.

[8] B. Lewis, *Revisiting the Pascal matrix*, Amer. Math. Monthly **117** (2010) 50–66.

[9] X.-G. Lv, T.-Z. Huang, Z.-G. Ren, *A new algorithm for linear systems of the Pascal type*, J. Comput. Appl. Math. **225** (2009) 309–315.

[10] J.H. Mathews, *Numerical Methods for Mathematics, Science, and Engineering*, second ed., Prentice-Hall, Englewood Cliffs, NJ, 1992.

[11] B. Psenicka, F. Garcia-Ugalde, A. Herrera-Camacho, *The bilinear Z transform by Pascal matrix and its application in the design of digital filters*, IEEE Signal Process. Lett. **9** (2002) 368–370.

[12] X. Wang, L. Lu, *A fast algorithm for solving linear systems of the Pascal type*, Appl. Math. Comput. **175** (2006) 441–451.

[13] X. Wang, J. Zhou, *A fast eigenvalue algorithm for Pascal matrices*, Appl. Math. Comput **183** (2006) 711–716.

 ISBN: 978-84-614-6167-7

Proceedings of the 11th International Conference on Computational and Mathematical Methods in Science and Engineering, CMMSE 2011 26–30 June 2011.

Building a library for solving structured matrix problems

Pedro Alonso-Jordá[1], Pablo Martínez-Naredo[2], F.J. Martínez-Zaldívar[3], José Ranilla[2] and Antonio M. Vidal[1]

[1] *Departamento de Sistemas Informáticos y Computación, Universitat Politècnica de València (Spain)*

[2] *Departmento de Informática, Universidad de Oviedo (Spain)*

[3] *Departamento de Comunicaciones, Universitat Politècnica de València (Spain)*

emails: palonso@dsic.uvp.es, pmnaredo@gmail.com, fjmartin@dcom.upv.es, ranilla@uniovi.es, avidal@dsic.upv.es

Abstract

This papers shows **StructPack**, a set of library subroutines and executable commands to solve structured linear systems. It is an evolving software package that nowadays contains a solver for symmetric and tridiagonal symmetric Toeplitz matrices. **StructPack** is freely-available and can be downloaded from http://www.inco2.upv.es.

Key words: Software library, structured matrices, Toeplitz

1 Introduction

Matrix Computation is especially useful in many problems of engineering and science. Often the types of matrix problems that appear in many of them are well-known standard problems. Thus the existence of matrix libraries is a useful tool that allows the specialist in a particular field to focus on solving his problem and to save hours of programming numerical routines with which this specialist is often not used.

In the market there are currently a large number of matrix libraries covering a wide field of scientific and technological applications. To cite just a few: LAPACK [6], ScaLAPACK [11], PETSc [10], SuperLU [14], ARPACK [1],... or commercial implementations such as Matlab [8], Mathematica [7], etc.

Many of the matrix libraries are designed for one or more classes of matrices. For example, LAPACK works on dense or band matrices and their routines are optimized for this type of matrices. Similarly, ARPACK is designed to work with sparse matrices.

In contrast, matrices arising in many scientific or technical problems often have an explicit external structure, i.e., they are called structured matrices. Typical examples

 ISBN: 978-84-614-6167-7

of structured matrices are Toeplitz matrices, Hankel, Vandermonde, Cauchy, circulant matrices, etc. There are many areas where you often see these types of matrices. Without wishing to be exhaustive we can include: image and signal processing, solving differential and integral equations, calculation of spline functions, time series analysis, Markov chains and queuing theory, polynomial and power series computations, etc. (see for example references [19, 21, 22, 18, 24])

There is an unfilled space in the field of matrix libraries: that for structures matrices. It is true that it represents a broad and ambiguous set of processing methods which are dependent on the type of matrix. It is difficult to conceive an efficient library for a large number of cases and with a certain unity of content in the using of computational technology. In other words the creation of a library of these features is a challenge both of scientific and technological importance.

The most significant precedents for this idea are the developments presented in Netlib [9] and SLICOT [12]. The first one is a set of routines to solve systems of linear equations, dating mainly from 1982 and became part of Netlib in the '90s. The second precedent is a set of subroutines included in the package SLICOT to solve systems of linear equations with general Toeplitz matrices, symmetric positive definite Toeplitz matrices, and block Toeplitz matrices.

Motivated by some applications we have encountered in the field of Signal Processing and being aware of the potential uses in many other fields, we have begun the task of developing a library for processing structured matrices. The library, named **StructPack** [13], aims at solving typical computational matrix problems on structured matrices. These problems are fundamentally solving systems of linear equations, solving least squares problems, calculation of eigensystems and calculation of singular value decomposition. Some of these problems have been tackled by the authors during the last years, [20, 23, 17, 16, 15].

Given the wide range of problems that **StructPack** tries to address is unthinkable to present it as a closed and finished product. Therefore, our design goals include a progressive development of different routines. This will also allow feedback needed to ensure the quality of developments.

In this paper we present the basic ideas that have helped to design the library and we give an overview of its functionality. We also describe the current state of present developments and what are the next ones to be included in the short term.

The rest of the paper is organized as follows: Section 2 describes some general features of the library and some problems which have already been solved by **StructPack** routines. Section 3 describes the main features of the webpage that allows access to the library. Section 4 is dedicated to showcasing the capabilities of the library and examples of use. Section 5 shows the future lines of work.

2 Solving structured problems with StructPack

StructPack is a library of numerical routines that solve Numerical Linear Algebra problems on structured matrices. The routines are written in Fortran90/95, but may

 ISBN: 978-84-614-6167-7

also be called from C, for which appropriate interfaces have been provided. Currently, the version `v0.1` is designed for Linux environments, and optimized for use on sequential CPU type and multicore architectures. For it, the OpenMP API has been used in its development. Versions for other operating systems like Windows, or for other programming paradigms, such as distributed memory or graphic accelerator units, shall be added in successive versions.

The problems to be solved with **StructPack** are the classical problems of Numerical Linear Algebra, that is, solving systems of linear equations, solving least squares problems, calculate eigenvectors and compute the singular value decomposition. Algorithms implemented in **StructPack** to solve such above problems, act on different types of matrices such as Toeplitz, Hankel, Vandermonde, or circulant matrices, ... and in general on matrices that present displacement structure [21]. This includes specific cases within each of the previous types as tridiagonal Toeplitz matrices, positive definite symmetric matrices, etc.

StructPack has been designed using efficient algorithms. Basically we are using algorithms that use the displacement structure of structured matrices, and we are optimizing them for multicore architectures. For simple operations, of linear or quadratic complexity, BLAS computational kernels [2] have been used. The codes also use libraries to calculate the FFT, either the MKL Intel library [5] (if provided by the user) or the FFT Pack [3]. The source code can be compiled either by using public domain compilers, like GNU gcc [4], or by commercial compilers like Intel compilers.

An important feature of the library is that it provides commands to resolve problems directly, simply by writing them in the command line. This provides primarily a great ease of use. **StructPack** also provides files `.mex`, allowing the use of routines in the Matlab environment.

Currently, the routines included in the library can be used to:

- Solve symmetric, tridiagonal, Toeplitz, linear systems.
- Calculate eigensystems and singular value decomposition of symmetric tridiagonal Toeplitz matrices.
- Solve symmetric Toeplitz linear systems of equations.
- Solve non-symmetric Toeplitz linear systems.

Routines to calculate eigenvalues and eigenvectors of symmetric Toeplitz matrices will be incorporated soon. All information on **StructPack** is public domain and is accessible at [13]. The following sections describe in detail the contents of this website and the main features of **StructPack**.

3 Website description

The website structure of **StructPack** is organized as usual manner. Thus, it has a "Main page" and some tabs for "Installation", "Documentation", "Test", "FAQs" and so on.

 ISBN: 978-84-614-6167-7

On the one hand, the main page shows a presentation of the website, the licensing and last version news. Besides, a brief description of the groups involved on the development of the package and the related projects are outlined. On the other hand, next items summarize the most relevant tabs:

- Documentation. Here, a link to the `Doxygen` generated documentation is shown. In this documentation we can find all the library API specifications and the available line commands.

- Test. In this tab it is shown how to run performance tests to obtain the timing and precision results of the installed package in the machine.

- Working Notes. The working notes of the package will be stored here. This section will collect all the detailed information related to the entire package, to some routines, to performance results or conclusions, etc. The chosen acronym is SPAWN: StructPAck Working Notes.

- FAQs. The "Frequently Asked Questions" tab includes general questions, installation questions, how to use or program with **StructPack**, questions or problems about different platforms or operating systems and miscellaneous questions.

- Third Party Software. This tab contains links to the websites of the used software to produce **StructPack**. The software is grouped as:

  - Compilers: `gcc`, `gfortran`, `icc` and `ifort`.
  - Software version control: `subversion`.
  - Software documentation: `doxygen`.
  - Text editing: `vi (vim, gvim, ...)`
  - Environment configuration: `libtool`, `autoconf`
  - ...

- References. In this tab we can find bibliographic references that have been used to produce **StructPack**. They include our previous work in these algebraic problems and other related publications. Besides, we include links to other websites with similiar contents.

- Installation. It contains three sections: "How to install", "Download" and "Installation test". In the "Download" section the last versions of the software can be chosen and downloaded, whilst "Installation test" tab shows the way to check that the installation is correct.

  In the "How to install" section all details of how to get a correct installation are shown. A typical installation process, consisting of three steps, is used: `$ ./configure`, `$ make` and `$ make install`.

  In the *configure* step the system where the package is being installed is tested, and the appropriate *Makefile* files for building the package are created. Several

 ISBN: 978-84-614-6167-7

options, i.e. installation path, compilers and libraries, etc., can be specified at configuration time in order to change the default installation settings. *Make* step compiles and links all libraries and programs of **StructPack**. Both static and shared versions of the libraries are built, unless otherwise stated at the configuration time. Finally, *make install* step moves software and documentation toward the desired folder.

By default, the source codes are compiled using the GNU compilers, though Intel compilers are currently supported.

StructPack has been designed using efficient algorithms which use BLAS computational kernels for operations of linear or quadratic order complexity. Therefore a generic BLAS implementation (the default choice), MKL Intel library or any compatible package must be installed. **StructPack** also includes and uses libraries to calculate the FFT. The FFT Pack [3] is part of **StructPack**, but users can use the solution provided by MKL.

Figure 1 shows partially the information shown in this tab. For a detailed description of the installation process see [13].

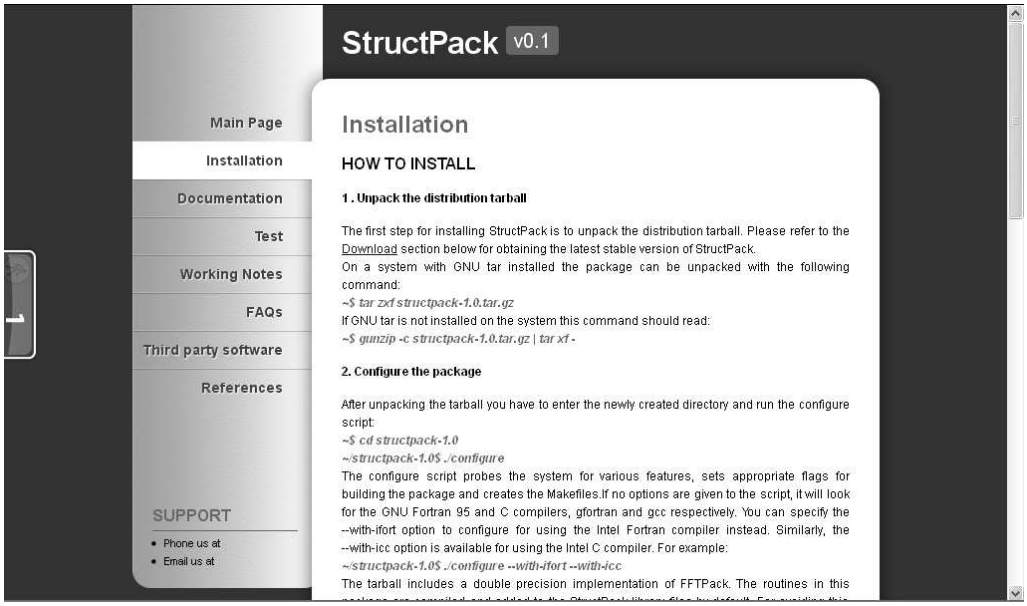


Figure 1: Installation instructions

4 Library usage and commands

Currently, **StructPack** offers solution to Toeplitz linear systems where the matrix can be symmetric tridiagonal or full symmetric. The user has two possibilities of using the package to solve this problem. He can implement his own application using the

 ISBN: 978-84-614-6167-7

library modules provided by **StructPack** or he can use the operating system commands available to solve the problem.

The most natural way to use our modules is by means of the design of a Fortran 90 application. Currently, the package provides modules `tpsysv_module` (full symmetric) and `tpsytrid_module` (tridiagonal symmetric). They can be use as shown in the following simplified example:

```
program tpsysv_test
  use tpsysv_module
  implicit none

  double precision, dimension(100) :: t, x, b
  integer :: nb = 20
  logical :: pivoting = .true.

  call random_number( t )
  call random_number( b )

  call tpsysv( t, x, b, nb, pivoting )

end program tpsysv_test
```

Routine `tpsysv` obtains the solution $\mathbf{x}$ of linear system $\mathbf{Tx} = \mathbf{b}$ given the first column (row) $\mathbf{t}$ of $\mathbf{T}$ and the right hand side vector $\mathbf{b}$. The application is obtained by linking with the library `structpack.a` provided by the package.

For its use in a C application, the package provides the header file `tpsysv_module.h` which must be included, and the Fortran 90 module `ctpsysv_module.F90` which the application must be linked with. Each **StructPack** module that solves a given problem has a counterpart module whose name is prefixed by `c`. This module interfaces C to make possible to call a Fortran 90 routine/function from the C application. It uses the `iso_c_binding` module which provides C/fortran interoperability. For example, routine `ctpsysv` can be called from a C source code keeping the name, number and type of arguments of the driver routine to be called described in the module as follows:

```
subroutine ctpsysv(n, t, x, b, nb, piv ) bind(c)
  integer(kind=c_int), value, intent(in) :: n
  real(c_double), dimension(n), intent(in) :: t
  real(c_double), dimension(n), intent(out) :: x
  real(c_double), dimension(n), intent(inout) :: b
  integer(kind=c_int), value, intent(in) :: nb
  integer(kind=c_int), value, intent(in) :: piv

  . . .
end subroutine ctpsysv
```

Two commands are currently provided by the package to allow the user to solve each one of the two problems tackled: `tpsysv` and `tpsytrid`. A calling example could

 ISBN: 978-84-614-6167-7

be `tpsysv -n 100 -p`. In this example, the command solves the problem of a symmetric Toeplitz matrix of order 100 randomly generated using diagonal pivoting (`-c` option). The independent vector is also randomly generated. The command returns the execution time. Option `--help` outputs the available options that allow the command to receive, e.g. input data, or save the solution in a file. Other useful outputs of the command are related to some accuracy error. For example, the following call

```
tpsysv -n 100 -p --raw-results --raw-headers --random-seed=123
```

returns some accuracy error statistics:

```
#      n     Time (sec.)   Forward error    Backward error
#============================================================
     100         0.19          1.16e-13         3.02e-16
```

5 Future of StructPack

One of the target of this package is to disseminate its existence among the scientific community and also to include its maintenance.

In the future the working team will consolidate the existing methods, solving the problems that may arise from use by other researchers. In the solution of linear systems we are currently working on non-symmetric, complex and Hermitian linear systems of Toeplitz matrices, all of them based on the same approach used for the solution of the already tackled symmetric Toeplitz linear system. The same mathematical background used for the efficient implementation of the symmetric solver in multicores can be extended to the solution of a wide range of Toeplitz-like problems like block-Toeplitz, Toeplitz-block, Toeplitz+Hankel, etc. Furthermore, the least squares problem with Toeplitz-like matrices can also be solved with similar techniques so it will be part of our solver. The extension to other non Toeplitz-like but also structured matrices like Vandermonde are also part of our objectives.

The solution of the eigenproblem and the singular value problem involving structured matrices is also an important objective of the project. Some proposals are based on the solution of linear systems such as those mentioned above, so it only remains to be incorporated into our package.

Acknowledgements

This work was financially supported by the Spanish Ministerio de Ciencia e Innovación and by FEDER (Projects TIN2010-14971, TIN2008-06570-C04-02, TEC2009-13741 and CAPAP-H3 TIN2010-12011-E), Universitat Politècnica de València through Programa de Apoyo a la Investigación y Desarrollo (PAID-05-10) and Generalitat Valenciana through project PROMETEO/2009/013

References

[1] Arpack. `http://www.caam.rice.edu/software/ARPACK/`.

[2] Blas. http://www.netlib.org/blas/.

[3] FFTPack. http://www.netlib.org/fftpack/.

[4] GNU gcc. http://gcc.gnu.org/.

[5] Intel MKL. http://software.intel.com/en-us/articles/intel-mkl/.

[6] Lapack. http://www.netlib.org/lapack/.

[7] Mathematica. http://www.wolfram.com/mathematica/.

[8] Matlab. http://www.mathworks.com/products/matlab/.

[9] Netlib. http://www.netlib.no/netlib/toeplitz/.

[10] Petsc. http://www.mcs.anl.gov/petsc/petsc-as/.

[11] Scalapack. http://www.netlib.org/scalapack/.

[12] Slicot. http://www.slicot.org/.

[13] StructPack. http://www.inco2.upv.es/structpack.php.

[14] SuperLU. http://crd.lbl.gov/~xiaoye/SuperLU/.

[15] Alonso, P., Badía, J., and Vidal, A. Solving the block-Toeplitz least squares problem in parallel. *Concurrency and Computation: Practice and Experience 17* (January 2005), 49–67.

[16] Alonso, P., and Vidal, A. Cauchy-like system solution on multicore platforms. Proceedings of PARA 2008, NTNU.

[17] Bernabeu, M., Alonso, P., and Vidal, A. A multilevel parallel algorithm to solve symmetric Toeplitz linear systems. *The Journal of Supercomputing 44* (June 2008), 237–256.

[18] Bini, D. Toeplitz matrices, algorithms and applications. *ERCIM News*, 22 (July 1995).

[19] Bunch., J. Stability of methods for solving Toeplitz systems of equations. *SIAM J. on Scientific and Statistical Computing 6*, 2 (April 1985), 349–364.

[20] Graciá, L., Alonso, P., and Vidal, A. Solution of symmetric Toeplitz linear systems in GPUs. In *CMMSE'09 International Conference on Computational and Mathematical Methods in Science and Engineering* (Gijón (Asturias), España, June 2009).

[21] T.Kailath, and A.H.Sayed. Displacement structure: Theory and applications. *SIAM Review 37*, 3 (September 1995), 297–386.

 ISBN: 978-84-614-6167-7

[22] T.Kailath, and A.H.Sayed, Eds. *Fast Reliable Algorithms for Matrices with Structure*. SIAM, Philadelphia, PA, 1999.

[23] Vidal, A., García, V., Alonso, P., and Bernabeu, M. Parallel computation of the eigenvalues of symmetric Toeplitz matrices through iterative methods. *Journal of Parallel and Distributed Computing 68* (August 2008), 1113–1121.

[24] V.Olshevsky, Ed. *Fast Algoritmhs for Structured Matrices: Theory and Applications*. SIAM, Philadelphia, 2003.

 ISBN: 978-84-614-6167-7

Proceedings of the 11th International Conference on Computational and Mathematical Methods in Science and Engineering, CMMSE 2011 26–30 June 2011.

A numerical technique of cleaning in solitary-wave simulations

I. Alonso-Mallo[1], Ángel Durán[1] and Nuria Reguera[2]

[1] *Department of Applied Mathematics, University of Valladolid, Spain*

[2] *Department of Mathematics and Computation, University of Burgos, Spain*

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

Abstract

The paper introduces a numerical method to improve the simulations of solitary-wave solutions of some nonlinear dispersive wave equations. The properties that characterize the proposal are a local basis in the spatial discretization, an invariant preserving time integrator, a dynamical cleaning procedure which allows a simpler implementation and the automation of the whole process.

Key words: Solitary waves, perturbations, stability, conserved quantities, conservative methods, cleaning procedures
MSC 2000: 65M20, 65M99, 35Q53, 76B25

1 Introduction

The purpose of the paper is the description of a numerical technique to improve the simulation of solitary-wave solutions of nonlinear dispersive wave equations of the general form

$$u_t + u_x + f(u)_x + Mu_t = 0, x \in \mathbb{R}, t > 0. \tag{1}$$

where $u = u(x,t)$ is a real-valued function of the two real independent variables x, t, f is a smooth, real-valued function of u, representing a nonlinear term and M is a linear, nonnegative, formally self-adjoint operator, characterized as a Fourier multiplier operator by its symbol

$$\widehat{Mv}(\xi) = \alpha(\xi)\widehat{v}(\xi), \tag{2}$$

where $\widehat{\ }$ denotes Fourier transform. Equations of the form (1) appear in many models about the propagation of small-amplitude, nonlinear, dispersive long waves, as an alternative to the KdV-type equations, see e. g. [1, 2]. One of the most important

 ISBN: 978-84-614-6167-7

cases included in (1) is the BBM equation and generalized versions [5, 18, 9] (with $f(s) = s^{p+1}/(p+1), p \geq 1, M = -\partial_{xx}$),

Solitary-wave solutions of (1) are of the form $u(x,t) = \phi(x - ct)$ for some positive c, approaching some constant (zero in this case) at infinity. The parameter $c > 1$ represents the velocity of the wave. Assuming that $M\phi \to 0$ at $\pm\infty$, the profile $\phi = \phi_c$, whose explicit form is generally unknown, satisfies the equation

$$cM\phi_c - f(\phi_c) + (c-1)\phi_c = 0, \tag{3}$$

The simulation of these solitary waves requires to pay attention to some numerical problems (see e. g. [3, 4] and references therein). The one we focus on here appears frequently when managing with small perturbations of the waves, in the context of studies of stability (see e. g. [21, 19]). In some perturbations, numerical experiments suggest the evolution of the perturbed wave into a main pulse (or a train of pulses) along with dispersive tails (see Figure 1). If one is interested in the analysis of the parameters of the pulses, or in a way to generate multipulses of (3), then the finite computational window forces to 'clean' the solution: isolating the pulses, eliminating somehow and sometime the tails, and leaving them alone to evolve. Several cleaning procedures to this problem can be seen in the literature [12, 13, 6, 11, 7, 8, 15]. The main new computational feature of the technique proposed here is the automation of the whole process, from the selection of the interval of cleaning to the dynamic way this is implemented through the evolution. In this sense, equations of the form (1) are taking as a case study, being the technique applicable to other models for which the dynamics of solitary waves is under study.

The paper is structured as follows. In Section 2 we remind some properties of the equations (1) and complete the theoretical preliminaries. The numerical technique, along with the scheme of approximation to (1) is described in Section 3. Finally, Section 4 is devoted to show, taking the BBM equation as a case study, some numerical experiments concerning the application of the method to small perturbations of solitary waves. Extensions to the work concern the application to other equations and the simulation of the dynamics of other structures, such as periodic travelling waves, generalized solitary waves or multipulses.

2 Preliminaries

Several hypotheses on the nonlinear term f and the symbol α are assumed (see [21] among others).

(H1) f is C^2 with $f'(s) \geq 0$ for $s \geq 0$ and $f(s) = O(s^2), s \to 0$.

(H2) There are positive constants $m \geq 1/2, A_1, A_2$ such that

$$A_2|\xi|^{2m} \leq \alpha(\xi) \leq A_1(1 + |\xi|^2)^m.$$

 ISBN: 978-84-614-6167-7

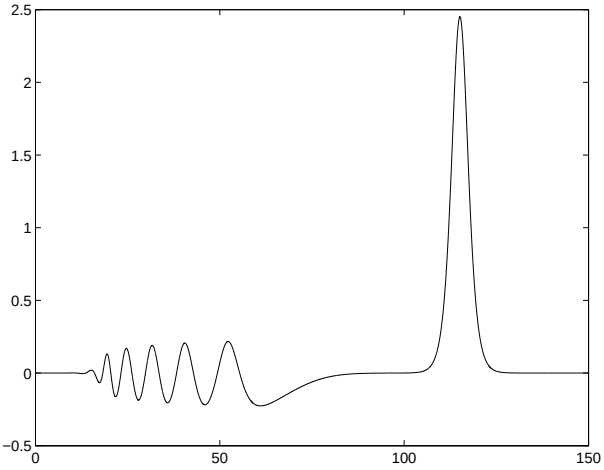


Figure 1: Solution from a perturbed solitary wave. BBM equation. Note the main wave and the generation of the tail behind.

They concern some well-posedness results of (1) in Sobolev spaces $X = H^s = H^s(\mathcal{R})$ with the corresponding norm

$$H^s = \{g, \int_{-\infty}^{\infty} (1+|\xi|^2)^s |\hat{g}(\xi)|^2 d\xi < \infty\}, \quad ||g||_s = \left(\int_{-\infty}^{\infty} (1+|\xi|^2)^s |\hat{g}(\xi)|^2 d\xi \right)^{1/2}$$

and $H^0 = L^2$ [1]. Note that condition (H2) implies that the operator M is nonnegative and self-adjoint. It also behaves essentially like $(-\partial_{xx})^m, m \geq 1/2$, which allows to define the operator L with symbol

$$\omega(\xi) = \alpha(\xi)/(i\xi), \quad \xi \neq 0, \quad \omega(0) = 0. \tag{4}$$

This will be used later on. On the other hand, for initial data in $H^s, s \geq s_0 > m$, the functionals

$$I(u) = \int_{\infty}^{\infty} u(x,t)dx, \tag{5}$$

$$V(u) = \int_{\infty}^{\infty} \frac{1}{2}(u(x,t)Mu(x,t) + u^2(x,t))dx, \tag{6}$$

$$H(u) = \int_{\infty}^{\infty} \left(\frac{1}{2}u^2(x,t) + F(u(x,t)) \right) dx, \tag{7}$$

are preserved by the solutions of (1), where $F' = f, F(0) = 0$ and the convergence of the integral that defines (5) is assumed. The quantity (7) determines the Hamiltonian structure of (1). As well, (6) is the Hamiltonian function of the Hamiltonian symmetry group of (1) consisting of translations in space [17]. This defines orbits of solutions of (3) $\{\phi_c(x - x_0) : x_0 \in \mathbb{R}\}$ and the corresponding solitary wave is obtained by applying the symmetry group with parameter ct to the profile $\phi_c(x - x_0)$.

 ISBN: 978-84-614-6167-7

3 Description of the numerical technique

Our proposal of cleaning involves the three stages of the numerical simulation: the generation of initial profiles, the spatial discretization and the time integration. First, the generation of some initial solitary wave profiles is necessary in order to define a support to guide the process. On the other hand, a local basis for the semidiscretization in space is required by the local character of the cleaning algorithm. Finally, the time integration must satisfy properties of conservation of invariants of the problem. This ensures [3] a more correct simulation of the parameters of the emerging solitary waves, with special emphasis on the amplitude and velocity. These are the main properties that the elements of the simulation must have to implement the cleaning procedure. In this sense, our specific choices below admit, under these requirements, other alternatives.

In order to adapt (1) to a finite computational window we consider the initial-periodic-boundary problem

$$\begin{array}{l} u_t + u_x + f(u)_x + Mu_t = 0, \quad x \in [0, L],\, t \geq 0, \\ u(0,t) = u(L,t) \\ u_x(0,t) = u_x(L,t) \\ u(x,0) = u_0(x), \quad x \in [0, L] \end{array} \tag{8}$$

In (8), the corresponding periodic version of the operator M is considered, defining (2) for the Fourier coefficients [20].

3.1 Generation of initial profiles

The cleaning procedure will require the generation of some initial solitary-wave profiles, by solving numerically equation (3). The operator M suggests to treat the problem in the Fourier space. Denoting by $\widehat{\phi_c}(k)$ the $k-$th Fourier coefficient of ϕ_c, the corresponding system for it, obtained from (3), has the form of fixed point condition

$$\widehat{\phi_c}(k) = \frac{\widehat{f(\phi_c)}(k)}{c - 1 + c\alpha(k)}, \quad k \in \mathbb{Z}, \tag{9}$$

There are several techniques in the literature to solve (9) iteratively. Here we choose the so-called Petviashvili's method, which has been proposed in several works as an efficient alternative to generate solitary wave profiles (see e. g. [16] and references therein). Note that if we multiply (3) by ϕ_c and integrate in $[0, L]$ then

$$K = K(\phi_c) = \frac{\int_0^L ((c - 1 + cM)\phi_c)\phi_c dx}{\int_0^L (f(\phi_c))\phi_c dx} = 1. \tag{10}$$

Now, using (10), the Petviashvili's method introduces a stabilizing factor and proposes the following algorithm to solve (9) iteratively

$$\widehat{\phi}(k)^{[\nu+1]} = K(\phi^{[\nu]})^{\gamma} \frac{\widehat{f(\phi^{[\nu]})}(k)}{c - 1 + c\alpha(k)}, \quad k \in \mathbb{Z}, \nu = 0, 1, \ldots, \tag{11}$$

 ISBN: 978-84-614-6167-7

where $\widehat{\phi}(k)^{[\nu]}$ is the ν-th iteration and γ is a free parameter chosen to make (11) be convergent [16].

For an even number N of nodes $x_j = jh, j = 0, \ldots, N$, with $h = L/N$, the practical implementation of (11) to approximate the corresponding nodal values of the profile can be done in the space S_h of N-th degree trigonometric polynomials defined on $(0, L)$. If Z_j denotes an approximation to $\phi_c(x_j)$, the p-th discrete Fourier coefficient

$$\widehat{Z}_p = \frac{1}{N} {\sum_{0\le j\le N}}'' Z_j e^{-ipj\tilde{h}}, \quad -\frac{N}{2} \le p \le \frac{N}{2}, \tilde{h} = 2\pi/N,$$

(the double prime in the sum means that the first and last terms are divided by two) is obtained by solving the iteration

$$\widehat{Z}_p^{[\nu+1]} = \widetilde{K}(Z^{[\nu]})^\gamma \frac{\widehat{f(Z^{[\nu]})}_p}{c - 1 + c\alpha(p)}, \quad -N/2 \le p \le N/2, \nu = 0, 1, \ldots, \tag{12}$$

where $\widetilde{K}(Z)$ is the approximation to the stabilizing factor K of the form

$$\widetilde{K}(Z) = \frac{\sum_{p=-N/2}^{N/2} (c - 1 + c\alpha(p))|\widehat{Z}_p|^2}{\sum_{p=-N/2}^{N/2} \widehat{(f(Z)}_p)\overline{\widehat{Z}_p}}.$$

(Recall that the implicit change of variables may affect the computation of the formulas by means of the scaling $2\pi/L$ [10]). Once (12) is iteratively solved with this technique, the approximation to the nodal values of the initial profile is recovered from the Fourier coefficients by evaluating, at the grid points, the trigonometric interpolation polynomial $Z_h \in S_h$ given by

$$Z_h(x) = {\sum_{-N/2\le p\le N/2}}'' \widehat{Z}_p e^{2\pi ipx/L},$$

in such a way that $Z_j = Z_h(x_j)$. Then it is ready to be used in the next stages of the simulation. Thus, the generation of the initial profile is performed in the Fourier space and then it can be adapted to the spatial discretization selected for (8). This implementation is not necessary if explicit formulas for the profiles are known as, for instance, in the case of the BBM equation and generalized versions.

3.2 Spatial discretization

The main property of the spatial discretization of (8), required by the cleaning technique, is the local character of the basis of discretization. In our case, for N, h and the nodes x_j previously defined in Subsection 3.1, we consider the space $\mathcal{V}_h$ of Hermite piecewise cubic polynomial functions on $[0, L]$ and satisfying periodic boundary conditions, see [3]. The Galerkin semidiscretization of the problem (8) is given by

$$\begin{array}{l} \langle (u_h)_t, w_h\rangle + \langle (Lu_h)_t, (w_h)_x\rangle = \langle u_h, (w_h)_x\rangle + \langle f(u_h), (w_h)_x\rangle, \quad \forall w_h \in \mathcal{V}_h,\ t \ge 0 \\ u_h(0) = u_{0,h} \end{array} \tag{13}$$

 ISBN: 978-84-614-6167-7

where the operator L is defined in (4). This spatial semidiscretization leaves to an ordinary differential system

$$R_h \frac{dU_h}{dt} = M_h U_h + F(U_h), \tag{14}$$

where the array $U_h = [u_0, \widetilde{u}_0, \ldots, u_{N-1}, \widetilde{u}_{N-1}]^T$ is formed by the approximations of the solution and its derivative at the nodes x_j. The form of M will sometimes require to treat the second inner product of the left hand side of (13) in an hybrid way: the operator L is implemented in the Fourier space, in order to obtain approximations to the corresponding nodal values (see Subsection 3.1) and then the result is projected on $\mathcal{V}_h$.

3.3 Time integration

The semidiscrete system (14) retains a Hamiltonian structure and has the corresponding discrete versions of the invariants (5)-(7) as conserved quantities [3]. This property plays a relevant role in the choice of the time integrator, since, in general, the use of geometric integration techniques ([14] and references therein) has influence to obtain good results when long time integration is required. In particular, the preservation of invariant quantities through the numerical integration has turned out to be an important tool to get a better simulation of the dynamics of solitary waves, including a more correct computation of the parameters characterizing the waves, such as the amplitude, velocity or phase (see e. g. [3, 4]).

In this sense, among all the possibilities cited in the literature, we have chosen the implicit midpoint rule as time integrator for the numerical experiments. This classical scheme [14] combines a good behaviour with respect to the invariants of (14), a simple implementation and a good adaptation to the cleaning algorithm. In this case, for a time step $\Delta t > 0$ and given an approximate value $U_h(t_n)$ to the solution of (14) at $t_n = n\Delta t$, the numerical solution at the next step $U_{h,n+1}$ is obtained by solving iteratively the nonlinear system

$$R_h G_{h,n} = \frac{\Delta t}{2} \left(M_h (U_{h,n} + G_{h,n}) + F_h (U_{h,n} + G_{h,n}) \right),$$

and then advancing with $U_{h,n+1} = U_{h,n} + 2G_{h,n}$.

3.4 Cleaning procedure

We now assume that the initial condition in (8) evolves into a main pulse along with some other small waves of possibly different nature and the experiment requires to isolate the pulse. This may occur, for instance, in numerical tests about the stability of solitary waves (see e. g. [9]). Our purpose is to automate, in an efficient way, a procedure to select dynamically a support of the wave and clean the solution out of it in a smooth way [6, 8]. The method assumes initially a relation between the velocity, the amplitude and the 'support ' of the solitary-wave solutions of (1). For us, the term 'support ' represents an interval where the profile is greater than a previously fixed

 ISBN: 978-84-614-6167-7

tolerance ϵ. This relation can be established, in general in approximate way, by using the techniques of generation of solitary waves for a range of velocities (Subsection 3.1).

We now describe the cleaning procedure from time t_{n-1} to t_n. The first step is to estimate, at time t_n, the amplitude of the main pulse and the point where this is attained. This can be done as follows:

1. A first initial point is taken from the velocity c_{n-1} of the numerical main wave and computed at the previous time t_{n-1}. This reduces the search of the point where the maximum of the numerical solution is attained to the nearest nodes (cf. [11]).

2. From this point, the value $x_{max}(n)$ where the cubic Hermite piecewise interpolant associated to the numerical solution reaches its maximum is computed analytically.

3. The nodes $x_{max}(n-1), x_{max}(n)$ and corresponding computed amplitudes can be used to estimate the velocity c_n at t_n.

The following step is to choose the cleaning region. To this end, we compute the interval $(\beta_1(n), \beta_2(n))$ centered at $x_{max}(n)$ where the values of the wave with velocity $c = c_n$ are greater than the tolerance ϵ, by using the previous numerical study with the solitary-wave profiles (notice that $\beta_1(n)$ y $\beta_2(n)$ could not be inside the computational window. In this case, we need to locate the suitable values in this window taking into account the periodic boundary conditions). The cleaning region is then the complementary of this interval in the computational window.

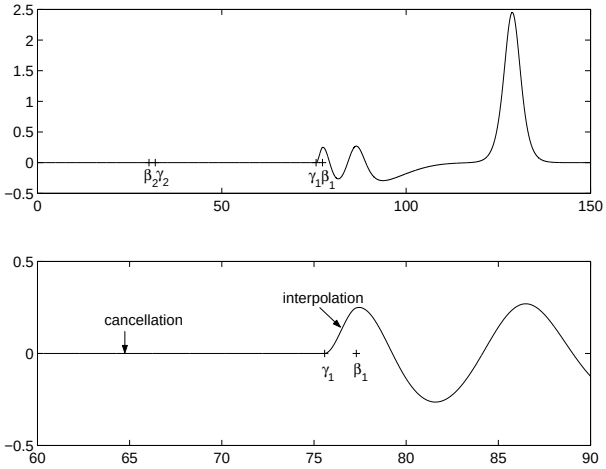


Figure 2: Procedure of cleaning with the smoothing technique. The figure at the bottom is a zoom of the one at the top.

Now, the cleaning algorithm may be completed in two ways [11]. The first one is setting all the nodal values u_j and $\tilde{u}_j$ located in this cleaning region equal to zero. This

 ISBN: 978-84-614-6167-7

seems quite efficient in many cases, specially for well separated waves. Sometimes it is necessary to clean in a smooth way (for instance, in the case of isolating a second pulse, see e. g. [11, 8]). Then, a new region $(\gamma_1(n), \gamma_2(n))$ where the solitary wave associated to the velocity c_n is greater than another tolerance $\epsilon_1 < \epsilon$ may be computed. Now, the solution out of $(\gamma_1(n), \gamma_2(n))$ is set equal to zero and in the intervals $(\gamma_1(n), \beta_1(n))$ and $(\beta_2(n), \gamma_2(n))$ a cubic interpolation is implemented. Figures 2 and 3 represent the smoothing process (here the values $\beta_1, \beta_2, \gamma_1, \gamma_2$ are not located in its natural order inside the computational window).

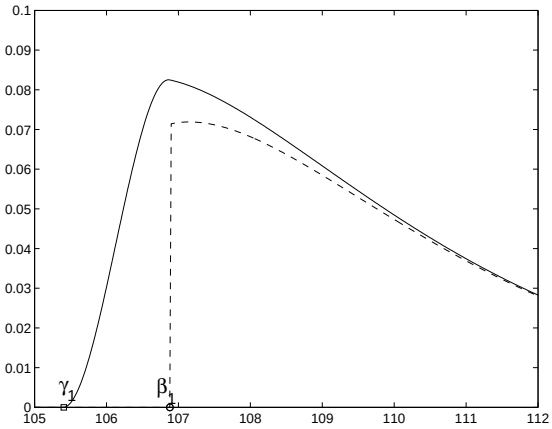


Figure 3: The two cleaning techniques. − Without smoothing, −− with smoothing.

4 Numerical experiments

In this section, we test the technique in perturbations of solitary-wave solutions of the BBM equation (equation (1) with $f(s) = s^2/2, M = -\partial_{xx}$). In this case, the relation between the amplitude and velocity is known and the explicit form of the solitary waves

$$u(x,t) = A\text{sech}^2(K(x - ct - L_0)), \quad A = 3(c-1), \quad K = \frac{1}{2}\sqrt{1 - \frac{1}{c}}, \tag{15}$$

allows to compute the support without the previous numerical generation of the profiles.

4.1 Perturbation of the parameters

We first consider, as initial condition, a perturbation in amplitude of a solitary wave

$$u_{0,j} = (1+\alpha)A\text{sech}^2(K(x_j - L/2)), \tag{16}$$

where A and K are given by (15), $c = 2$ and $\alpha = 0.5$. The evolution of this profile shows a type of experiments for which the cleaning may be necessary. The initial condition evolves into a main pulse, travelling to the right, with a small tail behind it. This is

 ISBN: 978-84-614-6167-7

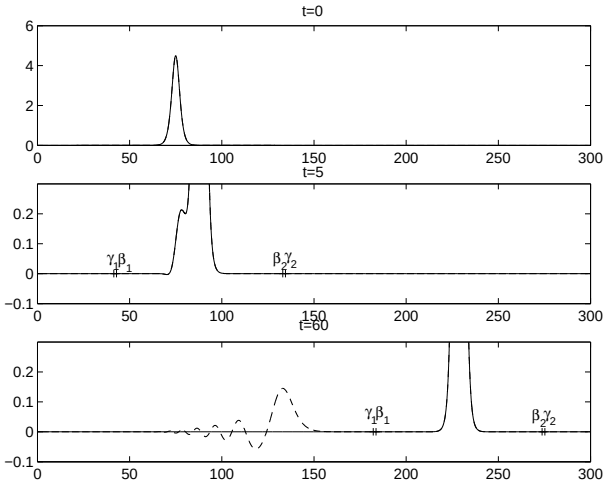


Figure 4: Modified initial profile (16). Numerical solution at different times: −− without cleaning, − cleaning with the smoothing technique.

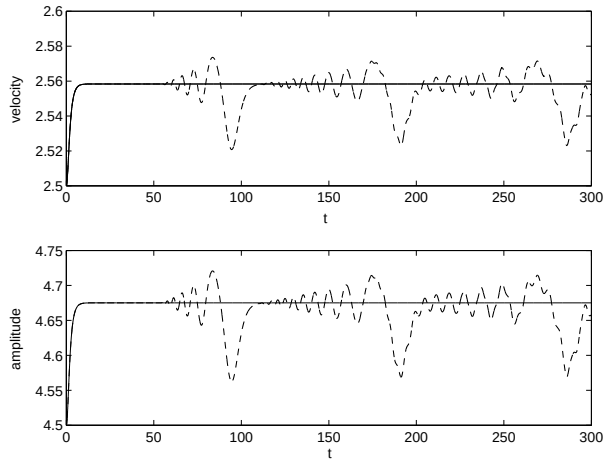


Figure 5: Modified initial profile (16). Time evolution of the velocity and the amplitude of the numerical solution: −− without cleaning, − cleaning with the smoothing technique.

 ISBN: 978-84-614-6167-7

observed in Figure 4, which shows the numerical solution at different times (without cleaning) and the corresponding cleaning process (in a smooth way). The application of the technique allows to analyze the perturbed main pulse in more detail. This is illustrated in Figure 5, which displays the differences in the computation of the velocity and amplitude of the uncleaned and (smoothly) cleaned numerical solution. Thus, the cleaning method provides a way to estimate the parameters of the main wave.

4.2 Perturbation of a solitary wave with a small gaussian profile

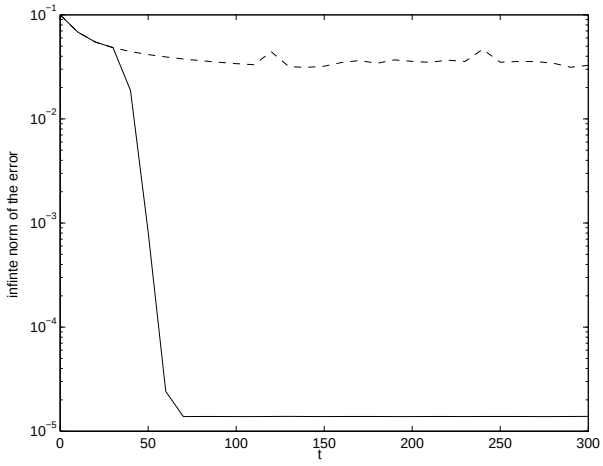


Figure 6: Modified initial profile (17) . Time evolution of error in the amplitude: $--$ without cleaning, $-$ cleaning with the smoothing technique.

The cleaning algorithm can also be applied when an exact solitary-wave profile is perturbed in the form

$$u_{0,j} = A\text{sech}^2(K(x_j - L/2)) + \alpha_1 \exp(-\alpha_2(x_j - L/2)^2), \tag{17}$$

where A and K are given by (15) with $c = 2$ and $\alpha_1 = 0.5$ and $\alpha_2 = 0.1$. The corresponding evolution also leads to a main pulse plus smaller tails behind. In Figure 6 we have calculated, for each t_n, the corresponding amplitude of the exact solitary wave with velocity $c = c_n$ and measured the evolution of the maximum difference with the numerical solution. This may give basic information about the size and type of the tails, as well as the evolution of the amplitude of the computed main wave. This complements Figure 7, which shows, for this experiment, the evolution of the velocity and amplitude of the cleaned and original numerical solution.

Acknowledgements

This research has been supported by MICINN project MTM2007-66343 cofinanced by FEDER funds and project MTM2010-19510/MTM.

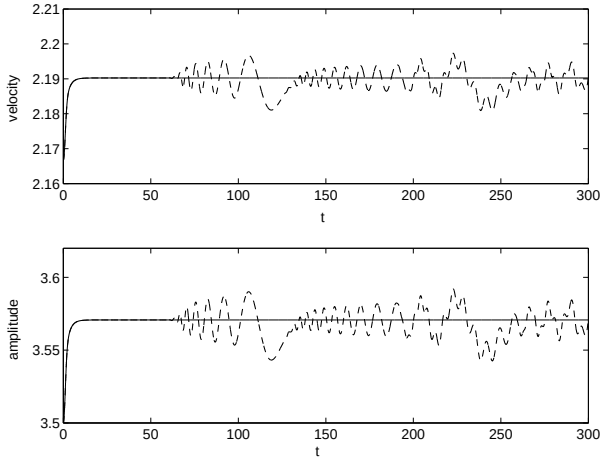


Figure 7: Modified initial profile (17). Time evolution of the velocity and the amplitude of the numerical solution: −− without cleaning, − cleaning with the smoothing technique.

References

[1] L. Abdelouhab, J. L. Bona, M. Felland and J. C. Saut, *Nonlocal models for nonlinear, dispersive waves*, Physica D **40** (1989) 360–392.

[2] J. P. Albert, J. L. Bona and J. C. Saut, *Model equations for stratified fluids*, Proc. R. Soc. London A **453** (1997) 1233–1260.

[3] I. Alonso-Mallo, A. Durán and N. Reguera, *Simulation of coherent structures in nonlinear Schrdinger-type equations*, J. Comput. Phys. **227** (2010), 8180–8198.

[4] J. Álvarez and A. Durán, *A numerical scheme for periodic travelling-wave simulations in some nonlinear dispersive wave models*, J. Comp. Appl. Math. **235** (2011), 1790–1797.

[5] T. B. Benjamin, J. L. Bona and J. J. Mahony, *Model equations for long waves in nonlinear dispersive systems*, Phil Trans R Soc Lond A **272** (1972), 47–78.

[6] J. J. Bona, M. Chen, *A Boussinesq system for two-way propagation of nonlinear dispersive waves*, Physica D **116** (1998), 191–224.

[7] J. L. Bona, V. A. Dougalis and D. E. Mitsotakis, *Numerical solution of KdV-KdV systems of Boussinesq equations I. The numerical scheme and generalized solitary waves*, Math. Comp. Simul. **74** (2007), 214–228.

 ISBN: 978-84-614-6167-7

[8] J. L. Bona, V. A. Dougalis and D. E. Mitsotakis, *Numerical solution of KdV-KdV systems of Boussinesq equations II. Generation and evolution of radiating solitary waves*, Nonlinearity **21** (2008), 2825–2848.

[9] J. L. Bona, W. R. McKinney and J. M. Restrepo, *Stable and unstable solitary-wave solutions of the generalized Regularized Long-Wave equation*, J Nonlinear Sci **10** (2000), 603–638.

[10] J. P. Boyd, *Chebyshev and Fourier Spectral Methods*, 2nd ed. Dover Publications, New York, 2000.

[11] V. A. Dougalis, A. Durán, M. A. López-Marcos and D. E. Mitsotakis, *A numerical study of the stability of solitary waves of the Bona-Smith family of Boussinesq systems*, J Nonlinear Sci **17** (2007), 569–607.

[12] J. D. Fenton and M. M. Rienecker, *A Fourier method for solving nonlinear water-wave problems: application to solitary-wave interactions*, J. Fluid Mech. **118** (1982) 411–443.

[13] B. Fornberg and G. B. Whitham, *A numerical and theoretical study of certain nonlinear wave phenomena*, Phil. Trans. Royal Soc. London A **289** (1978) 373–404.

[14] E. Hairer, C. Lubich and G. Wanner, *Geometric Numerical Integration, Structure-Preserving Algorithms for Ordinary Differential Equations*, Springer-Verlag, New York-Heidelberg-Berlin, 2002.

[15] M. Haragus, D. P. Nicholls and D. H. Sattinger, *Solitary wave interactions of the Euler-Poisson equations*, J. Math. Fluid Mech. **5** (2003), 92–118.

[16] T. I. Lakoba and J. Yang, *A generalized Petviashvili method for scalar and vector Hamiltonian equations with arbitrary form of nonlinearity*, J. Comput. Phys. **226** (2007) 1668–1692.

[17] P. J. Olver, *Applications of Lie Groups to Differential Equations*, 2nd ed. Springer-Verlag, New York-Heidelberg-Berlin, 1998.

[18] D. H. Peregrine, *Calculations of the development of an undular bore*, J Fluid Mech **25** (1996), 321–330.

[19] R. L. Pego and M. I. Weinstein, *Asymptotic stability of solitary waves*, Comm. Math. Phys. **164** (1994) 305–349.

[20] V. Thomée and A. S. Vasudeva Murthy, *A numerical method for the Benjamin-Ono equation*, BIT **38** (1998) 597–611.

[21] M. I. Weinstein, *Existence and dynamical stability of solitary-wave solutions of equations arising in long-wave propagation*, Comm Partial Diff Eq **12** (1987), 1133–1173.

 ISBN: 978-84-614-6167-7

Proceedings of the 11th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2011
26–30 June 2011.

On the influence of numerical preservation of invariants when simulating Hamiltonian relative periodic orbits

Jorge Álvarez[1] **and Ángel Durán**[1]

[1] *Department of Applied Mathematics, University of Valladolid, Spain*

emails: joralv@eii.uva.es, angel@mac.uva.es

Abstract

We study the structure of the error when simulating relative periodic solutions of Hamiltonian systems with symmetries. We identify the mechanisms for which the preservation, in the numerical integration, of the Hamiltonian and the invariants associated to the symmetry group, implies a better time behaviour of the error. A second consequence is a more correct simulation of the parameters that characterize the relative periodic orbit.

Key words: Hamiltonian relative periodic orbits, symmetry groups, geometric numerical integration, invariant preserving numerical methods
MSC 2000: 65L99, 65L05, 65Z05, 70H33, 37J15

1 Introduction

The aim of this paper is the analysis of long time numerical simulations of relative periodic solutions for canonical autonomous Hamiltonian systems

$$\dot{u} = J\nabla H(u), \quad u \in \Omega, \quad J = \begin{pmatrix} 0 & -I_n \\ I_n & 0 \end{pmatrix}, \tag{1}$$

where Ω is a domain of $\mathcal{R}^{2n}$, I_n is the identity matrix of order n and $H : \Omega \to \mathcal{R}$ is the Hamiltonian and ∇H denotes the gradient of H. The so-called geometric numerical integration is devoted to the construction and study of numerical methods specially designed to preserve physical properties of the system under study (see [7] and references therein for a modest representation of the literature on it). In this context, the paper studies the time behaviour of numerical approximations to Hamiltonian relative periodic solutions. These solutions may appear in Hamiltonian systems (1) admitting a group of transformations as a symmetry group [10]. This means, roughly speaking, that any element of the group takes solutions into other solutions. This symmetry group may have influence on the dynamics of the system by identifying points that belong to

 ISBN: 978-84-614-6167-7

the same group orbit and then considering the system for these group orbits. In this situation, solutions of the original system that project to special solutions of this reduced system are of interest in some applications. The literature pays specific attention to the relative equilibrium solutions (or RE, solutions that project to equilibria of the reduced system) and relative periodic solutions (or RPO, associated to periodic solutions of the reduced system), see [8, 9]. The long time simulation of RE has been studied in several papers [5, 3, 1]. They show the influence of the numerical preservation of invariant quantities of the system in the time propagation of the errors and the simulation of the parameters that characterize the RE. In this paper, we analyze the structure of this error when simulating RPO, studying the role of the reduction by group orbits and the periodicity in the reduced system in order to obtain a better long time simulation. We also establish the differences with respect to the numerical integration of RE.

The paper is structured as follows. The framework of Hamiltonian RPO is explained in Section 2. For simplicity, we will consider Hamiltonian one-parameter symmetry groups; the extension to the Abelian, Hamiltonian multi-parameter case is direct [5]. The mechanism of reduction by symmetries and the generation of RPO are described. In Section 3, the asymptotic behaviour of numerical approximations to RPO is analyzed and invariant preserving conditions for a better long time simulation are obtained. This section also contains some numerical illustrations of the results.

2 Hamiltonian relative periodic orbits (RPO)

We will describe the generation of Hamiltonian relative periodic solutions in a similar framework to that of [5, 1]. We suppose that (1) admits a first integral $I : \Omega \to \mathcal{R}$, different from the Hamiltonian H and which is not a distinguished function [10]. Thus, the Hamiltonian vector field $g = J\nabla I$ is the infinitesimal generator of a one-parameter symmetry group of (1), $\mathbf{G} = \{G_s = \sigma_{s,g} : s \in \mathcal{R}\}$, where, $\sigma_{s,g} = \exp(sg)$ denotes the flow associated to g. For simplicity, it is assumed that the domain of the flow φ_t of (1) and the diffeomorphisms $\sigma_{s,g}$ is the whole Ω. The condition for G to be a symmetry group can be established by using three equivalent conditions [5]:
(i) $\varphi_t(G_s(u)) = G_s(\varphi_t(u)), \quad s, t \in \mathcal{R}, u \in \Omega$.
(ii) $\{I, H\}(u) = \nabla I^T(u)(J\nabla H)(u) = 0, \quad u \in \Omega$.
(iii) $[J\nabla I, J\nabla H](u) = (J\nabla I)'(u)(J\nabla H)(u) - (J\nabla H)'(u)(J\nabla I)(u) = 0, \quad u \in \Omega$.
That is, the commutativity between the flow of the system and the elements of the group, the null Poisson bracket of the first integrals and the null Lie bracket of the corresponding vector fields. From now on, the prime denotes the corresponding Jacobian matrix.

2.1 Reduced system and RPOs

The generation of RPO is closely related to the action of the symmetry group G on the system. A solution $u(t) = \varphi_t(u_0)$ of (1) is a relative periodic solution or a relative periodic orbit (RPO) if there exists a positive $T_0 > 0$ (the relative period) such that

 ISBN: 978-84-614-6167-7

the solution at $t = T_0$ lies in the group orbit of the initial condition, that is

$$\varphi_{T_0}(u_0) = \sigma(u_0), \tag{2}$$

for some element $\sigma \in \mathbf{G}$, called phase-shift symmetry or drift symmetry [8]. Now, the symmetry group property implies that the solution must satisfy the condition of relative periodicity

$$\varphi_{t+T_0}(u_0) = \sigma(\varphi_t(u_0)), \quad t \in \mathcal{R}.$$

The Hamiltonian reduction [2, 10, 8] of (1) provided by the presence of the symmetry group G can be used in the search of RPO. This reduction of the system requires first foliating the phase space Ω with level sets of the invariant I. On each level set, the corresponding system has one fewer variable. Now, a second step identifies, on each level set, points that belong to the same group orbit, implying a reduction of one more variable. Thus, this reduced system is the Hamiltonian system for the group orbits in the corresponding level set. See e. g. [5] for a more detailed description of the process, including local coordinates and multi-parameter symmetry groups.

In a similar way that relative equilibrium solutions of (1) are related to the equilibria of the reduced system [2], the reduction connects relative periodic orbits with periodic solutions of the reduced system. This can be briefly described as follows (see [11] and references therein for details). Note first that the phase shift σ of (2) can be written as

$$\sigma_{\lambda_0 T_0, g}(u) = exp(T\xi)u, \quad \xi = \lambda_0 J \nabla I, \tag{3}$$

for some parameter λ_0. Then

$$\Phi_t(u_0) = exp(-t\xi)(\varphi_t(u_0)), \tag{4}$$

is a solution of

$$\dot{u} = J\left(\nabla H(u) - \lambda_0 \nabla I(u)\right), \quad u \in \Omega, \quad u(0) = u_0, \tag{5}$$

and it is T-periodic. Thus, every RPO is associated to a periodic solution of the Hamiltonian system (5). Thus, (4) can be interpreted as a change of variable to a frame moving uniformly with velocity ξ, where the relative period motion is transformed to a periodic motion. Inversely, it can be seen, by using the process of reconstruction [8], that any periodic solution of the reduced system corresponds to a RPO of the original one.

2.2 An example

The literature on the relative periodic solutions is very extensive. It includes many physical applications, such as rigid bodies, Celestial Mechanics, Molecular systems, Continuum Mechanics or Optics (for instance, the introduction of [12] includes an important number of references about it). As a modest example, we consider here the so-called Manev problem (for a description and references, se e. g. [11]), which is a

 ISBN: 978-84-614-6167-7

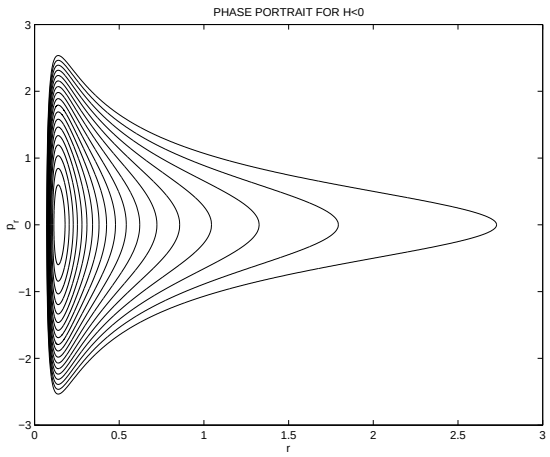


Figure 1: Phase portrait of a reduced system of the Manev problem. Note the cycles around an equilibrium (relative equilibrium of the original system). The presence of RPO near stable RE is very typical, see [12].

two-body problem (alternative to the classical Kepler problem) with a Hamiltonian of the form

$$H(p,q) = \frac{1}{2m}|p|^2 - V(q), \quad V(q) = \frac{A}{|q|} + \frac{B}{|q|^2},$$

where A, B are positive constants, $(p,q) = (p_1, p_2, q_1, q_2)$ and $m = M/(M+1)$, being M the mass of the (fixed) body and assuming that the other mass is one. Apart from the Hamiltonian, the corresponding system (1) admits, as the Kepler problem, the angular momentum $I(p,q) = p_1q_2 - p_2q_1$, as a second invariant. This is associated to the symmetry group of the system, consisting of rotations. In polar coordinates, the reduced Hamiltonian on the level set $\{I(p,q) = \mu\}$ has the form

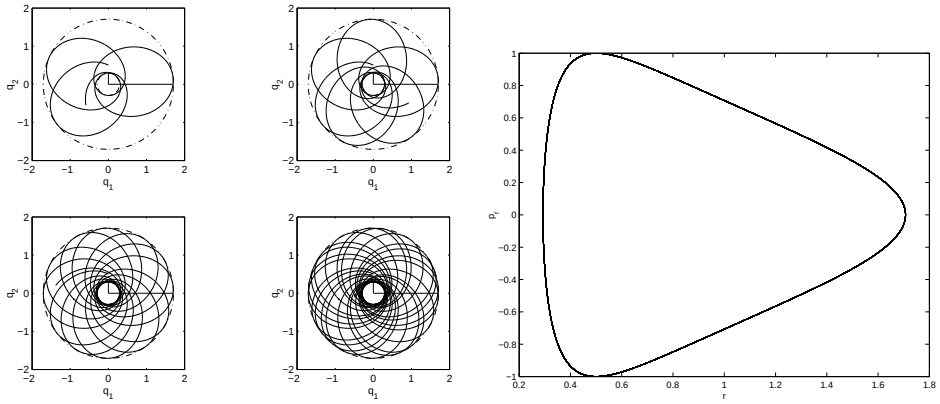
$$H_\mu(p_r, r) = \frac{p_r^2}{2m} - V_\mu(r), \quad V_\mu(r) = \frac{A}{r} + \frac{\mu^2 - 2mB}{2mr^2}.$$

It can be seen that when $\mu^2 - 2mB > 0$, bounded motions occur in the reduced system. In particular, for values of the energy $H \in (H_c, 0)$, with $H_c = -mA^2/2(\mu^2 - 2mB)$, the motion in the corresponding reduced phase space is periodic. This is illustrated in Figure 1, which shows the phase portrait of one of these reduced systems. Thus, the system describes periodic orbits, that lift to relative periodic motions in the original phase space. The details can be seen in e. g. [11], where the phase shift symmetry and the relative period are also computed

$$\sigma = e^{i\Delta\varphi}, \quad \varphi = \frac{2\pi\mu}{\sqrt{\mu^2 - 2mB}}, \quad T = \frac{A\sqrt{m}\pi}{\sqrt{2}|H|^{3/2}}. \tag{6}$$

The form of one relative periodic solution in the configuration space and its projection on the corresponding level set of I (the periodic orbit in the reduced phase space) can be seen in Figure 2.

 ISBN: 978-84-614-6167-7



(a) (b)

Figure 2: (a) RPO for the Manev problem and for several times: $\mu = 1$, $m = 1$, $A = 1$, $B = 1/8$. (b) Projection of the RPO on the level set $\{I = \mu\}$: periodic orbit of the reduced phase space.

3 Numerical behaviour in RPO simulations

In this section we will discuss the time propagation, at leading order approximation, of the errors when simulating relative periodic solutions. Assume that a one-step method of order $r \geq 1$

$$u_{n+1} \quad = \quad \psi_h(u_n), \quad n = 0, 1, \ldots, \tag{7}$$

for a mapping $\psi_h : \Omega \to \Omega$ and stepsize h, is used to approximate an RPO $u(t) = \varphi_t(u_0, T_0, \lambda_0)$ of (1), satisfying (2) with initial data $u_0 = u_0(\mu_0, H_0)$, relative period $T_0 = T_0(\mu_0, H_0)$ and phase shift σ of the form (3) for some $\lambda_0 = \lambda_0(\mu_0, H_0)$. The values of the momentum and the Hamiltonian at the initial data are denoted, respectively, by μ_0, H_0. The hypotheses on (7) are standard, including local and global error expansions at the solution and invariance of the function ψ_h by the symmetry group $\mathbf{G}$, see [5]. As for the global error expansion, this is written in the form

$$u_n - u(t_n) \quad = \quad h^r e(t_n) + h^r Q(t_n, h), \tag{8}$$

where e is the solution of the variational problem [6]:

$$\dot{e}(t) = JH''(u(t)) \cdot e(t) - l(u(t)), \quad e(0) = 0, \tag{9}$$

Q is a remainder that, for fixed t, tends to zero as $h \to 0$ and l is the leading term of the asymptotic expansion of the local error. On the other hand, the hypothesis that the mapping ψ_h is invariant by the symmetry group implies that [5]

$$l(G_s(u)) = G_s'(u) l(u), \quad s \in \mathcal{R}, \quad u \in \Omega. \tag{10}$$

 ISBN: 978-84-614-6167-7

3.1 Main result

Here we state the main result of the paper. The proof requires several steps.

Theorem 3.1 *Under previous conditions and assuming the hypotheses of Lemmas 3.2, 3.3, 3.4 and 3.5 below, the numerical approximation to the relative periodic orbit* $u(t) = \varphi_t(u_0, T_0, \lambda_0, \sigma_0)$ *at times* $t_N = NT_0$ *can be written as*

$$u_N = G_{t_N\tilde{\lambda}}(\Phi_{N\tilde{T}}(\tilde{u})) + h^r\Gamma(t_N) + h^r q(h, t_n), \tag{11}$$

where Φ_t *is given by (4) and*

$$\tilde{\lambda} = \lambda\left(\mu_0 - \frac{1}{2}\left(\frac{t_N}{T_0}+1\right)h^r\theta_1, H_0 - \frac{1}{2}\left(\frac{t_N}{T_0}+1\right)h^r\theta\right) + \frac{\alpha_1 h^r}{T_0}, \tag{12}$$

$$\tilde{T} = T\left(\mu_0 + \frac{1}{2}\left(\frac{t_N}{T_0}+1\right)h^r\theta_1, H_0 + \frac{1}{2}\left(\frac{t_N}{T_0}+1\right)h^r\theta\right) + \frac{\alpha h^r}{T_0}, \tag{13}$$

$$\tilde{u} = u\left(\mu_0 + \frac{t_N}{T_0}h^r\theta_1, H_0 + \frac{t_N}{T_0}h^r\theta\right), \tag{14}$$

for some constants $\alpha_1, \theta_1, \alpha, \theta$*. If the method satisfies the orthogonality conditions*

$$\nabla H(\sigma_0(u_0))^T e(T_0) = \nabla I(\sigma_0(u_0))^T e(T_0) = 0, \tag{15}$$

then $\theta_1 = \theta = 0$*. In particular, (15) holds if the method (7) preserves both quantities, that is,* $H(\psi_h(u_0)) = H(u_0)$*,* $I(\psi_h(u_0)) = I(u_0)$*.*

Furthermore, Γ *is a function that, if the RPO is linearly stable (as periodic solution of the reduced system) and* $\mathbf{G}$ *consists of isometries, is bounded for* $t \geq 0$*. The function* q *is a remainder that, for fixed* t*, tends to zero as* $h \to 0$*.*

Proof.

Step 1. Global error. We note first that, using (10) and the change of variables

$$e(t) = \exp(t\xi)' \left(\Phi_t(u_0)\right) \Delta(t), \tag{16}$$

then (9) can be transformed into

$$\dot{\Delta}(t, \epsilon) = J(H'' - \lambda I'')(\Phi_t(u_0)) \cdot \Delta(t) - l(\Phi_t(u_0)), \quad \Delta(0) = 0,$$

where $\Phi_t(u_0)$ is given by (4). Equation (16) can be seen as a linearization of (5) at the periodic function $\Phi_t(u_0)$ with a nonhomogeneous term. Thus, in order to study the time behaviour of $\Delta(t)$, it is sufficient to analyze its values at multiples of the period T_0. If $\Delta^{(N)} = \Delta(NT_0)$ for some integer $N \geq 1$, then [4]

$$\Delta^{(N)} = \left(\sum_{i=1}^{N-1} M^i\right)\Delta^{(1)}, \quad \Delta^{(1)} = \Delta(T) = \int_0^T \widetilde{M}(T, s) l(\Phi_s(u_0)) ds, \tag{17}$$

 ISBN: 978-84-614-6167-7

where $\widetilde{M}(t,s) = \Phi'_{t-s}(\Phi_s(u_0))$ and the monodromy matrix M has the form

$$M = \tilde{M}(T_0, 0) = \Phi'_{T_0}(u_0) = (\sigma'(u_0))^{-1}\varphi'_{T_0}(u_0). \tag{18}$$

Then, (16) and (17) imply

$$e^{(N)} = e(NT_0) = \exp(NT_0\xi)'\,(u_0)\,\Delta^{(N)}. \tag{19}$$

Therefore, the solution of the variational problem, at the RPO (9) evaluated at multiples of the relative period is a product of two terms growing with time: The Jacobian matrix $\exp(NT_0\xi)'\,(u_0)$ and the solution of (17) at $t_N = NT_0$.

Step 2. Structure of the monodromy matrix. The study of M can be made with the following technical results, see [12] and references therein for the details.

Lemma 3.2 *The vectors* $g_1(u_0) = J\nabla I(u_0)$, $g_2(u_0) = J(\nabla H(u_0) - \lambda_0 \nabla I(u_0))$ *are eigenvectors of* M *with eigenvalue* 1.

Note also that if the RPO is proper, that is, u_0 is not a relative equilibrium, then these vectors are independent.

Let V be now the unique $M-$ invariant supplementary subspace of the generalized eigenspace of M associated to the eigenvalue 1,

$$\mathcal{R}^{2n} = Ker_g(M - I) \oplus V. \tag{20}$$

Observe, on the other hand, that since $\varphi_t(u_0)$ projects to a periodic solution $\Phi_t(u_0)$ of the reduced system, we may consider the corresponding monodromy matrix M_R of its linearization at the periodic solution. Here we assume that the RPO is nondegenerate as periodic solution of the reduced system, in the sense that the algebraic multiplicity of one as eigenvalue of M_R is two [4]. Then, taking local coordinates, we have [12]

Lemma 3.3 *If* W *is the (unique) supplementary subspace satisfying*

$$\mathcal{R}^{2n-2} = Ker_g(M_R - I) \oplus W,$$

then the eigenvalues and Jordan structure of the restriction $M\Big|_V$ *coincide with those of* $M_R\Big|_W$.

Now, the following result is a consequence of a more general persistence theorem proved in [12]. It establishes that, near a RPO, a continuum of relative periodic orbits can be defined, with the family parameterized by the values of the invariants.

Lemma 3.4 *Under the above conditions, assume that* $P = \{g(\varphi_t(u_0)),\, g \in G,\, t \in \mathcal{R}\}$ *is nondegenerate as periodic solution of the reduced system. Let* $\mu_0 = I(u_0)$ *and* $H_0 = H(u_0)$. *Then, for* (μ, H) *close to* (μ_0, H_0), *there is a RPO* $P(\mu, H) = \{g(\varphi_t(u(\mu, H))),\, g \in G,\, t \in \mathcal{R}\}$, *with* $\varphi_{T(\mu,H)}(u(\mu, H)) = \sigma(\mu, H)(u(\mu, H))$, *for smooth* $\sigma(\mu, H)$, $T(\mu, H)$, $u(\mu, H)$ *satisfying*

$$\begin{aligned}
&u_0 = u(\mu_0, H_0), \quad T_0 = T(\mu_0, H_0),\\
&I(u(\mu, H)) = \mu, \quad H(u(\mu, H)) = H,\\
&\sigma(\mu, H) = \exp\left(\lambda(\mu, H) J\nabla I(u(\mu, H))\right), \quad \lambda_0 = \lambda(\mu_0, H_0).
\end{aligned}$$

 ISBN: 978-84-614-6167-7

Finally, Lemmas 3.2, 3.3, 3.4 complete the description of the structure of M [12] (see also [5] to compare with the case of relative equilibria).

Lemma 3.5 *With the notation and hypotheses of Lemmas 3.2, 3.3, 3.4, we assume that the matrix*

$$D = \begin{pmatrix} \frac{\partial \lambda}{\partial \mu}|_{\mu_0,H_0} & \frac{\partial \lambda}{\partial H}|_{\mu_0,H_0} \\ -\frac{\partial T}{\partial \mu}|_{\mu_0,H_0} & -\frac{\partial T}{\partial H}|_{\mu_0,H_0} \end{pmatrix},$$

is nonsingular. Then $dimKer(M-I) = 2$, $dimKer(M-I)^2 \backslash Ker(M-I) = 2$ *and the vectors* $\{g_1(u_0), g_2(u_0), g_3(u_0), g_4(u_0)\}$, *with* g_1, g_2 *given in Lemma 3.2 and*

$$g_3(u_0) = \frac{\partial u}{\partial \mu}|_{\mu_0,H_0}, \quad g_4(u_0) = \frac{\partial u}{\partial H}|_{\mu_0,H_0},$$

form a basis of $ker_g(M-I)$ *satisfying*

$$\begin{aligned} (M-I)g_3 &= T_0 \frac{\partial \lambda}{\partial \mu}|_{\mu_0,H_0} g_1(u_0) - \frac{\partial T}{\partial \mu}|_{\mu_0,H_0} g_2(u_0), \\ (M-I)g_4 &= T_0 \frac{\partial \lambda}{\partial H}|_{\mu_0,H_0} g_1(u_0) - \frac{\partial T}{\partial H}|_{\mu_0,H_0} g_2(u_0), \end{aligned}$$

and the biorthogonality conditions

$$\nabla I(u_0)^T g_3 = 1, \quad \nabla I(u_0)^T g_4 = 0, \quad \nabla H(u_0)^T g_3 = 0, \quad \nabla H(u_0)^T g_4 = 1. \tag{21}$$

Step 3. Application to the global error expansion. Proof of Theorem 3.1. Now, we apply these results to the linearized problem (17) and incorporate the conclusions to the global error expansion to prove Theorem 3.1. Following (20) and Lemma 3.5, we first decompose the term $\Delta^{(1)}$ in (17) in the form

$$\Delta^{(1)} = \Delta(T_0) = \alpha_1 g_1(u_0) + \alpha g_2(u_0) + \theta_1 g_3(u_0) + \theta g_4(u_0) + \Delta_V, \tag{22}$$

with $\Delta_V \in V$ and where (21) implies

$$\theta = \nabla H(u_0)^T \Delta^{(1)}, \quad \theta_1 = \nabla I(u_0)^T \Delta^{(1)}. \tag{23}$$

The substitution of (22) in (17) and the previous results lead to

$$\begin{aligned} \Delta^{(N)} &= N\left(\alpha_1 g_1(u_0) + \alpha g_2(u_0)\right) + \theta_1 \left(N g_3(u_0) + \frac{N(N-1)}{2}(M-I) g_3(u_0) \right) \\ &\quad + \theta \left(N g_4(u_0) + \frac{N(N-1)}{2}(M-I) g_4(u_0) \right) + \left(\sum_{i=1}^{N-1} M^i \right) \Delta_V. \end{aligned} \tag{24}$$

We denote $G'_{t_N \lambda_0} = \exp(N T_0 \xi)'(u_0)$. Using (19) and (24), we write

$$\begin{aligned} \varphi_{t_N}(u_0) + h^r e(t_N) &= \varphi_{t_N}(u_0) + G'_{t_N \lambda_0} \Delta^{(N)} = G_{t_N \lambda_0}\left(\Phi_{t_N}(u_0)\right) \\ &\quad + G'_{t_N \lambda_0} \left(\Delta^{(N)} - \left(\sum_{i=1}^{N-1} M^i \right) \Delta_V \right) + G'_{t_N \lambda_0} \left(\sum_{i=1}^{N-1} M^i \right) \Delta_V. \end{aligned} \tag{25}$$

 ISBN: 978-84-614-6167-7

Now, (22) and (4) prove that the first two terms differ from the term $G_{t_N\tilde{\lambda}}(\Phi_{N\tilde{T}}(\tilde{u}))$, with $\tilde{\lambda}$, $\tilde{T}$, $\tilde{u}$ given respectively by (12), (13) and (14), in $o(h^r)$ terms, that can be included in the remainder of (8). This leads to (11) if we take

$$\Gamma(t_N) = G'_{t_N\lambda_0}\left(\sum_{i=1}^{N-1} M^i\right)\Delta_V.$$

Note then that if the group consists of isometries, then $||G'_{t_N\lambda_0}|| = 1$. Furthermore, if the RPO is linearly stable, as periodic solution of the reduced system, then the other Floquet multipliers have modulus one and are simple; thus, the component of $\Delta^{(N)}$ in the supplementary subspace V is bounded in time. This implies the bounded behaviour in time of $\Gamma(t_N)$. The rest of the theorem comes from (23) and the fact that the preservation of the invariants implies (15) is standard, see e. g. [5]. □

Remarks

1. From (11)-(14), we observe that, in a leading term approximation (order $O(h^r)$) and at multiple values of the relative period, the numerical solution consists of three terms. The first one is a modified relative periodic orbit, with initial condition and relative period given, respectively, by (14) and (13). This factor is transformed by the element of the group with a new phase shift given by (12). Note that the difference with respect to the exact relative periodic solution at $t_N = NT_0$ grows, in general, quadratically with time, due to the terms $t_N\tilde{\lambda}$, $N\tilde{T}$. If the method satisfies both orthogonality conditions (15), this growth is reduced to linear. In particular, this holds if the integrator preserves I and H. This could be compared with the case of approximations to relative equilibria [1, 3, 5], where the preservation of only one of the invariants is sufficient to obtain linear error growth.
2. The second term in (11) consists of the differential of the group at the RPO and of the component of the error in the direction determined by the eigenvalues of the linearization which are different from one. As mentioned in the proof, if the RPO is linearly stable, then this element is bounded and the growth of the second term is controlled by the behaviour of the group. In the typical case of isometries (as in the example of Subsection 2.2), then the complementary term does not grow with time and the time behaviour is that of the modified RPO and the remainder.
3. Finally, this remainder includes $o(h^r)$ terms, whose behaviour is not uniform. This may affect the numerical solution, in the sense that, depending on N and h, it may limit the dominance of the behaviour of the modified RPO in (11).

3.2 A numerical illustration

We illustrate the previous results with a numerical example. The test problem will be the Manev problem, introduced in Subsection 2.2. Our goal is to observe the influence, in the numerical simulation of a RPO, of the preservation of the invariant quantities of the problem, by illustrating the results of Theorem 3.1. To this end, we will consider

 ISBN: 978-84-614-6167-7

three numerical integrators: the first one is the simply diagonally implicit Runge-Kutta method of order three and tableau

$$\begin{array}{c|cc} \frac{3+\sqrt{3}}{6} & \frac{3+\sqrt{3}}{6} & 0 \\ \frac{3-\sqrt{3}}{6} & \frac{-\sqrt{3}}{3} & \frac{3+\sqrt{3}}{6} \\ \hline & \frac{1}{2} & \frac{1}{2} \end{array} \tag{26}$$

which will be denoted as [SD]. This is taken as an example of nonconservative integrator, since it does not preserve any of the two invariants of the problem. This scheme has been modified to preserve the momentum I. Among all the techniques, presented in the literature, to do this, we have selected the so-called projection technique (see [6, 7] and references therein). The resulting method is denoted by [SDI]. We have also considered the corresponding scheme, obtained with the same technique but designed to preserve H. This method, denoted as [SDH], gives similar results to those of [SDI] and they will not be shown here (see comments in Subsection 3.2.2). Finally, a third method, [SDIH], is designed as the previous one, but with the projection technique ensuring the preservation of both quantities, momentum I and Hamiltonian H.

We have simulated the evolution of a RPO with parameters $m = 1$, $A = 1$, $B = 1/8$, $\mu_0 = 1$, $H_0 = -1/2$. The relative period and the phase shift are given by (6). The simulation is performed up to a final time of two hundred times the relative period $t_N = 200T_0$.

3.2.1 Error growth

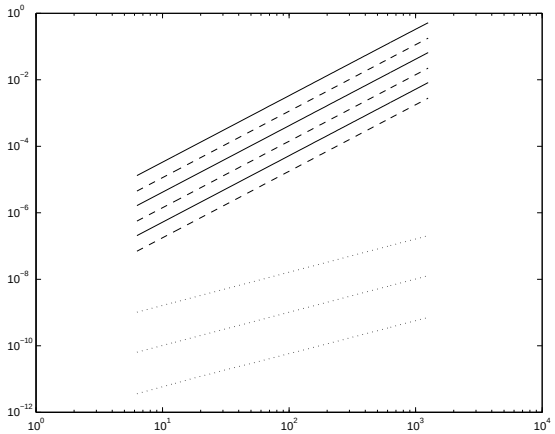


Figure 3: Error vs time for the Manev problem: [SD] (solid lines), [SDI] (dashed lines) and [SDIH] (dotted lines). The stepsizes are $h = 2E-03, 1E-03, 5E-04$.

A first point to show the influence of the preservation of the invariants is given in Figure 3. This shows, in logarithmic scale, the global error between the numerical

 ISBN: 978-84-614-6167-7

solution and the RPO, at multiples times of the relative period, as a function of time and for different values of the stepsize. We observe that for the nonconservative [SD] (solid lines) and the method [SDI], that only preserves the momentum (dashed lines), the slopes of the lines show that the growth of the error is quadratic with time. This is improved by [SDIH], that preserves both I and H (dotted lines) and gives linear error growth. This illustrates formulas (11) and (15) and may be compared with the behaviour when simulating relative equilibria [5]. It is therefore necessary to preserve both quantities to improve the time behaviour of the error.

3.2.2 Simulation of the parameters

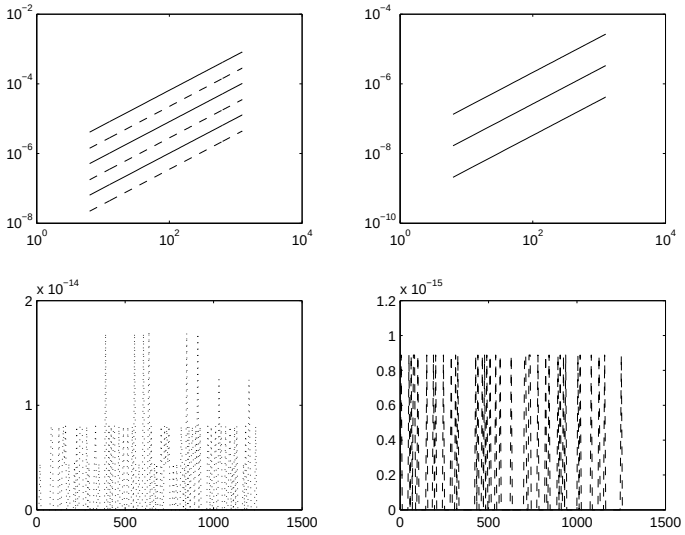


Figure 4: Manev problem. Left: error in relative period vs time. Up: [SD] (solid lines), [SDI] (dashed lines) with $h = 2E-03, 1E-03, 5E-04$ (log scale); down: [SDIH] with $h = 2E-03$. Right: error in the phase shift vs time. Up: [SD] (solid lines), with $h = 2E-03, 1E-03, 5E-04$ (log scale); down: [SDI] with $h = 2E-03$.

The illustration of formulas (12) and (13) is given by Figure 4. On the left, we measure (again in logarithmic scale) the time behaviour of the error in the relative period, provided by the three methods (with the same values and rules as those of Figure 3). Figure of the right corresponds to the behaviour of the error in the phase shift. In both cases the benefits of the preservation of both invariants are again observed, providing, for moderately long times, a better simulation of the parameters that characterize the relative periodic orbit. Thus, the numerical solution behaves essentially as a RPO with perturbed relative period and phase shift. The perturbation in the relative period grows linearly with time, with respect to the original parameters, in the case of nonconservative methods (like [SD]) or that preserve I, like [SDI] (Figure 4, left and up). The simulation of the parameters is more correct when the scheme conserves

 ISBN: 978-84-614-6167-7

the Hamiltonian, since the relative period only depends on H, see (6). This is also confirmed by the results of [SDIH] (left, down) or [SDH] (not shown here). On the contrary, since the phase shift in (6) only depends on μ, the linear in time perturbation provided by [SD] (Figure 4, right and up) is improved by the good simulation of [SDI] (right, down) and [SDIH] (not shown here), because both preserve the momentum.

Acknowledgements

This research has been supported by JCYL project VA060A09 and MICINN projects MTM2008-00700/MTM, MTM2010-19510/MTM.

References

[1] J. Álvarez and A. Durán, *Error propagation in numerical approximations near relative equilibria*, J. Comput. Appl. Math. **234** (2010) 3373–3386.

[2] V.I. Arnold, *Mathematical Methods of Classical Mechanics*, Springer, 1989.

[3] B. Cano and A. Durán, *A technique to improve the error propagation when integrating relative equilibria*, BIT **44** (2004) 215–235.

[4] B. Cano and J. M. Sanz-Serna, *Error growth in the numerical integration of periodic orbits, with application to Hamiltonian and reversible systems*, SIAM J. Numer. Anal., **34** (1997) 1391–1417.

[5] A. Durán and J.M. Sanz-Serna, *The numerical integration of relative equilibrium solutions. Geometric theory*, Nonlinearity **11** (1998) 1547–1567.

[6] E. Hairer, S.P. Norsett and G. Wanner, *Solving Ordinary Differential equations I. Nonstiff Problems*, 2nd ed. Springer Series in Computational Mathematics 8, Springer-Verlag, 1993.

[7] E. Hairer, Ch. Lubich and G. Wanner, *Geometric Numerical Integration. Structure-Preserving Algorithms for Ordinary Differential Equations*, Springer Series in Comput. Mathematics, Vol. 31, Springer-Verlag 2002.

[8] J.E. Marsden, *Lecture on Mechanics*, Cambridge University Press, 1992.

[9] J.E. Marsden and T.S. Ratiu, *Introduction to Mechanics and Symnmetry*, Springer, 1994.

[10] P.J. Olver, *Applications of Lie Groups to Differential Equations*, Springer, 1986.

[11] J.P. Ortega, *Symmetry, Reduction and Stability in Hamiltonian Systems*, Ph.D. Thesis, University of California, 1998.

[12] C. Wulff and M. Roberts, *Hamiltonian systems near relative periodic orbits*, SIAM J. Appl. Dyn. Syst. **1** (2002) 1-43.

 ISBN: 978-84-614-6167-7

Proceedings of the 11th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2011
26–30 June 2011.

An Efficient Java-Based Multhreaded and GPU port of an Implementation based On A Secure Multicast Protocol

J.A. Alvarez-Bermejo[1] and J.A. Lopez-Ramos[2]

[1] *Department of Architecture of Computers and Electronics, University of Almeria*

[2] *Department of Algebra and Analysis, University of Almeria*

emails: jaberme@ual.es, jlopez@ual.es

Abstract

We introduce a protocol for secure multicast using GPU. The protocol is based on an orthogonal system whose elements are distributed individually to every user as a ticket for accessing the encrypted information. We use GPU to generate the orthogonal system and to refresh cryptographic keys that protect the distributed information.

Key words: Gram-Schmidt, Java, Threaded, GPU, Secure Multicast

1 Introduction

Most of security protocols are mainly applicable when communications take place between two single parties. Confidentiality may be reached by means of a secret key that can be interchanged in a public channel using a public key cryptosystem. These are usually known as unicast communications. However there exist many different situations where the usual secure unicast protocols cannot be used, mainly due to the nature of the information to be transmitted. A typical example occurs when trying to deliver data from a sender to multiple receivers, especially when a huge amount of data needs to be delivered very quickly. Examples of these are video-conferencing, pay-per-view TV or internet radio. One of the most efficient ways to do this is the so-called multicast. In a multicast protocol a certain group of people receives the information and may also act as a source of data. This group is usually highly dynamic, where users join and leave the group constantly ([6]).

The typical approach to establish secure multicast communications is to agree on one or several symmetric encryption keys in order to encrypt messages. However, the key, or keys, must be renewed periodically to prevent outer or inner attacks.

Our aim in this paper is to deal with a centralized multicast scheme, i.e., a single entity distributes every cryptographic key. We can encounter different approximations

 ISBN: 978-84-614-6167-7

to the centralized schemes. On one hand we have those know as tree-based schemes. A very well-known protocol is *Hierarchical Tree Approach* (HTA) [9].

It uses a logical tree arrangement of the users in order to facilitate key distribution. The benefit of this idea is that the storage requirement for each client and the number of transmissions required for key renewal are both logarithmic in the number of members.

A second group of protocols known as secure multicast framework uses some intermediate participants that act as trustable agents. The most well-known of this type of protocols is the so-called IOLUS ([5]) where a large multicast group is partitioned into subgroups and employing a hierarchy of group security agents to "relay key and encrypt data.

Finally there is a third group of protocols based on number theoretic approaches. Examples of this type are *Secure Lock*, introduced in [1] or *Euclides* in [4], where the authors base their solution in the Chinese Remainder Theorem and in the existence of modular inverses respectively. In the case of Euclides authors show that scalability is better in every aspect when comparing with most of previous protocols, excepting the length of messages.

Most used protocols currently are those that combine the distribution of users by trees or groups and number theoretic approaches in order to deal with the main drawbacks of all of them (for instance cf. [7]). In the first case, the big amount of keys that have to be stored and secondly, the big computational overcomes or the huge length of the refreshing messages. For instance the computing time at the key server becomes problematic in Secure Lock as the number of member grows [2] or the the length of the refreshing messages grows linearly with the number of users in order to avoid factorization attacks [4].

The distribution by groups is in fact often beneficial and is used by most key managing protocols. A first benefit is the parallelization of the process which speeds up the rekeying operations. Secondly a compromised key in one of the groups does not affect the others and, last but not least, in most applications of secure multicast the group distribution is connected with the scalability of the system, i.e., the efficiency of the communication protocols concerning the rekeying process, with particular reference to leave and join operations. Groups are usually highly dynamic and the joining or the leaving of users implies a rekeying operation, and thus key refreshment due to this fact in one group does not affect the others.

Our aim in this paper is to deal with the implementation of a geometrical approach introduced in [3] that allows to distribute a cryptographic key group using only a refreshing message per group and where users store a single key in order to access at the common secret. We introduce an implementation on a GPU of this method. GPU have shown to be of undoubted help when managing problems with high computing cost as usually occur in cryptography (cf. [8] or [10], not only to obtain the result of the processing in much less time, but also to free up CPU resources, that can be crucial when using the contents distribution server also as key server itself.

 ISBN: 978-84-614-6167-7

2 The protocol

Through this section we introduce the protocol for secure multicast discussed in [3].

We will denote the users with the integers $1, \ldots, n$

1. Initialization step:

   Let $\mathbb{K}$ be a field and V be a $\mathbb{K}$ vector space of dimension $m \geq n$ with an inner (scalar) product, say $<,>$ and consider $B = \{e_1, \ldots, e_n\}$ a set of n mutually orthogonal vectors in V. We select then a family $\{x_i\}_{i=1}^n$ of random scalars in $\mathbb{K}$. Note that $B' = \{x_1 e_1, \ldots, x_n e_n\}$ spans the same subspace as B (unless $x_i = 0$ for some i). These two sets are kept secret by the server and each user i is assigned the vector $v_i = x_i e_i$.

2. Sending the information:

   Let $s \in \mathbb{K}$ be the secret to be distributed. Then we compute the vector $x_1 e_1 + \ldots + x_n e_n$ and multicast $c = s(x_1 e_1 + \ldots + x_n e_n)$.

3. Recovering the secret:

   Each user computes $h =< c, v_i >= s < v_i, v_i >$ and the secret s is recovered by computing $s = h < v_i, v_i >^{-1}$.

4. Key refreshment:

   (a) Join:

   If user j joins, then she is assigned one of the vectors in B' that is not being used by another user, say $x_j e_j$. The server selects a new secret $s' \in \mathbb{K}$ and multicasts $c' = s'(x_1 e_1 + \ldots + x_n e_n)$.

   (b) Leave:

   If user j leaves, then her vector $v_j = x_j e_j$ is deleted from B' and a new set B'' is considered formed by the same vectors in B' but substituting v_j with $v'_j = x'_j e_j$ where x'_j $(\neq x_j)$ is selected at random in $\mathbb{K}$. A secret s' is then distributed as before.

3 Key Sharing: Implementation details

Regarding the implementation of the above cited key sharing protocol, it requires to be developed under certain constrained bounds. One of the most limiting bounds is that the solution should be architecture neutral due to probable internal representation issues between clients and server. Another key to consider during development is that the solution should seize all the processing units available in the server in order to produce an efficient distribution scheme. Regarding this last issue, multicore platforms are driving a new direction in software development where multithreaded applications are elected as the style-to-follow development rules.

 ISBN: 978-84-614-6167-7

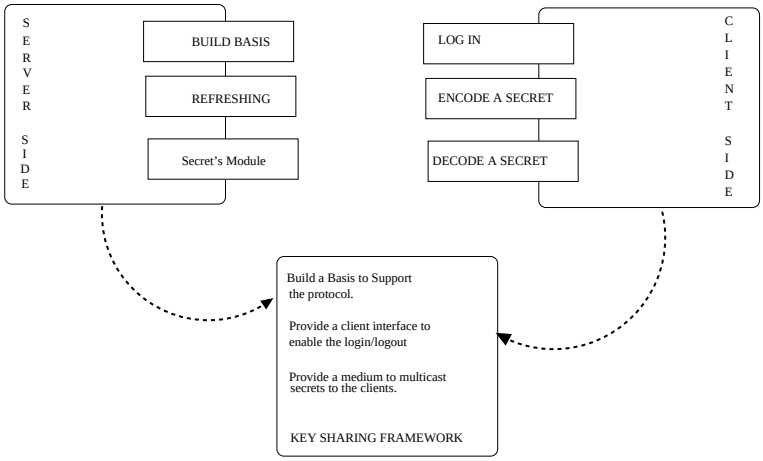


Figure 1: Sequential implementation schema

Another issue to evaluate is the use of hardware accelerators (i.e. GPU) to act as an intense matrix computing device, so the main memory and processing resources are reserved to management tasks. Java was selected as a first tool to develop and deploy the solution for several reasons: it is object oriented, which is an important issue if considering parallel schemes [12], because this methodoloy prints a natural parallel communication based on the flow of inter objects messages. As the multicore platform is involved the threading part of the language is also an important characteristic to consider. In this case, Java is a language but is also a virtual machine, so threads are not native, this means that the threading package model that Java uses is on the user side and is a light-weight package, this is directly translated into fast context-switches and user control.

3.1 Sequential implementation

We have implemented a sequential approach, which may be considered as the seed of the threaded versions, including the accelerator-based version. We may use this case to explain the implementation details. As sketched in Figure 1, there are three main objects, **the Key Sharing Framework object** (KSF from now) that acts as a controller during the application life cycle, this object is in charge of invoking services on the **Server object** which hosts the 2D-matrix that represents the vector space and on the **Client object**. In this case the KSF invokes services on the server and client sequentially, which is the main difference with the multithreaded version. The server object is the hotspot in terms of computation due to the huge matrix it hosts. Several matrix frameworks were tested such as EJML, JAMA, jampack and ujmp. EJML was selected due to a better behaviour when compared with the rest of frameworks, as it is shown in figure 2. We used several aspects inherited from EJML such as the condensed matrix navigation [13]. One of the key aspects we seized from EJML was

 ISBN: 978-84-614-6167-7

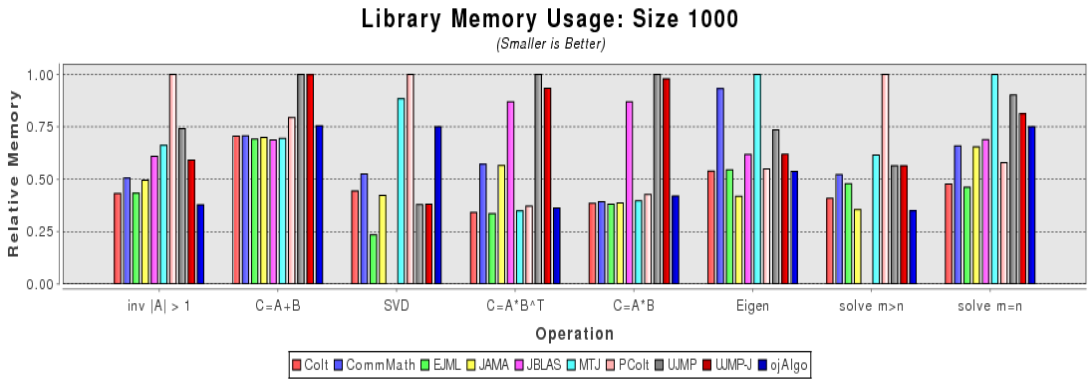


Figure 2: Memory foot print of the matrix model used (EJML Benchamrks).

the small memory foot print when managing huge and dense matrix data structures, this accelerates accessing it, and favours cache. The implementation was divided in several significative stages:

- *Configuration of the vector space*: this creates the 2D matrix, users will be provided each one of the 1D vectors contained in this data structure. This is a hot spot to be accelerated. This stage is composed by two important sub-stages: orthonormalization (hard to accelerate due to its data dependencies) and denormalization, stage used to accomodate the data to the key sharing protocol. The initial matrix, prior to orthogonalization, was built using a method inherited from the *Semantic Vectors* [14] package that accelerates the build in a 10% to 20% of the normal procedure.

- *Content Generator Coder*: this is the key (calculated as the reduction of each column in the main matrix (previous stage) that is used by the server (content provider) to code the content itself. Whenever a user is removed or a security issue is informed, this key must be refreshed. If the user is removed, its vector must be replaced in the main matrix and then the content generator coder is recalculated. If a security issue is reported, then the whole bunch of vectors are replaced, so the content generator coder is refreshed. This stage is clearly threadable.

- *User/Client login*: this stage consists in creating data structures for each user that claims a key to decode content provided by the server. In this phase we run an authentication mechanism (password based). Once the client is authorised to log in, we provide it the key needed to decode. This stage is clearly threadable.

- *Server initialization and startup*: Once the structure is ready, the KSF framework opens the server to accept petitions from clients, eliminate clients, accept new clients, key refreshing, etc.

 ISBN: 978-84-614-6167-7

3.2 Multicore enabled: Threaded Version

Taking Figure 1 as a base, all the computational workload was distributed between cores by the Operating System (OS in advance) itself. If we consider that we are operating on a Java Virtual Machine, it is no hard to assume that the OS is doing a coarse-grain distribution of tasks between cores. An easy way of making life easier to the OS is by creating threads that can be easily identified as an independent workload unit that can be attached to a different core. In this version, methods that were intially sequential in the previous subsection, are threaded now. So, to serve as an example, the KSF facilities regarding login/logout are threaded so clients can concurrently be added or deleted. This implies that the method in the server object regarding the key refreshment is to be reentrant and operated by rows.

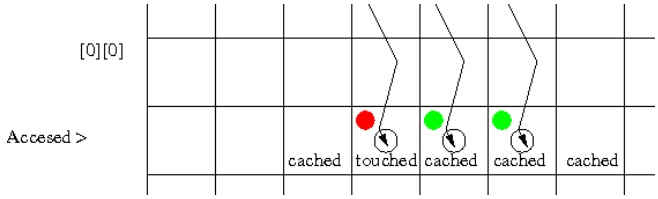


Figure 3: Threading the content generator coder

To enable the threading, we used a pool of threads, whose threads were waiting in the background their reactivation by filling a pool of tasks. This gave us enormous flexibility to enable concurrence. The code was modified by moving the methods to be executed in threads into a new method type, the *Task method*, the body of the method was marked as *Runnable* and the method was interpreted as a separate unit (like an object) that could be inserted as a task into the task's data structure. Doing this, we could thread the *content generator coder* in a easy way. As stated above, this key is a vector where each element is the reduction of the corresponding column in the 2D matrix. We built a task to traverse a matrix in column order (to host the 2D Matrix we used an ArrayList, protected as adviced in [11]) and created as many tasks of this type as columns we configured, see Figure 3, if threads are sharing caches then this method seizes the cache when reading components.

The User login facilities only needed a simple threading (reads) in row order, so we simply used the ArrayList in its conventional way to retrieve a row for each user or client. In this case, each request sent by a client was queued in the tasks list. Another benefit from our task list is that it is an heterogeneous task list, we can insert in it any method by simply renaming it to be a Task method. This version still has a hard to solve issue, it is how to accelerate the orthogonalization of the 2D Matrix. To do this one must consider that each vector (k) to be orthogonalized (Gram-Smichdt) depends on the orthogonalization of the previous *k-1* vectors. Guessing a way to split the work in such a way that we can feed any available core is hard. The results obtained were not significantly good, so as this process is calculated only the first time the KSF framework is started we decided to keep it single threaded.

 ISBN: 978-84-614-6167-7

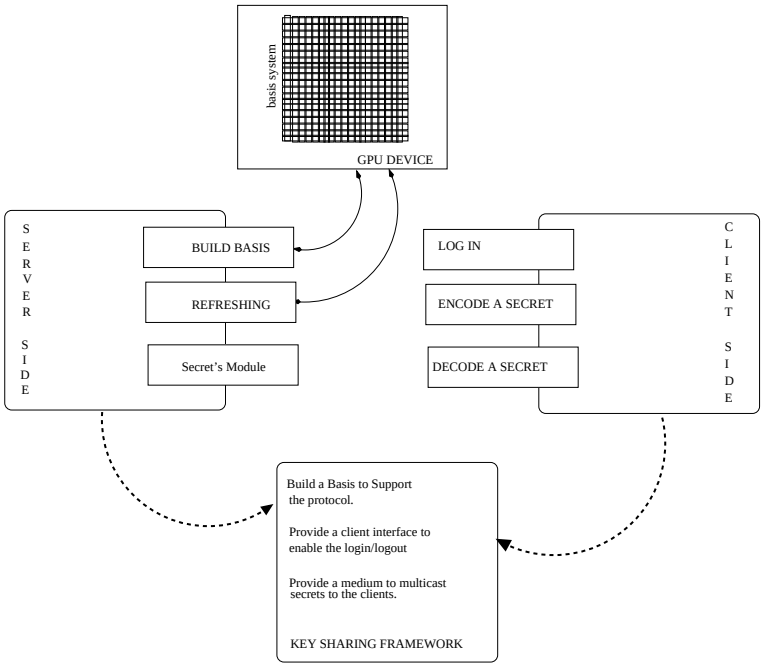


Figure 4: Cuda implementation schema

3.3 Cuda: Accelerator based version

When trying to scale such a system, it is advisable to leave the main processing unit in charge of the KSF object and let the Server object derive its computation to a GPU device that may act as a back-end service. When referring to matrix computations, GPUs come to mind. In this implementation we are using jcuda [15] to interface the GPU and impersonate it as a new computational object to which we can send computational demands, see Figure 4, and wait for the results without having to spend cpu cycles in this task. The NVIDIA GPU used is a GForce GTX-460, which is a Fermi's architecture graphics card. The 2D matrix used to build the vector's space is built randomly, in the sequential and multithreaded versions we used a method (short[] to float[] expansion) to accelerate the process but this method is not necessary when using GPUs. There were two options to have the 2D Matrix created in the device: build it in the CPU and send the data through the PCI to the NVIDIA device which was a bad idea because the program was not experiencing any improvement due to the latencies, or simply write a kernel that using a random number generator could allocate space for the 2D matrix and populate it. The experimental result was achieved with this last option. We used the Dynamic Creator Mersenne Twister random number generator. Using the GPU we reduced the orthogonalization in approximatedly 90% of the time.

 ISBN: 978-84-614-6167-7

4 Results

In this section we discuss the results obtained with the implementation of the key sharing protocol proposed. The tables below shows the most significative cases in everyone of the stages in which the protocol implementation was divided. We have run our first experimental tests on two platforms: A Intel Coreo 2-duo T9500 processor with 6MB cache L2, two hardware threads, a bus-core ratio of 13 and 4GB of RAM. The second platform used was a Intel Core-i7 Extreme-Edition Sandy Bridge, with 6 cores and 12 hardware threads, 12MB L2 cache. 8GB of RAM, and Intel TurboBoost technology. The BUS technology is QPI. This core-i7 is equipped with a NVIDIA GForce GTX 460, used to accelerate the orthogonalization process up to a 90%.

Table 2 shows the execution of the cases where we used a 10000vectors base (that is a 2D matrix of 10000x10000) that can host upto 10000 users. We have run tests from 10x10 to 10000x10000, and only the bigger cases are shown, smaller tests are always around zero. In this case it is significative that the real hot spot is, as anounced before, the orthogonalization, that as expected, have worst times as the 2D matrix grows. The following stages were threaded with java threads, the thread pool was designed to have *only 12 java threads* running at every instant so we could make a fair use of the 6 cores and the two hardware threads per core. If the Key refreshment stage (which is the phase where the 2D Matrix is updated row by row to get them multiplied by a random number so each row is converted into a potential key for a client) and the Content Generator coder stage (which is the phase where the 2D matrix is traversed column by column to get the sum of each component in each column) have similar times this means that the threading is acting properly, see the case in Table 2 with 10000 vectors and 5000 users and the case of 10000 vectors and 10000 users, the time is kept due to the pool of threads.

Table 1: Execution of the protocol in its sequential version on a core i7 sandybridge

stage	5000v x 5000u	10000v x 5000u	10000v x 10000u
Orthogonalization	114223	902736	912511
Key Refreshment	70	287	370
Generator Coder	70	287	370
Server activation	2	1	3
Client's setup	46	80	168
Bcast	52	67	73
Client removal + refresh	1	2	2

The sequential table (Table 1 is presented to compare the first method with the threaded implementation.

To test the architectural benefits of the architecture, where the tests in Table 2, we launched the server in a conventional processor: core 2 duo. Although the trend reflected in table 3 follows those studied in table 2, but times are worst. This is due to certain main reasons: the number of hardware threads per core is one, so no native

 ISBN: 978-84-614-6167-7

Table 2: Execution of the protocol in its threaded version on a core i7 sandybridge

stage	5000v x 5000u	10000v x 5000u	10000v x 10000u
Orthogonalization	114240	913866	913596
Key Refreshment	30	197	198
Generator Coder	30	197	198
Server activation	0	1	0
Client's setup	1	1	2
Bcast	12	16	33
Client removal + refresh	1	1	1

concurrence is available. The design of the pool of threads, only two active threads are running in any instant of time. Another significative difference is the bus (which is much slower than the IQP used for the processor in Table 2).

Table 3: Execution of the protocol in its threaded version on a dual core T9500

stage	5000v x 5000u	10000v x 5000u	10000v x 10000u
Orthogonalization	169909	1339753	1371083
Key Refreshment	66	231	235
Generator Coder	66	231	235
Server activation	2	1	3
Client's setup	46	80	168
Bcast	1	1	0
Client removal + refresh	45	300	556

One of the key aspects of this problem that we implemented is the client's removal and the time it takes to renew its key so it can be used by a new client logged in the system. This time is reflected in the stage named *Client's setup*. As it can be seen, the time employed to refresh a client is not dependent on the size of the problem. Threads can help this operation to scale if the server find bursts of client's removal operations.

5 Conclusions

The protocol presented is fast but relies on the computational representation of vectos, therefore the performance of the solution is on the computer architecture and memory organization underneath. Here we have presented a sequential technique to implement the proposed protocol and have improved the results by using techniques to accelerate matrix operations, techniques to thread code and keep threading operations stable by using pools of threads, mapped to hardware threads. Also we implemented a kernel for CUDA devices that expand the 2D matrix and implement the orthogonalization.

 ISBN: 978-84-614-6167-7

Acknowledgements

First contributing author is supported by grants from the Spanish Ministry of Science and Innovation (TIN2008-01117), and Junta de Andalucía (P08-TIC-3518), in part financed by the European Regional Development Fund (ERDF). Second contributing author is supported by the Spanish Ministry of Science and Innovation (TEC2009-13763-C02-02) and Junta de Andalucía (FQM 0211).

References

[1] G. Chiou and W. Chen, *Secure broadcasting using the secure lock*, IEEE Trans. Softw. Eng. **15**(8) (1989) 929–934.

[2] P. S. Kruus and J. P. Macker, *Techniques and issues in multicast security*, Proceedings of Military Communications Conference, MILCOM 1998, 1028–1032.

[3] J.A. Lopez-Ramos, J. Rosenthal, D. Schipani, *Managing key multicasting through orthogonal systems,* preprint.

[4] J. A. M. Naranjo, N. Antequera, L. G. Casado and J. A. Lopez-Ramos, *A suite of algorithms for key distribution and authentication in centralized secure multicast environments*, To appear in J. Comput. Appl. Math., doi:10.1016/j.cam.2011.02015.

[5] S. Mittra, Iolus, *A framework for Scalable Secure Multicasting,* Proc. ACM SIGCOMM97, 1997, 277-88.

[6] S. Rafaeli and D. Hutchison, *A Survey of Key Management for Secure Group Communication*, ACM Computing Surveys, **35**(3) (2003), 309–329.

[7] O. Scheikl, J. Lane, R. Boyer and M. Eltoweissy, *Multi-level secure multicast: the rethinking of secure locks*, Parallel Processing Workshops, 2002. Proceedings. International Conference on, 2002, 17–24.

[8] R. Szerwinski, T. Güneysu, *Exploiting the Power of GPUs for Asymmetric Cryptography,* Cryptographic Hardware and Embedded Systems. CHES 2008 Lecture Notes in Computer Science, 2008, Volume 5154/2008, 79–99.

[9] D. Wallner, E. Harder and R. Agee, *Key management for multicast: Issues and architectures*, RFC 2627, 1999.

[10] J. Yang, J. Goodman, *Symmetric Key Cryptography on Modern Graphics Hardware,* Advances in Cryptology. ASIACRYPT 2007 Lecture Notes in Computer Science, 2007, Volume 4833/2007, 249–264.

[11] Shavit, N. *Data Structures in the Multicore Age* Communications of the ACM**54**(3), 2011, 76–84

[12] J.A. Alvarez and J. Roca Piera and J.J. Fernandez *From structured to object oriented programming in parallel algorithms for 3D image reconstruction* Proceeding POOSC '09 Proceedings of the 8th workshop on Parallel/High-Performance Object-Oriented Scientific Computing ACM ISBN: 978-1-60558-547-5

[13] H. Ardnt *The Universal Java Matrix Package: Everything is a Matrix!* Proceeding POOSC '09 Proceedings of the 8th workshop on Parallel/High-Performance Object-Oriented Scientific Computing ICML/MLOSS, Haifa, 2010-06-25

[14] Widdows, D.; Cohen, T. *The Semantic Vectors Package: New Algorithms and Public Tools for Distributional Semantics* Semantic Computing (ICSC), 2010 IEEE Fourth International Conference on pp.9-15, 22-24 Sept. 2010

[15] Wenjun Fan; Xudong Chen; Xuefeng Li *Parallelization of RSA Algorithm Based on Compute Unified Device Architecture* Grid and Cooperative Computing (GCC), 2010 9th International Conference on , vol., no., pp.174-178, 1-5 Nov. 2010

Proceedings of the 11th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2011
26–30 June 2011.

Pairings and Secure Multicast

N. Antequera[1] and J.A. Lopez-Ramos[1]

[1] *Department of Algebra and Analysis, University of Almeria*

emails: nicolas.antequera@gmail.com, jlopez@ual.es

Abstract

We introduce a new for multicast distribution of secrets based on pairings, which allows low cost of communications and key storage, improving some alternatives existing for real time communications.

Key words: pairing, elliptic curve, secure communications
MSC 2000: AMS codes (optional)

1 Introduction

Traditional security measures are mainly applicable to a unicast environment, i.e. communications take place between two single parties. For instance, data confidentiality, one of the most important features in network security, can be offered in this environment by means of a pair of keys. However there exist many different situations where usual secure unicast protocols cannot be used, mainly due to the nature of the information to be transmitted. Multicast communications allow a host to simultaneously send information to a set of other hosts, avoiding the establishment of point-to-point connections with all of them. Applications of secure multicast are, among others, pay-per-view IPTV or P2PTV, private multiconferences (oriented to business, politics or even military affairs), or any private service that involves several participants or clients.

Key management is a crucial issue in secure multicast. Not only security is required, but also efficiency in the key management, which includes key storage and refreshment and perfect and backward secrecy, i.e., a new (resp. old) user should not be able to access the contents before (resp. after) she joins (leaves) the multicast group. Depending on how key distribution and management are carried out, secure multicast schemes are divided into centralized and distributed. Centralized schemes depend directly on a single entity to distribute every cryptographic key.

A very well-known protocol is *Hierarchical Tree Approach* (HTA), (cf. [1]) is the recommended option. It uses a logical tree arrangement of the users in order to facilitate key distribution. The benefit of this idea is that the storage requirement for each client

 ISBN: 978-84-614-6167-7

and the number of transmissions required for key renewal are both logarithmic in the number of members. Other key tree approaches and extensions are LKH [2], LKH++ [3], OFT [4] or ELK [5].

In [6] the so-called *Secure Lock* protocol is introduced. The authors approach the problem in a computational manner rather than a tree arrangement. It is based on the Chinese Remainder Theorem, being its main drawback the inefficient computations required at the key server side on each rekey operation: the computation time needed quickly becomes excessive when the number of members grows [7].

In [8], a divide-and-conquer extension to Secure Lock is proposed. It combines the Hierarchical Tree Approach and the Secure Lock: members are arranged in a HTA fashion, but Secure Lock is used to refresh keys on each tree level. Therefore, the number of computations required by Secure Lock is reduced.

Another computational approach is introduced in [9] with the particular application on Pay-TV but extendable to any other secure multicast application. It is based on an interpolator polynomial over some secret values in a finite field hold by the authorized users and making use of hash functions. The main drawback is that the implementation of this protocol needs changing the hash function used with every rekeying due to security matters in the definition of the aforementioned polynomial.

Finally and more recently, in [10], the authors introduce a novel method for distributing keys in a centralized secure multicast based on the Extended Euclidean Algorithm. It is shown that the behavior of this new approach is much better than the others in what respects to number of keys stored on both the server and the user and number of messages in a rekeying operation. However the length of a rekeying message grows linearly with respect to the sum of the every user's secret key length what represents a weakness for this. We have to note that similar situations occurs also in the other computational approaches.

The aim of this paper is to give a new computational solution based on [10] with the same good behavior concerning rekeying and reducing the length of the messages to refresh the key.

2 The protocol

We will start this section by recalling some mathematical background that is needed to introduce the proposed protocol. Although pairings are defined in different settings such as modules over a ring we will focus in the case of elliptic curves.

Given an elliptic curve over a finite field $\mathbb{F}_p$, say E, a pairing is a computable mapping $S : E \times E \to \mathbb{F}_{p^k}$ satisfying bilinearity, i.e., $S(P+Q,R) = S(P,R)S(Q,R)$ and $S(P,Q+R) = S(P,Q)S(P,R)$ for every P, Q and R in E and non-degeneracy, i.e., $S(P,P) \neq 1$, for every $P \in E$. Examples of these mappings are the well-known Weyl or Tate pairings and their use in public cryptography is extended not only for constructing new encryption protocols as Identity Based encryption or signature (for example cf. [11, Chapter X]), but also even with hacking purposes as the pairing attack by reducing the Elliptic Curve Discrete Logarithm Problem (ECDLP) to the Discrete

 ISBN: 978-84-614-6167-7

Logarithm Problem (DLP).

Initial setup: each client i is assigned a different element $x_i \in \mathbb{F}_{p^k}$ and a point $P_i \in E$. The pair (x_i, P_i) constitutes the *ticket*.

Rekeying message:

1. The Key Server selects a secret value: $K \in \mathbb{F}_{p^k}, H \in E$ and an integer n in the rank of $\mathbb{F}_p$.
2. The Key Server calculates $P = \sum_{i=1}^{m} P_i \in E$. P is kept private in the Key Server.
3. The Key Server calculates $S(H, P)$, $S(nH, P)$, $\alpha_i = S(H, P - P_i)$ for every $i = 1, \ldots, n$ and $p(x) = n + \Pi_{i=1}^{m}(x - x_i\alpha_i) \in \mathbb{F}_{p^k}[x]$.
4. The Key Server multicasts (makes public) $r = K + S(nH, P)$, H, $S(H, P)$ and $p(x)$ on plain text.

Getting the key K: each client i calculates

- $S(H, P_i)$
- $\alpha_i = S(H, P)S(H, P_i)^{-1}$
- $p(x_i\alpha_i) = n$
- $K = r - S(H, P)^n$

3 Security

A passive adversary should know n in order to get K, what it is impossible without knowing any right value in $K \in \mathbb{F}_{p^k}$ to substitute in $p(x)$.

A second option for this adversary is trying to get any ticket but this is twofold. Firstly suppose that, somehow, she is able to get K at some moment. Then she could try to factorize $p(x) - K = \Pi_{i=1}^{m}(x - x_i\alpha_i) \in \mathbb{F}_{p^k}[x]$ and she is successful. Then what she would get are the products $x_i\alpha_i,\ i = 1, \ldots, m$ that do not give any information on x_i and α_i since there is not unique factorization in $\mathbb{F}_{p^k}$. Even in case P is public, this does not let know anything on P_i's and so, it cannot be used to factorize $x_i\alpha_i,\ i = 1, \ldots, m$.

We note at this point that this extends the method used in [9]. In that case the authors use a polynomial as the one introduced in the protocol to let the users calculate the value of the exponent n. In that case, the server multicasts a polynomial $f(x) = K + \Pi_{i=1}^{m}(x - h(x_i))$. In order to avoid a factorization attack by an authorized user who can get all values $h(x_i$, the authors change the hash function with every refreshment of the key K. In our case, the polynomial changes with every refreshment of K since values interpolating the polynomial depend on a random point H of the curve E.

 ISBN: 978-84-614-6167-7

An active adversary, as usual, could be more dangerous. In this case we can assume that this is an authorized user that pretends to compromise the other clients' tickets. Suppose that we are using an elliptic curve with a prime number of points and user i is able to factorize two polynomials $p(x)$ and $p'(x)$ constructed for two different rekeying and using H and H' respectively. Suppose also that she is able to determine that $x_j\alpha_i$ and $x'_j\alpha'_j$ belongs to some other user j. Then the quotient

$$\frac{x_j\alpha_j}{x_j\alpha'_j} = \frac{\alpha_j}{\alpha'_j} = \frac{S(H,P)S(H',P_j)}{S(H',P')S(H,P_j)}$$

Recall that the same users do not have to be "connected" in two different refreshments, and so the sum of all the points corresponding to all users, P and P' may be different.

Therefore what user i is able to get using the above quotient is $S(H'-H,P_j)$ since $S(H,P)$ and $S(H',P')$ are public. Now, taking into account that the curve E is generated by any point in it (excepting the infinite point) she can consider Q any point in E as a generator. Then $P_j = aQ$ and therefore $S(H'-H,P_j) = S(H'-H,Q)^a$. Thus, solving DLP she would be able to get a. We recall that she knows H', H and Q. In this way, $aQ = P_j$ and calculating $S(H,P_j)$ she would get α_j and therefore, from $x_j\alpha_j$, the ticket (x_j,P_j) would be compromised.

There exist two different forms of avoiding this attack. The first one is to consider a finite field big enouth in order that DLP is still a hard problem but this means that rekeying messages would increase their length. A good alternative then could be using a Hash function in the construction of the polynomial $p(x)$ that allow the clients to calculate the exponent n. In this way the polynomial would be of the form $p(x) = K + \Pi_{i=1}^{m}(x - h(x_i\alpha_i))$, similar to that used in [9] but with the advantage that unlike in that case, we do not have to change the hash function with every rekeying message since the values where $p(x)$ interpolates change in every rekeying since they depend on a different random point H of the curve E.

4 Scalability

Scalability of a secure communications protocol concerns key storage, both in the server and in every client and number of rekeying messages. The protocol that we introduce in this paper has exactly the same behavior of that one that the authors introduced in [10] concerning rekeying messages and number of keys stored by everyone. The advantage that we present is the length of a rekeying message. We recall that security of the algorithm in [10] lies on the factorization of an integer which is a product of large prime numbers and the length of the message is of the same order of this integer. Thus, a huge number of users produces a message that can be unaffordable for a communication system and therefore it is needed (among other advantages that are shown in [10]) the division of the users in groups. From the analysis of the protocol that we are introducing we can observe that the length of a rekeying message is most of it due to the polynomial, but in this case, the length of this polynomial over a finite

 ISBN: 978-84-614-6167-7

field of standard size in these situations would produce, even when the audience is large, a message whose length is perfectly affordable.

Acknowledgements

This work has been supported by the Spanish Ministry of Science and Innovation (TEC2009-13763-C02-02) and Junta de Andalucía (FQM 0211).

References

[1] D. Wallner, E. Harder and R. Agee, *Key management for multicast: Issues and architectures*, RFC 2627, 1999.

[2] Chung Kei Wong, Mohamed Gouda, and Simon S. Lam, *Secure group communications using key graphs*, IEEE/ACM Transactions on Networking **8**(1) (2000) 16–30.

[3] Roberto Di Pietro and Luigi V. Mancini, *Efficient and Secure Keys Management for Wireless Mobile Communications*, Proceedings of the second ACM international workshop on Principles of mobile computing, 2002, 66–73.

[4] Alan T. Sherman and David A. McGrew, *Key establishment in large dynamic groups using one-way function trees*, IEEE Transactions on Software Engineering **29** (2003) 444–458.

[5] Adrian Perrig, Dawn Song, and J. D. Tygar, *Elk, a new protocol for efficient large-group key distribution*, Proceedings of IEEE Symposium on Security and Privacy (S& P), 2001, 247–262.

[6] Guang-huei Chiou and Wen-Tsuen Chen, *Secure broadcasting using the secure lock*, IEEE Trans. Softw. Eng. **15**(8) (1989) 929–934.

[7] Peter S. Kruus and Joseph P. Macker, *Techniques and issues in multicast security*, Proceedings of Military Communications Conference, MILCOM 1998, 1028–1032.

[8] O. Scheikl, J. Lane, R. Boyer and M. Eltoweissy, *Multi-level secure multicast: the rethinking of secure locks*, Parallel Processing Workshops, 2002. Proceedings. International Conference on, 2002, 17–24.

[9] B. Liu, W. Zhang and T. Jiang, *A Scalable Key Distribution Scheme for Conditional Access System in Digital Pay-TV System*, IEEE Consumer Electronics **50**(2) (2004) 632–637.

[10] J.A.M. Naranjo, N. Antequera, L.G. Casado and J.A. Lopez-Ramos, *A suite of algorithms for key distribution and authentication in centralized secure*

multicast environments, Journal of Computational and Applied Mathematics, In press, accepted manuscript. 2011. DOI: 10.1016/j.cam.2011.02.015

[11] I.F. Blake, G. Seroussi and N.P. Smart, *Advances in Elliptic Curve Cryptography*, London Mathematical Society LNS Series 317, Cambridge University Press, Cambridge, 2005.

 ISBN: 978-84-614-6167-7

Proceedings of the 11th International Conference on Computational and Mathematical Methods in Science and Engineering, CMMSE 2011 26–30 June 2011.

Numerical solution of an optimal investment problem with transaction costs *

Iñigo Arregui[1] **and Carlos Vázquez**[1]

[1] *Dpto. de Matemáticas, Universidad de La Coruña, Spain*

emails: arregui@udc.es, carlosv@udc.es

Abstract

In a theoretical work, Dai and Yi [1] present an optimal investment problem and its equivalence with a double obstacle problem, and prove some properties of the solution and the free boundaries. We now present some numerical techniques to approximate the solution of the obstacle problem, and show the results obtained in a realistic simulation.

Key words: Investment, transaction costs, obstacle problem, free boundaries, numerical methods

1 Introduction

This paper concerns the numerical solution of a continuous-time optimal investment problem carried out by a constant relative risk aversion investor facing a finite horizon when transactions costs are considered. In the absence of transaction costs the strategy proposed by Merton in [4] is based on maintaining a constant fraction of the total wealth in each asset. This cannot be applied in a real market with transaction costs due to the required continuous trading. For infinite horizon problems [2] an appropriate free boundary model for the case of transaction has been first proposed and the regularity of solution is proved, later [5] used the viscosity solution approach. The finite horizon case makes the free boundary time dependent (moving boundary); this problem has been addressed by using the relationship between the singular control problem and the obstacle problem in [1]. Thus, the original problem is transformed into an equivalent double obstacle problem associated to a nonlinear parabolic differential equation. This approach allows to apply the well developed techniques of double obstacle problems to obtain the existence of solution and several properties of the free boundary.

*Partially supported by Ministerio de Ciencia e Innovación (project MTM2010-21135-C02-01) and Xunta de Galicia (project INCITE09-105-339-PR)

 ISBN: 978-84-614-6167-7

In the present paper the authors propose a set of numerical techniques to solve this problem: time discretization based on characteristics method to cope with the convection dominated aspect, spatial discretization with piecewise linear finite elements, a Newton-Rapshon technique to solve the nonlinear term of the parabolic equation and a projection method to treat the nonlinearity associated to the free boundaries. The numerical results illustrate the performance of the method by verifying all the properties stated in [1], including a steady-state solution in certain cases. Moreover, the buy, sell and no transaction regions can be recovered and the optimal value in terms of the original economic variables can be obtained.

2 The continuous model

Let us suppose an investor who holds X_t and Y_t in bank and stock, respectively, expressed in monetary terms. In the presence of transaction costs, their evolutions are described by

$$\begin{cases} dX_t = rX_t\,dt - (1+\lambda)\,dL_t + (1-\mu)\,dM_t\,, & X_0 = x, \\ dY_t = \alpha Y_t\,dt + \sigma Y_t\,d\mathcal{B}_t + dL_t - dM_t\,, & Y_0 = y, \end{cases} \tag{1}$$

where $r > 0$ is the constant risk free interest rate, $\alpha > r$ is the constant expected rate of return and $\sigma > 0$ is the constant volatility of the stock. Also, $\mathcal{B}_t$ denotes a standard Brownian process on a filtered probability space. Moreover, L_t and M_t are right–continuous, nonnegative and nondecreasing processes representing cumulative money values for the purpose of buying and selling stock, respectively, while the constants $\lambda \in [0,\infty)$ and $\mu \in [0,1)$ account for proportional transaction costs incurred on purchase and sale of stock, respectively.

Due to transaction costs, the investor's net wealth in monetary terms at time t is given by:

$$W_t = \begin{cases} X_t + (1-\mu)Y_t\,, & \text{if } Y_t \geq 0, \\ X_t + (1+\lambda)Y_t\,, & \text{if } Y_t < 0, \end{cases}$$

and the solvency region is defined [2] as $\mathcal{S} = \{(x,y) \in \mathbb{R}^2 \,/\, x+(1+\lambda)y > 0,\ x+(1-\mu)y > 0\}$. Given an initial position $(X_t, Y_t) = (x,y) \in \mathcal{S}$, an investment strategy (L,M) is admissible if (X_s, Y_s) given by (1) is in $\mathcal{S}$ for all $s \in [t,T]$. Let $A_t(x,y)$ be the set of admissible investment strategies.

The investor's problem is to choose an admissible strategy so as to maximize, at initial time t, the expected utility of terminal wealth, that is,

$$\sup_{(L,M)\in A_t(x,y)} E_t^{x,y}[U(W_T)]$$

subject to (1), where $E_t^{x,y}$ denotes the conditional expectation at time t given an initial

 ISBN: 978-84-614-6167-7

endowment $(X_t, Y_t) = (x, y)$, and the utility function is given by

$$U(W) = \begin{cases} \dfrac{W^\gamma}{\gamma}\,, & \text{if } \gamma < 1,\ \gamma \neq 0, \\ \log W\,, & \text{if } \gamma = 0\,. \end{cases}$$

Thus, Dai and Yi [1] define the value function by

$$\varphi(x,y,t) = \sup_{(L,M)\in A_t(x,y)} E_t^{x,y}[U(W_T)]\,, \quad (x,y)\in\mathcal{S}, \quad t\in[0,T)\,.$$

We assume that $\lambda+\mu > 0$. The value function is the viscosity solution of the following Hamilton–Jacobi–Bellman equation [5]:

$$\min\{-\varphi_t - \widehat{\mathcal{L}}\varphi, -(1-\mu)\varphi_x + \varphi_y, (1+\lambda)\varphi_x - \varphi_y\} = 0\,, \quad (x,y)\in\mathcal{S}, \quad t\in[0,T)\,,$$

with the terminal condition

$$\varphi(x,y,T) = \begin{cases} U(x+(1-\mu)y)\,, & \text{if } y > 0\,, \\ U(x+(1+\lambda)y)\,, & \text{if } y \leq 0\,, \end{cases}$$

where

$$\widehat{\mathcal{L}}\varphi = \frac{1}{2}\sigma^2 y^2 \varphi_{yy} + \alpha y \varphi_y + rx\varphi_x\,.$$

Considering $y > 0$ (as short selling is always suboptimal), $V(x,t) = \varphi(x,1,t)$ is defined so that

$$\varphi(x,y,t) = \begin{cases} y^\gamma V(\frac{x}{y},t)\,, & \text{if } \gamma < 1,\ \gamma \neq 0\,, \\ V(\frac{x}{y},t) + \log y\,, & \text{if } \gamma = 0\,. \end{cases} \tag{2}$$

Then, by introducing

$$w(x,t) = \gamma^{-1}\log(\gamma V) \qquad \text{and} \qquad v(x,t) = w_x(x,t)\,, \tag{3}$$

in [1] it is proved that v satisfies the following parabolic double obstacle nonlinear problem in $\Omega\times[0,T)$:

$$\begin{cases} -v_t - \mathcal{L}v = 0 & \text{if } \dfrac{1}{x+1+\lambda} < v < \dfrac{1}{x+1-\mu}\,, \\ -v_t - \mathcal{L}v \geq 0 & \text{if } v = \dfrac{1}{x+1+\lambda}\,, \\ -v_t - \mathcal{L}v \leq 0 & \text{if } v = \dfrac{1}{x+1-\mu}\,, \\ v(x,T) = \dfrac{1}{x+1-\mu}\,. \end{cases} \tag{4}$$

where $\Omega = (-(1-\mu), +\infty)$ and

$$\mathcal{L}v = \frac{1}{2}\sigma^2 x^2 v_{xx} - \left(\alpha - r - (2-\gamma)\sigma^2\right) x v_x - \left(\alpha - r - (1-\gamma)\sigma^2\right) v + \gamma\sigma^2\left(x^2 v v_x + x v^2\right)\,.$$

Moreover, the existence, uniqueness and regularity of solution to this problem and some theoretical properties of the solution and the two free boundaries are proved.

 ISBN: 978-84-614-6167-7

3 Numerical methods

In the present work, we propose a numerical method to approximate the solution of the nonlinear problem (4). First, we use a localization procedure to consider the bounded domain $\Omega_N = (x^*, N)$ where $x^* > -(1-\mu)$ and N is large enough so that the solution is not affected in the region of economic interest. In this bounded domain we consider the following mixed boundary conditions

$$v(x^*,t) = \frac{1}{x^*+1-\mu}, \qquad v_x(N,t) = \frac{-1}{(N+1+\lambda)^2}\,. \tag{5}$$

Let us define $k_0 = \dfrac{1}{2}\sigma^2$, $k_1 = -\big(\alpha - r - (2-\gamma)\sigma^2\big)$, $k_2 = -\big(\alpha - r - (1-\gamma)\sigma^2\big)$ and $k_3 = -\dfrac{1}{2}\gamma\sigma^2 = -\gamma k_0$. Next, we introduce a time to maturity variable $\tau = T - t$ and take into account the identity $(x^2 v_x)_x = x^2 v_{xx} + 2x v_x$, so that the first equation in (4) can be written as:

$$v_\tau + (2k_0 - k_1)x v_x - k_0(x^2 v_x)_x - k_2 v + k_3(x^2 v^2)_x = 0\,.$$

Next, we use a characteristics method for time discretization, which has been first used in financial applications in [6]. For this purpose we introduce the material derivative $Dv/D\tau = v_\tau + (2k_0 - k_1)x v_x$ so that the previous equation can be written as

$$\frac{Dv}{D\tau} - k_0(x^2 v_x)_x - k_2 v + k_3(x^2 v^2)_x = 0\,.$$

For $M > 1$, let $\Delta\tau = T/M$ and $\tau^m = m\Delta\tau$, $m = 0, \ldots, M$. We approximate the total derivative at time τ^m so that the time discretized governing equation results to be

$$\mathcal{L}^{m+1} v^{m+1} = \frac{v^{m+1} - (v^m \circ \chi^m)}{\Delta\tau} - k_0(x^2 v_x^{m+1})_x - k_2 v^{m+1} + k_3\left(x^2 (v^{m+1})^2\right)_x = 0 \tag{6}$$

where $v^{m+1} \approx v(t^{m+1}, \cdot)$ and $\chi^m(x) = \chi(\tau^m)$ is obtained from the solution of the final value problem:

$$\frac{d\chi}{d\tau}(s) = (2k_0 - k_1)\chi(s)\,, \qquad \chi(\tau^{m+1}) = x$$

so that the following discretized two obstacle nonlinear problem in $(0,T)\times\Omega_N$ is obtained:

$$\begin{cases} \mathcal{L}^{m+1} v^{m+1} = 0 & \text{if } \dfrac{1}{x+1+\lambda} < v^{m+1} < \dfrac{1}{x+1-\mu}\,, \\ \mathcal{L}^{m+1} v^{m+1} \geq 0 & \text{if } v^{m+1} = \dfrac{1}{x+1+\lambda}\,, \\ \mathcal{L}^{m+1} v^{m+1} \leq 0 & \text{if } v^{m+1} = \dfrac{1}{x+1-\mu}\,, \\ v^{m+1}(x^*) = \dfrac{1}{x^*+1-\mu}, \quad v_x^{m+1}(N) = \dfrac{-1}{(N+1+\lambda)^2} \end{cases} \tag{7}$$

 ISBN: 978-84-614-6167-7

starting from $v^0(x) = \dfrac{1}{x+1-\mu}$.

After posing a variational formulation of (7), a piecewise linear finite elements scheme for spatial discretization is proposed. The nonlinear term is treated with a Newton–Raphson algorithm, leading to a discrete double obstacle problem at each iteration. For this problem we propose a projected relaxation method [3].

4 A numerical example

In order to illustrate the behavior of the solution and the performance of the numerical methods, we present in this section an example and the corresponding numerical results obtained with the previously described techniques. Note that due to the presence of the double obstacle, two free boundaries appear; we will use the notation x_s and x_b for the selling and buying free boundaries, which are related to the upper and lower obstacles, respectively.

Moreover, let us assume the following data set:

$$\begin{array}{lll} T = 4 & & \\ \sigma = 0.25 & r = 0.03 & \lambda = 0.08 \\ \gamma = 0.50 & \alpha = 0.10 & \mu = 0.02 \end{array}$$

According to [1] and the previous data set, the selling and buying free boundaries should verify:

$$x_s(t) \leq -0.5425\,, \qquad x_b(t) \geq -0.5979\,, \qquad \forall t \in [0,T]\,. \tag{8}$$

In Figure 1 we show the numerical solution at time $t = 0$, jointly with the lower and upper obstacles. The figure on the right is obtained by zoom from the one on the left, to better illustrate the results. For this numerical solution, a time step $\Delta t = -0.01$ and a finite element mesh size $\Delta x = 0.0195$ have been considered.

Figure 2(a) shows the evolution in time of the two free boundaries. We can see that they verify conditions (8), and all the other theoretical properties stated in [1] can also be numerically proved. The region between both curves is the *no transaction* (NT) region; the *selling region* (SR) and *buying region* (BR) are also represented.

In Figure 2(b) another representation of the solvency region is shown. In this case, the coordinate axes represent the amounts that the investor holds in bank (X) and stock (Y), respectively. The ratio of these amounts is an indicator of the optimal strategy: stock selling, no transaction or stock buying.

Once the function v has been computed, we can recover $w(x,t)$ and $V(x,t)$ from (3). Finally, the value function φ is obtained from (2). Figure 3 shows the value function over the solvency region.

 ISBN: 978-84-614-6167-7

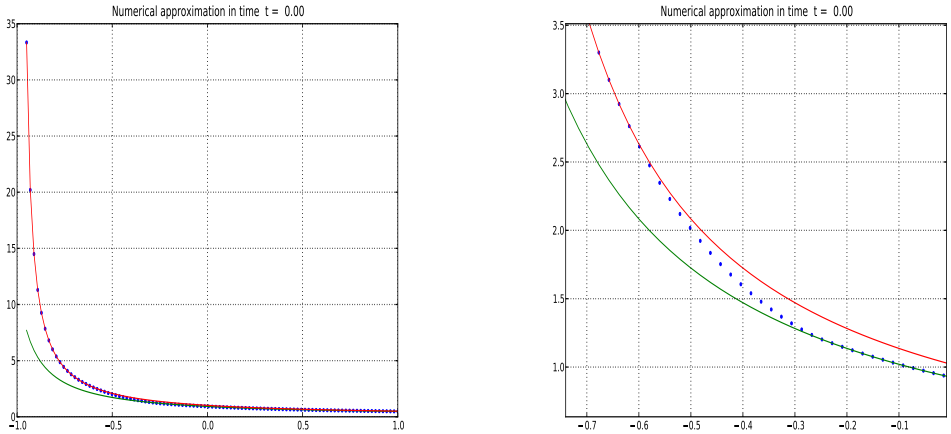


Figure 1: Numerical approximation of function v

References

[1] M. Dai and F. Yi, *Finite–horizon optimal investment with transaction costs: A parabolic double obstacle problem*, J. Differential Eq., **246** (2009), 1445–1469.

[2] M. H. A. Davis, A. R. Norman, *Portfolio selection with transaction costs*, Math. Methods Oper. Res., **15** (1990) 676–713.

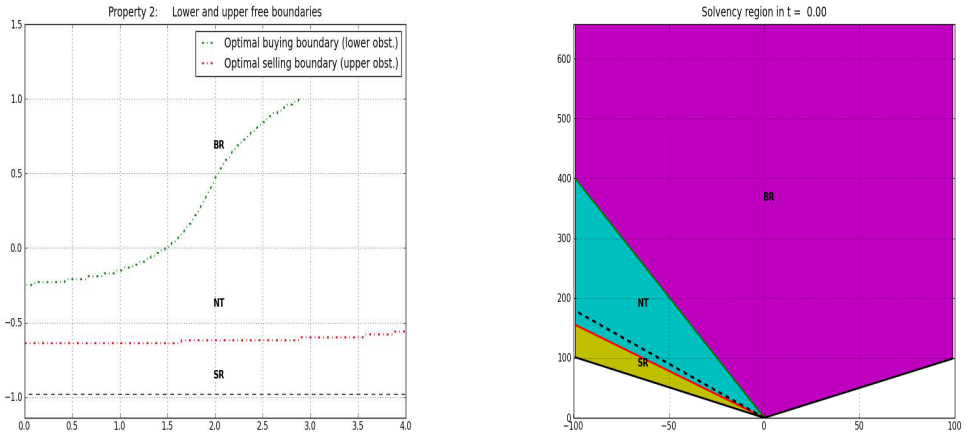
[3] R. Glowinski, J. L. Lions, R. Trémolières, *Analyse Numérique des Inéquations Variationnelles*, Dunod, Paris, 1976.

[4] R. Merton, *Optimal consumption and portfolio rules in a continuous time model*, J. Econom. Theory, **3** (1971) 373–413.

[5] S. E. Shreve, H. M. Soner, *Optimal investment and consumption with transaction costs*, Ann. Appl. Probab., **4** (1994) 609–692.

[6] C. Vázquez, *An upwind numerical approach for an American and European option pricing model*, Appl. Math. Comput., **97** (1998) 273–286.

 ISBN: 978-84-614-6167-7



(a) (b)

Figure 2: Free boundaries and solvency region, including the no transaction, selling and buying subregions

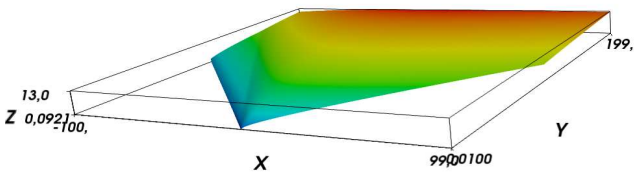


Figure 3: Value function, φ, over the solvency region

 ISBN: 978-84-614-6167-7

Proceedings of the 11th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2011
26–30 June 2011.

Solving competitive location problem with variable demand via parallel algorithms

A.G. Arrondo[1], J.L. Redondo[2], J. Fernández[1] and P.M. Ortigosa[2]

[1] *Department of Statistic and Operations research, University of Murcia, Spain*

[2] *Department of Computer Architecture and Electronics, University of Almería, Spain*

emails: agarrondo@um.es, jlredondo@ual.es, josefdez@um.es, ortigosa@ual.es

Abstract

A continuous location problem in which a firm wants to set up one new facility in a competitive environment is considered. Both the locations and the qualities of the new facility are to be found so as to maximize the profit obtained by the firm. In most location models, it is assumed that the demand is fixed, since such assumption reduces the computational effort to solve the problem. However, demand varies depending on prices, distances to the facilities, etc. and it influences the location decision very much. In this paper, a competitive location and design problem with variable demand is solved. The increase in the computational time is addressed using a shared memory parallel model. A computational study has been carried out to check the new programming method in terms of efficiency and effectiveness.

Key words: Evolutionary algorithm Parallelization Shared memory Continuous location Competition variable demand

1 Introduction

Location science deals with the location of one or more facilities in a way that optimizes a certain objective (minimization of transportation costs, minimization of social costs, maximization of the market share, etc.). For an introduction to the topic see [2, 5]. All location problems share several components, which leads to different models. The mathematical formulations and methods used to solve the problems vary substantially depending on the type of model.

Depending on whether a single player or multiple players in the market are considered, we can distinguish between *non-competitive* and *competitive* location models. A detailed

 ISBN: 978-84-614-6167-7

taxonomy can be found in the survey papers [4, 6, 14]. In many location models it is assumed that the decision maker, who plans the location of his facilities, faces an empty space without any similar or competing facilities. Nevertheless, in most of the cases, similar facilities already exist in the region and the task is to add new ones in an optimal way [4, 7, 8]. The existing facilities may belong to the decision maker's own chain or to a competitor's chain [3]. When a competition takes place, it may be *static*, which means that the competitors are already in the market, the owner of the new facility knows their characteristics and no reaction is expected from them, or *with foresight*, in which the competitors are assumed to react after the new facility enters. Furthermore, if the competitors can change their decisions, then we have a *dynamic model*, in which the existence of *equilibrium* situations is of major concern.

Regarding customers, one of the important features is the so called *demand*. Demand can be either *fixed* or *variable*. In the first case, the demand is known with certainty. This is usually the case when the goods are *essential* for the customers, and then, they will buy the goods independently of the distance to the facility or the price. In the second case, goods are *inessential* for customers, so, demand can vary depending on prices, distances to the facilities, etc. In most competitive location literature, it is assumed that the demand is fixed, regardless of the conditions of the market. This is mainly due to the difficulty of the problems to be solved: even with fixed demand, the corresponding location models may be hard-to-solve global optimization problems.

In this paper, we consider the planar competitive location and design problem with variable demand described in [9]. In that paper, the authors address the location problem using two global optimization techniques, i.e. the interval B&B method and the evolutionary algorithm UEGO. The first one is an exact method that can solve nearly any continuous optimization problem, although it can only solve small instances. The second one can generate solutions of large instances. However, to be effective, it requires to explore the search space deeply to obtain good solutions. This translates directly into larger computational times and larger resource requirements (memory, processors...). In such situation, a parallel machine together with a parallel model is required. Literature contains many examples of successful parallel implementations. Nevertheless, as nowadays new multicore systems are become common as personal computers, to develop models based on this paradigm could be appropriated.

In shared memory programming, the whole memory is directly accessible to all the processes with an intent to provide communication among themselves. Depending on context, programs may run on the same physical processor or on separate ones. Although there are several ways to deal with parallelism in a shared memory model, the application programming interface OpenMP (Open Multi-Processing) have been considered. It consists of a set of compiler directives, library routines, and environment variables that are used to express shared-memory parallelism.

 ISBN: 978-84-614-6167-7

In this work, to achieve a substantial reduction of the computing effort for UEGO, a shared memory programming model has been designed. To illustrate the performance of such model, a set of computational experiments has been carried out.

The rest of the paper is organized as follows: In Section 2 the location model is presented. In Section 3, the main ideas and particularities of the evolutionary algorithm UEGO are briefly described. The description of the parallel implementation is depicted in Section 4. It is Section 5 where the computational experiments to study the performance of the parallel algorithm are carried out. The paper ends with some conclusions in Section 6.

2 The continuous competitive location and design problem with variable demand

In [9] a planar competitive location and design problem was introduced, as well as a sensitivity analysis. We briefly describe the model.

A chain wants to locate a new single facility in a given area of the plane, where there already exist m facilities offering the same goods or product. The first k of those m facilities belong to the chain ($0 \leq k < m$). The demand is supposed to be *variable* and concentrated at n demand points, whose locations p_i given, as well as the location f_j and quality of the existing facilities. The following notation will be used throughout this paper:

Indices

- i index of demand points, $i = 1, \ldots, n$.
- j index of existing facilities, $j = 1, \ldots, m$.

Variables

- x location of the new facility, $x = (x_1, x_2)$.
- α quality of the new facility ($\alpha > 0$).
- nf variables of the problem, $nf = (x, \alpha)$.

Data

- p_i location of demand point i ($i = 1, \ldots, n$).
- $\widehat{w}_i$ demand (or buying power or total expenditure) at p_i.
- f_j location of existing facility j ($j = 1, \ldots, m$).
- d_{ij} distance between demand point p_i and facility f_j.
- a_{ij} quality of facility f_j as perceived by demand point p_i.
- $g_i(\cdot)$ a non-negative non-decreasing function.
- u_{ij} attraction that p_i feels for f_j (or utility of f_j perceived by the people at p_i), $u_{ij} = \dfrac{a_{ij}}{g_i(d_{ij})}$.
- γ_i weight for the quality of the new facility as perceived by demand point p_i.
- $d_i^{\min}$ minimum distance from p_i at which the new facility can be located.

 ISBN: 978-84-614-6167-7

$\alpha_{\min}$	minimum level of quality.
$\alpha_{\max}$	maximum level of quality.
S	region of the plane where the new facility can be located.

Miscellaneous

$d_i(x)$	distance between demand point p_i and the new facility.
u_{i0}	attraction that p_i feels for the new facility, $u_{i0} = \dfrac{\gamma_i \alpha}{g_i(d_i(x))}$.
$M(x, \alpha)$	market share captured by the chain.
$F(M(x, \alpha))$	expected sales obtained by the chain.
$G(x, \alpha)$	operating costs of the new facility.
$\Pi(x, \alpha)$	profit obtained by the chain.

We assume that $g_i(d_{ij}) > 0 \ \forall i, j$. In this paper demand is considered variable, that is, it varies depending on several factors. As it can be seen in [1], consumer expenditures on products or services offered by the facilities may increase for a variety of reasons related to the location of the new facility: opening new outlets may increase the overall utility of the product; the marketing expenditures resulting from the new facilities may increase the overall marketing presence of the product, leading to increased consumer demand; or some consumers who did not patronize any of the facilities, perhaps because none were close enough to their location, may now be induced to do so. On the other hand, the quality of the facilities may also affect consumer expenditures, since a better service usually leads to more sales. To be able to obtain good solutions to location problems, it is necessary to describe them as close to reality as possible, which may result in hard-to-solve optimization problems.

Therefore, the demand at p_i is affected by the perceived utility of the facilities, given by the vector $u_i = (u_{i0}, u_{i1}, \ldots, u_{im})$. Making the simplifying assumption that *the utility is additive*, then $U_i = u_{i0} + \sum_{j=1}^{m} u_{ij}$ represents the total utility perceived by a customer at p_i provided by all the facilities. Hence, it is natural to assume that the actual demand at p_i is a function of U_i, $w_i(U_i) = maxd_i \cdot e_i(U_i)$, where $maxd_i$ represents the maximun possible demand at p_i and $e_u(U_i) = c_i \cdot U_i$ is the utility given by a customer at p_i, with c_i a given constant such that $c_i \leq 1/U_i^{\max}$, where $U_i^{\max}$ is the maximum utility that can possibly be perceived by a customer at i, see [1].

Based on these assumptions the market share captured by the chain is

$$M(x, \alpha) = \sum_{i=1}^{n} w_i(U_i) \frac{u_{i0} + \sum_{j=1}^{k} u_{ij}}{u_{i0} + \sum_{j=1}^{m} u_{ij}}.$$

and the problem of profit maximisation is described by

$$\begin{cases} \max & \Pi(x, \alpha) = F(M(x, \alpha)) - G(x, \alpha) \\ \text{s.t.} & d_i(x) \geq d_i^{\min} \ \forall i \\ & \alpha \in [\alpha_{\min}, \alpha_{\max}] \\ & x \in S \subset \mathbb{R}^2 \end{cases} \tag{1}$$

 ISBN: 978-84-614-6167-7

where $F(\cdot)$ is a strictly increasing differentiable function which transforms the market share into expected sales, $G(x, \alpha)$ is a differentiable function which gives the operating cost of a facility located at x with quality α, and $\Pi(x, \alpha)$ is the profit obtained by the chain. The parameter $d_i^{\min} > 0$ is a given threshold, the parameters $\alpha_{\min}$ and $\alpha_{\max}$ are the minimum and maximum values, respectively, that the quality of a facility may take in practice. By S we refer to the region of the plane where the new facility can be located.

In this paper we assume function F to be linear, $F(M(x, \alpha)) = c \cdot M(x, \alpha)$, where c is the income per unit of goods sold. Function G should increase as x approaches one of the demand points, since it is rather likely that the operational cost of the facility will be higher around those locations (due to the value of land and premises, which will make the cost of buying or renting the location higher). On the other hand, G should be a convex function in the variable α, since the more quality we expect from the facility the higher the costs will be at an increasing rate. We assume G to be separable, in the form $G(x, \alpha) = G_1(x) + G_2(\alpha)$, where $G_1(x) = \sum_{i=1}^{n} \Phi_i(d_i(x))$, with $\Phi_i(d_i(x)) = \widehat{w}_i/((d_i(x))^{\phi_{i0}} + \phi_{i1})$, $\phi_{i0}, \phi_{i1} > 0$ and $G_2(\alpha) = e^{\frac{\alpha}{\beta_0}+\beta_1} - e^{\beta_1}$, with $\beta_0 > 0$ and β_1 given values.

3 The sequential optimization algorithm

UEGO has proved its ability at finding the global optimal solutions when solving different competitive location problems and also test functions described in literature (see references [9, 10, 11, 12, 13] and papers there in).

UEGO is an algorithm that works on a *set of species* (i.e. a population). Then, during the optimization process, a list of species is kept by UEGO. UEGO is, in fact, a method for managing this list (i.e. creating, deleting and optimizing species). In Algorithm 1, a global description of the evolutionary algorithm is given. At the beginning, a single species (the root species) exists, and as the algorithm evolves and applies genetic operators, new species can be created (*Create_specie* procedure). At every generation, UEGO performs a local optimizer operation on each species (*Optimize_species* procedure). It is important to highlight that UEGO, unlike other evolutionary algorithm, realizes two selection procedures during the optimization process. The first one is carried out after the new offspring is generated. This consist of the *Fuse_species* and *Shorten_species_list* procedures. The second one takes place after the optimization procedure, and only considers the *Fuse_species* procedure. The reader is referred to [12] for a more detailed description of the UEGO algorithm.

It is important to mention that the population can be separated into several isolated subpopulations, which can evolve to the local or global optima without participation of the remaining ones. Therefore, there exists an intrinsic parallelism that consists of dividing the population into the available processing elements. Notice that, in UEGO, a subpopulation is compounded by a single individual.

 ISBN: 978-84-614-6167-7

Algorithm 1: Algorithm UEGO

```
1 Init_species_list
2 Optimize_species
3 FOR i = 2 to L
4     Create_specie
5     Fuse_species
6     Shorten_species_list
7     Optimize_species
8     Fuse_species
```

4 Solving the location model in parallel

The parallel algorithm developed in this study considers a single population, which is stored in shared memory. The parallelism comes from the concurrent execution of both *creation* and *optimization* procedures. Notice that, in our case, the *selection* is a synchronization point, since it is necessary to have the whole population to proceed as the sequential version. Even so, partial selections are carried out concurrently, although finally a global one may be necessary, since it is required the parallel algorithm behaves as the sequential one.

In the parallel model, $MaxTh$ threads are created. This value refers to the maximum number of available process units to solve the problem. Basically, the algorithm distributes the species in the list among the $MaxTh$ threads, they apply the *Creation_species* or *Optimize_species* procedures, and writes back the evaluation results. Threads only receive the address memory of the corresponding species and they are in charge of either read or update through this value. Notice that the distribution is carried out so that a single species is assigned to each thread each time. When a particular thread finishes its task, another single species will be picked up for working on it.

In the following, the procedures, which have been either modified or parallelized, will be briefly described.

- *Init_species_list_paral*: For the parallel version, as many individuals as available processing elements, i.e. $MaxTh$, are created.

- *Create_specie_paral*: At each iteration of the algorithm, the current population is divided among $MaxTh$ threads. The distribution is carried out so that a single individual is assigned to each processing element in a sequential way. Then, the corresponding genetic operators are applied to each species to obtain a new offspring. When such procedures have finished, a new individual is picked up for working on it.

  As result, each processing elements have a sub-population, which contains a set of new candidate individuals. Then, partial selection is performed on such sub-list. The

 ISBN: 978-84-614-6167-7

selected individuals are added to the global population. Notice that this is a critical region and it is necessary to ensure that threads do not update the global population simultaneously. At the end of a parallel region, there is an unavoidable synchronization point.

- *Optimize_species_paral*: In this procedure, each individual of the population is optimized using a local optimizer. The distribution of individuals is similar to the one described for the previous procedure, that is, a dynamic schedule is considered: individuals are assigned to processing elements when they request them. Each one of them works on a single individual applying the local optimizer. This procedure is repeated while there exist individuals in the population that has not been assigned yet.

5 Computational studies

All the computational studies in this paper have been run in the Supercomputer BenArabi of Murcia, Spain. The shared-memory machine is a HP Integrity Superdome SX2000 with 128 cores and 1.5 TB of memory and the operating system is Linux. The algorithms have been implemented in C++ and shared-memory applications programming interface used is OpenMP.

In [9], a comprehensive computational study was carried out to compare the sequential UEGO with other algorithm from literature when solving the current location problem. In such a study, different types of problems, varying the number n of demand points, the number m of existing facilities, the number k of those facilities belonging to the chain were generated. For the current study, only some of the harder problems have been considered, and in particular, for this abstract, only the results for four instances with setting $n-m-k$ equal to $1000-5-0$, $1000-25-7$, $1000-50-0$ and $1000-50-30$ are presented. It is important to mention that, due to the stochastic nature of the algorithms, all the experiments were executed 5 times and average values were considered.

To ensure the solutions provided by our parallel algorithm reach the same function value than UEGO and to determine if our algorithm are efficient from a computational point of view, numerical values of *effectiveness* and *efficiency* were registered. As effectiveness measurement, we have computed the relative difference in objective value between the value obtained by the sequential UEGO $OptVal(UEGO)$ and the solution obtained by the parallel algorithms using P processing elements $OptVal(P)$,

$$DifObj(P) = \frac{OptVal(UEGO) - OptVal(P)}{OptVal(UEGO)}.$$

The closer to zero the value of $DifObj$, the better the effectiveness of the parallel model.

 ISBN: 978-84-614-6167-7

The efficiency of the parallel versions, which estimates how well-utilized the processors are in solving the problem, is computed as follows:

$$Eff(P) = \frac{T(1)}{P \cdot T(P)},$$

where $T(i)$ is the CPU time employed by the algorithm when i processing elements are used $(i = 1, 2, ..., P)$.

The effectiveness analysis showed that the proposed parallel algorithm is able to provide the same solution than the sequential UEGO for the complete set of problems, i.e. $DifObj \simeq 0$. Regarding the efficiency, Figure 1 summarizes the obtained results by the parallel version when it is executed with $P = 2, 4, 8$ processors.

As can be seen, the efficiency obtained by the parallel strategy is promising, since it is closer to the ideal case when the number of processing elements is less or equal than 4. Nevertheless, the efficiency tends to decrease when the number of processing elements increases. It could be due to those instances do not have enough computational load to be divided among such amount of processing elements. Then, for the present study, $P = 4$ seems to be appropriated.

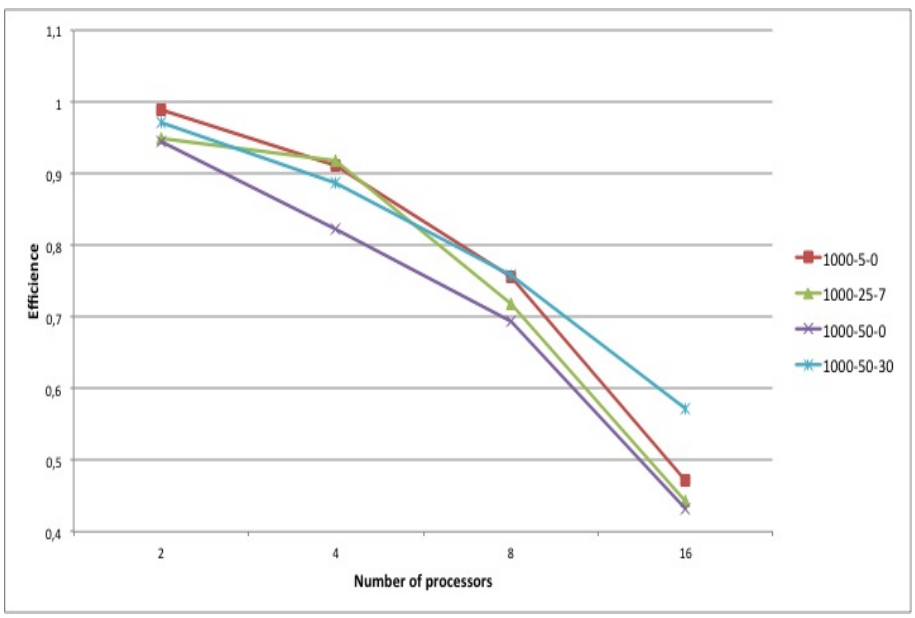


Figure 1: Efficiency obtained for different problems with $n = 1000$.

To determine if the parallel version is scalable or not, a set of three problems have been generated with larger numbers of the parameters $n - m - k$. In particular, two instances

 ISBN: 978-84-614-6167-7

with $20000 - 100 - 0$ and $20000 - 100 - 30$ have been executed. As can be seen in Figure 2, the obtained efficiency improves with regards to the previous case.

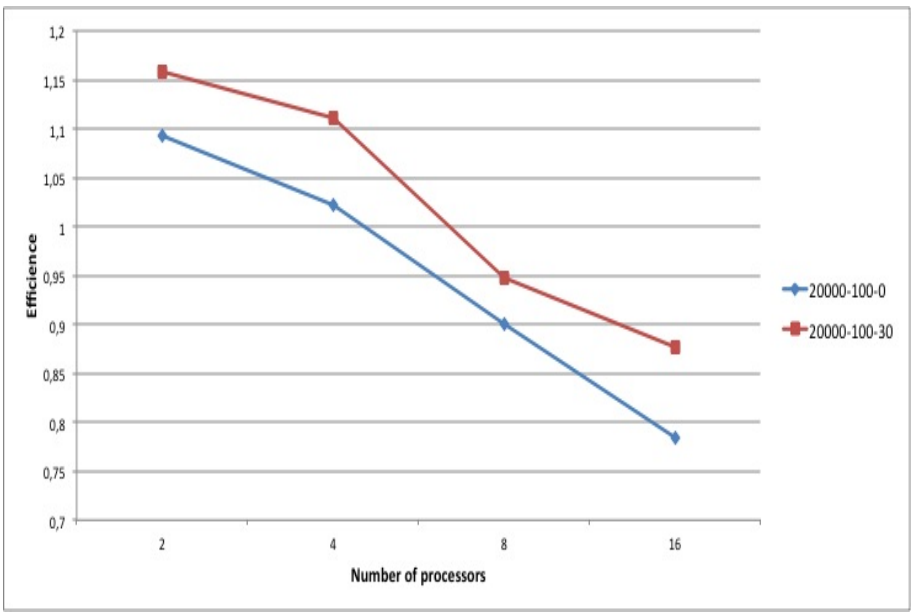


Figure 2: Efficiency obtained for different problems with $n = 20000$.

6 Conclusion

In this paper, a parallel algorithm for solving a single facility competitive location and design problem on the plane with variable demand, is considered. The objective is to maximize the profit obtained by the chain. The model is rather real, and as a consequence, it is also rather difficult to solve. In [9] the evolutionary algorithm UEGO was proposed to deal with this hard-to-solve optimization problem. The results showed that the evolutionary algorithm UEGO provides good solutions in terms of objective value, although it is time consuming.

In this paper, a parallel model devised to reduce the computing time of UEGO has been developed. This new method has a good behaviour in terms of effectiveness and acceptable in terms of efficiency.

In the near future, other parallel strategies based also on shared memory programming will be developed, evaluated and compared. But also some behaviours of the current parallel version will be studied deeper.

 ISBN: 978-84-614-6167-7

Acknowledgements

This work has been funded by grants from the Spanish Ministry of Science and Innovation (TIN2008-01117, ECO2008-00667/ECON), the Fundacion Séneca (00003/CS/10, 15254/PI/10) and Junta de Andalucía (P08-TIC-3518), in part financed by the European Regional Development Fund (ERDF). The authors are indebted to the Supercomputing Center of Fundación Parque Científico of Murcia, Spain, for the exceptional level of service provided.

References

[1] O. Berman and D. Krass. Locating multiple competitive facilities: Spatial interaction models with variable expenditures. *Annals of Operations Research*, 111(1):197–225, 2002.

[2] Z. Drezner and H.W. Hamacher. *Facility location. Applications and theory.* Springer, Berlin, 2002.

[3] H.A. Eiselt and G. Laporte. Objectives in location problems. In Z. Drezner, editor, *Facility Location: A Survey of Applications and Methods*, Springer Series in Operations Research and Financial Engineering, pages 151–180. Springer, Berlin, 1995.

[4] H.A. Eiselt, G. Laporte, and J.F. Thisse. Competitive location models: a framework and bibliography. *Transportation Science*, 27(1):44–54, 1993.

[5] R.L. Francis, L.F. McGinnis, and J.A. White. *Facility layout and location: an analytical approach.* Prentice Hall, Englewood Cliffs, 1992.

[6] H.W. Hamacher and S. Nickel. Classification of location models. *Location Science*, 6(1):229–242, 1998.

[7] M. Kilkenny and J.F. Thisse. Economics of location: a selective survey. *Computers and Operations Research*, 26(14):1369–1394, 1999.

[8] F. Plastria. Static competitive facility location: an overview of optimisation approaches. *European Journal of Operational Research*, 129(3):461–470, 2001.

[9] J.L. Redondo, J. Fernández, A. G. Arrondo, I. García, and P.M. Ortigosa. Fixed or variable demand? does it matter when locating a facility? *OMEGA-International journal of management science*, In press:DOI: 10.1016/j.omega.2011.02.007, 2011.

[10] J.L. Redondo, J. Fernández, I. García, and P.M. Ortigosa. Heuristics for the facility location and design (1|1)-centroid problem on the plane. *Computational Optimizations and Applications*, 2007. To appear, DOI: 10.1007/s10589-008-9170-0.

 ISBN: 978-84-614-6167-7

[11] J.L. Redondo, J. Fernández, I. García, and P.M. Ortigosa. A robust and efficient global optimization algorithm for planar competitive location problems. *Annals of Operations Research*, 167(1):87–106, 2009.

[12] J.L. Redondo, J. Fernández, I. García, and P.M. Ortigosa. Solving the multiple competitive location and design problem on the plane. *Evolutionary Computation*, 17(1):21–53, 2009.

[13] J.L. Redondo, P.M. Ortigosa, I. García, and J.J. Fernández. Image registration in electron microscopy. A stochastic optimization approach. *Lecture Notes in Computer Science, Proceedings of the International Conference on Image Analysis and Recognition, ICIAR 2004*, 3212(II):141–149, 2004.

[14] C.S. ReVelle and H.A. Eiselt. Location analysis: A synthesis and survey. *European Journal of Operational Research*, 165(1):1–19, 2005.

 ISBN: 978-84-614-6167-7

Proceedings of the 11th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2011
26–30 June 2011.

The Main Problem of the Satellite in Planar Motion: Topological Analysis of the Phase Flow

M. C. Balsas[1], E. S. Jiménez[1] and J. A. Vera[1]

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

emails: mariacarmen.balsas@murciaeduca.es, elena.jimenez@upct.es, juanantonio.vera3@upct.es

Abstract

In this paper we consider the Kepler problem with a perturbation. This is an approximation to the Main Problem of the artificial satellite. We are going to use the Liouville-Arnold theorem and a particular analysis of the momentum map in its critical points and we obtain a complete topological classification of the different invariant sets of the phase flow of this problem when we have two equilibrium points.

Key words: Hamiltonian system; Liouville-Arnold theorem; gyrostat; invariant manifolds; amended potential; Hill regions; artificial satellites theory.

1 Introduction

In this work we are going to consider a perturbation of the case of the Main Problem of the artificial satellite (*ASP, artificial satellite problem*, [7], [8]). We are in the case of an artificial satellite orbiting around a celestial body, in particular a planet, like for example the Earth. We are going to make an analytical, numerical and topology study of Hamiltonian dynamics for a simplified case where we only considered the first one and second dominant term of the gravitational potential.

 ISBN: 978-84-614-6167-7

The Hamiltonian to consider is:

$$H = \frac{1}{2}\left(p_r^2 + \frac{p_\theta^2}{r^2}\right) - \frac{1}{r} + \alpha p_\theta + \frac{\beta}{r^3}$$

and it corresponds to the dynamics of an artificial satellite orbiting around a celestial body, in particular a planet, to which we have introduced the effect associated to the rotation of the reference frame.

In accordance with [3]-[6] and [10], the dynamics of the movement comes given by the Hamiltonian one $\mathcal{H} : \mathbf{E} \longrightarrow \mathbb{R}$, being:

$$\mathcal{H} = \mathcal{H}_{Kepler} + \alpha p_\theta + \frac{\beta}{r^3}$$

where $\mathcal{H}_{Kepler}$ represents the Hamiltonian associated to the classical two-body problem and αp_θ is the effect associated to the rotating reference frame ($\alpha = \dfrac{2\pi}{T}$, with T = period of rotation of the satellite around the celestial body) and $\dfrac{\beta}{r^3}$ is the effect associated to the form of the celestial body. The parameters $\alpha > 0$ and $\beta \in \mathbb{R}$ are two structural constants of the system. Finally $\mathbf{E} = \mathbb{R}^+ \times S^1 \times \mathbb{R}^2$ is the phase space.

In order to do a qualitative study of the dynamics associated with the Hamiltonian system, similar to ([9]) we are going to consider the following sets:

$$E_h = \{(r, \theta, p_r, p_\theta) \in \mathbf{E} : \mathcal{H}(r, \theta, p_r, p_\theta) = h\}, h \in \mathbb{R},$$

$$J_k = \{(r, \theta, p_r, p_\theta) \in \mathbf{E} : p_\theta = k\}, k \in \mathbb{R},$$

$$I_{hk} = E_h \cap I_k,$$

with $z = (r, \theta, p_r, p_\theta) \in \mathbf{E}$ y $(h, k) \in \mathbb{R}^2$.

These sets are invariant by the flow associated with Hamiltonian, namely $\mathcal{H}$ and p_θ first integrals of motion, independent and in involution.

The main results of this paper are the description of the foliation of the phase space $\mathbf{E}$ by the invariant sets E_h, the energy sets E_h by the invariant sets I_{hk} and I_{hk} by the flow of the Hamiltonian system. This foliation provides a good description of the phase space when $(h; k) \in \mathbb{R}^2$ and depends on the different values of α and β. The main tool for this study is the Liouville-Arnold theorem ([2],[9]), applied to the momentum map $(\mathcal{H}; p_\theta) : \mathbb{E} \times \mathbb{R} \longrightarrow \mathbb{R}^2$

 ISBN: 978-84-614-6167-7

at regular values. A particular study of the sets E_h, J_k and I_{hk} for critical values of the momentum map is made. These values come given by the equilibrium points of $\mathcal{H}$ or by values where $p_\theta = k$ is a maximum or a minimum of the energy surface.

2 Amended potential. Hill regions

The amended potential, in polar symplectic coordinates is:

$$\tilde{\mathcal{V}}(r,\theta) = -\frac{\alpha^2 r^2}{2} - \frac{1}{r} + \frac{\beta}{r^3} \tag{1}$$

The Hill regions are determinate by the critical points of the amended potential $\tilde{\mathcal{V}}$. The value $\tilde{z}_i = (r_i, \theta_i)$ is a critical point of $\tilde{\mathcal{V}}$ if it is a extreme of the amended potential. These points are the real roots of the polynomial equation

$$\alpha^2 r^5 - r^2 + 3\beta = 0, \tag{2}$$

Using the Sturm algorithm we discuss, according to the values of the parameters α and β, the number of positive real roots respect to r of the equation (2).

When $3125\alpha^4\beta^4 - 4\beta > 0$ the potential function (1) does not have critical points; if $3125\alpha^4\beta^4 - 4\beta = 0$, $\tilde{\mathcal{V}}$ has one critical point (in this case we obtain a double root for the equation (2)); if $3125\alpha^4\beta^4 - 4\beta < 0$ y $\beta > 0$ thas two citical points; if $\beta < 0$, $\tilde{\mathcal{V}}$ has only one critical point (in this case we obtain a simple root of the equation (2)). And finally, if $\beta = 0$ has only one critical point.

Remark 1 *Note that if $\beta = 0$ the amended potential is*

$$\tilde{\mathcal{V}}(r,\theta) = -\frac{\alpha^2 r^2}{2} - \frac{1}{r}$$

and the term $\dfrac{\beta}{r^3}$, associated the effect of the form of the celestial body, disappears. In this case, the Hamiltonian of the Kepler problem in rotating reference frame is obtained, (see [3]).

Let $\pi : \mathbf{E} \longrightarrow \mathbb{R}^+ \times S^1$ be the natural projection of the phase space $\mathbf{E}$ in the configuration space $\mathbb{R}^+ \times S^1$. For each $h \in \mathbb{R}$ the Hill region R_h is defined by $R_h = \pi(I_h)$.

$$R_h = \{(r,\theta) \in \mathbb{R}^+ \times S^1 : \tilde{\mathcal{V}} \leq h\} \tag{3}$$

 ISBN: 978-84-614-6167-7

It easy to see that

$$R_h \approx \{r \in \mathbb{R}^+ : -\frac{\alpha^2 r^2}{2} - \frac{1}{r} + \frac{\beta}{r^3} \leq h\} \times S^1 \tag{4}$$

where $\approx$ means diffeomorphic to.

From this it follows that the values of the amended potential, in each one of the critical points r_i (2), will be denoted by $h_i = \widetilde{\mathcal{V}}(r_i)\,(i = 1, 2, 3, 4)$. The values $A_i\,(i = 1, 2, 3, 4)$, are the intersection points between the graph of the amended potential and $\widetilde{\mathcal{V}} = h$.

From now on, we are going to focus on the case of $3125\alpha^4\beta^4 - 4\beta < 0$ and $\beta > 0$, when the Hamiltonian has two equilibrium points. To a clearer understanding of the topology of the Hill regions the following figure is presented.

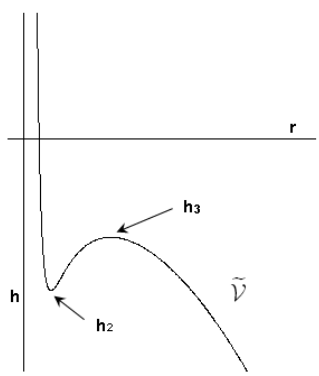


Figure 1: The amended potential for $3125\alpha^4\beta^4 - 4\beta < 0$ and $\beta > 0$ (There are two critical points).

Using the previous results, the topological classification of the Hill regions is:

 ISBN: 978-84-614-6167-7

$\beta > 0$ and $3125\alpha^4\beta^4 - 4\beta < 0$	$h < h_2$	$R_h \approx [A_1, +\infty) \times S^1$
	$h = h_2$	$R_h \approx \{\{h_2\} \cup [A_4, +\infty)\} \times S^1$
	$h_2 < h < h_3$	$R_h \approx [[A_1, A_2] \cup [A_3, +\infty)] \times S^1$
	$h = h_3$	$R_h \approx [[A_1, h_3] \cup [h_3, +\infty)] \times S^1$
	$h > h_3$	$R_h \approx [A_1, +\infty) \times S^1$

Table 1: Topological classification of the Hill regions for $3125\alpha^4\beta^4 - 4\beta < 0$ and $\beta > 0$ (two critical points). Correspond to the figure (1).

3 Qualitative study of the Hamiltonian flow

In this section we study the topology of the invariant manifolds $\mathcal{H}^{-1}(h) = E_h$ and I_{hk}. To give the topological classification of these invariant sets we need some notation and some new results.

For the study of the topology of $\mathcal{H}^{-1}(h) = E_h$ and I_{hk}, $(h, k) \in \mathbb{R}^2$ we must characterize the equilibrium points of the Hamiltonian $\mathcal{H}$, because the energy surfaces are regular when $h \neq h_i$, being $h_i = \mathcal{H}(z_{e_i})$ where z_{e_i} corresponds to each one of the equilibrium points of the Hamiltonian $\mathcal{H}$.

Note that $z_e = (r_e, \theta_e, p_{r_e}, p_{\theta_e}) \in \mathbf{E}$ is an equilibrium point of the Hamiltonian flow if and only if $z_e = (r_e, \theta_e)$ is a critical point of the amended potential. Moreover, $\pi(z_e) = \widetilde{z}_e$, where $\pi : \mathbf{E} \longrightarrow \mathbb{R}^+ \times S^1$ is the natural projection.

For this reason, when $3125\alpha^4\beta^4 - 4\beta > 0$, the Hamiltonian does not have equilibria; one family of equilibrium points (there is a double root) when $3125\alpha^4\beta^4 - 4\beta = 0$; two families of equilibrium points when $3125\alpha^4\beta^4 - 4\beta < 0$ and $\beta > 0$; one family of equilibrium points (simple root) when $\beta < 0$ and one family of equilibrium points when $\beta = 0$.

 ISBN: 978-84-614-6167-7

In which it follows, the values of the energy in each one of the families of equilibrium will be denoted by $h_i = \mathcal{H}\left(r_i, \theta, 0, p_{\theta_i}\right), (i = 1, 2, 3, 4)$.

The sets E_h and I_{hk} come given by:

$$E_h = \mathcal{H}^{-1}(h) = \{(r, \theta, p_r, p_\theta) \in \mathbf{E} : g\left(r, p_r, p_\theta\right) = h\} \approx g^{-1}(h) \times S^1,$$

$$J_k = \{z \in \mathbf{E} : p_\theta = k\}, k \in \mathbb{R},$$

$$I_{hk} = E_h \cap J_k \approx \left(g^{-1}(h) \cap \{p_\theta = k\}\right) \times S^1,$$

where $g : \mathbb{R}^+ \times \mathbb{R}^2 \longrightarrow \mathbb{R}$ is defined by $g\left(r, p_r, p_\theta\right) = \mathcal{H}\left(r, \theta, p_r, p_\theta\right)$. If $h \in \mathbb{R}$ is a regular value of the map g and $g^{-1}(h) \neq \emptyset$, then $g^{-1}(h)$ is a surface of $\mathbb{R}^+ \times \mathbb{R}^2$ called energy surface.

Remark 2 $g^{-1}(h) = E_h/S^1$.

The values $c_j, (j = 1, 2, 3)$ are the real roots respect to k of the equation:

$$18\beta k^2 + k^6 - 108\beta^2(h - \alpha k) - 12\beta\sqrt{12\beta + k^4} - k^4\sqrt{12\beta + k^4} = 0 \tag{5}$$

and correspond to the extremes of the energy surface $Q_j, (j = 1, 2, 3)$.

To classify the trajectories we need the equilibrium points $h_i\,(i = 1, 2, 3, 4)$, the values $c_j\,(i = 1, 2, 3)$ and a new value a_1. This last value a_1 is the real root respect to k of the equation:

$$-2\beta - k^2 r + 2r^2 - 2\left(h - \alpha k\right) r^3.$$

Remark 3 *From a_1 the orbits are not bounded.*

Finally, let S^{n-1} be the sphere in $\mathbb{R}^n$, with $n > 1$.

We obtained the topological classification for E_h and I_{hk} in the case of two equilibrium points. In this case $3125\alpha^4\beta^4 - 4\beta < 0$ and $\beta > 0$. The different subcases can be shown by means of the next figures:

4 Conclusions

In this paper we have considered the Kepler problem with a perturbation. This is an approximation to the Main Problem of the artificial satellite. Using the Liouville-Arnold

 ISBN: 978-84-614-6167-7

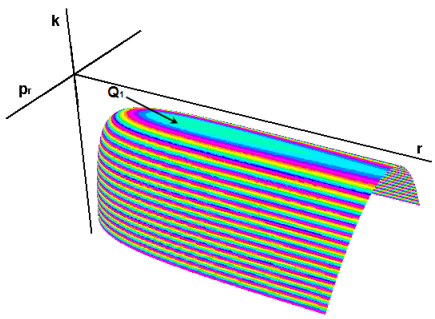


Figure 2: $g^{-1}(h) = E_h/S^1$, $h < h_2$ when $3125\alpha^4\beta^4 - 4\beta < 0$ and $\beta > 0$. Where Q_1 is a extreme of the energy surface and $k = p_\theta$.

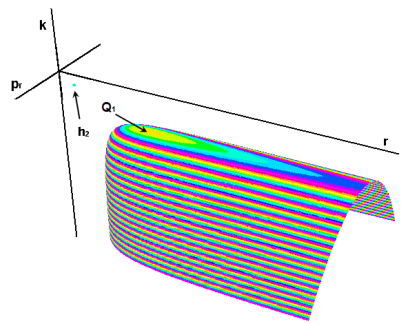


Figure 3: $g^{-1}(h) = E_h/S^1$, $h = h_2$ when $3125\alpha^4\beta^4 - 4\beta < 0$ and $\beta > 0$. Where Q_1 is a extreme of the energy surface, h_2 is an equilibrium point and $k = p_\theta$.

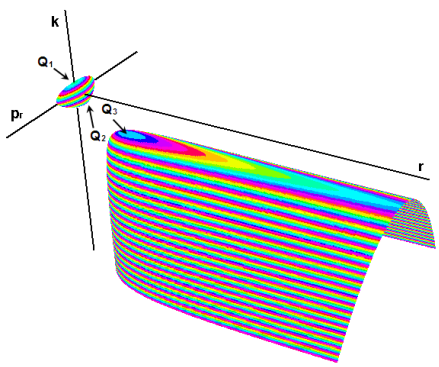


Figure 4: $g^{-1}(h) = E_h/S^1$, $h_2 < h < h_3$ when $3125\alpha^4\beta^4 - 4\beta < 0$ and $\beta > 0$. Where Q_1, Q_2 y Q_3 are three extremes of the energy surface and $k = p_\theta$.

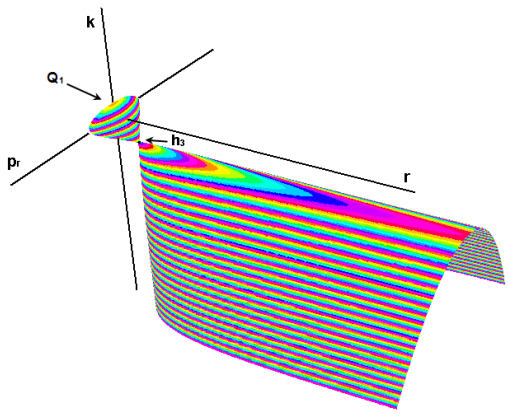


Figure 5: $g^{-1}(h) = E_h/S^1$, $h = h_3$ when $3125\alpha^4\beta^4 - 4\beta < 0$ and $\beta > 0$. Where Q_1 is an extreme of the energy surface, h_3 is an equilibrium point and $k = p_\theta$.

 ISBN: 978-84-614-6167-7

h	E_h	I_{hk}	
$h < h_2$	$S^3\backslash S^1$, see figure 2	$\emptyset$	$k > c_1$
		S^1	$k = c_1$
		$S^1 \times S^1$	$a_1 < k < c_1$
		$S^1 \times \mathbb{R}$	$k \leq a_1$
$h = h_2$	$\{S^1\} \cup \{S^3\backslash S^1\}$, see figure 3	$\emptyset$	$k > h_2$
		S^1	$k = h_2$
		$\emptyset$	$c_1 < k < h_2$
		$S^1 \times S^1$	$a_1 < k < c_1$
		$S^1 \times \mathbb{R}$	$k \leq a_1$
$h_2 < h < h_3$	$\{S^3\} \cup \{S^3\backslash S^1\}$, see figure 4	$\emptyset$	$k > c_1$
		S^1	$k = c_1$
		$S^1 \times S^1$	$c_2 < k < c_1$
		S^1	$k = c_2$
		$\emptyset$	$c_3 < k < c_2$
		S^1	$k = c_3$
		$S^1 \times S^1$	$a_1 < k < c_3$
		$S^1 \times \mathbb{R}$	$k \leq a_1$
$h = h_3$	$\{S^3\} \cup \{S^3\backslash S^1\}$, see figure 5	$\emptyset$	$k > c_1$
		S^1	$k = c_1$
		$S^1 \times S^1$	$h_3 < k < c_1$
		S^1	$k = h_3$
		$S^1 \times S^1$	$a_1 \leq k < h_3$
		$S^1 \times \mathbb{R}$	$k \leq a_1$
$h > h_3$	$S^3\backslash S^1$, see figure 6	$\emptyset$	$k > c_1$
		S^1	$k = c_1$
		$S^1 \times S^1$	$a_1 < k < c_1$
		$S^1 \times \mathbb{R}$	$k \leq a_1$

Table 2: Topological classification of E_h and I_{hk} when $3125\alpha^4\beta^4 - 4\beta < 0$ and $\beta > 0$, two equilibrium points.

 ISBN: 978-84-614-6167-7

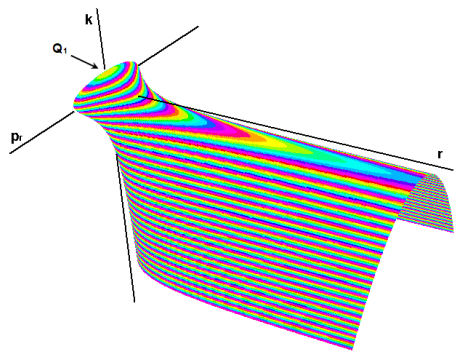


Figure 6: $g^{-1}(h) = E_h/S^1$, $h > h_3$ when $3125\alpha^4\beta^4 - 4\beta < 0$ and $\beta > 0$. Where Q_1 is a extreme of the energy surface and $k = p_\theta$.

theorem and a particular analysis of the momentum map in its critical points, we obtain a complete topological classification of the different invariant sets of the phase flow of this problem when we have two equilibrium points. The complete topological classification come be obtained using the same tools.

Acknowledgements

This work has been partially supported by the Departamento de Matemática Aplicada y Estadística of the UPCT and the Consejería de Educación y Cultura de la Comunidad Autónoma de la Región de Murcia (Spain) (Project 12001/PI/09).

References

[1] R. Abraham and J. E. Marsden, *Foundations and Mechanics*, Addison-Wesley (1978).

 ISBN: 978-84-614-6167-7

[2] V. I. Arnold et al, *Dynamical Systems III (Encyclopaedia of Mathematical Sciences*, (Encyclopaedia of Mathematical Sciences), Springer Verlag, Berlin.

[3] M. B. Balsas, E. S. Jiménez and J. A. Vera, *The Kepler Problem in rotating reference frames: Topological study of the Phase flow*, AIP Conference Proceedings **963** (2007) 1146–1449.

[4] M. B. Balsas, E. S. Jiménez, J. A. Vera and A. Vigueras, *The motion of a gyrostat in a central gravitational field: Phase portraits of an integrable case*, Journal Nonlinear on Mathematics and Physics, **15** (2008) 53–64.

[5] M. B. Balsas, E. S. Jiménez, J. A. Vera, *Qualitative analysis of the phase flow of an integrable approximation of a generalized rototranslatory problem*, Central European Journal of Physics, **7** (2009) 67–78.

[6] M. B. Balsas, J. L. Guirao, E. S. Jiménez, J. A. Vera, *Qualitative analysis of the phase flow of a Manev system in a rotating reference frame*, International Journal of Computer Mathematics, **7** (2009) 67–78.

[7] A. V. Bobylev, *On Vlasov–Manev equations. I: Foundations, properties, and non-global existence*, Journal of Statistical Physics **88** (1997) 885–911.

[8] S. L. Coffey et al, *Frozen orbits for satellites close to an Earth-Like Planet*, Celestial Mechanics and Dynamical Astronomy **59** (1993) 37-72.

[9] J. LLibre et al, *Phase portraits of the two-body problem with Manev potential*, J. of Physics A: Mathematical and General **34** (2001) 1919–1934.

[10] J. A. Vera and A. Vigueras, *Hamiltonian dynamics of a gyrostat in the n-body problem: relative equilibria*, Celestial Mechanics and Dynamical Astronomy, **94** (3) (2006) 289–315.

 ISBN: 978-84-614-6167-7

Proceedings of the 11th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2011
26–30 June 2011.

The Profit Maximization Problem in Economies of Scale

L. Bayón[1], J.A. Otero[1], M.M. Ruiz[1], P.M. Suárez[1] and C. Tasis[1]

[1] *Department of Mathematics, University of Oviedo. Spain*

emails: bayon@uniovi.es, jaurelio@uniovi.es, mruiz@uniovi.es, pedrosr@uniovi.es, ctasis@uniovi.es

Abstract

In this paper we present a generalization of the classic Firm's Profit Maximization Problem, using the linear model for the production function, considering a decreasing price $w_i(x_i) = b_i - c_i x_i$ and maximum constraints for the inputs or, equivalently, considering inputs that are in turn outputs in a economies of scale with quadratic concave cost functions. We formulate the problem by previously calculating the analytical minimum cost function in the quadratic concave case. This minimum cost function will be calculated for each production level via the infimal convolution of quadratic concave functions whose result is a piecewise quadratic concave function.

Key words: Concave Programming, Economies of Scale, Infimal Convolution, Piecewise Concave Functions, Algorithm Complexity
MSC 2000: 90C20, 90C26, 90C60

1 Introduction

Problems involving economies of scale (in production and sales) can often be formulated as concave quadratic programming problems [1], [2]. Consider a case in which n products are being produced, with x_i being the number of units of product i and w_i being the unit production cost of product i. As the number of units produced increases, the unit cost usually decreases. This can often be correlated by a linear functional

$$w_i(x_i) = b_i - c_i x_i \tag{1}$$

where $c_i > 0$. Thus, given constraints on production demands and the availability of each product and using the classic linear production function model [3], [4], the Firm's Cost Minimization (FCM) problem [5], [6] can be written as:

$$\begin{aligned} & C(y) = \min_{\mathbf{x}} \sum_{i=1}^{n} x_i w_i(x_i) \\ \text{s.t.} \quad & \sum_{i=1}^{n} a_i x_i = y;\ a_i \neq 0,\ i = 1, ..., n \\ & 0 \leq x_i \leq U_i;\ i = 1, ..., n \end{aligned} \tag{2}$$

 ISBN: 978-84-614-6167-7

where y is the output. This is a concave minimization problem. As well as representing a situation in which the inputs are acquired with a discount proportional to the amount, the affine function model for the prices (1) can also be interpreted as dealing with inputs that are in turn outputs of a prior production process of economies of scale with a quadratic cost: $x_i b_i - c_i x_i^2$. On the other hand, the linear production function is presented in a natural way when the output is the result of the sum of the inputs ($a_i = 1$) or, in general, a specific fraction of each of these.

Similarly, when the Firm's Profit Maximization (FPM) Problem is considered:

$$\begin{aligned} & \pi(p, \mathbf{w}) = \max_{\mathbf{x},y}(py - \sum_{i=1}^{n} x_i w_i(x_i)) \\ \text{s.t.} \quad & \sum_{i=1}^{n} a_i x_i = y;\ a_i \neq 0,\ i = 1, ..., n \\ & 0 \leq x_i \leq U_i;\ i = 1, ..., n \end{aligned} \tag{3}$$

the economies of scale dictate that the profit per unit rises linearly with the number of units produced. In this case, therefore, the problem becomes one of maximization of a convex functional.

To solve the FPM problem, we formulate the problem by previously calculating the analytical minimum cost function $C(y)$ and then maximizing over the output quantity:

$$\pi(p, \mathbf{w}) = \max_{y}(py - c(\mathbf{w},y)) = \max_{y}(py - C(y))$$

Concave programming [7], [8] constitutes one of the most fundamental and most widely studied problem classes in deterministic nonconvex optimization. Concave programming has a remarkably broad range of direct and indirect applications. Many of the mathematical properties of concave programming are even identical to the properties of linear programming. The goal in concave programming, or the concave minimization problem (CPM):

$$\begin{aligned} glob\min \quad & f(x) \\ \text{s.t.} \quad & x \in D \end{aligned}$$

is to find the global minimum value that f achieves over D, where D is a nonempty, closed convex set in $\mathbb{R}^n$ and f is a real-valued, concave function defined on some open convex set A in $\mathbb{R}^n$ that contains D. The application of standard algorithms designed for solving constrained convex programming problems will generally fail to solve CMP. Accordingly, in this paper we shall present an algorithm specifically designed for the problem we are going to solve that, as we shall see, presents very advantageous features.

To develop the algorithm which determines the optimal production level, we shall make use of the infimal convolution operator. This operator is well known within the context of convex analysis [9], [10] and [11]. However, convexity is only one desirable property so as to be able to resort to differential techniques to tackle its calculation and its use should definitely not be restricted to this context alone.

Definition 1. *Let $F, G : \mathbb{R} \longrightarrow \bar{\mathbb{R}}$ be two functions of $\mathbb{R}$ in $\bar{\mathbb{R}} := \mathbb{R} \cup \{+\infty, -\infty\}$. We denote as the Infimal Convolution of F and G the operation defined below:*

$$(F \bigodot G)(x) := \inf_{y \in \mathbb{R}} \{F(x) + G(y - x)\}$$

 ISBN: 978-84-614-6167-7

It is well known that $(\mathcal{F}(\mathbb{R},\bar{\mathbb{R}}),\bigodot)$ is a commutative semigroup. Furthermore, for every finite set $E\subset\mathbb{N}$, it is verified that

$$(\bigodot_{i\in E}F_i)(K)=\inf_{\sum_{i\in E}x_i=K}\sum_{i\in E}F_i(x_i)$$

When the functions are considered constrained to a certain domain, $Dom(F_i)=[m_i,M_i]$, the above definition continues to be perfectly valid redefining $F_i(x)=+\infty$ if $x\notin Dom(F_i)$. In this case, the definition may be expressed as follows:

$$(F_1\bigodot F_2)(\xi):=\min_{\substack{x_1+x_2=\xi\\ m_i\le x_i\le M_i}}(F_1(x_1)+F_2(x_2))=\min_{\substack{m_1\le x\le M_1\\ m_2\le\xi-x\le M_2}}((F_1(x)+F_2(\xi-x))$$

2 Statement of the Generalized Problem

We first consider the FCM problem (2). Using (1) and making these changes in the variables

$$\begin{array}{ll} a_ix_i & =z_i;\quad a_iU_i=M_i\\ \dfrac{b_i}{a_i} & =\beta_i;\quad \dfrac{c_i}{a_i^2}=\gamma_i\end{array}$$

the FCM problem may be re-written as follows:

$$\begin{array}{ll} & C(y)=\min_{\mathbf{z}}\sum_{i=1}^{n}\beta_iz_i-\gamma_iz_i^2\\ \text{s.t.} & \sum_{i=1}^{n}z_i=y\\ & 0\le z_i\le M_i;\ i=1,...,n\end{array}\tag{4}$$

which makes $C(y)$ the infimal convolution of the quadratic functions:

$$F_i(z_i):=\beta_iz_i-\gamma_iz_i^2$$

respectively constrained to the domains $[0,M_i]$; i.e.

$$C=F_1\bigodot F_2\bigodot\ldots\bigodot F_n$$

In this paper we shall demonstrate that $C(y)$ is piecewise concave such that the solution to the FPM problem:

$$\begin{array}{ll} & \max_y(py-C(y))\\ \text{s.t.} & \sum_{i=1}^{n}z_i=y\\ & 0\le z_i\le M_i;\ i=1,...,n\end{array}\tag{5}$$

cannot be tackled by means of marginalistic techniques (coinciding of marginal cost and price). In fact, the maximum profit will be obtained at a production level y^* where C is not differentiable, or at boundary values

$$y^*=0\ \text{ or }\ y^*=\sum_{i=1}^{n}M_i$$

 ISBN: 978-84-614-6167-7

3 The infimal convolution in the concave case

In this section we shall study the infimal convolution of two concave functions, which is crucial as the basis for the optimization algorithm.

Lemma 1. *Let F_1 and F_2 be concave functions with domains $[m_1, M_1]$ and $[m_2, M_2]$, respectively. We shall consider the following four functions:*

$$\begin{array}{lll} \Psi_1^-(x) := F_1(x - m_2) + F_2(m_2) & \textit{with domain} & [m_1 + m_2, M_1 + m_2] \\ \Psi_1^+(x) := F_1(x - M_2) + F_2(M_2) & \textit{with domain} & [m_1 + M_2, M_1 + M_2] \\ \Psi_2^-(x) := F_2(x - m_1) + F_1(m_1) & \textit{with domain} & [m_1 + m_2, m_1 + M_2] \\ \Psi_2^+(x) := F_2(x - M_1) + F_1(M_1) & \textit{with domain} & [M_1 + m_2, M_1 + M_2] \end{array}$$

then

$$(F_1 \bigodot F_2)(x) = \min\{\Psi_1^-(x), \Psi_1^+(x), \Psi_2^-(x), \Psi_2^+(x)\}$$

Proof. Due to the concavity of the functions involved, the minimum value of $F_1(x_1) + F(x_2)$ constrained to $x_1 + x_2 = \xi$ can only be achieved in those pairs (x_1, x_2) in which only one of the components can be inside the corresponding domain of F_i. In other words, the aforementioned minimum value can only be achieved in pairs of the following form

$$(m_1, \xi - m_2), (m_1, \xi - M_2), (m_2, \xi - m_1) \text{ and } (m_2, \xi - M_1)$$

Thus, for each value of ξ, we have that

$$\begin{aligned} (F_1 \bigodot F_2)(\xi) = \min\{&F_1(\xi - m_2) + F_2(m_2), F_1(x - M_2) + \\ &+ F_2(M_2), F_2(\xi - m_1) + F_1(m_1), F_2(\xi - M_1) + F_1(M_1)\} \end{aligned}$$

□

Unfortunately, the operator of the infimal convolution does not preserve the concave nature of the functions. In general, the result is a piecewise concave function. This means that the infimal convolution of more than two functions cannot be obtained by means of a simple reiteration of the aforestated lemma, but requires resorting to calculating the infimal convolution of several piecewise concave functions. To carry out this calculation, we shall interpret a piecewise concave function as the minimum function of several concave functions, preceding as shown in the following obvious lemma.

Lemma 2. *Let the function*

$$F(x) = \left\{ \begin{array}{ccc} F_1(x) & \textit{if} & x \in [m_1, M_1] \\ \ldots & \ldots & \ldots \\ F_k(x) & \textit{if} & x \in [m_k, M_k] \end{array} \right.$$

be piecewise concave (concave in each interval $[m_k, M_k]$). Thus,

$$F(x) = \min_{i \in \{1, \ldots k\}} F_i(x)$$

where, we have redefined each function $F_i(x)$ as

$$F_i(x) := \left\{ \begin{array}{ccc} F_i(x) & \textit{if} & x \in [m_i, M_i] \\ \infty & \textit{if} & x \notin [m_i, M_i] \end{array} \right. , i = 1, \ldots k$$

 ISBN: 978-84-614-6167-7

Once redefined in this way, the calculation of the infimal convolution of two piecewise concave functions requires a combinatorial exploration that is reflected in the following theorem.

Theorem 1. *Let* $F(x) := \min_{i\in A}(F_i(x))$ *and* $G(x) := \min_{i\in B}(G_i(x))$, *then:*

$$(F\bigodot G)(t) = \min_{(i,j)\in A\times B}(F_i\bigodot G_j)(t)$$

Proof.

$$\begin{aligned}(F\bigodot G)(t) &= \min_x(F(t-x)+G(x)) = \min_x(\min_{i\in A}(F_i(t-x)) + \min_{j\in B}(G_j(x)))\\ &= \min_x(\min_{(i,j)\in A\times B}(F_i(t-x)+G_j(x))) =\\ &= \min_{(i,j)\in A\times B}(\min_x(F_i(t-x)+G_j(x))) = \min_{(i,j)\in A\times B}(F_i\bigodot G_j)(t)\end{aligned}$$

□

This theorem justifies the construction of the infimal convolution of the two functions defined piecewise as the minimum function of all the possible infimal convolutions of "*pairs of pieces*".

Now, bearing in mind the associative nature of the infimal convolution operation, the infimal convolution may be calculated by means of a recursive process, carrying out n operations of infimal convolution considering the following recurrence:

$$H_1\bigodot H_2\bigodot\cdots\bigodot H_n = (H_1\bigodot H_2\bigodot\cdots\bigodot H_{n-1})\bigodot H_n$$

4 Algorithm and complexity

In this section we analyze the computational complexity of the previously proposed recursive algorithm for calculating the analytical solution for the piecewise concave quadratic functions. We first analyze the calculation of the minimum of a set of piecewise quadratic functions.

4.1 Algorithm

Let G be a quadratic function and let F be a piecewise quadratic function:

$$F(x) = \begin{cases} F_1(x) & \text{if} \quad x\in[m_1,M_1]\\ \ldots & \ldots \quad\quad \ldots\\ F_k(x) & \text{if} \quad x\in[m_k,M_k]\end{cases}$$

considering $F_j(x) := \infty$ if $x\notin[m_j,M_j]$ and $G(x) := \infty$ if $x\notin[\tilde{m},\tilde{M}]$. Hence,

$$F(x) = \min_{i\in A=\{1,\ldots k\}} F_i(x)$$

The calculation of the infimal convolution

$$(F\bigodot G)(x) = \min_i((F_i\bigodot G)(x))$$

 ISBN: 978-84-614-6167-7

is carried out in two phases:

PHASE (1) Calculation of $F_i \bigodot G$ for each i
PHASE (2) Calculate $\min_i (F_i \bigodot G)(x)$

4.2 Computational Complexity

The nature of the underlying problem in the calculation of the infimal convolution of piecewise concave functions suggests that the computational complexity of the algorithm is exponential seeing as it entails exploring all the combinations of intervals of concavity of the functions involved. In certain cases, this is effectively so; however, we shall see that the complexity is polynomial in some other cases.

Theorem 2. *Let* $\{F_i\}_{i=1}^n$, where $F_i(x) := \beta_i x - \gamma_i x^2$, *with* $\gamma_i > 0$, *with the same domain* $[0, M]$. *If* $F_i(x) \neq F_j(x)$ *for all* $0 \neq x \in [0, M]$, *then the computational complexity of the algorithm is cubic in order; i.e.* $T \in O(n^3)$.

5 Example

A program that solves the FPM problem was written using the Mathematica package and was then applied to one example using the previously developed model for the cost function

$$C(y) = \min_{\mathbf{z}} \sum_{i=1}^{n} \beta_i z_i - \gamma_i z_i^2$$

and maximum constraints for the $n = 4$ inputs.

$$\begin{aligned} & \max_{\mathbf{x},y} (py - \sum_{i=1}^{n} \left(\beta_i z_i - \gamma_i z_i^2\right)) \\ \text{s.t.} \quad & \sum_{i=1}^{n} z_i = y \\ & 0 \leq z_i \leq M_i;\ i = 1, ..., n \end{aligned}$$

The data on the inputs is summarized in Table 1.

Table 1. Example data.

i	1	2	3	4
β_i	1	2	3	4
γ_i	-0.01	-0.03	-0.03	-0.01
M_i	10	15	4	2

Applying the aforementioned algorithm, we have that the infimal convolution

$$C = (F_1 \bigodot F_2 \bigodot F_3 \bigodot F_4)$$

is a piecewise quadratic function:

$$C(y) = \begin{cases} y - 0.01y^2 & \text{if} \quad 0 \leq y \leq 10 \\ -14 + 2.6y - 0.03y^2 & \text{if} \quad 10 \leq y \leq 25 \\ -61.5 + 4.5y - 0.03y^2 & \text{if} \quad 25 \leq y \leq 29 \\ -80.64 + 4.58y - 0.01y^2 & \text{if} \quad 29 \leq y \leq 31 \end{cases}$$

 ISBN: 978-84-614-6167-7

Finally, considering different values of the price p, we calculate the solution to the FPM problem

$$\max_y(py - C(y))$$

The results are summarized in Table 2.

Table 2. Solution y^*.

p	2	1	$\frac{1}{2}$	5
y^*	25	10	0	31

As already mentioned, despite having the analytical cost expression, $C(y)$, the optimum level of output cannot be obtained via marginalistic techniques; i.e. $\partial C(y)/\partial y$ coincides with the price p. The maximum profit is always obtained with a level of output y^* in which either C is not differentiable or y^* is one of the extreme values of the interval $[0, \sum_{i=1}^{n} M_i]$.

In fact, for $p = 2 \to y^* = 25$ and for $p = 1 \to y^* = 10$, the solution is obtained from angle points of $C(y)$, whereas, as we have already seen, for $p = 1/2 \to y^* = 0$, i.e. production is not profitable, and for $p = 5 \to y^* = 31$, the maximum is produced at the technical maximum.

6 Conclusions

Concave quadratic problems often arise involving economies of scale. In this paper we present an algorithm for calculating the analytical solution for the classic firm's cost minimization problem in the case of economies of scale, with n inputs, maximum constraints for the inputs and a general output y (i.e. a family of monoparametric problems). The algorithm uses the infimal convolution of piecewise concave functions. For the firm's profit maximization problem, the solution cannot be obtained using derivatives and our method calculates the exact solution, without any kind of simplification, searching non-differentiable points of the analytical formulae of the cost or extreme values of the output.

References

[1] C. A. Floudas and V. Visweswaran, *Quadratic Optimization*, Handbook of Global Optimization (R. Horst and P.M. Pardalos, Editors), Kluwer Academic Publishers, (1994), 217-270.

[2] C. A. Floudas and P.M. Pardalos, Eds., *Encyclopedia of Optimization*, Kluwer Academic Publishers, 2001.

[3] G. A. Jehl and P. J. Reny, *Advanced Microeconomic Theory* (2nd Edition), Boston: Addison-Wesley, 2001.

[4] D. G. Luenberger, *Microeconomic Theory*, McGraw-Hill, 1995.

 ISBN: 978-84-614-6167-7

[5] H. R. Varian, *Intermediate Microeconomics*, W.W. Norton & Company, 7th Edition, 2005.

[6] W. Nicholson, *Microeconomic theory: basic principles and extensions,* South-Western/Thomson Learning, 2002.

[7] H. P. Benson, *Concave minimization: Theory, applications and algorithms*, in Horst, R. and Pardalos, P.M. (eds.), Handbook of Global Optimization, Kluwer, Dordrecht, (1995), 43-148.

[8] H. P. Benson, *Deterministic algorithms for constrained concave minimization: A unified critical survey*, Naval Research Logistics, **43** (1996), 765-795.

[9] J. J. Moreau, *Inf-convolution, sous-additivit e, convexit eriques*, J. Math. Pures et Appl. **49** (1970), 109-154.

[10] R. T. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, 1970.

[11] T. Stromberg, *The operation of infimal convolution*, Diss. Math. **352** (1996).

 ISBN: 978-84-614-6167-7

Proceedings of the 11th International Conference on Computational and Mathematical Methods in Science and Engineering, CMMSE 2011 26–30 June 2011.

Analysis of GPU thread structure in a multichannel audio application

José A. Belloch[1], F. J. Martínez-Zaldívar[3], Antonio M. Vidal[2] and Alberto González[3]

[1] *Audio and Communications Signal Processing Group. Instituto de Telecomunicaciones y Aplicaciones Multimedia, Universitat Politècnica de València (Spain)*

[2] *Departamento de Sistemas Informáticos y Computación, Universitat Politècnica de València (Spain)*

[3] *Departamento de Comunicaciones, Universitat Politècnica de València (Spain)*

emails: jobelrod@iteam.upv.es, fjmartin@dcom.upv.es, avidal@dsic.upv.es, agonzal@dcom.upv.es

Abstract

Audio effects like spatial sound, crosstalk cancellation, multiple equalizations, etc., are based on real time multichannel applications. They have experienced a major development in recent years due to its demand in theaters, funfairs, etc. Operations like convolutions or correlations are frequently used in multichannel audio applications and they require a high computing capability. Our target is to compute these operations efficiently. Graphics Processors Units (GPUs), a highly parallel programmable co-processors, offer the possibility of parallelizing these operations obtaining the desired efficiency. In this paper, different parallel programming environment parameters related to the GPU are analyzed in order to obtain the best performance in an audio multichannel application.

Key words: Convolution, FFT, GPU, Multichannel, Audio

1 Introduction

Multichannel applications have experienced a major development in recent years. Convolutions and lineal combinations among different signals are especially important in an audio application [6]. Also, a multichannel system [3] is configured by C_{in} inputs and C_{out} outputs. In case of an audio application, C_{in} is the number of audio sources and C_{out} the number of loudspeakers [4]. Figure 1 shows a multichannel system with two input signals and four output signals.

 ISBN: 978-84-614-6167-7

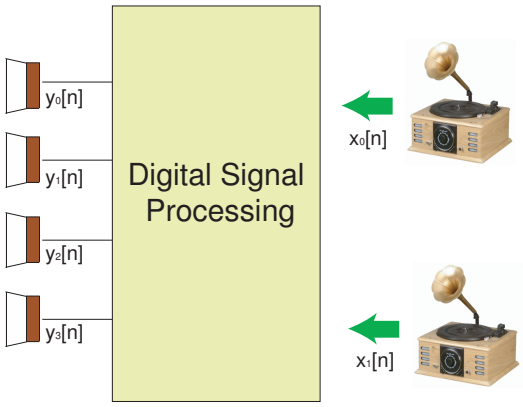


Figure 1: Multichannel system.

The block *Digital Signal Processing* is composed by different filters which modify several characteristics of the input signals. Here, multiple convolutions are carried out among input signals and filter impulse responses. Output signals from convolutions are combined among them according to the number of outputs. In general, any multichannel application output can be represented by Equation 1. Input signals (*sources* in the application) are represented by x_j, $j \in [0, C_{in} - 1]$) and output signals by y_i, $i \in [0, C_{out} - 1]$ (*loudspeakers*). The sequence $h_{ij}[n]$ represents the filter impulse response implemented between the *j-source* and *i-loudspeaker* and $*$ denotes the convolution operation. C_{tot} is the total number of necessary filters to execute the application ($C_{tot} = C_{in} \cdot C_{out}$)

$$y_i[n] = \sum_{j=0}^{M} (h_{ij}[n] * x_j[n]). \tag{1}$$

The convolution operation can be carried out using several FFT transformations and element-wise multiplications between blocks of signal samples and the correspondent filter [10]. Both of them are vectors in the frequency domain. FFT transformations are carried out efficiently by the NVIDIA FFT library, CUFFT [7]. However, carrying out multiple element-wise multiplications between different vectors (signal block and filter) will not be an easy task when multiple channels and large filters are involved in the application. In next sections, the optimum CUDA [2] parameters will be explained to obtain multiple element-wise multiplications in an efficient way.

2 Data Structure in a Multichannel System

In a real-time application, it is essential to give the response within the required time. In that sense, it must be taken into account, from where the input samples for every channel will be read. System response (*data processing*) must be at least as fast as the input-buffer filling, in order to avoid sample losses. Figure 2 shows a system with 4 sources and 2 loudspeakers. In a real-time application, it is clear that $tapp < tbuff$.

 ISBN: 978-84-614-6167-7

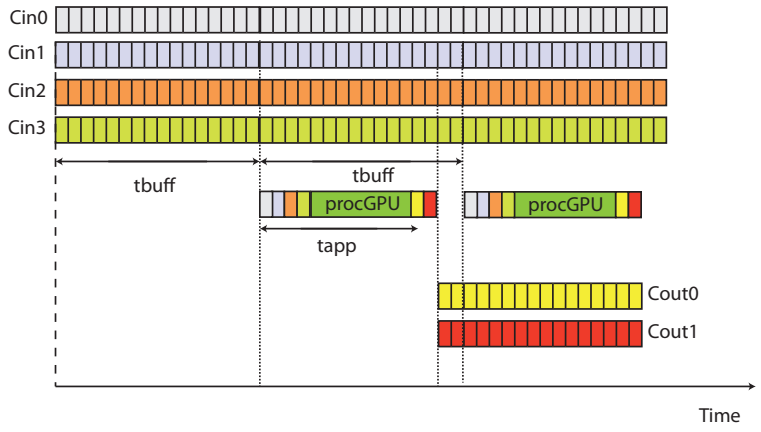


Figure 2: Response of a real-time system.

Also, when the size of h_{ij} is too large, this filter is usually fragmented into parts of equal size to the input buffer [5]. Following the overlap-save technique [1], samples in the input-buffer must be element-wise multiplied in the frequency domain for all the fragments of the filters. So, in case of 4 sources, and 2 loudspeakers, it will be necessary 8 filters (h_{00}, h_{01}, h_{02}, h_{03}, h_{10}, h_{11}, h_{12}, h_{13}). Using a tridimensional matrix structure with the fragments of the filters, data will be transferred to the GPU as Figure 3 shows, where L is the buffer size and F the number of fragments of the filter.

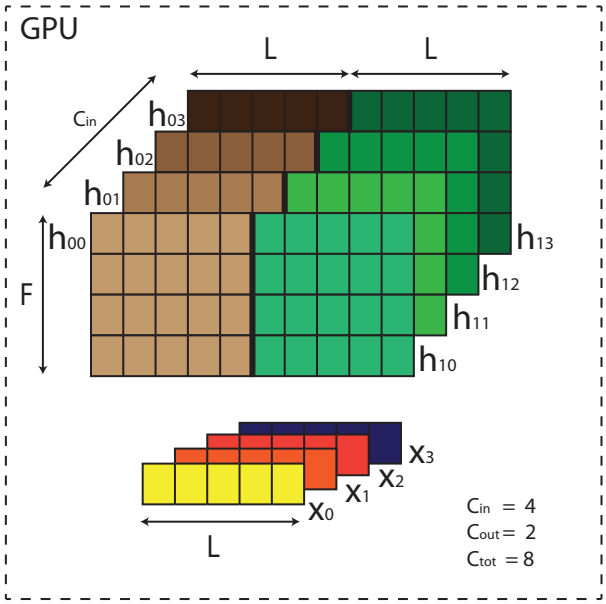


Figure 3: Data structure in the GPU.

 ISBN: 978-84-614-6167-7

3 Implementation on GPU

As incoming samples in the buffer must be element-wise multiplied with every fragment of the filter, they will be stored in the *shared-memory* in the GPU. Now a GPU-kernel must be configured. CUDA 4.0 [8] lets configure a tridimensional grid, composed of tridimensional blocks: (*ThreadIdx.x* , *ThreadIdx.y* and *ThreadIdx.z*), see Figure 4.

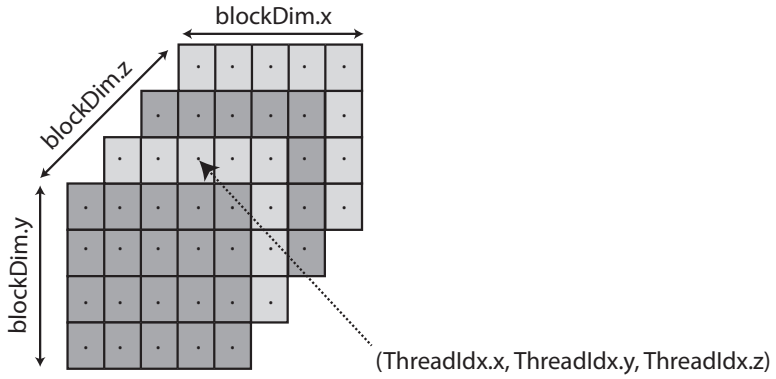


Figure 4: Tridimensional thread block.

Looking at Figure 3, CUDA grid is configured as:

```
<<< (Cout*L)/blockDim.x , F/blockDim.y , Cin/block.z >>>
```

Figure 5 clarifies the position of each threads.Each thread will access to its corresponding value in the GPU *shared-memory* and executes an element-wise multiplication obtaining the results in Figure 6.

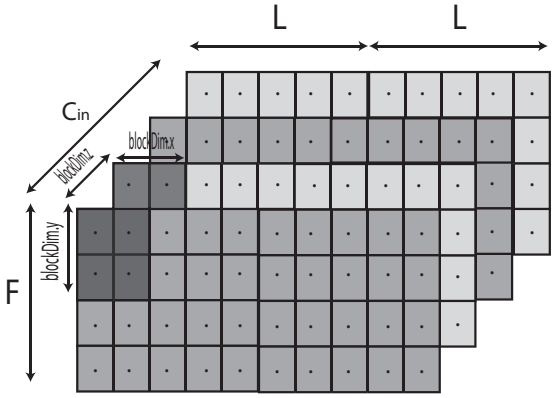


Figure 5: Thread block grid structure for *Cin*=4 and *Cout*=2.

The last step will be carried out for a bidimensional kernel which will accumulate all the values en the same plane. The second CUDA grid will be (see Figure 7):

```
<<< (Cout*L)/blockDim.x , F/blockDim.y , 1 >>>
```

 ISBN: 978-84-614-6167-7

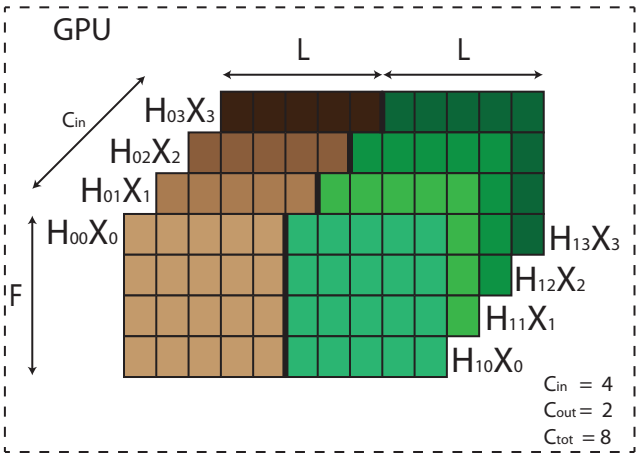


Figure 6: Thread Block Grid Structure for *Cin*=4 and *Cout*=2.

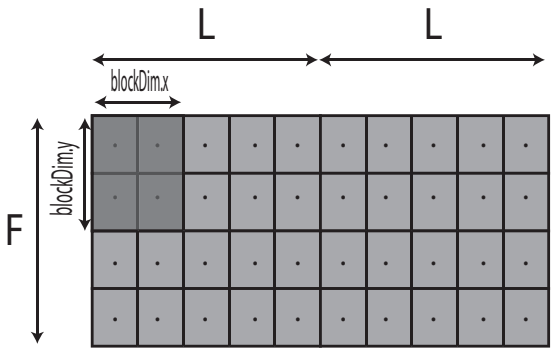


Figure 7: Thread Block Grid Structure for accumulating the resulting values

The second kernel will add up all the planes (Figure 8) according to the Equation 1.

4 Test System

There are a lot of possibilities for the values of *blockdim.x*, *blockdim.y* and *blockdim.z*. In our case, we have tested a typical audio system where every input-buffer size is 128 samples. Taking into account that one sample of one channel arrives every 1/44100 s, the *tbuff* showed in Figure 2 will be 2.9 ms. The filters used in the test have a length of 2048 coefficients. So, the value of F (rows of the tridimensional filters matrix) will depend on the value of *blockdim.y*.

The test executions evaluate the *tapp* (Figure 2), covering all the possiblities of the C_{in} and C_{out}, and try to look for the maximum number of filters C_{tot} that a GPU is able to manage in a real time application ($tapp < tbuff$). The test focuses mainly on different configurations varying the number of threads per block that NVIDIA recommends: 256, 512 and 1024. The device used for the experiments is a Fermi architecture

 ISBN: 978-84-614-6167-7

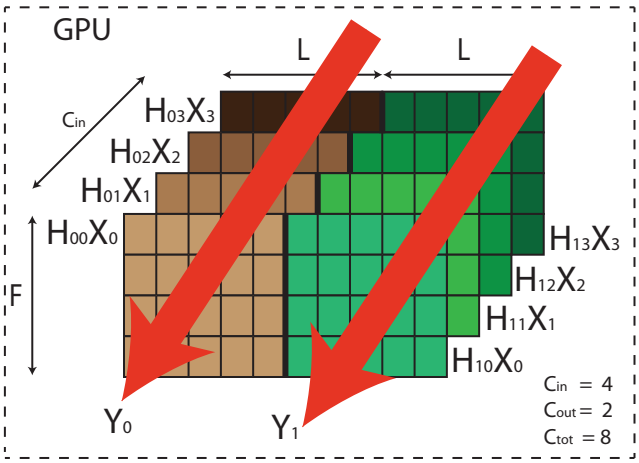


Figure 8: The second kernel will sum up all the planes

TESLA C2070 GPU[11].

5 Results and Conclusions

In this section, the maximum number of filters that different CUDA configurations can manage is analyzed. Table 1 shows different grid configurations with different number of threads per block that allows to manage a real time multichannel application (configurations with minimum *tapp*, which is lower than *tbuff*).

Num conf	C_{in}	C_{out}	C_{tot}	bdim.x	bdim.y	bdim.z	NumThreads per block	tapp
1	22	64	1408	16	8	2	256	2.6037ms
2	38	32	1216	8	16	2	256	2.7119ms
3	102	16	1632	32	8	2	512	2.7734ms
4	102	16	1632	16	16	2	512	2.889ms
5	22	64	1408	64	8	2	1024	2.7924ms
6	38	32	1216	32	16	2	1024	2.7422ms

Table 1: Maximum number of filters, (C_{tot}) which let manage a real-time multichannel application with different grid size. bdim.x, bdim.y and bdim.z refer to blockDim.x, blockDim.y and blockDim.z respectively.

As it is appreciated in Table 1, the configuration number 3 with *blockDim.x*=32, *blockDim.y*=8 and *blockDim.z*=2 (512 threads per block) achieves the larger number of filters (1632) that a GPU can manage in a real time multichannel application. Comparing configurations 1 and 5, we observe that the best results in time are not obtained

 ISBN: 978-84-614-6167-7

using the maximum number of threads that CUDA allows in a block. If the number of threads per block is high, there are not enough blocks for all the multiprocessors in the Fermi GPU. So, it is better to design the kernels launching a higher number of blocks without configuring the maximum number of threads possible per block, instead of setting maximum number of threads per block and less blocks. Also, it is noticeable (see configurations 3 and 4) that a larger number of threads in *blockDim.x* improves the performance because it provides coalescing access to *global memory*, even more when this number is multiple of 32 [9]. Although this property has not the importance of the previous architectures, it plays still a important role in the Fermi architecture. A good way to speedup the CUDA application is to find first the number of threads per block that achieves the best performance and then distribute the number of threads in the three dimensions. From an acoustic perspective, it is clear that every multichannel application will have its own optimum configuration of the CUDA parameters. Different sizes of input-buffer and lengths of filters could modify the values of *blockdim.x*, *blockdim.y* and *blockdim.z* to obtain the best performance. Therefore, the GPU can be used as a co-processor carrying out audio processing tasks. As a result, the GPU can free up resources of the CPU and/or increase the performance of the audio application.

Acknowledgements

This work has been supported by Spanish Ministry of Science and Innovation through grant TEC2009-13741, Regional Government Generalitat Valenciana through grant PROMETEO/2009/013, GV/2010/027 and NVIDIA through CUDA Community program.

References

[1] A. V. Oppenheim, A. S. Willsky and S. Hamid, *Signals and Systems*, Prentice Hall, 1997.

[2] David B. Kirk, and Wen-mei W. Hwu, *Programming Massively Parallel Processors*, Morgan Kaufman, 2010.

[3] E. Torick, *Highlights in the history of multichannel sound*, J. Audio. Eng. Soc. **46** (1998) 27–31.

[4] Xie Bo-Sun, Guan Shan-Qun, Liang Zhi-Qiang, Rao Dan, Yu Guang-Zheng, Zhan Cheng-Yun, and Zhong Xiao-Li, *Some Recent Works on Head-Related Transfer Functions and Virtual Auditory Display in China*, Audio Engineering Society Conference: 40th International Conference: Spatial Audio: Sense the Sound of Space.

[5] F. Wefers and J. Berg, *High-Performance real-time FIR-filtering using fast convolution on graphics hardware*, Proc. of the 13th Conference on Digital Audio Effects (2010).

 ISBN: 978-84-614-6167-7

[6] Y. Huang, J. Chen and J. Benesty, *Inmerse Audio Schemes*, IEEE Signal Processing Magazine **28** (2011) 20–32.

[7] CUFFT library: "http://developer.download.nvidia.com/compute/cuda/4_0_rc2 /toolkit/docs/CUFFT_Library.pdf"

[8] CUDA Toolkit 4.0: "http://developer.nvidia.com/cuda-toolkit-40"

[9] NVIDIA CUDA C Best Practices Guide: "http://developer.download.nvidia.com /compute/cuda/4_0_rc2/toolkit/docs/CUDA_C_Best_Practices_Guide.pdf"

[10] S. S. Soliman, and Mandyam D. Srinath, *Continuous and Discrete Signals and Systems*, *Prentice Hall*, 1997.

[11] NVIDIA FERMI Generation: "http://www.nvidia.com/content/PDF/ fermi_white_papers/NVIDIA_Fermi_Compute_Architecture_Whitepaper.pdf"

 ISBN: 978-84-614-6167-7

Proceedings of the 11th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2011
26–30 June 2011.

A GFDM with PML for seismic wave equation in heterogeneous media .

J.J. Benito[1], F. Ureña[2], L. Gavete[3] and E. Salete[1]

[1] *Departamento de Construcción y Fabricación, Universidad Nacional de Educación a Distancia*

[2] *Departamento de Matemática Aplicada, Universidad de Castilla-La Mancha*

[3] *Departamento de Matemática Aplicada a los Recursos Naturales, Universidad Politécnica de Madrid*

emails: jbenito@ind.uned.es, francisco.urena@uclm.es, lu.gavete@upm.es, esalete@ind.uned.es

Abstract

The interior of the Earth is heterogeneous with different material and it may have complex geometry. The free surface can also be uneven. Therefore, the use of a meshless method with the possibility of using and irregular grid-point distribution can be interest for modelling this kind of problem.
This paper shows the application of GFDM to the problem of seismic wave propagation in 2-D for homogeneous and heterogeneous media. To use this method in unbounded domains one must truncate the computational grid-point avoiding reflection from the edges. PML absorbing boundary condition has then been included in the numerical model proposed in this work.

Key words: meshless methods, generalized finite difference method, moving least squares, seismic waves, perfectly matched layer.
MSC 2000: 65M06, 65M12, 74S20, 80M20

1 Introduction

The GFDM [1, 6, 7] is a robust numerical method applicable to structurally complex media. Due to its relative accuracy and computational efficiency it is the dominant method in modelling earthquake motion [4]. The perfectly matched layer (PML) absorbing boundary

 ISBN: 978-84-614-6167-7

performs more efficiently and more accurately than most traditional or differential equation-based absorbing boundaries [2, 3, 5].
The paper is organized as follows. Section 1 is an introduction. Section 2 describes the GFDM obtaining the explicit generalized differences schemes for the seismic waves propagation and the heterogeneous media approach. Section 3 shows the recursive equations where the model include PML in once and two directions. In Section 4 some numerical results are included. Finally, in Section 5 some conclusions are given.

2 Explicit Generalized Differences Schemes for the seismic waves propagation problem for a perfectly elastic, heterogeneous and isotropic medium

2.1 Equation of motion

The equations of motion for a perfectly elastic, homogeneous, isotropic medium in 2-D are

$$\begin{cases} \dfrac{\partial^2 U_x(x,y,t)}{\partial t^2} = \alpha^2 \dfrac{\partial^2 U_x(x,y,t)}{\partial x^2} + \beta^2 \dfrac{\partial^2 U_x(x,y,t)}{\partial y^2} + (\alpha^2 - \beta^2) \dfrac{\partial^2 U_y(x,y,t)}{\partial x \partial y} \\ \dfrac{\partial^2 U_y(x,y,t)}{\partial t^2} = \beta^2 \dfrac{\partial^2 U_y(x,y,t)}{\partial x^2} + \alpha^2 \dfrac{\partial^2 U_y(x,y,t)}{\partial y^2} + (\alpha^2 - \beta^2) \dfrac{\partial^2 U_x(x,y,t)}{\partial x \partial y} \end{cases} \tag{1}$$

with the initial conditions

$$U_x(x,y,0) = f_1(x,y); U_y(x,y,0) = f_2(x,y); \frac{\partial U_x(x,y,0)}{\partial t} = f_3(x,y); \frac{\partial U_y(x,y,0)}{\partial t} = f_4(x,y) \tag{2}$$

and the boundary condition

$$\begin{cases} a_1 U_x(x_0,y_0,t) + b_1 \frac{\partial U_x(x_0,y_0,t)}{\partial n} = g_1(t) \\ a_2 U_y(x_0,y_0,t) + b_2 \frac{\partial U_y(x_0,y_0,t)}{\partial n} = g_2(t) \end{cases} \qquad en \quad \Gamma \tag{3}$$

where $f_1(x,y)$, $f_2(x,y)$, $f_3(x,y)$, $f_4(x,y)$, $g_1(t)$ y $g_2(t)$ are showed functions, $\alpha = \sqrt{\dfrac{\lambda + 2\mu}{\rho}}$, $\beta = \sqrt{\dfrac{\mu}{\rho}}$, ρ is the density, λ and μ are Lamé elastic coefficients and Γ is the boundary of Ω.

2.2 A GFDM Explicit Scheme

The aim is to obtain explicit linear expressions for the approximation of partial derivatives in the points of the domain. First of all, an irregular grid or cloud of points is generated in the domain $\Omega \cup \Gamma$. On defining the central node with a set of nodes surrounding that node, the star then refers to a group of established nodes in relation to a central node. Every

 ISBN: 978-84-614-6167-7

node in the domain has an associated star assigned to it.
Following [1, 6, 7], the explicit difference formulae for the spatial derivatives are obtained

$$\begin{cases} \frac{\partial^2 U_x(x_0,y_0,n\triangle t)}{\partial x^2} = -m_0 u^n_{x,0} + \sum_{j=1}^N m_j u^n_{x,j} & \frac{\partial^2 U_y(x_0,y_0,n\triangle t)}{\partial x^2} = -m_0 u^n_{y,0} + \sum_{j=1}^N m_j u^n_{y,j} \\ \frac{\partial^2 U_x(x_0,y_0,n\triangle t)}{\partial y^2} = -\eta_0 u^n_{x,0} + \sum_{j=1}^N \eta_j u^n_{x,j} & \frac{\partial^2 U_y(x_0,y_0,n\triangle t)}{\partial y^2} = -\eta_0 u^n_{y,0} + \sum_{j=1}^N \eta_j u^n_{y,j} \\ \frac{\partial^2 U_x(x_0,y_0,n\triangle t)}{\partial x \partial y} = -\zeta_0 u^n_{x,0} + \sum_{j=1}^N \zeta_j u^n_{x,j} & \frac{\partial^2 U_y(x_0,y_0,n\triangle t)}{\partial x \partial y} = -\zeta_0 u^n_{y,0} + \sum_{j=1}^N \zeta_j u^n_{y,j} \end{cases} \quad (4)$$

where N is the number of nodes in the star whose central node has the coordinates (x_0, y_0) (in this work $N = 8$ and the are selected by using the four quadrants criteria [7]).
m_0, η_0, ζ_0 are the coefficients that multiply the approximate values of the functions U and V at the central node for the time $n\triangle t$ (u^n_0 and v^n_0 respectively) in the generalized finite difference explicit expressions for the space derivatives.
m_j, η_j, ζ_j are the coefficients that multiply the approximate values of the functions U and V at the rest of the star nodes for the time $n\triangle t$ (u^n_j and v^n_j respectively) in the generalized finite difference explicit expressions for the space derivatives.
This scheme uses the central-difference formulae for the time derivative

$$\frac{\partial^2 U_x(x_0,y_0,n\triangle t)}{\partial t^2} = \frac{u^{n+1}_{x,0} - 2u^n_{x,0} + u^{n-1}_{x,0}}{(\triangle t)^2}; \frac{\partial^2 U_y(x_0,y_0,n\triangle t)}{\partial t^2} = \frac{u^{n+1}_{y,0} - 2u^n_{y,0} + u^{n-1}_{y,0}}{(\triangle t)^2} \quad (5)$$

The replacement in Eq. 1 of the explicit expressions obtained for the partial derivatives leads to the explicit difference scheme. As we are using a explicit method, we have studied the stability and we have obtained the star stability condition in [6], we have also studied in [6] for the phase and group velocities.

2.3 Heterogeneous formulation

If density ρ and Lame's coefficients λ and μ are functions of spatial coordinates, in agreement with [4], in seismological problems the homogeneous approach can be complicated and, then, the so-called heterogeneous formulation is preferred. Therefore, the same formulas are used for all points in the domain and material discontinuities are accounted by spatial variation of the material parameters.

 ISBN: 978-84-614-6167-7

3 Recursive Equations

3.1 Recursive equations for PML in x-direction.

For computational convenience, we split the second order equations of motion 1 into five coupled first order equations by introducing the new field variables $\gamma_{xx}, \gamma_{xy}, \gamma_{yy}$

$$\begin{cases} \rho\dfrac{\partial U_x(x,y,t)}{\partial t} = \dfrac{\partial \gamma_{xx}(x,y,t)}{\partial x} + \dfrac{\partial \gamma_{xy}(x,y,t)}{\partial y} \\ \rho\dfrac{\partial U_y(x,y,t)}{\partial t} = \dfrac{\partial \gamma_{xy}(x,y,t)}{\partial x} + \dfrac{\partial \gamma_{yy}(x,y,t)}{\partial y} \\ \dfrac{\partial \gamma_{xx}(x,y,t)}{\partial t} = (\lambda+2\mu)\dfrac{\partial U_x(x,y,t)}{\partial x} + \lambda\dfrac{\partial U_y(x,y,t)}{\partial y} \\ \dfrac{\partial \gamma_{xy}(x,y,t)}{\partial t} = \mu\dfrac{\partial U_x(x,y,t)}{\partial y} + \mu\dfrac{\partial U_y(x,y,t)}{\partial x} \\ \dfrac{\partial \gamma_{yy}(x,y,t)}{\partial t} = \lambda\dfrac{\partial U_x(x,y,t)}{\partial x} + (\lambda+2\mu)\dfrac{\partial U_y(x,y,t)}{\partial y} \end{cases} \tag{6}$$

We shall make two simplifications, we shall assume that the space far from the region of interest is homogeneous, linear and time invariant.Then, under these assumptions, the radiating solution in infinite space must be (superposition of plane waves):

$$\boldsymbol{\omega}(\boldsymbol{x},t) = \boldsymbol{W}(\boldsymbol{x},t)e^{i(\boldsymbol{\kappa}\cdot\boldsymbol{x}-wt)} \tag{7}$$

where $\boldsymbol{W}(\boldsymbol{x},t) = \{U_x, U_y, \gamma_{xx}, \gamma_{xy}, \gamma_{yy}\}^T$, w is the angular frequency and $\boldsymbol{\kappa}$ is the wavevector As $\boldsymbol{\omega}$ is an analytic function of $\boldsymbol{x}$,then we can analytically continue it, evaluating the solution at complex values of $\boldsymbol{x}$. Then, the solution is not changed in the region of interest and the reflections are avoided.

$$\begin{cases} U_x(x,y,t) = u_x(x,y)e^{-iwt} \Rightarrow \dot{U}_x(x,y,t) = -iwu_x(x,y)e^{-iwt} = -iwU_x(x,y,t) \\ U_y(x,y,t) = u_y(x,y)e^{-iwt} \Rightarrow \dot{U}_y(x,y,t) = -iwu_y(x,y)e^{-iwt} = -iwU_y(x,y,t) \\ \gamma_{xx}(x,y,t) = \Gamma_{xx}(x,y)e^{-iwt} \Rightarrow \dot{\gamma}_{xx}(x,y,t) = -iw\Gamma_{xx}(x,y)e^{-iwt} = -iw\gamma_{xx}(x,y,t) \\ \gamma_{xy}(x,y,t) = \Gamma_{xy}(x,y)e^{-iwt} \Rightarrow \dot{\gamma}_{xy}(x,y,t) = -iw\Gamma_{xy}(x,y)e^{-iwt} = -iw\gamma_{xy}(x,y,t) \\ \gamma_{yy}(x,y,t) = \Gamma_{yy}(x,y)e^{-iwt} \Rightarrow \dot{\gamma}_{yy}(x,y,t) = -iw\Gamma_{yy}(x,y)e^{-iwt} = -iw\gamma_{yy}(x,y,t) \end{cases} \tag{8}$$

Thus, we have a complex coordinate

$$\tilde{x} = x + if \tag{9}$$

As this complex coordinate is inconvenient, we have a change variables in this region (PML)

$$\partial\tilde{x} = (1 + i\frac{df}{dx})\partial x \tag{10}$$

 ISBN: 978-84-614-6167-7

In order to have an attenuation rate in the PML independent of frequency, we have

$$\frac{df}{dx} = \frac{\delta_x(x)}{w} \tag{11}$$

where δ_x is some function of x.
PML x-dir can be conceptually assumed up by a single transformation of the original equation. Then wherever an x derivative appears in the wave equations, it is replaced in the form

$$\frac{\partial}{\partial x} \rightarrow \frac{1}{1 + i\frac{\delta_x(x)}{w}} \frac{\partial}{\partial x} \tag{12}$$

The equations are frequency-dependent, and to advoid it a solution is to use an auxiliary differential equation (ADE) approach in the implementation of PML. The following equations are obtained

$$\begin{cases} \frac{\partial U_x(x,y,t)}{\partial t} = \frac{1}{\rho}[\frac{\partial \gamma_{xx}(x,y,t)}{\partial x} + \frac{\partial \gamma_{xy}(x,y,t)}{\partial y}] + \psi_1(x,y,t) - \delta_x U_x(x,y,t) \\ \frac{\partial U_y(x,y,t)}{\partial t} = \frac{1}{\rho}[\frac{\partial \gamma_{xy}(x,y,t)}{\partial x} + \frac{\partial \gamma_{yy}(x,y,t)}{\partial y}] + \psi_2(x,y,t) - \delta_x U_y(x,y,t) \\ \frac{\partial \gamma_{xx}(x,y,t)}{\partial t} = (\lambda + 2\mu)\frac{\partial U_x(x,y,t)}{\partial x} + \lambda \frac{\partial U_y(x,y,t)}{\partial y} + \psi_3(x,y,t) - \delta_x \gamma_{xx}(x,y,t) \\ \frac{\partial \gamma_{xy}(x,y,t)}{\partial t} = \mu \frac{\partial U_x(x,y,t)}{\partial y} + \mu \frac{\partial U_y(x,y,t)}{\partial x} + \psi_4(x,y,t) - \delta_x \gamma_{xy}(x,y,t) \\ \frac{\partial \gamma_{yy}(x,y,t)}{\partial t} = \lambda \frac{\partial U_x(x,y,t)}{\partial x} + (\lambda + 2\mu)\frac{\partial U_y(x,y,t)}{\partial y} + \psi_5(x,y,t) - \delta_x \gamma_{yy}(x,y,t) \\ \frac{\partial \psi_1(x,y,t)}{\partial t} = \frac{\delta_x}{\rho}\frac{\partial \gamma_{xy}(x,y,t)}{\partial y} \\ \frac{\partial \psi_2(x,y,t)}{\partial t} = \frac{\delta_x}{\rho}\frac{\partial \gamma_{yy}(x,y,t)}{\partial y} \\ \frac{\partial \psi_3(x,y,t)}{\partial t} = \lambda \delta_x \frac{\partial U_y(x,y,t)}{\partial y} \\ \frac{\partial \psi_4(x,y,t)}{\partial t} = \mu \delta_x \frac{\partial U_x(x,y,t)}{\partial y} \\ \frac{\partial \psi_5(x,y,t)}{\partial t} = (\lambda + 2\mu)\delta_x \frac{\partial U_y(x,y,t)}{\partial y} \end{cases} \tag{13}$$

Where the five last equations 13 are ADE approach and the new field variables

$$\begin{cases} \psi_1(x,y,t) = \frac{1}{\rho} i \frac{\delta_x}{w} \frac{\partial \gamma_{xy}(x,y,t)}{\partial y} \\ \psi_2(x,y,t) = \frac{1}{\rho} i \frac{\delta_x}{w} \frac{\partial \gamma_{yy}(x,y,t)}{\partial y} \\ \psi_3(x,y,t) = i\lambda \frac{\delta_x}{w} \frac{\partial U_y(x,y,t)}{\partial y} \\ \psi_4(x,y,t) = i\mu \frac{\delta_x}{w} \frac{\partial U_x(x,y,t)}{\partial y} \\ \psi_5(x,y,t) = i(\lambda + 2\mu) \frac{\delta_x}{w} \frac{\partial U_y(x,y,t)}{\partial y} \end{cases} \tag{14}$$

 ISBN: 978-84-614-6167-7

3.1.1 An scheme in GDFM for domain of interest.

Following [1, 6, 7], the explicit difference formulae for the spatial derivatives of a function are obtained,

$$\frac{\partial U_x(x_0,y_0,n\triangle t)}{\partial x} = -m_{1,0}u^n_{x,0} + \sum_{j=1}^{N} m_{1,j}u^n_{x,j}; \frac{\partial U_x(x_0,y_0,n\triangle t)}{\partial y} = -m_{2,0}u^n_{x,0} + \sum_{j=1}^{N} m_{2,j}u^n_{x,j} \tag{15}$$

and similarly for first spatial derivatives of the functions: $U_y, \gamma_{xx}, \gamma_{xy}, \gamma_{yy}, \psi_1, \psi_2, \psi_3, \psi_4, \psi_5$ Substituting Eq. 16 into Eq. 7 the explicit difference scheme in GFDM for elastic part is obtained

$$\begin{cases} u^{n+1}_{x,0} = u^n_{x,0} + \frac{\triangle t}{\rho}[-m_{1,0}\gamma^n_{xx,0} + \sum_{j=1}^{N} m_{1,j}\gamma^n_{xx,j} - m_{2,0}\gamma^n_{xy,0} + \sum_{j=1}^{N} m_{2,j}\gamma^n_{xy,j}] \\ u^{n+1}_{y,0} = u^n_{y,0} + \frac{\triangle t}{\rho}[-m_{1,0}\gamma^n_{xy,0} + \sum_{j=1}^{N} m_{1,j}\gamma^n_{xy,j} - m_{2,0}\gamma^n_{yy,0} + \sum_{j=1}^{N} m_{2,j}\gamma^n_{yy,j}] \\ \gamma^{n+1}_{xx,0} = \gamma^n_{xx,0} + \triangle t[(\lambda+2\mu)(-m_{1,0}u^n_{x,0} + \sum_{j=1}^{N} m_{1,j}u^n_{x,j}) \\ +\lambda(-m_{2,0}u^n_{y,0} + \sum_{j=1}^{N} m_{y,j}u^n_{2,j})] \\ \gamma^{n+1}_{xy,0} = \gamma^n_{xy,0} + \triangle t[\mu(-m_{2,0}u^n_{x,0} + \sum_{j=1}^{N} m_{2,j}u^n_{x,j}) \\ +\mu(-m_{1,0}u^n_{y,0} + \sum_{j=1}^{N} m_{1,j}u^n_{y,j})] \\ \gamma^{n+1}_{yy,0} = \gamma^n_{yy,0} + \triangle t[\lambda(-m_{1,0}u^n_{x,0} + \sum_{j=1}^{N} m_{1,j}u^n_{x,j}) \\ +(\lambda+2\mu)(-m_{2,0}u^n_{y,0} + \sum_{j=1}^{N} m_{2,j}u^n_{y,j})] \end{cases} \tag{16}$$

 ISBN: 978-84-614-6167-7

3.1.2 An scheme in GDFM for PML part.

Substituting Eqs. 15 into Eqs. 13 the explicit difference scheme in GFDM for PML part is obtained

$$
\begin{cases}
u_{x,0}^{n+1} = u_{x,0}^{n} + \dfrac{\triangle t}{\rho}[-m_{1,0}\gamma_{xx,0}^{n} + \sum_{j=1}^{N} m_{1,j}\gamma_{xx,j}^{n} - \\
m_{2,0}\gamma_{xy,0}^{n} + \sum_{j=1}^{N} m_{2,j}\gamma_{xy,j}^{n}] + \triangle t[\psi_{1,0}^{n} - \delta_x u_{x,0}^{n}] \\
u_{y,0}^{n+1} = u_{y,0}^{n} + \dfrac{\triangle t}{\rho}[-m_{1,0}\gamma_{xy,0}^{n} + \sum_{j=1}^{N} m_{1,j}\gamma_{xy,j}^{n} - \\
m_{2,0}\gamma_{yy,0}^{n} + \sum_{j=1}^{N} m_{2,j}\gamma_{yy,j}^{n}] + \triangle t[\psi_{2,0}^{n} - \delta_x u_{y,0}^{n}] \\
\gamma_{xx,0}^{n+1} = \gamma_{xx,0}^{n} + \triangle t[(\lambda+2\mu)(-m_{1,0}u_{x,0}^{n} + \sum_{j=1}^{N} m_{1,j}u_{x,j}^{n}) + \\
\lambda(-m_{2,0}u_{y,0}^{n} + \sum_{j=1}^{N} m_{2,j}u_{y,j}^{n})] + \triangle t[\psi_{3,0}^{n} - \delta_x \gamma_{xx,0}^{n}] \\
\gamma_{xy,0}^{n+1} = \gamma_{xy,0}^{n} + \triangle t[\mu(-m_{2,0}u_{x,0}^{n} + \sum_{j=1}^{N} m_{2,j}u_{x,j}^{n}) + \\
\mu(-m_{1,0}u_{y,0}^{n} + \sum_{j=1}^{N} m_{1,j}u_{y,j}^{n})] + \triangle t[\psi_{4,0}^{n} - \delta_x \gamma_{xy,0}^{n}] \\
\gamma_{yy,0}^{n+1} = \gamma_{yy,0}^{n} + \triangle t[\lambda(-m_{1,0}u_{x,0}^{n} + \sum_{j=1}^{N} m_{1,j}u_{x,j}^{n}) + \\
(\lambda+2\mu)(-m_{2,0}u_{y,0}^{n} + \sum_{j=1}^{N} m_{2,j}u_{y,j}^{n})] + \triangle t[\psi_{5,0}^{n} - \delta_x \gamma_{yy,0}^{n}] \\
\psi_{1,0}^{n+1} = \psi_{1,0}^{n} + \dfrac{\triangle t}{\rho}\delta_x[-m_{2,0}\gamma_{xy,0}^{n} + \sum_{j=1}^{N} m_{2,j}\gamma_{xy,j}^{n}] \\
\psi_{2,0}^{n+1} = \psi_{2,0}^{n} + \dfrac{\triangle t}{\rho}\delta_x[-m_{2,0}\gamma_{yy,0}^{n} + \sum_{j=1}^{N} m_{2,j}\gamma_{yy,j}^{n}] \\
\psi_{3,0}^{n+1} = \psi_{3,0}^{n} + \lambda\triangle t\delta_x[-m_{2,0}u_{y,0}^{n} + \sum_{j=1}^{N} m_{2,j}u_{y,j}^{n}] \\
\psi_{4,0}^{n+1} = \psi_{4,0}^{n} + \mu\triangle t\delta_x[-m_{2,0}u_{x,0}^{n} + \sum_{j=1}^{N} m_{2,j}u_{x,j}^{n}] \\
\psi_{5,0}^{n+1} = \psi_{5,0}^{n} + (\lambda+2\mu)\triangle t\delta_x[-m_{2,0}u_{y,0}^{n} + \sum_{j=1}^{N} m_{2,j}u_{y,j}^{n}]
\end{cases}
\tag{17}
$$

 ISBN: 978-84-614-6167-7

3.2 Recursive equations with PML in x-direction and y-direction.

In this case the transformation are

$$\begin{cases} \dfrac{\partial}{\partial x} \to \dfrac{\partial}{\partial x}(1+i\dfrac{\delta}{w})^{-1} \\ \dfrac{\partial}{\partial y} \to \dfrac{\partial}{\partial y}(1+i\dfrac{\delta}{w})^{-1} \end{cases} \tag{18}$$

and, if we perform these transformations, we obtain

$$\begin{cases} \dfrac{\partial U_x(x,y,t)}{\partial t} = \dfrac{1}{\rho}[\dfrac{\partial \gamma_{xx}(x,y,t)}{\partial x} + \dfrac{\partial \gamma_{xy}(x,y,t)}{\partial y}] - \delta U_x(x,y,t) \\ \dfrac{\partial U_y(x,y,t)}{\partial t} = \dfrac{1}{\rho}[\dfrac{\partial \gamma_{xy}(x,y,t)}{\partial x} + \dfrac{\partial \gamma_{yy}(x,y,t)}{\partial y}]) - \delta U_y(x,y,t) \\ \dfrac{\partial \gamma_{xx}(x,y,t)}{\partial t} = (\lambda+2\mu)\dfrac{\partial U_x(x,y,t)}{\partial x} + \lambda\dfrac{\partial U_y(x,y,t)}{\partial y} - \delta\gamma_{xx}(x,y,t) \\ \dfrac{\partial \tau_{xy}(x,y,t)}{\partial t} = \mu\dfrac{\partial U_x(x,y,t)}{\partial y} + \mu\dfrac{\partial U_y(x,y,t)}{\partial x} - \delta\gamma_{xy}(x,y,t) \\ \dfrac{\partial \gamma_{yy}(x,y,t)}{\partial t} = \lambda\dfrac{\partial U_x(x,y,t)}{\partial x} + (\lambda+2\mu)\dfrac{\partial U_y(x,y,t)}{\partial y} - \delta\gamma_{yy}(x,y,t) \end{cases} \tag{19}$$

3.2.1 An scheme in GDFM for elastic part.

The explicit difference scheme for the elastic part is given by Eq. 16

 ISBN: 978-84-614-6167-7

3.2.2 An scheme in GDFM for PML part.

Substituting Eqs. 15 into Eqs. 19 the explicit difference scheme in GFDM for PML part is obtained

$$\begin{cases} u_{x,0}^{n+1} = u_{x,0}^{n} + \dfrac{\triangle t}{\rho}[-m_{1,0}\gamma_{xx,0}^{n} + \sum_{j=1}^{N} m_{1,j}\gamma_{xx,j}^{n} - \\ m_{2,0}\gamma_{xy,0}^{n} + \sum_{j=1}^{N} m_{2,j}\gamma_{xy,j}^{n}] - \triangle t\delta u_{x,0}^{n} \\ u_{y,0}^{n+1} = u_{y,0}^{n} + \dfrac{\triangle t}{\rho}[-m_{1,0}\gamma_{xy,0}^{n} + \sum_{j=1}^{N} m_{1,j}\gamma_{xy,j}^{n} - \\ m_{2,0}\gamma_{yy,0}^{n} + \sum_{j=1}^{N} m_{2,j}\gamma_{yy,j}^{n}] - \triangle t\delta u_{y,0}^{n} \\ \gamma_{xx,0}^{n+1} = \gamma_{xx,0}^{n} + \triangle t[(\lambda+2\mu)(-m_{1,0}u_{x,0}^{n} + \sum_{j=1}^{N} m_{1,j}u_{x,j}^{n}) + \\ \lambda(-m_{2,0}u_{y,0}^{n} + \sum_{j=1}^{N} m_{y,j}u_{y,j}^{n})] - \triangle t\delta\gamma_{xx,0}^{n} \\ \gamma_{xy,0}^{n+1} = \gamma_{xy,0}^{n} + \triangle t[\mu(-m_{2,0}u_{x,0}^{n} + \sum_{j=1}^{N} m_{2,j}u_{x,j}^{n}) + \\ \mu(-m_{1,0}u_{y,0}^{n} + \sum_{j=1}^{N} m_{1,j}u_{y,j}^{n})] - \triangle t\delta\gamma_{xy,0}^{n} \\ \gamma_{yy,0}^{n+1} = \gamma_{yy,0}^{n} + \triangle t[\lambda(-m_{1,0}u_{x,0}^{n} + \sum_{j=1}^{N} m_{1,j}u_{x,j}^{n}) + \\ (\lambda+2\mu)(-m_{2,0}u_{y,0}^{n} + \sum_{j=1}^{N} m_{2,j}u_{y,j}^{n})] - \triangle t\delta\gamma_{yy,0}^{n} \end{cases} \tag{20}$$

4 Numerical Results

4.1 Discretization and wavelength

Table 1 shows the values of the global error, for $n = 500$, for several analytical solutions of the problem Eq.1, with $\Omega = [0,1] \times [0,1] \subset \mathbf{R}^2$, with Dirichlet boundary conditions and initial conditions

$$U_x(x,y,0) = \sin x \sin y; U_y(x,y,0) = \cos x \cos y; \frac{\partial U_x(x,y,0)}{\partial t} = 0; \frac{\partial U_y(x,y,0)}{\partial t} = 0 \tag{21}$$

using the irregular mesh with 121 nodes (see Fig. 1) with $IIC = 0.8944$, $n = 500$ and $\triangle t = 0.01$.
The weighting function is

$$w(h_{jx}, h_{jy}) = \frac{1}{(\sqrt{h_{jx}^2 + h_{jy}^2})^3} \tag{22}$$

 ISBN: 978-84-614-6167-7

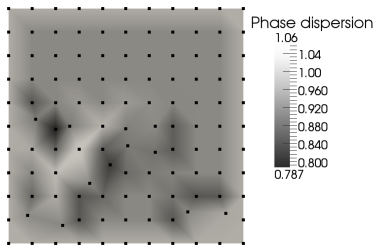


Figure 1: Dispersion of the waves P in the irregular mesh fig. 2 with $IIC = 0.8944$

and the criterion for the selection of star nodes is the quadrant criterion [1, 2, 4]. The global error is evaluated for each time increment, in the last time step considered, using the following formula

$$Global \quad error = \frac{\sqrt{\frac{\sum_{j=1}^{NT}(sol(j)-exac(j))^2}{NT}}}{|exac_{max}|} \tag{23}$$

where $sol(j)$ is the GFDM solution at the node j $exac(j)$ is the exact value of the solution at the node j, $exac_{max}$ is the maximum value of the exact solution in the cloud of nodes considered and NT is the total number of nodes of the domain.
IIC is the index of irregularity of the cloud and it is the smallest of the index of irregularity of the star (IIS) for all the nodes of domain, define by

$$IIS_{(x_0,y_0)} = \frac{\sqrt{5}(\sqrt{2}+1)}{\sqrt{3(|\overline{m_0}| + |\overline{\eta_0}| + \sqrt{(\overline{m_0}+\overline{\eta_0})^2 + \overline{\zeta_0}^2})}} \tag{24}$$

The analytical solutions are:

$$U_x(x,y,t) = \cos(\sqrt{2}\beta t)\sin x \sin y; \quad U_y(x,y,t) = \cos(\sqrt{2}\beta t)\cos x \cos y \tag{25}$$

$$\begin{cases} U(x,y,t) = \cos(0.5\sqrt{2}\beta t)\sin(0.5x)\sin(0.5y) \\ V(x,y,t) = \cos(0.5\sqrt{2}\beta t)\cos(0.5x)\cos(0.5y) \end{cases} \tag{26}$$

$$\begin{cases} U(x,y,t) = \cos(2\sqrt{2}\beta t)\sin(2x)\sin(2y) \\ V(x,y,t) = \cos(2\sqrt{2}\beta t)\cos(2x)\cos(2y) \end{cases} \tag{27}$$

$$\begin{cases} U(x,y,t) = \cos(4\pi\beta t)\sin(2\sqrt{2}\pi x)\sin(2\sqrt{2}\pi y) \\ V(x,y,t) = \cos(4\pi\beta t)\cos(2\sqrt{2}\pi x)\cos(2\sqrt{2}\pi y) \end{cases} \tag{28}$$

$$\begin{cases} U(x,y,t) = \cos(8\pi\beta t)\sin(4\sqrt{2}\pi x)\sin(4\sqrt{2}\pi y) \\ V(x,y,t) = \cos(8\pi\beta t)\cos(4\sqrt{2}\pi x)\cos(4\sqrt{2}\pi y) \end{cases} \tag{29}$$

 ISBN: 978-84-614-6167-7

$$\begin{cases} U(x,y,t) = \cos(16\pi\beta t)\sin(8\sqrt{2}\pi x)\sin(8\sqrt{2}\pi y) \\ V(x,y,t) = \cos(8\pi\beta t)\cos(8\sqrt{2}\pi x)\cos(8\sqrt{2}\pi y) \end{cases} \tag{30}$$

Table 1: Global errors for several analytical solutions of the problem Eq.1

Analytical Sol.	Error global U	Error global V
25	1.646×10^{-6}	1.778×10^{-6}
26	1.232×10^{-5}	2.443×10^{-5}
27	4.081×10^{-4}	2.001×10^{-4}
28	9.188×10^{-2}	8.647×10^{-2}
29	3.035×10^{-1}	3.275×10^{-1}
30	4.942×10^{-1}	4.999×10^{-1}

4.2 Dispersion and Irregularity

Figure 1 shows the dispersion of the waves P in each node of the irregular mesh with $IIC = 0.8944$.

4.3 GFDM with PML

Let us solve the Eq. 1, in $\Omega = [0,2] \times [0,1] \subset \mathbf{R}^2$, with homogeneous the Dirichlet boundary conditions and the initial conditions are given by Eq. 22, using the regular mesh with 861 nodes (see Fig. 2), the analytical solutions is given by Eq. 25 (see Fig. 3). The weighting function is given by Eq. 23 and the criterion for the selection of star nodes is the quadrant criterion.
Figure 5 shows the graphic the approximated solution of u_x, after 100 time steps, with PML in x-direction and y-direction for $1.4 \leq x \leq 2$ and $0.6 \leq y \leq 1$ (see Fig. 4).

Figure 7 shows the graphic the approximated solution of u_x, after 100 time steps, with PML in x-direction and y-direction for $\leq x \leq 0.6$ and $0 \leq y \leq 0.2$, $0.8 \leq y \leq 1$ (see Fig. 6).

Figure 9 shows the graphic the approximated solution of u_x, after 100 time steps, with PML in x-direction and y-direction for $\leq x \leq 0.6$ and $0 \leq y \leq 0.2$, $0.8 \leq y \leq 1$ and for a heterogeneous region of interest defined(see Fig. 8), and with the properties

 ISBN: 978-84-614-6167-7

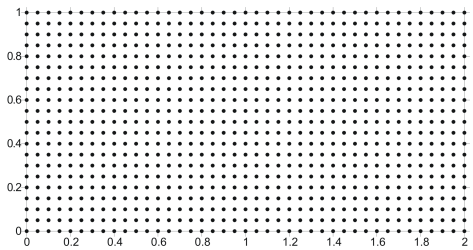


Figure 2 Regular mesh (861 nodes)

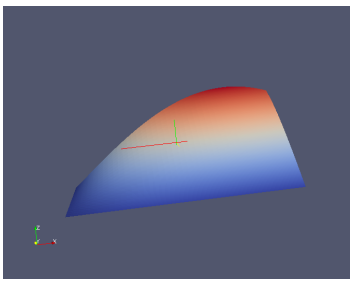


Figure 3 Exact solution U_x without PML.

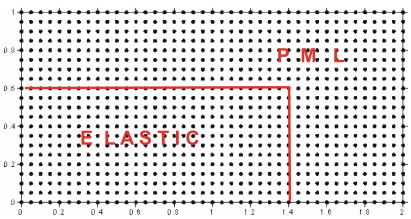


Figure 4: Regular mesh with PML region.

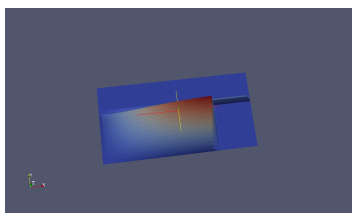


Figure 5: Approximated solution U_x with PML.

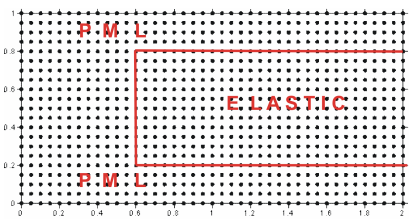


Figure 6: Regular mesh with PML region.

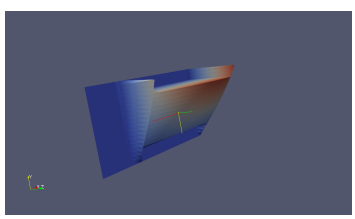


Figure 7: Approximated solution U_x with PML.

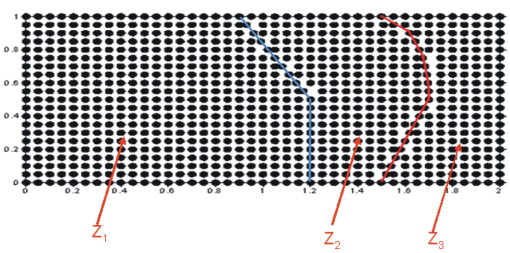


Figure 8: Regular mesh with PML region and heterogeneous dates.

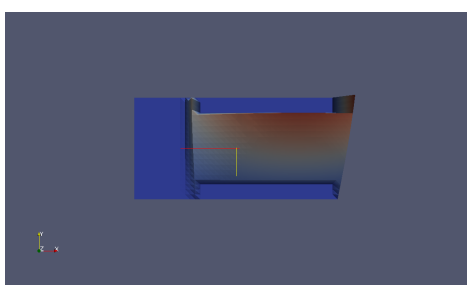


Figure 9: Approximated solution U_x with PML.

 ISBN: 978-84-614-6167-7

Table 2: properties of the region

	λ	μ	ρ	α	β
z_1	0.5	0.25	1.0	1.0	0.5
z_2	0.7	0.4	1.2	1.118	0.577
z_3	0.8	0.5	1.1	1.279	0.674

5 Conclusions

This paper shows a scheme in generalized finite differences, for seismic wave propagation in 2-D for homogeneous and heterogeneous media.
The formulation of the PML is compatible with GFDM and numerical results confirm that PML has an extraordinary performance in absorbing outgoing waves for homogeneous and heterogeneous media.

Acknowledgements

The authors acknowledge the support from Ministerio de Ciencia e Innovación of Spain, project CGL2008 – 01757/CLI.

References

[1] Benito, J.J., Ureña, F., Gavete, L., Alonso, B.: Solving parabolic and hyperbolic equations by Generalized Finite Difference Method. Journal of Computational and Applied Mathematics 209 Issue 2, 208–233 (2007).

[2] Berenger, J.P.: A perfectly matched layer for the absortion of electromagnetic. J. Comput. Physics , 114,185–200 (1994).

[3] Jonshon, S.G.: Notes on Perfectly Matched Layers (PMLs). Courses 18.369 and 18.336 at MIT. July (2008).

[4] Moczo, P.: Introduction to modelling seismic wave propagation by the finite difference method. Lectures Notes, Kyoto, (1998).

[5] Skelton E.A., Adams S.D.M., Craster R.V.: Guided elastic waves and perfectly matched layers. Wave motion, Elsevier, 44,573–592 (2007).

[6] Ureña, F., Benito, J.J., Salete, E., Gavete, L.: A note on the application of the generalized finite difference method to seismic wave propagation in 2-D. Journal of Computational and Applied Mathematics (2011), doi:10.1016/j.cam.2011.04.005

[7] Benito, J.J., Ureña, F., Gavete, L.: Leading-Edge Applied Mathematical Modelling Research (chapter 7). Nova Science Publishers, New York, (2008).

 ISBN: 978-84-614-6167-7

Proceedings of the 11th International Conference on Computational and Mathematical Methods in Science and Engineering, CMMSE 2011 26–30 June 2011.

Solving Differential Riccati Equations on Multi-GPU Platforms

Peter Benner[1], Pablo Ezzatti[2], Hermann Mena[3], Enrique S. Quintana-Ortí[4] and Alfredo Remón[4]

[1] *Max Planck Institute for Dynamics of Complex Technical Systems, Sandtorstr. 1, 39106 Magdeburg, (Germany)*

[2] *Centro de Cálculo-Instituto de Computación, Universidad de la República, 11.300-Montevideo (Uruguay)*

[3] *Departamento de Matemática, Escuela Politécnica Nacional, EC1701 Quito (Ecuador)*

[4] *Dpto. de Ingeniería y Ciencia de Computadores, Universidad Jaime I, 12.071-Castellón (Spain)*

emails: benner@mpi-magdeburg.mpg.de, pezzatti@fing.edu.uy, hermann.mena@epn.edu.ec, quintana@icc.uji.es, remon@icc.uji.es

Abstract

Differential Riccati Equations (DREs) arise in many scientific and engineering applications. Particularly, they play an important role in control problems, where a finite-time horizon of integration is considered. In this paper, we present several high-performance implementations of the Rosenbrock method for multi-core and graphic processors (GPUs). The Rosenbrock method for solving DREs is an iterative technique that requires the solution of a Lyapunov equation per step, which in our approach is solved via the highly parallel sign function method. Mainly, this is an iterative procedure, where the most time-consuming operation is the computation of a matrix inverse per step. Hence, an efficient implementation of the Rosenbrock method can be obtained providing an efficient matrix inversion kernel. We analyze two different approaches for the matrix inversion: the traditional method based on the LU factorization and the Gauss-Jordan elimination method. Numerical experiments show that the execution time can be drastically reduced by off-loading part of the computations to one or more GPUs.

Key words: Differential Riccati equations, Rosenbrock methods, matrix sign function, graphics processors, multi-core processors.

 ISBN: 978-84-614-6167-7

1 Introduction

Consider the symmetric differential Riccati equation (DRE)

$$\begin{array}{rcl} \dot{X}(t) & = & Q + X(t)A + A^T X(t) - X(t)SX(t) \equiv F(X(t)), \\ X(t_0) & = & X_0, \end{array} \tag{1}$$

where $t \in [a, b]$, $A \in \mathbb{R}^{n\times n}$, $Q \in \mathbb{R}^{m\times n}$, $S \in \mathbb{R}^{n\times m}$, and $X(t) \in \mathbb{R}^{m\times n}$. The solution of (1) exists and it is unique, e.g., [1, Thm. 4.1.6]. Symmetric DREs arise in linear-quadratic optimal control problems such as LQR and LQG design with finite-time horizon, in H_∞ control of linear-time varying systems as well as in differential games, e.g., [1, 2]. Unfortunately, in most control problems fast and slow modes occur. Then, the DRE will be fairly stiff, so implicit methods have to be used for solving the DRE numerically. Matrix-valued algorithms based on generalizations of the Rosenbrock methods have been proved to yield accurate solutions for large-scale DREs arising in optimal control problems for parabolic partial differential equations [4, 6]. Using Rosenbrock methods for solving DREs requires the solution of one Lyapunov equation per iterative step. The Lyapunov equation is usually solved by exploiting the structure of the matrices (sparsity, symmetry, low rank), see, e.g., [3]. However, in some applications, a large interval of integration has to be considered and/or a thiner mesh is required to describe the solution accurately, which turns these methods un affordable due to their high computational cost.

We will focus on the method of order one, i.e, the so-called linearly implicit Euler method, on equidistant meshes. For practical purposes, the method should be used with adaptive time steps as suggested in [4]. Here we will focus on the parallel perfomance of the computation of one time step, which is independet of the grid chosen. Thus, we keep things as simple as possible for this purpose. The resulting Lyapunov equation is solved via the highly parallel sign function method, where the most time-consuming operation is the computation of a matrix inversion per step. We analyze two different approaches for the matrix inversion: the traditional method based on the LU factorization and the Gauss-Jordan elimination method.

We present several high-performance implementations of the Rosenbrock method for a hybrid platform composed of multi-core processors and several graphics processors. The use of high-performance kernels of several linear algebra libraries, and several optimization techniques, like padding or the concurrent computation in all the devices of the platform, report a remarkable performance in the developed implementations.

This paper is organized as follows. In Section 2, we briefly describe the application of the Rosenbrock method of order one to DREs and the sign function method for solving Lyapunov equations. The different implementations are described in Section 3. Then, in Section 4, numerical experiments showing the performance of the proposed methods are included. Finally, conclusions and future work are pointed out.

 ISBN: 978-84-614-6167-7

2 Numerical solution of DREs

We focus on the solution of DREs arising in optimal control for ordinary differential equations. Using the Rosenbrock method of order one for solving an autonomous symmetric DRE of the form (1) yields:

$$\tilde{A}_k^T X_{k+1} + X_{k+1}\tilde{A}_k = -Q - X_k S X_k - \frac{1}{h} X_k \tag{2}$$

where $X_k \approx X(t_k)$ and $\tilde{A}_k = A - SX_k - \frac{1}{2h}I$; see [6, 4] for details. In addition, we assume that

$$\begin{array}{rcll} Q & = & C^T C, & C \in \mathbb{R}^{p\times n}, \\ S & = & BB^T, & B \in \mathbb{R}^{n\times m}, \\ X_k & = & Z_k Z_k^T, & Z_k \in \mathbb{R}^{n\times z_k}, \end{array}$$

with p, m, $z_k \ll n$. If we denote $N_k = [\, C^T,\, Z_k(Z_k^T B),\, \sqrt{h^{-1}} Z_k \,]$, then the Lyapunov equation in (2) results in

$$\tilde{A}_k^T X_{k+1} + X_{k+1}\tilde{A}_k = -N_k N_k^T, \tag{3}$$

where $\tilde{A}_k = A - B(Z_k(Z_k^T B))^T - \frac{1}{2h}I$. Observing that $\operatorname{rank}(N_k) \leq p + m + z_k \ll n$, we can efficiently solve (3) with the sign function method as described in the following subsection. This is stated in Algorithm 1

Algorithm 1: Rosenbrock method of order one for DREs

Data: $A \in \mathbb{R}^{n\times n}$, B, C, Z_0 satisfying (2), $t \in [a, b]$, and step size h.
Result: (Z_k, t_k) such that $X_k \approx Z_k Z_k^T$, $Z_k \in \mathbb{R}^{n\times z_i}$ with $z_i \ll n$.

```
1 begin
2     t_0 := a
3     for k := 0 to ⌈(b−a)/h⌉ do
4         N_k := [ C^T, Z_k(Z_k^T B), √(h^{-1}) Z_k ]
5         Compute Z_{k+1} such that the low rank factor product Z_{k+1} Z_{k+1}^T
          approximates the solution of Ã_k^T X_{k+1} + X_{k+1} Ã_k = −N_k N_k^T.
6         t_{k+1} := t_k + h.
7     end
8 end
```

2.1 The sign function method

The matrix sign function is an efficient tool to solve stable Lyapunov equations. There exist several iterative schemes for the computation of this matrix function. Among those, the Newton iteration described in Algorithm 2 is specially appealing for its simplicity, efficiency, parallel performance, and asymptotic quadratic convergence [7].

 ISBN: 978-84-614-6167-7

Algorithm 2: Matrix sign function for Lyapunov equations

Data: $A \in \mathbb{R}^{n\times n}$, $N \in \mathbb{R}^{j\times n}$.
Result: $\tilde{S}$ such that $\tilde{S}^T\tilde{S} \approx X$ and $A^TX + XA = NN^T$.

begin
  $A_0 := A$
  $\hat{S}_0 := N^T$
  $k := 0$
  repeat
    $A_{k+1} := \frac{1}{\sqrt{2}}\left(A_k/c_k + c_k {A_k}^{-1}\right)$
    Compute the rank-revealing QR (RRQR) decomposition
    $\frac{1}{\sqrt{2c_k}}\left[\tilde{S}_k,\ \ c_k\tilde{S}_k(A_k^{-1})^T\right] = Q_s\begin{bmatrix} U_s \\ 0 \end{bmatrix}\Pi_s.$
    $\tilde{S}_{k+1} := U_s\Pi_s$
    $k := k + 1$
  until $\sqrt{\frac{\|A_{k+1}+I\|_\infty}{n}} < \tau\|A_k\|_\infty$
end

Algorithm 2 roughly requires $2n^3$ floating-point arithmetic operations (flops) per iteration.

On convergence, $\tilde{S}$ is the factor of the approximated solution and satisfies $X \approx \tilde{S}^T\tilde{S}$. Convergence can be accelerated using several techniques [7]. In our approach, we employ a scaling defined by the parameter

$$c_k = \sqrt{\|{A_k}^{-1}\|_\infty / \|A_k\|_\infty}.$$

In the convergence test, τ is a tolerance threshold for the iteration that is usually set as a function of the problem dimension and the machine precision ϵ.

3 High-performance implementations

In this section we describe several high-performance codes for the solution of DREs. All the implementations use linear algebra libraries (e.g., MKL or CUBLAS) and employing single precision arithmetic. A double precision accurate solution can be obtained, at a low-computational cost, applying an iterative refinement technique like the one proposed in Benner et al. [8].

The solution of the Lyapunov equation is the most expensive part when solving DREs using the Rosenbrock method (see Algorithm 1). Particularly, most of the computational cost is due to the matrix inversion required at each iteration of the sign function method (Algorithm 2). Thus, we have optimized the matrix inversion process. The rest of operations are mainly matrix-matrix products of small matrices, which can be efficiently computed on the multi-core architecture invoking a multi-thread implementation of BLAS.

 ISBN: 978-84-614-6167-7

We present three different algorithms/implementations of the Rosenbrock method which, basically differ in the procedure to compute the matrix inverse. All the remaining steps in these algorithms/implementations are performed on the CPU and therefore, we do not go into detail.

3.1 Implementation on a multi-core CPU: `Ros(CPU)`

The traditional approach to compute the inverse of a matrix $A \in \mathbb{R}^{n \times n}$ is based on Gaussian elimination (i.e., the LU factorization), and consists of the following three steps:

1. Compute the LU factorization $PA = LU$, where $P \in \mathbb{R}^{n \times n}$ is a permutation matrix, and $L, U \in \mathbb{R}^{n \times n}$ are unit lower and upper triangular factors, respectively, see [9].
2. Invert the triangular factor $U \rightarrow U^{-1}$.
3. Solve the system $XL = U^{-1}$ using backward substitution for X.
4. Undo the permutations $A^{-1} := XP$.

LAPACK [10] is a high-performance linear algebra library, which provides efficient routines to compute the previous steps. In particular, routine `getrf` yields the LU factorization (with partial pivoting) of a nonsingular matrix (Step 1), while routine `getri` computes the inverse matrix of A using the LU factorization obtained by `getrf` (Steps 2–4). The computational cost of computing a matrix inverse following the previous four steps is $2n^3$ flops.

3.2 Implementation on a many-core GPU: `Ros(GPU)`

The traditional algorithm for matrix inversion (see Section 3.1) shows some limitations from the high-performance computing point of view. There is no possibility to compute concurrently several steps, so parallelism has to be extracted within each step. Also, step 2 and 3 are unbalanced, because they operate on triangular matrices. The Gauss-Jordan elimination algorithm (GJE) is a reordering of the computations performed by the Gaussian elimination procedure for matrix inversion. Thus, the arithmetic cost of matrix inversion using GJE is the same as the one based on the LU factorization. However, the GJE method is better suited for parallelization. We present an implementation of the GJE method on a GPU.

Algorithm 4 illustrates a blocked version of the GJE for matrix inversion. A description of the unblocked version (`GEINGJ`), called from inside the blocked one, can be found in [11]; for simplicity, the application of pivoting during the factorization is concealed; see [11]. The bulk of computations is cast in terms of matrix-matrix products; an operation which exhibits a high degree of concurrency. Therefore, GJE is a highly appealing method for matrix inversion on emerging architectures like GPUs, where many computational units are available and a highly tuned implementation of the matrix-matrix product is available (e.g., in the nVidia CUBLAS library).

 ISBN: 978-84-614-6167-7

Algorithm 3: Blocked Gauss-Jordan elimination algorithm for matrix inversion

Data: $A \in \mathbb{R}^{n \times n}$.
Result: $A := A^{-1}$.

begin
  $t_0 := a$
  for $k := 1$ *to* $\lceil \frac{n}{b} \rceil$ **do**
    $$A \rightarrow \left(\begin{array}{c|c|c} A_{00} & A_{01} & A_{02} \\ \hline A_{10} & A_{11} & A_{12} \\ \hline A_{20} & A_{21} & A_{22} \end{array}\right) \quad \text{where } A_{00} \in \mathbb{R}^{(k-1)b \times (k-1)b},\ A_{11} \in \mathbb{R}^{b \times b}$$
    $[A_{01},\ A_{11},\ A_{21}]^T :=$ `GEINGJ`$([A_{01},\ A_{11},\ A_{21}]^T)$
    $A_{00} := A_{00} + A_{01}A_{10}$
    $A_{20} := A_{20} + A_{21}A_{10}$
    $A_{10} := A_{11}A_{10}$
    $A_{02} := A_{02} + A_{01}A_{12}$
    $A_{22} := A_{22} + A_{21}A_{12}$
    $A_{12} := A_{11}A_{12}$
  end
end

3.3 Implementation on multi-GPU: `Ros(MGPU)`

`Ros(MGPU)` is in essence an extension of the `Ros(GPU)` solver which targets platforms equipped with multiple GPUs. As mentioned before, the critical operation of the Rosenbrock method is the matrix inversion. In this variant we employ a highly tuned implementation for this operation that targets a platform with several GPUs connected to a single CPU. This matrix inversion routine includes several optimization techniques, in particular, the use of optimized CUBLAS kernels, padding to accelerate the GPU-memory access, an optimized task schedule that permits the concurrent computation in all the devices, a look-ahead approach to minimize the negative impact from the critical path, a cyclic distribution that maximizes load balance, and the use of two block-sizes that allows to adapt the routine execution to the particularities of the CPU and the GPU architectures simultaneously. See [12] for more details on the matrix inversion routine.

4 Numerical Results

In this section we evaluate the performance of the implementations introduced in Section 3 on a hybrid platform composed of a multi-core CPU and several many-core GPUs. Particularly, the experiments are performed on a computer with two Intel Xeon QuadCore E5530 processors at 2.27GHz, connected to an nVidia Tesla C1060 (consisting of four nVidia Tesla S1060 GPUs) via a PCI-e bus (more details about the platform can be found in Table 4).

 ISBN: 978-84-614-6167-7

Algorithm 4: Blocked Gauss-Jordan elimination algorithm for matrix inversion

Require: $A \in \mathbb{R}^{n \times n}$

1: $t_0 := a$
2: **for** $k := 1$ to $\lceil \frac{n}{b} \rceil$ **do**
3: Partition $A \rightarrow \left(\begin{array}{c|c|c} A_{00} & A_{01} & A_{02} \\ \hline A_{10} & A_{11} & A_{12} \\ \hline A_{20} & A_{21} & A_{22} \end{array}\right)$ where $A_{00} \in \mathbb{R}^{(k-1)b \times (k-1)b}$, $A_{11} \in \mathbb{R}^{b \times b}$
4: $[A_{01},\ A_{11},\ A_{21}]^T := \texttt{GEINGJ}([A_{01},\ A_{11},\ A_{21}]^T)$
5: $A_{00} := A_{00} + A_{01}A_{10}$
6: $A_{20} := A_{20} + A_{21}A_{10}$
7: $A_{10} := A_{11}A_{10}$
8: $A_{02} := A_{02} + A_{01}A_{12}$
9: $A_{22} := A_{22} + A_{21}A_{12}$
10: $A_{12} := A_{11}A_{12}$
11: **end for**

Processors	#proc.	#cores (per proc.)	Frequency (GHz)	L2 cache (MB)	Memory (GB)
Intel QuadCore E5530	2	4	2.27	8	48
nVidia TESLA c1060	4	240	1.3	–	(4x4)16

Table 1: Hardware employed in the experiments.

A multi-thread version of the Intel MKL library (version 10.2) provided the necessary LAPACK and BLAS kernels for the CPU, and nVidia CUBLAS (version 2.1) for the GPU computations.

4.1 Test examples

We evaluate the performance of the implementations using two problems from the *Oberwolfach Model Reduction Benchmark Collection*[1]: the semi-discretized heat transfer problem for the optimal cooling of steel profiles (steel), and the butterfly gyro problem (gyro). In the following we briefly describe these two models.

4.1.1 STEEL

This model arises in a manufacturing method for steel profiles. The goal is to design a control that yields moderate temperature gradients when the rail is cooled down. The mathematical model corresponds to the boundary control for a 2-D heat equation. A finite element discretization, followed by adaptive refinement of the mesh, results

[1] `http://www.imtek.de/simulation/benchmark/`.

 ISBN: 978-84-614-6167-7

in several instances of this benchmark. We employed two instances of this problem, STEEL$_S$ and STEEL$_L$. For both DREs, $m = 7$ and $p = 6$. The order of the system (size of A) is n=5,177 for the STEEL$_S$ instance and 20,209 for the STEEL$_L$ instance.

4.1.2 GYRO

The Butterfly is a vibrating micro-mechanical gyro that has been proposed for inertial navigation applications. The model is a simplified version which includes the pure structural mechanics problem only. It is designed to test model reduction approaches. The dimension of the mechanical system described by a system of second-order differential equations is $n = 17,361$. Hence, for the numerical experiments we first need to transform the system into a first order one, the main dimension of the transformed system is thus $n = 34,722$.

4.2 Numerical Results

All the experiments were done using single-precision arithmetic, and all the reported execution times include the overhead introduced by data transfers between the CPU and the GPUs memory spaces.

We first study the STEEL case, because our three implementations can tackle this problem using the available hardware (the large dimension of the matrices involved in GYRO did not allow us to solve this problem using `Ros(GPU)`).

Tables 2 and 3 present the total time as well as the time spend in the sign function Lyapunov solver (F_{sign}) required to solve the DRE associated with both instances of STEEL on the $[0, 1]$ interval with a stepsize of $h = 0.1$.

`Ros(CPU)` Time		`Ros(GPU)` Time		Speed-up		`Ros(MGPU)` Time		Speed-up	
F_{sign}	Total	F_{sign}	Total	F_{sign}	Total	F_{sign}	Total	F_{sign}	Total
92.06	94.65	47.55	50.13	1.94	1.89	26.44	28.99	3.48	3.27

Table 2: Execution time (in seconds) and speed-up obtained for the STEEL$_S$ benchmark.

`Ros(CPU)` Time		`Ros(GPU)` Time		Speed-up		`Ros(MGPU)` Time		Speed-up	
F_{sign}	Total	F_{sign}	Total	F_{sign}	Total	F_{sign}	Total	F_{sign}	Total
8703.2	9061.6	3406.5	3712.7	2.56	2.44	1338.2	1688.5	6.50	5.37

Table 3: Execution time (in seconds) and speed-up obtained for the STEEL$_L$ benchmark.

The experimental results in the tables show that most of the time (approximately 97%) is dedicated to the computation of the sign function method. They also shown how this computation can be drastically accelerated using graphics processors. The

 ISBN: 978-84-614-6167-7

hybrid CPU-GPU implementation accelerates the computation time of F_{sign} between 1.94× and 2.56×. The use of the four GPUs available on the platform enhances this acceleration factor to 3.48× for the STEEL$_S$ problem and 6.5× for the STEEL$_L$ case.

In a second experiment we evaluate the performance of the CPU and the multi-GPU implementations (`Ros(CPU)` and `Ros(MGPU)`) using the GYRO example. Table 4 summarizes the results obtained for this benchmark. Once more, most of the time is spent in the computation of F_{sign} and important time reductions are obtained from the use of the four GPUs. In this problem routine `Ros(MGPU)` computes F_{sign} approximately 9× faster than the `Ros(CPU)` implementation.

`Ros(CPU)` Time		`Ros(MGPU)` Time		Speed-up	
F_{sign}	Total	F_{sign}	Total	F_{sign}	Total
44919.69	46136.36	4986.36	6141.64	9.01	7.51

Table 4: Execution time (in seconds) and speed-up obtained for the GYRO benchmark.

From the results we can conclude that the hybrid CPU-GPU implementation reports an important reduction of the computational time, but the amount of memory limits its application to small and medium problems. The multi-GPU permits to target larger problems while simultaneously providing a notorious performance, e.g., accelerating the execution of the GYRO problem 7.51×.

Finally, we remark that in the `Ros(CPU)` implementation based on the LAPACK routines for the matrix inversion, approximately the 97% from the execution time is dedicated to the computation of the sign function. Despite the effort to optimize this operation in the GPU-based implementations (`Ros(GPU)` and `Ros(MGPU)`), the sign function still concentrates a 80% of the total time for the fastest implementation and the largest problem studied in this work.

5 Conclusions and future work

The numerical results show the dramatic acceleration when using GPUs for solving DREs. The single GPU-based implementation reports a speed-up of 2.5× over the multi-core CPU implementation based on LAPACK. The multi-GPU implementation, executed on four GPUs, increases this ratio up to 7.5× Another remarkable advantage from the use of several GPUs is the increment of the aggregated memory. As each device includes its own memory, an increase in the number of devices reports an increment in the amount of available memory and, hence, also the dimension of the affordable problems. This is specially important in many optimal control problems, where the dimension of the related mathematical models is extremely large.

The use of four GPUs reduces the percentage of time dedicated to compute matrix inverses from 97% to 80%. The use of more GPUs will probably keep decreasing this percentage. In general, GPUs show an excellent relationship between cost and com-

 ISBN: 978-84-614-6167-7

puting power, achieving more FLOPS per dollar than the traditional high performance architectures in many application areas, e.g., in the execution of dense linear algebra operations.

The acceleration of other stages involved in Algorithm 1, the implementation of higher order methods and a stepsize control will be discussed in future works.

Acknowledgments

Enrique S. Quintana-Ortí and Alfredo Remón were supported by projects PROMETEO 2009/013 and CICYT TIN2008-06570-C04. This work was partially done while Hermann Mena was visiting the Universidad Jaime I with the support from the program "Pla de suport a la investigació 2009" from Universidad Jaime I, and continued while Pablo Ezzatti, Hermann Mena, and Alfredo Remón were visiting the Max Plank Institute (MPI) in Magdeburg. Pablo Ezzatti, Hermann Mena, and Alfredo Remón gratefully acknowledge support received from the MPI.

References

[1] H. Abou-Kandil, G. Freiling, V. Ionescu and G. Jank, *Matrix Riccati Equations in Control and Systems Theory*, Birkhäuser, Basel, Switzerland, 2003.

[2] A. Ichikawa and H. Katayama, *Remarks on the time-varying H_∞ Riccati equations*, Sys. Cont. Lett. 37(5):335-345, 1999.

[3] P. Benner, J-R. Li and T. Penzl, *Numerical Solution of Large Lyapunov Equations, Riccati Equations, and Linear-Quadratic Control Problems*, Numerical Linear Algebra with Applications, Vol. 15, No. 9, pp. 755-777, 2008.

[4] H. Mena, *Numerical Solution of Differential Riccati Equations Arising in Optimal Control Problems for Parabolic Partial Differential Equations*, PhD thesis, Escuela Politécnica Nacional, 2007.

[5] P. Benner and H. Mena, *Rosenbrock methods for solving differential Riccati equations*, Max Planck Institute Magdeburg, Preprints, 2011, in preparation.

[6] P. Benner and H. Mena, *Numerical solution of large scale differential Riccati Equations arising in optimal control problems*, Max Planck Institute Magdeburg, Preprints, 2011, in preparation.

[7] N.J. Higham, *Functions of Matrices: Theory and Computation*, SIAM, Philadelphia, USA, 2008.

[8] P. Benner, P. Ezzatti, D. Kressner, E. S. Quintana-Ortí and A. Remón, *A mixed-precision algorithm for the solution of Lyapunov equations on hybrid CPU-GPU platforms*, Parallel Computing, to appear.

 ISBN: 978-84-614-6167-7

[9] G. Golub and C. V. Loan, *Matrix Computations, 3rd Edition*, The Johns Hopkins University Press, Baltimore, 1996.

[10] E. Anderson, Z. Bai, J. Demmel, J. E. Dongarra, J. DuCroz, A. Greenbaum, S. Hammarling, A. E. McKenney, S. Ostrouchov and D. Sorensen, *LAPACK Users' Guide*, SIAM 1992

[11] E.S. Quintana-Ortí, G. Quintana-Ortí, X. Sun and R.A. van de Geijn, *A note on parallel matrix inversion*, SIAM J. Sci. Comput., vol. 22, pp. 1762-1771, 2001.

[12] P. Ezzatti, E. S. Quintana-Ortí and A. Remón, *High Performance Matrix Inversion on a Multi-core Platform with Several GPUs*, Proceedings of the 19th Euromicro International Conference on Parallel, Distributed and Network based Processing, pp. 87-93, 2011.

 ISBN: 978-84-614-6167-7

Proceedings of the 11th International Conference on Computational and Mathematical Methods in Science and Engineering, CMMSE 2011 26–30 June 2011.

The Galerkin method for a generalized Lax–Milgram theorem

M. I. Berenguer[1] and M. Ruiz Galán[1]

[1] *Department of Applied Mathematics, University of Granada (Spain)*

emails: maribel@ugr.es, mruizg@ugr.es

Abstract

A recent version of the Lax–Milgram theorem allows us to show that a wide class of linear elliptic boundary value problems, whose data belong to reflexive Banach spaces that are not necessarily Hilbert, admits a weak or variational solution. In addition we provide the associated Galerkin scheme. The generation of the corresponding finite–dimensional subspaces for concrete examples of boundary value problems is developed by means of certain biorthogonal systems in the reflexive Banach spaces in question. In particular, we obtain a numerical solution for some of these problems for which the classical results of Hilbertian nature do not apply.

Key words: Lax–Milgram theorem, variational formulations, Galerkin methods, biorthogonal systems, elliptic boundary value problems.

MSC 2000: 65N30, 58E30, 49J40.

1 Introduction

The celebrated Lax–Milgram theorem [7] is a fundamental tool in the solvability theory for linear elliptic partial differential equations, as well as for their numerical solution. Recently, in [8, Theorem 1.2] an extension of this result has been stated in the setting of real locally convex spaces (our statements are equally valid in the complex case). The aim of this work is to show how this result, including its particular case for real reflexive Banach spaces, extends the class of linear elliptic boundary value problems that admit a variational formulation with a solution and in addition, is able of be solved numerically through the associated Galerkin method.

2 The discrete problem and numerical testing

In order to recall the mentioned version of the Lax–Milgram theorem, we present some standard notation: given real linear spaces E and F, a bilinear form $a : E \times F \longrightarrow \mathbb{R}$ and $x_0 \in E$, $y_0 \in F$, $a(\cdot, y_0)$ denotes the linear functional on E

$$x \in E \longmapsto a(x, y_0) \in \mathbb{R},$$

 ISBN: 978-84-614-6167-7

whereas $a(x_0,\cdot)$ stands for the analogous linear functional on F. Besides, given a real normed space E, we write E^* for its topological dual space. Finally, $()_+$ is the positive part, i.e., for $t\in\mathbb{R}$, $(t)_+=\max\{t,0\}$.

Proposition 1 ([8]). *Assume that E is a real reflexive Banach space and that F is a real normed space, $y_0\in F^*$, $a:E\times F\longrightarrow\mathbb{R}$ is bilinear and that C is a nonempty convex subset of F such that for all $y\in C$, $a(\cdot,y)\in E^*$. Then*

$$\text{there exits } x_0\in E \text{ such that for all } y\in C,\quad y_0^*(y)\le a(x_0,y)$$

if, and only if,

$$\text{there exists } \alpha>0 \text{ such that for all } y\in C,\quad y_0^*(y)\le \alpha\|a(\cdot,y)\|.$$

In addition, if one of these equivalent statements is satisfied and for some $y\in C$ we have that $a(\cdot,y)\neq 0$, then

$$\min\{\|x_0\| : x_0\in E \text{ and for all } y\in C,\ y_0^*(y)\le a(x_0,y)\}=\left(\sup_{y\in C,\ a(\cdot,y)\neq 0}\frac{y_0^*(y)}{\|a(\cdot,y)\|}\right)_+.$$

As a consequence of the uniform boundedness theorem we arrive at the corresponding global version of Proposition 1 (the functional $y_0^*\in F^*$ is arbitrary), which clearly generalizes the classical Lax–Milgram theorem:

Theorem 2. *If E is a real reflexive Banach space, F a real normed space, $a:E\times F\longrightarrow\mathbb{R}$ a bilinear form and C a nonempty convex subset of F such that for all $y\in C$ we have that $a(\cdot,y)\in E^*$, then*

$$\text{for all } y_0^*\in F^* \text{ there exists } x_0\in X \text{ such that } y\in C\Rightarrow y_0^*(y)\le a(x_0,y) \tag{1}$$

if, and only if,

$$\text{existe } \alpha>0 \text{ such that } y\in C\Rightarrow \alpha\|y\|\le\|a(\cdot,y)\|. \tag{2}$$

Furthermore, if one of these equivalent assertions holds and $C\neq\{0\}$, then, for all $y_0^*\in F^*$,

$$\min\{\|x_0\| : x_0\in E \text{ and for all } y\in C,\ y_0^*(y)\le a(x_0,y)\}=\left(\sup_{y\in C,\ y\neq 0}\frac{y_0^*(y)}{\|a(\cdot,y)\|}\right)_+. \tag{3}$$

Let us note that when $C=Y$, inequality (1) becomes an equality and the stability condition (3) is more natural:

$$\min\{\|x_0\| : x_0\in E \text{ and } y_0^*=a(x_0,\cdot)\}\le\frac{\|y_0^*\|}{\alpha}.$$

 ISBN: 978-84-614-6167-7

Uniqueness in the variational inequality (1) is nothing more than that of the corresponding homogeneous problem if C is balanced: if $C = -C$ and the variational equation

$$\text{find } x_0 \in X \text{ such that } y \in C \Rightarrow y_0^*(y) = a(x_0, y)$$

has a solution, then it is unique if, and only if,

$$x \in E \text{ and for all } y \in C,\ a(x, y) = 0 \Rightarrow x = 0.$$

Now we show a simple but illustrative elliptic boundary value problem that does not admit a classical variational formulation in terms of a continuous and coercive bilinear form and a linear functional on a Hilbert space, but it can be weakly solved by means of the Lax–Milgram theorem above:

Example 3. Let $1 < p < \infty$, let $f \in L^p(0,1)$ and consider the Poisson's problem with homogeneous Dirichlet boundary conditions

$$\left\{ \begin{array}{rcl} -x'' = & f & \text{in } (0,1) \\ x(0) = x(1) = & 0 & \end{array} \right. .$$

For $p = 2$, and only for $p = 2$, the classic Lax–Milgram theorem guarantees that it admits a unique weak solution. Otherwise we can apply Theorem 2 (or Proposition 1). More specifically, if q is the exponent conjugate to p, and we multiply by a test function $y \in W_0^{1,q}(0,1)$ the equation $-x'' = f$ in $(0,1)$ and integrate by parts, then we arrive at the following variational formulation of the boundary value problem above:

$$\text{find } x_0 \in W_0^{1,p}(0,1) \text{ such that } y \in W_0^{1,q}(0,1) \Rightarrow \int_0^1 x_0' y' = \int_0^1 fy, \tag{4}$$

that clearly generalizes the usual one with $f \in L^2(0,1)$. Let us consider the reflexive Banach spaces $E := W_0^{1,p}(0,1)$, $F := W_0^{1,q}(0,1)$, the continuous linear functional $y_0^* : F \longrightarrow \mathbb{R}$ given by

$$y_0^*(y) := \int_0^1 fy, \qquad (y \in F)$$

and the continuous bilinear form $a : E \times F \longrightarrow \mathbb{R}$ defined as

$$a(x, y) := \int_0^1 x'y', \qquad (x \in E,\ y \in F).$$

It is not difficult to prove that a satisfies the inf–sup condition (2), making use of the Poincaré's inequality [1, Theorem 6.30]. Therefore Theorem 2 guarantees the existence of a solution of variational problem (4) and its uniqueness turns out to be obvious thanks to the comment given after Theorem 2. □

Now we deal with the Galerkin scheme related to the variational problem (1), as well as with its stability. In particular we have proven the following result that includes, a an special case, a generalization of Céa's inequality:

 ISBN: 978-84-614-6167-7

Proposition 4. *Let E be a real reflexive Banach space, let F be a real normed space, let $y_0^* \in F^*$ and let $a : E \times F \longrightarrow \mathbb{R}$ be a continuous bilinear form such that and*

$$\text{there exists a unique } x_0 \in E \text{ such that } a(x_0, \cdot) = y_0^*.$$

Suppose in addition that for all $n \geq 1$, E_n and F_n are finite–dimensional subspaces of E and F, respectively, and that

$$\text{there exists a unique } x_n \in E_n \text{ such that } a(x_n, \cdot)_{|F_n} = y_{0|F_n}^*.$$

If moreover for each $n \geq 1$ there exists $\mu_n > 0$ such that

$$y \in F_n \Rightarrow \mu_n \|y\| \leq \|a(\cdot, y)_{|E_n}\|$$

and

$$\mu := \inf_{n \geq 1} \mu_n > 0,$$

then, for all $n \geq 1$ we have that

$$\|x_0 - x_n\| \leq \left(1 + \frac{\|a\|}{\mu}\right) \operatorname{dist}(x_0, E_n).$$

The adequate use of biorthogonal systems (see [6]) in certain real Sobolev reflexive spaces allows us to generate finite–dimensional subspaces that satisfy the mentioned above stability condition which, together with a immediate aproximation property of such systems, imply the convergence of the Galerkin method. Let us emphasize that the biorthogonal systems also become a fundamental tool in the develop of a wide family of numerical methods for solving differential, integral and integro–differential problems ([2, 3, 5]).

Example 5. Let $\{f_n\}$ be the Schauder basis in $W_0^{1,p}(0,1)$ and $W_0^{1,q}(0,1)$ given in [4, Proposition 4.8.]. Let $\{E_n\}$ and $\{F_n\}$ be sequences of finite–dimensional subspaces of E and F where $E_n = F_n := \text{Span}\{f_1, ..., f_n\}$. The Galerkin scheme associated with Example 3 is to find $x_n \in E_n$ such that for all $y_n \in F_n$, $a(x_n, y_n) = y_0^*(y_n)$.

The next table illustrates the statements above for $p = \frac{3}{2}$ and $f(t) = t^{-\frac{3}{5}}$. We exhibit $x_m(t)$ for certain t of $[0,1]$ and x_m has been chosen in such a way that, letting

$$\max\left\{\left|1 - \frac{x_m\left(\frac{1}{j}\right)}{x_{m+1}\left(\frac{1}{j}\right)}\right| : j = 1, 2, \ldots, 9\right\} < 10^{-2}$$

	$t = 0.1$	$t = 0.2$	$t = 0.3$	$t = 0.4$	$t = 0.5$	$t = 0.6$	$t = 0.7$	$t = 0.8$	$t = 0.9$
$x_{32}(t)$	0.1071	0.16922	0.20450	0.21904	0.21619	0.1979	0.16605	0.12185	0.066240

Acknowledgements

Research partially supported by Junta de Andalucía Grant FQM 359.

 ISBN: 978-84-614-6167-7

References

[1] R.A. Adams and J.J.F. Fournier, *Sobolev spaces*, 2nd Edition, Elsevier, 2003.

[2] M.I. Berenguer, D. Gámez, A.I. Garralda–Guillem, M. Ruiz Galán and M.C. Serrano Pérez, *Biorthogonal systems for solving Volterra integral equation systems of the second kind*, J. Comput. Appl. Math. **235** (2011), 1875–1883.

[3] M.I. Berenguer, A.I. Garralda–Guillem and M. Ruiz Galán, *Biorthogonal systems approximating the solution of the nonlinear Volterra integro–differential equation*, Fixed Point Theory Appl., Volume **2010**, Article ID 470149, (2010), 9 pp.

[4] S. Fučik, *Fredholm alternative for nonlinear operators in Banach spaces and its applications to differential and integral equations*, Časopis Pěst. Mat. **96** (1971), 371–390.

[5] D. Gámez, A.I. Garralda Guillem and M. Ruiz Galán, *High order nonlinear initial–value problems countably determined*, J. Comput. Appl. Math. **228** (2009), 77–82.

[6] P. Hájek et al., *Biorthogonal systems in Banach spaces*, CMS Books in Mathematics, Springer, 2008.

[7] P.D. Lax and A.N. Milgram, *Parabolic Equations, Contributions to the Theory of the Partial Differential Equations*, Annals of Math. Studies, (1954) **33** 167–190, Princeton Univ. Press, Princeton, N.Y.

[8] M. Ruiz Galán, *A version of the Lax–Milgram Theorem for locally convex spaces*, J. Convex Anal. **16** (2009), 993–1002.

 ISBN: 978-84-614-6167-7

Proceedings of the 11th International Conference on Computational and Mathematical Methods in Science and Engineering, CMMSE 2011 26–30 June 2011.

A perturbation solution of Michaelis-Menten kinetics in a total quasi-steady-state framework

Alberto Maria Bersani[1] **and Guido Dell'Acqua**[1]

[1] *Dipartimento di Scienze di Base e Applicate per l'Ingegneria (S.B.A.I.), "Sapienza" University, Via A. Scarpa 16, 00161 - Rome, Italy*

emails: bersani@dmmm.uniroma1.it, dellacqua@dmmm.uniroma1.it

Abstract

In this paper we expand the equations governing Michaelis– Menten kinetics in a total quasi-steady-state setting, finding the first order uniform expansions. Our results improve previous approximations and work well especially in presence of an enzyme excess.

Key words: Michaelis-Menten kinetics, quasi-steady state approximations, asymptotic expansions
MSC 2000: 41A58, 41A60, 92C45

1 Introduction

Since the pioneering papers by Bodenstein and Underhill and Chapman in 1913 [1, 2] the quasi-steady state approximation (QSSA) has represented a very important tool in the mathematical modeling of biochemical reactions. It brings to a simplification of the model and allows the qualitative analysis of the reaction, in terms of time scales separation, asymptotic behavior etc., which any numerical analysis could not, in general, capture.
In enzyme kinetics, the standard QSSA (sQSSA), or Michaelis-Menten-Briggs-Haldane approximation [3, 4] was introduced in order to describe the phase after the short transient, where the catalytic enzyme and the substrate rapidly form complexes at high concentrations.
From the Sixties of the last century, mathematicians have interpreted the QSSA in terms of leading order in asymptotic expansions with respect to an appropriate parameter ε, which must be supposed sufficiently small. Heineken, Tsuchiya and Aris [5] use $\varepsilon = \frac{E_T}{S_T}$, where E_T and S_T are respectively the total catalyst concentrations and the substrate concentrations; Segel and Slemrod [6] use $\varepsilon = \frac{E_T}{S_T+K_M}$, where K_M is the *Michaelis constant* or *affinity constant*, showing that the sQSSA is valid in a more

 ISBN: 978-84-614-6167-7

extended parameter range than the one supposed by biochemists.
The technique of singular perturbation allows us to mathematically reproduce the boundary layer in the temporal evolution of the complex concentrations and the separation between the two characteristic time scales, related to the rapid complex formation and to the substrate depletion.
Laidler in 1955 [7] and Borghans, deBoer and Segel in 1996 [8] have approached enzyme kinetics from a different point of view, which is known as total QSSA (tQSSA) and is valid in a wider range of parameters (see also [9]). The tQSSA has been applied to several enzyme reactions [10, 11, 12, 13, 14, 15] even in a stochastic framework [16].
In the papers [8, 9, 14] the tQSSA has been approached requiring some conditions which simplify equations, without any formal tool in terms of asymptotic expansions.
In 2002 Schnell and Maini [17] studied the tQSSA by means of aggregation or lumping techniques, which reduce the number of differential equations describing the system [18]. They nondimensionalized the system of differential equations governing the reaction and introduced the perturbation parameter $\varepsilon = \frac{KE_T}{(K_M+E_T+S_T)^2}$, where K is the van Slyke-Cullen constant. In 2008 Dingee and Anton [19] developed a two-parameter singular perturbation analysis which curiously, at the leading order, does not reproduce the approximated solutions given by Laidler and Borghans, deBoer and Segel, but the zero-order approximation of the tQSSA, obtained by Tzafriri [9] with respect to the parameter

$$\varepsilon = \frac{K}{2S_T}\left(\frac{E_T + K_M + S_T}{\sqrt{(E_T + K_M + S_T)^2 - 4E_T\,S_T}} - 1\right)$$

In this paper we find the tQSSA as the leading order of an asymptotic expansion, obtained with respect to the parameter $\varepsilon = \frac{KE_T}{(K_M+E_T+S_T)^2}$, we find the first order corrections of the inner and the outer solutions (reproducing, respectively, the transient and the QSSA phase) and finally the uniform approximations.

2 Model equations and nondimensionalization

Let us consider the classical system of (dimensional) equations describing the Michelis–Menten kinetics:

$$\begin{aligned} \frac{dS}{dt} &= -k_1(E_T - C)S + k_{-1}\,C, \\ \frac{dC}{dt} &= k_1(E_T - C)S - (k_{-1} + k_2)C, \end{aligned} \tag{1}$$

with initial conditions

$$S(0) = S_T, \quad C(0) = 0, \tag{2}$$

and conservation laws

$$E + C = E_T, \quad S + C + P = S_T\ . \tag{3}$$

 ISBN: 978-84-614-6167-7

Introducing the total substrate $\overline{S} = S + C$, we obtain

$$\begin{aligned} \frac{d\overline{S}}{dt} &= -k_2\, C, \\ \frac{dC}{dt} &= k_1 \left[C^2 - (E_T + \overline{S} + K_M)\, C + E_T\, \overline{S}\right] \end{aligned} \tag{4}$$

with initial conditions

$$\overline{S}(0) = S_T, \quad C(0) = 0, \tag{5}$$

and conservation laws

$$E + C = E_T, \quad \overline{S} + P = S_T\ . \tag{6}$$

If we adopt the change of variables

$$\overline{S} = \alpha\, \overline{s}\,, \quad C = \beta\, c\,, \quad t = \gamma\, \tau$$

we find that eq(s). (4) become

$$\begin{aligned} &\frac{\alpha}{\gamma}\frac{d\overline{s}}{d\tau} = -k_2 \beta\, c \\ &\frac{\beta}{\gamma}\frac{dc}{d\tau} = k_1 \left[\beta^2\, c^2 - (E_T + K_M + \alpha\, \overline{s})\beta\, c + E_T\, \alpha \overline{s}\right] \end{aligned} \tag{7}$$

We should first scale the inner variables, since they are supplemented by the initial conditions (5) that give us, when they are nonzero, information about the magnitude of the variables involved. Thus it follows immediately that $\alpha = S_T$. Therefore the second equation of (7) becomes

$$\frac{\beta}{\gamma}\frac{dc}{d\tau} = k_1 \left[\beta^2\, c^2 - (E_T + K_M + S_T\, \overline{s})\beta\, c + E_T\, S_T \overline{s}\right] \tag{8}$$

Now we attempt to ensure that all the terms on the right hand side of (8) are of the same magnitude, supposing that both c and $\overline{s}$ are $O(1)$. Proceeding as in [6] and [8], neglecting the term in c^2 and then setting for scaling purposes $\overline{s} = 1$ and $c = 1$ we find

$$(E_T + K_M + S_T)\, \beta = E_T\, S_T\,, \quad \text{i.e.,} \quad \beta = \frac{E_T S_T}{E_T + K_M + S_T} \tag{9}$$

while γ is determined by requiring that the left side of (8) is of same magnitude of the right side, i.e.,

$$\gamma = \frac{\beta}{k_1 E_T S_T} = \frac{1}{k_1 (E_T + K_M + S_T)} \tag{10}$$

The parameter γ corresponds to the time scale t_c of the complex formation [8, 19].

 ISBN: 978-84-614-6167-7

3 Asymptotic expansions

Substituting in (7) we have the inner equations:

$$\begin{aligned} \frac{d\overline{s}}{d\tau} &= -\varepsilon\, c \\ \frac{dc}{d\tau} &= \sigma\,\eta c^2 - (\eta + \chi_M)\, c - \sigma\,\overline{s}\, c + \overline{s} \end{aligned} \tag{11}$$

with initial conditions $\overline{s}(0) = 1$ and $c(0) = 0$, where

$$\sigma = \frac{S_T}{K_M + E_T + S_T}\,, \quad \eta = \frac{E_T}{K_M + E_T + S_T}\,, \quad \chi_M = \frac{K_M}{K_M + E_T + S_T}$$

and

$$\varepsilon = \frac{K\, E_T}{(K_M + E_T + S_T)^2} \tag{12}$$

where $K = \frac{k_2}{k_1}$ is the Van Slyke–Cullen constant.
The parameter ε, appearing in the right hand side of the first equation (11), arises as the natural perturbation parameter of our asymptotic expansion.
Let us remark that, with our scaling argumentation, we obtain the same perturbation parameter as in [17, 19]. Moreover, for any set of kinetic parameters and initial conditions, $\varepsilon \leq \frac{1}{4}$ [19].
Observe that

$$\sigma + \chi_M + \eta = 1 \tag{13}$$

Let us expand the solutions of (11) in the form

$$\overline{s} = \Sigma_0 + \varepsilon\,\Sigma_1 + o(\varepsilon)\,, \quad c = \Gamma_0 + \varepsilon\,\Gamma_1 + o(\varepsilon)\,.$$

Substituting in (11) and taking into account the initial conditions, we find at order 0 that $\Sigma_0 = const. = 1$ and

$$\frac{d\Gamma_0}{d\tau} = \sigma\,\eta\,\Gamma_0^2 - \Gamma_0 + 1 \tag{14}$$

whose solution, complying with (5), is easily found as

$$\Gamma_0(\tau) = \frac{\exp(\sqrt{1 - 4\sigma\,\eta}\;\tau) - 1}{\sigma\,\eta\,[\Gamma_0^+\,\exp(\sqrt{1 - 4\sigma\,\eta}\;\tau) - \Gamma_0^-]} \tag{15}$$

where $\Gamma_0^{\pm} = \dfrac{1 \pm \sqrt{1 - 4\,\sigma\,\eta}}{2\,\sigma\,\eta}$. At order 1 we have

$$\frac{d\Sigma_1}{d\tau} = -\Gamma_0 \tag{16}$$

$$\frac{d\Gamma_1}{d\tau} = \Gamma_1\,(2\sigma\,\eta\,\Gamma_0 - 1) - \sigma\,\Sigma_1\,\Gamma_0 + \Sigma_1 \tag{17}$$

with homogeneous initial conditions, which give

 ISBN: 978-84-614-6167-7

$$\Sigma_1 = \frac{1}{\sigma\,\eta}\log\left(\frac{\Gamma_0^+\,\exp(\sqrt{1-4\sigma\,\eta}\,\tau)-\Gamma_0^-}{\Gamma_0^+-\Gamma_0^-}\right)-\Gamma_0^+\,\tau \tag{18}$$

and the corresponding Γ_1. Details for this latter fuction are given in the Appendix (A).

Now we turn our attention to the outer solutions of (7). We only need to change the timescale γ; to this aim let us focus on the first equation of (7). In the slow, quasi-steady state phase the total variable $\overline{s}$ cannot anymore be considered roughly constant: it decreases monotonically from S_T to zero. Hence, to balance the left hand side with the right one we set

$$\gamma = \frac{\alpha}{k_2\,\beta} = \frac{E_T + K_M + S_T}{v_{max}} \tag{19}$$

where $v_{max} = k_2\,E_T$ is the maximal reaction velocity. In this case, γ represents the time scale $t_{\overline{s}}$, related to the total substrate depletion [8, 19]. Note that, having denoted by t_c the time scale in the fast, pre-steady phase and by $t_{\overline{s}}$ the time scale in the slow, quasi-steady phase, we get

$$\frac{t_c}{t_{\overline{s}}} = \varepsilon\,. \tag{20}$$

Setting $T = \gamma\,t$ and substituting (19) in (7) we obtain

$$\begin{aligned} &\frac{d\overline{s}}{dT} = -c \\ &\varepsilon\frac{dc}{dT} = \sigma\,\eta c^2 - (\eta+\chi_M)\,c - \sigma\,\overline{s}\,c + \overline{s} \end{aligned} \tag{21}$$

Let us expand the solutions of (21) in the form

$$\overline{s} = \overline{s}_0 + \varepsilon\,\overline{s}_1 + o(\varepsilon)\,, \quad c = c_0 + \varepsilon\,c_1 + o(\varepsilon)\,.$$

Upon substitution in (21) we find, at leading order,

$$\begin{aligned} &\frac{d\overline{s}_0}{dT} = -c_0 \\ &\sigma\,\eta c_0^2 - (\eta+\chi_M+\sigma\,\overline{s}_0)\,c_0 + \overline{s}_0 = 0 \end{aligned} \tag{22}$$

which correspond to the equations obtained in the tQSSA [8].
The second equation above is algebraic in c_0 with solutions

$$c_0^{\pm} = \frac{\eta+\chi_M+\sigma\,\overline{s}_0 \pm \sqrt{(\eta+\chi_M+\sigma\,\overline{s}_0)^2 - 4\sigma\,\eta\overline{s}_0}}{2\,\sigma\eta}$$

and it is easy to see that only c_0^- is admissible. Note that, being $\overline{s}_0(0) = \lim_{\tau\to\infty}\Sigma_0(\tau) = 1$, we have automatically that $c_0(0) = c_0^-(0) = \Gamma_0^-$.

 ISBN: 978-84-614-6167-7

From (21) it is found that the first correction terms in the outer solutions are given by

$$\begin{aligned} \frac{d\bar{s}_1}{dT} &= -c_1 \\ c_1 &= \frac{c_0' + \bar{s}_1\,(\sigma\, c_0 - 1)}{2\,\eta\,\sigma\, c_0 - \sigma\,\bar{s}_0 - \eta - \chi_M}\ . \end{aligned} \tag{23}$$

4 Figures and discussion

We have solved numerically (1) and we have compared the results with our approximations at leading order and first order. This is shown in fig.(s) (1–2), where we have changed only two kinetic parameters, in order to have different values of ε. In both cases our uniform expansion gives very good results.
In fig.(s) 2–4 we observe a less accurate approximation around the matching point, since in this case the value of $\varepsilon = 0.1856$ is close to the bound $\frac{1}{4}$.
Observe that we have chosen values for S_T and E_T such that the sQSSA approximates the dynamics with very low accuracy. In fact this latter approximation, in general, works well when there is a substrate excess, or when $K_M \gg E_T$ [5, 6]. When there is an enzyme excess, as in our figures, the perturbation techniques rely on the same parameter ε given in (12), but in a QSSA setting [20, 17, 21]. In fig.(s) (3–4) we compare our first order uniform approximations with the corresponding ones given in [21]. Also in these cases the results are very good.

A Solution of eq.(17)

Let us set $R = \sqrt{1 - 4\eta\sigma}$.

The equation for the term Γ_1 is

$$\frac{d\Gamma_1}{d\tau} = (2\sigma\eta\Gamma_0 - 1)\Gamma_1 + (1 - \sigma\Gamma_0)\Sigma_1 \tag{24}$$

i.e.

$$\begin{aligned} &\tfrac{d\Gamma_1}{d\tau} + \left\{1 + 2\left[\tfrac{1-e^{R\tau}}{\Gamma_0^+ e^{R\tau} - \Gamma_0^-}\right]\right\}\Gamma_1 = \\ &\left\{\tfrac{1}{\sigma\eta}\log\left[\tfrac{\Gamma_0^+ e^{R\tau} - \Gamma_0^-}{\Gamma_0^+ - \Gamma_0^-}\right] - \Gamma_0^+\tau\right\}\left\{1 + \left[\tfrac{1-e^{R\tau}}{\eta\left(\Gamma_0^+ e^{R\tau} - \Gamma_0^-\right)}\right]\right\} \end{aligned} \tag{25}$$

whose formal solution is

$$\Gamma_1(\tau) = e^{-A(\tau)}.\int_0^\tau \left\{\frac{1}{\sigma\eta}\log\left[\frac{\Gamma_0^+ e^{Rw} - \Gamma_0^-}{\Gamma_0^+ - \Gamma_0^-}\right] - \Gamma_0^+ w\right\}\left\{1 + \left[\frac{1 - e^{Rw}}{\eta\left(\Gamma_0^+ e^{Rw} - \Gamma_0^-\right)}\right]\right\} e^{A(w)}\, dw \tag{26}$$

where

$$e^{A(\tau)} = exp\left(\int_0^\tau \left\{1 + 2\left[\frac{1 - e^{Rw}}{\Gamma_0^+ e^{Rw} - \Gamma_0^-}\right]\right\} dw\right) = \left[\frac{\Gamma_0^+ e^{R\tau} - \Gamma_0^-}{\Gamma_0^+ - \Gamma_0^-}\right]^2 \cdot e^{-R\tau}\ . \tag{27}$$

 ISBN: 978-84-614-6167-7

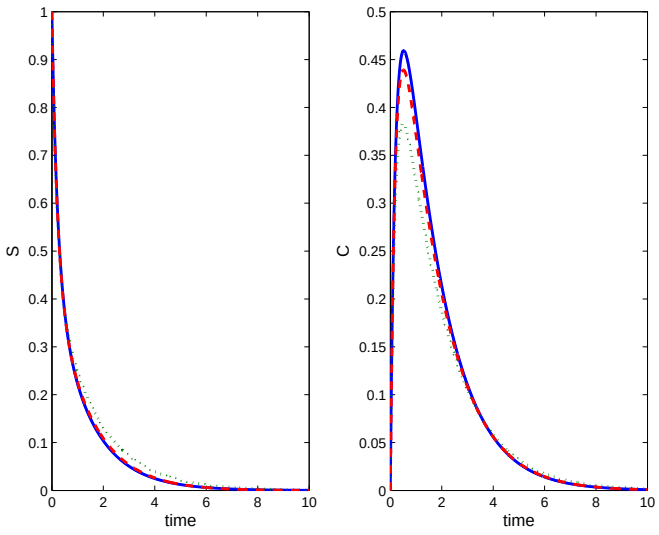


Figure 1: Dynamics of S (left panel) and of C (right panel): full system (solid), leading order uniform approximation (dotted), first order uniform approximation (dashed). Kinetic parameters: $E_T = 3$, $S_T = 1$, $k_1 = 1$, $k_{-1} = 1$, $k_2 = 1$, $\varepsilon = 0.0833$.

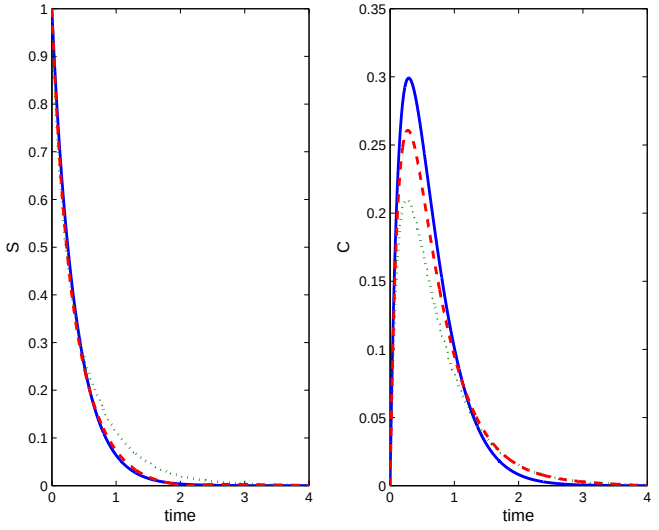


Figure 2: Dynamics of S (left panel) and of C (right panel): full system (solid), leading order uniform approximation (dotted), first order uniform approximation (dashed). Kinetic parameters: $E_T = 3$, $S_T = 1$, $k_1 = 1$, $k_{-1} = 0.04$, $k_2 = 4$, $\varepsilon = 0.1856$.

 ISBN: 978-84-614-6167-7

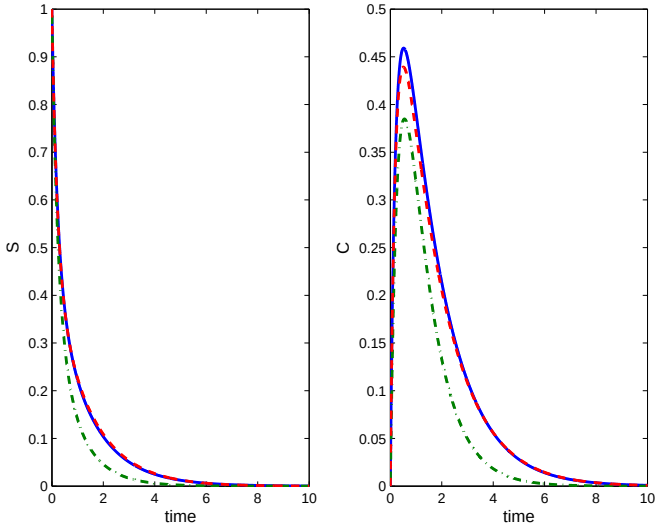


Figure 3: Dynamics of S (left panel) and of C (right panel): full system (solid), first order uniform approximation as in [21] (dashed-dotted), first order uniform approximation (dashed). Kinetic parameters: $E_T = 3$, $S_T = 1$, $k_1 = 1$, $k_{-1} = 1$, $k_2 = 1$, $\varepsilon = 0.0833$.

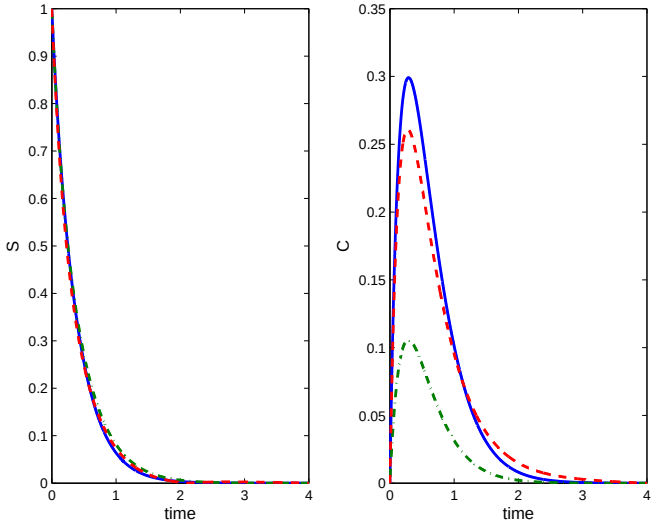


Figure 4: Dynamics of S (left panel) and of C (right panel): full system (solid), first order uniform approximation as in [21] (dashed-dotted), first order uniform approximation (dashed). Kinetic parameters: $E_T = 3$, $S_T = 1$, $k_1 = 1$, $k_{-1} = 0.04$, $k_2 = 4$, $\varepsilon = 0.1856$.

 ISBN: 978-84-614-6167-7

Solving the integrals and setting

$$M(\tau) = \frac{e^{R\tau}}{R \cdot \left(\Gamma_0^+ e^{R\tau} - \Gamma_0^-\right)^2} \quad ; \quad N(\tau) = -\frac{\Gamma_0^+}{R\eta} M(\tau) \quad ; \quad Q(\tau) = \frac{1}{\sigma\eta^2} M(\tau) \tag{28}$$

we have

$$\begin{aligned}
&\Gamma_1(\tau) = N(\tau) \cdot \{e^{R\tau}(R\tau - 1)(\eta\Gamma_0^+ - 1)\Gamma_0^+ + \tfrac{e^{2R\tau}}{2}\left(\Gamma_0^+ + \Gamma_0^- - 2\Gamma_0^+\Gamma_0^-\eta\right) + \Gamma_0^-(\eta\Gamma_0^- - 1) \cdot \\
&\left[-e^{-R\tau}(R\tau + 1) + 1\right]\} + Q(\tau)\{\Gamma_0^+(\eta\Gamma_0^+ - 1)\left[e^{R\tau}\log\left(\tfrac{\Gamma_0^+ e^{R\tau} - \Gamma_0^-}{\Gamma_0^+ - \Gamma_0^-}\right) - e^{R\tau} + 1\right] + \\
&(\Gamma_0^- - \Gamma_0^+)\log\left(\tfrac{\Gamma_0^+ e^{R\tau} - \Gamma_0^-}{\Gamma_0^+ - \Gamma_0^-}\right) + \Gamma_0^-(\eta\Gamma_0^- - 1)\left[-e^{-R\tau}\log\left(\tfrac{\Gamma_0^+ e^{R\tau} - \Gamma_0^-}{\Gamma_0^+ - \Gamma_0^-}\right) - \tfrac{\Gamma_0^+}{\Gamma_0^-}R\tau\right] + \\
&(\Gamma_0^+ + \Gamma_0^- - 2\eta\Gamma_0^+\Gamma_0^-)\int_1^{e^{R\tau}} \tfrac{1}{z}\log\left(\tfrac{\Gamma_0^+ z - \Gamma_0^-}{\Gamma_0^+ - \Gamma_0^-}\right)\, dz\} \ .
\end{aligned} \tag{29}$$

References

[1] M. Z. Bodenstein, "Eine theorie der photochemischen reaktionsgeschwindigkeiten," *Z. Phys. Chem.*, vol. 85, pp. 329–397, 1913.

[2] D. L. Chapman and L. K. Underhill, "The interaction of chlorine and hydrogen. the influence of mass," *J. Chem. Soc. Trans.*, vol. 103, pp. 496–508, 1913.

[3] L. Michaelis and M. L. Menten, "Die kinetik der invertinwirkung.," *Biochem. Z.*, vol. 49, pp. 333–339, 1913.

[4] G. Briggs and J. Haldane, "A note on the kinetics of enzyme action," *Biochem. J.*, vol. 19, pp. 338–339, 1925.

[5] F. G. Heineken, H. M. Tsushiya, and R. Aris, "On the mathematical status of the pseudo-steady state hypothesis of biochemical kinetics," *Math. Biosc.*, vol. 1, pp. 95–113, 1967.

[6] L. A. Segel and M. Slemrod, "The quasi steady-state assumption: a case study in perturbation.," *Siam Rev.*, vol. 31, pp. 446–477, 1989.

[7] K. J. Laidler, "Theory of the transient phase in kinetics, with special reference to enzyme systems," *Can. J. Chem.*, vol. 33, pp. 1614–1624, 1955.

[8] J. Borghans, R. de Boer, and L. Segel, "Extending the quasi-steady state approximation by changing variables," *Bull.Math.Biol.*, vol. 58, pp. 43–63, 1996.

[9] A. Tzafriri, "Michaelis-Menten kinetics at high enzyme concentrations," *Bull. Math. Biol.*, vol. 65, pp. 1111–1129, 2003.

 ISBN: 978-84-614-6167-7

[10] M. M. G. Pedersen, A. M. Bersani, and E. Bersani, "The total quasi steady-state approximation for fully competitive enzyme reactions," *Bull. Math. Biol.*, vol. 2005, p. 69.

[11] M. G. Pedersen, A. M. Bersani, and E. Bersani, "Steady-state approximations in intracellular signal transduction – a word of caution," *J. Math. Chem.*, vol. 43, pp. 1318–1344, 2008.

[12] M. G. Pedersen, A. M. Bersani, E. Bersani, and G. Cortese, "The total quasi-steady state approximation for complex enzyme reactions," *MATCOM*, vol. 79, p. 2008, 2008.

[13] M. Pedersen and A. M. Bersani, "The total quasi-steady state approximation simplifies theoretical analysis at non-negligible enzyme concentrations: Pseudo first-order kinetics and the loss of zero-order ultrasensitivity," *J. Math. Biol.*, vol. 60, pp. 267–283, 2010.

[14] A. R. Tzafriri and E. R. Edelman, "The total quasi-steady-state approximation is valid for reversible enzyme kinetics," *J. Theor. Biol.*, vol. 226, pp. 303–313, 2004.

[15] A. Ciliberto, F. Capuani, and J. J. Tyson, "Modeling networks of coupled anzymatic reactions using the total quasi-steady state approximation," *PLoS Comput. Biol.*, vol. 3, pp. 463–472, 2007.

[16] S. MacNamara, A. M. Bersani, K. Burrage, and R. B. Sidje, "Stochastic chemical kinetics and the total quasi-steady-state assumption: application to the stochastic simu- lation algorithm and chemical master equation," *J. Chem. Phys.*, vol. 129, pp. 1–13, 2008.

[17] S. Schnell and P. K. Maini, "Enzyme kinetics far from the standard quasi-steady-state and equilibrium approximations," *Math Comp. Model.*, vol. 35, pp. 137–144, 2002.

[18] J. Wei and J. C. W. Kuo, "A lumping analysis in monomolecular reaction systems: analysis of the exactly lumped system," *Ind. Eng. Chem. Fundam.*, vol. 8, pp. 114–123, 1969.

[19] J. W. Dingee and A. B. Anton, "A new perturbation solution to the Michaelis-Menten problem," *AIChE J.*, vol. 54, pp. 1344–1357, 2008.

[20] S. Schnell and P. K. Maini, "Enzyme kinetics at high enzyme concentration," *Bull. Math. Biol.*, vol. 62, pp. 483–499, 2000.

[21] A. M. Bersani and G. Dell'Acqua, "Asymptotic expansions in enzyme reactions with high enzyme concentration," *Math. Method. Appl. Sci. (to appear)*, 2011.

 ISBN: 978-84-614-6167-7

Proceedings of the 11th International Conference on Computational and Mathematical Methods in Science and Engineering, CMMSE 2011 26–30 June 2011.

Metaecoepidemics with migration of and disease in the predators.

Federica Bianco[1], Elena Cagliero[1], Marino Gastelurrutia[1] and Ezio Venturino[1]

[1] *Dipartimento di Matematica "Giuseppe Peano", Università di Torino, via Carlo Alberto 10, 10123 Torino, Italy*

emails: fede.03.87@alice.it, elenacagliero87@yahoo.it, mgastelu@gmail.com, ezio.venturino@unito.it

Abstract

In this contribution we outline the basic ideas behind metaecoepidemics. We consider a simple predator-prey system with two possible habitats and where an epidemic spreads by contact among the predators, in the patch in which prey are absent. Only the sound predators can freely move from one environment to another. The equilibria of the system are analyzed for stability. Theoretical and simulation results show that modifying the environment in which interacting populations live may entail unforeseen consequences. In particular circumstances however, it could constitute also a possible way of eradicating pests.

Key words: Eco-epidemiology; Local Stability; Metapopulations; Habitat
MSC 2000: AMS codes (92D25, 92D30)

1 Background information

Population theory started from the famous paper by Malthus at the end of the seventeenth century, [26], stating that populations either disappear or grow exponentially and therefore more and more resources are needed to sustain this growth, which is ultimately mot possible. These ideas were improved in the following century by Verhulst, who introduced the so-called logistic growth, [38, 39, 40]. But it is only in the twenties of the past century that the works of Lotka and Volterra on interacting species, [24, 41], originated many fruitful subsequent developments of this branch of science.

In more recent times, after many years of investigations involving not just two populations but entire food webs, researchers formulated a new concept, metapopulation. By this word namely, populations are considered living in different environments, among which possibly they freely migrate, [42].

 ISBN: 978-84-614-6167-7

The metapopulation tool is a good way of modeling real life situations in which human activiy, or natural causes, fragment the landscape and lead therefore to heterogeneous environments. Especially human activity, breaking the landscape via artifacts such as buildings, roads, or simply by clearing the vegetation from wild areas in order to find new fields for agriculture, is a major cause for the loss of natural habitat. Ultimately, this landscape reshaping might threaten the existence of wild populations, and for environmentalists and conservationists this has become a major issue, [43]. In fact, the original population settled in the unperturbed environment, in view of this landscape fragmentation, becomes separated into subpopulations. Each of them continues to independently thrive in each patch, but now it is more sensitive to perturbations, such as adverse climatic conditions. This fact is ultimately responsible for species extinction.

It has been observed that in fragmented environments in general the population persists globally, although sometimes the populations become locally extinct [8, 13, 16, 42, 45]. This positive finding is attributed to the migration possibility between patches.

The role of mathematical models, which are able to predict possible outcomes of specific situations [21], reveals to be essential in this situation, since data collections constitute a major problem, and are generally not gathered, [8, 13, 21]. Metapopulations represent a tool for assessing population dynamics in patched environments, [17, 31]. The classical Levins model [23] assumes that colonization depends only the part of the environment in which populations settle. Here only the habitats providing the most favorable living conditions become populated, while the ground that connects them is only used for interpatch migrations.

Two real life situations in which metapopulation dynamics has been successfully used are represented by the spotted owl (*Strix occidentalis*) and the mountain sheep (*Ovis canadensis*) populations, [14].

A further more complex instance is the study of *Melitaea cinxia* [18]. This butterfly in Finland, as well as other species [27, 28], has been studied with the help of incidence functions. For the butterfly, observations led to the conclusion that variations in local populations are related to the interaction with a specialist braconid parasitoid, *Cotesia melitaearum* [22]. This is a clear indication that there is a need for a metapopulation approach which accounts also for a host-parasitoid metapopulation dynamics [19]. With this background a first metapopulation model accounting for diseases in the above type of models has been proposed, [37]. To set up such models, we need to backstep and describe the idea of ecoepidemiology, a field of investigations dealing with population systems among which diseases spread, see [25] for an introduction.

From the first studies which assume mostly quadratic predator-prey models to model rather simple situations, [15, 11, 3, 32, 33, 35], more complex hypotheses have been considered, [6, 1]. But other kinds of demographic interactions have also been studied, for instance competition and symbiotic interactions, [34, 36].

By considering ecoepidemic systems in fragmented habitats, we obtain metaecoepidemics. This new field of study, [37] therefore investigates environments that are physically separated, but among which disease-affected populations can migrate.

The paper is organized as follows. In the next Section we provide some real life

 ISBN: 978-84-614-6167-7

examples in which these situations arise. The basic model studied here is presented in Section 3. Then in the next Section we analytically determine its equilibria and study their stability in Section 5. Numerical simulations are then provided and with a final discussion they conclude the paper.

2 Biological examples

The biological examples provided in the previous Section fall also in the framework of the metaecoepidemic models. In fact, there are well known diseases that affect these populations.

Ovis canadensis has various predators among which we mention the wolf (*Canis lupus*), coyote (*Canis latrans*), bear (*Ursus*), Canada lynx (*Lynx canadensis*), mountain lion (*Puma concolor*), golden eagle (*Aquila chrysaetos*), [9]. Bighorn sheep also host several parasites, such as nematode lungworms, *Protostrongylus stilesi* and *P. rushi*, [10].

Strix occidentalis is affected by helmints and it is hunted by the great horned owl, *Bubo virginianus*, which hunts the spotted owl either for feeding purposes but also to fight it as a competitor for resources. The two species in fact have several prey in common, [20]. Note that the prey of *Strix occidentalis*, mainly small rodents, play an important role, as they appear to be the vector by which the spotted owl gets the parasites, i.e. the infection, [20].

Predators of Lepidoptera are mainly birds, bats, parasitoids, small mammals, reptiles and other insects, especially ants and dragonflies. In particular the larvae are preferred in the diet of *Parus caerulens, P. major*, [7]. Some species of bats can eat Lepidoptera up to half their weight per night, [4, 5]. Hymenoptera and Diptera either kill, immobilize or implant in Lepidoptera eggs, that later will kill the host, [12, 30]. Viruses, e.g. nuclear polyhedrosis virus (NPV), cytoplasmatic polyhedrosis virus (CPV), granulosis virus (GV), entomopox virus (EPV), small RNA viruses also affect Lepidoptera, [29]. The common bacterium *Bacillus thuringiensis* var. kurstaki kills caterpillars when they assume it, by poisoning their digestive tract. Butterflies use several ways of defending themselves from predators, using shapes, producing colors or sounds which scare them, [30]. Another means of defense is myrmecophily, by which the caterpillars associate with ants, finding safe refuge near the ant nests, [2].

In addition to the three cases mentioned above, we provide another example assessing the need of studying populations living in fragmented habitats, which are also affected by epidemics. It is represented by the predator-prey interaction of red fox *Vulpes vulpes* (L.) and rabbits *Oryctolagus cuniculus* (L.), affected by the *Myxoma* virus. This is a classical case, since the wild rabbits in Australia have been inoculated with myxomatosis by humans to try to control their population, [44].

The necessity of studying epidemics-affected populations living in fragmented habitats is thus clear from the mentioned situations. Having outlined the basic ideas behind metaecoepidemics, in this contribution we specifically investigate a kind of dual situation that has been examined in [37]. In this context, we consider two patches in

 ISBN: 978-84-614-6167-7

which predators live, in the first one of which they have a prey population to hunt, in the second one however no resource to survive upon. In addition, we assume that in this second patch an epidemics spreads. Finally, only healthy predators can migrate between the two patches. This assumption is meaningful in view of the fact that migration might require an effort, and the weakest individuals, namely the disease-affected ones, are unable to exert it.

3 The model

We denote the two patches by A_1 and A_2. In A_1 there are prey $N(t)$ and the healthy predators $S_1(t)$. In A_2 there are only predators, among which an epidemic spreads. This is a harsher environment, with no resources for the survival of predators. We denote by $S_2(t)$ the heatlhy individuals and by $I(t)$ the infected ones. We assume that only the predators can migrate between the two patches, and in particular only if they are strong enough. Thus, infected individuals, being weakened by the unrecoverable disease, are confined to the second patch. This also implies that the epidemic cannot spread to the first environment. The model reads as follows

$$\begin{aligned} \frac{dN}{dt} &= rN\left(1-\frac{N}{K}\right) - aNS_1, \\ \frac{dS_1}{dt} &= bNS_1 - mS_1 - n_{21}S_1 + n_{12}S_2, \\ \frac{dS_2}{dt} &= -\gamma S_2 I - mS_2 + n_{21}S_1 - n_{12}S_2, \\ \frac{dI}{dt} &= \gamma S_2 I - \mu I. \end{aligned} \tag{1}$$

Here, a denotes the hunting rate on prey, r their reproduction rate and K their carrying capacity. Thus the first equation models logistic growth for the prey and the hunting on them by the predators in A_1. In the second equation we describe the dynamics of the predators in the first patch. Let $b < a$ denote the reward they get from hunting, m their mortality rate and n_{21} and n_{12} their migration rates from and into patch A_1 respectively. The second equation then states the fact that predators reproduce when they hunt successfully, are subject to natural mortality, and increase and decrease their population size based also on migrations out and into the patch. In the third equation we begin to describe patch A_2. Let γ denote the disease incidence and μ the natural plus disease-related mortality. The susceptible predators possibly catch the disease by interaction with infected ones, die from natural causes and migrate into and out from this environment. The last equation contains the infected predators dynamics. Individuals enter into this class from the healthy class by catching the disease, and die by biological or disease-related causes.

 ISBN: 978-84-614-6167-7

4 Equilibria

In this Section we study the model's equilibria. We easily find the origin, E_1, since the system (1) is homogeneous. We investigate at first the boundary equilibria, in which at least one population vanishes. Only two such points turn out to be feasible. We thus find the point $E_2 = (K, 0, 0, 0)$. Further, if we seek points for which $I = 0$, and solve for S_2 the third equilibrium equation, we find

$$S_2 = \frac{n_{21}}{m + n_{12}} S_1 \tag{2}$$

and substituting into the second one, we have

$$\widetilde{N} = \frac{m(m + n_{12} + n_{21})}{b(m + n_{12})}. \tag{3}$$

From the first equation we find then

$$\widetilde{S}_1 = \frac{r[Kb(m + n_{12}) - m(m + n_{12} + n_{21})]}{aKb(m + n_{12})} \tag{4}$$

which in turn upon substitution into (2) provides the value of S_2

$$\widetilde{S}_2 = \frac{rn_{21}[Kb(m + n_{12}) - m(m + n_{12} + n_{21})]}{aKb(m + n_{12})^2}. \tag{5}$$

We have thus found the equilibrium point $E_3 = \left(\widetilde{N}, \widetilde{S}_1, \widetilde{S}_2, 0\right)$, where the population values are explicitly given in (3),(4) e (5).

It is easily established that

$$bK(m + n_{12}) > m(m + n_{12} + n_{21}) \tag{6}$$

represents the only feasibility condition for the equilibrium E_3. It amounts to say that the prey carrying capacity for the A_1 environment must exceed the number of prey at equilibrium.

Next we examine the case for which all populations do not vanish, namely the metasystem exhibits coexistence. In this case we solve first from the last equation to get

$$\hat{S}_2 = \frac{\mu}{\gamma}. \tag{7}$$

Substitution into the second one gives

$$S_1 = \frac{-n_{12}\mu}{\gamma\left(bN - m - n_{21}\right)}. \tag{8}$$

From the first equation we then obtain the following quadratic equation in N

$$r\gamma bN^2 - r\gamma\left(bK + m + n_{21}\right)N - K(an_{12}\mu - r\gamma m - r\gamma n_{21}) = 0, \tag{9}$$

 ISBN: 978-84-614-6167-7

the discriminant of which is always positive, since

$$\begin{aligned}\Delta &= r^2\gamma^2(bK+m+n_{21})^2 + 4r\gamma bK\left(an_{12}\mu - r\gamma m - r\gamma n_{21}\right) \\ &= \left[r\gamma(bK-m-n_{21})\right]^2 + 4r\gamma bKan_{12}\mu > 0.\end{aligned} \tag{10}$$

Both roots are then real,

$$N_1 = \frac{r\gamma\left(m+bK+n_{21}\right) + \sqrt{\left[r\gamma(bK-m-n_{21})\right]^2 + 4r\gamma bKan_{12}\mu}}{2rb\gamma},$$
$$N_2 = \frac{r\gamma\left(bK+m+n_{21}\right) - \sqrt{\left[r\gamma(bK-m-n_{21})\right]^2 + 4r\gamma bKan_{12}\mu}}{2r\gamma b}.$$

Substitution of N_1 into (8) leads to the expression

$$S_1 = \frac{-2rn_{12}\mu}{r\gamma(bK-m-n_{21}) + \sqrt{\left[r\gamma(bK-m-n_{21})\right]^2 + 4r\gamma bKan_{12}\mu}}$$

in which the denominator is positive while the numerator is instead negative, so that this value is $S_1 < 0$ and therefore it is not acceptable. We thus consider only

$$\hat{N} = N_2 = \frac{r\gamma\left(bK+m+n_{21}\right) - \sqrt{\left[r\gamma(bK-m-n_{21})\right]^2 + 4r\gamma bKan_{12}\mu}}{2r\gamma b}, \tag{11}$$

from which $S_1 > 0$ is acceptable. Final substitution of this expression into (8) gives

$$\hat{S}_1 = \frac{-2rn_{12}\mu}{r\gamma\left(bK-m-n_{21}\right) - \sqrt{\left[r\gamma(bK-m-n_{21})\right]^2 + 4r\gamma bKan_{12}\mu}}. \tag{12}$$

From the third equation we find

$$I = \frac{\left(m+n_{12}\right)S_2 - n_{21}S_1}{-\gamma S_2}.$$

Substituting the values of S_2 and S_1 into (7) and (8), we finally have

$$\hat{I} = \frac{-2rn_{12}n_{21}}{r\gamma\left(bK-m-n_{21}\right) - \sqrt{\left[r\gamma(bK-m-n_{21})\right]^2 + 4r\gamma bKan_{12}\mu}} - \frac{m+n_{12}}{\gamma}. \tag{13}$$

We have thus found the components of $E_4 = \left(\hat{N}, \hat{S}_1, \hat{S}_2, \hat{I}\right)$, i.e. the coexistence equilibrium point.

We discuss now how to ensure feasibility. Since $\hat{S}_1 > 0$ and $\hat{S}_2 > 0$, we need to require

$$r\gamma\left(bK+m+n_{21}\right) > \sqrt{\left[r\gamma(bK-m-n_{21})\right]^2 + 4r\gamma bKan_{12}\mu}$$

which reduces to

$$m + n_{21} > \frac{an_{12}\mu}{r\gamma}. \tag{14}$$

As far as $\hat{I} > 0$ is concerned, the study is more complicated. It is equivalent to

$$\frac{-2rn_{12}n_{21}}{r\gamma\left(bK-m-n_{21}\right) - \sqrt{\left[r\gamma(bK-m-n_{21})\right]^2 + 4r\gamma bKan_{12}\mu}} > \frac{m+n_{12}}{\gamma}.$$

 ISBN: 978-84-614-6167-7

Note that the denominator on the left is negative, so that the condition reduces to

$$\frac{2r\gamma n_{12}n_{21}}{m+n_{12}} + r\gamma(bK - m - n_{21}) > \sqrt{[r\gamma(bK - m - n_{21})]^2 + 4r\gamma bKan_{12}\mu},$$

which in turn gives

$$\begin{cases} [r\gamma(bK - m - n_{21})]^2 + 4r\gamma bKan_{12}\mu \geq 0, \\ \dfrac{2r\gamma n_{12}n_{21}}{m+n_{12}} + r\gamma(bK - m - n_{21}) > 0, \\ \left[\dfrac{2r\gamma n_{12}n_{21}}{m+n_{12}} + r\gamma(bK - m - n_{21})\right]^2 > [r\gamma(bK - m - n_{21})]^2 + 4r\gamma bKan_{12}\mu. \end{cases}$$

In view of (10) the first above inequality is always satisfied. The second one gives

$$bK - m - n_{21} > -\frac{2n_{12}n_{21}}{m+n_{12}}$$

and the third one

$$bK - m - n_{21} > \frac{bKa\mu(m+n_{12})}{r\gamma n_{21}} - \frac{n_{12}n_{21}}{m+n_{12}}.$$

The system reduces to

$$\begin{cases} bK - m - n_{21} > -\dfrac{2n_{12}n_{21}}{m+n_{12}}, \\ bK - m - n_{21} > \dfrac{bKa\mu(m+n_{12})}{r\gamma n_{21}} - \dfrac{n_{12}n_{21}}{m+n_{12}}. \end{cases}$$

Since the condition

$$-\frac{2n_{12}n_{21}}{m+n_{12}} < \frac{bKa\mu(m+n_{12})}{r\gamma n_{21}} - \frac{n_{12}n_{21}}{m+n_{12}}$$

holds true inconditionally, the positivity of $\hat{I}$ reduces to only

$$m + n_{21} < bK + \frac{n_{12}n_{21}}{m+n_{12}} - \frac{bKa\mu(m+n_{12})}{r\gamma n_{21}}. \tag{15}$$

To sum up, E_4 is feasible if the following condition holds

$$\frac{an_{12}\mu}{r\gamma} - n_{21} < m < bK + \frac{n_{12}n_{21}}{m+n_{12}} - \frac{bKa\mu(m+n_{12})}{r\gamma n_{21}} - n_{21}. \tag{16}$$

5 Local stability analysis

The Jacobian $J \equiv J(N, S_1, S_2, I)$ of the system (1) at a generic point is given by

$$J = \begin{pmatrix} -aS_1 + r\left(1 - \dfrac{N}{K}\right) - \dfrac{rN}{K} & -aN & 0 & 0 \\ bS_1 & bN - m - n_{21} & n_{12} & 0 \\ 0 & n_{21} & -\gamma I - m - n_{12} & -\gamma S_2 \\ 0 & 0 & \gamma I & \gamma S_2 - \mu \end{pmatrix}.$$

 ISBN: 978-84-614-6167-7

5.1 Local stability of E_1

$J_0 \equiv J(0,0,0,0)$ has two explicit eigenvalues, r and $-\mu$, while the remaining two are the roots of the quadratic

$$(-m-n_{21}-\lambda)(-m-n_{12}-\lambda)-n_{21}n_{12}=0$$

which has the roots

$$-m-n_{12}-n_{21}, \quad -m.$$

In view of the first eigenvalue being positive, the origin is an unstable equilibrium.

5.2 Local stability of E_2

Again, the Jacobian factors to provide immediately two explicit eigenvalues, namely $-\mu$ and $-r$, while the remaining ones are the roots of the quadratic

$$(bK-m-n_{21}-\lambda)(-m-n_{12}-\lambda)-n_{21}n_{12}=0,$$

or, rearranging,

$$\lambda^2+(-bK+2m+n_{21}+n_{12})\lambda-bKm-bKn_{12}+n_{12}m+n_{21}m+m^2=0.$$

Its roots turn out to be

$$\frac{bK-2m-n_{21}-n_{12}\pm\sqrt{b^2K^2+n_{21}^2+n_{12}^2+2bKn_{12}-2bKn_{21}+2n_{21}n_{12}}}{2}.$$

Observing that

$$(bK-n_{12}-n_{21})^2=b^2K^2+n_{12}^2+n_{21}^2-2bKn_{12}-2bKn_{21}+2n_{12}n_{12}$$

the argument of the root becomes

$$(bK-n_{12}-n_{21})^2+4bn_{12}$$

and the eigenvalues simplify to the form

$$\lambda_3=\frac{bK-n_{12}-n_{21}-\sqrt{(bK-n_{12}-n_{21})^2+4bn_{12}}}{2}-m=c_0-m$$

with $c_0<0$ and

$$\lambda_4=\frac{bK-n_{12}-n_{21}+\sqrt{(bK-n_{12}-n_{21})^2+4bn_{12}}}{2}-m=c_1-m$$

In summary, we obtain conditional stability for this equilibrium, since $\lambda_i<0$ for $i=1,2,3$ and $\lambda_4<0$ for $m>c_1$, while $\lambda_4>0$ for $0<m<c_1$.

Summarizing,

Existence	Stability	Equilibrium
—	$1>\frac{1}{2m}\left[bK-n_{12}-n_{21}+\sqrt{(bK-n_{12}-n_{21})^2+4bn_{12}}\right]$	E_2
true	true	stable
true	false	unstable

 ISBN: 978-84-614-6167-7

5.3 Local stability of E_3

In this case the Jacobian turns out to be a tridiagonal matrix. One eigenvalue factors out, namely $\gamma\widetilde{S}_2 - \mu$, see (5). The remaining ones are those of the matrix of order 3

$$J_3' = \begin{pmatrix} \dfrac{-rm(m+n_{12}+n_{21})}{bK(m+n_{12})} & \dfrac{-am(m+n_{12}+n_{21})}{b(m+n_{12})} & 0 \\ \dfrac{r\left[bK(m+n_{12})-m(m+n_{12}+n_{21})\right]}{aK(m+n_{12})} & \dfrac{-n_{12}n_{21}}{m+n_{12}} & n_{12} \\ 0 & n_{21} & -m-n_{12} \end{pmatrix}$$

which gives a cubic characteristic equation

$$-P_{J_3'}(\lambda) = \lambda^3 - \mathrm{tr}(J_3')\lambda^2 + Z(J_3')\lambda - \det(J_3'), \tag{17}$$

with

$$Z\left(J_3'(\lambda)\right) = \frac{rm(m+n_{12}+n_{21})[n_{12}n_{21}+bK(m+n_{12})]}{bK(m+n_{12})^2} + \\ + \frac{rm(m+n_{12}+n_{21})[-m(m+n_{12}+n_{21})+(m+n_{12})^2]}{bK(m+n_{12})^2}.$$

To investigate the roots of this equation, we use the Routh-Hurwitz criterion. Note that since $-\mathrm{tr}(J_3') \geq 0$ and $-\det(J_3') \geq 0$ it is enough to study the sign of D_2, where $D_3 = -\det(J_3')D_2$. Explicitly, we have

$$D_2 = -\mathrm{tr}(J_3')Z(J_3') + \det(J_3') = \\ = \frac{rm(m+n_{12}+n_{21})[rm(m+n_{12}+n_{21})+bKn_{12}n_{21}][n_{12}n_{21}+(m+n_{12})^2]}{K^2b^2(m+n_{12})^3} + \\ + \frac{rbKm(m+n_{12}+n_{21})(m+n_{12})^2[n_{12}n_{21}+(m+n_{12})^2]}{K^2b^2(m+n_{12})^3},$$

so that also $D_2 \geq 0$. Therefore the Routh-Hurwitz criterion ensures that all the roots of $P_{J_3'}(\lambda)$ have negative real part. Hence stability is regulated by the first eigenvalue. Explicitly, we then find the stability conditon

$$\mu > \frac{r\gamma n_{21}[bK(m+n_{12})-m(m+n_{12}+n_{21})]}{Kab(m+n_{12})^2}. \tag{18}$$

In summary

Existence	Stability	Equilibrium
$\dfrac{bK(m+n_{12})}{m(m+n_{12}+n_{21})} \geq 1$	$1 > \dfrac{r\gamma n_{21}[bK(m+n_{12})-m(m+n_{12}+n_{21})]}{\mu aKb(m+n_{12})^2}$	E_3
true	true	stable
true	false	unstable
false	true	infeasible
false	false	infeasible

 ISBN: 978-84-614-6167-7

5.4 Local stability analysis of E_4

Also in this case the Jacobian is a tridiagonal matrix. Its entries are rather complicated. However, it is possible to ascertain the sign of each one of them, as follows. Firstly, in addition to the zero entries due to the matrix structure, in this case also $J(E_4)_{44} = 0$. Furthermore, $J(E_4)_{23} = n_{12}$, $J(E_4)_{32} = n_{21}$, $J(E_4)_{34} = -\mu$. Letting $\Omega = [r\gamma(bK + m + n_{21})]^2 + 4r\gamma bK(an_{12}\mu - r\gamma m - r\gamma n_{21}) \equiv [r\gamma(bK - m - n_{21})]^2 + 4r\gamma bKan_{12}\mu$ we have

$$J(E_4)_{11} = \frac{r\sqrt{\Omega} + r^2\gamma(m + n_{21}) - 2ran_{12}\mu - r^2\gamma bK}{-2r\gamma(m + n_{21}) + r\gamma(bK + m + n_{21}) - \sqrt{\Omega}} \equiv \frac{N_{11}}{D_{11}}, \tag{19}$$

$$J(E_4)_{12} = -a\frac{r\gamma(bK + m + n_{21}) - \sqrt{\Omega}}{2r\gamma b}, \tag{20}$$

$$J(E_4)_{21} = \frac{-2r\gamma n_{12}\mu b}{\gamma\left[r\gamma(bK + m + n_{21}) - \sqrt{\Omega} - 2r\gamma(m + n_{21})\right]}, \tag{21}$$

$$J(E_4)_{22} = \frac{r\gamma(bK + m + n_{21})}{2r\gamma} + \frac{-\sqrt{\Omega}}{2r\gamma} - m - n_{21}, \tag{22}$$

$$J(E_4)_{33} = \frac{2r\gamma n_{12}n_{21}}{r\gamma\,(bK - m - n_{21}) - \sqrt{\Omega}}, \tag{23}$$

$$J(E_4)_{43} = \frac{-2r\gamma n_{12}n_{21}}{r\gamma\,(bK - m - n_{21}) - \sqrt{\Omega}} - m - n_{12}. \tag{24}$$

We now proceed to study the sign of the Jacobian entries. Easily, $J_{12} = -a\hat{N} < 0$, $J_{21} = b\hat{S}_1 > 0$, $J_{23} = n_{12} > 0$, $J_{32} = n_{21} > 0$, $J_{33} = -\gamma\hat{I} - m - n_{12} < 0$, $J_{34} = -\mu < 0$, $J_{43} = \gamma\hat{I} > 0$. Moreover, for J_{11} we have

$$D_{11} = r\gamma\,(bK - m - n_{21}) - \sqrt{\Omega} < 0 \tag{25}$$

while $N_{11} < 0$ if and only if

$$\sqrt{\Omega} < 2an_{12}\mu + r\gamma bK - r\gamma(m + n_{21})$$

which then gives the following system of inequalities

$$\begin{cases} \Omega \geq 0, \\ 2an_{12}\mu + r\gamma bK - r\gamma(m + n_{21}) > 0, \\ \Omega < \left[2an_{12}\mu + r\gamma bK - r\gamma(m + n_{21})\right]^2. \end{cases}$$

In view of (10) the first inequality always holds. From the remaining ones we get

$$m < \frac{an_{12}\mu}{r\gamma} - n_{12}$$

 ISBN: 978-84-614-6167-7

which gives $\gamma r(m+n_{21}) < an_{12}\mu$, and using (14) the latter is never satisfied when E_4 is feasible. It follows that $N_{11} > 0$ and therefore $J_{11} > 0$.

For J_{22} we have

$$J_{22} = -2r\gamma(m+n_{21}) + r\gamma(bK+m+n_{21}) - \sqrt{\Omega} = r\gamma\left(bK - m - n_{21}\right) - \sqrt{\Omega} < 0.$$

In summary, the Jacobian has the following sign structure

$$J(E_4) = \begin{pmatrix} - & - & 0 & 0 \\ + & - & + & 0 \\ 0 & + & - & - \\ 0 & 0 & + & 0 \end{pmatrix} \tag{26}$$

Now the characteristic polynomial is

$$\begin{aligned} P_{J_4}(\lambda) \equiv \sum_{i=0}^{4} p_i\lambda^i = \lambda^4 - (J_{33}+J_{22}+J_{11})\lambda^3 \\ +(J_{33}J_{22} - J_{43}J_{34} - J_{32}J_{23} + J_{33}J_{11} - J_{21}J_{12} + J_{22}J_{11})\lambda^2 \\ +[(J_{22}+J_{11})J_{43}J_{34} + J_{33}J_{21}J_{12} - J_{33}J_{22}J_{11} + J_{32}J_{23}J_{11}]\lambda \\ +(J_{21}J_{12} - J_{22}J_{11})J_{43}J_{34}. \end{aligned}$$

The polynomial is monic and $p_3 = -J_{33} - J_{22} - J_{11} > 0$. Furthermore since $J_{33}J_{22} = n_{21}n_{12}$, we find that

$$p_2 = -J_{43}J_{34} + n_{21}n_{12} + J_{33}J_{11} - J_{21}J_{12} + J_{22}J_{11} > 0.$$

Also, since $J_{32}J_{23}J_{11} = J_{33}J_{22}J_{11}$, we have

$$p_1(J_{22}+J_{11})J_{43}J_{34} + J_{33}J_{21}J_{12} - J_{33}J_{22}J_{11} + J_{32}J_{23}J_{11} > 0$$

which is always satisfied since (26) implies that every term is positive.

Finally, $p_0 = (J_{21}J_{12} - J_{22}J_{11})J_{43}J_{34} > 0$, in view of (26). Thus every coefficient of the characteristic equation is positive, $p_i > 0$, for all $i = 0, \ldots, 4$. Thus the Routh-Hurwitz criterion since $D_1 = p_3 > 0$ requires only

$$D_3 = \det \begin{pmatrix} p_3 & p_1 & 0 \\ p_4 & p_2 & p_0 \\ 0 & p_3 & p_1 \end{pmatrix} > 0.$$

The latter explicitly gives

$$D_3 = -p_3^2p_0 + p_1p_2p_3 - p_1^2p_4 > 0. \tag{27}$$

Summarizing, the only feasible cases are the following ones

Existence	Stability	Equilibrium
C_1	C_2	E_4
true	true	stable
true	false	unstable

where C_1 is the condition given by (16) and C_2 is the condition (27).

 ISBN: 978-84-614-6167-7

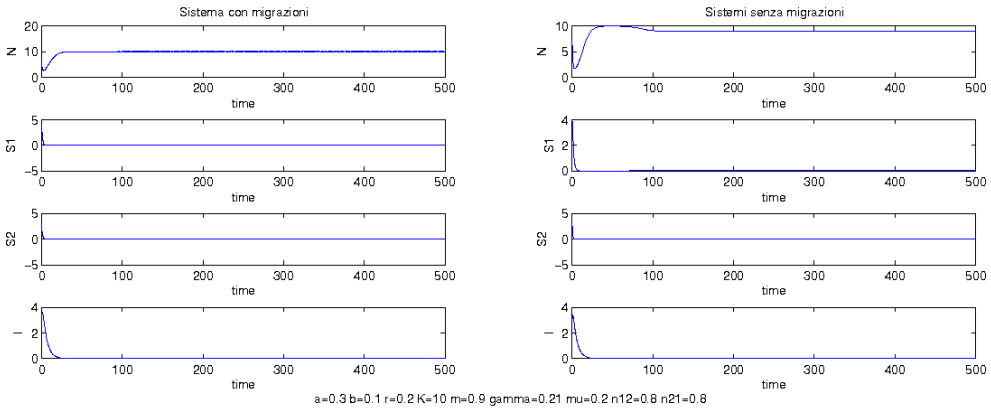


Figure 1: The prey-only equilibrium E_2 is attained for the parameter values .

6 Simulations and interpretation

We have used Matlab to simulate the system's behavior. In the figures we report the system's behavior, left column, compared to the two uncoupled models, right column. In the latter, it is apparent that the predators die out in patch A_2 because of lack of resources, in view of our assumptions.

Figure 1 shows the system behavior at equilibrium E_2. In both the coupled and uncoupled models, sound predators are wiped out in both patches.

In Figure 2 we discover instead an interesting phenomenon. While in the uncoupled patches predators disappear in the harsh environment, as stated above, the fact that they are allowed to migrate from one to the other one makes them survive also in the unfavorable territory. Viewed from the converse point of view, to interfere with the environment in a way as to cut out the possibility for animals to move freely between two patches, may make one population that formerly thrives in it to vanish altogether.

Figure 3 shows the same phenomenon, in which however also the disease persists in the unfavorable environment in addition to the healthy predators. This is the co-existence equilibrium E_4. In this case instead, if the aim is to fight the disease and possibly eradicate it, the measure of breaking the contacts with the environment in which resources are available could constiture a good way of controlling the epidemic.

In Figure 4 we find another interesting system's behavior. Namely, persistent limit cycles in the uncoupled system are damped when they two subystems are connected together. On one hand this shows a stabilizing behavior of the larger more complex unified system. On the other hand, by breaking a thriving metaecosystem, one may risk not only to wipe out the predator's population from the harsher environment, but also

 ISBN: 978-84-614-6167-7

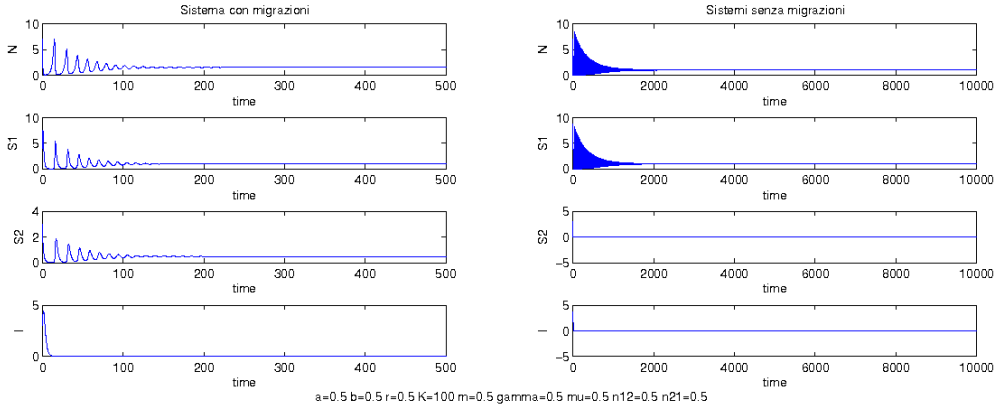


Figure 2: The disease-free equilibrium E_3 is attained for the parameter values .

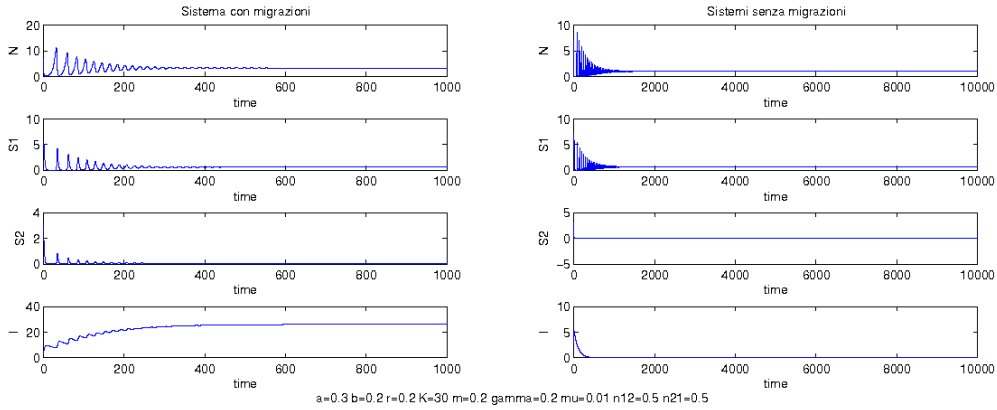


Figure 3: The coexistence equilibrium E_4 is attained for the parameter values .

 ISBN: 978-84-614-6167-7

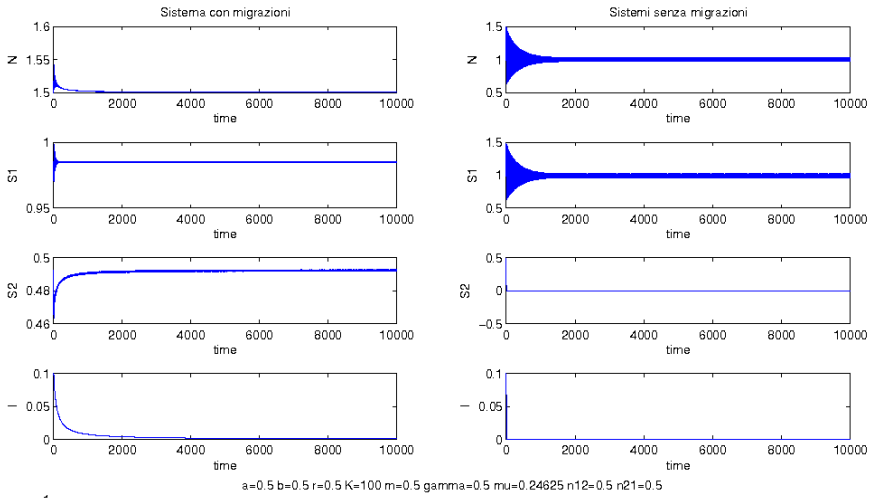


Figure 4: Limit cycles are obtained for the parameter values $a = 0.5$, $b = 0.5$, $r = 0.5$, $K = 100$, $m = 0.5$, $\gamma = 0.5$, $\mu = 0.24625$, $n_{12} = 0.5$, $n_{21} = 0.5$.

to trigger persistent oscillations in the predator-prey environment. It is well-known that sometimes these oscillations in view of stochastic environmental fluctuations in particularly unfavorable circumstances may possibly lead to the disappearance of one or both populations. This is therefore one of the risks in which one delves by tampering with the natural environment. A similar behavior is obtained in Figure 5, but here also the infected individuals survive in the uncoupled system. In Figure 6 we find instead that the persistent oscillations arise in both coupled and uncoupled systems.

In Figure 7 we find that by coupling the two patches, the total prey settle to a different value than the one they have in the uncoupled system, the former being about 50% higher than the latter. Figure 8 shows instead that in the coupled ecosystem both species thrive, while in the separate patches the predators are wiped out. This is another possible consequence to be weighted before undertaking activities that may change the physical shape of the territory. On the other hand, if the predators are a population that needs to be fought, the measure of breaking a larger environment into smaller patches could be a suitable measure to get rid of it. In Figure 9 however, predators vanish in both types of environments.

References

[1] O. Arino, A. El Abdllaoui, J. Mikram, J. Chattopadhyay, *Infection on prey population may act as a biological control in ratio-dependent predator-prey model*, Nonlinearity **17** (2004) 1101–1116.

 ISBN: 978-84-614-6167-7

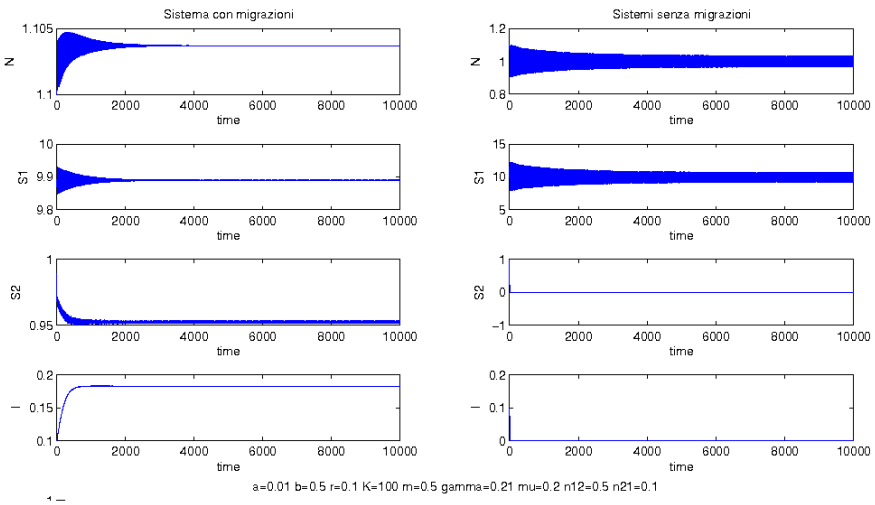


Figure 5: Limit cycles are obtained for the parameter values $a = 0.01$, $b = 0.5$, $r = 0.1$, $K = 100$, $m = 0.5$, $\gamma = 0.21$, $\mu = 0.2$, $n_{12} = 0.5$, $n_{21} = 0.1$.

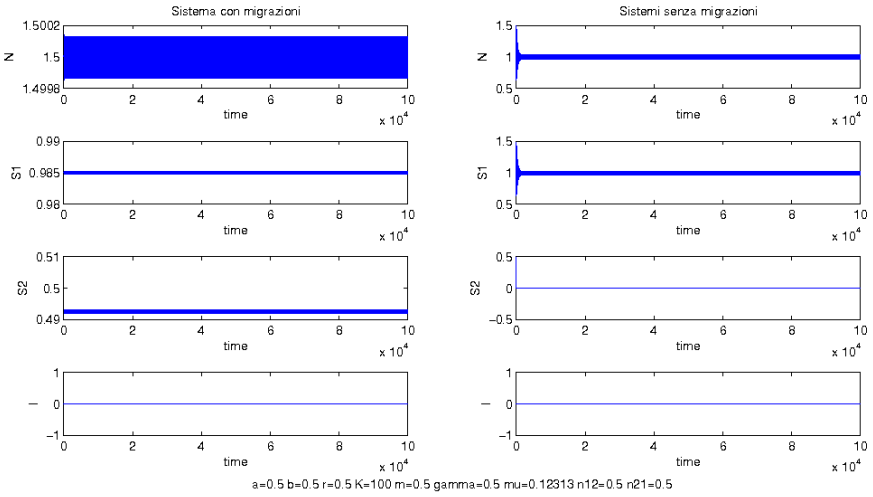


Figure 6: Limit cycles are obtained for the parameter values $a = 0.5$, $b = 0.5$, $r = 0.5$, $K = 100$, $m = 0.5$, $\gamma = 0.5$, $\mu = 0.12313$, $n_{12} = 0.5$, $n_{21} = 0.5$.

 ISBN: 978-84-614-6167-7

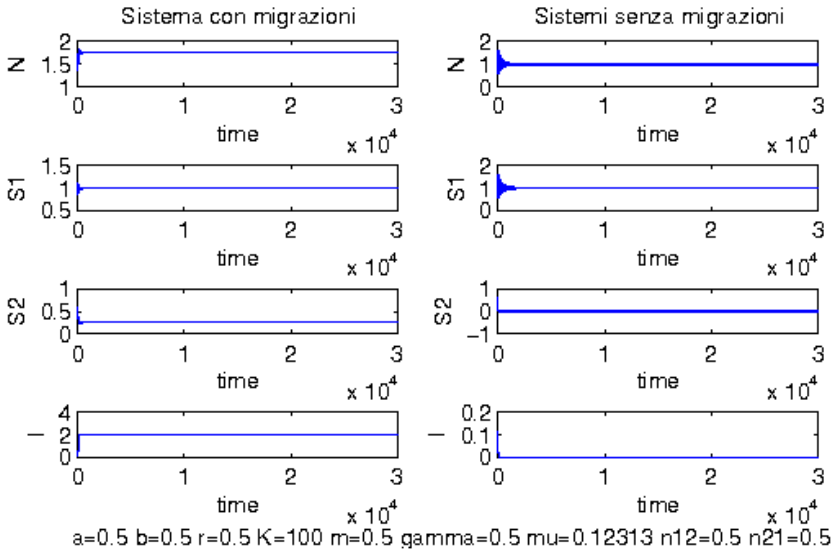


Figure 7: The prey population settles at different levels in the two environments, for the parameter values $a = 0.5$, $b = 0.5$, $r = 0.5$, $K = 100$, $m = 0.5$, $\gamma = 0.5$, $\mu = 0.12313$, $n_{12} = 0.5$, $n_{21} = 0.5$.

[2] P. R. Atsatt, *Lycaenid butterflies and ants: selection for enemy-free space*, The American Naturalist **118** (1981) 638–654.

[3] E. Beltrami, T. O. Carroll, *Modelling the role of viral disease in recurrent phytoplankton blooms*, J. Math. Biol. **32** (1994) 857–863.

[4] H. L. Black, *A North Temperate Bat Community: Structure and Prey Populations*, Journal of Mammalogy **55** (1974) 138–157.

[5] V. Brack, Jr., R. K. LaVal, *Food Habits of the Indiana bat in Missouri*, Journal of Mammalogy **66** (1985) 308–315.

[6] J. Chattopadhyay, O. Arino, *A predator-prey model with disease in the prey*, Nonlinear Analysis **36** (1999) 747–766.

[7] R. J. Cowie, S. A. Hinsley, *Feeding Ecology of Great Tits (Parus major) and Blue Tits (Parus caeruleus), breeding in suburban gardens*, Journal of Animal Ecology **57** (1988) 611–626.

[8] J. T. Cronin, *Movement and spatial population structure of a prairie planthopper*, Ecology **84** (2003) 1179–1188.

 ISBN: 978-84-614-6167-7

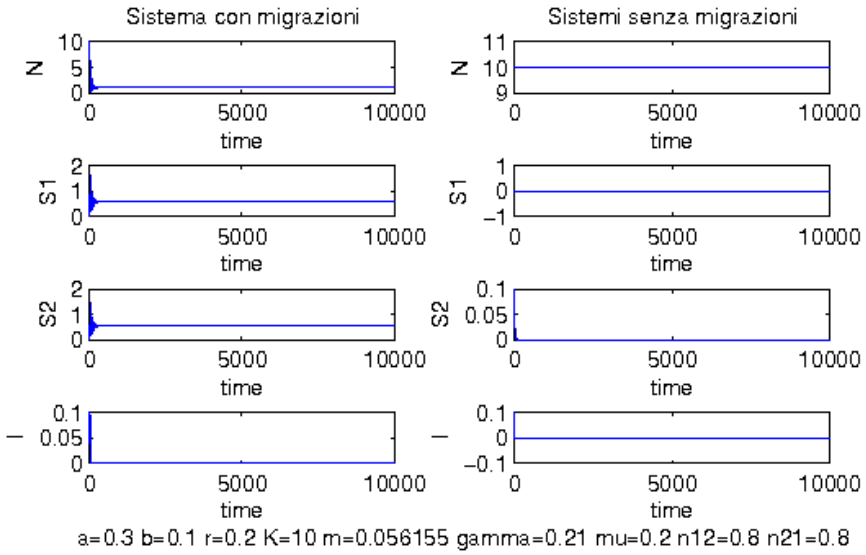


Figure 8: The predators population is wiped out in the uncoupled system, for the parameter values $a = 0.3$, $b = 0.1$, $r = 0.2$, $K = 10$, $m = 0.056155$, $\gamma = 0.21$, $\mu = 0.2$, $n_{12} = 0.8$, $n_{21} = 0.8$.

[9] T. Dewey, L. Ballenger, *"Ovis canadensis" (On-line)*, Animal Diversity Web, Accessed December 04, 2009 at http://animaldiversity.ummz.umich.edu/site/accounts/information/Ovis_canadensis.html.

[10] M. Festa-Bianchet, *Bighorn sheep*, in D.E. Wilson, S. Ruff (Editors) The Smithsonian book of North American mammals, Washington: Smithsonian Institution Press (1999) 348-350.

[11] H. I. Freedman, *A model of predator-prey dynamics as modified by the action of parasite*, Math. Biosci. **99** (1990) 143–155.

[12] I. D. Gauld, *Evolutionary patterns of host utilization by ichneumonoid parasitoids (Hymenoptera: Ichneumonidae and Braconidae)*, Biological Journal of the Linnean Society **35** (1988) 351–377.

[13] E. J. Gustafson, R. H. Gardner, *The effect of landscape heterogeneity on the probability of patch colonization*, Ecology **77** (1996) 94–107.

[14] R. J. Gutiérrez, S. Harrison, *Applying metapopulation theory to spotted owl management: a history and critique*, in D. R. McCollough (Ed.) Metapopulations and wildlife conservation, Washington: Island Press, (1996) 167–185.

 ISBN: 978-84-614-6167-7

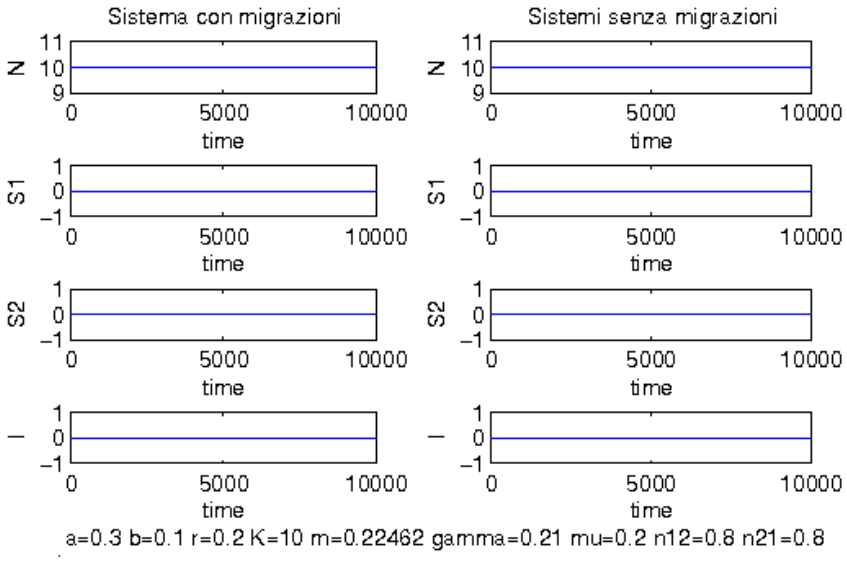


Figure 9: The predators population is wiped out in the both coupled and uncoupled systems, for the parameter values $a = 0.3$, $b = 0.1$, $r = 0.2$, $K = 10$, $m = 0.22462$, $\gamma = 0.21$, $\mu = 0.2$, $n_{12} = 0.8$, $n_{21} = 0.8$.

[15] K. P. Hadeler, H. I. Freedman, *Predator-prey populations with parasitic infection*, J. Math. Biology **27** (1989) 609–631.

[16] I. Hanski, *Single-species spatial dynamics may contribute to long-term rarity and commonness*, Ecology **66** (1985) 335–343.

[17] I. Hanski, M. Gilpin (Ed.s) *Metapopulation biology: ecology, genetics and evolution*, London: Academic Press, 1997.

[18] I. Hanski, A. Moilanen, T. Pakkala, M. Kuussaari, *Metapopulation persistence of an endangered butterfly: a test of the quantitative incidence function model*, Conservation Biology **10** (1996) 578–590.

[19] S. Harrison, A. Taylor, *Empirical evidence for metapopulation dynamics*, in I. Hanski, M. Gilpin (Ed.s) Metapopulation biology: ecology, genetics and evolution. London: Academic Press (1997) 27–42.

[20] E. P. Hoberg, G. S. Miller, E. Wallner-Pendleton, O. R. Hedstrom, *Helminth parasites of northern spotted owls (Strix occidentalis caurina) from Oregon*, Journal of Wildlife Diseases **25** (1989) 246–251.

 ISBN: 978-84-614-6167-7

[21] P. Kareiva, *Population Dynamics in Spatially Complex Environments: theory and data*, Philosophical Transactions of the Royal Society of London B **330** (1990) 175–90.

[22] G. Lei, I. Hanski *Metapopulation structure of Cotesia melitaearum, a parasitoid of the butterfly Melitaea cinxia*, Oikos **78** (1997) 91–100.

[23] R. Levins, *Some demographic and genetic consequences of environmental heterogeneity for biological control*, Bulletin of the Entomological Society America **15** (1969) 237–240.

[24] A. J. Lotka, *Elements of Mathematical Biology*, Dover, New York, 1956.

[25] H. Malchow, S. Petrovskii, E. Venturino, *Spatiotemporal patterns in Ecology and Epidemiology*, Boca Raton: CRC, 2008.

[26] T. R. Malthus, *An essay on the principle of population* J. Johnson in St. Paul's Churchyard, London, 1798.

[27] A. Moilanen, I. Hanski, *Habitat destruction and competitive coexistence in a spatially realistic metapopulation model*, Journal of Animal Ecology **64** (1995) 141–144.

[28] A. Moilanen, A. Smith, I. Hanski, *Long-term dynamics in a metapopulation of the American pika*, American Naturalist **152** (1998) 530–542.

[29] *L. D. Rothman, J. H. Myers*, *Debilitating effects of viral diseases on host lepidoptera*, J. of Invertebrate Pathology **67** (1996) 1–10.

[30] M. J. Scoble, *The lepidoptera: form, function, and diversity*, Oxford: Oxford University Press, 1992.

[31] D. Tilman, P. Kareiva, (Ed.s) *Spatial ecology*, Princeton: Princeton University Press, 1997.

[32] E. Venturino, *The influence of diseases on Lotka Volterra systems*, Rocky Mountain Journal of Mathematics **24** (1994) 381–402.

[33] E. Venturino, *Epidemics in predator-prey models: disease in prey*, in O. Arino, D. Axelrod, M. Kimmel, M. Langlais (Ed.s) Mathematical Population dynamics 1: Analysis of heterogeneity, Winnipeg: Wuertz (1995) 381–393.

[34] E. Venturino, *The effects of diseases on competing species*, Math. Biosc. **174** (2001) 111–131.

[35] E. Venturino, *Epidemics in predator-prey models: disease in the predators*, IMA Journal of Mathematics Applied in Medicine and Biology **19** (2002) 185–205.

[36] E. Venturino, *How diseases affect symbiotic communities*, Math. Biosc. **206** (2007) 11–30.

 ISBN: 978-84-614-6167-7

[37] E. VENTURINO, *Simple metaecoepidemic models*, Bulletin of Mathematical Biology **73** (2011) 917–950.

[38] P. F. VERHULST, *Notice sur la loi que la population suit dans son accroissement*, in *Correspondance Mathématique et Physique* Publiée par A. QUÉTELET **10** (1838) 113-121.

[39] P. F. VERHULST, *Recherches mathématiques sur la loi d'accroissement de la population*, Mém. Acad. Roy. Bruxelles **18** (1845).

[40] P. F. VERHULST, *Recherches mathématiques sur la loi d'accroissement de la population*, Mém. Acad. Roy. Bruxelles **20** (1847).

[41] V. VOLTERRA, U. D'ANCONA, *La concorrenza vitale tra le specie dell'ambiente marino*, VIIe Congr. Int. acquicult et de pêche, Paris (1931) 1-14.

[42] J. A. WIENS, *Wildlife in patchy environments: metapopulations, mosaics, and management*, in D. R. McCullough (Ed.) Metapopulations and Wildlife Conservation, Washington: Island Press (1996) 53–84.

[43] J. A. WIENS, *Metapopulation dynamics and landscape ecology*, in I. A. Hanski, M. E. Gilpin (Ed.s), Metapolulation BiologyEcology Genetics and Evolution San Diego: Academic Press (1997) 43-62.

[44] R. T.WILLIAMS, J. D.DUNSMORE,I.PARER, *Evidence for the Existence of Latent Myxoma Virus in Rabbits (Oryctolagus cuniculus (L.))*, *Nature* **238** (1972) 99–101.

[45] J. WU, *Modeling dynamics of patchy landscapes: linking metapopulation theory, landscape ecology and conservation biology*, in Yearbook in Systems Ecology (English edition) Beijing: Chinese Academy of Sciences, 1994.

 ISBN: 978-84-614-6167-7

Proceedings of the 11th International Conference on Computational and Mathematical Methods in Science and Engineering, CMMSE 2011 26–30 June 2011.

Segmentation of blood cells images with the use of wavelet denoising and mathematical morphology

Macarena Boix[1] and Begoña Cantó[1]

[1] *Departament de Matemàtica Aplicada, Universitat Politècnica de València*

emails: mboix@mat.upv.es, bcanto@mat.upv.es

Abstract

Image segmentation is one of the fundamental approaches of digital image processing. In this work we propose an efficient segmentation method for medical blood cell images analysis. We propose a method that combines the wavelet transform with the morphological operations for segmentation of significant cells. We show that an efficient computational technique based on a wavelet thresholding technique can be use to smooth the original image while filtering out noise and prepare it for suitable segmentation. We propose an algorithm based in Matlab environment to segment the biomedical image and the results we have achieved with this algorithm are very encouraging.

Key words: image segmentation, Wavelet Transform, blood cell, morphological operations.

MSC 2000: 68, 90.

1 Introduction

Image segmentation is a technique widely employed in many applications such that object detection [7], object tracking [9], image retrieval [1] and medical image [4], for example.

The need for accurate segmentation tools in medical applications is driven by the increased capacity of the imaging devices. Tools, such as segmentation, can aid the medical staff in browsing through such large images by highlighting objects of particular importance.

Image segmentation technique is the first step in the analysis image. It is the process of partitioning image pixels connected subsets, called regions, on the basis of homogeneity criteria. That is, with segmentation technique, an image is separated into disjoint regions which the union of them gets the whole picture.

Cell classification has high interest for laboratories. For example, blood cell segmentation is used to study the diagnosis, treatment planning or locate tumors and

 ISBN: 978-84-614-6167-7

other pathologies [2]. Segmentation of blood cell images is an important step for automatic cell analysis, because their final classification depends on the correct use of this technique.

On the recent years, segmentation technique is used to extract useful information on medical images. This technique is a noninvasive pre-processing step for biomedical treatment. For example we cite three of them. An iterative approach based on circular histogram used to detect white blood is given in [8]. An approach for color image segmentation using entropy to determine the threshold is discussed in [6]. Segmentation method of red blood cell using different algorithms based in gray level thresholding, gradient-in method and morphological operators is developed in [5].

This paper is about blood diseases. Some diseases can be detected by analyzing the white blood cells (leukemia, for example). Cells with a disease have a specific structure and we can segment them in order to label them.

The aim of this paper is to segment cell blood images to obtain leukocytes. For that, we show a modified method of the segmentation technique to improve the results desired. In this method we integrate the wavelet transform to binarize the images. Also, we use morphological operations that allow us to improve the segmentation technique of images of blood cells. This algorithm is fully automated, fast and accurate.

We applied this method to 47 images of blood cells with a high success rate. And the experiment shows that this method can obtain the boundary of the desired part of the biomedical images quickly and reliably. It can show isolated cells and obtained the area and perimeter of them. It has practical value in biomedical image analysis.

The set of images used in this work are provided by the UAB Pathology Department PEIR Digital Library and by the webpage of Peter G. Anderson. These databases are selected according to the variety cases included.

In our method to isolate the diseased cells of the blood in order to label them, we use MatLab software. And the procedure is defined as follows.

a) Transforming color image to grayscale image.
To transform RGB image to grayscale image is to calculate gray pixel value I of gray image by 3 components R, G and B of a pixel.

b) Improving contrast of images.
To make this stage we use the histogram equalization. This function involves transforming the intensity values.

c) Binary image processing.
Binary image means that the whole image is only black and white. For that we apply wavelet analysis.

In our case, we apply the universal threshold given by $\sigma\left(\sqrt{2\log\left(m\times n\right)}\right)$, where σ is the estimated noise variance and $m\times n$ is the number of pixels in the image. We use this threshold because is a soft threshold and it has some advantages. For example, it makes algorithms mathematically more tractable (see [3] for more information).

 ISBN: 978-84-614-6167-7

On wavelet families, we have chosen Daubechies (dbN), $N = 1, 2, 3$, Coiflets (coifN), $N = 1, 2, 3$, and Symlets (symN), $N = 2, 3$, being N the order.

The choice of these families is because they share some characteristics such as orthogonality, biorthogonality, compact support, a large number of zero moments and we can make their analysis using the Discrete Wavelet Transform or using the Continuous Wavelet Transform. Next, we chose the adequate level to analysis wavelet.

d) Image processing fundamental operations.
We use two fundamental operations. First, we use the complement of the image. So, in the output image, the objects take the value 1 (white) and the background takes the value 0 (black). Next, we fill the holes in this image.

e) Removing small objects.
When the image is binarized we remove small objects using a morphological technique.

f) Segmented image.
Finally, the algorithm traces region boundaries in the image and label the image for later use. Thus, the image is segmented.

This method is represented, for example, in figure 1.

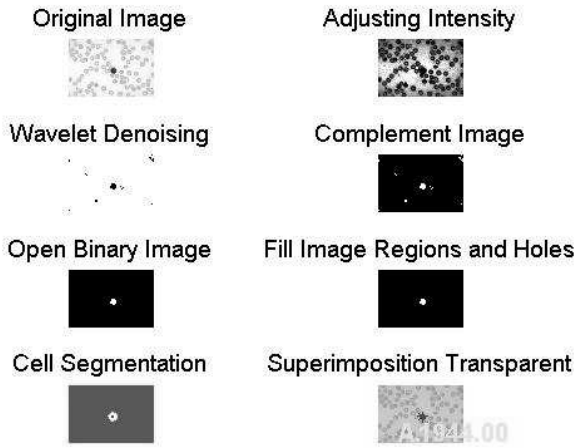


Figure 1: Segmentation of cell blood image that contains a small lymphocyte.

If we apply the above algorithm to a set of images, we can conclude that the accuracy of the segmentation depends strongly on which wavelet is chosen.

 ISBN: 978-84-614-6167-7

We prove that if we use the correct wavelet family and the correct decomposition level, we obtain good results to segment the images. Moreover, we obtain the area and perimeter of the segmented section.

Thus, the algorithm combines automatic threshold selection with the use of morphological operations to segment blood cell images obtaining an excellent result.

Acknowledgements

The authors wish to thank Dr. Nivaldo Medeiros and the UAB Pathology Department PEIR Digital Library (Dr. Peter G. Anderson) since they are the source of all images and the commentaries of them in this paper.

References

[1] C. C. Chiang, Y. P. Hung and G. C. Lee, *A learning state-space model for image retrieval*, IEEE. Trans. Mult. **10**(2) (2008) 1–10.

[2] L. Costrarido, *Medical Image Analysis Methods: Medical-image Processing and Analysis for CAD Systems*, Taylor and Francis, USA, 2005.

[3] D. L. Donoho, *An ideal spatial adaptation by wavelet shrinkage*, Biometrika. **81**(3) (1994) 425–455.

[4] V.Grau, A. U. Mewes, M. Alcáñiz, R. Kikinis, S., K. Warfield, *Improved watershed transform for medical image segmentation using prior information*, IEEE. Trans. Med. Imaging **23**(4) (2004) 447–458.

[5] K. B. How, A. S. Kok Bin, N. T. Siong, K. K. Soo , *Red Blood Cell Segmentation Utilizing Various Image Segmentation Techniques*, Proceedings of International Conference on Man-Machine Systems, Malaysia, 2006.

[6] R. S. Kumar, A. Verma, J. Singh, *Color Image Segmentation and Multi-Level Thresholding by Maximization of Conditional Entrophy*, Int. Sig. Processing **3**(2) (2007) 91–99.

[7] D. Liu, T. Chen, *DISCOV: a framework for discovering objects in video*, Int. Trans. Multimedia **10**(2) (2008) 200–208.

[8] J. Wu, P. Zeng, Y. Zhou, C. Olivier, *A Novel Color Segmentation Method and Its Application to White Blood Cell Image Analysis*, IEEE Proceeding, ICSP 2006, 2006.

[9] J. Y. Zhou, E. P. Ong, C. C. Ko, *Video object segmentation and tracking for content-based video coding*, Proceedings of IEEE International Conference on Multimedia and Expo, ICME 2000, USA, 2000.

Proceedings of the 11th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2011
26–30 June 2011.

Scalability in Parallel Applications with Unbalanced Workload

Jose Luis Bosque[1], Oscar D. Robles[2], Pablo Toharia[2] and Luis Pastor[2]

[1] *Dpto. de Electrnica y Computadores, Universidad de Cantabria, Spain*

[2] *Dpto. de ATC y CCIA , Universidad Rey Juan Carlos, Spain*

emails: joseluis.bosque@unican.es, oscardavid.robles@urjc.es, pablo.toharia@urjc.es, luis.pastor@urjc.es

Abstract

This paper presents a new formulation for the isoefficiency function which can be applied to parallel systems executing balanced or unbalanced workloads. For that purpose, an unbalanced workload model is proposed first. Using this model and the theoretical properties of homogeneous systems with unbalanced workloads, a new proposal for the isoefficiency function is posed. This new formulation permits analyzing the scalability of homogeneous parallel systems under either balanced or unbalanced workloads. Last, the validity of this new metric is evaluated using some synthetic benchmarks. The results produced in these tests allow assessing the importance of considering the specific problems that unbalanced workloads introduce while analyzing the scalability of parallel systems.

Key words: Scalability analysis, isoefficiency, workload imbalance

1 Introduction

Thanks to the multicore architecture, the number of cores available in supercomputers has been dramatically increasing during the last years. The last top500 list (November 2010 [1]) shows 12 machines with more than 100,000 cores. This fact has turned scalability into a factor of increasing importance in the design and implementation of parallel applications, being currently even more important than performance.

On the other hand, it is widely known that load balancing has a deep impact on the performance of a parallel system, and therefore, on its efficiency [3]. Following the isoefficiency function [4], the scalability of a parallel system (computer + application) can be defined based on how much the workload has to be increased so that the efficiency remains constant when the system is scaled up by introducing more processors. Then, if

 ISBN: 978-84-614-6167-7

scalability is defined as a function of the problem size, an interesting aspect to consider is the effect of load imbalance on the system's scalability.

All work done up to now that models scalability on parallel computing systems [4, 2, 6, 5, 7] consider, either implicitly or explicitly, a workload which is perfectly balanced among all of the system nodes. This means that every node receives a workload proportional to its computational power, in such a way that all processors will finish its work simultaneously, if communication overheads are ignored.

This is a non-realistic hypothesis for conventional parallel applications. First, because it means that the workload can be considered continuous and infinitely divisible. Although this assumption can be valid whenever the differences among the work packages assigned to each node fall below certain threshold. And second, because parallel systems have multiple sources of imbalance, such as: (a) a poor initial workload distribution which does not take into account, for example, initial data communication times, and (b) to be non-dedicated systems, in which computational power of nodes can change during the execution of an application. It can happen when additional tasks appear in the system. In order to illustrate the problem, let us consider a parallel application without any communication overhead (i. e., an embarrassing parallel application). Its isoefficiency function should be constant; that is $O(K)$, where $K \in \Re$, and the system should be perfectly scalable. Figure 1 shows the isoefficiency function, that is, the evolution of the workload in front of the number of processors so as to keep the efficiency constant and make the system scalable, labeled as "Theoretical isoefficiency".

However, once the application has been implemented and its isoefficiency function has been experimentally measured, the results obtained can be seen in Figure 1 under the label "Real isoefficiency". The figure shows an isoefficiency function which is linear with respect to the number of processors. It means that to have a scalable system, the problem size has to grow linearly with respect to the number of processors.

Which is the source of this difference between theoretical predictions and real results measured? The answer is simple: the application has a workload showing constant imbalance for all configurations and system sizes evaluated. Then, for improved accuracy the workload imbalance has to be included in the isoefficiency theoretical model.

Hence, it can be seen the importance of imbalance to obtain an accurate scalability model, since not taking imbalance into account yields poor predictions. This paper presents a new expression for the isoefficiency model that includes a model for unbalanced workload, i. e. a more general isoefficiency function that can be applied to parallel systems with or without load balancing, taking into account theoretical properties of the systems. To the author's knowledge, this is the first work that highlights this problem and suggests a suitable solution. Although this result is eminently theoretical, it has a great impact in the design and implementation of parallel applications.

Using the new isoefficiency function, a number of theoretical examples are considered

 ISBN: 978-84-614-6167-7

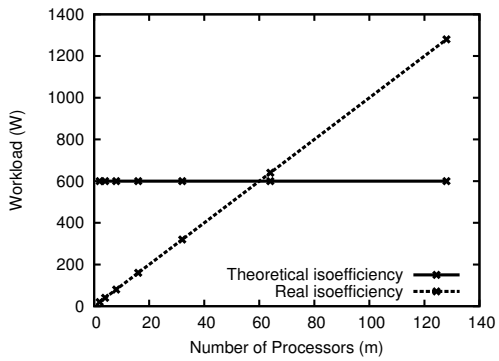


Figure 1: Theoretical and real Isoefficiency function: unbalanced parallel application

in this paper, studying different aspects of the scalability of parallel systems which include the communication overhead as well as the overhead originated by workload imbalance. Finally, an experimental validation has been carried out in order to verify the validity and correctness of the proposed model.

The rest of the paper is organized as follows: Section 2 presents the workload imbalance model and its inclusion into the isoefficiency function. Section 3 analyzes the influence of unbalanced workloads in parallel applications, while Section 4 introduces communication and imbalance overheads. Section 5 shows the experimental evaluation and finally Section 6 shows the conclusions and future work.

2 Scalability of unbalanced parallel systems

2.1 Imbalance workload model

Up to a first approximation, a parallel homogeneous system can be seen as a set of m interconnected nodes, $N = n_1, \ldots, n_m$. If the system is homogeneous, all of the nodes will be identical, offering the same performance. The workload to be executed within a specific parallel application can also be characterized by a set of basic operations which can be executed in parallel; it will be assumed that the workload size is W and that it can be decomposed into a set of computational units.

In order to parametrize the system's performance while executing this application, the computational power of each node within this environment (p) can be defined as the number of basic operations the system is capable of execute per time unit. Similarly, the system's overall computational power (p_t) can be defined as the total number of basic computational operations the whole system can execute per time unit, $p_t = m \cdot p$.

If the workload is continuous and divisible *ad infinitum*, the system can achieve a perfect load balancing by assigning the same workload to every node: $w_i = \frac{W}{m}$. Therefore, assuming that there are not any overheads due to synchronization or communication operations, the

 ISBN: 978-84-614-6167-7

execution time of all of the nodes will remain the same, given by the expression: $T_{CPU} = \frac{W}{p_t} \; \forall \; n_i \; with \; i \; \in 1...m$

However, in real applications, the workload can not be divided continuously, due to the intrinsic grain size of the problem to be solved. This forces to assign an integer number of jobs to each node, leading to an unbalanced load distribution situation not covered by the previous expression, since the computation time is not equal for every node. Moreover, some applications might have a complex data structure that affect grain size, making these differences significantly larger.

In consequence, in unbalanced systems there will be some nodes executing less workload while others process than the optimal. Let's define Δw_i as the difference between the optimal and the actual workload of node i. Note that this value can be both positive or negative, if that node has more or less workload than the optimal. Taking into account this fact, the previous equation can be rewritten as follows:

$$w_i = \frac{W}{m} + \Delta w_i \tag{1}$$

Hence, the execution time of a single node, assuming no communication overhead (i. e., only CPU time), is given by the following expression:

$$T_{CPU} = \frac{\frac{W}{m} + \Delta w_i}{p} \Rightarrow T_{CPU} = \frac{W}{m \cdot p} + \frac{\Delta w_i}{p} \tag{2}$$

Let T_{qi} represent the time needed by processor N_i for computing the additional workload. Then, $T_{qi} = \frac{\Delta w_i}{p}$. Also, the maximum deviation between ideal and real execution times because of the additional workload will be $T_q = \max_{i=1}^{m}\{\frac{\Delta w_i}{p}\}$. The processor achieving this maximum will be the last one to finish its computation, assuming no communication or synchronizing overhead.

The overhead introduced by this fact, as it will be shown later on, can be a constant value, but it can also depend both on the total workload W and on the number of processors m, so it is redefined as $T_q(W, m)$. In the first case, the imbalance introduced will depend on the total size of the problem to be solved, i. e., increasing the size of the problem will lead to an increase or decrease of the imbalance proportional to W. Taking into consideration an overhead time $T_o(W, m)$, the response time will be:

$$T_R = \frac{W}{p_t} + T_q(W, m) + T_o(W, m) \tag{3}$$

It is important to remark that while T_o increases when the number of processors does (or, at least, it remains constant), T_q might decrease when m increases. This is possible because of the way both grain size and number of processors affect imbalance, and in consequence, T_q.

 ISBN: 978-84-614-6167-7

2.2 Isoefficiency function

As explained above, the response time of a parallel system is: $T_R = \frac{W}{p_t} + T_q(W,m) + T_o(W,m)$. Thus, based on the same parameters, the sequential time needed to solve the same problem with the same input size would be: $T_S = \frac{W}{p}$, which is equivalent to solving the problem by executing the whole workload in a single processor with computational power p. Therefore, this concept can be applied to the efficiency expression defined by $E = \frac{T_S}{T_R \cdot m}$ obtaining the following expression:

$$E = \frac{T_S}{T_R \cdot m} = \frac{\frac{W}{p}}{(\frac{W}{p \cdot m} + T_o + T_q) \cdot m} = \frac{\frac{W}{p}}{\frac{W}{p} + m \cdot (T_o + T_q)} = \frac{1}{1 + \frac{m \cdot p \cdot (T_o + T_q)}{W}}$$

Hence, for scalable parallel systems, the efficiency can be kept at a desired value if the ratio $\frac{m \cdot (T_q + T_o)}{W}$ in the expression above can be kept at a constant value. In order to maintain a specific efficiency figure, the following expression gives how large the workload W has to be:

$$\frac{m \cdot p \cdot (T_q + T_o)}{W} = \frac{1 - E}{E} \Rightarrow W = \frac{E}{1 - E} \cdot m \cdot p \cdot (T_q + T_o)$$

Let $K = p \cdot \frac{E}{1-E}$ be a constant (K) depending on the efficiency. Then the isoefficiency function with load imbalance support can be written as:

$$W = K \cdot m \cdot (T_q + T_o) \tag{4}$$

This means the effect of the workload imbalance can be modeled as an additional overhead of the system, the same way as communication time can be dealt with. In this case the factor $T_q = \frac{\Delta w_m}{p}$ is the additional time that the processor that gets the largest workload chunk needs in order to process it in the CPU, i. e. without taking into consideration communication and synchronization times. This factor multiplied by the computational power gives the global wasted time and thus the achieved efficiency.

Moreover, another remarkable fact that can be pointed out from this expression is that in every parallel system, even those which present no communication overheads, the presence of an unbalanced workload means that the system can not be perfectly scaled. As it can be seen, the imbalance is modeled as an additional overhead that has to be added to the other ones (communication and synchronization, typically). Besides, it has an important specific behavior: as overheads always increase with the number of nodes, the imbalance can change the other way around, since an unbalanced system might become more balanced when scaled up.

3 Influence of the imbalance workload in the scalability

Equation 4 means that the imbalance can vary when the system size, the problem size or both of them change. Then, a system would be perfectly scalable if $T_q = T_o = 0$. That is, if

 ISBN: 978-84-614-6167-7

there is not any imbalance and synchronization/communication overhead at all the system is completely scalable, and accordingly it is not needed to increase the size of the problem when the number of processors grows (satisfying that every processor has a part $w_i > 0$ of the workload W).

This section presents some interesting examples of the scalability of parallel systems around the variation of T_q and assuming $T_o = 0$. It is a study of the way the evolution of an unbalanced workload affects the system as its size is scaled. In all cases the starting point is a parallel system defined by $S(m, W)$ and a scaled system $S'(m', W')$, being $T_o = 0$ in both of them. All the assumptions made are useful to isolate the effect of imbalance on the theoretical cases studied.

If each system's workload is distributed among its processors being $T_q = c$, i. e. there is a constant imbalance in both S and S', and it is independent of m and W, then the system is scalable if and only if there is an increase of workload proportional to the number of processors in the system, that is, W is $O(m)$. This situation can happen whenever the problem size is relatively coarse, and the number of available work-packages is not a multiple of the number of processors.

Let us define $T_q = c \in \Re$, constant independent of p and W. In this case the isoefficiency function is:

$$W = K \cdot m \cdot T_q \Rightarrow W = K' \cdot m, \quad \text{where} \quad K' = K \cdot c \tag{5}$$

Then $W = K' \cdot m$, that is the growth of W must be linear to the system's number of processors and computational power, independently of the system's imbalance.

Also, if the imbalance depends on the number of processors in a linear way, that is $T_q = c_1 \cdot m + c_2$, with $c_1, c_2 \in \Re$, and there is not any additional overhead, the system will be scalable if and only if the problem size grows as a function of $O(m^2)$. The proof is straightforward by replacing T_q in Equation 5 for the new expression.

If the imbalance is inversely proportional to the number of processors, that is $T_q = \frac{c_1}{m}$, with $c_1 \in \Re$, then the system will be perfectly scalable since the isoefficiency function remains constant independently of the number of processors that are added to the system, that is W is $O(K)$. It can be noticed that, when $W > p$, this situation brakes the lower bound established by Grama et al. [4] for the isoefficiency function.

Given $T_q = c_1 \cdot W$, with $c_1 \in \Re$, i. e. the imbalance is linearly dependent of the size of the problem, then the system is non scalable, since an increase in the workload size produces also an increase in the imbalance. In consequence, the efficiency can not remain constant. A similar situation occurs when the imbalance can be given as a percentage of the size of the problem.

Finally, if the imbalance is inversely proportional to the size of the problem, that is $T_q = \frac{c_1}{W}$, with $c_1 \in \Re$, it is straightforward to prove that the system will be scalable if an only if the size of the problem grows as a function of $O(\sqrt{m})$. Once again, it brakes the lower bound established by Grama et al. for perfectly balanced systems, since the growth

 ISBN: 978-84-614-6167-7

obtained here is under the linear one.

4 Combining overhead and imbalance

A number of theoretical cases regarding the isoefficiency of unbalanced systems were given in the previous section. These cases did not take into consideration the overhead due to communication or synchronization, being valid only for embarrassing parallel application. A theoretical study of how these two factors affect the overall scalability will be presented in this section.

Given a parallel system $S(m, W)$ and a scaled version of it $S'(m', W')$, let's assume its workload is always distributed proportionally to the number of processors, with an imbalance T_q. In order to analyze its system scalability, its imbalance and overhead behavior have to be modeled. For example, if both the imbalance and the overhead are assumed to be constant and independent of m and W, i.e., $T_q = c_1$ and $T_o = c_2$, being $c_1, c_2 \in \Re$, the system is scalable if and only if the size of the problem W grows proportionally to the number of processors, that is, W is $O(m)$. This can be easily demonstrated using the isoefficiency function previously presented in Section 2.2:

$$W = K \cdot m \cdot (T_q + T_o) \Rightarrow W = K \cdot m \cdot (c_1 + c_2) \Rightarrow W = K' \cdot m$$

where

$$K' = K \cdot (c_1 + c_2)$$

This means that W is $O(m)$, i.e., the computational power of the system must grow linearly with m, independently of the system imbalance, just as if it was perfectly balanced.

Thus, if the system has an overhead due to process communication or synchronization and the imbalance remains constant with size, then the imbalance has no effect on the system's scalability. It only has impact on the maximum scalability the system can achieve but not on the evolution of the efficiency when increasing the system size. Therefore, in this case W has to grow linearly with the number of nodes, which is the lower bound posed by Grama *et al.* [4].

One last remarkable example of this type of *a priori* study would be a system where the imbalance grows inversely with the number of processors ($T_q = \frac{c_1}{m}$), and where the overhead grows linearly with the number of processors ($T_o = c_2 \cdot m$, being c_1 y $c_2 \in \Re$). In this case it can be deduced, using the same expression as previous examples, that W would be $O(m^2)$. Then, in order to get a scalable system, W has to grow quadratically with the number of processors.

 ISBN: 978-84-614-6167-7

5 Model evaluation

A number of experiments was carried out in order to test empirically the validity of the proposed model. The main goals behind the tests were: (1) To verify the hypothesis posed in this paper about the influence of an unbalanced workload in the scalability of parallel systems from an empirical point of view. (2) To validate the scalability model proposed in Section 2 in situations in which there is an unbalanced workload. Also, to verify the properties stated for scenarios with no communication overhead but with unbalanced workloads. (3) To verify the correctness of the model stated in Section 4, by showing how a simple application with communication overheads behaves when the proportion of workload imbalance changes. Altamira, the cluster from the Universidad de Cantabria used in these tests, is composed of 18 eServer BladeCenter, with 256 JS20 nodes (512 processors) linked by a Myrinet network with 1 Gbps of bandwidth.

A number of tests have been performed changing the system and workload sizes. For each of them, 4 repetitions have been done, and response and communication times have been measured. The means of these values have been computed and efficiency values are thus obtained for each system and workload. Then, for every system's size a curve showing the evolution of the efficiency for different workloads has been obtained.

5.1 Validation of imbalance model

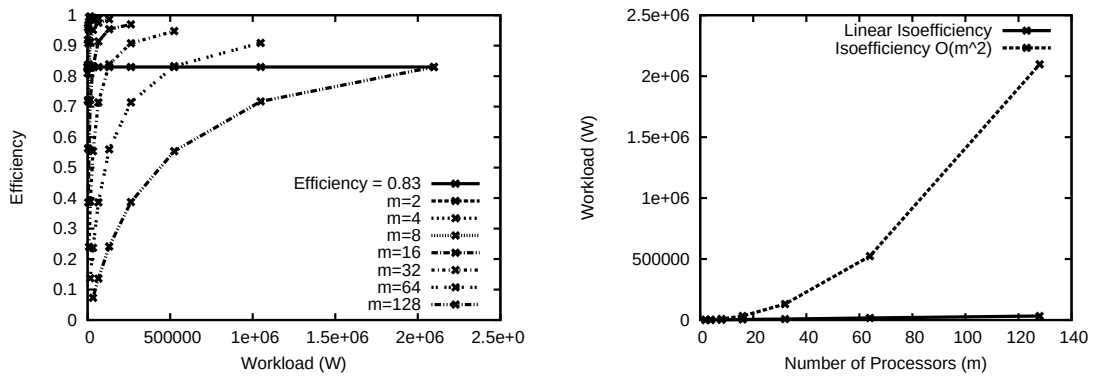
In order to validate the model of scalability in the presence of unbalanced workloads a number of *ad hoc* benchmarks without any communication overhead have been developed. This situation appears quite often in *embarrassing* parallel applications. %

The starting point is a basic benchmark in which each node works on local data and obtains its own solution without any communication or synchronization with the remaining nodes, which means that the communication overhead is null and therefore its efficiency is constant and equal to 1.0, with independence of the workload and the system's number of nodes. Then, this benchmark should be fully scalable since its isoefficiency function is $\mathrm{O}(K)$, with $K \in \Re$. This statement is true if the workload is perfectly balanced, but the isoefficiency function changes when different levels of imbalance are introduced in the system, as the experimental results show.

The experiments have been arranged as follows: First, there is an initial efficiency value which is computed for a specific benchmark on a system with the minimum number of nodes. For all the system configurations used in the tests, this efficiency value should remain constant. This way, it will be shown whether experimental results verify the model proposed in this paper.

All the constants used on the experiments depend on the nodes' computational power and on the problem's nature. They have been measured in the sequential implementation of the benchmarks. Four different workload distributions are considered:

 ISBN: 978-84-614-6167-7



(a) Evolution of the efficiency. (b) Isoefficiency function

Figure 2: Isoefficiency function for *proportional_m* benchmark

- Imbalance proportional to the number of processors, i. e. $T_q = c \cdot m$, where $c \in \Re$. It will be labeled *proportional_m*.
- Imbalance inversely proportional to the number of processors, i. e. $T_q = \frac{c}{m}$, where $c \in \Re$. It will be label-led *inverse_m*.
- Imbalance proportional to the workload, i. e. $T_q = c \cdot W$, where $c \in \Re$. It will be label-led *proportional_w*.
- Imbalance inversely proportional to the workload, i. e. $T_q = \frac{c}{W}$, where $c \in \Re$. It will be label-led *inverse_w*.

Figure 2(a) shows the results obtained in the *proportional_m* test, presenting the evolution of the efficiency as the workload size grows, considering also different system's sizes. It can be observed that the initial efficiency (0.83 in this case) is reached in all cases, showing that the system is scalable. Figure 2(b) shows the isoefficiency obtained from these results, being the lower curve the linear isoefficiency. As it can be seen, the isoefficiency function follows a parabolic curve; in this case the problem size should grow quadratically (W^2). This situation matches the prediction made by the theoretical model in Section 3.

Table 1 shows the results obtained regarding the evolution of the efficiency for the *inverse_m* test. It can be seen that the efficiency remains constant, around 0.50, with independence of the problem's size and the number of nodes. Therefore, the isoefficiency function in this situation is constant and the system is fully scalable, as predicted by the model.

Figure 3 shows the variation of the efficiency in front of the workload for different system's sizes for the *proportional_w* benchmark. In this case it can be seen that given a certain system's size, the efficiency remains almost constant when the workload grows. On the other hand, it can be seen that increasing the system's size it is not possible to keep the efficiency constant, whatever the workload is. Therefore, the isoefficiency function says the system is not scalable.

 ISBN: 978-84-614-6167-7

Table 1: Evolution of the efficiency as a function of the problem's size and the number of nodes (*inverse_m benchmark*).

Workload	2 nodes	4 nodes	8 Nodes	16 Nodes	32 nodes	64 nodes	128 nodes
1024	0.500	0.503	0.499	0.496	0.497	0.495	0.485
4096	0.500	0.503	0.502	0.499	0.498	0.498	0.498
16384	0.500	0.501	0.499	0.501	0.502	0.499	0.496
65536	0.500	0.498	0.499	0.501	0.498	0.500	0.502
262144	0.500	0.501	0.502	0.500	0.500	0.503	0.500

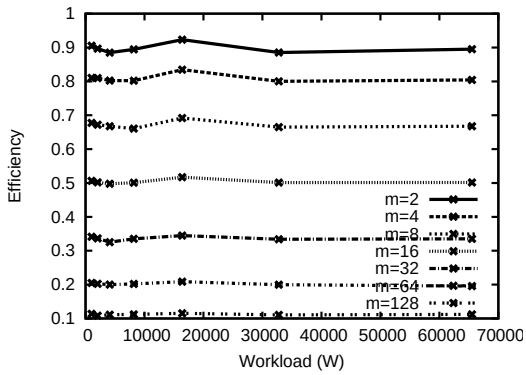
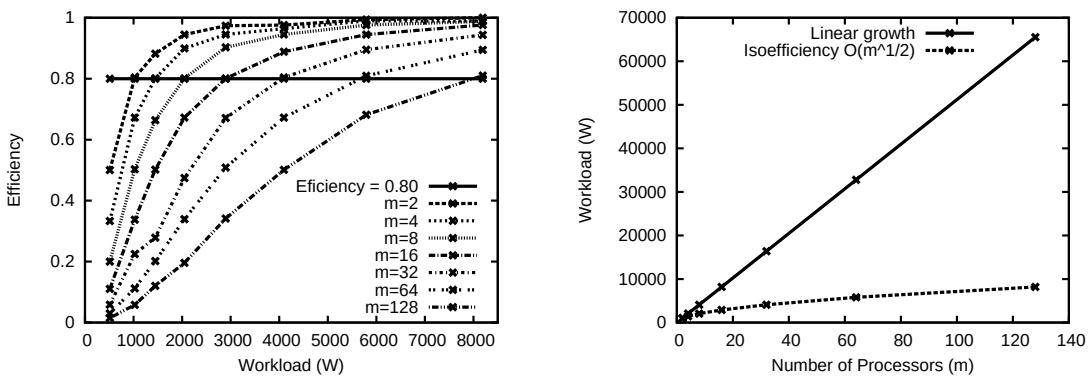


Figure 3: Isoefficiency variation vs. workload for different system's sizes (*proportional_w* benchmark)



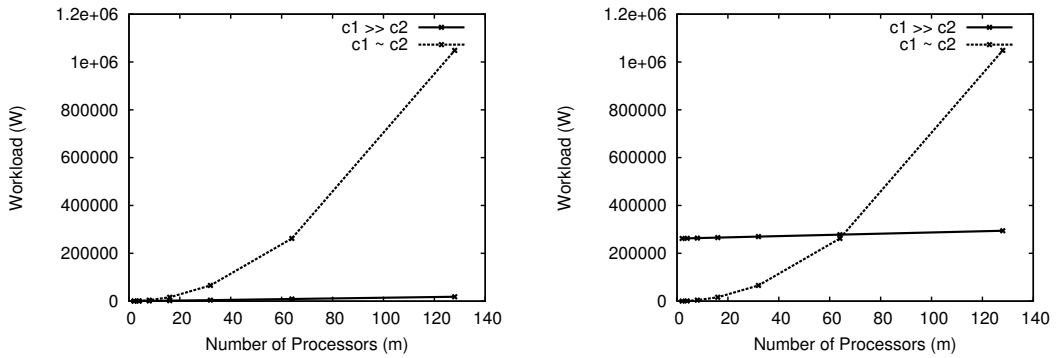
(a) Evolution of efficiency.

(b) Isoefficiency function

Figure 4: Isoefficiency function for *inverse_w* benchmark.

Last, Figure 4 shows the variation of both efficiency (Figure 4(a)) and the isoefficiency function (Figure 4(b)) with an imbalance that changes inversely to the workload (the upper line in Figure 4(a) corresponds to the function of linear isoefficiency). In this case (*proportional_w* benchmark) it can be sen that in all of the tests the initial efficiency is reached, so it can be said that the system is scalable. The isoefficiency function obtained is $O(\sqrt{m})$, under a linear growth.

 ISBN: 978-84-614-6167-7



(a) Constant imbalance ($T_q = c_1$)

(b) Imbalance inversely proportional to the number of processors ($T_q = \frac{c_1}{m}$)

Figure 5: Isoefficiency: benchmark with both communication and imbalance overheads.

Some important conclusions can be extracted from these cases. First, it is clear the great impact that an unbalanced workload has on the scalability of a parallel system. The results obtained allow concluding that the same parallel system can have quite different scalability behaviors, based on the workload distribution, being able to even turn into a non-scalable system, even if it does not have any communication overhead.

On the other hand, the four cases presented here allow to asses the correctness of the model posed in this paper. They have been analyzed from a theoretical point of view in Section 4, and the results obtained match perfectly the theoretical predictions.

5.2 Influence of the imbalance in systems with overhead

Finally, this section presents some experimental results obtained from a benchmark that introduces both imbalance and communication overheads. The communication overhead introduced has been proportional to the number of processors. In this case, the classic isoefficiency function, which does not take into account the workload imbalance, would be $O(m)$. In this paper two different imbalance scenarios are presented as examples:

- Constant imbalance, i.e, $T_q = c_1$, where $c_1 \in \Re$.
- Imbalance inversely proportional to the number of processors, i.e., $T_q = \frac{c_1}{m}$, where $c_1 \in \Re$.

In both cases, the communication overhead is proportional to the number of processors, being $T_o = c_2 \cdot m$. Therefore, taking into account both overhead sources, the scalability model presented in this paper predicts an isoefficiency function $O(m^2)$. Figure 5 shows the collected experimental results showing the isoefficiency function achieved for both cases.

Figure 5(a) shows two results achieved for the isoefficiency function regarding the existing relationship between the values of T_o and T_q. These results show the influence of both c_1 and c_2 constants in the scalability properties. If these values are similar (in the same

 ISBN: 978-84-614-6167-7

order of magnitude) both communication and imbalance overheads have similar weight in the isoefficiency function. Therefore, it can be seen that the isoefficiency function grows quadratically with the number of processors, i.e., it is $O(m^2)$ as the presented model has predicted. Nevertheless, if $c_1 \gg c_2$ the influence of the communication overhead is almost negligible and has no impact on the isoefficiency function. In this case the isoefficiency function is linear (instead of quadratic), although the slope is slightly over 1, that would be the ideal scalability. This behavior is due to the relationship between both overheads when $T_q \gg T_o$, which makes the linear growth of T_o negligible compared to T_q. Hence, the effect it produces to the efficiency is also almost negligible, being the value of T_q the one that actually controls the efficiency behavior, and thus the isoefficiency.

A similar effect can be observed in Figure 5(b) for the case where $T_q = \frac{c_1}{m}$. Again if $c_1 \gg c_2$, the same value controls the behavior of the system and thus, the obtained isoefficiency remains constant. If both c_1 and c_2 are similar the isoefficiency function is quadratic, as predicted by the theoretical model presented here.

These results point out the importance of taking into account the workload imbalance when estimating the scalability of a parallel system. Additionally, it remarks the importance of the constant parameters. In general, when complexity studies are carried out, these parameters are assumed to have a small influence and only "big picture" tendencies are taken into account. Nevertheless, in these cases two different effects have to be added up and it is important not to forget the weight of each of them in the final efficiency value. This weight is determined by the complexity of the expression but also by the constant values. Not taking into account this values might lead to an unexpected system behavior.

6 Conclusions and Future Work

Since long ago there is evidence that unbalanced workloads is one of the aspects that have the biggest impact on the performance of parallel applications. This paper gives the proof for the very first time, as far as the authors know, that this statement can also be applied to scalability. Thus, the main contribution of this paper is that the evaluation of the scalability of a parallel system without considering the workload imbalance leads to potentially erroneous predictions. Although many authors have proposed scalability models before, none of them considers load imbalance.

Once it is clear the great importance of workload imbalance, it has to be included in all the evaluation tools. This paper proposes a simple mathematical model for node imbalance in parallel systems. The model allows considering unbalanced workloads as another overhead in the system, in a conception quite similar to the communication overhead. This way, it is quite simple to introduce this factor in the isoefficiency function, in order to successfully predict the scalability of unbalanced parallel systems.

The new isoefficiency function proposed here makes it possible, as it has been done in this paper, to perform a deep analysis of the influence of imbalance on the scalability

 ISBN: 978-84-614-6167-7

of parallel systems, as well as its relationship with the communication overhead. This way, a number of theoretical analysis have been presented with one remarkable result: if the variation of imbalance is inversely proportional to the workload or to the number of processors, the system's scalability can be under linear. This result brakes the lower bound established by Grama et al. [4] for the isoefficiency function.

Both the model and its application to scalability studies have been experimentally validated using some synthetic benchmarks. In all of the experiments carried out done the correlation between theoretical predictions and empirical results is excellent, and therefore, it can be stated that the model is quite accurate. Another important conclusion from the experiments is the importance of the relative value of both communication and imbalance overheads. Thus, since the isoefficiency function is a complexity analysis, it only collects the function's tendency as the number of processors grows. This is valid if all sources of imbalance are homogeneous and therefore they all present similar constants. But the results achieved show that if one of the overheads is much bigger than the others (for example, an imbalance on the order of seconds while the communication overhead is on the order of milliseconds) then the first one can clearly control the systems' behavior, even when they have hundreds or thousands of nodes.

Finally, the next step is the use of this model with real applications in order to obtain a methodology that facilitates modeling unbalanced workloads as proposed here. Also, an extension of this model to heterogeneous computing systems in which the nodes have different computing and communication capabilities has to be done.

Acknowledgements

This work has been partially supported by the Spanish Ministry of Education and Science (grants TIN2010-21289, TIN2010-21291-C02-02, Consolider CSD2007-00050 and Cajal Blue Brain project) as well as by the HiPEAC European Network of Excellence.

References

[1] The top500 project. November 2010. http://www.top500.org.

[2] J. Chen and V. Taylor. Mesh partitioning for distributed systems. *Proceedings of Seventh IEEE International Symposium on High Performance Distributed Computing*, July 1998.

[3] Ananth Grama, Anshul Gupta, George Karypis, and Vipin Kumar. *Introduction to Parallel Computing (Second Edition)*. PEARSON - Addison-Wesley, Redwood City, CA, 2003.

[4] Ananth Y. Grama, Anshul Gupta, and Vipin Kumar. Isoefficiency: measuring the scalability of parallel algorithms and architectures. *IEEE parallel and distributed technology: systems and applications*, 1(3):12–21, August 1993.

 ISBN: 978-84-614-6167-7

[5] A. Ya. Kalinov. Scalability of heterogeneous parallel systems. *Programming and Computer Software*, 32(1):1–7, 2006.

[6] Luis Pastor and Jose L. Bosque. Efficiency and scalability models for heterogeneous clusters. In *Third IEEE International Conference on Cluster Computing,*, pages 427–434, Los Angeles, California, Octubre 2001. IEEE Computer Society Press.

[7] Y., X.-H. Sun, and M. Wu. Algorithm-system scalability of heterogeneous computing. *Journal of Parallel and Distributed Computing*, 68(11):1403–1412, 2008.

 ISBN: 978-84-614-6167-7

Proceedings of the 11th International Conference on Computational and Mathematical Methods in Science and Engineering, CMMSE 2011 26–30 June 2011.

Memory in mathematical modeling of highly diffusive tumors

João R. Branco[1], José A. Ferreira[2] and Paula de Oliveira[2]

[1] *CMUC & Department of Physics and Mathematics, Coimbra Institute of Engineering, Coimbra, Portugal*

[2] *CMUC, Department of Matehmatics, University of Coimbra, Coimbra, Portugal*

emails: jrbranco@isec.pt, ferreira@mat.uc.pt, poliveir@mat.uc.pt

Abstract

Gliomas are diffusive and highly invasive brain tumors. Even with aggressive surgical resection and radiotherapy and/or chemotherapy, gliomas almost always recur, with fatal consequences. The median survival for patients with glioma doesn't go beyond 1 year. Thus, due to their highly invasive and recurrent behaviour, effective therapeutic strategies for gliomas are extremely important to improve survival time.

Although more clinical trials are necessary to determine the optimal treatment strategies, the development of mathematical models to address this questions is also appropriate and timely. Carefully devised and validated mathematical models might be useful for developing hypotheses to be tested in future clinical trials, and for optimizing the design of future trials.

The aim of this paper is to present an overview on mathematical models for gliomas growth going from the simplest model described by the classical reaction diffusion equation to a complex system characterized by an integro-differential equation where a certain memory effect is introduced.

Key words: Tumor growth, glioma, mathematical modeling, radiation, resection, chemotherapy.

1 Introduction

Cancer is a complex disease which leads to the uncontrolled growth of abnormal cells, destruction of normal tissues and invasion of vital organs. There are different stages at tumor development of varying duration, starting from genetic changes of the cell level and finishing with detachment of metastases and invasion. Tumor cell transport and proliferation are the main contributors to the malignant dissemination ([14]). This

 ISBN: 978-84-614-6167-7

evolution, related to collective or macroscopic behaviour of cells, is described by kinetic approaches ([11]).

Extensive investigations have been done to modeling cancerous growth, especially on solid tumors, in which growth primarily comes from cellular proliferation. However the understanding of malignant gliomas is much less complete, mostly because gliomas proliferates as solid tumors invade the surrounding brain parenchyma actively. Proliferation and specially migration of gliomas provide a significant challenge for modeling.

Gliomas are diffusive and highly invasive brain tumors accounting for about 50% of all primary brain tumors and, unfortunately, the prognosis for patients with gliomas is very poor. Median untreated survival time for high grade gliomas ranges from 6 months to 1 year and even lower grade gliomas can rarely be cured. Theorists and experimentalists believe that inefficiency of treatments results of the highly motile capacity from glioma cells. Additionally gliomas can exhibit very high proliferation rates.

This overview is organized in five sections. In Section 2 we present the most significant mathematical models described in the literature. In Section 3 a mathematical description of of chemotherapy and radiation are introduced. Resection is addressed in Section 4. Plots illustrating the evolution of gliomas obtained using the mathematical models described in the previous sections are presented in Section 5. In Section 6 we present some conclusions.

2 Mathematical modeling of tumor growth

Mathematical modeling is a powerful tool to analyze biological problems allowing researchers to develop and test hypotheses which can lead to a better understanding of the processes involved. Upon comparison with real life results, the models can be modified to more accurately emulate the phenomena. This iterative process of simulating model results and making biological comparisons can continue to the point at which the model suggests appropriate experiments to clarify portions of the biological mechanism not yet understood and to make realistic predictions ([11]).

Cancer research has been a fertile ground for mathematical modeling, beginning with the early concept of simple exponential growth of solid tumors doubling at a constant rate. The introduction of logistic or gompertzian growth (there is increased doubling time and decreased growth fraction as a function of time) allowed to slow the growth in the later stages. With the recognition that tumor cells might spread outside the grossly visible mass, invading locally and metastasizing distantly, and that some cells die during the development process, the mathematical concepts necessarily became more complicated than the ones used in the original simple models for solid tumors.

An exponential growth

The initial answer to the question of how we can measure the growth of an infiltrating glioma was provided by Murray in the early 90s. He formulated the problem as

 ISBN: 978-84-614-6167-7

a conservation law where the rate of change of tumor cell population is define by the summation of the mobility (diffusion) and the net proliferation of tumor cells. Mathematically, this law, for untreated gliomas, can be reasonably quantified by the partial differential equation

$$\frac{\partial c}{\partial t} + \nabla J = f(c)\,, \tag{1}$$

where $c(x,t)$ denotes the tumor cell density at location x and time t, $f(c)$ denotes net proliferation of tumor cells, and ∇ defines the spatial gradient operator. Under the assumption of the classical Fick's law for the diffusion,

$$J = -D\,\nabla c\,, \tag{2}$$

where D is the diffusion coefficient, the model can be written as

$$\frac{\partial c}{\partial t} = \nabla(D\,\nabla c) + f(c)\,. \tag{3}$$

The mathematical model is completed by boundary conditions which impose no migration of cells beyond the brain boundary

$$\nabla c.\eta = 0 \text{ on the boundary},$$

where η denotes the unitary exterior normal to the brain region, and by initial conditions $c(x,0) = c_0(x)$, where $c_0(x)$ defines the initial spatial distribution of malignant cells. Net proliferation rate and invasiveness are defined histologically but practically never defined accurately. As gliomas consists of mobile cells that can migrate as well as proliferate, the invasiveness makes it almost impossible to define growth as a classical volume-doubling time, even in the ideal case where as least two scans, magnetics resonance image (MRI) or computational tomography (CT), are analyzed at different times without treatment intervening. In general gliomas are not encapsulated and consequently the boundary between tumor and normal tissue is not sharp and the number of cells in the normal tissue is not determinable.

Brain geometry and diffusion coefficient

In the first stage of the mathematical modeling of gliomas it was assumed that the brain tissue is homogeneous being the diffusion coefficient D, that defines random mobility of the glioma cells, constant and uniform throughout the brain. Tumor growth is generally assumed to be exponential, so that the cell growth term is given by $f(c) = \rho\, c$, where the net proliferation rate ρ is constant. However, logistic and gompertzian growths are also possible and have been considered but found to be unnecessary in the time frames considered for gliomas. The parameters D and ρ can be calculated for an individual tumor from two pre-treatment MRIs, eliminating the estimation and exploration of large parameter domains, and providing an immediately applicable metric for glioma growth and invasion for any given tumor in vivo. Current data show a range of $6 - 324\ mm^2/year$ for D and $1 - 32\ /year$ for ρ ([9]).

An interesting consequence of the basic model assumptions is that the profile of the concentration of tumor cells depends on the ratio of the growth rate ρ and the diffusion

 ISBN: 978-84-614-6167-7

coefficient D. For ρ/D fixed, the geometry of the tumor growth and invasion remains the same, only the time scale on which the growth and invasion occurs changes. This illustrates the clinical difficulty of using only one MRI/CT observation of the lesion and proceeding with treatment without really knowing the expected pattern of growth of the untreated lesion. When ρ/D is large, the tumor is predominantly growing, so most of the tumor is detectable by an imaging technique. As ρ/D decreases, the relative contribution of the motility of the cells increases, spreading the density profile. The more diffusive tumor is less likely to be accurately identified on medical images. Burgess et al. ([3]) studied this model in three-spatial dimensions with spherical symmetry.

To apply the modeling approach to specific patients, a more realistic look at the brain geometry and structure was necessary. Swanson et al. ([13]) introduced the complex geometry of the brain and allowed diffusion to be a function of the spatial variable x to reflect the observation that glioma cells exhibit higher motility in the white matter than in grey matter ([6]). Mass conservation equation (1) still applied but Fick's law for the mass flux involves a spatially varying diffusion coefficient $D(x)$

$$D(x) = \begin{cases} D_g, & x \text{ in grey matter} \\ D_w, & x \text{ in white matter}\,, \end{cases} \tag{4}$$

where D_g and D_w are constants such that $D_w > D_g$. Estimates of the difference in the diffusion coefficients in grey and white matter have ranged from 2 to 100 fold ([13]).

To accurately analyze the dynamics of this model under the influence of the heterogeneous structure of the human brain, a detailed description of the grey and white matter distribution throughout the brain was necessary. This was made possible by the neuro-anatomical atlas available on the BrainWeb database ([4]).

To compare the numerical simulation obtained from the mathematical modeling with practical clinical measures and to quantify the effectiveness of treatment, it was necessary to model the concept of survival time. Analysis of observations in live patients and in dead revealed that gliomas are detectable on enhanced CT at an average diameter of 3 cm (based on a sphere of equal volume to the tumor) and fatal at an average diameter of 6 cm. Given estimates of model parameters for a particular virtual patient, the expected survival time could be calculated as the time that it takes to growth from 3 to 6 cm in average diameter.

Since most of the information regarding gliomas of specific patients comes from medical images of various types, it is necessary to translate the model results in terms of their manifestations on CT, MRI, macro and microscopic examinations. No presently available medical image will show the entire tumor, including individual cells, because only that portion of the tumor above the threshold of detection will appear on the image. To approximate the threshold of detection associated with the detectable boundary of a glioma on enhanced CT, postmortem microscopic analysis of the patient's brain with the treated glioma was compared with the apparent tumor edge defined on enhanced CT, producing an estimate of the threshold of detection on enhanced CT of 400 cells/mm^2.

Tumor cell's migration

 ISBN: 978-84-614-6167-7

Since the tumor cell's migration is the most critical feature of brain cancer, causing treatment failure, the transport process has to be properly understood. In [5] Fedotov and Iomin proposed an alternative approach for the migration-proliferation dichotomy, employing a two-component continuous time random walk, assuming that the glioma cells are of two phenotypes. In state 1 (migratory phenotype) the cells randomly move but there is no cell fission. In state 2 (proliferation phenotype) the cancer cells do not migrate and only proliferation takes place. The exact mechanism of switching between the two phenotypes is not known. Many of the proposed models are too complex because the switching mechanism involves many parameters. Fedotov and Iomim proposed a stochastic approach for the proliferation-migration switching involving only two parameters ([5]). They assumed that a cell of type 1 remains in state 1 during a waiting time τ_1 and then switches to a cell of type 2. After a waiting time τ_2, spent in state 2, it switches back to a cell of type 1. Both waiting times τ_1 and τ_2 are mutually independent random variables and exponentially distributed with parameters β_1 and β_2, were β_1 is the switching rate from state 1 to 2, while β_2 determines the transition rate from state 2 to 1. Based in the previous considerations, and assuming that the density probability function j associated with the cell jumps of the migratory cells satisfies $\int z^n j(z)dz = 0$ when $n \in \mathbb{N} - \{2\}$ and taking $D = \frac{1}{2}\int z^2 j(z)dz$ the following system partial differential equations

$$\begin{cases} \dfrac{\partial u}{\partial t} = D\displaystyle\int_0^t \alpha(t-s)\frac{\partial^2 u}{\partial x^2}\,ds - \beta_1 u + \beta_2 v \\ \dfrac{\partial v}{\partial t} = \rho v(1-\dfrac{v}{K}) + \beta_1 u - \beta_2 v\,, \end{cases} \tag{5}$$

was proposed in [5]. In (5) $u(t,x)$ and $v(t,x)$ represent the density of the cells of migratory and proliferation phenotypes, respectively, the diffusion coefficient D is determined from the random variable associated with migration jump length, ρ denotes the cell proliferation rate, K represents the carrying capacity of the environment. The kernel $\alpha(t)$ is determined by the probability density of waiting times between jumps ψ in terms of its Laplace transform

$$\mathcal{L}\{\alpha\}(s) = \frac{(s+\beta_1)\ \mathcal{L}\{\psi\}(s+\beta_1)}{1-\mathcal{L}\{\psi\}(s+\beta_1)}\,. \tag{6}$$

In most cases its impossible to find an explicit expression for the memory kernel $\alpha(t)$ for arbitrary choices of waiting time function $\psi(t)$. However, if the random waiting time ψ s given by $\psi(t) = \lambda e^{-\lambda t}$ then $\mathcal{L}(\alpha)(s) = \lambda$. Consequently $\alpha(t) = \lambda\delta(t)$ and (5) assumes the form

$$\begin{cases} \dfrac{\partial u}{\partial t} = \lambda D\dfrac{\partial^2 u}{\partial x^2} - \beta_1 u + \beta_2 v \\ \dfrac{\partial v}{\partial t} = \rho v(1-\dfrac{v}{K}) + \beta_1 u - \beta_2 v\,. \end{cases} \tag{7}$$

Otherwise, if ψ is the Gamma distribution, $\psi(t) = \dfrac{\lambda^m t^{m-1} e^{-\lambda t}}{\Gamma(m)}$, then

$$\mathcal{L}(\alpha)(s) = \frac{\lambda^2}{2\lambda + s + \beta_1}.$$

 ISBN: 978-84-614-6167-7

Consequently $\alpha(t) = \lambda^2 e^{-(2\lambda+\beta_1)t}$ and (5) is given by

$$\begin{cases} \dfrac{\partial u}{\partial t} = \lambda D \displaystyle\int_0^t e^{-(2\lambda+\beta_1)(t-s)} \frac{\partial^2 u}{\partial x^2} ds - \beta_1 u + \beta_2 v \\ \dfrac{\partial v}{\partial t} = \rho v(1 - \dfrac{v}{K}) + \beta_1 u - \beta_2 v \,. \end{cases} \tag{8}$$

The first equation of (8) can be deduced considering a modified Fick's law for the mass flux. In fact, replacing (2) by

$$\frac{\partial J}{\partial t} + \frac{1}{\tau} J = -\hat{D} \frac{\partial u}{\partial x} \tag{9}$$

with $\hat{D} = D\tau$ and $\tau = \dfrac{1}{2\lambda + \beta_1}$ we get

$$J(x,t) = -\hat{D} \int_0^t e^{-\frac{t-s}{\tau}} \frac{\partial u}{\partial x} \, ds \tag{10}$$

where $J(x,0) = 0$. Relation (9) is a first order approximation of the equality

$$J(x, t+\tau) = -\hat{D} \frac{\partial u}{\partial x}(x,t) \tag{11}$$

which establish that the mass flux at time t is related with the gradient of the concentration u at a delayed time. This observation means that the system (8) incorporates a certain memory effect presented in the behaviour of migratory cells.

We remark that as the authors observed in [1] and [2], neither Fick's law, as used in (3), or generalized Fick's law , as used in (8), are an accurate description of the tumor evolution. In fact a more accurate description should be obtained by a "mix" of the two diffusion models, as presented in [1]. In this case a generalized model of type

$$\frac{\partial c}{\partial t} = D_1 \int_0^t \Delta c \, ds + \nabla(D_2 \, \nabla c) + f(c) \,. \tag{12}$$

is obtained.

3 Modeling treatment

The most popular treatment used to combat gliomas are chemotherapy and radiation. Treated patients data base is obviously much longer than untreated ones. From this data base realistic parameter estimates can be obtained.

Chemotherapy

Cancer chemotherapy involves the use of drugs to disrupt the cell cycle and so block proliferation. The success of chemotherapy agents varies widely, depending on cell type and the type of drug being used. The effectiveness of a particular drug is

 ISBN: 978-84-614-6167-7

dependent on the concentration of drug reaching the tumor, the duration of exposure and the sensitivity of the tumor cells to the drug.

Tracqui et al. ([15]) incorporated chemotherapy into the spatially homogeneous model equation (3) by introducing cell death as a loss term. If $G(t)$ defines the time profile of the chemotherapy treatments then, assuming a loss proportional to the amount of therapy at a given time, model (3) can be re-written as

$$\frac{\partial c}{\partial t} = \nabla(D\,\nabla c) + \rho\, c - G(t)\, c\,, \tag{13}$$

where

$$G(t) = \begin{cases} k, & \text{when chemotherapy is being administered} \\ 0, & \text{otherwise}\,. \end{cases} \tag{14}$$

Here k describes the rate of cell death due to exposure to the drug. We point out that the diffusion coefficient D can be constant or given by (4).

For a tumor to decrease in size during chemotherapy, k must be larger than the growth rate ρ of the cell population. The model developed by Tracqui et al. ([15]) included drug-sensitive and drug-resistant tumor cell subpopulations. This work strongly suggested that multiple tumor cell subpopulations could be modeled to respond differently to treatment and could be responsible for the treatment failure. El-Kareh and Secomb ([10]) demonstrated that mathematical models accounting for the kinetics of metabolic and cellular processes can effectively lead to a more rational basis optimizing drug administration in chemotherapy treatments.

If we assume that the cancer cells population are divided into two phenotypes and the chemotherapy has the same effect in both cell types, then, for instance, the system (8) is replaced by

$$\begin{cases} \dfrac{\partial u}{\partial t} = \lambda D \displaystyle\int_0^t e^{-(2\lambda+\beta_1)(t-s)} \frac{\partial^2 u}{\partial x^2} ds - \beta_1 u + \beta_2 v - G(t)\, u \\ \dfrac{\partial v}{\partial t} = \rho v(1 - \dfrac{v}{K}) + \beta_1 u - \beta_2 v - G(t)\, v\,. \end{cases} \tag{15}$$

However, if the resistance of the cells is different during the period of administration depending on the phenotype then the system (8) should be modified according this difference.

Radiation

Radiation therapy is used as a treatment for gliomas because of the precision with which it targets the tumor region, and its ability to increase survival as much as two fold. Assessing response to therapy in gliomas has historically focused on visible changes in gross tumor as measured on MRI. Using the classical linear-quadratic model ([8]) for the radiation efficacy, Rockne et al. ([12]) suggested an extension of the basic model to include delivery and effect of radiation therapy.

Rockne *et al* considered the following extension of the classical model (3)

$$\frac{\partial c}{\partial t} = \nabla(D(x)\,\nabla c) + \rho\, c - R(x,t)\, c\,, \tag{16}$$

 ISBN: 978-84-614-6167-7

where $R(x,t)$ represents the effect of XRT at location x and time t and it is given by

$$R(x,t) = \begin{cases} 0\,,\ t \notin \text{therapy}, \\ \\ 1 - S(x,t)\,, t \in \text{therapy}. \end{cases} \tag{17}$$

In (17) S denotes the survival probability of the cancer cells which is given by $S(x,t) = e^{-\gamma_1 n\, d(x,t)\, eff(x,t)}$, where n is the number of dosage fractions, $d(x,t)$ is dosage fraction distribution. The relative effectiveness eff is given by $eff(x,t) = 1 + \frac{d(x,t)}{1+\frac{\gamma_1}{\gamma_2}}$. In the previous definitions, γ_1, γ_2 are measured parameter.

The advantage of this mathematical model lies in the ability to observe tumor response at any point during therapy, and to virtually alter the treatment schedule and dose delivery, options not otherwise available in vivo.

In [12] Rockne *et al* presented numerical simulation of the growth of a virtual tumor under a variety of treatment schedules and dose distributions. The authors also investigated the cancer response to the treatments using different metrics. Based in their results, they suggested that the conventional course of treatment, involving radiation dose administrated on a per day basis, is much more effective than several treatments per day, and that an optimal response is produced by a low frequency and high dose scheme.

To take into account the radiation therapy when it is assumed that the cancer cells are differentiated into two phenotypes, the system (8) for instance should be modified incorporating in each equation the term R, that is, the system (8) is replaced by

$$\begin{cases} \dfrac{\partial u}{\partial t} = \lambda D \displaystyle\int_0^t e^{-(2\lambda+\beta_1)(t-s)} \frac{\partial^2 u}{\partial x^2} ds - \beta_1 u + \beta_2 v - R(x,t)\, u \\ \\ \dfrac{\partial v}{\partial t} = \rho v (1 - \dfrac{v}{K}) + \beta_1 u - \beta_2 v - R(x,t)\, v\,. \end{cases} \tag{18}$$

Clearly, if the radiation effect is different for each type of cells then the previous system should be adapted.

4 Modeling resection

Resection, the surgical removal of an accessible tumor, is one of the usual treatments for gliomas even though it has shown only limited success, because recurrence of tumor growth at the resection boundary is a very common phenomenon. Experimentalists and theoreticians believe that the distantly invaded cells are the responsible for tumor reappearance following surgery. The modeling framework suggests that, since the density of cancerous cells remaining at this after resection is highest at the resection boundary, reappearance at this location seems most likely. Several experimental results support that to extend the range of resection, to apply radiotherapy or other localized treatments is not going to be generally successful due to the very diffusive nature of gliomas. Mathematical modeling shows that at the time of resection many tumor cells have already migrated not only well beyond the margin of the resection region but also

 ISBN: 978-84-614-6167-7

beyond the radiation effects. Although resection may have succeeded in reducing the pressure effects of the bulk tumor, it is the diffusely invaded tumor cells well beyond the margin that continue to grow and migrate and damage the normal brain tissue, ultimately causing death.

To simulate surgical resection, the tumor cell density is set zero inside the resection bed. The basic conservation law (1), that is (3) with diffusion coefficient D constant or non constant given by (4) still applies before and after resection. If we consider different phenotypes in the cells cancer the we should consider (8).

The model supports the concept that gliomas infiltrate so extensively that they cannot be cured by resection alone. Increasing the size of the resection does increase life expectancy but significantly. Besides the minimal increase in life expectancy, the model has made no effort to differentiate eloquent regions of the brain that must be spared during such surgical procedures. The theoretical analysis combined with the reality of human brain surgery suggests that resection can never be the sole solutions to the treatment of these lesion. Realistic mathematical modeling can be helpful in highlighting and demonstrating the fact that any local treatment of a diffusely invading glioma will fail.

5 Numerical simulation

In this section we present some numerical simulations of the previous models obtained using standard numerical methods. We consider a square homogeneous domain with $[0, 15\,cm] \times [0, 15\,cm]$ with diffusion coefficient $D = 0.05\,cm^2/day$, exponential growth $f(c) = \rho c$ with $\rho = 0.05\,/year$, initial condition defined by 10^6 tumor cells at the point $(7.5, 7.5)$ and we observe the behavior of the tumor cells during the following 100 days. For the $2D$ model we obtain the results presented at Figure 1. We observe a decreasing on the highest values of the tumor cells concentration at initial times followed by an increasing and very intense spreading of cells.

In Figure 2 we plot numerical solutions when chemotherapy is administered with a protocol of one cycle between days 20 and 34, with $G = 0.15\,Gy$. The results are as expected: between day 20 and 34 we observe a decreasing on the tumor cell concentration. The increasing behaviour of tumor cells restarts when the drug administration stopped. The numerical experiments show us that this behaviour occurs for each chemotherapy application. From figures 1 and 2, at day 50, we observe that chemotherapy application leads to reduction of tumor cells although we do not observe a significant reduction of tumor's area.

Finally, in Figure 3 we present the numerical simulations obtained with models (7) and (8). Here we considered $\beta_1 = 10^{-6},\ \beta_2 = 0.036$ and $2\lambda + \beta_1 = 1$. At initial times the migratory cells defined by the classical model (7) presents a delay when compared with the same phenotype defined by the non Fickian model (8). At large times we observe an inversion on the behaviour of the two models, that is, migratory phenotype cells defined by the non Fickian model (8) presents a higher concentration levels than the migratory phenotype cells defined by the Fickian model (8).

 ISBN: 978-84-614-6167-7

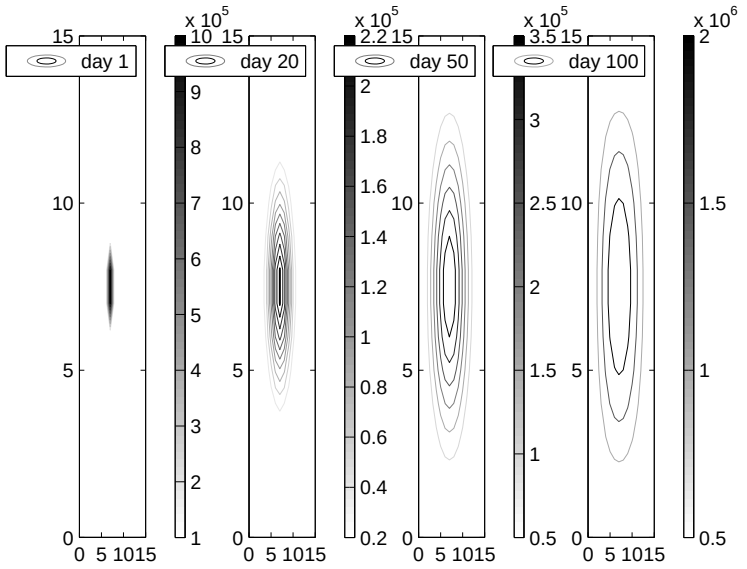


Figure 1: Numerical results obtained with 2D model (3).

6 Conclusions

In this paper we presented a review of some of the most significant mathematical models of brain tumors, specially on gliomas. To improve accuracy mathematical models should be ideally include a huge array of complex processes to simulate the evolution of tumors. However compromise should exist between the complexity of the model and a large set of parameters that cannot be directly measured. The models presented have the advantage of needing a relatively small number of parameters and input data necessary to run the model (D, ρ and 2 pretreatments MRI). The mathematical basis of all models is the mass conservation equation (1) combined with Fick's law (2) for the classical models and modified Fick's law (9) for models presenting memory effect. To include chemotherapy and radiation therapies the diffusion models were adapted modifying the reaction terms.

References

[1] A. Araújo, J.R. Branco, J.A. Ferreira, 2009, *On the stability of a class of splitting methods for integro-differential equations*, Applied Numerical Mathematics, Vol.59, 436-453.

[2] J.R. Branco, J.A. Ferreira, P. de Oliveira, 2007, *Numerical methods for the generalized Fisher-Kolmogorov-Petrovskii-Piskunov equation*, Applied Numerical Mathematics, Vol.57, No.1, 89-102.

 ISBN: 978-84-614-6167-7

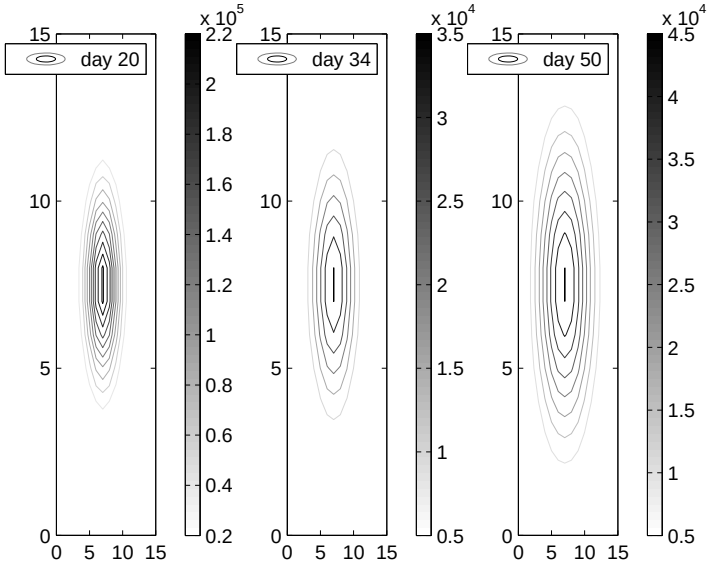


Figure 2: Numerical results obtained with 2D model (13)-(14) and $G = 0.15$.

[3] P.K. Burgess, P.M. Kulesa, J.D. Murray, E.C. Alvord Jr, 1997, *The interaction of growth rates and diffusion coefficients in a three-dimensional mathematical model of gliomas*, Journal of Neuropathology and Experimental Neurology, 56:704-713.

[4] C.A. Cocosco, V. Kollokian, R.K.S. Kwan, A.C. Evans, 1997, *Brainweb: online interface to a 3D MRI simulated brain database*, Neuroimage, 5:S425.

[5] S. Fedotov, A. Iomin, 2007, *Migration and proliferation dichotomy in tumor-cell invasion*, Physical Review Letters, 98:118110(1)-(4).

[6] A. Giese, L. Kluwe, B. Laube, H. Meissner, M. Berens, M. Westphal, 1996, *Migration of human glioma cells on myelin*, Neurosurgery, 38:755-764.

[7] S. Habib, C. Molina-París, T. Deisboeck, 2003, *Complex dynamics of tumors: modeling an energing brain tumor system with coupled reaction-diffusion equations*, Physica A, 327:501-524.

[8] E. Hall, 1994, *Radiobioly for the radiologist.*, Lippincott, Philadelphia, 478-480.

[9] H.L. Harpold, E.C. Alvord Jr, K.R. Swanson, 2007, *The evolution of mathematical modeling of glioma proliferation and invasion*, Journal of Neuropathology and Experimental Neurology, 66:1-9.

[10] A. El-Kareh, T.W. Secomb, 2000, *A mathematical model for comparasion of bolus injection, continuos infusion and liposomal delivery of dozorubicin to tumor cells*, Neoplasia, 2:325-338.

[11] J.D. Murray, 2002,*Mathematical Biology*, Springer.

 ISBN: 978-84-614-6167-7

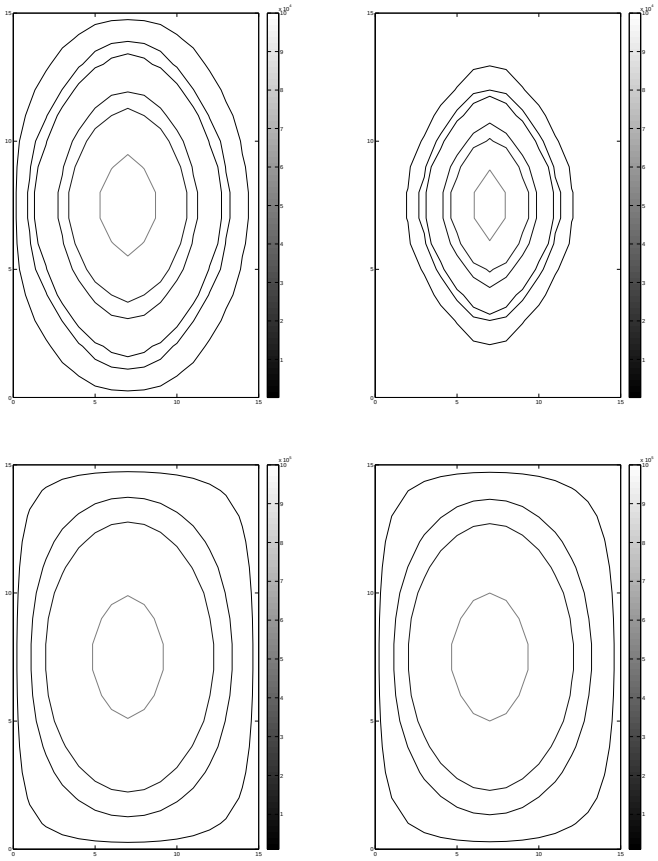


Figure 3: Numerical results obtained with 2D models (7), left, and (8), right, at days 20 (top) and 200 (bottom).

[12] R. Rockne, E.C. Alvord Jr, J.K. Rockhill, K.R. Swanson, 2009, *A mathematical model for brain tumor response to radiation therapy*, Journal of Mathematical Biology, 58:561-578.

[13] K.R. Swanson, E.C. Alvord Jr, J.D. Murray, 2000, *A quantitative model for differential motility of gliomas in grey and white matter*, Cell Proliferation, 2000:317-330.

[14] K.R. Swanson, C. Bridge, J.D. Murray, E.C. Alvord Jr, 2003, *Virtual and real brain tumors: using mathematical modelig to quantify glioma growth and invasion*, Journal of the Neurological Sciences, 216:1-10.

[15] P. Tracqui, G.C. Cruywagen, D.E. Woodward, G.T. Bartoo, J.D. Murray, E.C. Alvord Jr, 1995, *A mathematical model of glioma growth: the effect of chemotherapy on spatio-temporal growth*, Cell Proliferation, 28:17-31.

 ISBN: 978-84-614-6167-7

Proceedings of the 11th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2011
26–30 June 2011.

Theoretical and Computational Aspects of Flow Modeling on Graphs: Traffic on Complex Networks

Buslaev A.P.[1], **Lebedev A.A.**[2] **and Yashina M.V.**[2]

[1] *Department of Mathematics, Moscow State Automobile & Road Tech. University*

[2] *Department of Math. Cybernetics, Moscow Tech. University of Communication and Informatics*

emails: apal2006@yandex.ru, lebed@yandex.ru, yash-marina@yandex.ru

Abstract

Traffic flow theory appeared in the 1930s, however it has not yet reached a level that would satisfy both authors and consumers. Thus, the traffic on complex traffic networks remains a stimulus to search for "the traffic jam formula" that would be as simple as a quadratic equation and provide a miracle cure for jams.

Key words: Nonlinear ordinary differential equation on networks, qualitative and imitation methods, flow modeling with control, traffic flow
MSC 2000: AMS codes 34L30, 90B10, 90C35, 60K30

1 Introduction

Active development of the motor transport, a "tsunami wave" of traffic, is overwhelming the planet. In many of the world's largest cities road networks now constantly operate at capacity and traffic jams have become part of everyday life.

Traffic flow theory appeared in the 1930s, however it has not yet reached a level that would satisfy both authors and consumers. Thus, the traffic on complex traffic networks remains a stimulus to search for "the traffic jam formula" that would be as simple as a quadratic equation and provide a miracle cure for jams.

 ISBN: 978-84-614-6167-7

2 Traffic Theories and Experiments

With performance capabilities of computers constantly increasing, incessant attempts are made to perform multi-agent modeling of traffic, [1]. However, to follow the trains of thought of a large number of drivers is apparently impossible. Modeling necessarily involves some reasonable averagings and abstraction; latest advancements in the field of automatic traffic monitoring must also be taken into account.

The overwhelming majority of theoretical works are devoted to describing behaviour of the traffic on a road section: follow-the-leader models, hydrodynamical analogies, cellular automata models, etc., [2].

Research is still going on, as no universally recognized approach to modeling local traffic flow behaviour in the entire spectrum of states, from free to congested, has yet been found. Modern means of measurement reveal the insufficiency of theoretical conceptions of traffic flow even on an elementary carrier, i.e. a single road section [3].

3 NODE Model

If we consider the problem of traffic on a complex city network, it is reasonable to describe the behavior of the flow on an individual road section with a state function, which accounts for the dependence of speed on density, number of lanes and control mode. The mathematical model of a network is a bidirectional graph with corresponding state functions. In the elementary case there is an interchange in each node, and a dynamic system on the graph is considered, with each edge described with a particle density function and each node assigned a Markov mixing matrix.

The simplest cases of the problem, for which exact solutions can be found, were published in [4], and a general study of the qualitative properties of solutions with respect to the stability of stationary points is carried out in [5].

Let G be a directed graph with M nodes and $N = M \cdot (M-1)$ be the maximum possible number of edges (for a complete graph). For each edge a dependence of velocity on density is defined, which is used as the state function $f_i(\rho_i)$. E.g., $f_i(\rho_i) = v_{max\,i}\,(1 - \rho_i/\rho_{max\,i})$, where $\rho_{max\,i}$ is the maximum density on the i-th edge, $0 \le \rho_i \le \rho_{max\,i}$, $v_{max\,i}$ is the maximum velocity, $F_i(\rho_i) = \rho_i f_i(\rho_i)$ is flow intensity (volume).

If l_i is the length of the i-th edge, then the system of ordinary differential equations

$$\frac{d\,(l_i\rho_i)}{dt} = \sum_{j=1}^{N} \alpha_{ij} F_j(\rho_j) - F_i(\rho_i), \quad i = 1, \ldots, N, \tag{1}$$

is called the NODE model. Here $\alpha = (\alpha_{ij})_{i,j=1}^{N}$ is the mixing matrix, i.e. the stochastic matrix that determines how the flow entering a node from the i-th edge is distributed among

 ISBN: 978-84-614-6167-7

all the j-th edges connected in the node, $\sum_{i=1}^{N} \alpha_{i,j} = 1, \quad j = 1, \ldots, N$.

Then $\alpha_{i,j} = 0$, if the i-th and j-th edges do not have a common node. If α is time-independent, then by renormalizing time we can reduce (1) to (2):

$$\frac{d\rho_i}{dt} = \sum_{j=1}^{N} \alpha_{ij} F_j(\rho_j) - F_i(\rho_i), \quad i = 1, \ldots, N. \tag{2}$$

From (2) we obtain the law of conservation of mass for the flow

$$\rho_1 + \cdots + \rho_N \equiv C. \tag{3}$$

A simulation of the dynamic system (2), (3) is planned to present at the Conference, [4].

4 Flow Model with Control (NODEC)

Traffic light system is an essential part of real traffic. However, any traffic lights cause the input flow density to reach maximum while the "stop" signal is on. Therefore the flow on the i-th edge is described with the following state parameters:

y_i is the length of the congestion in front of the traffic lights;
ρ_i is flow density on the part of the edge before the congestion;
$x_i = (l_i - y_i)$ is the length of that part of the edge;
s_i is the intensity of dissipation of the congestion in front of the traffic lights.

It is assumed that control is only carried out in the nodes of the graph; if not, a new node is added.

The problem is stated and some results for simple graphs are given in [5]. " Some theoretical results and the NODEC computer model described here will be presented.

5 The Problem of Identification

Edge state functions, mixing matrices and traffic control in the nodes are model parameters that have to be adjusted and tested for non-stationarity from time to time.

5.1. Each edge of the graph is, in general, a model of a multi-lane road. The outside lanes may be occupied by parked vehicles. With this in mind, state functions cannot be considered time-independent. Centralized traffic monitoring usually is not total. GPS tracking system based on a server–client structure (smartphones) has been developed under the project. The system enables one to establish the necessary parameters of edge state functions, accounting for random events on the road such as accidents, breakdowns, etc.

 ISBN: 978-84-614-6167-7

5.2. For a complex road network model, the mixing matrix in the system (1), (2) is sparse, since the number of nodes amounts to thousands while the number of edges connected by a single node is sure to be under 10. A mixing matrix for an intersection equipped with intelligent monitoring systems such as traffic cameras is recovered in the natural way; as a rule, using special procedures.

For other intersections, signal-controlled as well as uncontrolled, a client–server system and related software were developed. The algorithm is based on synchronized measurement of series of input flow rate vectors $\bar{Q}_{in}$ and output flows $\bar{Q}_{out}$. The interval of generation of flow rate vectors must be long enough so that it is possible to average the stationarity of the mixing matrix on the control period of the signal-controlled intersection. At the same time it has to be short enough so that the non-stationarity of input flow is not overlooked in consecutive series of measurements. The square matrix $\bar{Q}_{in}$ obtained from series of observations must have full rank. Otherwise, regularization methods are applied.

6 Mobile Distributed System for Traffic Synchronization

Behavior of millions of drivers on a road network cannot be accurately modeled unless advanced methods of data exchange between observers, controllers, and objects of observation and control are developed. Modern means of communication, smartphones, permit synchronization of traffic on road networks using special-purpose software and involving existing and proven techniques and tools, from long-term route planning to rigid traffic control using local client–server data exchange in the neighborhood of bottlenecks ("soft traffic lights").

Acknowledgements

This work has been supported by Ministry of Education and Science of the Russian Federation, project No.14.740.11.0397, and grant of RFBR No.11-07-00622-a.

References

[1] Panasuk Y.S., Maluh V.A., Manuilov V.A., Dudinov I.K., Chernyak G.M. *Agent modeling of traffic flows*, Proc. of the 53th Conference of MIPT Moscow (2010) **Part 5** 130–131.

[2] Treiber M., Hennecke A., Helbing D. , *Congested Traffic States in Empirical Observations and Microscopic Simulations*, Physical Review E. **62** (2000) 1805–1824.

[3] B. S. Kerner , *The Phisics of Traffic Empirical Freeway Pattern Features, Engineering Applications and Theory*, Springer-Verlay Berlin Heidelberg, 2004.

 ISBN: 978-84-614-6167-7

[4] Buslaev A.P., Tatashev A.G., Yashina M.V., *On Stability of Flow on a Ring with Three Links*, In book: C. Appert-Rolland etc. (eds) Traffic and Granular Flow 2007, Springer, Berlin Heidelberg New York, (2009) 265–272.

[5] A.I.Nazarov, *On stability of the stationary modes in one system of nonlinear ODE appearing at traffic modeling* Vestnik St-Peterburg University **S.1, V.3**, (2006) 35–42.

[6] A. P.Buslaev, A. A.Lebedev, M. V.Yashina, *Modeling of flows on graphs: theoretical and computational aspects. Part 1. NODE - traffic model*, - M. Publ. MADI, 2011. - 117 p.

[7] A.P.Buslaev, M.V.Yashina, *About flows on a traffic flower with control*, The 2009 World Congress in Computer Science, Computer Engineering, and Applied Computing (WORLDCOMP09). Las Vegas, Nevada, USA (July 13-16, 2009) in Proc. of the 2009 Int.Conf. on Modelling, Simulation and Vizualization, CSREA Press, (2009) p.254-257

 ISBN: 978-84-614-6167-7

Proceedings of the 11th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2011
26–30 June 2011.

Residuated operations in hyperstructures: residuated multilattices

I. P. Cabrera[1], P. Cordero[1], G. Gutiérrez[1], J. Martínez[1] and M. Ojeda-Aciego[1]

[1] *Department of Applied Mathematics, University of Málaga, Spain*

emails: ipcabrera@uma.es, pcordero@uma.es, gloriagb@ctima.uma.es, javim@ctima.uma.es, aciego@uma.es

Abstract

We initiate the exploration of the residuated operations in the framework of hyperstructures. We focus on the case of a multilattice as underlying algebraic structure, introduce the notion of residuated multilattice and study some of its properties, among which we have shown that the idempotency of the monoidal operation characterises the subclass of Heyting algebras.

Key words: hyperstructures, pocrim, multilattices, residuation.

1 Introduction and preliminary definitions

Residuation has a prominent role in the algebraic study of logical systems, which usually are partially ordered sets together with some operations reflecting the properties of the connectives. This work is related to the use of residuated implication in the framework of hyperstructures and fuzzy logic reasoning.

Although the most used structure in this context is that of residuated lattice, there are reasons which suggest to weaken some of its properties, leading to a more general class of algebraic structures for computation. A commonly considered algebraic structure is that of partially ordered commutative residuated integral monoid [2].

Definition 1 *A tuple* $\langle A, \to, *, \top, \leq \rangle$ *is said to be a* partially ordered commutative residuated integral monoid*, briefly a* ***pocrim****, if, for every* $a, b, c \in A$*, the following properties hold:*

 ISBN: 978-84-614-6167-7

- $\langle A, *, \top \rangle$ *is a commutative monoid with neutral element* $\top$
- $\langle A, \leq \rangle$ *is a partially ordered set which is compatible with* $*$ *(i.e.,* $a \leq b$ *implies* $a * c \leq b * c$*) and* $\top$ *is the maximum of* $\langle A, \leq \rangle$
- $\langle A, \leq \rangle$ *has the* residuum property*, that is* $a * c \leq b$ *if and only if* $c \leq a \to b$*.*

If some extra properties hold, we obtain other well-known structures, such as those given below:

Definition 2

- *A pocrim* $\langle A, \to, *, \top, \leq \rangle$ *is said to be a* ***residuated lattice*** *if, in addition,* $\langle A, \leq \rangle$ *is a lattice.*
- *A residuated lattice in which* $*$ *coincides with the meet operation is said to be a* ***Heyting algebra****.*

It is well-known that residuated lattices are considered to be the algebraic structures of substructural logics [8], which are logics without some of the structural rules of logic: weakening, contraction, or associativity.

We focus here on some extensions of the previously defined notions, by considering a partially-ordered set together with two non-deterministic operations which generalize the supremum and the infimum by weakening the restrictions imposed on a (complete) lattice, namely, the "existence of least upper bounds and greatest lower bounds" is relaxed to the "existence of *minimal* upper bounds and *maximal* lower bounds". Specifically, a *multisupremum of* a and b is defined as a minimal element of the set of upper bounds of a and b, we write $a \sqcup b$ to refer to *the set of all the multi-suprema of* a *and* b; the notion of multiinfimum $a \sqcap b$ is introduced similarly. Now, we can proceed with the formal definition of multilattice and related structures.

Definition 3

- *A poset* $(M, \leq)$ *is said to be a* ***multilattice*** *if for all* $a, b, x \in M$ *with* $a \leq x$ *and* $b \leq x$*, there exists*[1] $z \in a \sqcup b$*, such that* $z \leq x$*; and, similarly, for all* $a, b, x \in M$ *with* $a \geq x$ *and* $b \geq x$*, there exists* $z \in a \sqcap b$*, such that* $z \geq x$*.*
- *A multilattice is said to be* ***full*** *if* $a \sqcup b \neq \varnothing$ *and* $a \sqcap b \neq \varnothing$ *for all* $a, b \in M$*.*

[1]Note that the definition is consistent with the existence of two incomparable elements *without* any multisupremum. In other words, $a \sqcup b$, and also $a \sqcap b$, can be empty.

 ISBN: 978-84-614-6167-7

The notion of multilattice was introduced originally by Benado [1], and further studied by Hansen [4], who proposed an algebraic equivalent definition of multilattice. More recently, another algebraic formalisation of the notion of multilattice was introduced in [5,6] as a theoretical tool to deal with some problems in the theory of mechanised deduction in temporal logics. Multilattices arise as well in other research areas, such as fuzzy extensions of logic programming [7]: for instance, one of the hypotheses of the main termination result for sorted multi-adjoint logic programs [3] can be weakened only when the underlying set of truth-values is a multilattice (the question of providing a counter-example on a lattice remains open).

Definition 4 *A **residuated multilattice** is a pocrim whose underlying poset is a multilattice. If, in addition, there exists a bottom element, we say that the residuated multilattice is **bounded**.*

It is convenient to remark that any finite poset is actually a multilattice, hence the only proper examples of pocrims not multilattices have to be infinite. The following example, taken from [9], shows a proper residuated multilattice, in that its carrier is not a lattice.

Example 1 *Let $\mathbb{Z}$, $\mathbb{Z}^-$ and $\mathbb{Z}^+$ denote, respectively, the sets of all integers, of all non-positive integers, and of all non-negative integers. Given $\bot, \top \notin \mathbb{Z}$, a pocrim A with carrier*

$$A = \Big(\{\bot\} \times \mathbb{Z}^+ \Big) \cup \Big(\mathbb{Z}^+ \times \mathbb{Z} \Big) \cup \Big(\{\top\} \times \mathbb{Z}^- \Big)$$

Let $\leq$ be the partial ordering on A depicted in Figure 1 and note that

$$\langle \alpha, i \rangle \leq \langle \beta, j \rangle \quad \textit{iff} \quad i + |\alpha - \beta| \leq j$$

The operation $*$ on A is defined as follows:

$$\begin{aligned}
&x * y = y * x \\
&\langle \top, i \rangle * \langle \top, j \rangle = \langle \top, i + j \rangle && (i, j \leq 0) \\
&\langle \top, i \rangle * \langle \alpha, j \rangle = \langle \alpha, i + j \rangle && (i \leq 0) \\
&\langle \top, i \rangle * \langle \bot, j \rangle = \langle \bot, \max\{0, i + j\} \rangle && (i \leq 0 \leq j) \\
&\langle \alpha, i \rangle * \langle \beta, j \rangle = \langle \bot, \max\{0, i + j + |\alpha - \beta|\} \rangle \\
&\langle \alpha, i \rangle * \langle \bot, j \rangle = \langle \bot, k \rangle * \langle \bot, j \rangle = \langle \bot, 0 \rangle && (0 \leq j, k)
\end{aligned}$$

 ISBN: 978-84-614-6167-7

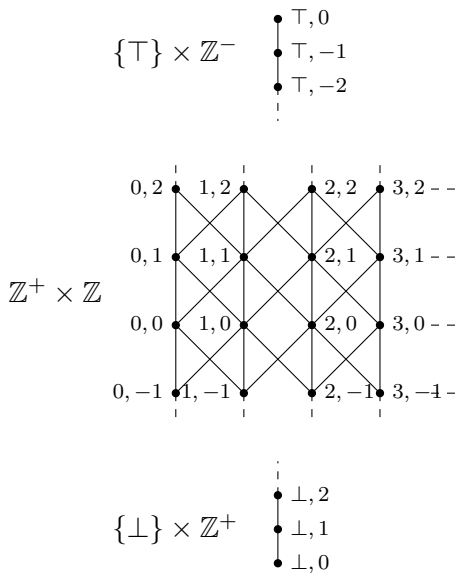


Figure 1: Hasse Diagram of $\langle A; \leq \rangle$

This makes $(A; *, \langle \top, 0 \rangle)$ *to be residuated multilattice when considering the following residuate implication.*

$$x \leq y \quad \textit{iff} \quad x \to y = \langle \top, 0 \rangle$$

$$\langle \top, i \rangle \to \langle \top, j \rangle = \langle \top, \min\{0, j - i\} \rangle \qquad (i, j \leq 0)$$

$$\langle \top, i \rangle \to \langle \alpha, j \rangle = \langle \alpha, j - i \rangle \qquad (i \leq 0)$$

$$\langle \top, i \rangle \to \langle \bot, j \rangle = \langle \bot, j - i \rangle \qquad (i \leq 0 \leq j)$$

$$\langle \alpha, i \rangle \to \langle \beta, j \rangle = \langle \top, \min\{0, j - i - |\alpha - \beta|\} \rangle$$

$$\langle \alpha, i \rangle \to \langle \bot, j \rangle = \langle \alpha, j - i \rangle \qquad (0 \leq j)$$

$$\langle \bot, i \rangle \to \langle \bot, j \rangle = \langle \top, \min\{0, j - i\} \rangle \qquad (0 \leq i, j)$$

2 Algebraic properties of residuated multilattices

We study here some properties of the structures defined above.

Lemma 1 *Every residuated multilattice is full.*

 ISBN: 978-84-614-6167-7

Proof: For all $a, b \in M$ we have that $a, b \leq \top$ and, therefore, $a \sqcup b \neq \varnothing$. Furthermore, $a * b \leq a$, and $a * b \leq b$, hence $a \sqcap b \neq \varnothing$. □

Lemma 2 *Let M be a residuated multilattice, then the following items hold:*

1. $a * b \sqcup a * c = minimals\{a * (b \sqcup c)\}$ *for all* $a, b, c \in M$.
2. $a * (b \sqcap c) \subseteq (a * b \sqcap a * c)\downarrow$ *for all* $a, b, c \in M$.
3. *There exists* $c \in a \sqcap b$ *such that* $a * (a \to b) \leq c$, *for all* $a, b \in M$.
4. *There exists* $c \in a \sqcap b$ *such that* $a * b \leq c$, *for all* $a, b \in M$.

Proof: For item 1, we firstly prove that $a * b \sqcup a * c \subseteq a * (b \sqcup c)$. Let $x \in a * b \sqcup a * c$. Since $a * b, a * c \leq x$, then $b, c \leq a \to x$ and, hence, there exists $y \in b \sqcup c$ such that $y \leq a \to x$ and, thus, $a * y \leq x$. Moreover, by monotonicity of $*$, we have that $a * b \leq a * y$ and $a * c \leq a * y$ and, by definition of $\sqcup$, $x = a * y \in a * (b \sqcup c)$.

Finally, since any element in $a * (b \sqcup c)$ is an upper bound of $a * b$ and $a * c$, the equality $a * b \sqcup a * c = minimals\{a * (b \sqcup c)\}$ holds.

Items 2, 3 and 4 are immediate consequence of basic properties of pocrims and the definition of multilattice.

□

Example 2 *The previous example illustrates the fact that we cannot get rid of the computation of the minimals in the first item of Lemma 2, but $a * b \sqcup a * c \neq a * (b \sqcup c)$ because, for instance,*

$$\langle 0,0\rangle \sqcup \langle 1,0\rangle = \{\langle 0,1\rangle, \langle 1,1\rangle\}$$

$$\langle 2,0\rangle * (\langle 0,0\rangle \sqcup \langle 1,0\rangle) = \{\langle 2,0\rangle * \langle 0,1\rangle, \langle 2,0\rangle * \langle 1,1\rangle\} = \{\langle \bot, 3\rangle, \langle \bot, 2\rangle\}$$

$$\langle 2,0\rangle * \langle 0,0\rangle \sqcup \langle 2,0\rangle * \langle 1,0\rangle = \langle \bot, 2\rangle \sqcup \langle \bot, 1\rangle = \langle \bot, 2\rangle$$

Proposition 1 *Let M be a residuated multilattice such that $a * b \in a \sqcap b$ for all $a, b \in M$, then M is a Heyting algebra.*

Proof: Given $x \in a \sqcap b$, since $x \leq a$, then $x = a \sqcap x = a * x$ and the same for b. Thus $a * b * x = a * x = x$ which implies that $x \leq a * b$. As $x, a * b \in a \sqcap b$, then $x = a * b$. We have obtained that, for all $a, b \in M$, $a * b = a \sqcap b$, in particular, there exists the infimum for all a and b. Being M full (see Lemma 1), there also exists the supremum of a and b, by [5,6]. □

Lemma 3 *Let M be a residuated multilattice with idempotent product, then, for all $a, b \in M$,*

 ISBN: 978-84-614-6167-7

1. *If $x \in a \sqcup b$, then $a * x = a$.*
2. *$a \leq b$ if and only if $a * b = a$.*
3. *$a * b \in a \sqcap b$*

Proof:

1. Observe that $a = a \sqcup a * b = a * a \sqcup a * b = minimals\{a * (a \sqcup b)\}$. If $x \in a \sqcup b$, then $a \leq a * x$. Since, by monotonicity of $*$, $a * x \leq a$, we have $a * x = a$.
2. By monotonicity of the product, if $a \leq b$, then $a*b \leq a*\top = a$ and $a = a*a \leq a*b$ and, hence, $a*b = a$. On the other hand, if $a*b = a$, then $\top = a \to a = a \to a * b \leq a \to b$ which implies $a \leq b$.
3. By item 4 of Lemma 2, there exists $c \in a \sqcap b$ such that $a*b \leq c$, and so $a*b*c = a*b$. On the other hand, from item 2, since $c \leq b$ and $c \leq a$, we have that $a*b*c = a*c = c$. Therefore, $a * b = c \in a \sqcap b$.

□

Theorem 1 *Any idempotent residuated multilattice is a Heyting algebra.*

Proof: It is a direct consequence of the previous lemma and proposition.

□

Sometimes, in connection to an algebraic structure with a binary operation $*$, the following relation so-called *natural preordering* has been considered:

$$a \sqsubseteq b \quad \text{if and only if} \quad a * b = a$$

In the framework of residuated multilattices, the operation $*$ is assumed to be both associative and commutative, and this implies anti-symmetry and transitivity of $\sqsubseteq$. Moreover, this relation is included in $\leq$. That is, $a \sqsubseteq b$ implies $a \leq b$ (it is due to item 4 in Lemma 2). Note, finally, that $\sqsubseteq$ is reflexive if and only if the product is idempotent. Specifically, $\sqsubseteq$ is a partial ordering relation (in a residuated multilattice) exactly in the subclass of Heyting algebras.

Example 3 *Let us consider the meet-semilattice $\langle A; \leq \rangle$ depicted in Figure 2, the product being the meet operator and the residuated implication $\to$ defined by*

$$\begin{array}{ll} x \to y = \top & \textit{iff } x \leq y \\ c_i \to x = x & \textit{for all } x \leq c_i \\ a \to \bot = a \to b = b & \\ b \to \bot = b \to a = a & \end{array}$$

 ISBN: 978-84-614-6167-7

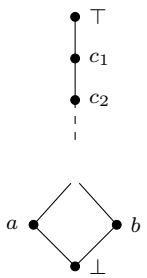


Figure 2: Hasse Diagram of $\langle A; \leq \rangle$

then $\langle A, \rightarrow, *, \top, \leq \rangle$ *is an idempotent pocrim, but it is not a lattice (elements* a *and* b *do not have a supremum) and, hence, is not a Heyting algebra.*

Note that this example shows that, in general, the presence of idempotency in a pocrim is not a sufficient condition to guarantee the structure of Heyting algebra.

3 Conclusions and future work

The algebraic structure of residuated multilattice has been defined between those of partially ordered commutative residuated integral monoids (pocrims) and residuated lattices. All finite pocrims are trivial examples of residuated multilattices, an instance of an infinite pocrim not being a residuated multilattice has been shown.

Preliminary algebraic properties of this new structure have been studied and, specifically, we have shown that the idempotency of the monoidal operation characterises the subclass of Heyting algebras.

Future work will focus on the study of the ideals and filters, which turn out to be specially important in relation to the algebraic semantics of logical systems.

Acknowledgements

Partially supported by projects TIN2009-14562-C05-01 (Science Ministry of Spain), and P09-FQM-5233 (Junta de Andalucía).

References

[1] M. Benado. Les ensembles partiellement ordonnés et le théorème de raffinement de Schreier. I. *Čehoslovack. Mat. Ž.*, 4(79):105–129, 1954.

 ISBN: 978-84-614-6167-7

[2] W. Blok and J. Raftery. Varieties of commutative residuated integral pomonoids and their residuation subreducts. *J. Algebra*, 190:280–328, 1997.

[3] C. Damásio, J. Medina, and M. Ojeda-Aciego. Termination of logic programs with imperfect information: applications and query procedure. *Journal of Applied Logic*, 5(3):435–458, 2007.

[4] D. J. Hansen. An axiomatic characterization of multilattices. *Discrete Math.*, 33(1):99–101, 1981.

[5] J. Martínez, G. Gutiérrez, I. P. de Guzmán, and P. Cordero. Generalizations of lattices via non-deterministic operators. *Discrete Math.*, 295(1-3):107–141, 2005.

[6] J. Martínez, G. Gutiérrez, I. P. de Guzmán, and P. Cordero. Multilattices via multi-semilattices. In *Topics in applied and theoretical mathematics and computer science*, Math. Comput. Sci. Eng., pages 238–248. WSEAS, Athens, 2001.

[7] J. Medina, M. Ojeda-Aciego, and J. Ruiz-Calviño. Fuzzy logic programming via multilattices. *Fuzzy Sets and Systems*, 158(6):674–688, 2007.

[8] H. Ono. Substructural logics and residuated lattices—an introduction. In *50 years of Studia Logica*, volume 50 of *Trends in Logic*, pages 177–212. Kluwer, 2003.

[9] J. Raftery. On the variety generated by involutive pocrims. *Reports on Mathematical Logic*, 42:71–86, 2007.

 ISBN: 978-84-614-6167-7

Proceedings of the 11th International Conference on Computational and Mathematical Methods in Science and Engineering, CMMSE 2011 26–30 June 2011.

Combinatorial structures of three vertices and Lie algebras

J. Cáceres[1], M. Ceballos[2], J. Núñez[2], M.L. Puertas[1] and Á.F. Tenorio[3]

[1] *Dpto. de Estadística y Matemática Aplicada, Universidad de Almería.*

[2] *Dpto. 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: jcaceres@ual.es, mceballos@us.es, jnvaldes@us.es, mpuertas@ual.es, aftenorio@upo.es

Abstract

In this paper, we characterize digraphs of 3 vertices associated with Lie algebras according to isomorphism classes of these associated Lie algebras. At this respect, we introduce and implement two algorithmic methods: the first is devoted to draw the digraph associated with a given Lie algebra and the second allows us to determine if a given digraph is associated or not with a Lie algebra.

Key words: Digraph, Lie algebra, Isomorphism class, Algorithm.
MSC 2000: 17B60, 05C25, 05C20, 05C90, 17B20, 17B30.

1 Introduction

Finding new links among different fields of Mathematics has always been one of the most interesting challenges in mathematical research, since it allows to use alternative techniques to solve open problems, improve known theories and reveal new ones. This paper is devoted to link Lie Theory and Graph Theory. On one hand, research on Graph Theory is running in a high level, being used as a very useful tool to deal with other knowledge fields. Regarding this, this work continues the line opened in [1], where a mapping between Lie algebras and combinatorial structures was introduced in order to translate properties of Lie algebras into the language of Graph Theory and vice versa.

 ISBN: 978-84-614-6167-7

On the other hand, applications of Lie Theory are being considered in fields like Engineering, Physics and Applied Mathematics, for instance, being its research very extensive both theoretically and practically. However, several topics are still unsolved and new alternatives are welcome to work them. In this sense, determining which isomorphism classes there exist for nilpotent and solvable Lie algebras is nowadays an important open problem, especially if we take into account that other types of Lie algebras (like semisimple and simple) were completely classified in 1890.

The main goal of this paper is to make progress in the relation between graphs and Lie algebras, carrying on with previous papers like [1, 2, 3]. The structure is the following: firstly, we determine all isomorphism classes of Lie algebras admitting configurations of 3 vertices described in [1]. In fact, we characterize the different configurations that correspond to the same isomorphism class. Secondly, we introduce and implement two new algorithmic methods based on the relation between graphs and Lie algebras: one to obtain the digraph associated with a given n-dimensional Lie algebra and another to determine if a given digraph is associated with a Lie algebra or not.

In our opinion, the procedures introduced here allow us to advance, make easier and improve the characterization of Lie-algebra isomorphism classes by means of the classification of their associated combinatorial structures (graphs, in this case).

2 Preliminaries of Lie algebras

Some preliminary concepts of Lie algebras are recalled, bearing in mind that the reader can consult [4] for a general overview. In this paper, we consider $\mathbb{K} = \mathbb{R}$ or $\mathbb{C}$ and $\mathbb{K}^* = \mathbb{K} \setminus \{0\}$.

Definition 1 *A* Lie algebra $\mathfrak{g}$ *is a vector space with a second bilinear composition law* $[\cdot,\cdot]$ *called the bracket product, which satisfies two conditions:* $[X,X] = 0$, $\forall X \in \mathfrak{g}$ *and* $[[X,Y],Z] + [[Y,Z],X] + [[Z,X],Y] = 0$, $\forall X,Y,Z \in \mathfrak{g}$. *This last condition is called the* Jacobi identity *and denoted by* $J(X,Y,Z) = 0$.

Definition 2 *The Lie algebra* $\mathfrak{g}$ *is* semisimple *if it does not contain any proper abelian ideals. If* $\mathfrak{g}$ *is non-abelian with no non-trivial ideals, then it is* simple.

Definition 3 *The* commutator central series *and the* lower central series *of a finite-dimensional Lie algebra* $\mathfrak{g}$ *are, respectively,*

$$\mathcal{C}_1(\mathfrak{g}) = \mathfrak{g},\ \mathcal{C}_2(\mathfrak{g}) = [\mathfrak{g},\mathfrak{g}],\ \ldots,\ \mathcal{C}_k(\mathfrak{g}) = [\mathcal{C}_{k-1}(\mathfrak{g}), \mathcal{C}_{k-1}(\mathfrak{g})],\ \ldots \quad \text{and}$$

$$\mathcal{C}^1(\mathfrak{g}) = \mathfrak{g},\ \mathcal{C}^2(\mathfrak{g}) = [\mathfrak{g},\mathfrak{g}],\ \ldots,\ \mathcal{C}^k(\mathfrak{g}) = [\mathcal{C}^{k-1}(\mathfrak{g}), \mathfrak{g}],\ \ldots$$

Hence, $\mathfrak{g}$ *is* $(m-1)$-step solvable *if there exists* $m \in \mathbb{N}$ *such that* $\mathcal{C}_m(\mathfrak{g}) \equiv \{0\}$ *and* $\mathcal{C}_{m-1}(\mathfrak{g}) \neq \{0\}$. *Analogously,* $\mathfrak{g}$ *is* $(m-1)$-step nilpotent *if there exists* $m \in \mathbb{N}$ *such that* $\mathcal{C}^m(\mathfrak{g}) \equiv \{0\}$ *and* $\mathcal{C}^{m-1}(\mathfrak{g}) \neq \{0\}$

 ISBN: 978-84-614-6167-7

3 Associating combinatorial structures with Lie algebras

Given an n-dimensional Lie algebra $\mathfrak{g}$ with basis $\mathcal{B} = \{e_i\}_{i=1}^n$, the law of $\mathfrak{g}$ with respect to $\mathcal{B}$ is given by its structure constants $c_{i,j}^k$ as follows: $[e_i, e_j] = \sum_{k=1}^n c_{i,j}^k e_k$. The pair $(\mathfrak{g}, \mathcal{B})$ can be associated with a combinatorial structure by the following method, introduced in [1]:

a) For each $e_i \in \mathcal{B}$, one vertex labelled as index i is drawn.

b) Given three vertices $i < j < k$, the full triangle can be drawn. The weights $c_{i,j}^k$, $c_{j,k}^i$, and $c_{i,k}^j$ are assigned to the edges $\{i,j\}$, $\{j,k\}$ and $\{i,k\}$, respectively.

   b1) If $c_{i,j}^k = c_{j,k}^i = c_{i,k}^j = 0$, then the triangle is not drawn.

   b2) If a structure constant is zero, its corresponding edge is drawn using a discontinuous line and called *ghost edge*.

   b3) If two triangles of vertices $\{i,j,k\}$ and $\{i,j,l\}$ with $1 \leq i < j < k < l \leq n$ satisfy $c_{i,j}^k = c_{i,j}^l$, then the edge $\{i,j\}$ is shared.

c) Given two vertices $i < j$, draw a directed edge from j to i if $c_{i,j}^i \neq 0$ or from i to j if $c_{i,j}^j \neq 0$.

4 Digraphs of 3-vertices associated with Lie algebras

There exist only 4 digraphs of 3 vertices associated with 3-dimensional Lie algebras according to Lemma 3.1 in [1] (see Figure 1). We study their isomorphism classes.

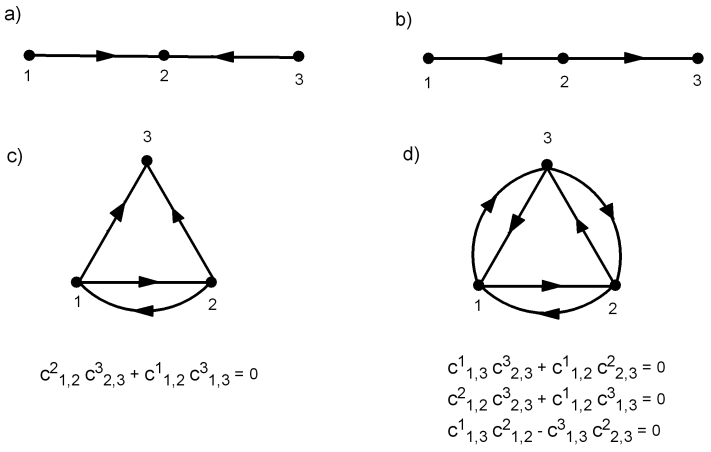


Figure 1: Digraphs of 3 vertices associated with Lie algebras.

 ISBN: 978-84-614-6167-7

Proposition 1 *Let us consider the Lie algebra* $\mathfrak{g} = \langle w_1, w_2, w_3\rangle : [w_1, w_2] = c^2_{1,2}w_2$; $[w_2, w_3] = c^2_{2,3}w_2$, *with* $(c^2_{1,2}, c^2_{2,3}) \in \mathbb{K}^2 \setminus \{(0,0)\}$ *associated with the configuration a) in Figure 1. Then*

i) There exists a basis $\{e_i\}_{i=1}^3$ *of* $\mathfrak{g}$ *with respect to which:* $[e_1, e_2] = e_2$; $[e_2, e_3] = e_2$.

ii) There exists a basis $\{v_i\}_{i=1}^3$ *of* $\mathfrak{g}$ *with respect to which:* $[v_1, v_2] = p_1v_1 + p_2v_2$ *with* $(p_1, p_2) \in \mathbb{K}^2 \setminus \{(0,0)\}$.

Proof: For i), it suffices to consider the basis change $\phi : \mathfrak{g} \to \mathfrak{g}$ given by $e_1 = \phi(w_1) = \frac{1}{c^2_{1,2}}w_1$; $e_2 = \phi(w_2) = w_2$; $e_3 = \phi(w_3) = \frac{1}{c^2_{2,3}}w_3$.

To prove ii), we consider an arbitrary basis change from an arbitrary basis to the basis given in i) as follows: $e_i = \sum_{j=1}^3 a_{i,j}v_j$, with $[v_i, v_j] = \sum_{k=1}^3 d^k_{i,j}e_k$. Imposing the law given in i) and solving the resulting system, we obtain the law expressed in ii). □

Remark 1 *Statement ii) in Proposition 1 means that there exists an isomorphism between the two structures of Figure 2, independently of the weights.*

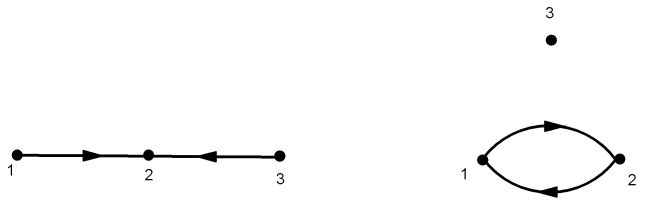


Figure 2: Isomorphism from Proposition 1.

Corollary 1 *Lie algebras associated with configuration a) constitutes a unique isomorphism class* $\mathfrak{g}_1$*:* $[e_1, e_2] = e_2$, $[e_2, e_3] = e_2$. *This class also contains 3-dimensional Lie algebras with center of dimension 1 (associated with a graph having a unique isolated vertex)."*

Proposition 2 *Let us consider the Lie algebra* $\mathfrak{g} = \langle w_1, w_2, w_3\rangle : [w_1, w_2] = c^1_{1,2}w_1$; $[w_2, w_3] = c^3_{2,3}w_3$, *with* $c^1_{1,2}, c^3_{2,3} \in \mathbb{K}^*$ *associated with the configuration b) in Figure 1. Then, there exists a basis* $\{e_i\}_{i=1}^3$ *of* $\mathfrak{g}$ *verifying* $[e_1, e_2] = e_1$; $[e_2, e_3] = pe_3$ *with* $p \in \mathbb{K}^*$.

Proof: It is sufficient to consider the basis change $\phi : \mathfrak{g} \to \mathfrak{g}$ defined by $e_1 = \phi(w_1) = w_1$; $e_2 = \phi(w_2) = \frac{1}{c^1_{1,2}}w_2$; $\phi(w_3) = w_3$ and denote $p = \frac{c^3_{2,3}}{c^1_{1,2}}$. □

Remark 2 *From here on,* $\mathfrak{g}_2(p)$ *with* $p \in \mathbb{K}^*$, *will denote the 3-dimensional Lie algebra of law:* $[e_1, e_2] = e_1$; $[e_2, e_3] = pe_3$, *obtained in Proposition 2.*

 ISBN: 978-84-614-6167-7

Proposition 3 *Let us consider the Lie algebra* $\mathfrak{g} = \langle w_1, w_2, w_3\rangle : [w_1, w_2] = c^1_{1,2}w_1 + c^2_{1,2}w_2$; $[w_1, w_3] = c^3_{1,3}w_3$; $[w_2, w_3] = c^3_{2,3}w_3$, *with* $c^1_{1,2}, c^2_{1,2}, c^3_{1,3}, c^3_{2,3} \in \mathbb{K}^*$ *associated with the configuration c) in Figure 1. Then, there exists a basis* $\{e_i\}_{i=1}^3$ *of* $\mathfrak{g}$ *with respect to which* $[e_1, e_2] = p(e_1 - e_2)$; $[e_1, e_3] = e_3$; $[e_2, e_3] = e_3$, *with* $p \in \mathbb{K}^*$.

Proof: It is sufficient to consider the basis change $\phi : \mathfrak{g} \to \mathfrak{g}$ defined by $e_1 = \phi(w_1) = \frac{1}{c^3_{1,3}}w_1$; $e_2 = \phi(w_2) = \frac{1}{c^3_{2,3}}w_2$; $e_3 = \phi(w_3) = w_3$, denote $p = \frac{c^1_{1,2}}{c^3_{2,3}}$ and keep in mind that $c^1_{1,2}c^3_{1,3} + c^2_{1,2}c^3_{2,3} = 0$, due to the Jacobi identity. □

Remark 3 *From here on,* $\mathfrak{g}_3(p)$ *with* $p \in \mathbb{K}^*$, *will denote the 3-dimensional Lie algebra of law* $[e_1, e_2] = p(e_1 - e_2)$; $[e_1, e_3] = e_3$; $[e_2, e_3] = e_3$, *obtained in Proposition 3.*

Proposition 4 *Let us consider the Lie algebra* $\mathfrak{g} = \langle w_1, w_2, w_3\rangle : [w_1, w_2] = c^1_{1,2}w_1 + c^2_{1,2}w_2$; $[w_1, w_3] = c^1_{1,3}w_1 + c^3_{1,3}w_3$; $[w_2, w_3] = c^2_{2,3}w_2 + c^3_{2,3}w_3$, *with* $c^1_{1,2}, c^2_{1,2}, c^1_{1,3}, c^3_{1,3}, c^2_{2,3}, c^3_{2,3} \in \mathbb{K}^*$ *associated with the configuration d) in Figure 1. Then*

i) There exists a basis $\{e_i\}_{i=1}^3$ *of* $\mathfrak{g}$ *verifying* $[e_1, e_2] = -\frac{p_1}{p_2}(e_1 - e_2)$; $[e_1, e_3] = p_1e_1 + e_3$; $[e_2, e_3] = p_2e_2 + e_3$, *with* $p_1, p_2 \in \mathbb{K}^*$.

ii) There exists a basis $\{v_i\}_{i=1}^3$ *of* $\mathfrak{g}$ *verifying* $[e_1, e_2] = -p(e_1 - e_2)$; $[e_1, e_3] = e_1 + e_3$; $[e_2, e_3] = \frac{1}{p}e_2 + e_3$, *with* $p \in \mathbb{K}^*$.

Proof: For i), it is sufficient to consider the basis change $\phi : \mathfrak{g} \to \mathfrak{g}$ given by $e_1 = \phi(w_1) = \frac{1}{c^3_{1,3}}w_1$; $e_2 = \phi(w_2) = \frac{1}{c^3_{2,3}}w_2$; $e_3 = \phi(w_3) = \frac{1}{c^2_{1,2}}w_3$ and the Jacobi identity $J(e_1, e_2, e_3) = 0$, as well as denoting $p_1 = \frac{c^1_{1,3}}{c^2_{1,2}}$ and $p_2 = \frac{c^2_{2,3}}{c^2_{1,2}}$. Starting from this law and considering the basis change $\psi : \mathfrak{g} \to \mathfrak{g}$ with $e_1 = \psi(w_1) = w_1$; $e_2 = \psi(w_2) = w_2$; $e_3 = \psi(w_3) = \frac{1}{p_1}w_3$, we obtain the law stated in ii) after denoting $p = \frac{p_1}{p_2}$. □

Remark 4 *From here on,* $\mathfrak{g}_4(p)$ *with* $p \in \mathbb{K}^*$, *will denote the 3-dimensional Lie algebra of law* $[e_1, e_2] = -p(e_1 - e_2)$; $[e_1, e_3] = e_1 + e_3$; $[e_2, e_3] = \frac{1}{p}e_2 + e_3$, *obtained in Proposition 4.*

Proposition 5 *The dimension of the derived Lie algebra* $\mathcal{D}(\mathfrak{g}_i) = [\mathfrak{g}_i, \mathfrak{g}_i]$ *is*

$$dim(\mathcal{D}(\mathfrak{g}_i)) = \begin{cases} 1, \text{ if } i = 1; \\ 2, \text{ if } i = 2, 3 \vee (i = 4 \wedge p = 1); \\ 3, \text{ if } i = 4 \text{ with } p \neq 1. \end{cases}$$

Proof: In virtue of Propositions 1, 2, 3 and 4, we only need to study $\mathcal{D}(\mathfrak{g}_4) = \langle -p(e_1 - e_2), e_1 + e_3, \frac{1}{p}e_2 + e_3\rangle$. The coefficient matrix is

$$\begin{pmatrix} -p & 1 & 0 \\ p & 0 & \frac{1}{p} \\ 0 & 1 & 1 \end{pmatrix}$$

whose rank equal to 2 if and only if $p = 1$. □

 ISBN: 978-84-614-6167-7

Corollary 2 *The following two statements are verified*

1. $\mathfrak{g}_1$ *is not isomorphic to* $\mathfrak{g}_2(p)$, $\mathfrak{g}_3(p)$ *or* $\mathfrak{g}_4(p)$, *for* $p \in \mathbb{K}^*$. *Consequently, the last three are not isomorphic to Lie algebras associated with configurations having an isolated vertex.*

2. *Given* $p \in \mathbb{K}^* \setminus \{1\}$, $\mathfrak{g}_4(p)$ *is not isomorphic to* $\mathfrak{g}_2(q)$, $\mathfrak{g}_3(q)$ *either* $\mathfrak{g}_4(1)$, *for* $q \in \mathbb{K}^*$.

Proposition 6 *Given* $p_1, p_2 \in \mathbb{K}^*$ *and* $i \in \{2,3,4\}$, *the Lie algebras* $\mathfrak{g}_i(p_1)$ *and* $\mathfrak{g}_i(p_2)$ *are isomorphic if and only if* $p_1 = p_2$ *or* $p_1 \cdot p_2 = 1$.

Proof: Fixed and given $i \in \{2,3,4\}$, the Lie algebras $\mathfrak{g}_i(p_1)$ and $\mathfrak{g}_i(p_2)$ are isomorphic if and only if there exists a basis change leading from the law of $\mathfrak{g}_i(p_1)$ to the one of $\mathfrak{g}_i(p_2)$. Let $\{e_j\}_{j=1}^3$ and $\{w_j\}_{j=1}^3$ be the bases giving the law of $\mathfrak{g}_i(p_1)$ and $\mathfrak{g}_i(p_2)$, respectively and let us consider a general basis change given by $w_1 = \sum a_{1,j}e_j$, $w_2 = \sum a_{2,j}e_j$, $w_3 = \sum a_{3,j}e_j$. Imposing the law of $\mathfrak{g}_i(p_1)$ over $\mathfrak{g}_i(p_2)$, we obtain a system whose solutions are $p_1 = p_2$ or $p_1 \cdot p_2 = 1$. □

Proposition 7 *Given* $p \in \mathbb{K}^*$, *Lie algebra* $\mathfrak{g}_2(p)$ *is isomorphic to Lie algebra* $\mathfrak{g}_3(p)$.

Proof: The isomorphism is $\phi : \mathfrak{g}_2(p) \to \mathfrak{g}_3(p)$ defined by $w_1 = \phi(e_1) = -e_2 + e_3$; $w_2 = \phi(e_2) = -e_2 - e_3$; $w_3 = \phi(e_3) = e_1$; where $\{e_i\}_{i=1}^3$ and $\{w_i\}_{i=1}^3$ are the respective bases of $\mathfrak{g}_2(p)$ and $\mathfrak{g}_3(p)$. □

Remark 5 *Proposition 7 involves that configurations of Figure 3 are associated with the same Lie algebra for a given p.*

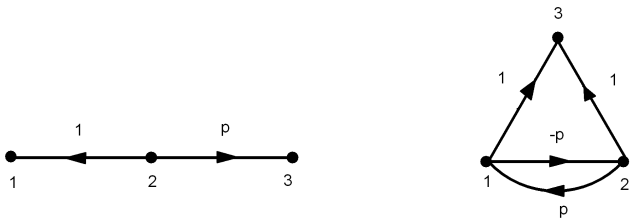


Figure 3: Isomorphism from Proposition 7.

Proposition 8 *Lie algebra* $\mathfrak{g}_2(-1)$ *is isomorphic to Lie algebra* $\mathfrak{g}_4(1)$.

Proof: The isomorphism is given by $\phi : \mathfrak{g}_2(-1) \to \mathfrak{g}_4(1)$, where $w_1 = \phi(e_1) = e_1 - e_2; w_2 = \phi(e_2) = -e_2 + e_3; w_3 = \phi(e_3) = e_1 + e_2 + e_3$. □

 ISBN: 978-84-614-6167-7

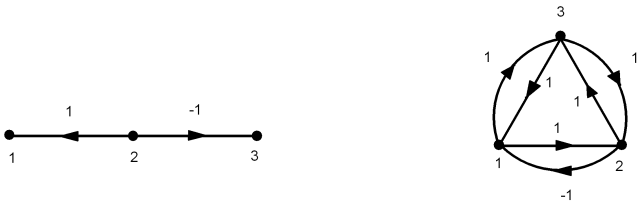


Figure 4: Isomorphism from Proposition 8.

Remark 6 *Proposition 8 establishes that configurations in Figure 4 correspond to the same Lie algebra.*

Proposition 9 $\mathfrak{sl}_2(\mathbb{K})$ *is isomorphic to* $\mathfrak{g}_4(p)$ *if and only if* $p = -1$.

Proof: Given $p \in \mathbb{K}^* \setminus \{1\}$, we consider the Lie algebra $\mathfrak{g}_4(p) = \langle e_1, e_2, e_3 \rangle$ with law $[e_1, e_2] = -p(e_1 - e_2)$; $[e_1, e_3] = e_1 + e_3$; $[e_2, e_3] = \frac{1}{p}e_2 + e_3$ and Lie algebra $\mathfrak{sl}_2(\mathbb{K}) = \langle w_1, w_2, w_3 \rangle$ with law $[w_1, w_2] = 2w_2$; $[w_1, w_3] = -2w_3$; $[w_2, w_3] = w_1$. When defining an arbitrary basis change $e_i = \sum_{j=1}^{3} a_{i,j} w_j$, for $i = 1, 2, 3$ and imposing the laws of $\mathfrak{g}_4(p)$ and $\mathfrak{sl}_2(\mathbb{K})$, we obtain a system of equations such that every solution involves $p = -1$, which concludes the proof. □

Remark 7 *Proposition 9 implies that configurations in Figure 5 comes from the same Lie algebra.*

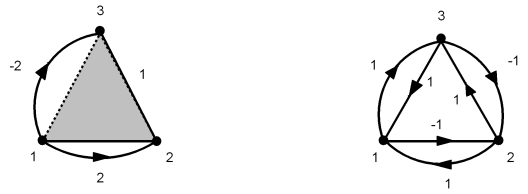


Figure 5: Isomorphism from Proposition 9.

All the previous results can be summarized as follows

Theorem 1 *The isomorphism classes of 3-dimensional Lie algebras are the following*

a) $\mathfrak{g}_1$.

b) $\mathfrak{g}_2(-1) \cong \mathfrak{g}_3(-1) \cong \mathfrak{g}_4(1)$.

 ISBN: 978-84-614-6167-7

c) $\mathfrak{g}_2(p) \cong \mathfrak{g}_2(\frac{1}{p}) \cong \mathfrak{g}_3(p) \cong \mathfrak{g}_3(\frac{1}{p}), \forall p \in \mathbb{K}^* \setminus \{-1\}$.

d) $\mathfrak{g}_4(-1) \cong \mathfrak{sl}_2(\mathbb{K})$.

e) $\mathfrak{g}_4(p) \cong \mathfrak{g}_4(\frac{1}{p}), \forall p \in \mathbb{K}^* \setminus \{-1, 1\}$.

Moreover, the algebras belonging to the first three classes are 2-step solvable and non-nilpotent, while the corresponding with the fourth class is simple.

5 Algorithmic methods

In this section we show two algorithms dealing with converse questions: the first is devoted to obtain the digraph associated with a given Lie algebra starting from its law; and the second is useful to determine if a weighted digraph is associated with a Lie algebra or not.

5.1 Algorithm to obtain the digraph associated with a Lie algebra

Given an n-dimensional Lie algebra $\mathfrak{g}$ with basis $\mathcal{B}_n = \{e_i\}_{i=1}^n$, its law consists only of brackets $[e_i, e_j] = c_{i,j}^i e_i + c_{i,j}^j e_j$. This is because of dealing with digraphs and not with full triangles.

To implement the algorithm, we have used the symbolic computation package MAPLE 12, loading the libraries `linalg`, `GraphTheory` and `Maplets[Elements]`. The first two libraries allow us to apply commands of Linear Algebra and Graph Theory, respectively; whereas the last is used to display a message so that the user introduces the required input in the first subroutine, devoted to define the law of the Lie algebra $\mathfrak{g}$. The algorithm to obtain the digraph associated with $\mathfrak{g}$ considers the following two steps:

1. Entering the law of $\mathfrak{g}$ by means of a routine computing the Lie bracket between two arbitrary basis vectors in $\mathcal{B}_n$.

2. Defining the digraph associated with $\mathfrak{g}$ using the method reviewed in Section 3.

The first routine, named `law`, receives two natural numbers as inputs. These numbers represent the subindexes of two basis vectors in $\mathcal{B}_n$. The subroutine returns the result of the bracket between these two vectors. In addition, conditional sentences are inserted to determine the non-zero brackets and the skew-symmetry property. Since the user has to complete the subroutine inserting the non-zero brackets of $\mathfrak{g}$, we have also added a sentence at the beginning of the implementation, reminding this fact. Note that before running any other sentence, we must restart all the variables and delete all the computations saved for previous law. Additionally, we must update the value of variable `dim` with the dimension of $\mathfrak{g}$.

 ISBN: 978-84-614-6167-7

```
> restart:
> maplet:=Maplet(AlertDialog("Don't forget to introduce non-zero brackets of the algebra
and its dimension in subroutine law",'onapprove'=Shutdown("Continue"),
'oncancel'=Shutdown("Aborted"))):
> Maplets[Display](maplet):
> assign(dim,...):
> law:=proc(i,j)
> if i=j then return 0; end if;
> if i>j then return -law(j,i); end if;
> if (i,j)=... then return ...; end if;
> if ....
> else return 0; end if;
> end proc;
```

The ellipsis in command `assign` corresponds to write the dimension of $\mathfrak{g}$. The following two suspension points are associated with the computation of $[e_i, e_j]$: first, the value of the subindexes (i, j) and second, the result of $[e_i, e_j]$ with respect to $\mathcal{B}_n$. The last ellipsis denotes the rest of non-zero brackets. For each non-zero bracket, a new sentence `if` has to be included in the cluster.

Next, we implement the second step of the algorithm with the routine `drawdigraph`, receiving the dimension `n` of $\mathfrak{g}$ as input. This routine draws the digraph associated with $\mathfrak{g}$. To do so, two local variables `V` and `E` have been defined: `V` is a list with the vertices of the digraph and `E` is a set containing the edges. Hence, several loops are programmed to include all the directed, weighted edges in the set `E` according to the non-zero brackets saved in the subroutine `law`.

```
> drawdigraph:=proc(n)
> local E,V; E:={};V:=[];
> for x from 1 to n do
> V:=[op(V),x]; end do;
> for i from 1 to n do
> for j from i+1 to n do
> if coeff(law(i,j),e[j])<>0 then
> E:={op(E),[[i,j],coeff(law(i,j),e[j])]}; end if;
> if coeff(law(i,j),e[i])<>0 then
> E:={op(E),[[j,i],coeff(law(i,j),e[i])]};
> end if; end do; end do;
> G:=Digraph(V,E);
> return DrawGraph(G);
> end proc:
```

Example 1 *To illustrate this algorithm, we apply it to the* 6*-dimensional Lie algebra with law* $[e_1, e_3] = 2e_3$, $[e_1, e_4] = -e_4$, $[e_1, e_6] = e_6$, $[e_2, e_3] = -e_3$, $[e_2, e_4] = e_4$, $[e_2, e_5] = e_5$. *First, we complete the routine* `law` *as follows:*

```
> restart:
```

 ISBN: 978-84-614-6167-7

```
> maplet:=Maplet(AlertDialog("Don't forget to introduce non-zero brackets of the algebra
 and its dimension in subroutine law",'onapprove'=Shutdown("Continue"),
 'oncancel'=Shutdown("Aborted"))):
> Maplets[Display](maplet):
> assign(dim,6):
> law:=proc(i,j)
> if i=j then return 0;end if;
> if i>j then return -law(j,i);end if;
> if (i,j)=(1,3) then return 2*e[3];end if; if (i,j)=(1,4) then return -e[4];end if;
> if (i,j)=(1,6) then return e[6];end if; if (i,j)=(2,3) then return -e[3];end if;
> if (i,j)=(2,4) then return e[4];end if; if (i,j)=(2,5) then return e[5];
> else return 0;
> end if;
> end proc:
```

Next, we run the routine `drawdigraph` *and execute the sentence* `drawdigraph(dim)` *obtaining the digraph in Figure 6.*

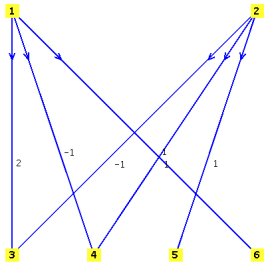


Figure 6: output from Example 1.

5.2 Algorithm to decide if a digraph is associated with a Lie algebra

We show an algorithmic procedure to determine if a given digraph is associated or not with a Lie algebra. The algorithm consists of two steps: a) generating the law candidate to be a Lie algebra using the construction reviewed in Section 3; and b) checking if the Jacobi identities are satisfied for this law. To implement the algorithm, we need load the libraries `DifferentialGeometry, LieAlgebras` and `GraphTheory` to activate commands related to Lie algebras and Graph Theory.

First, we build a vector space associated with the digraph using the routine `program`, receiving two inputs: a list `V` with the vertices of the digraph and a set `E` with the directed, weighted edges. As output, we obtain a vector space with basis $\{e_i\}_{i=1}^n$ where e_i corresponds to vertex i from the list `V` and the brackets associated with the edges in the set `E`. To implement this routine we define two local variables: `B` and `L`, where `B` saves the basis $\{e_i\}_{i=1}^n$ and `L` is a list containing the indexes of the structure constants from the non-zero brackets.

 ISBN: 978-84-614-6167-7

```
> program:=proc(V,E)
> local B, L;
> B:=[]; L:=[];
> for x from 1 to nops(V) do
> B:=[op(B),e[x]];
> end do;
> for i from 1 to nops(E) do
> L:=[op(L),[[op(E[i][1]),E[i][1][2]],E[i][2]]];
> end do;
> return _DG([["LieAlgebra",Alg1,[nops(V)]],L]);
> end proc:
```

Next, the vector space having such basis and law is generated when evaluating the sentence

```
> DGsetup(program(V,E));
```

After defining this vector space `Alg1`, we can operate over it. More concretely, we check if Jacobi identities hold or not for `Alg1`:

```
Alg1 > Query(Alg1,"Jacobi");
```

The vector space `Alg1` defined by the output of `program` is a Lie algebra if and only if the answer is `true` for this question.

Example 2 *Consider the digraph in Figure 7. After running the routine* `program`*, we define the list* `V` *of vertices and the set* `E` *of edges. Then the routine* `program` *generates the vector space associated with the graph and finally Jacobi identities are checked*

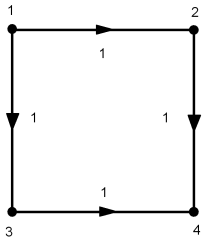


Figure 7: Digraph of Example 2.

```
> V:=[1,2,3,4];
> E:={[[1,2],1],[[1,3],1],[[2,4],1],[[3,4],1]};
> DGsetup(program(V,E));
Alg1 > Query(Alg1,"Jacobi");
>            false
```

Since the answer is `false`*, the digraph in Figure 7 is not associated with a* 4*-dimensional Lie algebra.*

 ISBN: 978-84-614-6167-7

Acknowledgements

This work has been partially supported by MTM2010-19336 and FEDER.

References

[1] A. Carriazo, L.M. Fernández and J. Núñez, *Combinatorial structures associated with Lie algebras of finite dimension*, Linear Algebra and Applications. **389** (2004), 43–61.

[2] M. Ceballos, J. Núñez and A.F. Tenorio, *Complete triangular structures and Lie algebras*, International Journal of Computer Mathematics (2011), 1–13. DOI: 10.1080/00207161003767994.

[3] M. Ceballos, J. Núñez and A.F. Tenorio, *Study of Lie algebras by using combinatorial structures*, Linear Algebra and its Applications (2011), DOI: 10.1016/j.laa.2010.11.030.

[4] V.S. Varadarajan, *Lie Groups, Lie Algebras and Their Representations*, Springer, 1984.

[5] H.S. Wilf, *Algorithms and Complexity*, Prentice Hall, 1986.

Proceedings of the 11th International Conference on Computational and Mathematical Methods in Science and Engineering, CMMSE 2011 26–30 June 2011.

Permutations and entropy on individual orbits

Jose S. Cánovas[1]

[1] *Department of Applied Mathematics and Statistics , Technical University of Cartagena*

emails: jose.canovas@upct.es

Abstract

A definition of permutation entropy for two dimensional maps is introduced. It seems to be a useful notion for analyzing the complexity of discrete models. We use it to analyze an economic model.

Key words: entropy, chaos, permutations
MSC 2000: 26A18, 37E40.

1 Introduction and main definitions

Let $I = [a, b]$ be a compact subinterval and let $f : I \to I$ be a continuous map. The topological entropy of f (see [4] for a definition) is a non negative number, $h(f)$, which can be taken as a measure of the dynamical complexity of the map f. For instance, if $h(f) > 0$, then the map f is chaotic in the sense of Li and Yorke (see [1, Chapter 4]). However, it is a conjugacy invariant which is very difficult to compute in practical examples.

Hence, we consider some techniques for time series analysis (see eg. [7] or [3]) which allow us to obtain estimations of the topological entropy. Namely, let $\mathcal{S}_n$ be the set of permutations of length n and a sequence $(x_n)_{n=0}^{\infty}$. Let $\mathcal{A}_n \subset \mathcal{S}_n$ be the subset of permutations with the property that for any $\pi \in \mathcal{A}_n$, there is $k \in \mathbb{N}$, such that $x_{k+\pi(1)} < x_{k+\pi(2)} < ... < x_{k+\pi(n)}$ and define the permutation entropy of the sequence as

$$h^*((x_n)_{n=0}^{\infty}) = \limsup_{n\to\infty} \frac{1}{n} \log \#\mathcal{A}_n.$$

In addition, we can consider the Shannon permutation entropy which is defined as

$$h_S((x_n)_{n=0}^{\infty}) = \limsup_{n\to\infty} -\frac{1}{n} \sum_{\pi\in\mathcal{A}_n} p(\pi) \log p(\pi),$$

 ISBN: 978-84-614-6167-7

where $p(\pi)$ is the relative frequency of the symbol π. It is clear that both, permutation entropy and Shannon entropy can be estimated by finite time series, that is, for $(x_n)_{n=0}^N$, $N \in \mathbb{N}$.

Note that easily we obtain that $h^*((x_n)_{n=0}^\infty) = h_S((x_n)_{n=0}^\infty) = 0$ when the sequence is periodic or have some monotonicity properties. The connection between topological entropy and permutation entropy was recently pointed out in [2] by proving the following result. For any $\pi \in \mathcal{S}_n$, define the partition $\mathcal{P}_\pi = \{x \in I : f^{\pi(1)}(x) < f^{\pi(2)}(x) < ... < f^{\pi(n)}(x)\}$. Then, if the map f is piecewise monotone (even non continuous), we have that

$$h(f) = \lim_{n\to\infty} \frac{1}{n} \log \#\{\pi \in \mathcal{S}_n : \mathcal{P}_\pi \neq \emptyset\}.$$

It is just a simple observation that $\mathcal{P}_\pi \neq \emptyset$ provided there is $x \in I$ for which its orbit $\mathrm{Orb}(x, f) = (f^n(x))_{n=0}^\infty$ contains π as a permutation. This is the idea for proving the following result (see [5]). Let $f : I \to I$ be continuous and piecewise monotone. Then, for any $y \in I$

$$h^*(\mathrm{Orb}(y, f)) \leq h(f) = \sup_{x\in I} h^*(\mathrm{Orb}(x, f)).$$

For two dimensional discrete dynamical systems (Ω, F), $\Omega \subset \mathbb{R}^2$ and $F : \Omega \to \Omega$ continuous, we can define permutation and Shannon permutation entropies as follows. For $(x, y) \in \Omega$, let

$$(x_n)_{n=0}^\infty = p_1(\mathrm{Orb}((x, y), F))$$

and

$$(y_n)_{n=0}^\infty = p_2(\mathrm{Orb}((x, y), F)),$$

where $(x_0, y_0) = (x, y)$ and p_i are the natural projections from the real plane to the real line. Hence, define $\mathcal{P}_n = \{(\pi_1, \pi_2)$:there is $k \in \mathbb{N}$, such that $x_{k+\pi_1(1)} < x_{k+\pi_1(2)} < ... < x_{k+\pi_1(n)}$ and $y_{k+\pi_2(1)} < y_{k+\pi_2(2)} < ... < y_{k+\pi_2(n)}\}$ and hence, define the topological permutation entropy by

$$h^*((x_n, y_n)_{n=0}^\infty) = \limsup_{n\to\infty} \frac{1}{n} \log \#\mathcal{P}_n$$

and the Shannon permutation entropy as

$$h_S((x_n, y_n)_{n=0}^\infty) = \limsup_{n\to\infty} -\frac{1}{n} \sum_{\pi_1,\pi_2\in\mathcal{P}_n} p(\pi_1, \pi_2) \log p(\pi_1, \pi_2),$$

where $p(\pi_1, \pi_2)$ is the relative frequency of the symbol (π_1, π_2).

Then, define the topological and Shannon permutation entropies of F as

$$h^*(F) = \sup\{h^*(\mathrm{Orb}((x, y), F)) : (x, y) \in \Omega\}$$

and

$$h_S(F) = \sup\{h_S(\mathrm{Orb}((x, y), F)) : (x, y) \in \Omega\}$$

We will test the above definitions by using a triopoly economical model.

 ISBN: 978-84-614-6167-7

2 An example

As an example, we consider the dynamics of a triopoly given in [6] given by

$$\begin{cases} q_1(t+1) = \max\left\{0, \sqrt{2q(t)/c_1} - 2q(t)\right\}, \\ q(t+1) = \max\left\{0, \sqrt{(q_1(t)+q(t))/c} - q_1(t) - q(t)\right\}, \end{cases}$$

where c_1 is the constant marginal cost of firm one, $c_2 = c_3 = c$, are the constant marginal costs of firms two and three, and additionally all the initial conditions are taken in the subset

$$\Delta = \{(q_1, q, q) : q_1, q \geq 0\}.$$

It is well known that one can consider either $c_1 = 1$ or $c = 1$. When $c = 1$ the dynamics is simple, while if $c_1 = 1$, the model seems to be a change from simple dynamics to complicated dynamics when c ranges the interval $(3.5, 4.5)$. On this interval, the Cournot equilibrium point, which plays an important role in economic dynamics, goes from stability to instability. In particular, when $c = 4.5$ simulations show the existence of a dense orbit on a positive two–dimensional Lebesgue measure.

Figures 1–7 shows several experiments with different permutation size and data length.

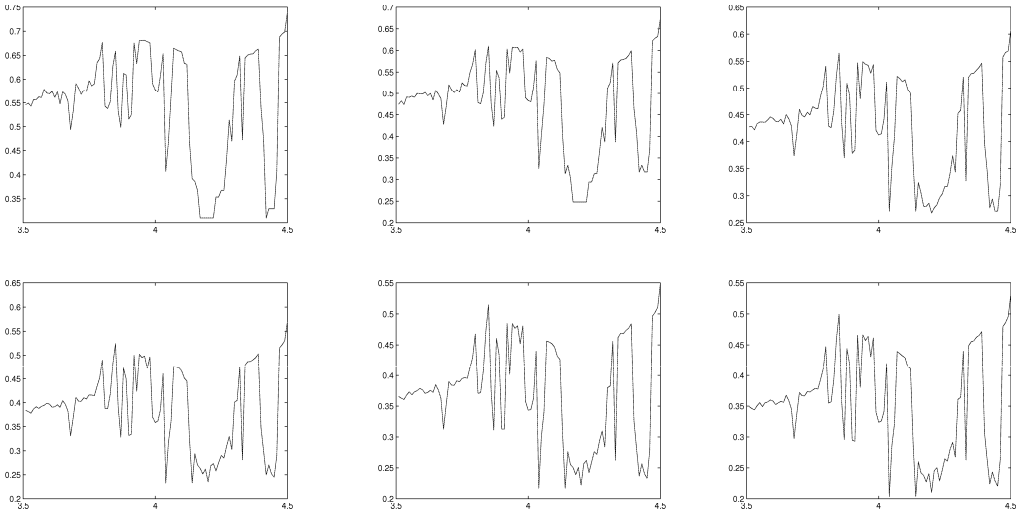


Figure 1: *Topological permutation entropies for permutation length m=8, 10, 12, 14, 15 and 16 (from left and top to right and down). In all the cases, we generate an orbit starting from the initial conditions* $(0.045, 0.015)$. *We increase the data length when m increases.*

 ISBN: 978-84-614-6167-7

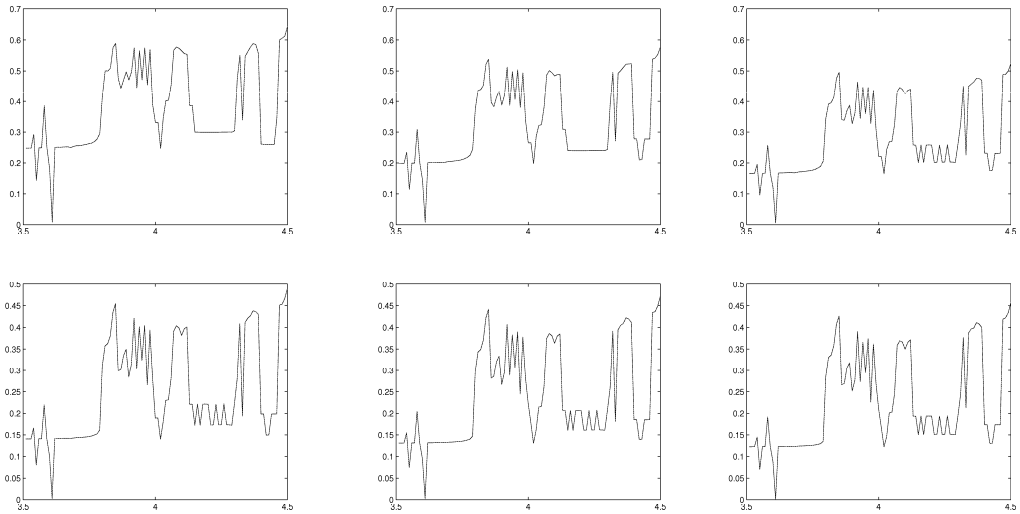


Figure 2: *Shannon permutation entropies for permutation length m=8, 10, 12, 14, 15 and 16 (from left and top to right and down). In all the cases, we generate an orbit starting from the initial conditions* $(0.045, 0.015)$. *We increase the data length when m increases.*

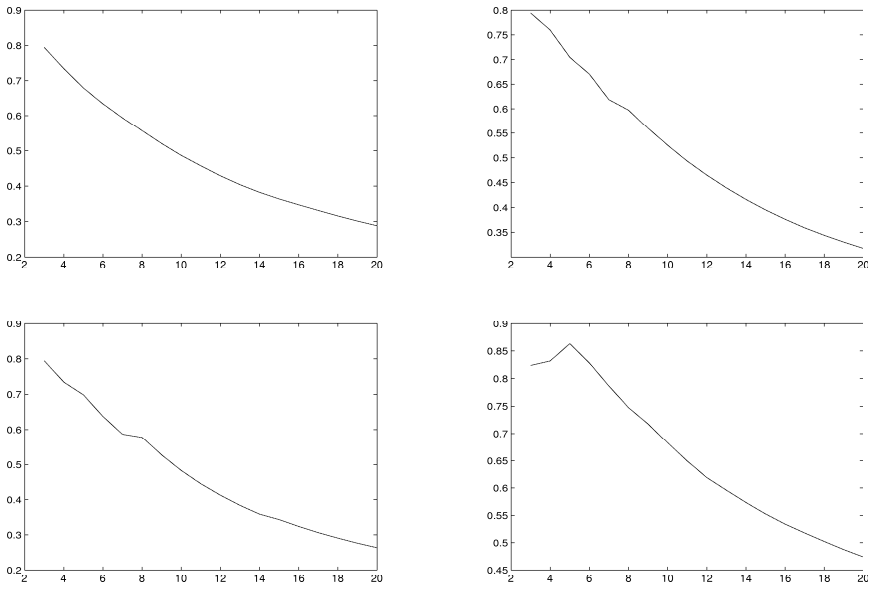


Figure 3: *Topological permutation entropy for fixed parameter values c=3.5, 3.75, 4 and 4.5 (from left and top to right and down). The permutation length m moves from 3 to 20. We increase the data length when m increases.*

 ISBN: 978-84-614-6167-7

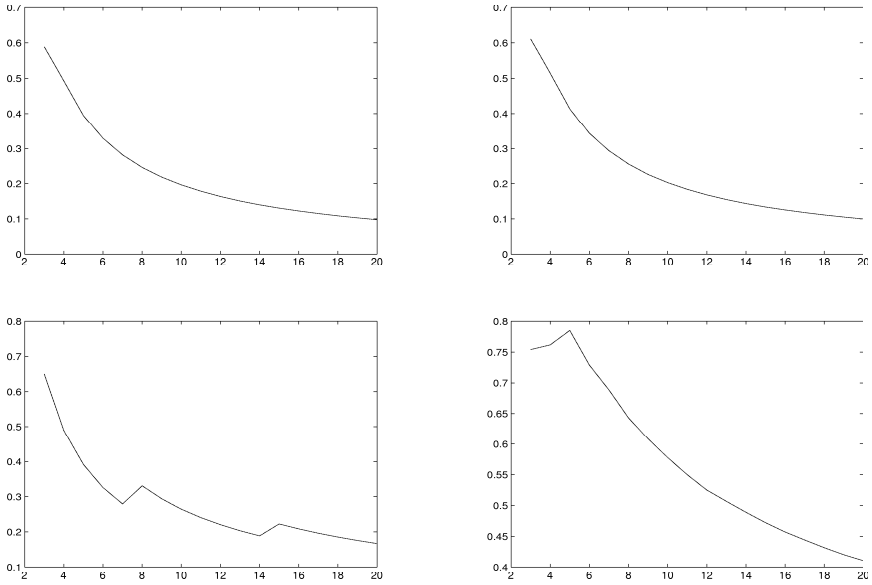


Figure 4: *Shannon permutation entropy for fixed parameter values c=3.5, 3.75, 4 and 4.5 (from left and top to right and down). The permutation length m moves from 3 to 20. We increase the data length when m increases.*

Acknowledgements

This work has been supported by the grants MTM2008-03679/MTM [MCI (Ministerio de Ciencia e Innovación) and FEDER (Fondo Europeo de Desarrollo Regional)] and 08667/PI/08 [Fundación Séneca, CARM].

References

[1] L. Alsedá, J. Llibre and M. Misiurewicz, *Combinatorial dynamics and entropy in dimension one,* World Scientific Publishing (1993).

[2] C. Bandt, G. Keller and B. Pompe, *Entropy of interval maps via permutations*, Nonlinearity **15** (2002), 1595–1602.

[3] C. Bandt and B. Pompe, *Permutation entropy-a natural complexity measure for time series*, Phys. Rev. Lett. **88** (2002), 174102.

[4] R. Bowen, *Entropy for group endomorphism and homogeneous spaces,* Trans. Amer. Math. Soc. **153** (1971), 401–414.

[5] J. S. Cánovas, *Estimating topological entropy from individual orbits*, Int. J. Comput. Math. **86** (2009), 1901–1906.

 ISBN: 978-84-614-6167-7

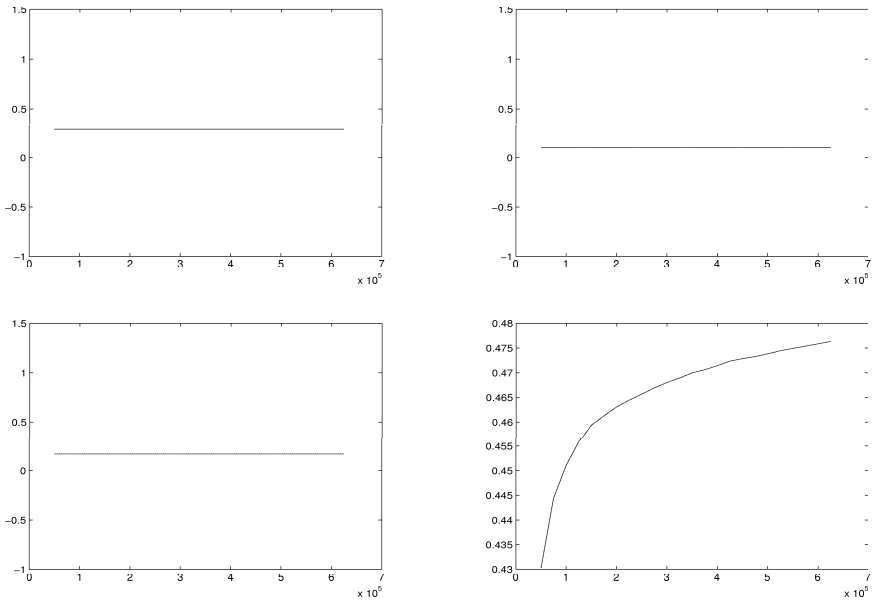


Figure 5: *Topological permutation entropy for fixed parameter values c=3.5, 3.75, 4 and 4.5 (from left and top to right and down) and fixed permutation length* $m = 20$. *We increase the data length by adding 25000 new points in each computation.*

[6] T. Puu, *The chaotic duopolist revisited*, Journal of Economic Behavior & Organization **33** (1998), 385–394.

[7] M. Small, *Applied nonlinear time series analysis. Applications in physics, physiology and finance*, World Scientific Series on Nonlinear Science. Series A: Monographs and Treatises, **52**. World Scientific Publishing Co. Pte. Ltd., Hackensack, NJ, 2005.

[8] P. Walters, *An introduction to ergodic theory,* Springer Verlag, New York (1982).

 ISBN: 978-84-614-6167-7

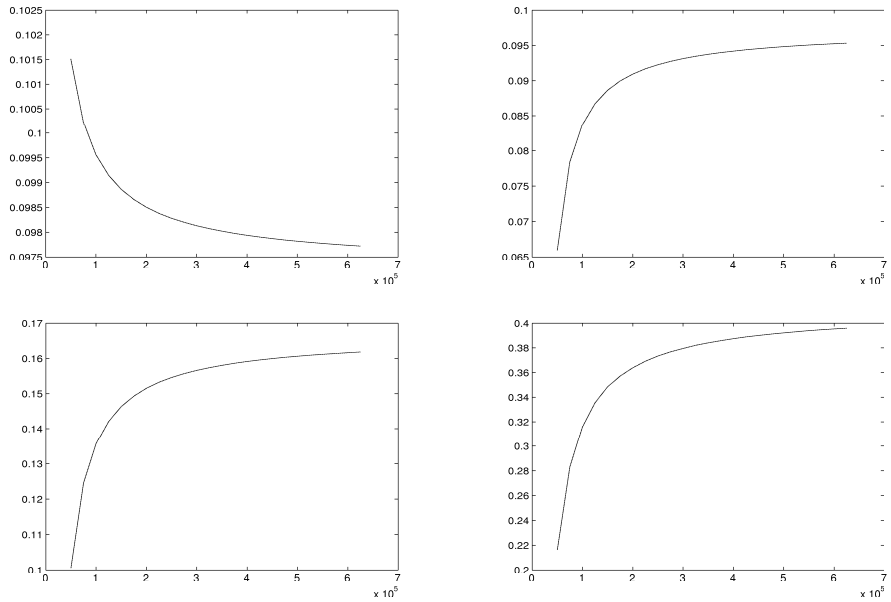


Figure 6: *Shannon permutation entropy for fixed parameter values c=3.5, 3.75, 4 and 4.5 (from left and top to right and down) and fixed permutation length* $m = 20$. *We increase the data length by adding 25000 new points in each computation.*

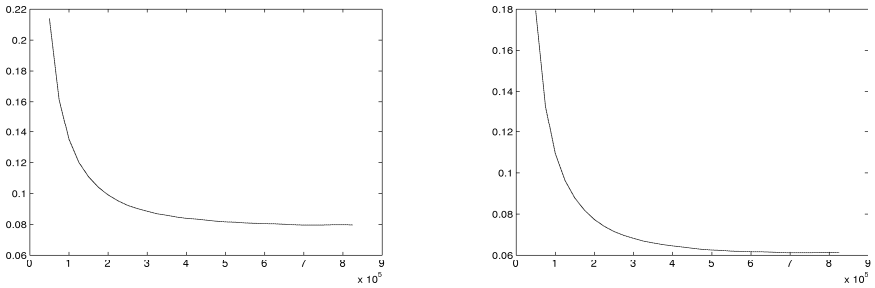


Figure 7: *Topological permutation entropy minus Shannon permutation entropy for fixed parameter value* $c = 4.5$ *and fixed permutation lengths* $m = 20$ *and* 25. *We increase the data length by adding 25000 new points in each computation. We see that the difference seems to go to zero, which could be motivated by the variational principle for classical topological entropy (see [8, Chapter 8])*

 ISBN: 978-84-614-6167-7

Proceedings of the 11th International Conference on Computational and Mathematical Methods in Science and Engineering, CMMSE 2011 26–30 June 2011.

Optimal control in dynamic gas-liquid reactors

B.Cantó[1], S.C. Cardona[2], C.Coll[1], J. Navarro-Laboulais[2] and E. Sánchez[1]

[1] *Institut de Matemàtica Multidisciplinar, Universitat Politècnica de València*

[2] *Departament d'Enginyeria Química i Nuclear, Universitat Politècnica de València*

emails: bcanto@mat.upv.es, scardona@iqn.upv.es, mccoll@mat.upv.es, jnavarla@iqn.upv.es, esanchezj@mat.upv.es

Abstract

We consider a dynamic gas-liquid transfer model without chemical reaction based on unsteady film theory. Using an explicit representation of the model, we study the quadratic optimal control problem. For that, some results on controllability,observability and stability criteria and the relation between these properties and the parameters of the model are shown. After, the steady-state solution of the Riccatti equation is used. And finally, the optimal control problem with a quadratic cost functional is solved.

Key words: controllability, observability, stability, optimal control, quadratic cost.

MSC 2000: 34, 93

1 Introduction

Optimal control theory is a mature mathematical discipline with numerous applications in both science and engineering. The objective of optimal control theory is to determine the control signals that will cause a process to satisfy the physical constraints and at the same time minimize (or maximize) some performance criterion. A control problem includes a cost functional that is a function of state and control variables. An optimal control is a set of differential equations describing the paths of the control variables that minimize the cost functional.

One useful part of optimal control theory for ordinary differential equations is the theory of optimal control of linear differential systems with a quadratic cost criterion. This theory is also the most complete, both for systems evolving in a finite-time interval as well as over an infinite-time interval. It is well known that in the finite-time case the optimal control can be expressed in linear feedback form, where the "feedback gains"

 ISBN: 978-84-614-6167-7

satisfy a matrix differential equation of Riccati type. In the infinite-time case by using the theory of controllability, the asymptotic behavior of the controlled system can be studied and a rather complete solution to the problem is available, see [5] for more details.

In this paper we solve the optimal control problem when we have a dynamic gas-liquid reactor. Bubble column reactors are widely used in chemical, petrochemical, biochemical and metallurgical industries. The absence of moving parts, their low operating and maintenance costs and the excellent mass and heat transfer rates explain the large number of applications developed with this kind of reactor against the others [3], [4]. However, the design and scale-up of bubble columns are difficult because of the complexity of the gas and liquid flow patterns coupled with mass transfer and chemical reactions. Key factors such as gas hold-up, ϵ, volumetric mass transfer coefficient, k_La, specific interfacial area, a, bubble size, r_{32}, and kinetic rate constants, k_n, and how those parameters are related is fundamental for the proper design and the operational control of gas-liquid reactors.

The classical description of mass transfer processes in gas-liquid reactors makes use of two spatial-temporal scales which one refers to the physical mass-transfer at the gas-liquid interface level, i.e. the microscopic model, and the other time scale is referred to the modelling of the reactor configuration which considers the mixing processes and the chemicals distribution in the whole volume of the reactor, i.e. the macroscopic model. Assuming the two-film theory to model the gas-liquid mass transfer at the microscopic level and an ideal mixing flow reactor model to describe the macroscopic behaviour of the gas and the liquid phases in the reactor, the simplest gas-liquid mass transfer model is given by the following system of partial differential equations:

$$\begin{aligned}
\frac{dy(t)}{dt} &= \frac{RT}{PV}\frac{1-\epsilon}{\epsilon}F_1\left(\frac{y_0}{1-y_0}U(t)-\frac{y(t)}{1-y(t)}\right)+\frac{RT}{P}\frac{1-\epsilon}{\epsilon}aD\left(\frac{\partial C(z,t)}{\partial z}\right)_{z=0} && (1)\\
\frac{\partial C(z,t)}{\partial t} &= D\frac{\partial^2 C(z,t)}{\partial z^2} \quad \forall z\in[0,\delta] && (2)\\
\frac{dC^b(t)}{dt} &= -aD\left(\frac{\partial C(z,t)}{\partial z}\right)_{z=\delta}
\end{aligned}$$

with the initial and boundary conditions:

$$\begin{aligned}
t=0 \quad & \left|\begin{array}{c} y(0)=0 \\ C(z,0)=0 \\ C^b(0)=0 \end{array}\right. && (3)\\
t>0 \quad & C(0,t)=\frac{P}{H}y(t)\\
t>0 \quad & C(\delta,t)=C^b(t)
\end{aligned}$$

where the state variables of the system are $y(t)$, the gas phase concentration, $C(t)$, the concentration in the interface of the transferred substance and $C^b(t)$ the concentration at the liquid bulk. Considering the Method of Lines as an approximate solution of the

 ISBN: 978-84-614-6167-7

partial differential equation (2) applied to the spatial coordinate z, the system (1-3) is transformed to the linear system:

$$\begin{aligned} \dot{x}_1 &= \frac{K_1 p_1}{K_2 - 1} U(t) - \frac{K_1 p_1}{K_2} x_1 - N K_3 \frac{p_1}{p_2} \frac{p_3}{p_4} p_5 (3x_1 - 4x_2 + x_3) \\ \dot{x}_i &= N^2 \frac{p_3}{p_4^2} (x_{i-1} - 2x_i + x_{i+1}) \\ \dot{x}_{N+1} &= -\frac{N}{2} \frac{p_3}{p_4} p_5 (x_{N-1} - 4x_N + 3x_{N+1}) \end{aligned} \tag{4}$$

where the parameters of the model are defined as:

$$p_1 = \frac{1-\epsilon}{\epsilon} \quad p_2 = H \quad p_3 = D \quad p_4 = \delta \quad p_5 = a$$

together with the constants:

$$K_1 = \frac{RT}{PV} \frac{F_1}{y_0} \quad K_2 = y_0^{-1} \quad K_3 = \frac{1}{2} RT$$

All the initial conditions of the equations (4) are set to zero, $x_i(0) = 0, \ \forall i$.

So, the present paper is concerned with a study of the optimal feedback control problem for the model given in (1) with a quadratic cost using the approximate system (4). The theory is currently being completed in order to show the relation of the theory of controllability and observability to the infinite-time quadratic cost problem.

One of the most remarkable results in linear control theory and design is that if the cost criterion is quadratic, and the optimization is over an infinite horizon, the resulting optimal control law has many nice properties, including that of closed loop stability.

For that, we note that optimal control theory is intimately connected no only to system structural properties of controllability and observability but also to the property that relates to the system response to inputs or disturbances, the stability. If we want to study the optimal control problem, the system must be controllable and observable. The goal in optimal control theory is to transfer a system from an arbitrary initial state to the origin while minimizing some performance measure. On the other hand, it is important to note that although practical optimal control systems that minimize quadratic performance indexes are almost always asymptotically stable.

Remind that a system is controllable, if given two state $x_0 = x(t_0) \in \mathbb{R}^n$ and $x_f = x(t_f) \in \mathbb{R}^n$, there exists a time t_f, with $t_0 < t_f$ and a control vector $u(t)$ defined on de interval $[t_0, t_f]$ which takes the state vector state from x_0 to x_f. A system is said to be observable, if there exist a time t_f, with $t_0 < t_f$ such that given the vectors $u(t)$ and $y(t)$ over the interval $[t_0, t_f]$ it is possible to deduce the initial state-vector $x(t_0)$, see [7] for more details.

Finally, in a chemical process, when two compartments are in contact, they can achieve equilibrium points. An equilibrium point of system (1) denoted by $x^* \in \mathbb{R}^n$, verifies $x^* = A(p)x^*$, being $A(p)$ the system matrix of model (4). An equilibrium point x^* is said to be asymptotically stable if every trajectory starting in a neighborhood of it is around the $x^*$and converges on x^*.

 ISBN: 978-84-614-6167-7

To apply optimal control it is necessary that the model given in (1) being identifiable. The problem of the structural identifiability of the model consists of the determination of all parameter sets which give the same input-output structure. In the last years many papers of different fields were published in structural identification, [1] and [2]. A identifiability relation of the proposed model is given in [6] and this property gave a relation between some of those parameters.

The main contribution of this paper is a characterization of the sequence of control to appropriate optimal control. We proposed an approach for solving optimal control problem for the model described in (1). In particular, we derived the derivatives of the cost and use optimization techniques to locate the optimal control. To solve this problem, the system must be controlable, observable and asymptotically stable, we also study these properties and the relation between them and optimal control problem. Additionally, we give some conditions on the parameters in order to verify these properties. The results provide a theoretical foundation for extending the use of numerical algorithms.

Acknowledgements

This work has been partially supported by PAID-05-10-003-295.

References

[1] B. Cantó, C. Coll and E. Sánchez, *Structural identifiability of a model of dialysis* , Math. Comp. Modelling **50** (2009) 733–737.

[2] B. Cantó, C. Coll and E. Sánchez, *Identifiability of a class of discretized linear partial differential algebraic equations*, Math. Problems Eng. **2011** (2011) 1–12.

[3] W. D. Deckwer, *Bubble column reactors*, John Wiley and Sons, Chichester, 1992.

[4] N. Kantarci, F. Borak and K. O. Ulgen, *Bubble column reactors* , Proc. Biochem. **40**(7) (2005) 2263–2283.

[5] H. Kawakernaak and R. Sivan, *Linear Optimal Control Systems*, Wiley-Interscience, New York, 1972.

[6] J. Navarro-Laboulais, S. C. Cardona, J. I. Torregrosa, A. Abad and F. López, *Practical identifiability analysis in dynamic gas-liquid reactors. Optimal experimental design for mass-transfer parameters determination* , Comp. and Chem. Eng. **32** (2008) 2382–2394.

[7] R.Patel and N. Munro, *Multivariable Systen. Theory and Design*, Pergamon Press, New York, 1982.

 ISBN: 978-84-614-6167-7

Proceedings of the 11th International Conference on Computational and Mathematical Methods in Science and Engineering, CMMSE 2011 26–30 June 2011.

MDS array codes based on superregular matrices

Sara D. Cardell[1], **Joan-Josep Climent**[1] **and Verónica Requena**[1]

[1] *Departament d'Estadística i Investigació Operativa, Universitat d'Alacant*

emails: s.diaz@ua.es, jcliment@ua.es, vrequena@ua.es

Abstract

In this paper we introduce a new construction of MDS array codes. In order to obtain a code with this property, we construct the parity-check matrix just using a superregular matrix by blocks composed by powers of the companion matrix of a primitive polynomial.

Key words: Array code, MDS code, superregular matrix, companion matrix, primitive polynomial

1 Introduction

Array codes are a class of error control codes. They have several applications in communication and storage systems [3, 14, 15] and they have been widely studied [1, 2, 4]. Array codes are very useful to dynamic high-speed storage applications since they have low-complexity decoding algorithms over small fields and low update complexity when small changes are applied to the stored data [3]. In general, Reed-Solomon codes have none of these properties; thus, they are more efficient than Reed-Solomon codes in computational complexity terms [3, 6]. Furthermore, if they are MDS, they provide the maximum protection against device failure for a given amount of redundancy [2]. It is possible to find some constructions of this kind of codes in [1, 2, 5, 10, 14, 15]. In this paper we propose a new construction to obtain array codes which are MDS.

The rest of the paper is organized as follows. In Section 2 we introduce some notation and recall some properties and definitions. In Section 3 we introduce the construction of an array code using a superregular matrix and the companion matrix of a primitive polynomial. Finally, we present some conclusions in Section 4.

2 Preliminaries

Let $\mathbb{F}_q$ be the Galois field of q elements and consider b a positive integer. If $\mathcal{C}$ is a code of length n over $\mathbb{F}_q^b$, we can consider the codewords of $\mathcal{C}$ as codewords of length nb over

 ISBN: 978-84-614-6167-7

$\mathbb{F}_q$. Then, a code $\mathcal{C}$ is said to be a **linear array code** of length n over $\mathbb{F}_q^b$ if it is a linear code of length nb over $\mathbb{F}_q$ (see [5]). If $\dim \mathcal{C}$ is the dimension of $\mathcal{C}$ as a vector space over $\mathbb{F}_q$, then $k = \frac{\dim \mathcal{C}}{b}$ is the **normalized dimension** of $\mathcal{C}$, and thus, the parameters of the code over $\mathbb{F}_q^b$ are $[n, k, d]$, where d is the minimum distance.

In order for the code to be MDS over $\mathbb{F}_q^b$, the Singleton bound

$$d \leq n - k + 1$$

must be attained [5, 11].

The following two theorems are useful to check whether an array code is MDS or not.

Theorem 1 ([5]): *Let $H = (H_1, H_2, \ldots, H_n)$ be an $rb \times nb$ systematic parity-check matrix of an $\mathbb{F}_q$-linear $[n, n-r]$ code $\mathcal{C}$ over $\mathbb{F}_q^b$, where each H_i is an $rb \times b$ submatrix of H. Then $\mathcal{C}$ is MDS if and only if the rb columns of any r distinct submatrices H_i form a linearly independent set over $\mathbb{F}_q$.*

Theorem 2 ([5]): *Let $H = (A, I_{rb})$ be an $rb \times nb$ systematic parity-check matrix of an $\mathbb{F}_q$-linear $[n, n-r]$ code $\mathcal{C}$ over $\mathbb{F}_q^b$ and write $A = (A_{i,j})_{i,j=1}^{r,n-r}$, where each $A_{i,j}$ is a $b \times b$ block submatrix of A. Then $\mathcal{C}$ is MDS if and only if every square submatrix of A consisting of full blocks submatrices $A_{i,j}$ is nonsingular.*

Several constructions of MDS block codes based on superregular matrices have been proposed [8, 13]. Our purpose is to extend these constructions using the characterization given in Theorem 2 in order to obtain array codes which are also MDS.

Definition 1 ([12]): A matrix A is said to be a **superregular matrix** if every square submatrix of A is nonsingular.

It is worth pointing out that some authors have used the term superregular to define a related but different type of matrices, see for instance [7]. This type of matrices is not suitable to contruct MDS array codes using Theorem 2, as we will see in Example 2 below.

3 Main results

Let $p(x) = p_0 + p_1 x + \cdots + p_{b-1}x^{b-1} + x^b \in \mathbb{F}_q[x]$. Remember that the companion matrix of $p(x)$ is given by

$$C = \begin{pmatrix} 0 & 0 & \cdots & 0 & -p_0 \\ 1 & 0 & \cdots & 0 & -p_1 \\ 0 & 1 & \cdots & 0 & -p_2 \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & -p_{b-2} \\ 0 & 0 & \cdots & 1 & -p_{b-1} \end{pmatrix}.$$

 ISBN: 978-84-614-6167-7

Moreover, if $p(x)$ is a primitive polynomial, it is well known that (see [9])

$$\mathbb{F}_q[C] = \{0, I, C, C^2, \ldots, C^{q^b-2}\} \approx \mathbb{F}_{q^b}. \tag{1}$$

The isomorphism $\psi : \mathbb{F}_{q^b} \to \mathbb{F}_q[C]$, with $\psi(\alpha) = C$, where $\alpha \in \mathbb{F}_{q^b}$ is a primitive element, can be extended to a ring isomorphism

$$\Psi : \mathrm{Mat}_{m\times t}(\mathbb{F}_{q^b}) \longrightarrow \mathrm{Mat}_{m\times t}(\mathbb{F}_q[C])$$

in the following way: if $A = (\alpha_{ij}) \in \mathrm{Mat}_{m\times t}(\mathbb{F}_{q^b})$, then $\Psi(A) = (\psi(\alpha_{ij}))$ where $\psi(\alpha_{ij}) \in \mathbb{F}_q[C]$ for $i = 1, 2, \ldots, m$, $j = 1, 2, \ldots, t$.

Theorem 3: *If $A = (\alpha_{ij}) \in \mathrm{Mat}_{m\times t}(\mathbb{F}_{q^b})$ is a superrregular matrix, then $H = (\mathbf{A}, \mathbf{I}_{bm})$, where $\mathbf{A} = (\psi(\alpha_{ij}))$, is the parity check-matrix of an $[m + t, t, m + 1]$ MDS array code $\mathcal{C}$ over $\mathbb{F}_q^b$.*

Proof: It is sufficient to consider that $A \in \mathrm{Mat}_{m\times t}(\mathbb{F}_{q^b})$ is superregular if and only if $\mathbf{A} = \Psi(A) \in \mathrm{Mat}_{m\times t}(\mathbb{F}_q[C])$ is superregular by blocks. □

Remember that if we have an MDS block code, the dual code is MDS as well [11]. This result can be extended for array codes. Therefore, the dual code $\mathcal{C}^\perp$ of the code constructed in Theorem 3 is an $[m + t, m, t + 1]$ MDS array code over $\mathbb{F}_q^b$.

The following example helps us to understand this construction.

Example 1: Let $p(x) = x^3 + x + 1 \in \mathbb{F}_2[x]$ then

$$C = \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}.$$

The matrix $A = \begin{pmatrix} 1 & \alpha \\ 1 & \alpha^3 \end{pmatrix}$ where $\alpha \in \mathbb{F}_{2^3}$ is a primitive element, is a superregular matrix over $\mathbb{F}_{2^3}$. Then, according to Theorem 3 the parity-check matrix of $\mathcal{C}$ is given by

$$H = \left(\begin{array}{cc|c} I_3 & C \\ I_3 & C^3 \end{array}\right| I_6 \Bigg) = \left(\begin{array}{ccc|ccc|ccc|ccc} 1&0&0&0&0&1&1&0&0&0&0&0 \\ 0&1&0&1&0&1&0&1&0&0&0&0 \\ 0&0&1&0&1&0&0&0&1&0&0&0 \\ \hline 1&0&0&1&0&1&0&0&0&1&0&0 \\ 0&1&0&1&1&1&0&0&0&0&1&0 \\ 0&0&1&0&1&1&0&0&0&0&0&1 \end{array}\right).$$

As a code over $\mathbb{F}_2$, the parameters of $\mathcal{C}$ are $n = 12$, $k = 6$ and $d = 3$. If we consider this code as an array code over $\mathbb{F}_2^3$, the parameters of the code are $n = 4$, $k = 2$ and $d = 3$. Thus, the code is MDS.

 ISBN: 978-84-614-6167-7

Furthermore, the generator matrix of $\mathcal{C}$, or equivalently the parity-check matrix of $\mathcal{C}^\perp$, is

$$G = \left(\begin{array}{c|cc} I_6 & \begin{matrix} I_3 \\ C^T \end{matrix} & \begin{matrix} I_3 \\ (C^3)^T \end{matrix} \end{array} \right) = \left(\begin{array}{ccc|ccc|ccc|ccc} 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 \\ \hline 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 \end{array} \right)$$

so, the dual code $\mathcal{C}^\perp$ is also an MDS array code over $\mathbb{F}_2^3$. Its parameters are $n = 4$, $k = 2$ and $d = 3$ as well. □

Next example shows that the concept of superregular matrix in the sense of [7] is not suitable for this construction.

Example 2: Let $p(x) = x^3 + x^2 + 1 \in \mathbb{F}_2[x]$ then

$$C = \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 1 \end{pmatrix}.$$

The matrix $A = \begin{pmatrix} 1 & 0 \\ 1 & \alpha \end{pmatrix}$ where $\alpha \in \mathbb{F}_{2^3}$ is a primitive element, is a superregular matrix over $\mathbb{F}_{2^3}$ in the sense of [7]. According to Theorem 3 the parity-check matrix of $\mathcal{C}$ is given by

$$H = \left(\begin{array}{cc|c} \begin{matrix} I_3 \\ I_3 \end{matrix} & \begin{matrix} O \\ C \end{matrix} & I_6 \end{array} \right) = \left(\begin{array}{ccc|ccc|ccc|ccc} 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ \hline 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \end{array} \right).$$

In this case, the parameters of the code over $\mathbb{F}_2^3$ are $n = 4$, $k = 2$ and $d = 2$. Consequently, the code is not MDS. □

As we said before, the construction of MDS codes using superregular matrices has been studied by several authors [8, 13]. We can use some of these constructions to extend these results for MDS array codes. For example, a Cauchy matrix over $\mathbb{F}_q$ is an $m \times t$ matrix $A = (\alpha_{ij})$ with $\alpha_{ij} = (x_i - y_j)^{-1}$ where x_i, $y_j \in \mathbb{F}_q$ satisfying the following conditions for $i \in \{1, 2, \ldots, m\}$ and $j \in \{1, 2, \ldots, t\}$:

- $x_i \neq y_j$
- $x_i \neq x_k$ for $k \in \{1, 2, \ldots, m\} \backslash \{i\}$
- $y_j \neq y_l$ for $l \in \{1, 2, \ldots, t\} \backslash \{j\}$

 ISBN: 978-84-614-6167-7

This kind of matrices are superregular matrices [12, 13].

Consider $(u_1, u_2, \ldots, u_m)$ and $(v_1, v_2, \ldots, v_t)$, satisfying the following properties for $i \in \{1, 2, \ldots, m\}$ and $j \in \{1, 2, \ldots, t\}$:

- $u_i, v_j \in \{0, 1, \ldots, q^b - 2\}$
- $u_i \neq v_j$
- $u_i \neq u_k$ for $k \in \{1, 2, \ldots, m\} \backslash \{i\}$
- $v_j \neq v_l$ for $l \in \{1, 2, \ldots, t\} \backslash \{j\}$

If C is the companion matrix of a primitive polynomial of degree b, according to Theorem 3, the matrix $(\mathbf{M}, \mathbf{I}_{bm})$, where $\mathbf{M} = (M_{ij})$, with

$$M_{ij} = (C^{u_i} - C^{v_j})^{-1}$$

for $i = 1, 2, \ldots, m$, $j = 1, 2, \ldots, t$, is the parity-check matrix of a $[m + t, t, m + 1]$ MDS array code $\mathcal{C}$ over $\mathbb{F}_q^b$.

Example 3: For $q = 2$ and $b = 3$, consider $(1, 2)$ and $(4, 5, 6)$. The matrix $\mathbf{M}$ is given by

$$\mathbf{M} = \begin{pmatrix} (C - C^4)^{-1} & (C - C^5)^{-1} & (C - C^6)^{-1} \\ (C^2 - C^4)^{-1} & (C^2 - C^5)^{-1} & (C^2 - C^6)^{-1} \end{pmatrix}.$$

As a result, the matrix $(\mathbf{M}, \mathbf{I}_6)$ is the parity-check matrix of a $[5, 3, 3]$ MDS array code over $\mathbb{F}_2^3$. □

4 Conclusions

In this paper we have introduced a construction of MDS array codes based on superregular matrices. The main idea is replace the elements of a superregular matrix by powers of the companion matrix of a primitive polynomial. The resultant matrix helps us to construct the parity-check matrix of an MDS array code.

5 Acknowledgements

This work was partially supported by Spanish grants MTM2008-06674-C02-01 of the Ministerio de Ciencia e Innovación of the Gobierno de España and ACOMP/2011/005 of the Generalitat Valenciana. The work of the first author was also supported by a grant for research students from the Generalitat Valenciana with reference BFPI/2008/138.

References

[1] M. Blaum, J. Brady, J. Bruck and J. Menon. EVENODD: An efficient scheme for tolerating double disk failures in RAID architectures. *IEEE Transactions on Computers*, **42(2)**: 192–202 (1995).

[2] M. Blaum, J. Bruck and A. Vardy. MDS array codes with independent parity symbols. *IEEE Transactions on Information Theory*, **42(2)**: 529–542 (1996).

[3] M. Blaum, P. G. Farrell and H. C. A. van Tilborg. Array codes. In V. S. Pless and W. C. Huffman (editors), *Handbook of Coding Theory*, pages 1855–1909. Elsevier, North-Holland, 1998.

[4] M. Blaum and R. M. Roth. New array codes for multiple phased burst correction. *IEEE Transactions on Information Theory*, **39(1)**: 66–77 (1993).

[5] M. Blaum and R. M. Roth. On lowest density MDS codes. *IEEE Transactions on Information Theory*, **45(1)**: 46–59 (1999).

[6] Y. Cassuto and J. Bruck. Cyclic lowest density MDS array codes. *IEEE Transactions on Information Theory*, **55(4)**: 1721–1728 (2009).

[7] R. Hutchinson, R. Smarandache and J. Trumpf. On superregular matrices and MDP convolutional codes. *Linear Algebra and its Applications*, **428**: 2585–2596 (2008).

[8] G. Kéri. Types of superregular matrices and the number of n-arcs and complete n-arcs in $PG(r, q)$. *Journal of Combinatorial Designs*, **14(5)**: 363–390 (2006).

[9] R. Lidl and H. Niederreiter. *Introduction to Finite Fields and Their Applications*. Cambridge University Press, New York, NY, 1994.

[10] E. Louidor and R. M. Roth. Lowest density MDS codes over extension alphabets. *IEEE Transactions on Information Theory*, **52(7)**: 46–59 (2006).

[11] F. J. MacWilliams and N. J. A. Sloane. *The Theory of Error-Correcting Codes*. North-Holland, Amsterdam, 6th edition, 1988.

[12] R. M. Roth and A. Lempel. On MDS codes via Cauchy matrices. *IEEE Transactions on Information Theory*, **35(6)**: 1314–1319 (1989).

[13] R. M. Roth and G. Seroussi. On generator matrices of MDS codes. *IEEE Transactions on Information Theory*, **31(6)**: 826–830 (1985).

[14] L. Xu and J. Bruck. X-code: MDS array codes with optimal encoding. *IEEE Transactions on Information Theory*, **45(1)**: 272–276 (1999).

[15] V. Xu, Lihao Bohossian, J. Bruck and D. G. Wagner. Low-density MDS codes and factors of complete graphs. *IEEE Transactions on Information Theory*, **45(6)**: 1817–1826 (1999).

 ISBN: 978-84-614-6167-7

Proceedings of the 11th International Conference on Computational and Mathematical Methods in Science and Engineering, CMMSE 2011 26–30 June 2011.

Varying Laguerre–Sobolev–type orthogonal polynomials: a first approach

Laura Castaño–García[1] **and Juan José Moreno–Balcázar**[1]

[1] *Departamento de Estadística y Matemática Aplicada, Universidad de Almería*

emails: lcastano@ual.es, balcazar@ual.es

Abstract

In this article we summarize some new results about the asymptotic behaviour of the zeros of Sobolev–type orthogonal polynomials with respect to a varying nonstandard inner product. The main result is illustrated with some numerical examples.

Key words: Laguerre polynomials, Sobolev–type orthogonal polynomials, zeros, Bessel functions
MSC 2000: 33C47

1 Introduction

Classical Laguerre polynomials, $L_n^{(\alpha)}(x) = \frac{(-1)^n}{n!}x^n + \ldots$, are orthogonal with respect to the inner product

$$(p,q) = \frac{1}{\Gamma(\alpha+1)}\int_0^\infty p(x)q(x)x^\alpha e^{-x}dx, \quad \alpha > -1. \tag{1}$$

We know many properties of the sequence of orthogonal polynomials $\{L_n^{(\alpha)}\}_n$. One of them is the Mehler–Heine formula (see [5, p.199]), i.e.,

$$\lim_{n\to\infty}\frac{L_n^{(\alpha)}(x/n)}{n^\alpha} = x^{-\alpha/2}J_\alpha(2\sqrt{x}), \tag{2}$$

uniformly on compact subsets of $\mathbb{C}$, where J_α is the Bessel function of the first kind given by

$$J_\alpha(x) = \sum_{n=0}^\infty \frac{(-1)^n}{n!\,\Gamma(n+\alpha+1)}\left(\frac{x}{2}\right)^{2n+\alpha}.$$

 ISBN: 978-84-614-6167-7

Thus, if we denote by $x_{n,i}$ the zeros of the classical Laguerre polynomials $L_n^{(\alpha)}(x)$ ordered as $0 < x_{n,1} < x_{n,2} < \ldots < x_{n,n}$, and we apply Hurwitz's Theorem in (2), then we can deduce

$$\lim_{n\to\infty} n x_{n,i} = \frac{j_{\alpha,i}^2}{4}\,,\ i \geq 1,$$

where $j_{\alpha,i}$ are the positive zeros of J_α ordered as $0 < j_{\alpha,1} < j_{\alpha,2} < \ldots$.

On the other hand, in the early 90s Koekoek and Meijer began to modify the standard inner product (1) by adding derivatives of Dirac delta functions at the point $x = 0$ (see [3] or [4]). Thus, we can consider the nonstandard inner product which is called Sobolev–type inner product

$$(p,q) = \frac{1}{\Gamma(\alpha+1)} \int_0^\infty p(x)q(x)x^\alpha e^{-x}dx + Np'(0)q'(0), \quad N > 0, \quad \alpha > -1. \tag{3}$$

Asymptotic properties of these polynomials were studied in [1], where Mehler–Heine type formulas got relevance because they are able to describe precisely the asymptotic behaviour of the polynomials around the point where we have posed the perturbation. Let us denote by $L_n^{(\alpha,N)}$ the orthogonal polynomials with respect to (3). Then, it was proved in [1] that

$$\lim_{n\to\infty} \frac{L_n^{(\alpha,N)}(x/n)}{n^\alpha} = \frac{1}{\alpha+2}(g_2(x) - (\alpha+2)g_1(x) - g_0(x)), \tag{4}$$

uniformly on compact subsets of $\mathbb{C}$, where $g_i(x) := x^{-\alpha/2} J_{\alpha+2i}(2\sqrt{x})$. It was also proved in [1] that the function

$$h_\alpha(x) = \frac{1}{\alpha+2}(g_2(x) - (\alpha+2)g_1(x) - g_0(x)) \tag{5}$$

has only one negative real zero. Thus, denoting by $y_{n,1} < y_{n,2} < \ldots < y_{n,n}$ the zeros of $L_n^{(\alpha,N)}$, we have

$$\lim_{n\to\infty} n y_{n,i} = h_{\alpha,i},$$

where $h_{\alpha,i}$ denotes the i–th real zero of h_α.

Now, we wonder about what happens if we pose a sequence of masses $\{N_n\}_n$ in the inner product (3) instead of a fixed mass N. Thus, the rest of this paper is devoted to show the changes produced in the Mehler–Heine type formula and in the asymptotic behaviour of the zeros of the orthogonal polynomials with respect to a varying Sobolev inner product.

2 Varying Laguerre–Sobolev orthogonal polynomials

We consider a sequence of nonnegative numbers, $\{N_n\}_n$, such that

$$\lim_{n\to\infty} N_n n^\gamma = N > 0, \quad \gamma \in \mathbb{R}. \tag{6}$$

 ISBN: 978-84-614-6167-7

Then, we introduce the varying inner product

$$(p,q)_n = \frac{1}{\Gamma(\alpha+1)}\int_0^\infty p(x)q(x)x^\alpha e^{-x}dx + N_n p'(0)q'(0), \quad \alpha > -1. \tag{7}$$

We denote by $\{L_n^{(\alpha,N_n)}\}_n$ the sequence of orthogonal polynomials with respect to (7), and we call them varying Laguerre–Sobolev orthogonal polynomials. The leading coefficient of the polynomial $L_n^{(\alpha,N_n)}$ is taken as $\frac{(-1)^n}{n!}$.

Following the paper [4], it is easy to prove

Proposition 1 *We have,*

$$L_n^{(\alpha,N_n)}(x) = A_0(n)L_n^{(\alpha)}(x) + A_1(n)xL_{n-1}^{(\alpha+2)}(x) + A_2(n)x^2L_{n-2}^{(\alpha+4)}(x),\ n \geq 0,$$

where

$$A_i(n) = \frac{B_i(n)}{B_0(n) - nB_1(n) + n(n-1)B_2(n)}, \quad i = 0,1,2, \tag{8}$$

and

$$B_0(n) = 1 - \frac{N_n}{\alpha+1}\binom{n+\alpha+1}{n-2},$$

$$B_1(n) = -\frac{(\alpha+2)N_n}{(\alpha+1)(\alpha+3)}\binom{n+\alpha}{n-2},$$

$$B_2(n) = \frac{N_n}{(\alpha+1)(\alpha+2)(\alpha+3)}\binom{n+\alpha}{n-1}.$$

Therefore, taking limits in (8) and using (6) we get the following result.

Lemma 1 *It holds*

$$\lim_{n\to\infty} A_0(n) = \begin{cases} -\dfrac{1}{\alpha+2}, & \text{if } \gamma < \alpha+3, \\ \dfrac{(\alpha+1)\Gamma(\alpha+4) - N}{D(\alpha,N)}, & \text{if } \gamma = \alpha+3, \\ 1, & \text{if } \gamma > \alpha+3. \end{cases}$$

$$\lim_{n\to\infty} nA_1(n) = \begin{cases} -1, & \text{if } \gamma < \alpha+3, \\ \dfrac{-N(\alpha+2)}{D(\alpha,N)}, & \text{if } \gamma = \alpha+3, \\ 0, & \text{if } \gamma > \alpha+3. \end{cases}$$

$$\lim_{n\to\infty} n^2A_2(n) = \begin{cases} \dfrac{1}{\alpha+2}, & \text{if } \gamma < \alpha+3, \\ \dfrac{N}{D(\alpha,N)}, & \text{if } \gamma = \alpha+3, \\ 0, & \text{if } \gamma > \alpha+3, \end{cases}$$

where $D(\alpha,N) = (\alpha+1)\Gamma(\alpha+4) + N(\alpha+2)$.

 ISBN: 978-84-614-6167-7

3 Mehler–Heine type formula and zeros

The results in the previous section allow us to obtain the Mehler–Heine type asymptotics for the polynomials $L_n^{(\alpha,N_n)}$.

Theorem 1 *Let $\{L_n^{(\alpha,N_n)}\}_n$ be the sequence of orthogonal polynomials with respect to (7), and let $\{N_n\}_n$ be the sequence of nonnegative numbers satisfying (6). We denote $g_i(x) := x^{-\alpha/2}\, J_{\alpha+2i}(2\sqrt{x})$. Then, we get*

$$\lim_{n\to\infty} \frac{L_n^{(\alpha,N_n)}(x/n)}{n^\alpha} =$$
$$\begin{cases} -\dfrac{g_0(x) + (\alpha+2)\, g_1(x) - g_2(x)}{\alpha+2}, & \text{if } \gamma < \alpha+3, \\ \dfrac{((\alpha+1)\Gamma(\alpha+4) - N)\, g_0(x) - N(\alpha+2)g_1(x) + Ng_2(x)}{D(\alpha,N)}, & \text{if } \gamma = \alpha+3, \\ g_0(x), & \text{if } \gamma > \alpha+3, \end{cases}$$

where $D(\alpha,N) = (\alpha+1)\Gamma(\alpha+4) + N(\alpha+2)$.

Remark. There are important differences with respect to the particular case $N_n = N$, for all n. We recover the result for this case taking $\gamma = 0$. For the varying case we have three cases which are related to the asymptotic behaviour of the sequence of masses $\{N_n\}_n$.

- **Case** $\gamma < \alpha+3$. Here we obtain the same Mehler–Heine type formula as the one obtained in [1], that is, we have (4).

- **Case** $\gamma > \alpha+3$. The sequence of the masses does not play any role in this type of asymptotic behaviour, and the Mehler–Heine type formula is the same as the one we know for classical Laguerre orthogonal polynomials given in (2).

- **Case** $\gamma = \alpha+3$. That is the most interesting situation, because it is the transition case between the other two ones. In fact,

$$\lim_{N\to 0} \frac{((\alpha+1)\Gamma(\alpha+4) - N)\, g_0(x) - N(\alpha+2)g_1(x) + Ng_2(x)}{D(\alpha,N)} = g_0(x),$$
$$\lim_{N\to\infty} \frac{((\alpha+1)\Gamma(\alpha+4) - N)\, g_0(x) - N(\alpha+2)g_1(x) + Ng_2(x)}{D(\alpha,N)} = -\frac{g_0(x) + (\alpha+2)\, g_1(x) - g_2(x)}{\alpha+2}.$$

  Thus, in the limit situations ($N \to 0$, and $N \to \infty$) this case transforms itself in one of the other cases in a *smoothing* way.

 ISBN: 978-84-614-6167-7

Therefore, the asymptotic behaviour of the sequence of masses $\{N_n\}_n$ influences on this type of asymptotics.

We denote by $s_{n,1} < s_{n,2} < \ldots < s_{n,n}$ the zeros of $L_n^{(\alpha,N_n)}$. Note that the zeros, $y_{n,i}$, of $L_n^{(\alpha,N)}$ are a particular case of these zeros. Now, applying Hurwitz's Theorem in Theorem 1, we obtain

Corollary 1 *We have,*

- *If* $\gamma < \alpha + 3$,
$$\lim_{n\to\infty} n s_{n,i} = h_{\alpha,i},$$
where $h_{\alpha,i}$ *denotes the i–th real zero of the function* h_α *defined in (5).*

- *If* $\gamma = \alpha + 3$,
$$\lim_{n\to\infty} n s_{n,i} = k_{\alpha,i},$$
where $k_{\alpha,i}$ *denotes the i–th real zero of the function*
$$k_\alpha(x) := ((\alpha+1)\Gamma(\alpha+4) - N)\, g_0(x) - N(\alpha+2)g_1(x) + N g_2(x).$$

- *If* $\gamma > \alpha + 3$,
$$\lim_{n\to\infty} n s_{n,i} = \frac{j_{\alpha,i}^2}{4},$$
where $j_{\alpha,i}$ *are the positive zeros of* J_α.

All these results appearing here have been obtained for the Doctoral Dissertation of L. Castaño–García (see [2]). To finish, we show some numerical experiments. In all the tables $N_n = \frac{N}{n^\gamma}$.

Table 1: $n\, s_{n,i}$, for $\alpha = 2$, $N = 2.5$, and $\gamma = 2$.

	$ns_{n,1}$	$ns_{n,2}$	$ns_{n,3}$	$ns_{n,4}$
n=50	-4.67417494	11.78361837	27.67607672	48.18603357
n=100	-4.65064916	11.99470471	28.09657336	48.86487599
n=150	-4.64142765	12.06761219	28.24354141	49.10595374
n=300	-4.63183332	12.14180148	28.39403289	49.35478356
Limit	$\mathbf{h_{2,1}} = -4.62207114$	$\mathbf{h_{2,2}} = 12.21727721$	$\mathbf{h_{2,3}} = 28.54812610$	$\mathbf{h_{2,4}} = 49.61159809$

Acknowledgements

This research was supported by MICINN of Spain under Grant MTM2008-06689-C02-01 and Junta de Andalucía (FQM229 and P09–FQM–4643).

 ISBN: 978-84-614-6167-7

Table 2: $n\, s_{n,i}$, for $\alpha = 2$, $N = 2.5$, and $\gamma = 5$.

	$ns_{n,1}$	$ns_{n,2}$	$ns_{n,3}$	$ns_{n,4}$
n=50	6.32483175	17.08681030	32.67713219	53.09435429
n=100	6.41954598	17.33635506	33.13808389	53.81364179
n=150	6.45191350	17.42227844	33.29867796	54.06832509
n=300	6.48469330	17.50963439	33.46292274	54.33088113
Limit	$\mathbf{k_{3,1}} = 6.51789433$	$\mathbf{k_{3,2}} = 17.5984621$	$\mathbf{k_{3,3}} = 33.63093373$	$\mathbf{k_{3,4}} = 54.60158096$

Table 3: $n\, s_{n,i}$, for $\alpha = 2$, $N = 2.5$, and $\gamma = 6$.

	$ns_{n,1}$	$ns_{n,2}$	$ns_{n,3}$	$ns_{n,4}$
n=50	6.40166849	17.20456306	32.80572808	53.22742477
n=100	6.49586617	17.45226673	33.26452025	53.94442724
n=150	6.52805175	17.53756902	33.42438700	54.19834120
n=300	6.56064474	17.62430027	33.58790080	54.46012418
Limit	$\frac{\mathbf{j_{2,1}^2}}{\mathbf{4}} = 6.59365410$	$\frac{\mathbf{j_{2,2}^2}}{\mathbf{4}} = 17.71249972$	$\frac{\mathbf{j_{2,3}^2}}{\mathbf{4}} = 33.75517721$	$\frac{\mathbf{j_{2,4}^2}}{\mathbf{4}} = 54.73004728$

References

[1] R. Álvarez–Nodarse, J. J. Moreno–Balcázar, *Asymptotic properties of generalized Laguerre orthogonal polynomials*, Indag. Math. (N.S.) **15**(2) (2004) 151–165.

[2] L. Castaño–García, *Aportaciones a la teoría asintótica de polinomios ortogonales de Sobolev*, Universidad de Almería, 2010.

[3] R. Koekoek, *Generalizations of Laguerre polynomials*, J. Math. Anal. Appl. **153**(2) (1990) 576–590.

[4] R. Koekoek, H. G. Meijer, *A generalization of Laguerre polynomials*, SIAM J. Math. Anal. **24**(3) (1993) 768–782.

[5] G. Szegő, *Orthogonal Polynomials*, Amer. Math. Soc. Colloq. Publ. vol. 23, Amer. Math. Soc., Providence, RI, 1975.

 ISBN: 978-84-614-6167-7

Proceedings of the 11th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2011
26–30 June 2011.

An Efficient Locality P2P Computing Architecture*

Damià Castellà[1], Francesc Solsona[1] and Francesc Giné[1]

[1] *Computer Science Department, University of Lleida*

emails: {dcastella, francesc, sisco}@diei.udl.cat

Abstract

This paper proposes a distributed computing architecture, named *DisCoP*, based on the P2P paradigm. Our proposal gathers the peers into markets according to their available multi-attribute computational resources. A Hilbert function is used to arrange multi-attribute markets in an ordered and mono-dimensional space. Each market is internally grouped in an N-tree, which are linked by a Bruijn graph. The tree topology allows efficient searching of available resources in a specific market, while Bruijn provides good scalability as searching complexity does not depend on the number of markets. This way, the proposed architecture exploits the Bruijn and Hilbert functions to provide the system with a good locality. The locality feature over a market overlay allows a lack of resources in a given market to be satisfied quickly by any other market with similar resources, whenever they are closer to each other. Consequently, locality concern arises as an essential challenge. According to this, a new procedure to measure locality is carried out, together with an extensive analysis of the DisCoP locality in relation to others overlays. Our results reveal the good performance of our proposal with gains of up to 80% compared with others.

Key words: P2P computing, P2P topologies, locality.

1 Introduction

P2P computing is a distributed computing paradigm that uses Internet to connect thousands or even millions of users into a single large virtual computer based on the sharing of computational resources [4]. Although the P2P paradigm can not hope to serve as a totally

*This work was supported by the MEyC-Spain under contract TIN2008-05913, TIN2010-12011-E and the CUR of DIUE of GENCAT and the European Social Fund

 ISBN: 978-84-614-6167-7

general-purpose efficient parallel computer, it can still serve as an excellent platform with unlimited computational resources for solving a wide variety of computational applications. In order to schedule and execute these applications efficiently, a P2P computing platform needs a mechanism to search and manage the set of peers, whose available computational resources (i.e. CPU, Memory or Bandwidth) fit the requirements of the application to be executed. Taking into account both the large-scale and the mutable amount of computational resources provided by each peer, resource management in P2P computation becomes a research challenge [9].

This paper proposes a reliable and scalable platform with a three-layer architecture, orientated to P2P computing, named DisCoP (Distributed Computing Platform). The top layer is a Hilbert SFC, which allows peers to be classified into markets according to their computational attributes. Markets are interconnected by means of a Bruijn graph in the second layer. Finally, the nodes of the same market are arranged with a tree topology in the bottom layer.

Grouping peers with similar computational attributes in markets significantly reduces the search time for the requested computational resources and minimizes unnecessary outgoing queries throughout the network. Unfortunately, the lookup queries are totally unbalanced, as those markets made up of peers with the most popular computational resources are those required most by the platform's users. Thus, the peers of the most popular markets are rapidly busied and, as a consequence, a huge avalanche of unsuccessful queries appears in the system. P2P computing platforms should accordingly provide alternative resources with capabilities as close as possible to those requested. So locality, thought of in terms of market proximity, is a key aspect to consider in the design of structured networks for P2P computing systems. We are interested on the locality defined in [8] as a value that indicates the closeness of the markets with similar computational resources (called *similar markets* henceforth) mapped into a multi-dimensional space using the distances provided by a specific overlay.

To this end, our paper analyzes the benefits of providing proximity or locality to a P2P computing environment. In order to do so, a procedure to measure the degree of locality of any topology is also provided. Based on this, we analyze the locality features of our proposal in relation to the most widely used topologies, such as Chord [17]. In addition, we optimize the linkage of similar markets over the DisCoP architecture by means of adding extra-links, which allow the Hilbert keys (markets) to run throughout the Bruijn topology. Our simulation results reveal that the interaction between the three levels of DisCoP architecture (Hilbert, Bruijn and Tree levels) together with the added extra-links allows the best degree of market locality to be achieved. It means that our proposal locates similar computing resource providers (markets) as close as possible.

The outline of this paper is as follows; the related work is described in Section 2. Section 3 introduces the proposed DisCoP architecture. The metrics and a procedure for obtaining

 ISBN: 978-84-614-6167-7

overlay locality is proposed in Section 4. Finally, the main conclusions and future work are explained in Section 5.

2 Related Work

Recently, several researchers have dedicated their efforts to developing new frameworks oriented to P2P computing [11, 6, 15, 14, 1]. To our knowledge, the CompuP2P architecture proposed by Gupta et al. [6] is the most closely related to our work. CompuP2P creates dynamic markets of network accessible computing resources, connected by means of a Chord network [17]. Chord is a distributed lookup algorithm based on DHT to locate a node that stores a particular resource. Unlike the Bruijn graph used in our proposal, the Chord overlay network does not have a constant degree and as a consequence, the communication speed is degraded. Likewise, the Chord protocol is not well suited to multi-attribute range queries since hashing does not preserve locality of data.

To address the locality problem, P2P systems such as SkipNet [7], Graps [16] and Brocade [18] have emerged [8]. In SkipNet [7], the storing position of resources was limited and a locality-based routing mechanism proposed. Because the locality information is stored into the resource name, it is not physical location-based locality but a logical location-based locality. Like in SkipNet, the nodes of the DisCoP platform have logical locations determined by the Hilbert ordering.

In Graps [16], the authors proposed a hierarchical virtual network to support lookup services. The hierarchical virtual network consists of sub-networks composed of physically close nodes. The leaders of each sub-network are interconnected by means of a super-network. As the resources and nodes in a sub-network increase, the leaders of such sub-network sacrifice load-balancing in the super-network. To solve imbalance, the DisCoP platform arranges the peers into markets according to their available computational resources, independently of their physical location. But, when the markets are filled, the nodes are inserted into similar markets according to the locality metric proposed in this paper.

In Brocade [18], a secondary overlay that exploits knowledge of the underlying network features, such as bandwidth and capacity, is presented. The secondary overlay builds a location layer between super-nodes. By associating local nodes with their nearby super-node, messages delivered across the wide-area can take advantage of the highly connected network infrastructure between these super-nodes, greatly reducing point-to-point routing distance and network bandwidth usage. We expand the Graps and Brocade idea by connecting the super-peer nodes though a Bruijn graph. In our proposal, super-peers act as managers of the the bottom layer. The existence of super-peers, who manage a group of peers (in our case, a market tree) and facilitates trade-off between peers on structured P2P networks, is an essential issue to be considered for optimizing the locality feature [5].

 ISBN: 978-84-614-6167-7

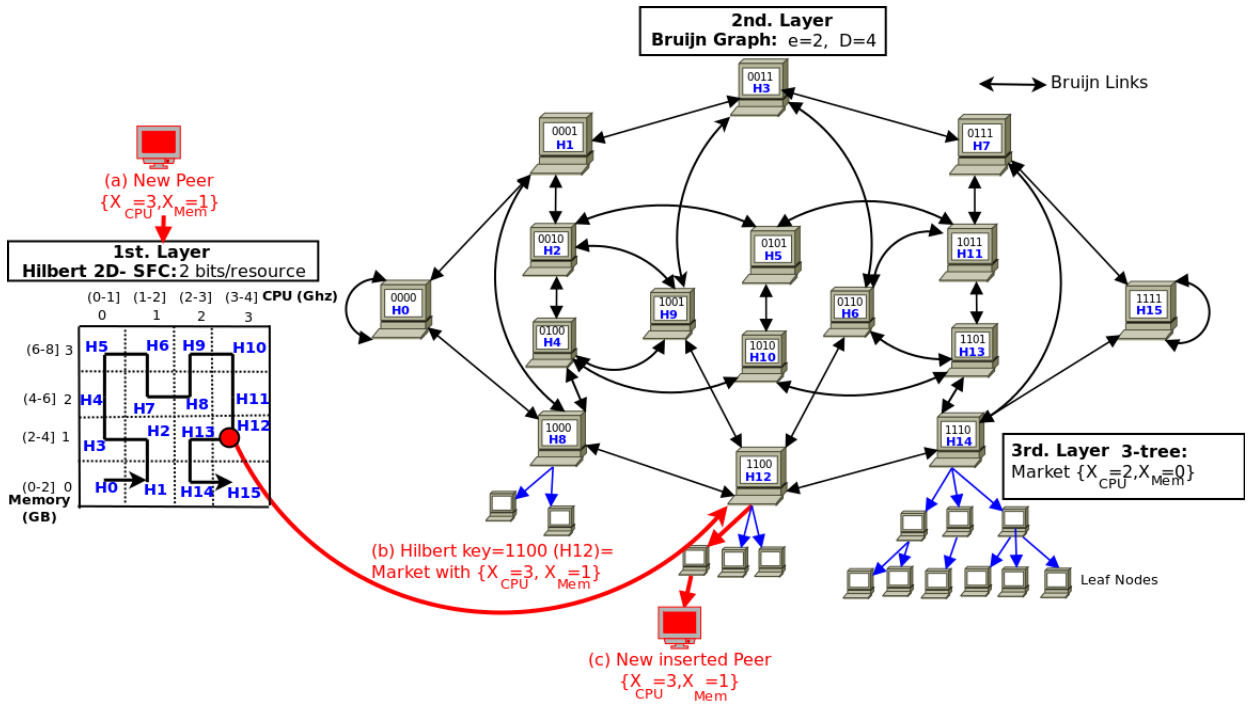


Figure 1: DisCoP system architecture.

3 The DisCoP Architecture

Our proposal, named DisCoP (Distributed Computing Platform), is made up of three different layers (Hilbert, Bruijn and Tree). This is depicted in Fig.1. The top layer is a Hilbert SFC, which allows peers to be classified into markets according to their computational resources. In general, each market M_i of the overlay is identified by k-dimensional coordinates $(X_1, .., X_i, .., X_k)$, where k is the number of computational attributes and $X_i \in \mathbb{Z}^+$ is the coordinate of the i^{th} attribute in the range $[0, X_i^{max}]$. Likewise, it is worth pointing out that each coordinate X_i has an associated range of values V_i^j of such i^{th} attribute, where $0 < V_i^j \leq V_i^{max}$, V_i^{max} being the maximum value to be taken by the i^{th} attribute. Fig. 1 (first layer) shows an example of the top layer with $k = 2$ attributes, CPU and Memory, with a range of values for CPU and Memory of $(0, 4Ghz]$ and $(0, 8GB]$, respectively. Therefore, V_{CPU}^{max} and V_{Memory}^{max} would be $4Ghz$ and $8GB$, respectively. As we can see in the same Figure, the top layer of DisCoP associates a Hilbert key H_i to each market M_i. Next, these Hilbert keys are mapped into the second layer of DisCoP through a Bruijn graph. Finally, a market of nodes, which are arranged by means of a tree topology, hangs from each Bruin node. The choice of each level of the DisCoP overlay is discussed and justified below.

The first layer of DisCoP uses a **Hilbert Multi-Dimensional Space-Filling Curve**

 ISBN: 978-84-614-6167-7

Graph	Degree	Diameter
Baton	k+1	$2log_kN$
Chord	log_2N	log_2N
Pastry	$(b-1)log_bN$	log_bN
Bruijn	k	log_kN
Viceroy	k	$2log_kN(1-o(1))$

Table 1: Degree-diameter complexities.

(SFC) to map the k-dimensional coordinate of each market into a 1D domain identifier, H_i. As Fig. 1 (first layer) shows, an SFC is a thread that goes through all the points in a space while visiting each point once. In the P2P computing case, it means that the k-attributes (computational resources) provided by a node are mapped into a single key, which represents the identification of the market where the requested resources are available. A simple example is also depicted in Fig. 1 (first layer). A new peer characterized by the 2-tuple $\{X_{CPU} = 3, X_{Mem} = 1\}$ wishes to be inserted into the system and the Hilbert SFC returns the $key = H_{12}$, which identifies such a market with nodes characterized by the 2-tuple $\{3, 1\}$ (see Fig. 1 (second layer)). It is worth pointing out that our system assumes that the attribute values are obtained with a specific benchmark tool. The Hilbert SFC was chosen for our purposes due to the fact that it achieves a better clustering property [13] than other SFCs widely used in the literature, such as Sweep, Scan, Peano and Gray [12]. Note that clustering means that the locality between objects in the multi-dimensional space is also preserved in the linear space. Thus, the SFC key-mapping in the Bruijn graph is made easier.

The second layer of DisCoP uses a **Bruijn graph** to arrange markets. The Bruijn graph is a directed graph with e outgoing and incoming edges at each node (market). Each node has a unique and fixed length key H_i. The maximum number of markets is $N = e^D$, where D is the diameter (maximum distances between any two markets). A classic Bruijn for $e = 2$ and $N = 16$ is shown in Fig.1 (second layer). From this Figure, we can see the routing algorithm followed by Bruijn. A node with key H_i is linked to two nodes with keys $2H_i$ and $2H_i + 1$. Bruijn was chosen from among the most widely used P2P topologies, [9, 10], such as Chord, Baton, Pastry, Bruijn and Viceroy, because it offers the best diameter-degree and connectivity. This is shown in Tables 1 and 2, where Table 1 shows the asymptotic degree-diameter properties of the different graphs and Table 2 its graph diameter for a maximum number of nodes $N = 10^6$.

The third level of DisCoP is made up by **N-ari Trees**, which are used to arrange the nodes belonging to the same market. This topology was chosen due to its low look-up complexity ($\log(N_{Tree})$), where N_{Tree} is the total number of peers, its constant degree and its hierarchical topology for managing and maintaining the system-growing capacity. Thus,

 ISBN: 978-84-614-6167-7

Graph/k	2	3	4	10	20
Baton	-	40	26	13	10
Chord	-	-	-	-	20
Pastry	-	-	-	-	20
Bruijn	20	13	10	6	5
Viceroy	31	20	16	10	8

Table 2: Diameter ($N = 10^6$ nodes).

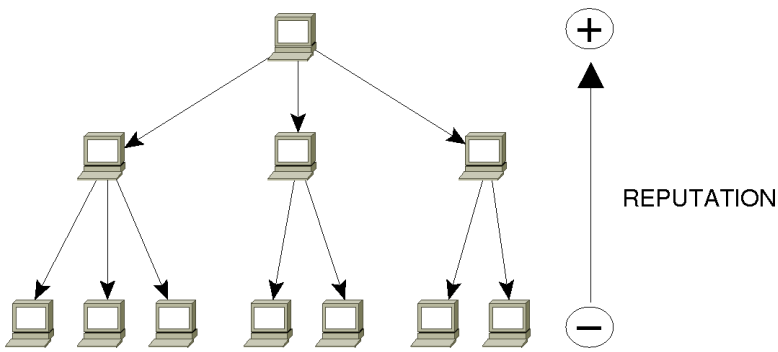


Figure 2: 3-Tree.

the good scalability of the system is favored. Likewise, as we can see in Fig. 2, nodes with high reputation are located on the higher levels of the tree, close to the root, whereas nodes with low reputation are located instead on the lower levels. We define reputation as the time elapsed since the entry of a node into the system. Note that the root node of the tree obtains information about all the available resources in the market, which is provided by the DisCoP maintenance system, explained in detail in [2]. Their main drawback is the congestion of the nodes near the root. Because of this, our architecture assumes that the size of the tree (market size) is limited. In the same way, our architecture assumes a huge number of different markets connected through the Bruijn graph.

Given that, our previous works tackled the main functionalities related to the management of the peers (insertion, maintenance and departure of peers) at the tree level [3, 2], the current paper is mainly focused on the interaction between the Hilbert SFC and the Bruijn graph.

3.1 Adding extra-links

In order to better map the Hilbert keys (market keys) over the Bruijn graph, some extra-links have been added in the Bruijn overlay. The three layers described above, together

 ISBN: 978-84-614-6167-7

with the extra-links make up the proposed system architecture of the *DisCoP* platform.

The Hilbert function assigns contiguous keys to markets with similar resources. For instance, according to the example in Fig. 1, the market with tuple {0,1} has a key H_3, whereas the market with tuple {0,2} has a key H_4. In order to maintain this locality in the Bruijn graph, a new bi-directional link (called extra-link) was added between two markets with contiguous Hilbert keys. It means that market H_2 will be linked ($\longleftrightarrow$) to market H_3 and $H_3 \longleftrightarrow H_4 \longleftrightarrow H_5 \longleftrightarrow H_6 \longleftrightarrow$ .... and so on. This way, the platform can perform searches for similar attributes with very few hops, taking advantage of the clustering Bruijn properties.

4 Locality

One of the most desired properties of a P2P overlay is its locality, given that a high locality will increase the efficiency in the searching of resources throughout the overlay. According to this, this section is devoted to defining a procedure for obtaining the *overlay locality* metrics in a given P2P topology. This procedure is applied to measuring and analyzing the locality of our proposal (Hilbert and Bruijn) in relation to other two-level overlays, composed by Hilbert and Chord. Moreover, we will analyze the locality performance of these overlays when extra-links, described in Section 3.1, are added.

4.1 Overlay Locality Metrics

Taking the described concepts into account, the procedure for obtaining the *Locality metrics* for a given overlay $\mathcal{O}$ is as as follows:

1. For each market M_i, calculate its set of **Similar Markets** SM_i. Each market M_i has an associated set of Similar Markets, denoted as SM_i, so that two markets are *similar* whenever their coordinates differ by exactly one unit in a single attribute. Thus, each market M_i will have two similar markets (*previous* and *next*) for each attribute, except when the attribute takes an extreme value, 0 or V^{max}, that in both cases only have one similar market, next or previous, respectively. This is depicted in Table 3.

2. For each market M_i, obtain the **Similar Market Distance** (SMD_i), defined as:

$$SMD_i = \sum_{k=0}^{|SM_i|} \frac{d_{\mathcal{O}}(M_i, M_k)}{|SM_i|}, \tag{1}$$

   where $|SM_i|$ is the size of the set SM_i and $d_{\mathcal{O}}(M_i, M_k)$ is the distance in hops of the overlay $\mathcal{O}$ between market M_i and its similar market $M_k \in SM_i$. It is worth pointing out that this metric depends totally on the overlay $\mathcal{O}$. This way, whenever

 ISBN: 978-84-614-6167-7

Previous similar market for the i$^{\text{th}}$ attribute
$\begin{cases} (X_1, .., X_i - 1, .., X_k) & if\ X_i > 0 \\ (X_1, .., 0, .., X_k) & if\ X_i = 0 \end{cases}$

Next similar market for the i$^{\text{th}}$ attribute
$\begin{cases} (X_1, .., X_i + 1, .., X_k) & if\ X_i < V_i^{max} \\ (X_1, .., V_i^{max}, .., X_k) & if\ X_i = V_i^{max} \end{cases}$

Table 3: Obtaining the *previous* and *next* similar markets for the i$^{\text{th}}$ attribute of market $(X_1, .., X_i, .., X_k)$.

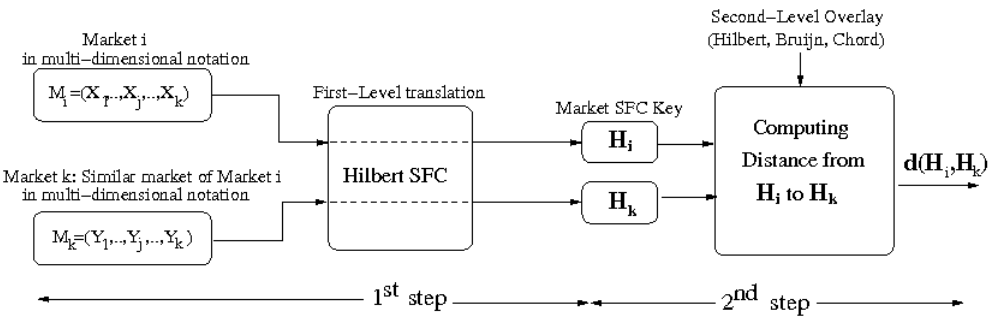


Figure 3: Locality process for a two-level overlay.

the overlay is composed of two levels, such as the DisCoP case, this calculation will imply two different steps. The first one converts the multi-dimensional coordinates of markets M_i and M_k into one-dimensional coordinates, H_i and H_k respectively, by means of a SFC Hilbert function. Next, the distance between H_i and H_k, $d_{\mathcal{O}}(H_i, H_k)$, is calculated according to the topology of the second overlay. This process is depicted in Fig. 3.

3. For a given overlay $\mathcal{O}$, calculate the **Mean Market Distance** ($MMD_{\mathcal{O}}$), defined as the mean of the Similar Market Distances (SMD_i). This metric provides a way to assess the absolute locality in function of the distance between similar markets of an overlay. Thus, the locality increases with the Mean Market Distance. Formally:

$$Mean\ Market\ Distance_{\mathcal{O}} = \sum_{\text{i}} \frac{SMD_i}{N_{\mathcal{O}}}, \tag{2}$$

where $N_{\mathcal{O}}$ is the number of markets in the overlay $\mathcal{O}$ and $1 \leq i \leq N_{\mathcal{O}}$.

4. Compute the **Relative Locality** ($RL_{\mathcal{O}}$) of an overlay, defined as:

$$Relative\ Locality_{\mathcal{O}} = \frac{MMD_{\mathcal{O}}}{D_{\mathcal{O}}}, \tag{3}$$

where $D_{\mathcal{O}}$ is the maximum distance between any pair of markets in the overlay $\mathcal{O}$. This is the overlay diameter. Note that the $RL_{\mathcal{O}}$ is not normalized because this metric varies in the range $\left[\frac{1}{D_{\mathcal{O}}}, 1\right]$.

5. Calculate the **Normalized Locality** ($NL_{\mathcal{O}}$) of an overlay $\mathcal{O}$, defined as:

$$Normalized\,Locality_{\mathcal{O}} = 1 - \frac{RL_{\mathcal{O}} - \frac{1}{D_{\mathcal{O}}}}{1 - \frac{1}{D_{\mathcal{O}}}}. \tag{4}$$

Therefore, the values of the $NL_{\mathcal{O}}$ metric are normalized in the range $[0, 1]$; this is:

- $NL_{\mathcal{O}} = 0$ means that the locality degree of the overlay $\mathcal{O}$ is null. In this case, all elements would be at a distance between them equal to $D_{\mathcal{O}}$.
- $NL_{\mathcal{O}} = 1$ means that the degree of locality is maximum. It indicates that the distance between a given market M_i and all its similar markets is equal to 1.

It is worth pointing out that $MMD_{\mathcal{O}}$ gives the real proportions of the magnitude of locality, while $NL_{\mathcal{O}}$ is better for comparing performance between different overlays or combinations of these.

4.2 Locality Results

In this section, we show how to apply the procedure described above to analyzing the locality of the DisCoP two-level overlay, made up of the Hilbert and Bruijn overlays, in relation to other referential two-level topologies, composed of overlays of the Hilbert and Chord. In addition, we analyze the improvement of the DisCoP locality with the addition of the extra-links, described in Section 3.1, to the Bruijn graph. Fig. 4 shows an example of the analyzed overlays with 4 nodes each and 2 links per node. Note that the nodes are represented by the tuple X_1X_2/H_i, where X_1X_2 represents the 2-dimensional notation and H_i the order in the uni-dimensional Hilbert space. These overlays are scaled by means of increasing the number of attributes, from 2 to 5, and bits per attribute, between 1 and 5. So, the maximum number of markets was $N = 3355443 = (2^5)^5$ markets. The locality performance of each overlay is measured by using the **Mean Market Distance** (**MMD**) and the **Normalized Locality** (**NL**) metrics.

Fig. 5(right) and (left) show the MMD results obtained for the HB (Hilbert+Bruijn) two-level overlay, with and without applying the extra-links, respectively. Notice that HB topology with extra-links corresponds to the DisCoP case. Comparing both figures shows how DisCoP obtains an average improvement of 36.7% due to the use of the extra-links.

 ISBN: 978-84-614-6167-7

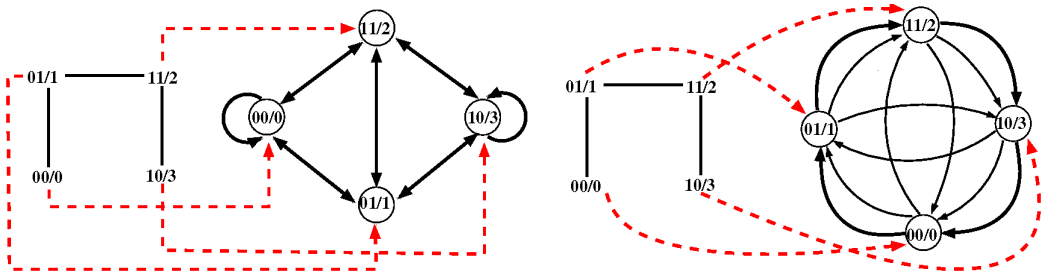


Figure 4: Two-level overlay: (left) Hilbert and Bruijn, (right) Hilbert and Chord.

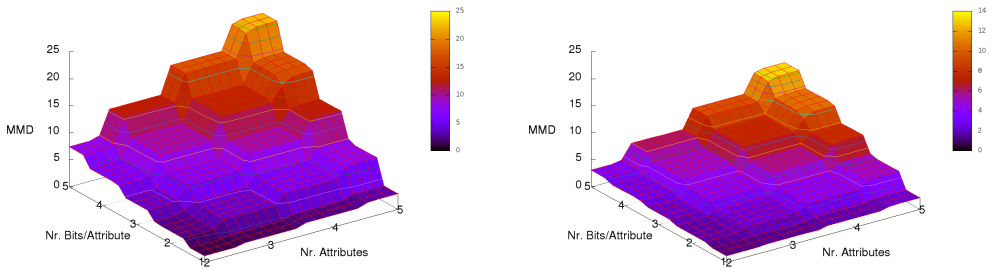


Figure 5: $MMD_{HB/DisCoP}$ Locality: (left) HB topology and (right) DisCoP (HB with extra-links) topology.

Fig. 6(right) and (left) show the obtained MMD results for HC (Hilbert+Chord) two-level overlay, with and without applying the extra-links, respectively; and under the same conditions described above. In this case, the application of the extra-links again increases the locality performance by an average of 45.7%.

Finally, the average gains for the NL metric, between all the two-level overlays described above, are shown in Table 4. Thus, each of the overlays in each row is compared with all those in each of the columns. In general, we can see that the Hilbert+Bruijn (HB) topology always behaves better than the Hilbert+Chord (HC) combination. This proves that similar multi-attributes are always better mapped in Bruijn than Chord. In addition, the Bruijn gain is obtained with lower links per node, 2, than with Chord, which has $log_2 N_{HC}$ links per node. In this sense, it is worth remarking that Bruijn locality improves drastically with the addition of new links per node (Bruijn degree). Likewise, the application of the extra-links produces a remarkable gain for both topologies; although this effect is slightly higher in the HC case, which means that Hilbert locality counteracts the lower Chord locality.

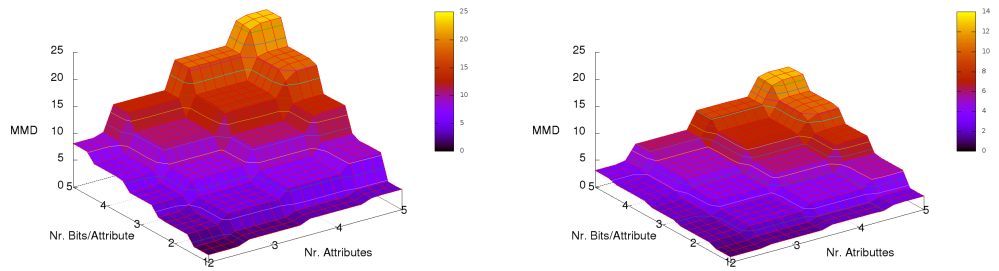


Figure 6: MMD_{HC} Locality. (left) HC topology, (right) HC topology with extra-links.

Row vs Column	Normalized Locality (NL) Gain (in %)			
	HB	HB with extra-links	HC	HC with extra-links
HB		-64,2%	14,2%	-62,9%
HB with extra-links	36,7%		46,3%	0,98%
HC	-17,5%	-91,15%		-89,0%
HC with extra-links	35,8%	-1,3%	45,7%	

Table 4: Normalized Locality gains between different two-level overlays.

5 Conclusions and Future Work

In this paper, a computing resource management system oriented to P2P computing (DisCoP) is presented. Our proposal is based on an architecture made up of three different layers: the top layer is a *Hilbert function*, which classifies the nodes into markets according to their computational multi-attribute resources, the medium layer is a *Bruijn graph*, which links the different markets into a cohesive system and the bottom layer is a set of *trees* (markets), where each tree gathers those nodes with similar resources. In order to optimize the Hilbert-Bruijn interaction, some extra-links between two markets with contiguous Hilbert keys were also added to the Bruijn graph.

A procedure to obtain the locality in P2P topologies is also presented. In doing so, two metrics to measure the locality degree are proposed. Both proposals have been used to measure the good election of the DisCoP architecture in terms of locality, defined in this case in terms of neighboring proximity of the Bruijn nodes. This in turn is equivalent to measure the proximity of markets. Likewise, our results reveal the locality improvement produced by the addition of the extra-links to our two-level architecture.

The future trend is directed towards achieving market balancing depending on the popularity of using of computational resources. Market trees and consequently the third level of DisCoP architecture would therefore be balanced.

References

[1] N. Al-Dmour and W. Teahan. Parcop: a decentralized peer-to-peer computing system. In *Parallel and Distributed Computing, 2004. Third International Symposium on/Algorithms, Models and Tools for Parallel Computing on Heterogeneous Networks*, pages 162 – 168, 2004.

[2] D. Castellà, J. Blanco, F. Giné, and F. Solsona. A computing resource discovery mechanism over a p2p tree topology. In *Proceedings of the 9th international conference on High performance computing for computational science*, VECPAR'10, pages 366–379, 2011.

[3] D. Castellà, J. Rius, I. Barri, F. Giné, and F. Solsona. Codip2p: a new p2p architecture for distributed computing. In *Conference on Parallel, Distributed and Network-based Processing (PDP 2009)*, pages 323–329, 2009.

[4] I. Foster and A. Iamnitchi. On death, taxes, and the convergence of peer-to-peer and grid computing. In *2nd International Workshop on Peer-to-Peer Systems (IPTPS'03)*, pages 118–128, 2003.

[5] L. Guo, S. Jiang, L. Xiao, and X. Zhang. Exploiting content localities for efficient search in p2p systems. In *Proceedings of the 18th International Symposium on Distributed Computing, (DISC 2004)*, pages 729–742, 2004.

[6] R. Gupta, V. Sekhri, and A. K. Somani. Compup2p: An architecture for internet computing using peer-to-peer networks. *IEEE Transactions Parallel Distributed Systems*, 17(11):1306–1320, 2006.

[7] N. J. A. Harvey, M. B. Jones, S. S. adn M. Theimer, and A. Wolman. Skipnet: A scalable overlay network with practical locality properties. In *Proceedings of 4th USITS*, pages 113–126, 2003.

[8] K. il Jeong, U. H. Yoon, J.-Y. Han, J.-M. Ahn, J.-H. Song, and S.-D. Kim. Rnet: A hierarchical p2p overlay network for improving locality in a mobile environment. *Networked Computing and Advanced Information Management, International Conference on*, 1:623–630, 2008.

 ISBN: 978-84-614-6167-7

[9] E. K. Lua, J. Crowcroft, M. Pias, R. Sharma, and S. Lim. A survey and comparison of peer-to-peer overlay network schemes. *IEEE Communications Surveys and Tutorials*, 7:72–93, 2005.

[10] E. Meshkova, J. Riihijärvi, M. Petrova, and P. Mähönen. A survey on resource discovery mechanisms, peer-to-peer and service discovery frameworks. *Comput. Netw.*, 52(11):2097–2128, 2008.

[11] J. Mishra and S. Ahuja. P2pcompute: A peer-to-peer computing system. In *Collaborative Technologies and Systems, (CTS 2007)*, pages 169 –176, May 2007.

[12] M. F. Mokbel, W. G. Aref, and I. Kamel. Analysis of multi-dimensional space-filling curves. *Geoinformatica*, 7(3):179–209, 2003.

[13] B. Moon, H. V. Jagadish, C. Faloutsos, and J. H. Saltz. Analysis of the clustering properties of the Hilbert space-filling curve. *IEEE Transactions on Knowledge and Data Engineering*, 13:2001, 1996.

[14] K. N., M. Vapa, M. Weber, J. Toyryla, and J. Vuori. P2pdisco - java distributed computing for workstations using chedar peer-to-peer middleware. In *Proceedings of the 19th IEEE International Parallel and Distributed Processing Symposium (IPDPS'05) - Workshop 5 - Volume 06*, 2005.

[15] L. Senger, M. de Souza, and D. Foltran. Towards a peer-to-peer framework for parallel and distributed computing. In *Computer Architecture and High Performance Computing (SBAC-PAD)*, pages 127 –134, 2010.

[16] K. Shin, S. Lee, G. Lim, H. Yoon, and J. Ma. Sgrapes: topology-based hierarchical virtual network for peer-to-peer lookup services. In *Proceedings of International Conference on Parallel Processing Workshops*, pages 159–156, 2002.

[17] I. Stoica, R. Morris, D. Karger, M. F. Kaashoek, and H. Balakrishnan. Chord: A scalable peer-to-peer lookup service for internet applications. In *SIGCOMM 2001*, pages 149–160, August 2001.

[18] B. Y. Zhao, Y. Duan, and L. Huang. Brocade: Landmark routing on overlay networks. In *Proceedings of the 1st International Workshop on Peer-to-Peer Systems (IPTPS)*, pages 34–44, 2002.

 ISBN: 978-84-614-6167-7

Proceedings of the 11th International Conference on Computational and Mathematical Methods in Science and Engineering, CMMSE 2011 26–30 June 2011.

Normal S-P Plots and Distribution Curves

Sonia Castillo-Gutiérrez[1], Emilio Lozano-Aguilera [1] and María Dolores Estudillo-Martínez[1]

[1] *Department of Statistics and Operations Research, University of Jaen*

emails: socasti@ujaen.es, elozano@ujaen.es, mdestudi@ujaen.es

Abstract

In this paper a procedure to construct distribution curves on Normal S-P Plots is provided. These distribution curves are useful to identify viable alternative probability models if the proposed distribution for the sample observations is rejected.

Key words: Normal S-P Plots, Distribution Curves

1 Introduction

The Normal Stabilized-Probability Plot or Normal S-P Plot is a probability plot used to assess the normality of a data set.

If the hypothesis of normality is rejected, it would be useful to have a procedure to determine an alternative probability model for the sample data. For this reason the distribution curves appear.

2 Normal S-P Plots

Let $\{x_1, x_2, ..., x_n\}$ be a simple random sample of size n from a distribution $F(x)$ and the ordered observations $x_{(i)}$ $i = 1, ..., n$. Let Φ be the standard normal distribution function, that is, that with which the distribution of observations is compared.

The aim of the S-P Plots or Stabilized-Probability Plot [1] is to stabilize the variance of the points in some probability plots, like P-P Plots.

The S-P Plot appears as a transformation of the P-P Plots to stabilize the variance of the plotted points, i.e., in this type of plot the above mentioned variances are approximately equal.

In S-P Plots we represent the values

$$r_i = \left(\frac{2}{\pi}\right) \arcsin(\sqrt{p_i}) \quad i = 1, ..., n$$

 ISBN: 978-84-614-6167-7

against

$$s_i = \left(\frac{2}{\pi}\right) \arcsin(\sqrt{u_i}) \quad i = 1, ..., n$$

where

$$u_i = \Phi\left(\frac{x_{(i)} - \mu}{\sigma}\right) \quad i = 1, ..., n$$

and p_i is an appropriate plotting position.

If the theoretical distribution (normal distribution) is a good approximation of the empirical distribution, plotted points are arranged around the bisectrix of the first quadrant, i.e., about the line $y = x$ defined between 0 and 1.

In this paper, the plotting position p_i proposed by Weibull [2] in (1939) is used[1]:

$$p_i = \frac{i}{n+1} \quad i = 1, ..., n$$

3 Distribution Curves for S-P Plots

As stated above, by using the Normal S-P Plot we can assess whether a set of observations has a normal distribution. If the plotted points do not follow a straight configuration, that indicates that the sample data do not have a normal distribution.

To determine an alternative probability model for the data set, Gan, Koehler and Thompson (1991) [3] constructed the so-called "distribution curves"applied to the P-P Plots. These curves allow us to propose an alternative probability model for observations when the normal distribution is rejected. The idea is to include in the graph different distribution curves and to study whether the plotted points are close to any of those curves, so that, it is possible to visually choose an appropriate probability model.

In this paper, we propose to extend the concept of distribution curves to the S-P Plots.

The procedure to construct a G distribution curve in a Normal S-P Plot for a location-scale family is the following [4]:

1. Construct a Normal S-P Plot with the sample observations.

2. Select a number k and compute the k plotting position p_i, for example, by using the following expression:

$$p_i = \frac{i}{k+1} \quad i = 1, ..., k$$

3. Apply the transformation to stabilize the variance:

$$r_i^* = \left(\frac{2}{\pi}\right) \arcsin(\sqrt{p_i}) \quad i = 1, ..., k$$

[1]Although in this paper we have chosen to use the definition proposed by Weibull, other definitions may also be used. See [4] and [5] for more details.

 ISBN: 978-84-614-6167-7

4. Obtain the values $G^{-1}(r_i^*)$ for each $i = 1, ...k$.
5. Use the values obtained in the previous step to calculate the estimates of location-scale parameters of normal distribution, $\widehat{\mu}$ and $\widehat{\sigma}$.
6. Calculate values
$$y_i = \Phi\left(\frac{G^{-1}(r_i^*) - \widehat{\mu}}{\widehat{\sigma}}\right) \quad i = 1, ..., k$$
7. Apply arcsine transformation to stabilize the variance:
$$s_i^* = \left(\frac{2}{\pi}\right) \arcsin(\sqrt{y_i}) \quad i = 1, ..., k$$
8. Plot the pairs of points (r_i^*, s_i^*) $i = 1, ..., k$ on the Normal S-P Plot of step 1, joining them to get a smooth curve.

4 Numerical Example

An application of the procedure explained above is presented below by providing the following example.

We have a simple random sample from a distribution $F(x)$ and we want to know if the sample observations have a normal distribution.

Figure 1 shows a Normal S-P Plot together with the sample observations and four distribution curves for four different probability distributions: Exponential, Uniform, Cauchy and Gumbel distributions. In this example a value of k = 100000 has been used.

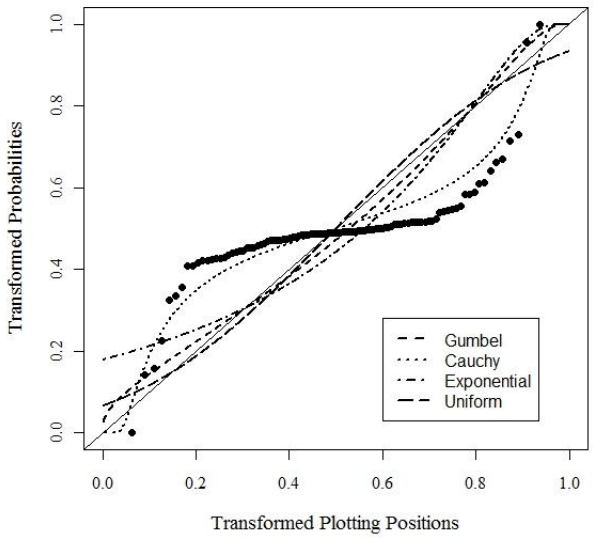


Figure 1: Normal S-P Plot and Distribution Curves

 ISBN: 978-84-614-6167-7

First, it can be clearly seen that the sample data do not have a rectilinear configuration in the S-P Plot, so these observations are not normally distributed.

By observing the different distribution curves, we conclude that the Cauchy distribution would be a good candidate as an alternative probability model for the sample data.

References

[1] John R. Michael, *The Stabilized Probability Plot*, Biometrika **Vol. 70, N 1** (1983) 11–17.

[2] W. Weibull, *The Phenomenon of Rupture in Solids*, Ingeniors Vetenskaps Akademien Handlingar **17** (1939).

[3] F. F. Gan, Kenneth J. Koehler and John C. Thompson, *Probability Plots and Distribution Curves for Assessing the Fit of Probability Models*, The American Statistician **Vol. 45, N 1** (1991) 14–21.

[4] Sonia Castillo-Gutiérrez, *Gráficos de Probabilidad: Una Herramienta para el Análisis y el Contraste de Normalidad*, Doctoral Thesis, Spain, 2011.

[5] Sonia Castillo-Gutiérrez, Emilio Lozano-Aguilera, *On Plotting Positions on Normal Q-Q Plots. R Script*, Proceedings of the International Conference on Computational and Mathematical Methods in Science and Engineering, CMMSE 2010 Spain, 2010. 1067–1070.

 ISBN: 978-84-614-6167-7

Proceedings of the 11th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2011
26–30 June 2011.

A first approach to an axiomatic model of multi-measures

Elena E. Castiñeira[1], Tomasa Calvo[2] and Susana Cubillo[1]

[1] *Department of Applied Mathematics, Technical University of Madrid (UPM)*

[2] *Department of Computer Sciences, University of Alcalá de Henares (UAH)*

emails: ecastineria@fi.upm.es, tomasa.calvo@uah.es, scubillo@fi.upm.es

Abstract

In this paper, we establish an axiomatic model of multi-measures, capturing some classes of measures studied in the fuzzy sets literature, where they are applied to only one or two arguments. Specifically, we look at multi-measures for determining the degrees of incompatibility and supplementarity between any number of fuzzy sets. Additionally, we introduce multi-measures for ranking their opposite properties, that is, compatibility and unsupplementarity.

Key words: Lattice, aggregation function, t-norm, t-conorm, incompatibility multi-measure, compatibility multi-measure, supplementarity multi-measure, unsupplementarity multi-measure.

MSC 2000: 03B52, 68T27, 68T30

1 Introduction

Traditionally, a measure has been defined as a real-valued set function on Boolean algebras or on σ-algebras of classical sets and, more recently, on fuzzy sets (like vagueness or ambiguity measures, etc.). However there are other measurable properties that it makes sense to apply on not only one but also two or more sets. Several works concerning how to measure the contradiction, the incompatibility or the supplementarity, among other properties, between two fuzzy sets have already been published in this respect. Nevertheless, it might be worth studying these properties for more than two fuzzy sets, since we are left wondering how incompatible or supplementary a set of n fuzzy premises is. This is why multi-argument functions should be addressed.

Of the multi-argument functions, aggregation functions [3, 6, 9] deserve a special mention. Their purpose is to combine several inputs into a single output. The nature of the

 ISBN: 978-84-614-6167-7

inputs depends on the context: they can be degrees of membership in fuzzy sets, degrees of preference, and so on. They play a significant role in many applications, which has driven their constant growth. They are useful in different areas like multi-criteria decision making, group decision making, fuzzy logic and rule-based systems, etc.

Other works have dealt with other kinds of multi-argument functions. In this respect, consider Martín and Mayor's recent papers on multi-distances [10] introducing a way to measure "how separated" the points of a collection of more than two elements are.

The main aim of this paper is to establish a general axiomatic model of multi-argument measures in order to capture some measures that we have examined previously and, in particular, incompatibility measures [7, 8] and supplementarity measures [8].

The remainder of the paper is organized as follows. Section 2 provides the axioms or conditions of multi-measures and describes some examples. Section 3 focuses on two multi-measures on the lattice of fuzzy sets with the order induced by the order of real numbers, whereas Section 4 studies two multi-measures on the lattice again composed of fuzzy sets but with the order induced by the reverse order of the real numbers. Finally, the paper ends with a summary of the results, and some future lines of research.

2 Multi-argument measures on lattices

Let $\mathcal{L} = (L, \leq_L, 0_L, 1_L)$ (or simply $(L, \leq_L)$) be a bounded lattice [4, 5] whose minimum and maximum elements are denoted by 0_L and 1_L, respectively. For each $n \in \mathbb{N}$, let us consider the set

$$L^n = \{(a_1, \ldots, a_n) \mid a_i \in L, \forall i \in \{1, \ldots, n\}\}$$

and the order relation $\leq_n$ induced by $\leq_L$, that is, given $\bar{a} = (a_1, \ldots, a_n), \bar{b} = (b_1, \ldots b_n) \in L^n$,

$$\bar{a} \leq_n \bar{b} \iff a_i \leq_L b_i, \forall i \in \{1, \ldots, n\}.$$

We have that L^n with the order relation $\leq_n$ is also a bounded lattice, whose minimum element is $0_{L^n} = (0_L, \overset{n)}{\ldots}, 0_L)$ and whose maximum element is $1_{L^n} = (1_L, \overset{n)}{\ldots}, 1_L)$, and we say that $\mathcal{L}^n = (L^n, \leq_n, 0_{L^n}, 1_{L^n})$ is induced by $\mathcal{L}$. Moreover, if $\mathcal{L}$ is complete, then $\mathcal{L}^n$ is also complete.

Definition 2.1. Let $\mathcal{L} = (L, \leq_L, 0_L, 1_L)$ be a bounded lattice. Also, for each $n \in \mathbb{N}$, let $\mathcal{L}^n = (L^n, \leq_n, 0_{L^n}, 1_{L^n})$ be the lattice induced by $\mathcal{L}$. Consider the bounded and complete lattice of real numbers $([0, 1], \leq, 0, 1)$. A map $M : \bigcup_{n \in \mathbb{N}} L^n \to [0, 1]$ is said to be a *multi-argument $\leq_L$-measure or multi-measure on* $(L, \leq_L)$ (or, simply, on L if there is not likely to be confusion) if, for each $n \in \mathbb{N}$, the function restriction of M to L^n, $M|_{L^n}$, satisfies:

i) (*Boundary conditions*) $M(0_{L^n}) = 0$ and $M(1_{L^n}) = 1$.

 ISBN: 978-84-614-6167-7

ii) (*Monotony condition*) For all $\bar{a}, \bar{b} \in L^n$ such that $\bar{a} \leq_n \bar{b}$, $M(\bar{a}) \leq M(\bar{b})$; that is, for each n, M is increasing with respect to the orders of the lattices $(L^n, \leq_n)$ and $([0,1], \leq)$.

Moreover,

iii) M is *increasing with respect to the argument* n or n*-increasing* if $M(a_1, \ldots, a_n) \leq M(a_1, \ldots, a_n, a_{n+1})$ holds for all $n \in \mathbb{N}$ and for all $a_1, \ldots, a_n, a_{n+1} \in L$.

iv) M is *decreasing with respect to the argument* n or n*-decreasing* if $M(a_1, \ldots, a_n) \geq M(a_1, \ldots, a_n, a_{n+1})$ holds for all $n \in \mathbb{N}$ and for all $a_1, \ldots, a_n, a_{n+1} \in L$.

Remark 2.2. If M is a multi-measure on $(L, \leq_L)$, note that:

1. M is n-increasing if and only if $M(\bar{a}) \leq M(\bar{a}, \bar{b})$ holds for all $n, m \in \mathbb{N}$ and for all $\bar{a} = (a_1, \ldots, a_n) \in L^n$ and $\bar{b} = (b_1, \ldots, b_m) \in L^m$, where $(\bar{a}, \bar{b})$ denotes $(a_1, \ldots, a_n, b_1, \ldots, b_m) \in L^{n+m}$.

2. M is n-decreasing if and only if $M(\bar{a}) \geq M(\bar{a}, \bar{b})$ holds for all $n, m \in \mathbb{N}$ and for all $\bar{a} \in L^n$ and $\bar{b} \in L^m$.

Example 2.3. Let X be a non-empty and finite set, and let $\mathcal{P}(X)$ denote the set of all subsets of X, that is, the power set of X. Consider the bounded lattice $(\mathcal{P}(X), \subseteq, \emptyset, X)$, which is, in fact, a Boolean algebra, and let us define two multi-argument measures on $(\mathcal{P}(X), \subseteq)$.

a) Let $M_I : \bigcup_{n \in \mathbb{N}} \mathcal{P}(X)^n \to [0,1]$ be the map defined for each $(A_1, \ldots, A_n) \in \mathcal{P}(X)^n$ as

$$M_I(A_1, \ldots, A_n) = \frac{|A_1 \cap \cdots \cap A_n|}{|X|},$$

where $|A|$ means cardinal of the set A. Then M_I satisfies:

i) $M_I(\emptyset, \ldots, \emptyset) = 0$ for all n-tuple of empty sets; and $M_I(X, \ldots, X) = 1$ for all n-tuple of elements X.

ii) $M_I(\bar{A}) \leq M_I(\bar{B})$ holds for all $n \in \mathbb{N}$ and for all $\bar{A} = (A_1, \ldots, A_n), \bar{B} = (B_1, \ldots, B_n) \in \mathcal{P}(X)^n$ such that $A_i \subseteq B_i$ for each $i \in \{1, \ldots, n\}$.

iv) $M_I(A_1, \ldots, A_n, A_{n+1}) \leq M_I(A_1, \ldots, A_n)$ for all $n \in \mathbb{N}$ and for all $A_1, \ldots, A_n, A_{n+1} \in \mathcal{P}(X)$.

 ISBN: 978-84-614-6167-7

Hence, M_I is an n-decreasing multi-argument $\subseteq$-measure on $\mathcal{P}(X)$; it provides a measure of the size of the intersection of any finite family of subsets of X.

b) Let $M_U : \bigcup_{n\in\mathbb{N}} \mathcal{P}(X)^n \to [0,1]$ be the map defined for each $(A_1,\ldots,A_n) \in \mathcal{P}(X)^n$ as

$$M_U(A_1,\ldots,A_n) = \frac{|A_1 \cup \cdots \cup A_n|}{|X|}.$$

Then, M_U also satisfies axioms i and ii of the multi-argument measure and, moreover, axiom iii, therefore M_U is an n-increasing multi-argument $\subseteq$-measure on $\mathcal{P}(X)$; it provides a measure of the size of the union of any finite family of subsets of X. $\lhd$

Example 2.4. Any aggregation function is a multi-argument measure on $([0,1],\leq)$. Indeed, recall that an aggregation function [3, 6, 9] is a map $\mathcal{A} : \bigcup_{n\in\mathbb{N}} [0,1]^n \to [0,1]$ such that

1. $\mathcal{A}(0,\ldots,0) = 0$ and $\mathcal{A}(1,\ldots,1) = 1$.
2. $\mathcal{A}(a) = a$ for all $a \in [0,1]$.
3. For all $n \in \mathbb{N}$ and for all $\bar{a} = (a_1,\ldots,a_n), \bar{b} = (b_1,\ldots,b_n) \in [0,1]^n$ such that $a_i \leq b_i$ with $i = 1,\ldots,n$, $\mathcal{A}(\bar{a}) \leq \mathcal{A}(\bar{b})$ holds. $\lhd$

Thus, the occurrence of symmetric aggregation functions suggests the following definition. We denote $S_n = \{\pi : \{1,\ldots,n\} \to \{1,\ldots,n\} \,|\, \pi \text{ is a bijection}\}$, that is, S_n is the set of *permutations* of $\{1,\ldots,n\}$.

Definition 2.5. A multi-argument measure M on a bounded lattice $(L,\leq_L)$ is *symmetric* if, for each $n \in \mathbb{N}$, the function $M|_{L^n}$ is symmetric, that is, $M(a_1,\ldots,a_n) = M(a_{\pi(1)},\ldots,a_{\pi(n)})$ holds for any $\pi \in S_n$ and for any $(a_1,\ldots,a_n) \in L^n$.

Example 2.6. The maps M_I and M_U defined in Example 2.3 are both symmetric multi-measures on $\mathcal{P}(X)$. $\lhd$

Example 2.7. The aggregation functions $\mathrm{Max}, \mathrm{Min} : \bigcup_{n\in\mathbb{N}} [0,1]^n \to [0,1]$, defined as $\mathrm{Max}(a_1,\ldots,a_n) = \max\{a_1,\ldots,a_n\}$ and $\mathrm{Min}(a_1,\ldots,a_n) = \min\{a_1,\ldots,a_n\}$ for each $(a_1,\ldots,a_n) \in [0,1]^n$, are symmetric multi-measures on $([0,1],\leq)$.

Nevertheless, for each $k \in \mathbb{N}\setminus\{1\}$, the function $\mathcal{A}_k : \bigcup_{n\in\mathbb{N}} [0,1]^n \to [0,1]$, defined as

$$\mathcal{A}(a_1,\ldots,a_n) = a_1 \prod_{i=2}^{n} a_i^k$$

for each $(a_1,\ldots,a_n) \in [0,1]^n$, is a non-symmetric multi-measure on $([0,1],\leq)$

 ISBN: 978-84-614-6167-7

In what follows, we study the particular instance of multi-measures on fuzzy set lattices: if X is a non-empty set, the set of all fuzzy sets on X is identified with the set of all its membership functions, $L = [0,1]^X$, with an order relation $\preceq$ such that $\mathcal{L} = ([0,1]^X, \preceq, \mu_\wedge, \mu_\vee)$ is a bounded lattice, where $\mu_\wedge$ and $\mu_\vee$ denote the minimum and maximum elements, respectively. For each $n \in \mathbb{N}$, if $\mathbb{I}_n$ denotes $[0,1]^n$, lattice $\mathcal{L}$ induces the bounded lattice $\mathcal{L}^n = (\mathbb{I}_n^X, \preceq_n, \bar{\mu}_\wedge, \bar{\mu}_\vee)$ as $\mathbb{I}_n^X = ([0,1]^n)^X = [0,1]^X \times \overset{n)}{\ldots} \times [0,1]^X$. Thus, a multi-measure on $([0,1]^X, \preceq)$ is a multi-argument function $M : \bigcup_{n\in\mathbb{N}} \mathbb{I}_n^X \to [0,1]$ where i) $M(\bar{\mu}_\wedge) = 0$ and $M(\bar{\mu}_\vee) = 1$; and ii) $M(\bar{\mu}) \leq M(\bar{\sigma})$ holds for all $\bar{\mu}, \bar{\sigma} \in \mathbb{I}_n^X$ such that $\bar{\mu} \preceq_n \bar{\sigma}$.

3 Multi-measures on lattice $([0,1]^X, \leq)$

In this section, we deal with multi-measures on $([0,1]^X, \preceq)$ when $\preceq$ is the order induced by the usual order on the real line, that is, if $\mu, \sigma \in [0,1]^X$, $\mu \preceq \sigma$ if and only if $\mu(x) \leq \sigma(x)$ for all $x \in X$; and we naturally denote this by $\mu \leq \sigma$. In this case, $\mu_\wedge = \mu_\emptyset$ and $\mu_\vee = \mu_X$, and thus $\mathcal{L} = ([0,1]^X, \leq, \mu_\emptyset, \mu_X)$.

Let us look at two types of multi-measures on $([0,1]^X, \leq)$: multi-measures that evaluate how compatible a set of fuzzy sets is and multi-measures that evaluate how supplementary the set is. Remember that, in classical logic, two statements are compatible if they can both be true at the same time. As we can identify a statement on a universe X with the set of elements of X that satisfy that statement, we can translate this concept to set theory: $A, B \subset X$ are compatible if $A \cap B \neq \emptyset$. On the other hand, supplementarity can, in a sense, be understood as a symmetric property of incompatibility: A an B are supplementary if $A \cup B = X$. These concepts are extended to the fuzzy set framework and studied in the following sections.

3.1 Compatibility multi-measures on fuzzy sets

In order to define compatible fuzzy sets, we need a function that models the intersection of fuzzy sets, that is, a t-norm. Remember that a *t-norm* [1, 2, 11] is a binary aggregation function T on the unit interval $[0,1]$, which is commutative, associative, monotone increasing with respect to the usual order on the real line, and whose neutral element is 1. As in the classical case, given a t-norm T, two fuzzy sets on X, or their membership functions $\mu, \sigma \in [0,1]^X$, are T-compatible if $T(\mu, \sigma) \neq \mu_\emptyset$, where $T(\mu, \sigma) \in [0,1]^X$ is defined by $T(\mu, \sigma)(x) = T(\mu(x), \sigma(x))$ for each $x \in X$. This can be generalized similarly as follows.

Definition 3.1. Given $X \neq \emptyset$ and a t-norm T, then

1. $\{\mu\} \subset [0,1]^X$ is said to be T-*compatible* if $\mu \neq \mu_\emptyset$.
2. If $n > 1$, $\{\mu_1, \ldots, \mu_n\} \subset [0,1]^X$ is said to be T-*compatible* if $T(\mu_1, \ldots, \mu_n) \neq \mu_\emptyset$.

 ISBN: 978-84-614-6167-7

The following definition determines the conditions that a multi-argument function must satisfy to fittingly assign a degree of compatibility to every $\{\mu_1, \ldots, \mu_n\} \subset [0,1]^X$.

Definition 3.2. Let T be a t-norm and $X \neq \emptyset$. A function $\mathcal{C}_T : \bigcup_{n\in\mathbb{N}} \mathbb{I}_n^X \to [0,1]$ is a *T-compatibility multi-measure* on $[0,1]^X$ if it is a symmetric and n-decreasing multi-measure on $([0,1]^X, \leq)$ satisfying $\mathcal{C}_T(\mu_1, \ldots, \mu_n) = 0$, provided that $\{\mu_1, \ldots, \mu_n\} \subset [0,1]^X$ satisfies $T(\mu_1, \ldots, \mu_n) = \mu_\emptyset$.

Remark 3.3. Note that $\mathcal{C}_T$ is a T-compatibility multi-measure on $[0,1]^X$ if and only if it satisfies the following axioms:

ci) $\mathcal{C}_T(\mu_X, \overset{n)}{\ldots}, \mu_X) = 1$ for each $n \in \mathbb{N}$.

c.ii) If $\{\mu_1, \ldots, \mu_n\} \subset [0,1]^X$ is not T-compatible, then $\mathcal{C}_T(\mu_1, \ldots, \mu_n) = 0$.

c.iii) $\mathcal{C}_T(\mu_1, \ldots, \mu_n) = \mathcal{C}_T(\mu_{\pi(1)}, \ldots, \mu_{\pi(n)})$ holds for all $\pi \in S_n$ and $\mu_1, \ldots, \mu_n \in [0,1]^X$.

c.iv) If $\mu_1, \ldots, \mu_n, \sigma_1, \ldots, \sigma_n \in [0,1]^X$ satisfy $\mu_i \leq \sigma_i$ for all $i \in \{1, \ldots, n\}$, then $\mathcal{C}_T(\mu_1, \ldots, \mu_n) \leq \mathcal{C}_T(\sigma_1, \ldots, \sigma_n)$.

c.v) $\mathcal{C}_T(\mu_1, \ldots, \mu_{n+1}) \leq \mathcal{C}_T(\mu_1, \ldots, \mu_n)$ holds for all $n \in \mathbb{N}$ and $\mu_1, \ldots, \mu_{n+1} \in [0,1]^X$.

As is well known, if $\varphi \in \mathcal{A}([0,1]) = \{\psi : [0,1] \to [0,1] \,|\, \psi \text{ is an increasing bijection}\}$ and T is a t-norm, then $T^\varphi : [0,1]^2 \to [0,1]$ defined for each $(a,b) \in [0,1]^2$ by $T^\varphi(a,b) = \varphi^{-1}(T(\varphi(a), \varphi(b)))$ is also a t-norm, and we say that it is the t-norm φ-conjugated with T. One of the main t-norms is the so-called Lukasiewicz t-norm that is defined for each $(a,b) \in [0,1]^2$ by $T_L(a,b) = \max\{0, a+b-1\}$. Moreover, if T_L^φ is the t-norm φ-conjugated with T_L, then $T_L^\varphi(a_1, \ldots, a_n) = \varphi^{-1}(\max\{0, \varphi(a_1) + \cdots + \varphi(a_n) - (n-1)\})$ for all $a_1, \ldots, a_n \in [0,1]$. For more details about t-norms see [1, 2].

Given $\{\mu_1, \ldots, \mu_n\} \subset [0,1]^X$, since $T_L^\varphi(\mu_1, \ldots, \mu_n) \neq \mu_\emptyset$ if and only if there exists $x \in X$ such that $\sum_{i=1}^n \varphi(\mu_i(x)) > n-1$, then a natural way to measure the T_L^φ-compatibility of $\{\mu_1, \ldots, \mu_n\}$ could be to conveniently take into account the difference between $\sum_{i=1}^n \varphi(\mu_i(x))$ and $n-1$, as follows.

Proposition 3.4. *Let $X \neq \emptyset$ and $\varphi \in \mathcal{A}([0,1])$. The function $\mathcal{C}_L^\varphi : \bigcup_{n\in\mathbb{N}} \mathbb{I}_n^X \to [0,1]$ defined for each $\bar{\mu} = (\mu_1, \ldots, \mu_n) \in \mathbb{I}_n^X$, by*

$$\mathcal{C}_L^\varphi(\bar{\mu}) = \max\left\{0, \sup_{x\in X} \sum_{i=1}^n \varphi(\mu_i(x)) - (n-1)\right\}$$

is a T_L^φ-compatibility multi-measure on $[0,1]^X$

 ISBN: 978-84-614-6167-7

Proof. First, note that $\mathcal{C}_L^{\varphi}$ is well defined; indeed, since $\sum_{i=1}^{n}\varphi(\mu_i(x)) \leq n$ then $\sum_{i=1}^{n}\varphi(\mu_i(x))-(n-1) \leq 1$.

Axioms c.i and c.iii follow straightforwardly from the definition of $\mathcal{C}_L^{\varphi}$. Regarding axiom c.ii, if $\mu = \mu_{\emptyset}$, then $\mathcal{C}_L^{\varphi}(\mu) = 0$; if $\bar{\mu} \in \mathbb{I}_n^X$ satisfies $T_L^{\varphi}(\mu_1,\ldots,\mu_n) = \mu_{\emptyset}$, then $\sum_{i=1}^{n}\varphi(\mu_i(x)) \leq n-1$ for all $x \in X$ and thus $\mathcal{C}_L^{\varphi}(\bar{\mu}) = 0$. $\mathcal{C}_L^{\varphi}$ is monotonic increasing: given $\bar{\mu} = (\mu_1,\ldots,\mu_n), \bar{\sigma} = (\sigma_1,\ldots,\sigma_n) \in \mathbb{I}_n^X$ such that $\mu_i \leq \sigma_i$ for every $i \in \{1,\ldots,n\}$, as φ is increasing, then $\sum\limits_{i=1}^{n}\varphi(\mu_i(x)) \leq \sum\limits_{i=1}^{n}\varphi(\sigma_i(x))$ and thus $\mathcal{C}_L^{\varphi}(\bar{\mu}) \leq \mathcal{C}_L^{\varphi}(\bar{\sigma})$.

Finally, $\mathcal{C}_L^{\varphi}$ is n-decreasing: given $n \in \mathbb{N}$ and $\mu_1, \ldots, \mu_n, \mu_{n+1} \in [0,1]^X$, it follows that $\mathcal{C}_L^{\varphi}(\mu_1,\ldots,\mu_{n+1}) \leq \mathcal{C}_L^{\varphi}(\mu_1,\ldots,\mu_n)$ as $\sum\limits_{i=1}^{n+1}\varphi(\mu_i(x)) \leq \sum\limits_{i=1}^{n}\varphi(\mu_i(x)) + 1$. □

3.2 Supplementary multi-measures on fuzzy sets

As applies in the case of compatible fuzzy sets, we need a tool to model the union of fuzzy sets in order to define supplementary fuzzy sets, and t-conorms are suitable functions for this purpose. Remember that a *t-conorm* [1, 2] is a binary aggregation function S on the unit interval $[0,1]$, which is commutative, associative, monotone increasing with respect to the usual order on the real line, whose neutral element is 0. Given a t-conorm S, two fuzzy sets on X or their membership functions $\mu, \sigma \in [0,1]^X$ are S-supplementary [8] if $S(\mu,\sigma) = \mu_X$, where $S(\mu,\sigma) \in [0,1]^X$ is defined by $S(\mu,\sigma)(x) = S(\mu(x),\sigma(x))$ for each $x \in X$. This can be generalized similarly as follows.

Definition 3.5. Given $X \neq \emptyset$ and a t-conorm S, then

1. $\{\mu\} \subset [0,1]^X$ is said to be *S-supplementary* if $\mu = \mu_X$.
2. If $n>1$, $\{\mu_1,\ldots,\mu_n\} \subset [0,1]^X$ is said to be *S-supplementary* if $S(\mu_1,\ldots,\mu_n) = \mu_X$.

Definition 3.6. Let S be a t-conorm and $X \neq \emptyset$. A function $\mathcal{S}_S : \bigcup_{n\in\mathbb{N}}\mathbb{I}_n^X \to [0,1]$ is an *S-supplementarity multi-measure* on $[0,1]^X$ if it is a symmetric and n-increasing multi-measure on $([0,1]^X,\leq)$ satisfying $\mathcal{S}_S(\mu_1,\ldots,\mu_n) = 0$ provided that $\{\mu_1,\ldots,\mu_n\} \subset [0,1]^X$ is not S-supplementary.

Remark 3.7. Note that $\mathcal{S}_S$ is an S-supplementarity multi-measure on $[0,1]^X$ if and only if:

- s.i $\mathcal{S}_S(\mu_X,\overset{n)}{\ldots},\mu_X) = 1$ for each $n \in \mathbb{N}$.
- s.ii If $\{\mu_1,\ldots,\mu_n\} \subset [0,1]^X$ is not S-supplementary, then $\mathcal{S}_S(\mu_1,\ldots,\mu_n) = 0$.
- s.iii $\mathcal{S}_S(\mu_1,\ldots,\mu_n) = \mathcal{S}_S(\mu_{\pi(1)},\ldots,a_{\pi(n)})$ holds for all $\pi \in S_n$ and $\mu_1,\ldots,\mu_n \in [0,1]^X$.
- s.iv If $\mu_1,\ldots,\mu_n,\sigma_1,\ldots,\sigma_n \in [0,1]^X$ satisfy $\mu_i \leq \sigma_i$ for all $i \in \{1,\ldots,n\}$, then $\mathcal{S}_S(\mu_1,\ldots,\mu_n) \leq \mathcal{S}_S(\sigma_1,\ldots,\sigma_n)$.

 ISBN: 978-84-614-6167-7

s.v $\mathcal{S}_S(\mu_1,\ldots,\mu_n) \leq \mathcal{S}_S(\mu_1,\ldots,\mu_{n+1})$ holds for all $n \in \mathbb{N}$ and $\mu_1,\ldots,\mu_{n+1} \in [0,1]^X$.

As in the t-norm case, if $\varphi \in \mathcal{A}([0,1])$ and S is a t-conorm, the function defined for each $(a,b) \in [0,1]^2$ by $S^\varphi(a,b) = \varphi^{-1}(S(\varphi(a),\varphi(b))$ is also a t-conorm, the t-conorm φ-conjugated with S. The Lukasiewicz t-norm has a dual t-conorm defined for each $(a,b) \in [0,1]$ by $S_L(a,b) = \min\{1, a+b\}$. Moreover, if S_L^φ is the t-conorm φ-conjugated with S_L, then $S_L^\varphi(a_1,\ldots,a_n) = \varphi^{-1}(\min\{1,\varphi(a_1)+\cdots+\varphi(a_n)\})$ for all $a_1,\ldots,a_n \in [0,1]$.

To find a way to measure the S_L^φ-supplementarity of $\{\mu_1,\ldots,\mu_n\} \subset [0,1]^X$, note that $S_L^\varphi(\mu_1,\ldots,\mu_n) = \mu_X$ if and only if $\sum_{i=1}^n \varphi(\mu_i(x)) \geq 1$ for all $x \in X$; hence we can fittingly use the difference between 1 and $\sum_{i=1}^n \varphi(\mu_i(x))$. So, we can prove the following result.

Proposition 3.8. *Let $X \neq \emptyset$ and $\varphi \in \mathcal{A}([0,1])$. Let $\mathcal{S}_L^\varphi : \bigcup_{n\in\mathbb{N}} \mathbb{I}_n^X \to [0,1]$ be the function defined:*

1. *For each $\mu \in [0,1]^X$, by $\mathcal{S}_L^\varphi(\mu) = \begin{cases} 1 & \text{if } \mu = \mu_X \\ 0 & \text{if } \mu \neq \mu_X, \end{cases}$*
2. *For each $\bar{\mu} = (\mu_1,\ldots,\mu_n) \in \mathbb{I}_n^X$ with $n > 1$, by*

$$\mathcal{S}_L^\varphi(\bar{\mu}) = \min\left\{1, \max\left\{0, \inf_{x\in X} \sum_{i=1}^n \varphi(\mu_i(x)) - 1\right\}\right\}.$$

Then $\mathcal{S}_L^\varphi$ is an S_L^φ-supplementarity multi-measure on $[0,1]^X$.

Remark 3.9. We have that $\mathcal{S}_L^\varphi(\mu_1,\mu_2) = \max\{0, \inf_{x\in X}(\varphi(\mu_1(x)) + \varphi(\mu_2(x))) - 1\}$ for each $(\mu_1,\mu_2) \in \mathbb{I}_2^X$, thus the restriction of $\mathcal{S}_L^\varphi$ to $\mathbb{I}_2^X = [0,1]^X \times [0,1]^X$ is an S_L^φ-supplementarity measure regarding the definition reported in [8].

4 Multi-measures on lattice $([0,1]^X, \geq)$

In this section, we deal with multi-measures on $([0,1]^X, \preceq)$ when $\preceq$ is the order induced by the usual reverse order of the real line, that is, if $\mu,\sigma \in [0,1]^X$, $\mu \preceq \sigma$ if and only if $\mu(x) \geq \sigma(x)$ for all $x \in X$; and we naturally denote this by $\mu \geq \sigma$. In this case, $\mu_\wedge = \mu_X$ and $\mu_\vee = \mu_\emptyset$, where, for all $x \in X$, $\mu_X(x) = 1$ and $\mu_\emptyset(x) = 0$, then the lattice $\mathcal{L}$ is $([0,1]^X, \geq, \mu_X, \mu_\emptyset)$.

Let us look at two types of multi-measures on $([0,1]^X, \geq)$: multi-measures that evaluate how incompatible a set of fuzzy sets is and multi-measures that evaluate how unsupplementary the set is, where the concepts of incompatibility and unsupplementarity are opposite to compatibility and supplementarity, respectively. That is, given a t-norm T and a t-conorm S, $\{\mu_1,\ldots,\mu_n\} \subset \mathbb{I}_n^X$ is *T-incompatible* if it is not T-compatible, and it is *S-unsupplementary* if it is not S-supplementary.

 ISBN: 978-84-614-6167-7

4.1 Incompatibility multi-measures on fuzzy sets

Although the concepts of compatibility and incompatibility are opposites, the negation of a compatibility measure cannot be used to assign degrees of incompatibility. Indeed, let $\mathcal{C}_T$ be a non-trivial T-compatibility multi-measure, that is, at least it takes a value $a \in (0,1)$, and let N be any strong negation [12] (i.e., $N : [0,1] \to [0,1]$ is an involutive and decreasing bijection); if a is achieved on $\bar{\mu} \in \mathbb{I}_n^X$, then $0 < \mathcal{C}_T(\bar{\mu}) = a < 1$, and it follows from axiom ii of Remark 3.3 that $\bar{\mu}$ is T-compatible, and also $0 = N(1) < N(\mathcal{C}_T(\bar{\mu})) < N(0) = 1$ holds. Thus $N(\mathcal{C}_T(\bar{\mu}))$ cannot be considered as a degree of the T-incompatibility of $\bar{\mu}$ since the incompatibility measure of compatible sets should be 0. Therefore, it makes sense to propose a mathematical model for the study of the incompatibility.

Definition 4.1. Let T be a t-norm and $X \neq \emptyset$. A function $\mathcal{I}_T : \bigcup_{n\in\mathbb{N}} \mathbb{I}_n^X \to [0,1]$ is a *T-incompatibility multi-measure* on $[0,1]^X$ if it is a symmetric and n-increasing multi-measure on $([0,1]^X, \geq)$ satisfying $\mathcal{I}_T(\mu_1, \ldots, \mu_n) = 0$, provided that $\{\mu_1, \ldots, \mu_n\} \subset [0,1]^X$ is T-compatible.

Remark 4.2. Note that $\mathcal{I}_T$ is a T-incompatibility multi-measure on fuzzy sets on X if and only if:

ic.i $\mathcal{I}_T(\mu_\emptyset, \overset{n)}{\ldots}, \mu_\emptyset) = 1$ for each $n \in \mathbb{N}$.

ic.ii If $\{\mu_1, \ldots, \mu_n\} \subset [0,1]^X$ is T-compatible, then $\mathcal{I}_T(\mu_1, \ldots, \mu_n) = 0$.

ic.iii $\mathcal{I}_T(\mu_1, \ldots, \mu_n) = \mathcal{I}_T(\mu_{\pi(1)}, \ldots, \mu_{\pi(n)})$ holds for all $\pi \in S_n$ and $\mu_1, \ldots, \mu_n \in [0,1]^X$.

ic.iv If $\mu_1, \ldots, \mu_n, \sigma_1, \ldots, \sigma_n \in [0,1]^X$ satisfy $\mu_i \leq \sigma_i$ for all $i \in \{1, \ldots, n\}$, then $\mathcal{I}_T(\sigma_1, \ldots, \sigma_n) \leq \mathcal{I}_T(\mu_1, \ldots, \mu_n)$.

ic.v $\mathcal{I}_T(\mu_1, \ldots, \mu_n) \leq \mathcal{I}_T(\mu_1, \ldots, \mu_{n+1})$ holds for all $n \in \mathbb{N}$ and $\mu_1, \ldots, \mu_{n+1} \in [0,1]^X$.

As in the case of compatibility, if T_L^φ is the t-norm φ-conjugated with the Lukasiewicz t-norm, taking into account that $T_L^\varphi(\mu_1, \ldots, \mu_n) = \mu_\emptyset$ if and only if $\sum_{i=1}^n \varphi(\mu_i(x)) \leq n-1$, we can find a T_L^φ-incompatibility multi-measure by fittingly considering the difference between $n-1$ and $\sum_{i=1}^n \varphi(\mu_i(x))$, and thus we can prove the following result.

Proposition 4.3. *Let $X \neq \emptyset$ and $\varphi \in \mathcal{A}([0,1])$. Let $\mathcal{I}_L^\varphi : \bigcup_{n\in\mathbb{N}} \mathbb{I}_n^X \to [0,1]$ be the function defined:*

1. *For each $\mu \in [0,1]^X$, by $\mathcal{I}_L^\varphi(\mu) = \begin{cases} 1 & \text{if } \mu = \mu_\emptyset \\ 0 & \text{if } \mu \neq \mu_\emptyset, \end{cases}$*
2. *For each $\bar{\mu} = (\mu_1, \ldots, \mu_n) \in \mathbb{I}_n^X$ with $n > 1$, by*

$$\mathcal{I}_L^\varphi(\bar{\mu}) = \min\left\{1, \max\left\{0, (n-1) - \sup_{x\in X} \sum_{i=1}^n \varphi(\mu_i(x))\right\}\right\}.$$

 ISBN: 978-84-614-6167-7

Then $\mathcal{I}_L^{\varphi}$ is a T_L^{φ}-incompatibility multi-measure on $[0,1]^X$.

Remark 4.4. We have that $\mathcal{I}_L^{\varphi}(\mu,\mu_2) = \max\{0, 1 - \sup_{x\in X}(\varphi(\mu_1(x)) + \varphi(\mu_2(x)))\}$ for each $(\mu_1,\mu_2) \in \mathbb{I}_2^X$, thus $\mathcal{I}_L^{\varphi}|_{[0,1]^X\times[0,1]^X}$ is a T_L^{φ}-incompatibility measure regarding the definition reported in [7].

4.2 Unsupplementarity multi-measures on fuzzy sets

As for incompatibility, although unsupplementary is the opposite to supplementary, it is not possible to assign degrees of unsupplementarity by means of a negation of a supplementarity multi-measure. Hence we establish a mathematical model to measure the unsupplementarity property.

Definition 4.5. Let S be a t-conorm and $X \neq \emptyset$. A function $\mathcal{U}_T : \bigcup_{n\in\mathbb{N}} \mathbb{I}_n^X \to [0,1]$ is an *S-unsupplementarity multi-measure* on $[0,1]^X$ if it is a symmetric and n-decreasing multi-measure on $([0,1]^X, \geq)$ satisfying $\mathcal{U}_S(\mu_1,\ldots,\mu_n) = 0$, provided that $\{\mu_1,\ldots,\mu_n\} \subset [0,1]^X$ is S-supplementary.

Remark 4.6. Note that $\mathcal{U}_S$ is an S-supplementarity multi-measure on $[0,1]^X$ if and only if:

us.i $\mathcal{U}_S(\mu_\emptyset, \overset{n)}{\ldots}, \mu_\emptyset) = 1$ for each $n \in \mathbb{N}$.

us.ii If $\{\mu_1,\ldots,\mu_n\} \subset [0,1]^X$ is S-supplementary, then $\mathcal{U}_S(\mu_1,\ldots,\mu_n) = 0$.

us.iii $\mathcal{U}_S(\mu_1,\ldots,\mu_n) = \mathcal{U}_S(\mu_{\pi(1)},\ldots,\mu_{\pi(n)})$ holds for all $\pi \in S_n$ and $\mu_1,\ldots,\mu_n \in [0,1]^X$.

us.iv If $\mu_1,\ldots,\mu_n,\sigma_1,\ldots,\sigma_n \in [0,1]^X$ satisfy $\mu_i \leq \sigma_i$ for all $i \in \{1,\ldots,n\}$, then $\mathcal{U}_S(\sigma_1,\ldots,\sigma_n) \leq \mathcal{U}_S(\mu_1,\ldots,\mu_n)$.

us.v $\mathcal{U}_S(\mu_1,\ldots,\mu_{n+1}) \leq \mathcal{U}_S(\mu_1,\ldots,\mu_n)$ holds for all $n \in \mathbb{N}$ and $\mu_1,\ldots,\mu_{n+1} \in [0,1]^X$.

As in the case of supplementarity, if S_L^{φ} is the t-conorm φ-conjugated with the Lukasiewicz t-conorm, taking into account that $S_L^{\varphi}(\mu_1,\ldots,\mu_n) = \mu_X$ if and only if $\sum_{i=1}^n \varphi(\mu_i(x)) \geq 1$ for all $x \in X$, we can use the difference between $\sum_{i=1}^n \varphi(\mu_i(x))$ and 1 to assign degrees of S_T^{φ}-unsupplementarity. So, we can prove the following result.

Proposition 4.7. *Let $X \neq \emptyset$ and $\varphi \in \mathcal{A}([0,1])$. The function $\mathcal{U}_L^{\varphi} : \bigcup_{n\in\mathbb{N}} \mathbb{I}_n^X \to [0,1]$ defined for each $\bar{\mu} = (\mu_1,\ldots,\mu_n) \in \mathbb{I}_n^X$, by*

$$\mathcal{U}_L^{\varphi}(\bar{\mu}) = \max\left\{0, 1 - \inf_{x\in X} \sum_{i=1}^n \varphi(\mu_i(x))\right\}$$

is an S_L^{φ}-unsupplementarity multi-measure on $[0,1]^X$

 ISBN: 978-84-614-6167-7

Conclusions

In this paper, we first introduced an axiomatic model of multi-measures on the bounded lattice structure, illustrating some examples in lattices of classical sets. Then, we studied two types of multi-argument measures in the particular case of the lattice of fuzzy sets $([0,1]^X, \leq)$, multi-measures of T-compatibility and S-supplementarity. Finally, we tackled two multi-measures on $([0,1]^X, \geq)$, capable of ranking their opposite properties, that is, compatibility and unsupplementarity. They all assume the classical case. We identified four types of multi-measures, taking any t-norm T or t-conorm S conjugated with the respective Lukasiewicz t-norms or t-conorms.

In the future, we intend to study the possible relations between given pairs of multi-measures, as well as searching other measures referred to t-norms and t-conorms that are not examined in this article.

Acknowledgements

This work has been partially supported by DGI (Spain) under Projects TIN2008-06890-C02-01, TIN2009-07901, MTM 2009-10962 and by UPM-CAM

References

[1] C. Alsina, M. Frank and B. Schweizer, *Associative Functions: Triangular Norms and Copulas*, World Scientific, Singapore, 2006.

[2] E. P. Klement, R. Mesiar and E. Pap, *Triangular Norms*, Kluwer Academic Publisher, Dordrecht, 2000.

[3] G. Beliakov, A. Pradera and T. Calvo, *Aggregation Functions: A Guide for Practitioners*, Springer-Verlag, Berlin, 2007.

[4] G. Birkhoff, *Lattice Theory*, American Mathematical Society, Providence, 1940.

[5] T.S. Blyth, *Lattices and Ordered Algebraic Structures*, Springer-Verlag, London, 2005.

[6] T. Calvo, A. Kolesarová, M. Komorníková and R. Mesiar, *Aggregation operators: properties, Classes and Construction Methods* in *Aggregation operators: New trends and applications.* Eds. T. Calvo, R. Mesiar and G. Mayor, Physica-Verlag, Heilderberg, (2002) 1–104.

 ISBN: 978-84-614-6167-7

[7] E. Castiñeira, S. Cubillo and P. L. Fernández, *Medidas de incompati-bilidad entre conjuntos borrosos*, Proc. XIII Congreso Español sobre Tecnologías y Lógica Fuzzy (in Spanish), Ciudad Real, Spain, (2006) 65–70.

[8] S. Cubillo, E. Castiñeira and W. Montilla, *Supplementarity measures on fuzzy sets*, Proc. of 7th Conference of the European Society for Fuzzy Logic and Technology (EUSFLAT-2011), Aix-Les-Bains (France), accepted.

[9] M. Grabisch, J. L. Marichal R. Mesiar and E. Pap, *Aggregation Functions*, Cambridge University Press, Cambridge, 2009.

[10] J. Martín and G. Mayor, *Multi-argument distances*, Fuzzy Sets and Systems **167** (2010) 92–100.

[11] B. Schweizer and A. Sklar, *Associative functions and statistical triangle inequalities*, Public. Math. Debrecen **8** (1961) 169–186.

[12] E. Trillas, *Sobre funciones de negación en los conjuntos difusos (in Spanish)*, Stochastica **3** (1) (1979) 47-60. Reprinted (English version) in *Advances of Fuzzy Logic*. Eds. S. Barro et alri (Universidad de Santiago de Compostela), (1998) 31–43.

 ISBN: 978-84-614-6167-7

Proceedings of the 11th International Conference on Computational and Mathematical Methods in Science and Engineering, CMMSE 2011 26–30 June 2011.

Minimal Faithful Unitriangular Matrix Representation of Filiform Lie Algebras

M. Ceballos[1], J. Núñez[1] and Á.F. Tenorio[2]

[1] *Dpto. de Geometría y Topología, Facultad de Matemáticas. Universidad de Sevilla.*

[2] *Dpto. de Economía, Métodos Cuantitativos e Historia Económica, Escuela Politécnica Superior. Universidad Pablo de Olavide.*

emails: mceballos@us.es, jnvaldes@us.es, aftenorio@upo.es

Abstract

Minimal faithful unitriangular matrix representations of filiform Lie algebras are computed in this paper. These representations are obtained by using nilpotent Lie algebras $\mathfrak{g}_n$, of $n \times n$ strictly upper-triangular matrices. This family of algebras allows to represent a given (filiform) nilpotent Lie algebra as a subalgebra of $\mathfrak{g}_n$ for some $n \in \mathbb{N} \setminus \{1\}$. In this sense, for a given filiform Lie algebra, we search the lowest natural integer n such that the Lie algebra $\mathfrak{g}_n$ contains this algebra as a subalgebra. In addition, we compute a representative of these representations too. It is convenient to note that all the computations in this paper have been carried out by using MAPLE 12.

Key words: Filiform Lie Algebra, Minimal Faithful Unitriangular Matrix Representation, Algorithm.

MSC 2000: 17B30, 17B15, 17B45, 68W30, 68W05.

1 Introduction

Filiform Lie algebras constitute a very special subclass of nilpotent Lie algebras. In fact, they are the most structured Lie algebras in the nilpotent class and were introduced by Vergne [10] in 1966. A well-known result about nilpotent Lie algebras states: Given a finite-dimensional nilpotent Lie algebra $\mathfrak{g}$, there exists $n \in \mathbb{N}$ such that $\mathfrak{g}$ is isomorphic to a subalgebra of the algebra $\mathfrak{g}_n$, of $n \times n$ strictly upper-triangular matrices [9, Theorem 3.6.6]. Therefore, a very interesting question for a given finite-dimensional filiform Lie algebra is the following: Determine the minimal $n \in \mathbb{N}$ such that $\mathfrak{g}_n$ contains this

 ISBN: 978-84-614-6167-7

Lie algebra $\mathfrak{g}$ as a subalgebra (i.e. obtain the minimal faithful unitriangular matrix representation of the algebra).

At this respect, Benjumea et al. [1] already obtained an algorithmic method to compute explicitly minimal unitriangular matrix representations of filiform Lie algebras and their associated Lie groups. However, they only computed representations explicitly for giving examples of application of the method in dimension less than 6. Later, Benjumea et al. [2] gave the list of minimal faithful unitriangular matrix representations for nilpotent Lie algebras of dimension less than 6. Nevertheless, they left open the following question: What Lie algebras have an n-dimensional representation for arbitrary dimensions? In this paper, we will study which filiform Lie algebras satisfy that property, giving minimal faithful unitriangular representations in the case of both model algebras and non-model ones.

Other authors, like Burde [4] or Ghanam et al. [8], studied the minimal dimension $\mu(\mathfrak{g})$ for the representations of a given Lie algebra $\mathfrak{g}$. However, these authors considered any faithful $\mathfrak{g}$-module instead of the family of Lie algebras $\mathfrak{g}_n$. Consequently, the value of $\mu(\mathfrak{g})$ is less than or equal to the dimension to be computed and determined in this paper. In particular, Ghanam et al. [8] computed matrix representations for low dimensional nilpotent Lie algebras, but their minimality was not studied. In fact, some representations in [8] were not minimal.

Independently, Echarte et al. [5] introduced some invariants of filiform Lie algebras, improving them in [6]. In this paper, we use these invariants to express and classify the law of filiform Lie algebras.

The structure of this paper is as follows: after reviewing some well-known results on Lie Theory in Section 2, Section 3 is devoted to show the general method used to compute a minimal faithful unitriangular matrix representation for filiform Lie algebras. Due to reasons of length, we only compute explicitly minimal faithful unitriangular matrix representations for filiform Lie algebras of dimension less than nine, although the method can be applied to any arbitrary finite-dimensional filiform Lie algebra provided its law is known, which is not easy for higher dimensions. Remember that the classifications of filiform Lie algebras are only known up to dimension 12, inclusive.

2 Preliminaries

Some preliminary concepts on Lie algebras are recalled in this section. For a general overview, the reader can consult [9]. Let us note that only finite-dimensional Lie algebras over the complex number field $\mathbb{C}$ are considered from here on.

2.1 Lie Algebras

The *lower central series* of a given Lie algebra $\mathfrak{g}$ is defined by

$$\mathcal{C}^1(\mathfrak{g}) = \mathfrak{g},\ \mathcal{C}^2(\mathfrak{g}) = [\mathfrak{g}, \mathfrak{g}],\ \mathcal{C}^3(\mathfrak{g}) = [\mathcal{C}^2(\mathfrak{g}), \mathfrak{g}],\ \ldots,\ \mathcal{C}^k(\mathfrak{g}) = [\mathcal{C}^{k-1}(\mathfrak{g}), \mathfrak{g}],\ \ldots$$

The Lie algebra $\mathfrak{g}$ is *nilpotent* when there exists a natural integer m such that $\mathcal{C}^m(\mathfrak{g}) \equiv 0$.

 ISBN: 978-84-614-6167-7

Let $\mathfrak{h}$ be a subalgebra of a Lie algebra $\mathfrak{g}$. The centralizer of $\mathfrak{h}$ in $\mathfrak{g}$ is the set of elements of $\mathfrak{g}$ which commutes with all the elements of $\mathfrak{h}$.

Related to the lower central series associated with a subalgebra of $\mathfrak{g}$, the following result holds:

Proposition 1 *Let $\mathfrak{h}$ be a subalgebra of a Lie algebra $\mathfrak{g}$. Then $\mathcal{C}^k(\mathfrak{h}) \subseteq \mathcal{C}^k(\mathfrak{g})$, $\forall\, k \in \mathbb{N}$.*

Let us denote by $\mathfrak{g}_n$ the nilpotent matrix algebra of $n \times n$ strictly upper-triangular matrices, where $n \in \mathbb{N} \setminus \{1\}$. The expression of the vectors in $\mathfrak{g}_n$ is the following

$$g_n(x_{r,s}) = \begin{pmatrix} 0 & x_{1,2} & \cdots & x_{1,n-1} & x_{1,n} \\ 0 & 0 & \cdots & x_{2,n-1} & x_{2,n} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & x_{n-1,n} \\ 0 & 0 & \cdots & 0 & 0 \end{pmatrix} \qquad (x_{i,j} \in \mathbb{C}).$$

The dimension of $\mathfrak{g}_n$ is $\frac{n(n-1)}{2}$. Fixed i and j such that $1 \leq i < j \leq n$, a basis of $\mathfrak{g}_n$ is $\mathcal{B}_n = \{X_{i,j} = g_n(x_{r,s}) \,|\, [x_{r,s} = 1 \Leftrightarrow (r,s) = (i,j)] \wedge [x_{r,s} = 0 \Leftrightarrow (r,s) \neq (i,j)]\}_{1 \leq i < j \leq n}$ with the law: $[X_{i,j}, X_{j,k}] = X_{i,k}$, for $1 \leq i < j < k \leq n$. Consequently, the dimension of each term in the lower central series of $\mathfrak{g}_n$ is

$$(\dim(\mathfrak{g}_n),\, \dim(\mathfrak{g}_{n-1}),\, \dim(\mathfrak{g}_{n-2}), \ldots,\, \dim(\mathfrak{g}_2), 0) \tag{1}$$

A particular family of nilpotent Lie algebras is formed by abelian Lie algebras. A Lie algebra $\mathfrak{g}$ is said to be *abelian* if $[\mathbf{v}, \mathbf{w}] = 0$, for all $\mathbf{v}$, $\mathbf{w} \in \mathfrak{g}$. An equivalent condition is the following: $Z(\mathfrak{g}) = \mathfrak{g}$, where

$$Z(\mathfrak{g}) = \{X \in \mathfrak{g} \quad | \quad [X, Y] = 0\,,\ \forall Y \in \mathfrak{g}\}$$

We also consider a second subclass of nilpotent Lie algebras in this paper: Filiform Lie algebras. An n-dimensional Lie algebra $\mathfrak{g}$ is *filiform* if its lower central series satisfies the following

$$\dim(\mathcal{C}^1(\mathfrak{g})) = n,\ \dim(\mathcal{C}^2(\mathfrak{g})) = n-2,\ \dim(\mathcal{C}^3(\mathfrak{g})) = n-3,\ \ldots, \dim(\mathcal{C}^n(\mathfrak{g})) = 0. \tag{2}$$

A basis $\{e_i\}_{i=1}^n$ of the filiform Lie algebra $\mathfrak{g}$ is called an *adapted basis* if these relations are verified

$$\begin{array}{ll} [e_1, e_h] = e_{h-1}, & \text{for } 3 \leq h \leq n; \\ [e_2, e_h] = 0, & \text{for } 1 \leq h \leq n; \\ [e_3, e_h] = 0, & \text{for } 2 \leq h \leq n. \end{array} \tag{3}$$

Remark 1 *If $\{e_i\}_{i=1}^n$ is an adapted basis of an n-dimensional filiform Lie algebra $\mathfrak{g}$, then the vector e_2 belongs to the center $Z(\mathfrak{g})$ of the algebra.*

A filiform Lie algebra $\mathfrak{g}$ is called *model* if the only nonzero brackets in its law are $[e_1, e_h] = e_{h-1}$, for $3 \leq h \leq n$.

 ISBN: 978-84-614-6167-7

2.2 Invariants of Filiform Lie Algebras

This subsection is devoted to recall the definitions of two invariants for filiform Lie algebras given in [6]. First, the invariant z_1 is defined as follows

$$z_1 = \max\{k \in \mathbb{N} \,|\, C_{\mathfrak{g}}(\mathcal{C}^{n-k+2}(\mathfrak{g})) \supset \mathcal{C}^2(\mathfrak{g})\},$$

where $C_{\mathfrak{g}}(\mathfrak{h})$ is the centralizer of a given subalgebra $\mathfrak{h}$ of $\mathfrak{g}$. Let us note that the set in the previous definition may be empty. In this case, it is easy to see that $\mathfrak{g}$ is a model filiform Lie algebra. Besides, the definition of z_1 means that the ideal $\mathcal{C}^{n-i+2}(\mathfrak{g})$ is the largest whose centralizer contains $\mathcal{C}^2(\mathfrak{g})$. Let us note that the previous definition is equivalent to the following: $z_1 = \min\{k \geq 2 \,|\, [e_k, e_n] \neq 0\}$, which is more convenient to be used in the practice, and where $\{e_i\}_{i=1}^n$ is an adapted basis of $\mathfrak{g}$.

The invariant z_2 is defined as

$$z_2 = \max\{k \in \mathbb{N} \,|\, \mathcal{C}^{n-k+1}(\mathfrak{g}) \text{ is abelian}\}.$$

An immediate consequence of this definition is that the ideal $\mathcal{C}^{n-j+1}(\mathfrak{g}) \equiv \langle e_2, \ldots, e_j \rangle$ is the largest abelian subalgebra in the lower central series of $\mathfrak{g}$.

3 Computing Minimal Matrix Representations

In this section we obtain a minimal faithful unitriangular matrix representation for each model filiform Lie algebra. Additionally, we also introduce a method for obtaining such representations for non-model filiform Lie algebras. Finally, we apply our method to compute minimal faithful unitriangular matrix representations of filiform Lie algebras with dimension less than 9.

Given a Lie algebra $\mathfrak{g}$, a *representation* of $\mathfrak{g}$ in $\mathbb{C}^n$ is a Lie-algebra homomorphism $\phi : \mathfrak{g} \to \mathfrak{gl}(\mathbb{C}^n) = \mathfrak{gl}(\mathbb{C}, n)$. The natural integer n is called the *dimension* of this representation. Ado's theorem states that every finite-dimensional Lie algebra over a field of characteristic zero (as in the case of $\mathbb{C}$) has a linear injective representation on a finite-dimensional vector space, that is, a *faithful representation*.

Usually, representations are defined by using an arbitrary n-dimensional vector space V (like in [7]) and homomorphisms of Lie algebras from $\mathfrak{g}$ to $\mathfrak{gl}(V)$ of endomorphisms of V; that is, by using $\mathfrak{g}$-*modules*.

With respect to minimal representations of Lie algebras, Burde [4] introduced the following invariant value for an arbitrary Lie algebra $\mathfrak{g}$

$$\mu(\mathfrak{g}) = \min\{\dim(M) \mid M \text{ is a faithful } \mathfrak{g}\text{-module}\}.$$

In this section, matrix representations of filiform Lie algebras are studied. Moreover, we are interested in minimal matrix representations of these algebras with a particular restriction: the representations have to be contained in $\mathfrak{g}_n$. In this way, given a filiform Lie algebra $\mathfrak{g}$, we want to compute the minimal value n such that $\mathfrak{g}_n$ contains a subalgebra isomorphic to $\mathfrak{g}$. This value is the invariant of $\mathfrak{g}$ expressed as follows

$$\bar{\mu}(\mathfrak{g}) = \min\{n \in \mathbb{N} \mid \exists \text{ subalgebra of } \mathfrak{g}_n \text{ isomorphic to } \mathfrak{g}\}.$$

 ISBN: 978-84-614-6167-7

Let us note that the invariants $\mu(\mathfrak{g})$ and $\bar{\mu}(\mathfrak{g})$ can be different from each other.

Proposition 2 *Let $\mathfrak{g}$ be an n-dimensional filiform Lie algebra. Then $\bar{\mu}(\mathfrak{g}) \geq n$.*

Proof: For a given n-dimensional filiform Lie algebra $\mathfrak{g}$, we have to prove that it is not possible to find a subalgebra of $\mathfrak{g}_{n-1}$ isomorphic to $\mathfrak{g}$.

First, we write the vectors of an adapted basis $\{e_i\}_{i=1}^n$ of $\mathfrak{g}$ as linear combinations of the vectors in the basis $\mathcal{B}_{n-1}$ of $\mathfrak{g}_{n-1}$

$$e_k = \sum_{1 \leq i < j \leq n-1} \lambda_{i,j}^k X_{i,j}, \text{ for } 1 \leq k \leq n.$$

We will prove that all the coefficients $\lambda_{i,j}^2$ of $e_2 \in Z(\mathfrak{g})$ have to be zero.

From $[e_1, e_h] = e_{h-1}$, for $3 \leq h \leq n$, the following relations are obtained

$$\left.\begin{array}{l} \lambda_{\beta,\beta+1}^{h-1} = 0, \\ \lambda_{\beta,\alpha_\beta}^{h-1} = \sum_{\beta<p<\alpha_\beta} (\lambda_{\beta,p}^1 \lambda_{p,\alpha_\beta}^h - \lambda_{p,\alpha_\beta}^1 \lambda_{\beta,p}^h), \end{array}\right\} \text{ for } \left\{\begin{array}{l} 1 \leq \beta \leq n-2; \\ \alpha_\beta \geq \beta + 2. \end{array}\right. \tag{4}$$

From $[e_1, e_3] = e_2$, we can conclude that $\lambda_{\beta,\beta+1}^2 = 0$, for $1 \leq \beta \leq n-2$. Now, we have to prove that $\lambda_{l,\alpha_l}^2 = 0$, for $1 \leq l \leq n-3$. To do so, we are going to prove that $\lambda_{p,\alpha_\beta}^3 = \lambda_{\beta,p}^3 = 0$ in each case.

From $[e_1, e_k] = e_{k-1}$, for $3 \leq k \leq n-1$, we can affirm that $\lambda_{\beta,\beta+1}^{k-1} = 0$, for $1 \leq \beta \leq n-2$. This implies that $\lambda_{p,q}^3 = 0$, when $q - p < n - 4$.

If we consider the bracket $[e_1, e_n] = e_{n-1}$, we conclude that $\lambda_{\beta,\beta+1}^{n-1} = 0$, and, therefore, $\lambda_{p,q}^3 = 0$, where $q - p = n - 3$. Consequently, all the coefficients of e_2 are null and this is a contradiction. □

3.1 Model Filiform Lie Algebras

The law of a fixed n-dimensional model filiform Lie algebra $\mathfrak{g}$ with an adapted basis $\{e_i\}_{i=1}^n$ is the following

$$[e_1, e_h] = e_{h-1}, \quad \text{for } 3 \leq h \leq n. \tag{5}$$

Now, we will construct the n-dimensional subalgebra $\mathfrak{f}_n'$ of $\mathfrak{g}_n$ and whose law is exactly the same of the model filiform Lie algebra $\mathfrak{g}$. Just define a basis $\{e_i\}_{i=1}^n$ of this subalgebra as linear combinations of the vectors in the basis $\mathcal{B}_n$ of the Lie algebra $\mathfrak{g}_n$

$$e_1 = \sum_{i=1}^{n-2} X_{i,i+1}, \; e_2 = X_{1,n}, \; e_3 = X_{2,n}, \; \ldots, \; e_n = X_{n-1,n} \tag{6}$$

Consequently, we have defined the subalgebra $\mathfrak{f}_n'$ whose elements are the following

$$f_n'(x_k) = \begin{pmatrix} 0 & x_1 & 0 & \cdots & 0 & x_2 \\ 0 & 0 & x_1 & \cdots & 0 & x_3 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & x_1 & x_{n-1} \\ 0 & 0 & 0 & \cdots & 0 & x_n \\ 0 & 0 & 0 & \cdots & 0 & 0 \end{pmatrix} \qquad (x_k \in \mathbb{C}, \text{ for } k = 1, \ldots, n).$$

 ISBN: 978-84-614-6167-7

With respect to this faithful matrix representation, the adapted basis $\{e_i\}_{i=1}^n$ of $\mathfrak{f}'_n$ is given as follows

$$e_h = f'_n(x_k), \quad \text{with } x_k = \begin{cases} 1, & \text{if } k = h; \\ 0, & \text{if } k \neq h. \end{cases}$$

According to the reasoning just given above and Proposition 2, we can affirm the following result

Proposition 3 Every n-dimensional model filiform Lie algebra has an n-dimensional minimal faithful unitriangular matrix representation. Moreover, a representation of this type is the Lie algebra $\mathfrak{f}'_n$ given in (6). □

3.2 Non-Model Filiform Lie Algebras

If the filiform Lie algebra $\mathfrak{g}$ is non-model, then the invariants z_1 and z_2 exist. Hence, there exist some additional nonzero brackets to $[e_1, e_h] = e_{h-1}$, for $3 \leq h \leq n$. Consequently, non-model filiform Lie algebras cannot be represented by the algebras $\mathfrak{f}'_n$.

Now, we show an algorithmic method to compute minimal faithful unitriangular matrix representations for non-model filiform Lie algebras. These representations are minimal in the following sense: Finding a faithful matrix representation of a given Lie algebra $\mathfrak{g}$ in $\mathfrak{g}_n$, but no representations of $\mathfrak{g}$ can be obtained in $\mathfrak{g}_{n-1}$.

To do so, we give a step-by-step explanation of the method used to determine these minimal representations for a given filiform Lie algebra $\mathfrak{g}$ of dimension $n > 4$.

1. According to Proposition 1, we have to compute the first natural integer l such that the lower central series of $\mathfrak{g}_l$ is compatible with the one associated with $\mathfrak{g}$. By bearing in mind Proposition 2, we can start considering $l = n$, for n-dimensional filiform Lie algebras. Hence, we have ruled out the Lie algebra $\mathfrak{g}_l$ with $l < n$.

2. Now, we search a subalgebra of $\mathfrak{g}_l$ isomorphic to $\mathfrak{g}$, where $l \geq n$ and l is as little as posible. To do so, an adapted basis $\{e_i\}_{i=1}^n$ of $\mathfrak{g}$ is considered and its vectors are expressed as linear combinations of the basis $\mathcal{B}_l$

$$e_h = \sum_{1 \leq i < j \leq l} \lambda_{i,j}^h X_{i,j}, \quad \text{for } 1 \leq h \leq n.$$

3. The next step is to impose the brackets given in (5), obtaining again the equations shown in (4), but with respect to the algebra $\mathfrak{g}_l$ considered by the method.

4. After solving the system of equations resulting from the previous step, we solve a new system obtained by imposing the rest of the brackets in the law of $\mathfrak{g}$.

Obviously, the solutions of this system depend on the particular Lie algebra studied in each moment. Hence, we have generalized some cases by using the invariants z_1 and z_2 in Section 4. The solutions of the system given in Step 4 have been computed

 ISBN: 978-84-614-6167-7

by using the command `solve` of the symbolic computation package Maple 12. This command works efficiently with polynomial equations, receives as inputs the list of equations and the list of variables, and returns as output the algebraic expression of the solutions.

After checking the existence of representations in the Lie algebra $\mathfrak{g}_l$, we search a natural representative in the sense of considering the following conditions: $e_2 \in \langle X_{1,l} \rangle$ and there exist the greatest possible number of null-parameters. The first condition is due to the fact that $Z(\mathfrak{g}_l) = \langle X_{1,l} \rangle$ and is also in concordance with Proposition 2.

Another point to consider is the number of solutions for the system. This number can be studied by defining the set F of polynomial expressions and by using the command `is_finite`, which determines if the number of solutions is or not finite for the system defined by the input set F. Likewise, Noether's normalization lemma is also very useful to describe the elements in an algebraic variety.

Furthermore, in order to compute a particular solution of the previous system, we have searched one whose number of null coefficients is as greater as possible. In this way, the coefficients could be assumed equal to zero when they do not appear in the relations obtained. This will be a natural representative of the Lie algebra $\mathfrak{g}$.

4 Minimal Representations of Filiform Lie Algebras of Dimension less than 9

This section is devoted to show explicit representatives for the minimal faithful unitriangular matrix representation of filiform Lie algebras with dimension less than 9 and the generalization for two families of n-dimensional filiform Lie algebras. In Tables 1–3, we write those representatives by using the classification given in [3]. Let us note that, for these algebras, we have only written down the nonzero brackets not given by filiformity; i.e. we are assuming implicitly Equation (5). In virtue of these tables, we can state the following result:

Proposition 4 If $\mathfrak{g}$ is a filiform Lie algebra of dimension $n < 9$, then $\bar{\mu}(\mathfrak{g}) = n$ (i.e. the minimal faithful matrix representation of $\mathfrak{g}$ is a subalgebra of $\mathfrak{g}_n$). Moreover, such representations can be obtained with a natural representative. □

 ISBN: 978-84-614-6167-7

Table 1: Minimal faithful matrix representations for dimension ≤ 7

	Law	**Representation**
$\mathfrak{f}_3^1$		$e_1 = X_{1,2}, e_2 = X_{1,3}, e_3 = X_{2,3}.$
$\mathfrak{f}_4^1$		$e_1 = X_{1,2} + X_{2,3}, e_2 = X_{1,4}, e_3 = X_{2,4}, e_4 = X_{3,4}.$
$\mathfrak{f}_5^1$		$e_1 = X_{1,2} + X_{2,3} + X_{3,4}, e_2 = X_{1,5},$ $e_3 = X_{2,5}, e_4 = X_{3,5}, e_5 = X_{4,5}.$
$\mathfrak{f}_5^2$	$[e_4, e_5] = e_2.$	$e_1 = X_{1,2} + X_{2,3} + X_{3,4}, e_2 = X_{1,5}, e_3 = X_{2,5},$ $e_4 = X_{1,4} + X_{3,5}, e_5 = X_{2,4} + X_{4,5}.$
$\mathfrak{f}_6^1$		$e_1 = X_{1,2} + X_{2,3} + X_{3,4} + X_{4,5}, e_2 = X_{1,6}, e_3 = X_{2,6},$ $e_4 = X_{3,6}, e_5 = X_{4,6}, e_6 = X_{5,6}.$
$\mathfrak{f}_6^2$	$[e_5, e_6] = e_2.$	$e_1 = \sum_{i=1}^{4} X_{i,i+1}, e_2 = X_{1,6}, e_3 = X_{2,6},$ $e_4 = X_{3,6}, e_5 = X_{1,5} + X_{4,6}, e_6 = X_{2,5} + X_{4,6}.$
$\mathfrak{f}_6^3$	$[e_4, e_6] = e_2,$ $[e_5, e_6] = e_3.$	$e_1 = \sum_{i=1}^{4} X_{i,i+1}, e_2 = X_{1,6}, e_3 = X_{2,6},$ $e_4 = X_{3,6} + X_{1,5}, e_5 = X_{2,5} + X_{4,6}, e_6 = X_{3,5} + X_{5,6}.$
$\mathfrak{f}_6^4$	$[e_4, e_5] = e_2$ $[e_4, e_6] = e_3,$ $[e_5, e_6] = e_4.$	$e_1 = X_{1,2} + X_{2,3} + X_{3,4} + X_{4,5}, e_2 = X_{1,6},$ $e_3 = \frac{-1}{2}X_{1,5} + X_{2,6}, e_4 = \frac{1}{2}X_{1,4} + \frac{1}{3}X_{1,5} + X_{3,6},$ $e_5 = \frac{-1}{2}X_{1,3} + \frac{1}{3}X_{2,5} + X_{4,6}, e_6 = \frac{-1}{2}X_{1,2} +$ $\frac{1}{3}X_{1,3} - X_{2,3} + \frac{1}{3}X_{2,4} - X_{3,4} - \frac{1}{3}X_{3,5} - X_{4,5} + X_{5,6}$
$\mathfrak{f}_6^5$	$[e_4, e_5] = e_2$ $[e_4, e_6] = e_3 + e_2,$ $[e_5, e_6] = e_4 + e_3.$	$e_1 = X_{1,2} + X_{2,3} + X_{3,4} + X_{4,5}, e_2 = X_{1,6}, e_3 = \frac{-1}{2}X_{1,5}$ $+X_{2,6}, e_4 = \frac{1}{2}X_{1,4} + X_{3,6}, e_5 = -\frac{1}{2}X_{1,3} + X_{2,5} + X_{4,6},$ $e_6 = -\frac{1}{2}X_{1,2} - X_{1,3} - X_{2,3} - X_{3,4} - X_{4,5} + X_{5,6}.$
$\mathfrak{f}_7^1$		$e_1 = \sum_{i=1}^{5} X_{i,i+1}, e_2 = X_{1,7}, e_3 = X_{2,7},$ $e_4 = X_{3,7}, e_5 = X_{4,7}, e_6 = X_{5,7}, e_7 = X_{6,7}.$
$\mathfrak{f}_7^2$	$[e_6, e_7] = e_2$	$e_1 = \sum_{i=1}^{5} X_{i,i+1}, e_2 = X_{1,7}, e_3 = X_{2,7}, e_4 = X_{3,7},$ $e_5 = X_{4,7}, e_6 = X_{1,6} + X_{5,7}, e_7 = X_{2,6} + X_{6,7}.$
$\mathfrak{f}_7^3$	$[e_5, e_7] = e_2,$ $[e_6, e_7] = e_3$	$e_1 = \sum_{i=1}^{5} X_{i,i+1}, e_2 = X_{1,7}, e_3 = X_{2,7}, e_4 = X_{3,7},$ $e_5 = X_{4,7}, e_6 = X_{5,7}, e_7 = -X_{1,4} - X_{2,5} - X_{3,6} + X_{6,7}.$
$\mathfrak{f}_7^4$	$[e_5, e_7] = e_2,$ $[e_6, e_7] = e_2 + e_3$	$e_1 = \sum_{i=1}^{5} X_{i,i+1}, e_2 = X_{1,7}, e_3 = X_{2,7}, e_4 = X_{3,7}, e_5 = X_{1,6}$ $+X_{4,7}, e_6 = X_{2,6} + X_{5,7}, e_7 = -X_{1,5} - X_{2,6} + X_{3,6} + X_{6,7}.$
$\mathfrak{f}_7^5$	$[e_4, e_7] = [e_5, e_7] = e_2$ $[e_5, e_6] = -e_2,$ $[e_6, e_7] = e_3$	$e_1 = \sum_{i=1}^{5} X_{i,i+1}, e_2 = X_{1,7}, e_3 = X_{2,7}, e_4 = -X_{1,6} + X_{3,7},$ $e_5 = X_{1,6} - X_{2,6} + X_{4,7}, e_6 = X_{1,4} + X_{2,5} + X_{2,6} + X_{5,7},$ $e_7 = -2X_{1,3} - X_{2,4} + X_{3,6} + X_{6,7}.$
$\mathfrak{f}_7^6$	$[e_4, e_7] = e_2,$ $[e_5, e_7] = e_3$ $[e_6, e_7] = e_2 + e_4$	$e_1 = \sum_{i=1}^{5} X_{i,i+1}, e_2 = X_{1,7}, e_3 = X_{2,7}, e_4 = X_{3,7},$ $e_5 = X_{4,7}, e_6 = X_{5,7}, e_7 = -X_{1,3} - X_{1,5} - X_{2,4} -$ $X_{2,6} - X_{3,5} - X_{4,6} - X_{5,6} + X_{6,7}.$
$\mathfrak{f}_7^7$	$[e_5, e_7] = e_3,$ $[e_4, e_7] = e_2$ $[e_6, e_7] = e_4$	$e_1 = \sum_{i=1}^{5} X_{i,i+1}, e_2 = X_{1,7}, e_3 = X_{2,7}, e_4 = X_{3,7},$ $e_5 = X_{4,7}, e_6 = X_{5,7}, e_7 = -X_{1,3} - X_{2,4} - X_{3,5} -$ $X_{3,5} - X_{4,6} - X_{5,6} + X_{6,7}.$
$\mathfrak{f}_7^8$	$[e_4, e_7] = \alpha e_2,$ $[e_5, e_6] = e_2$ $[e_5, e_7] = (1+\alpha)e_3,$ $[e_6, e_7] = (1+\alpha)e_4$	$e_1 = X_{1,3} + X_{2,3} - X_{2,4} + X_{3,4} + \beta i X_{3,6} + X_{4,5} + iX_{5,6} + X_{6,7},$ $e_2 = -iX_{1,7}, e_3 = iX_{1,6} - iX_{1,7}, e_4 = -X_{1,5} + iX_{1,6} +$ $(\frac{1}{2}i\beta - 1 - \frac{1}{2}i)X_{1,7} + \frac{1}{2}i(\beta+1+2\alpha)X_{2,7}, e_5 = X_{1,4} - X_{1,5}$ $+X_{1,6} - (1+\alpha)iX_{2,6} + \frac{1}{2}i(\beta+3+2\alpha)X_{2,7} + \frac{1}{2}i(\beta-1)X_{3,7},$ $e_6 = -X_{1,3} + X_{1,4} + iX_{1,5} + (\alpha - \beta + 1)X_{2,5} - i(1+\alpha)X_{2,6} +$ $(\beta + \frac{4}{3}\alpha i + 2 + i + \alpha + \frac{1}{3}i\alpha^2)X_{2,7} - i\beta X_{3,6} - \frac{1}{2}i(\beta+1)X_{4,7},$ $e_7 = (\beta - i)X_{1,4} + X_{1,2} + (2\beta - 1 - \alpha)X_{2,4} + (1+\alpha)X_{2,5} -$ $(\beta + \frac{4}{3}i\alpha - 2 - i - \alpha - \frac{1}{3}i\alpha^2)X_{2,6} + \beta X_{3,5} - \frac{1}{2}i(\beta+1)X_{5,7},$ β is a root of $3Z^2 - 2\alpha^2 - 5\alpha - 3 - 3\alpha Z.$

 ISBN: 978-84-614-6167-7

Table 2: Minimal faithful matrix representations for dimension 8 (I)

	Law	Representation
$\mathfrak{f}_8^1$		$e_1=\sum_{i=1}^{6} X_{i,i+1}, e_2=X_{1,8}, e_3=X_{2,8}, e_4=X_{3,8}, e_5=X_{4,8}, e_6=X_{5,8}, e_7=X_{6,8}, e_8=X_{7,8}$.
$\mathfrak{f}_8^2$	$[e_7,e_8]=e_2$	$e_1=\sum_{i=1}^{6} X_{i,i+1}, e_2=X_{1,8}, e_3=X_{2,8}, e_4=X_{3,8}, e_5=X_{4,8}, e_6=X_{5,8}, e_7=X_{1,7}+X_{6,8}, e_8=X_{2,7}+X_{7,8}$.
$\mathfrak{f}_8^3$	$[e_6,e_8]=e_2$, $[e_7,e_8]=e_3$	$e_1=\sum_{i=1}^{6} X_{i,i+1}, e_2=X_{1,8}, e_3=X_{2,8}, e_4=X_{3,8}, e_5=X_{4,8}, e_6=X_{5,8}, e_7=X_{6,8}, e_8=-X_{1,5}-X_{2,6}-X_{3,7}+X_{7,8}$.
$\mathfrak{f}_8^4$	$[e_6,e_8]=e_2$, $[e_7,e_8]=e_2+e_3$	$e_1=\sum_{i=1}^{6} X_{i,i+1}, e_2=X_{1,8}, e_3=X_{2,8}, e_4=X_{3,8}, e_5=X_{4,8}, e_6=X_{5,8}, e_7=X_{6,8}, e_8=-X_{1,5}-X_{1,6}-X_{2,6}-X_{2,7}-X_{3,7}+X_{7,8}$.
$\mathfrak{f}_8^5$	$[e_5,e_8]=e_2$, $[e_6,e_8]=e_3$, $[e_7,e_8]=e_2+e_4$	$e_1=\sum_{i=1}^{6} X_{i,i+1}, e_2=X_{1,8}, e_3=X_{2,8}, e_4=X_{3,8}, e_5=X_{4,8}, e_6=X_{5,8}, e_7=X_{6,8}, e_8=-X_{1,4}-X_{1,6}-X_{2,5}-X_{2,7}-X_{3,6}-X_{4,7}+X_{7,8}$
$\mathfrak{f}_8^6$	$[e_5,e_8]=e_2$, $[e_6,e_8]=e_2+e_3$, $[e_7,e_8]=\alpha e_2+e_3+e_4$	$e_1=\sum_{i=1}^{6} X_{i,i+1}, e_2=X_{1,8}, e_3=X_{2,8}, e_4=X_{3,8}, e_5=X_{4,8}, e_6=X_{5,8}, e_7=X_{6,8}, e_8=-X_{1,4}-X_{1,5}-\alpha X_{1,6}+X_{1,8}-X_{2,5}-X_{2,6}-\alpha X_{2,7}-X_{3,6}-X_{3,7}-X_{4,7}+X_{7,8}$
$\mathfrak{f}_8^7$	$[e_5,e_8]=\alpha e_2$, $[e_6,e_7]=e_2$, $[e_6,e_8]=(1+\alpha)e_3$, $[e_7,e_8]=(1+\alpha)e_4$	$e_1=\sum_{i=1}^{6} X_{i,i+1}, e_2=X_{1,8}, e_3=X_{2,8}, e_4=X_{3,8}, e_5=-X_{1,7}+X_{4,8}, e_6=X_{1,6}+X_{5,8}, e_7=X_{2,6}+X_{3,7}+X_{6,8}, e_8=-(1+\alpha)X_{1,4}+X_{1,8}-(1+\alpha)X_{2,5}-\alpha X_{3,6}+(1-\alpha)X_{4,7}+X_{7,8}$
$\mathfrak{f}_8^8$	$[e_5,e_8]=\alpha e_2$, $[e_6,e_7]=e_2$, $[e_6,e_8]=(1+\alpha)e_3+e_2$, $[e_7,e_8]=(1+\alpha)e_4+e_3$	$e_1=\sum_{i=1}^{6} X_{i,i+1}, e_2=X_{1,8}, e_3=X_{2,8}, e_4=X_{3,8}, e_5=-X_{1,7}+X_{4,8}, e_6=X_{1,6}+X_{5,8}, e_7=X_{2,6}+X_{3,7}+X_{6,8}, e_8=-(1+\alpha)X_{1,4}-X_{1,5}+X_{1,8}-(1+\alpha)X_{2,5}-X_{2,6}-\alpha X_{3,6}-X_{3,7}+(1-\alpha)X_{4,7}+X_{7,8}$.
$\mathfrak{f}_8^9$	$[e_4,e_8]=e_2$, $[e_5,e_8]=e_3$, $[e_6,e_8]=e_4$, $[e_7,e_8]=e_5$	$e_1=\sum_{i=1}^{6} X_{i,i+1}, e_2=X_{1,8}, e_3=X_{2,8}, e_4=X_{1,7}+X_{3,8}, e_5=X_{2,7}+X_{4,8}, e_6=-\frac{1}{2}X_{1,5}-\frac{1}{2}X_{2,6}+\frac{1}{2}X_{3,7}+X_{5,8}, e_7=-\frac{1}{2}X_{2,5}-X_{3,6}-\frac{1}{2}X_{4,7}+X_{6,8}, e_8=-\frac{1}{2}X_{3,5}-\frac{3}{2}X_{4,6}-2X_{5,7}+X_{7,8}$.
$\mathfrak{f}_8^{10}$	$[e_4,e_8]=e_2$, $[e_5,e_8]=e_3$, $[e_6,e_8]=e_4$, $[e_7,e_8]=e_2+e_5$	$e_1=\sum_{i=1}^{6} X_{i,i+1}, e_2=X_{1,8}, e_3=X_{2,8}, e_4=X_{1,7}+X_{3,8}, e_5=X_{2,7}+X_{4,8}, e_6=-\frac{1}{2}X_{1,5}-\frac{1}{2}X_{2,6}+\frac{1}{2}X_{3,7}+X_{5,8}, e_7=-\frac{1}{2}X_{2,5}-X_{3,6}-\frac{1}{2}X_{4,7}+X_{6,8}, e_8=-X_{1,6}-X_{2,7}-\frac{1}{2}X_{3,5}-\frac{3}{2}X_{4,6}-2X_{5,7}+X_{7,8}$.
$\mathfrak{f}_8^{11}$	$[e_4,e_8]=e_2$, $[e_5,e_8]=e_3$, $[e_6,e_8]=e_2+e_4$, $[e_7,e_8]=\alpha e_2+e_3+e_5$	$e_1=\sum_{i=1}^{6} X_{i,i+1}, e_2=X_{1,8}, e_3=X_{2,8}, e_4=X_{1,7}+X_{3,8}, e_5=X_{2,7}+X_{4,8}, e_6=-\frac{1}{2}X_{1,5}-\frac{1}{2}X_{2,6}+\frac{1}{2}X_{3,7}+X_{5,8}, e_7=-\frac{1}{2}X_{2,5}-X_{3,6}-\frac{1}{2}X_{4,7}+X_{6,8}, e_8=-X_{1,5}-\alpha X_{1,6}-X_{2,6}-\alpha X_{2,7}-\frac{1}{2}X_{3,5}-X_{3,7}-\frac{3}{2}X_{4,6}-2X_{5,7}+X_{7,8}$.
$\mathfrak{f}_8^{12}$	$[e_4,e_8]=e_2$, $[e_5,e_8]=e_3$, $[e_6,e_7]=e_2$, $[e_6,e_8]=\alpha e_2+e_3+e_4$, $[e_7,e_8]=\alpha e_3+e_4+e_5$	$e_1=\sum_{i=1}^{6} X_{i,i+1}, e_2=X_{1,8}, e_3=X_{2,8}, e_4=X_{1,7}+X_{3,8}, e_5=-2X_{1,7}+X_{2,7}+X_{4,8}, e_6=-\frac{1}{2}X_{1,5}+X_{1,6}-\frac{1}{2}X_{2,6}-X_{2,7}+\frac{1}{2}X_{3,7}+X_{5,8}, e_7=-\frac{1}{2}X_{2,5}+X_{2,6}-X_{3,6}-\frac{1}{2}X_{4,7}+X_{6,8}, e_8=-2X_{1,4}-\alpha X_{1,5}-2X_{2,5}-\alpha X_{2,6}-\frac{1}{2}X_{3,5}-X_{3,6}-\alpha X_{3,7}-\frac{3}{2}X_{4,6}-X_{4,7}-2X_{5,7}+X_{7,8}$.
$\mathfrak{f}_8^{13}$	$[e_4,e_8]=-e_2$, $[e_6,e_8]=e_3+e_4$, $[e_6,e_7]=e_2+e_3$, $[e_5,e_7]=e_2$, $[e_7,e_8]=e_4+e_5$	$e_1=\sum_{i=1}^{6} X_{i,i+1}, e_2=X_{1,8}, e_3=X_{2,8}, e_4=X_{3,8}, e_5=X_{1,6}+X_{2,7}+X_{4,8}, e_6=-X_{1,5}+\frac{1}{2}X_{1,6}+\frac{1}{2}X_{2,7}+X_{3,7}+X_{5,8}, e_7=-\frac{1}{2}X_{1,5}-X_{2,5}-X_{3,6}+\frac{1}{2}X_{3,7}+X_{6,8}, e_8=X_{1,3}+X_{2,4}-\frac{1}{2}X_{2,5}-\frac{1}{2}X_{3,6}-X_{4,6}-X_{5,7}+X_{7,8}$.
$\mathfrak{f}_8^{14}$	$[e_4,e_8]=\alpha e_2$, $[e_5,e_7]=e_2$, $[e_5,e_8]=-e_3$, $[e_6,e_7]=e_3$	$e_1=\sum_{i=1}^{6} X_{i,i+1}, e_2=X_{1,8}, e_3=X_{2,8}, e_4=-2X_{1,7}+X_{3,8}, e_5=X_{1,6}-X_{2,7}+X_{4,8}, e_6=X_{2,6}+X_{5,8}, e_7=X_{3,6}+X_{4,7}+X_{6,8}, e_8=X_{4,6}+2X_{5,7}+X_{7,8}$.

 ISBN: 978-84-614-6167-7

Table 3: Minimal faithful matrix representations for dimension 8 (II)

	Law	**Representation**
$\mathfrak{f}_8^{15}$	$[e_4,e_8]=-2e_2$, $[e_5,e_7]=e_2$ $[e_5,e_8]=-e_3+e_2$, $[e_6,e_7]=e_3$ $[e_6,e_8]=e_3$, $[e_7,e_8]=e_4$	$e_1=\sum_{i=1}^{6} X_{i,i+1}, e_2=X_{1,8}, e_3=X_{2,8}$, $e_4=-2X_{1,7}+X_{3,8}, e_5=X_{1,6}+3X_{1,7}-$ $X_{2,7}+X_{4,8}, e_6=X_{2,6}+3X_{2,7}+X_{5,8}$, $e_7=X_{3,6}+X_{3,7}+X_{4,7}+X_{6,8}$, $e_8=2X_{1,4}+2X_{2,5}+2X_{3,6}+$ $X_{4,6}+5X_{4,7}+2X_{5,7}+X_{7,8}$.
$\mathfrak{f}_8^{16}$	$[e_4,e_7]=e_2$, $[e_5,e_6]=-e_2$ $[e_4,e_8]=e_3$ $[e_5,e_8]=e_4$, $[e_6,e_8]=e_5$, $[e_7,e_8]=e_6$.	$e_1=\sum_{i=1}^{6} X_{i,i+1}, e_2=X_{1,8}, e_3=-\frac{1}{2}X_{1,7}+$ $X_{2,8}, e_4=\frac{1}{2}X_{1,6}+X_{3,8}, e_5=-\frac{1}{2}X_{1,5}+$ $X_{4,8}, e_6=\frac{1}{2}X_{1,4}+X_{5,8}, e_7=-\frac{1}{2}X_{1,3}$ $+X_{6,8}, e_8=-\frac{1}{2}X_{1,2}-X_{2,3}-X_{3,4}-X_{4,5}$ $-X_{5,6}-X_{6,7}+X_{7,8}$.
$\mathfrak{f}_8^{17}$	$[e_4,e_7]=e_2$, $[e_4,e_8]=e_3$ $[e_5,e_6]=-e_2$, $[e_5,e_8]=e_4$, $[e_6,e_7]=e_2$, $[e_6,e_8]=e_3+e_5$ $[e_7,e_8]=e_4+e_6$.	$e_1=\sum_{i=1}^{6} X_{i,i+1}, e_2=X_{1,8}, e_3=-\frac{1}{2}X_{1,7}$ $+X_{2,8}, e_4=\frac{1}{2}X_{1,6}+X_{3,8}, e_5=-\frac{1}{2}X_{1,5}-$ $3X_{1,7}+X_{4,8}, e_6=\frac{1}{2}X_{1,4}+X_{1,6}$ $-2X_{2,7}+X_{5,8}, e_7=-\frac{1}{2}X_{1,3}+X_{2,6}-X_{3,7}$ $+X_{6,8}, e_8=-3X_{1,4}-\frac{1}{2}X_{1,2}-X_{2,3}-$ $3X_{2,5}-X_{3,4}-2X_{3,6}-X_{4,5}-$ $3X_{4,7}-X_{5,6}-X_{6,7}+X_{7,8}$.
$\mathfrak{f}_8^{18}$	$[e_4,e_7]=e_2$, $[e_4,e_8]=e_3$ $[e_5,e_6]=-e_2$, $[e_5,e_8]=e_4$, $[e_6,e_8]=e_2+e_5$ $[e_7,e_8]=e_3+e_6$.	$e_1=\sum_{i=1}^{6} X_{i,i+1}+\frac{3}{40}X_{7,8}, e_2=X_{1,8}$, $e_3=-\frac{10}{3}X_{1,7}+\frac{3}{4}X_{2,8}, e_4=\frac{4}{3}X_{1,6}-2X_{2,7}$ $+\frac{3}{5}X_{3,8}, e_5=-\frac{4}{3}X_{1,5}-2X_{3,7}+\frac{9}{20}X_{4,8}$, $e_6=\frac{4}{3}X_{1,4}-2X_{4,7}+\frac{3}{10}X_{5,8}, e_7=-\frac{4}{3}X_{1,3}$ $-2X_{5,7}+\frac{3}{20}X_{6,8}, e_8=\frac{1}{3}X_{1,2}$ $-5X_{1,5}-X_{2,3}-5X_{2,6}-X_{3,4}-5X_{3,7}$ $-X_{4,5}-\frac{3}{8}X_{4,8}-X_{5,6}-3X_{6,7}-\frac{3}{40}X_{7,8}$.
$\mathfrak{f}_8^{19}$	$[e_4,e_7]=e_2, [e_4,e_8]=e_3$ $[e_5,e_6]=-e_2, [e_5,e_8]=e_4$, $[e_6,e_7]=e_2$, $[e_6,e_8]=e_2+e_3+e_5$ $[e_7,e_8]=e_3+e_4+e_6$.	$e_1=X_{1,2}+X_{1,3}-X_{2,3}+164X_{2,5}-X_{3,4}-$ $164X_{3,6}+\frac{164}{3}X_{3,7}-164X_{4,5}+164X_{4,7}$ $-X_{5,6}-X_{6,7}-\frac{1}{328}X_{7,8}, e_2=X_{1,8}, e_3=$ $164X_{1,7}+\frac{1}{2}X_{2,8}, e_4=164X_{1,6}+164X_{1,7}-$ $\frac{1}{2}X_{3,8}, e_5=164X_{1,5}+164X_{1,6}-X_{1,8}+$ $\frac{1}{2}X_{4,8}, e_6=X_{1,4}+164X_{1,5}+164X_{1,6}-$ $\frac{164}{3}X_{1,7}-\frac{1}{3}X_{2,8}-\frac{1}{2}X_{3,8}-\frac{1}{328}X_{5,8}$, $e_7=X_{1,3}+X_{1,4}+\frac{820}{3}X_{1,6}-\frac{328}{3}X_{2,7}$ $+\frac{1}{328}X_{6,8}, e_8=X_{2,3}-\frac{164}{3}X_{2,6}+$ $X_{3,4}+328X_{3,6}+164X_{4,5}+X_{5,6}+X_{6,7}$.
$\mathfrak{f}_8^{20}$	$[e_4,e_7]=e_2, [e_4,e_8]=e_2+e_3$ $[e_5,e_6]=-e_2, [e_5,e_7]=-\frac{2}{5}e_2$, $[e_5,e_8]=e_4+\frac{3}{5}e_3$, $[e_6,e_7]=-\frac{2}{5}e_3$, $[e_6,e_8]=e_5+\frac{1}{5}e_4$, $[e_7,e_8]=e_6+\frac{1}{5}e_5$.	$e_1=X_{1,2}+X_{1,3}-X_{2,3}+\frac{3}{5}X_{2,4}-\frac{37}{25}X_{2,5}$ $-X_{3,4}+\frac{2}{5}X_{3,5}-\frac{2}{5}X_{3,6}+\frac{1}{25}X_{3,7}$ $-\frac{3}{250}X_{3,8}-X_{4,5}-\frac{2}{5}X_{4,6}-\frac{2}{25}X_{4,7}$ $-X_{5,6}-\frac{3}{5}X_{5,7}-X_{6,7}-\frac{1}{2}X_{7,8}$, $e_2=X_{1,8}, e_3=X_{1,7}-X_{1,8}+\frac{1}{2}X_{2,8}$, $e_4=X_{1,6}+\frac{3}{5}X_{1,8}-\frac{1}{2}X_{2,8}-\frac{1}{2}X_{3,8}$, $e_5=X_{1,5}+\frac{3}{5}X_{1,6}+\frac{1}{5}X_{1,7}-\frac{3}{25}X_{1,8}$ $-\frac{6}{25}X_{2,7}+\frac{3}{10}X_{2,8}+\frac{1}{5}X_{3,8}+\frac{1}{2}X_{4,8}$, $e_6=X_{1,4}+\frac{4}{5}X_{1,5}+\frac{1}{5}X_{1,6}+\frac{6}{25}X_{1,7}+$ $\frac{3}{125}X_{1,8}-\frac{3}{5}X_{2,6}-\frac{29}{25}X_{2,7}-\frac{1}{25}X_{2,8}+$ $\frac{3}{5}X_{3,7}-\frac{1}{5}X_{3,8}-\frac{1}{10}X_{4,8}-\frac{1}{2}X_{5,8}$, $e_7=X_{1,3}+X_{1,4}+\frac{7}{25}X_{1,6}+\frac{1}{5}X_{2,5}-$ $\frac{7}{5}X_{2,6}-\frac{2}{25}X_{2,7}+\frac{3}{250}X_{2,8}+\frac{4}{5}X_{3,6}+$ $\frac{1}{5}X_{4,7}+\frac{1}{2}X_{6,8}, e_8=X_{2,3}-\frac{1}{25}X_{2,6}$ $+X_{3,4}+\frac{8}{25}X_{3,6}+X_{4,5}+X_{5,6}+X_{6,7}$.

 ISBN: 978-84-614-6167-7

Now, we show two results which determine a representative for the minimal faithful unitriangular matrix representations of filiform Lie algebras in two specific cases where the derived ideal $\mathcal{C}^2(\mathfrak{g}) = [\mathfrak{g}, \mathfrak{g}]$ is abelian.

Proposition 5 *Let $\mathfrak{f}$ be a n-dimensional filiform Lie algebra verifying $z_1 = z_2 = n-1$. Then, $\bar{\mu}(\mathfrak{f}) = n$ and a natural representative of $\mathfrak{f}$ is determined by the following vectors*

$$e_1 = \sum_{i=1}^{n-2} X_{i,i+1}; \quad e_k = X_{k-1,n}, \forall\, 2 \leq k \leq n-2;$$

$$e_{n-1} = X_{1,n-1} + X_{n-2,n}; \quad e_n = X_{2,n-1} + X_{n-1,n}.$$

Proof: To prove this result, it suffices to apply our algorithmic procedure taking into consideration that non-zero brackets for the case $z_1 = z_2 = n-1$ are the following

$$[e_1, e_h] = e_{h-1}, \forall\, 3 \leq h \leq n; \quad [e_{n-1}, e_n] = e_2.$$

□

Proposition 6 *Let $\mathfrak{f}$ be a n-dimensional filiform Lie algebra verifying $z_1 = n-2$ and $z_2 = n-1$. Then, $\bar{\mu}(\mathfrak{f}) = n$ and a natural representative of $\mathfrak{f}$ is determined by the following vectors*

$$e_1 = \sum_{i=1}^{n-2} X_{i,i+1}; \quad e_k = X_{k-1,n}, \forall\, 2 \leq k \leq n-2; \quad e_{n-2} = X_{1,n-1} + X_{n-3,n};$$

$$e_{n-1} = X_{2,n-1} + X_{n-2,n}; \quad e_n = X_{3,n-1} + X_{n-1,n}.$$

Proof: To prove this result, it is sufficient to apply our algorithmic procedure taking into consideration that the non-zero brackets for the case $z_1 = n-2$ and $z_2 = n-1$ are the following

$$[e_1, e_h] = e_{h-1}, \forall\, 3 \leq h \leq n; \quad [e_{n-2}, e_n] = e_2; \quad [e_{n-1}, e_n] = e_3.$$

□

5 Conclusions

At present, researchers on Lie and Representation Theories need to deal with examples of Lie algebras in higher dimensions. The representation of these algebras is a difficult task due to computing time and memory used. Besides, solvable, nilpotent and filiform complex Lie algebras are classified only up to dimension 6, 7 and 12, respectively (as can be seen in [3]).

In this paper we have shown an algorithmic method to obtain a minimal faithful unitriangular matrix representation of a given finite-dimensional filiform Lie algebra, introducing its law. In our opinion, this constitutes a new little step forward to tackle the classification problem of these algebras and the obtainment of their representations.

 ISBN: 978-84-614-6167-7

6 Acknowledgments

The authors want to thank Professor Francisco J. Castro, from the University of Seville, for some useful suggestions which helped to improve this paper.
This work has been partially supported by MTM2010-19336 and FEDER.

References

[1] J.C. Benjumea, F.J. Echarte, J. Núñez and A.F. Tenorio, *A method to obtain the Lie group associated with a nilpotent Lie algebra*, Comput. Math. Appl. **51** (2006), 1493–1506.

[2] J.C. Benjumea, J. Núñez and A.F. Tenorio, *Minimal linear representations of the low-dimensional nilpotent Lie algebras*, Math. Scand. **102** (2008), 17–26.

[3] L. Boza, E.M. Fedriani and J. Núñez, *A new method for classifying complex filiform Lie algebras*, Appl. Math. Comput. **121** (2001), 169–175.

[4] D. Burde, *On a refinement of Ado's Theorem*, Arch. Math.(Basel) **70** (1998), 118–127.

[5] F.J. Echarte, J. Núñez and F. Ramírez, *Relations among invariants of complex filiform Lie algebras*, Appl. Math. Comput. **147** (2004), 365–376.

[6] F.J. Echarte, J. Núñez and F. Ramírez, *Description of some families of filiform Lie algebras*, Houston J. Math. **34** (2008), 19–32.

[7] W. Fulton and J. Harris, *Representation theory: a first course.* Springer-Verlag, New York, 1991.

[8] R. Ghanam, I. Strugar and G. Thompson, *Matrix representations for low dimensional Lie algebras*, Extracta Math. **20** (2005), 151-184.

[9] V.S. Varadarajan, *Lie Groups, Lie Algebras and Their Representations*, Springer, 1984.

[10] 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.

 ISBN: 978-84-614-6167-7

Proceedings of the 11th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2011
26–30 June 2011.

A uniformly convergent hybrid scheme for one dimensional time-dependent reaction–diffusion problems

C. Clavero[1] and J.L. Gracia[1]

[1] *Department of Applied Mathematics, University of Zaragoza*

emails: clavero@unizar.es, jlgracia@unizar.es

Abstract

In this work we consider the numerical approximation of the solution of a time-dependent initial-boundary value problem of reaction-diffusion type, for which the diffusion parameter can be arbitrary small. We construct a finite difference scheme combining the classical implicit Euler method to discretize in time, which is defined on a uniform mesh, together with a hybrid finite difference scheme based on a HODIE compact fourth order method and the standard central finite difference approximation to discretize in space, which is defined on a mesh of Vulanovic type condensing the grid points in the boundary layer regions. We prove that the fully discrete method is uniformly convergent with respect to the singular perturbation parameter having first order in time and almost fourth order in space. The proof of the uniform convergence is based on splitting the contribution to the error from the time and the space discretization, and it uses the asymptotic behavior of the solution of the semidiscrete problems resulting after the time discretization. We show some numerical results obtained for a test problem, which confirm in practice the good results and the efficiency of the method.

Key words: time dependent reaction-diffusion problems, hybrid finite difference scheme, Vulanovic mesh, uniform convergence
MSC 2000: 35K20, 65N06, 65N12

1 Introduction

In this work we are interested in the numerical approximation of the following time-dependent initial-boundary value problem of reaction-diffusion type

$$\begin{cases} u_t + L_{x,\varepsilon} u = f(x,t), \quad (x,t) \in Q = \Omega \times (0,T] \equiv (0,1) \times (0,T], \\ u(x,0) = 0, \ x \in \overline{\Omega}, \ u(0,t) = u(1,t) = 0, \ t \in (0,T], \end{cases} \tag{1}$$

 ISBN: 978-84-614-6167-7

where the spatial differential operator is given by $L_{x,\varepsilon}u \equiv -\varepsilon u_{xx} + b(x,t)u$. We assume that $0 < \varepsilon \leq 1$ and it can be very small, the reaction term satisfies $b(x,t) \geq \beta > 0$ for all $(x,t) \in \overline{Q}$, and the solution is smooth enough to complete the forthcoming analysis (see [7]).

In general, the solution of (1) has a boundary layer at $x = 0, 1$ of width $O(\sqrt{\varepsilon}|\ln \varepsilon|)$ (see [5, 9]) and applying the classical theory (see [7]) for partial differential equations it is possible to prove that the solution of (1) satisfies the crude bounds

$$|u^{(k,m)}(x,t)| \leq C\varepsilon^{-k/2}, \quad 0 \leq k + 2m \leq 8. \tag{2}$$

Nevertheless, these bounds do not reflect the presence of boundary layers in the solution; so, it is necessary to have a more precise information of the asymptotic behavior, with respect to ε, of the solution and its derivatives. Following [6] it is possible to prove that the solution of (1) satisfies

$$|u^{(k,m)}(x,t)| \leq C(1 + \varepsilon^{-k/2}B_\varepsilon(x)), \quad 0 \leq k + 2m \leq 6, \tag{3}$$

where $B_\varepsilon(x) = e^{-\sqrt{\beta/\varepsilon}x} + e^{-\sqrt{\beta/\varepsilon}(1-x)}$. These estimates clearly show the presence of two boundary layers. Moreover, the solution can decomposed into a regular and a singular component, $u = \phi + \psi$, where the regular component ϕ satisfies

$$|\phi^{(k,m)}(x,t)| \leq C(1 + \varepsilon^{2-k/2}), \quad 0 \leq k + 2m \leq 6, \tag{4}$$

and the singular component ψ satisfies

$$|\psi^{(k,m)}(x,t)| \leq C\varepsilon^{-k/2}B_\varepsilon(x), \quad 0 \leq k + 2m \leq 6. \tag{5}$$

In this work we follows the ideas of [1, 2] to deduce the uniform convergence of the numerical method, which is based on two steps, providing one after the other the contribution to the error of the time and space discretizations. In previous papers the analysis of the uniform convergence of the space discretization uses some auxiliary problems arising in the definition of the method after the time discretization. Nevertheless, in [3] the uniform convergence was proved without the use of any auxiliary problems. The analysis requires to know the asymptotic behavior of the exact solution of the semidiscrete problems resulting from the time discretization, which is based on an inductive argument and the well–known behavior of steady singularly perturbed reaction-diffusion problems (see [9]). In [3] the implicit Euler method was considered to approximate the time variable and the central finite difference approximation for the spatial variable. Here we show that the idea can be extended to a HODIE compact fourth order scheme [8], proving that the fully discrete method is first order uniformly convergent in time and almost fourth order convergent in space.

Henceforth, C denotes a generic positive constant independent of the diffusion parameter ε and also of the discretization parameters N and M.

 ISBN: 978-84-614-6167-7

2 The fully discrete method: uniform convergence

The first step to construct the fully discrete method is the time discretization. We use the backward Euler method, which on the uniform mesh

$$\overline{\omega}^M = \{t_k = k\tau,\ 0 \le k \le M, \tau = T/M\},$$

is given by

$$\left\{ \begin{array}{l} z^0 = 0, \\ \left\{ \begin{array}{l} (I + \tau L_{x,\varepsilon}) z(x,t_n) = \tau f(x,t_n) + z(x,t_{n-1}), \quad 1 \le n \le M, \\ z(0,t_n) = z(1,t_n) = 0. \end{array} \right. \end{array} \right. \tag{6}$$

In [4] it was proved that the global error associated to the method (6), defined by $e_n = u(x,t_n) - z(x,t_n)$, satisfies

$$\|e_n\|_{\overline{\Omega}} \le C\tau, \quad 1 \le n \le M, \tag{7}$$

and therefore the Euler method is a first order uniformly convergent method.

For the analysis of the uniform convergence of the space discretization we need to know the asymptotic behavior of the exact solution $z(x,t_n)$ of (6). Following similar ideas to these ones given in [3] it is possible to prove the following result.

Theorem 1 *Let $z(x,t_n)$ be the solution of problem (6) at the time level n. Then, it holds*

$$|z^{(k)}(x,t_n)| \le C(1 + \varepsilon^{-k/2} B_\varepsilon(x)), \quad 0 \le k \le 6, \quad 1 \le n \le M. \tag{8}$$

Nevertheless these bounds are not sufficient and it is necessary to decompose $z(x,t_n)$ in a similar way to the continuous problem. Then, we have $z(x,t_n) = v(x,t_n) + w(x,t_n),\ 0 \le n \le M$, where the regular component v is the solution of the problem

$$\left\{ \begin{array}{l} v(x,t_0) = 0, \\ \dfrac{v(x,t_n) - v(x,t_{n-1})}{\tau} + L_{x,\varepsilon} v(x,t_n) = f(x,t_n), \quad x \in [0,1], \quad 1 \le n \le M, \end{array} \right.$$

with appropriate values in the boundary. The singular component w is the solution of the problem

$$\left\{ \begin{array}{ll} w(x,t_0) = 0, & \\ \dfrac{w(x,t_n) - w(x,t_{n-1})}{\tau} + L_{x,\varepsilon} w(x,t_n) = 0, & x \in [0,1], \\ w(0,t_n) = -v(0,t_n),\ w(1,t_n) = -v(1,t_n), \quad 1 \le n \le M. & \end{array} \right. \tag{9}$$

Lemma 1 *Let assume that $b, f \in \mathbf{C}^{(8,4)}(\overline{Q})$; then, the regular component v satisfies*

$$|v^{(k)}(x,t_n)| \le C(1 + \varepsilon^{2-k/2}), \quad 0 \le k \le 6, \quad 1 \le n \le M.$$

Moreover, the singular component w satisfies

$$|w^{(k)}(x,t_n)| \le C\varepsilon^{-k/2} B_\varepsilon(x), \quad 0 \le k \le 6, \quad 1 \le n \le M.$$

 ISBN: 978-84-614-6167-7

To discretize (6) in space, we use a hybrid finite difference scheme which combines the central difference and a HODIE type operator constructed on a special mesh of Vulanovic type (see [10]), which condenses the grid points in the boundary layer regions, which is constructed as follows. Let $N = 4k$ be, where k is a positive integer; then, we divide [0,1] into three intervals $[0, \sigma], [\sigma, 1-\sigma], [1-\sigma, 1]$, where σ is

$$\sigma = \min\{1/4, 4\sqrt{\varepsilon/\beta}\ln N\}. \tag{10}$$

The grid points are defined by $x_j = \aleph(j/N),\ j = 0, 1, \ldots, N/2$, with $\aleph \in C^2[0, 1/2]$ and

$$\aleph(z) = \begin{cases} 4\sigma/N, & z \in [0, 1/4], \\ p(z-1/4)^3 + 4\sigma(z-1/4) + \sigma, & z \in [1/4, 1/2]. \end{cases} \tag{11}$$

The coefficient p is such that $\aleph(1/2) = 1/2$ and the mesh is symmetric with respect to the point $x = 1/2$. Note that in $[0, \sigma]$ and $[1-\sigma, 1]$ the mesh is uniform; otherwise, in $[\sigma, 1-\sigma]$ it is nonuniform and the step sizes satisfy

$$|h_{j+1} - h_j| \le CN^{-2}, \quad j = N/4, \ldots, 3N/4. \tag{12}$$

Note that if $\sigma = 1/4$ the mesh is uniform in $[0, 1]$ and therefore a classical analysis could be made; so, here we are only interested in the case $\sigma = 4\sqrt{\varepsilon/\beta}\ln N$.

On this mesh, the finite difference scheme is given by (see [2]),

$$L_\varepsilon^{N,M} U_j^n \equiv \Gamma[\frac{U_j^n - U_j^{n-1}}{\tau}] - \varepsilon\delta_x^2 U_j^n + \Gamma[b_j^n U_j^n] = \Gamma[f_j^n], \quad 0 < j < N, \quad U_0^n = U_N^n = 0, \tag{13}$$

where

$$\Gamma[v_j^n] = \begin{cases} v_j^n, \quad \text{if } N/4 \le j \le 3N/4, \text{ and } h_{max}^2\left(\|b\|_\infty + \dfrac{1}{\tau}\right) \ge 6\varepsilon, \\ \dfrac{1-\gamma_j^-}{6} v_{j-1}^n + \dfrac{4+\gamma_j^- + \gamma_j^+}{6} v_j^n + \dfrac{1-\gamma_j^+}{6} v_{j+1}^n, \text{ otherwise.} \end{cases}$$

with

$$\gamma_j^- = \frac{h_{j+1}^2}{2h_j\hbar_j}, \quad \gamma_j^+ = \frac{h_j^2}{2h_{j+1}\hbar_j},$$

and

$$\delta_x^2 U_j^n := \frac{1}{\hbar_j}\left(\frac{U_{j+1}^n - U_j^n}{h_{j+1}} - \frac{U_j^n - U_{j-1}^n}{h_j}\right),$$
$$h_j = x_j - x_{j-1},\ j = 1, \cdots, N,\ \hbar_j = \frac{h_j + h_{j+1}}{2},\ j = 1, \cdots, N-1.$$

Lemma 2 *Let $N \ge N_0$ be, where $N_0 > 0$ is a positive integer independent of ε such that*

$$64(\|b\|_\infty + \frac{1}{\tau}) < \frac{3\beta N_0^2}{\ln^2 N_0}. \tag{14}$$

Then, the scheme (13) is of positive type, it satisfies a discrete maximum principle and it is ε-uniform stable in the maximum norm.

 ISBN: 978-84-614-6167-7

Proof. The proof is trivial using that the matrix associated with the discrete operator is an M-matrix (see [9]).

Theorem 2 *Let assume that $N \geq N_0$ and the space and time discretization parameters N and τ satisfy the following relation*

$$\frac{1}{\tau} \leq \ln^4 N. \tag{15}$$

Then, the error associated with the hybrid finite difference scheme (13) satisfies

$$\|[U - z]_j^n\|_\infty \leq C(N^{-1} \ln N)^4. \tag{16}$$

Then, from (7) and (16) it follows

$$\|[u - U]_j^n\|_\infty \leq C\left(\tau + (N^{-1} \ln N)^4\right), \tag{17}$$

proving the uniform convergence of first order in time and the almost fourth order in space.

3 Numerical experiments

In this section we show the numerical results obtained for the test problem

$$\begin{cases} u_t - \varepsilon u_{xx} + (1 + xe^{-t})u = f(x,t), \quad (x,t) \in (0,1) \times (0,1], \\ u(x,0) = 0, \ x \in \overline{\Omega}, \ u(0,t) = 0, \ u(1,t) = 0, \ t \in (0,1], \end{cases} \tag{18}$$

where f is taken such that the exact solution is

$$u(x,t) = t\left(\frac{e^{-x/\sqrt{\varepsilon}} + e^{-(1-x)/\sqrt{\varepsilon}}}{1 + e^{-1/\sqrt{\varepsilon}}} - \cos^2(\pi x)\right).$$

Figure 1 shows the numerical solution for $\varepsilon = 10^{-6}$; from it we clearly see the boundary layers at both edges $x = 0$ and $x = 1$.

From the error $E_{j,n}^{\varepsilon,N,M} = |U_{j,n}^{\varepsilon,N,M} - u(x_j, t_n)|$, at each grid point, we compute the maximum global errors and the numerical orders of convergence by $E^{\varepsilon,N,M} = \max\limits_{j,n} E_{j,n}^{\varepsilon,N,M}$, $p = \log\left(E^{\varepsilon,N,M}/E^{\varepsilon,2N,2M}\right)/\log 2$.

We denote the ε-uniform errors and the ε-uniform orders of convergence by $E^{N,M} = \max_{\varepsilon \in S_\varepsilon} E^{\varepsilon,N,M}$, $p_{uni} = \log\left(E^{N,N}/E^{2N,2M}\right)/\log 2$, where $S_\varepsilon = \{2^0, 2^{-2}, \ldots, 2^{-30}\}$ has been chosen in the numerical experiments performed.

Table 1 displays the maximum uniform errors and the computed uniform order of convergence of method (13) on the Vulanovic mesh. From it we can deduce the almost fourth order convergence of the method in agreement with (17).

Finally, we compare the maximum errors associated to method (13) with those ones for the basic method given in [3], which combines the implicit Euler method in time with the central difference approximation in space. Table 2 displays the results obtained with that method and we see that the maximum errors are considerably larger than in Table 1. So, we can conclude that the hybrid method (13) is more efficient than the method used in [3].

 ISBN: 978-84-614-6167-7

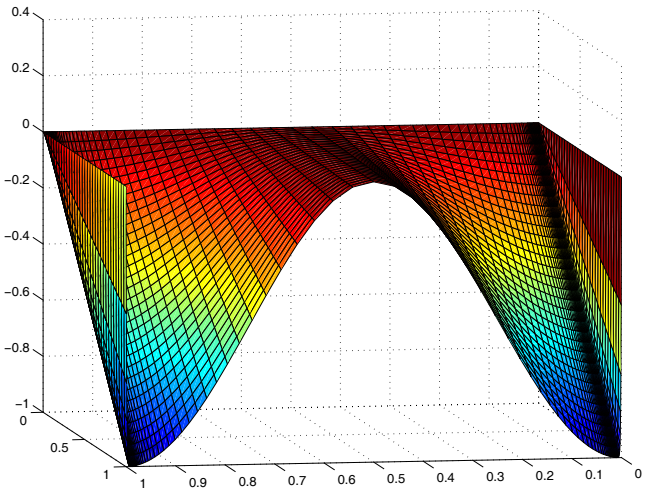


Figure 1: Solution of problem (18) for $\varepsilon = 10^{-6}$

Table 1: Maximum uniform errors and uniform orders of convergence for method (13)

	N=64 M= 8	N=128 M= 16	N=256 M= 32	N=512 M= 64	N=1024 M= 128	N=2048 M= 256
$E^{N,M}$	0.521E-3	0.615E-4	0.672E-5	0.678E-6	0.645E-7	0.590E-8
p_{uni}	3.084	3.193	3.311	3.392	3.450	

Acknowledgements

This research was partially supported by the project MEC/FEDER MTM 2010-16917 and the Diputación General de Aragón.

References

[1] B. Bujanda, C. Clavero,J.L. Gracia, J.C. Jorge, *A high order uniformly convergent alternating direction scheme for time dependent reaction-diffusion singularly perturbed problems*, Num. Math. **107** (2007) 1–25.

[2] C. Clavero, J.L. Gracia, *High order methods for elliptic and time dependent reaction-diffusion singularly perturbed problems*, Appl. Math. Comp. **168** (2005) 1109–1127.

 ISBN: 978-84-614-6167-7

Table 2: Maximum uniform errors and uniform orders of convergence using Euler and central differences

	N=64 M= 8	N=128 M= 16	N=256 M= 32	N=512 M= 64	N=1024 M= 128	N=2048 M= 256
$E^{N,M}$	0.239E-2	0.849E-3	0.280E-3	0.892E-4	0.275E-4	0.832E-5
p_{uni}	1.495	1.600	1.651	1.697	1.725	

[3] C. Clavero, J.L. Gracia, *On the uniform convergence of a finite difference scheme for time dependent singularly perturbed reaction–diffusion problems*, Appl. Math. Comp. **216** (2010) 1478–1488.

[4] C. Clavero, J.C. Jorge, F. Lisbona, G.I. Shishkin, *An alternating direction scheme on a nonuniform mesh for reaction–diffusion parabolic problem*, IMA J. Numer. Anal. **20** (2000) 263–280.

[5] P.A. Farrell, A.F. Hegarty, J.J.H. Miller, E. O'Riordan, E., G.I. Shishkin, *Robust computational techniques for boundary layers*, Chapman & Hall (2000).

[6] P.W. Hemker, G.I. Shishkin, L.P. Shiskina, *ε-uniform schemes with high-order time accuracy for parabolic singular perurbation problems*, IMA J. Numer. Anal. **20** (2000) 99–121.

[7] O.A. Ladyzhenskaya, V.A. Solonnikov, N.N. Ural'tseva, *Linear and quasilinear equations of parabolic type*, Transactions of Mathematical Monographs, **23**, American Mathematical Society, (1968).

[8] T. Linss, *Robust convergence of a compact fourth order finite difference scheme for reaction–diffusion problems*, Numer. Math. **(111)** (2008) 239–249.

[9] H.-G. Roos, M. Stynes, L. Tobiska, *Robust numerical methods for singularly perturbed differential equations*, Springer Series in Computational Mathematics **24**, Springer-Verlag, Berlin, 2008.

[10] R. Vulanovic, *An almost sixth-order finite-difference method for semilinear singular perurbation problems*, Comput. Meth. Appl. Math. **4** (2004) 368–383.

 ISBN: 978-84-614-6167-7

Proceedings of the 11th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2011
26–30 June 2011.

Construction of bent functions of n variables from a basis of $\mathbb{F}_2^n$

Joan-Josep Climent[1], **Francisco J. García**[2] **and Verónica Requena**[1]

[1] *Departament d'Estadística i Investigació Operativa, Universitat d'Alacant*

[2] *Departament de Fonaments de l'Anàlisi Econòmica, Universitat d'Alacant*

emails: jcliment@ua.es, francisco.garcia@ua.es, vrequena@ua.es

Abstract

In this paper we use a basis of $\mathbb{F}_2^n$ (with n even) to establish an iterative construction of some sets of $\mathbb{F}_2^n$ which are the support of bent functions of n variables.

Key words: Boolean function, bent function, balanced function, support

1 Introduction

A Boolean function maps a number of input bits into a single bit. Boolean functions are widely used in different types of cryptographic applications such as block ciphers, stream ciphers and hash functions [3, 4, 13], in coding theory [2, 7], among others. A cryptographic function should have high nonlinearity in order to prevent attacks based on linear approximation [1, 6, 11, 14]. The functions achieving the maximal possible nonlinearity possess the best resistance to the linear attack and they are called bent functions [16, 18]. Bent functions have been the subject of some interest in coding theory [9, 10], in logic synthesis [20] and in cryptography [13].

A general method for generating all bent functions is not known to exist yet, except for some particular cases ($n = 2, 4, 6$); the classification for $n \geq 8$ is still an open problem, although recently Langevin and Leander [8] provided the number of bent functions of 8 variables. The origin of bent functions goes back to a theoretical article of McFarland [12] on sets of finite differences in finite non-cyclic groups. One year after, Dillon [5] systematized and extended the ideas of McFarland, proving a great quantity of properties. The name *bent* for these functions is due to Rothaus [15].

 ISBN: 978-84-614-6167-7

The rest of the paper is organized as follows. Firstly, in Section 2 we introduce some basic definitions and notations that are used hereafter. In Section 3, starting with a bases of $\mathbb{F}_2^{2k}$ we introduce some sets of $\mathbb{F}_2^{2k}$ with the property that they are the supports of bent functions of $2k$ variables. Finally, in Section 4 we present some conclusions.

2 Preliminary results

We denote by $\mathbb{F}_2$ the Galois field of two elements, 0 and 1, with the addition (denoted by $\oplus$) and the multiplication (denoted by juxtaposition). For any positive integer n, it is well-known that $\mathbb{F}_2^n$ is a linear space of dimension n over $\mathbb{F}_2$ with the usual addition (denoted also by $\oplus$). If for $i = 0, 1, 2, \ldots, 2^n - 1$, we denote by $\boldsymbol{i}$ the binary expansion of i of n digits, then $\mathbb{F}_2^n = \{\boldsymbol{i} \mid 0 \leq i \leq 2^n - 1\}$. Furthermore, we denote by $\mathrm{Span}\{\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_k\}$ the linear subspace of $\mathbb{F}_2^n$ generated by $\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_k \in \mathbb{F}_2^n$. Moreover, if $S \subseteq \mathbb{F}_2^n$ and $\boldsymbol{a} \oplus S = \{\boldsymbol{a} \oplus \boldsymbol{u} \mid \boldsymbol{u} \in S\}$ for $\boldsymbol{a} \in \mathbb{F}_2^n$, then

$$\mathrm{card}\,(\boldsymbol{a} \oplus S) = \mathrm{card}\,(S)\,.$$

A **Boolean function** of n variables is a map $f : \mathbb{F}_2^n \longrightarrow \mathbb{F}_2$. The set of all Boolean functions of n variables is denoted by $\mathcal{B}_n$; it is well known that $\mathcal{B}_n$, with the usual addition of functions (that we also denote by $\oplus$), is a linear space of dimension 2^n over $\mathbb{F}_2$. If $f \in \mathcal{B}_n$, we call **support** of f, denoted by $\mathrm{Supp}\,(f)$, the set of vectors of $\mathbb{F}_2^n$ whose image by f is 1, that is,

$$\mathrm{Supp}\,(f) = \{\boldsymbol{a} \in \mathbb{F}_2^n \mid f(\boldsymbol{a}) = 1\}\,.$$

If $f \in \mathcal{B}_n$, we call **weight** of f, and we write $w(f)$, the number of elements of $\mathrm{Supp}\,(f)$, i.e., $w(f) = \mathrm{card}\,(\mathrm{Supp}\,(f))$. We said that f is **balanced** if $w(f) = 2^{n-1}$.

The following result gives us a characterization of a bent function.

Theorem 1 ([17, 18]): *Let $f(\boldsymbol{x})$ be a Boolean function of n variables with n even. Then $f(\boldsymbol{x})$ is a bent function if and only if the Boolean function $f(\boldsymbol{x}) \oplus f(\boldsymbol{a} \oplus \boldsymbol{x})$ is balanced for all $\boldsymbol{a} \in \mathbb{F}_2^n \setminus \{\boldsymbol{0}\}$.*

As a consequence of the previous result we have the following characterization of a bent function that we will use in the rest of the paper.

Corollary 1: *Assume that $S \subseteq \mathbb{F}_2^n$ with n even. Then S is the support of a bent function of n variables if and only if $S\,\Delta(\boldsymbol{a} \oplus S)$ is the support of a balanced function of n variables for all $\boldsymbol{a} \in \mathbb{F}_2^n \setminus \{\boldsymbol{0}\}$, where Δ is the symmetric difference of sets.*

Proof: Let $f \in \mathcal{B}_n$. If $S = \mathrm{Supp}\,(f)$, then $S\,\Delta(\boldsymbol{a} \oplus S)$ is the support of the Boolean function $f(\boldsymbol{x}) \oplus f(\boldsymbol{a} \oplus \boldsymbol{x})$. So the result follows by Theorem 1. □

 ISBN: 978-84-614-6167-7

3 Main results

From now on, we assume that $n = 2k$ and that $\mathcal{U} = \{\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_{2k-1}, \boldsymbol{u}_{2k}\}$ is a basis of $\mathbb{F}_2^n$. For $i = 1, 2, \ldots, k$, consider the linear subspaces

$$G_i = \operatorname{Span}\{\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_{2i-1}, \boldsymbol{u}_{2i}\} \quad \text{and} \quad H_i = \operatorname{Span}\{\boldsymbol{u}_{2i-1}, \boldsymbol{u}_{2i}\}$$

of $\mathbb{F}_2^n$. Clearly, $\dim G_i = 2i$ and $\dim H_i = 2$. Furthermore, if we consider $G_0 = \{\mathbf{0}\}$, then

$$G_i = G_{i-1} \oplus H_i \quad \text{and} \quad G_{i-1} \cap H_i = \{\mathbf{0}\},$$

and therefore, G_i is the direct sum of G_{i-1} and H_i. For convenience in the notation, we refer the elements of H_i, for $i = 1, 2, \ldots, k$, as

$$\boldsymbol{a}_0^{(i)} = \mathbf{0}, \quad \boldsymbol{a}_1^{(i)} = \boldsymbol{u}_{2i-1}, \quad \boldsymbol{a}_2^{(i)} = \boldsymbol{u}_{2i} \quad \text{and} \quad \boldsymbol{a}_3^{(i)} = \boldsymbol{u}_{2i-1} \oplus \boldsymbol{u}_{2i}.$$

For $p \in \{0, 1, 2, 3\}$ consider the sets

$$B(p) = \left\{\boldsymbol{a}_p^{(1)}\right\} \quad \text{and} \quad \widehat{B}(p) = \bigcup_{\substack{q=0 \\ q \neq p}}^{3} \left\{\boldsymbol{a}_q^{(1)}\right\}.$$

It is evident that

$$G_1 = B(p) \cup \widehat{B}(p) \quad \text{and} \quad B(p) \cap \widehat{B}(p) = \emptyset. \tag{1}$$

Furthermore, if $r, s \in \{0, 1, 2, 3\}$ with $r \neq s$, then $B(r) \neq B(s)$.

Now, let $(p_1, p_2, \ldots, p_{i-1}, p_i) \in \{0, 1, 2, 3\}^i$ and assume that we have defined the sets $B(p_1, p_2, \ldots, p_{i-1})$ and $\widehat{B}(p_1, p_2, \ldots, p_{i-1})$. Then, we define

$$\begin{aligned} &B(p_1, p_2, \ldots, p_{i-1}, p_i) \\ &= \left(\boldsymbol{a}_{p_i}^{(i)} \oplus \widehat{B}(p_1, p_2, \ldots, p_{i-1})\right) \cup \bigcup_{\substack{q=0 \\ q \neq p_i}}^{3} \left(\boldsymbol{a}_q^{(i)} \oplus B(p_1, p_2, \ldots, p_{i-1})\right), \end{aligned} \tag{2}$$

$$\begin{aligned} &\widehat{B}(p_1, p_2, \ldots, p_{i-1}, p_i) \\ &= \left(\boldsymbol{a}_{p_i}^{(i)} \oplus B(p_1, p_2, \ldots, p_{i-1})\right) \cup \bigcup_{\substack{q=0 \\ q \neq p_i}}^{3} \left(\boldsymbol{a}_q^{(i)} \oplus \widehat{B}(p_1, p_2, \ldots, p_{i-1})\right). \end{aligned} \tag{3}$$

Our goal is to prove that the sets $B(p_1, p_2, \ldots, p_k)$ and $\widehat{B}(p_1, p_2, \ldots, p_k)$, for all $(p_1, p_2, \ldots, p_k) \in \{0, 1, 2, 3\}^k$, are the supports of two bent functions of $2k$ variables, such that one is the complementary function of the other. However, we need to introduce beforehand a technical lemma which will simplify the proof of the above mentioned result.

 ISBN: 978-84-614-6167-7

Lemma 1: *For $i = 1, 2, \ldots, k$, if $(p_1, p_2, \ldots, p_i) \in \{0, 1, 2, 3\}^i$ and $\boldsymbol{u} \in G_i \setminus \{\mathbf{0}\}$, then:*

1. $G_i = B(p_1, p_2, \ldots, p_i) \cup \widehat{B}(p_1, p_2, \ldots, p_i)$
2. $B(p_1, p_2, \ldots, p_i) \cap \widehat{B}(p_1, p_2, \ldots, p_i) = \emptyset$
3. $\operatorname{card}\left(B(p_1, p_2, \ldots, p_i)\right) = 2^{2i-1} - 2^{i-1}$
4. $\operatorname{card}\left(\widehat{B}(p_1, p_2, \ldots, p_i)\right) = 2^{2i-1} + 2^{i-1}$
5. $\operatorname{card}\left(B(p_1, p_2, \ldots, p_i) \cap \left(\boldsymbol{u} \oplus B(p_1, p_2, \ldots, p_i)\right)\right) = 2^{2i-2} - 2^{i-1}$,
6. $\operatorname{card}\left(\widehat{B}(p_1, p_2, \ldots, p_i) \cap \left(\boldsymbol{u} \oplus \widehat{B}(p_1, p_2, \ldots, p_i)\right)\right) = 2^{2i-2} + 2^{i-1}$,
7. $\operatorname{card}\left(B(p_1, p_2, \ldots, p_i) \cap \left(\boldsymbol{u} \oplus \widehat{B}(p_1, p_2, \ldots, p_i)\right)\right) = 2^{2i-2}$,
8. $\operatorname{card}\left(\widehat{B}(p_1, p_2, \ldots, p_i) \cap \left(\boldsymbol{u} \oplus B(p_1, p_2, \ldots, p_i)\right)\right) = 2^{2i-2}$.

We are now able to prove that the sets $B(p_1, p_2, \ldots, p_k)$ and $\widehat{B}(p_1, p_2, \ldots, p_k)$ are the supports of a bent function of $2k$ variables and its complementary, respectively.

Theorem 2: *For all $(p_1, p_2, \ldots, p_k) \in \{0, 1, 2, 3\}^k$ the sets*

$$B(p_1, p_2, \ldots, p_k) \quad \text{and} \quad \widehat{B}(p_1, p_2, \ldots, p_k)$$

are the supports of two bent functions of $2k$ variables so that one is the complementary function of the other.

Proof: Assume that $\boldsymbol{u} \in \mathbb{F}_2^{2k} \setminus \{\mathbf{0}\}$. Since $\mathbb{F}_2^{2k} = G_k$, by Lemma 1 we have that

$$\begin{aligned}
&\operatorname{card}\left(B(p_1, p_2, \ldots, p_k) \,\Delta\, (\boldsymbol{u} \oplus B(p_1, p_2, \ldots, p_k))\right) \\
&\quad = \operatorname{card}\left(B(p_1, p_2, \ldots, p_k)\right) + \operatorname{card}\left(\boldsymbol{u} \oplus B(p_1, p_2, \ldots, p_k)\right) \\
&\qquad\quad - 2\operatorname{card}\left(B(p_1, p_2, \ldots, p_k) \cap (\boldsymbol{u} \oplus B(p_1, p_2, \ldots, p_k))\right) \\
&\quad = 2^{2k-1} - 2^{k-1} + 2^{2k-1} - 2^{k-1} - 2\left(2^{2k-2} - 2^{k-1}\right) = 2^{2k-1},
\end{aligned}$$

and therefore, $B(p_1, p_2, \ldots, p_k)\,\Delta(\boldsymbol{u} \oplus B(p_1, p_2, \ldots, p_k))$ is the support of a balanced function of $2k$ variables. So, by Corollary 1, we have that $B(p_1, p_2, \ldots, p_k)$ is the support of a bent function $f(\boldsymbol{x})$ of $2k$ variables.

Moreover, from parts 1 and 2 of Lemma 1 we have that $\widehat{B}(p_1, p_2, \ldots, p_k) = G_k \setminus B(p_1, p_2, \ldots, p_k)$ and, consequently, the set $\widehat{B}(p_1, p_2, \ldots, p_k)$ is the support of the complementary function $1 \oplus f(\boldsymbol{x})$. □

 ISBN: 978-84-614-6167-7

To count the number of bent functions provided by Theorem 2, we need the following result.

Theorem 3: *For $i = 1, 2, \ldots, k$, let $(p_1, p_2, \ldots, p_i), (q_1, q_2, \ldots, q_i) \in \{0, 1, 2, 3\}^i$. If $p_1 = q_1$, $p_2 = q_2$, $\ldots$, $p_l = q_l$, but $p_l \neq q_l$ for some $l \in \{1, 2, \ldots, i-1\}$, then*

$$B(p_1, p_2, \ldots, p_l, p_{l+1}, p_{l+2}, \ldots p_{l+m}) \neq B(p_1, p_2, \ldots, p_l, q_{l+1}, q_{l+2}, \ldots q_{l+m}),$$
$$\widehat{B}(p_1, p_2, \ldots, p_l, p_{l+1}, p_{l+2}, \ldots p_{l+m}) \neq \widehat{B}(p_1, p_2, \ldots, p_l, q_{l+1}, q_{l+2}, \ldots q_{l+m})$$

for $m = 1, 2, \ldots, i - l$.

Thus, as a consequence of this result we have that the number of bent functions of $2k$ variables that we can construct using a basis $\mathcal{U} = \{\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_{2k-1}, \boldsymbol{u}_{2k}\}$ is

$$2 \operatorname{card}\left(\{0, 1, 2, 3\}^k\right) = 2^{2k+1}.$$

Since the number of different basis in $\mathbb{F}_2^{2k}$ (see [19, page 46]) is $\prod_{i=0}^{2k-1}(2^{2k} - 2^i)$ we can construct

$$2^{2k+1} \prod_{i=0}^{2k-1} (2^{2k} - 2^i)$$

bent functions of $2k$ variables. For example, for $k = 2$, we can construct 645 120 bent functions. Nevertheless, the number of different bent functions of 4 variables is 896.

So there are different bases that provide the same bent functions. For example, it is not difficult to see that the bases $\mathcal{U} = \{\mathbf{1}, \mathbf{2}, \mathbf{4}, \mathbf{8}\}$ and $\mathcal{V} = \{\mathbf{6}, \mathbf{7}, \mathbf{9}, \mathbf{13}\}$ of $\mathbb{F}_2^4$ provide both the supports of the same bent functions of 2 variables.

4 Conclusions

In this paper we use a basis of $\mathbb{F}_2^{2k}$ to construct 2^{2k+1} sets in $\mathbb{F}_2^{2k}$ that are the supports of bent functions. Half of the bent functions obtained are the complementary functions of the other half. Since different bases can produce the same bent functions, we need to establish under what conditions two different bases provide the same bent functions.

Acknowledgements

This work was partially supported by Spanish grants MTM2008-06674-C02-01 of the Ministerio de Ciencia e Innovación of the Gobierno de España and ACOMP/2011/005 of the Generalitat Valenciana.

 ISBN: 978-84-614-6167-7

References

[1] C. M. Adams. Constructing symmetric ciphers using the CAST design procedure. *Designs, Codes and Cryptography*, **12**: 283–316 (1997).

[2] Y. Borissov, A. Braeken, S. Nikova and B. Preneel. On the covering radii of binary Reed-Muller codes in the set of resilient Boolean functions. *IEEE Transactions on Information Theory*, **51(3)**: 1182–1189 (2005).

[3] A. Braeken, V. Nikov, S. Nikova and B. Preneel. On Boolean functions with generalized cryptographic properties. In A. Canteaut and K. Viswanathan (editors), *Progress in Cryptology – INDOCRYPT 2004*, volume 3348 of *Lecture Notes in Computer Science*, pages 120–135. Springer-Verlag, Berlin, 2004.

[4] C. Carlet and Y. Tarannikov. Covering sequences of Boolean functions and their cryptographic significance. *Designs, Codes and Cryptography*, **25**: 263–279 (2002).

[5] J. F. Dillon. *Elementary Hadamard Difference Sets.* PhD Thesis, University of Maryland, 1974.

[6] K. C. Gupta and P. Sarkar. Improved construction of nonlinear resilient S-boxes. *IEEE Transactions on Information Theory*, **51(1)**: 339–348 (2005).

[7] K. Kurosawa, T. Iwata and T. Yoshiwara. New covering radius of Reed-Muller codes for t-resilient functions. *IEEE Transactions on Information Theory*, **50(3)**: 468–475 (2004).

[8] P. Langevin and G. Leander. Counting all bent functions in dimension eight. Submitted (Presented at the 2009 International Workshop on Coding and Cryptography. May 10–15, 2009. Ullensvang (Norway)).

[9] V. V. Losev. Decoding of sequences of bent functions by means of a fast Hadamard transform. *Soviet Journal of Communications Technology and Electronics*, **32(10)**: 155–157 (1987).

[10] F. J. MacWilliams and N. J. A. Sloane. *The Theory of Error-Correcting Codes.* North-Holland, Amsterdam, 6th edition, 1988.

[11] M. Matsui. Linear cryptanalysis method for DES cipher. In T. Helleseth (editor), *Advances in Cryptology – EUROCRYPT '93*, volume 765 of *Lecture Notes in Computer Science*, pages 386–397. Springer-Verlag, Berlin, 1994.

[12] R. L. McFarland. A family of difference sets in non-cyclic groups. *Journal of Combinatorial Theory (Series A)*, **15**: 1–10 (1973).

 ISBN: 978-84-614-6167-7

[13] W. Meier and O. Staffelbach. Nonlinearity criteria for cryptographic functions. In J. Quisquater and J. Vandewalle (editors), *Advances in Cryptology – EUROCRYPT'89*, volume 434 of *Lecture Notes in Computer Science*, pages 549–562. Springer-Verlag, Berlin, 1990.

[14] K. Nyberg. Perfect nonlinear S-boxes. In D. W. Davies (editor), *Advances in Cryptology – EUROCRYPT '91*, volume 547 of *Lecture Notes in Computer Science*, pages 378–386. Springer-Verlag, Berlin, 1991.

[15] O. S. Rothaus. On "bent" functions. *Journal of Combinatorial Theory (Series A)*, **20**: 300–305 (1976).

[16] P. Sarkar and S. Maitra. Construction of nonlinear Boolean functions with important cryptographic properties. In B. Preneel (editor), *Advances in Cryptology – EUROCRYPT 2000*, volume 1807 of *Lecture Notes in Computer Science*, pages 485–506. Springer-Verlag, Berlin, 2000.

[17] J. Seberry and X.-M. Zhang. Constructions of bent functions from two known bent functions. *Australasian Journal of Combinatorics*, **9**: 21–35 (1994).

[18] J. Seberry, X.-M. Zhang and Y. Zheng. Nonlinearity and propagation characteristics of balanced Boolean functions. *Information and Computation*, **119**: 1–13 (1995).

[19] S. A. Vanstone and P. C. van Oorschot. *An Introduction to Error Correcting Codes with Applications*. Kluwer Academic Publishers, Boston, MA, 2000.

[20] R. Yarlagadda and J. E. Hershey. Analysis and synthesis of bent sequences. *IEE Proceedings*, **136(2)**: 112–123 (1989).

 ISBN: 978-84-614-6167-7

Proceedings of the 11th International Conference on Computational and Mathematical Methods in Science and Engineering, CMMSE 2011 26–30 June 2011.

Key exchange protocols over noncommutative rings. The case $\mathrm{End}(\mathbb{Z}_p \times \mathbb{Z}_{p^2})$

Joan-Josep Climent[1], Pedro R. Navarro[2] and Leandro Tortosa[2]

[1] *Departament d'Estadística i Investigació Operativa, Universitat d'Alacant*

[2] *Departament de Ciència de la Computació i Intel·ligència Artificial, Universitat d'Alacant*

emails: jcliment@ua.es, prnr@alu.ua.es, tortosa@ua.es

Abstract

In this paper we introduce some key exchange protocols over noncommutative rings. These protocols uses some polynomials with coefficients in the center of the ring as part of the private keys. We give some examples over the ring $\mathrm{End}(\mathbb{Z}_p \times \mathbb{Z}_{p^2})$ where p is a prime number.

Key words: Key exchange protocol, noncommutative ring, center of a ring, polynomial, public key cryptography.

1 Introduction and preliminaries.

Most common public key cryptosystems and public key exchange protocols presently in use, are number theory based and hence theoretically depend on the structure of abelian groups. Their robustness is based on the difficulty of solving certain problems over finite commutative algebraic structures. One of these is the Integer Factorization Problem over the ring $\mathbb{Z}_n$, being n the product of two large prime numbers; the well known cryptosystem RSA [8] is based in this problem. The second classical problem is the Discrete Logarithm Problem over a finite field $\mathbb{Z}_p$, being p a large prime; the ElGamal protocol [4] and all its variants (see, for example, [7]) is based on this problem.

It is believed that the strength of computing machinery has made these techniques less secure. As a consequence of this there exists an active field of research named as noncommutative algebraic cryptography [6], with the aim to develop and analyze new cryptosystems and key exchange protocols based on noncommutative cryptographic platforms. Several authors have used nonabelian groups for public key exchange. In [1, 5], the authors suggested the braid groups as platform groups for their respective protocols.

 ISBN: 978-84-614-6167-7

The main idea of this work is the design of some public key exchange protocols over noncommutative rings, in particular over the ring of endomorphisms of $\mathbb{Z}_p \times \mathbb{Z}_{p^2}$ where p is a prime number p. Bergman [2] showed that this ring has p^5 elements and it is semilocal, but it cannot be embebed in matrices over any commutative ring. Nevertheless, Climent, Navarro, and Tortosa [3] established that

$$\mathrm{End}(\mathbb{Z}_p \times \mathbb{Z}_{p^2}) = \left\{ \begin{bmatrix} a & b \\ pc & d \end{bmatrix} \mid a, b, c \in \mathbb{Z}_p \text{ and } d \in \mathbb{Z}_{p^2} \right\}$$

where the addition and multiplication are given, respectively, by

$$\begin{bmatrix} a_1 & b_1 \\ pc_1 & d_1 \end{bmatrix} + \begin{bmatrix} a_2 & b_2 \\ pc_2 & d_2 \end{bmatrix} = \begin{bmatrix} (a_1 + a_2) \bmod p & (b_1 + b_2) \bmod p \\ p(c_1 + c_2) \bmod p^2 & (d_1 + d_2) \bmod p^2 \end{bmatrix}$$

and

$$\begin{bmatrix} a_1 & b_1 \\ pc_1 & d_1 \end{bmatrix} \cdot \begin{bmatrix} a_2 & b_2 \\ pc_2 & d_2 \end{bmatrix} = \begin{bmatrix} (a_1a_2) \bmod p & (a_1b_2 + b_1d_2) \bmod p \\ p(c_1a_2 + b_1c_2) \bmod p^2 & (pc_1b_2 + d_1d_2) \bmod p^2. \end{bmatrix}$$

We denote for simplicity $\mathrm{End}(\mathbb{Z}_p \times \mathbb{Z}_{p^2})$ by E_p.

The center of the ring E_p is the set

$$\mathcal{Z}(E_p) = \left\{ \begin{bmatrix} x & 0 \\ 0 & py + x \end{bmatrix} \mid x, y \in \mathbb{Z}_p \right\}.$$

It is easy to check that the number of elements of $\mathcal{Z}(E_p)$ is p^2 which coincides with the characteristic of E_p.

2 A modification of Stickel's Protocol over nonconmutative rings

From the Stickel's key exchange protocol [10] we propose the following protocol over a noncommutative ring R.

Protocol 1: The elements M, $N \in R$ are public.

Step 1: Alice and Bob choose their private keys $(r, s), (u, v) \in \mathbb{N}^2$ respectively.

Step 2: Alice computes her public key $P_A = M^r N M^s$ and sends it to Bob.

Similarly, Bob computes his public key $P_B = M^u N M^v$ and sends it to Alice.

Step 3: Alice and Bob compute S_A and S_B respectively as

$$S_A = M^r P_B M^s \quad \text{and} \quad S_B = M^u P_A N^v.$$

The shared secret is $S_A = S_B$ as we can see in the following theorem. □

Theorem 1: *With the above notation, it follows that* $S_A = S_B$.

 ISBN: 978-84-614-6167-7

Proof: The result follows from the fact that $M^k M^l = M^l M^k$ for all $k, l \in \mathbb{N}$. □

Note that if $MN = NM$ then

$$P_A = M^r N M^s = N M^r M^s \quad \text{and} \quad P_B = M^u N M^v = N M^u M^v,$$

therefore

$$NS_A = P_A P_B = NS_B,$$

and the shared secret $S_A = S_B$ may be easily obtained by an unauthorized part, since N, P_A and P_B are public.

Then, we need that $MN \neq NM$; therefore, from now on we will assume that $N \notin \mathcal{Z}(R)$.

Thus, the security of this protocol is based on achieving an element M with large order. However, using the ideas of Shpilrain [9] it is easy to cryptanalyze the above protocol because the element M is public.

In order to avoid this weakness in the protocol, we propose in the next section two new protocols in which, instead of considering the element M directly, we consider the elements $f(M)$ and $g(M)$ obtained from M and two polynomials $f(X), g(X) \in \mathcal{Z}(R)[\mathsf{X}]$.

3 Key exchange protocols using polynomials over a non-commutative ring

Assume that R is a noncommutative ring. If we consider $f(X), g(X) \in \mathcal{Z}(R)[\mathsf{X}]$ and $k, l \in \mathbb{N}$, although R is not commutative, we have that

$$f(M)^k g(M)^l = g(M)^l f(M)^k, \quad \text{for all } M \in R. \tag{1}$$

This property allows us to establish the following protocol.

Protocol 2: The elements $M \in R$ and $N \in R \setminus \mathcal{Z}(R)$, are public.

Step 1: Alice choose her private key $f(\mathsf{X}) \in \mathcal{Z}(R)[\mathsf{X}]$ and $r, s \in \mathbb{N}$.

Bob choose his private key $g(\mathsf{X}) \in \mathcal{Z}(R)[\mathsf{X}]$ and $u, v \in \mathbb{N}$.

Step 2: Alice computes her public key $P_A = f(M)^r N f(M)^s$, and sends it to Bob.

Analogously, Bob computes his public key $P_B = g(M)^u N g(M)^v$, and sends it to Alice.

Step 3: Alice and Bob compute S_A and S_B respectively as,

$$S_A = f(M)^r P_B f(M)^s \quad \text{and} \quad S_B = g(M)^u P_A g(M)^v. \qquad \square$$

As in Protocol 1, the shared secret is $S_A = S_B$, as we can see in the following theorem.

 ISBN: 978-84-614-6167-7

Theorem 2: *With the above notation, it follows that* $S_A = S_B$.

Proof: The result follows from expression (1). □

In the next example we show how to share a secret using the above protocol over the ring E_{11}.

Example 1: The starting point of the protocol consists on the sharing the elements

$$M = \begin{bmatrix} 5 & 8 \\ 44 & 102 \end{bmatrix} \in E_{11} \quad \text{and} \quad N = \begin{bmatrix} 10 & 3 \\ 77 & 37 \end{bmatrix} \in E_{11} \setminus \mathcal{Z}(E_{11}).$$

Now, we run the steps of the protocol.

Step 1: Alice chooses her private key $r = 3$, $s = 5$ and

$$f(\mathsf{X}) = \begin{bmatrix} 3 & 0 \\ 0 & 47 \end{bmatrix} + \begin{bmatrix} 3 & 0 \\ 0 & 80 \end{bmatrix} \mathsf{X} + \begin{bmatrix} 9 & 0 \\ 0 & 108 \end{bmatrix} \mathsf{X}^2 + \begin{bmatrix} 5 & 0 \\ 0 & 49 \end{bmatrix} \mathsf{X}^3 \in \mathcal{Z}(E_{11})[\mathsf{X}].$$

Then, $f(M) = \begin{bmatrix} 10 & 8 \\ 44 & 74 \end{bmatrix}$.

Bob chooses his private key $u = 2$, $v = 4$ and

$$g(\mathsf{X}) = \begin{bmatrix} 9 & 0 \\ 0 & 86 \end{bmatrix} + \begin{bmatrix} 6 & 0 \\ 0 & 72 \end{bmatrix} \mathsf{X} + \begin{bmatrix} 5 & 0 \\ 0 & 38 \end{bmatrix} \mathsf{X}^2 \in \mathcal{Z}(E_{11})[\mathsf{X}].$$

Then $g(M) = \begin{bmatrix} 10 & 5 \\ 88 & 39 \end{bmatrix}$.

Step 2: Alice computes her public key P_A as

$$P_A = f(M)^r N f(M)^s = \begin{bmatrix} 10 & 8 \\ 44 & 74 \end{bmatrix}^3 \begin{bmatrix} 10 & 3 \\ 77 & 37 \end{bmatrix} \begin{bmatrix} 10 & 8 \\ 44 & 74 \end{bmatrix}^5 = \begin{bmatrix} 10 & 5 \\ 110 & 9 \end{bmatrix},$$

and sends it to Bob.

Similarly, Bob computes his public key P_B as

$$P_B = g(M)^u N g(M)^v = \begin{bmatrix} 10 & 5 \\ 88 & 39 \end{bmatrix}^2 \begin{bmatrix} 10 & 3 \\ 77 & 37 \end{bmatrix} \begin{bmatrix} 10 & 5 \\ 88 & 39 \end{bmatrix}^4 = \begin{bmatrix} 10 & 2 \\ 99 & 31 \end{bmatrix},$$

and sends it to Alice.

Step 3: Alice computes S_A as

$$S_A = f(M)^r P_B f(M)^s = \begin{bmatrix} 10 & 8 \\ 44 & 74 \end{bmatrix}^3 \begin{bmatrix} 10 & 2 \\ 99 & 31 \end{bmatrix} \begin{bmatrix} 10 & 8 \\ 44 & 74 \end{bmatrix}^5 = \begin{bmatrix} 10 & 1 \\ 99 & 56 \end{bmatrix}.$$

Bob computes S_B as

$$S_B = g(M)^u P_A g(M)^v = \begin{bmatrix} 10 & 5 \\ 88 & 39 \end{bmatrix}^2 \begin{bmatrix} 10 & 5 \\ 110 & 9 \end{bmatrix} \begin{bmatrix} 10 & 5 \\ 88 & 39 \end{bmatrix}^4 = \begin{bmatrix} 10 & 1 \\ 99 & 56 \end{bmatrix}.$$

As we established in Theorem 2, the shared secret is $S_A = S_B$.

 ISBN: 978-84-614-6167-7

E_p	Degree of the polynomial									
p \ n	2	3	4	5	$\cdots$	12	13	$\cdots$	20	$\cdots$
2	12	16	20	24	$\cdots$	52	56	$\cdots$	84	$\cdots$
3	27	36	45	54	$\cdots$	117	126	$\cdots$	189	$\cdots$
5	75	100	125	150	$\cdots$	325	350	$\cdots$	525	$\cdots$
7	147	196	245	294	$\cdots$	637	686	$\cdots$	1029	$\cdots$
11	363	484	605	726	$\cdots$	1573	1694	$\cdots$	2541	$\cdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$		$\vdots$	
97	28227	37636	47045	56454	$\cdots$	122317	131726	$\cdots$	197589	$\cdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$		$\vdots$	

Table 1: Number of polynomials for different degrees n and primes p

Note that an attacker knows the element M because it is public, but the elements $f(\mathsf{X}), g(\mathsf{X}) \in \mathcal{Z}(E_{11})[\mathsf{X}]$ are unknown. Consequently, the following elements are also unknown

$$f(M)^r = \begin{bmatrix} 10 & 8 \\ 44 & 74 \end{bmatrix}^3 = \begin{bmatrix} 10 & 5 \\ 88 & 72 \end{bmatrix} \quad \text{and} \quad f(M)^s = \begin{bmatrix} 10 & 8 \\ 44 & 74 \end{bmatrix}^5 = \begin{bmatrix} 10 & 0 \\ 0 & 76 \end{bmatrix}.$$

as well as

$$g(M)^u = \begin{bmatrix} 10 & 5 \\ 88 & 39 \end{bmatrix}^2 = \begin{bmatrix} 1 & 3 \\ 77 & 25 \end{bmatrix} \quad \text{and} \quad g(M)^v = \begin{bmatrix} 10 & 5 \\ 88 & 39 \end{bmatrix}^4 = \begin{bmatrix} 1 & 1 \\ 66 & 9 \end{bmatrix}.$$

To recover the secret, an attacker must consider, for each user, the resolution of the equation

$$X^k N X^l = P,$$

where only N and P are known while X, k and l are unknown. □

We could think about a brute force attack on the set of polynomials with coefficients in the center of the ring. However, an attack of this kind is not viable because the number of polynomials of degree n and coefficients in $\mathcal{Z}(E_p)$, is $(n+1)p^2$. It is enough to take n or p sufficiently big. Table 1 shows the values of $(n+1)p^2$ for different values of n and p.

Note that Protocol 2 presents some symmetry in the sense that Alice and Bob uses the same polynomial to multiply both on the right and the left. To avoid this symmetry we introduce two polynomials for each user.

Protocol 3: The elements $M \in R$, $N \in R \setminus \mathcal{Z}(R)$, are public.

Step 1: Alice chooses her private key $f_1(\mathsf{X}), f_2(\mathsf{X}) \in \mathcal{Z}(R)[\mathsf{X}]$ and $r, s \in \mathbb{N}$.

Bob chooses his private key $g_1(\mathsf{X}), g_2(\mathsf{X}) \in \mathcal{Z}(R)[\mathsf{X}]$ and $u, v \in \mathbb{N}$.

Step 2: Alice computes her public key $P_A = f_1(M)^r N f_2(M)^s$ and sends it to Bob.

Similarly, Bob computes his public key $P_B = g_1(M)^u N g_2(M)^v$, and sends it to Alice.

 ISBN: 978-84-614-6167-7

Step 3: Alice and Bob compute S_A and S_B respectively as

$$S_A = f_1(M)^r P_B f_2(M)^s \quad \text{and} \quad S_B = g_1(M)^u P_A g_2(M)^s.$$

By a similar argument as in Protocol 2, it follows that $S_A = S_B$. □

In the next example, we show how to share a secret using the above protocol.

Example 2: We consider again the elements M and N of E_{11} as in Example 1.

Step 1: Alice chooses her private key $r = 3$, $s = 5$ and $f_1(\mathsf{X})$, $f_2(\mathsf{X})$ as

$$f_1(\mathsf{X}) = \begin{bmatrix} 3 & 0 \\ 0 & 47 \end{bmatrix} + \begin{bmatrix} 3 & 0 \\ 0 & 80 \end{bmatrix} \mathsf{X} + \begin{bmatrix} 9 & 0 \\ 0 & 108 \end{bmatrix} \mathsf{X}^2 + \begin{bmatrix} 5 & 0 \\ 0 & 49 \end{bmatrix} \mathsf{X}^3 \in \mathcal{Z}(E_{11})[\mathsf{X}],$$

$$f_2(\mathsf{X}) = \begin{bmatrix} 4 & 0 \\ 0 & 81 \end{bmatrix} + \begin{bmatrix} 3 & 0 \\ 0 & 14 \end{bmatrix} \mathsf{X}^2 + \begin{bmatrix} 1 & 0 \\ 0 & 34 \end{bmatrix} \mathsf{X}^3 + \begin{bmatrix} 1 & 0 \\ 0 & 56 \end{bmatrix} \mathsf{X}^4 \in \mathcal{Z}(E_{11})[\mathsf{X}].$$

Then, $f_1(M) = \begin{bmatrix} 10 & 8 \\ 44 & 74 \end{bmatrix}$ and $f_2(M) = \begin{bmatrix} 4 & 10 \\ 55 & 29 \end{bmatrix}$.

Bob chooses his private key $u = 2$, $v = 4$ and $g_1(\mathsf{X})$, $g_2(\mathsf{X})$ as

$$g_1(\mathsf{X}) = \begin{bmatrix} 9 & 0 \\ 0 & 86 \end{bmatrix} + \begin{bmatrix} 6 & 0 \\ 0 & 72 \end{bmatrix} \mathsf{X} + \begin{bmatrix} 5 & 0 \\ 0 & 38 \end{bmatrix} \mathsf{X}^2 \in \mathcal{Z}(E_{11})[\mathsf{X}],$$

$$g_2(\mathsf{X}) = \begin{bmatrix} 4 & 0 \\ 0 & 70 \end{bmatrix} + \begin{bmatrix} 5 & 0 \\ 0 & 49 \end{bmatrix} \mathsf{X} + \begin{bmatrix} 10 & 0 \\ 0 & 87 \end{bmatrix} \mathsf{X}^2 + \begin{bmatrix} 10 & 0 \\ 0 & 109 \end{bmatrix} \mathsf{X}^3 \in \mathcal{Z}(E_{11})[\mathsf{X}].$$

Then, $g_1(M) = \begin{bmatrix} 10 & 5 \\ 88 & 39 \end{bmatrix}$ and $g_2(M) = \begin{bmatrix} 8 & 6 \\ 33 & 56 \end{bmatrix}$.

Step 2: Alice computes her public key P_A as

$$P_A = f_1(M)^r N f_2(M)^s = \begin{bmatrix} 10 & 8 \\ 44 & 74 \end{bmatrix}^3 \begin{bmatrix} 5 & 8 \\ 44 & 88 \end{bmatrix} \begin{bmatrix} 4 & 10 \\ 55 & 29 \end{bmatrix}^5 = \begin{bmatrix} 1 & 2 \\ 99 & 20 \end{bmatrix},$$

and sends it to Bob.

Similarly, Bob computes his public key P_B as

$$P_B = g_1(M)^u N g_2(M)^v = \begin{bmatrix} 10 & 5 \\ 88 & 39 \end{bmatrix}^2 \begin{bmatrix} 10 & 3 \\ 77 & 37 \end{bmatrix} \begin{bmatrix} 8 & 6 \\ 33 & 56 \end{bmatrix}^4 = \begin{bmatrix} 7 & 3 \\ 77 & 56 \end{bmatrix},$$

and sends it to Alice.

Step 3: Alice computes S_A as

$$S_A = f_1(M)^r P_B f_2(M)^s = \begin{bmatrix} 10 & 8 \\ 44 & 74 \end{bmatrix}^3 \begin{bmatrix} 7 & 3 \\ 77 & 56 \end{bmatrix} \begin{bmatrix} 4 & 10 \\ 55 & 29 \end{bmatrix}^5 = \begin{bmatrix} 4 & 8 \\ 11 & 82 \end{bmatrix}.$$

Bob computes S_B as

$$S_B = g_1(M)^u P_A g_2(M)^v = \begin{bmatrix} 10 & 5 \\ 88 & 39 \end{bmatrix}^2 \begin{bmatrix} 1 & 2 \\ 99 & 20 \end{bmatrix} \begin{bmatrix} 8 & 6 \\ 33 & 56 \end{bmatrix}^4 = \begin{bmatrix} 4 & 8 \\ 11 & 82 \end{bmatrix}.$$

 ISBN: 978-84-614-6167-7

E_p	Degree of the polynomials					
p \ $[m,n]$	$[2,3]$	$[3,4]$	$[4,5]$	$\cdots$	$[9,10]$	$\cdots$
2	192	320	480	$\cdots$	1760	$\cdots$
3	972	1620	2430	$\cdots$	8910	$\cdots$
5	7500	12500	18750	$\cdots$	68750	$\cdots$
7	28812	48020	72030	$\cdots$	264110	$\cdots$
11	175692	292820	439230	$\cdots$	1610510	$\cdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$	
97	1062351372	1770585620	2655878430	$\cdots$	9738220910	$\cdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$	

Table 2: Number of polynomials for different values of the degrees and p

The shared secret is then $S_A = S_B$.

Note that an attacker knows M because it is public, but $f_1(\mathsf{X}), f_2(\mathsf{X}), g_1(\mathsf{X}), g_2(\mathsf{X}) \in \mathcal{Z}(E_{11})[\mathsf{X}]$ remain unknown. Consequently,

$$f_1(M)^r = \begin{bmatrix} 10 & 8 \\ 44 & 74 \end{bmatrix}^3 = \begin{bmatrix} 10 & 5 \\ 88 & 72 \end{bmatrix} \quad \text{and} \quad f_2(M)^s = \begin{bmatrix} 4 & 10 \\ 55 & 29 \end{bmatrix}^5 = \begin{bmatrix} 1 & 8 \\ 44 & 98 \end{bmatrix},$$

are unknown, as well as

$$g_1(M)^u = \begin{bmatrix} 10 & 5 \\ 88 & 39 \end{bmatrix}^2 = \begin{bmatrix} 1 & 3 \\ 77 & 25 \end{bmatrix} \quad \text{and} \quad g_2(M)^v = \begin{bmatrix} 8 & 6 \\ 33 & 56 \end{bmatrix}^4 = \begin{bmatrix} 4 & 1 \\ 66 & 78 \end{bmatrix}.$$

Following a similar argument as for Protocol 2, an attacker who want to discover the shared secret must solve the equation

$$X^k N Y^l = P,$$

where N and P are public and X, Y, k, l are unknown. □

In this protocol a user needs two polynomials with degrees m and n, respectively; therefore the number of possible polynomials becomes $(m+1)(n+1)p^4$. Table 2 shows the values of $(m+1)(n+1)p^4$ for different values of m, n and p.

4 Conclusion

In this paper we propose two new key exchange protocols based on noncommutative rings that avoid the weaknesses of the Stickel's protocol. The central idea underlying these protocols is the use of polynomials with coefficients in the center of the ring; these polynomials become a part of the private key for each user. Thus, an attacker who wants to recover the shared secret must solve an equation of the form

$$X^k N X^l = P \quad \text{or} \quad X^k N Y^l = P$$

where only $N, P \in R$ are known.

These protocols have been designed to work, in general, with any noncommutative ring. In our case, they have been applied to the particular case of the ring $\mathrm{End}(\mathbb{Z}_p \times \mathbb{Z}_{p^2})$.

 ISBN: 978-84-614-6167-7

5 Acknowledgements

This work was partially supported by Spanish grants MTM2008-06674-C02-01 of the Ministerio de Ciencia e Innovación of the Gobierno de España and ACOMP/2011/005 of the Generalitat Valenciana.

References

[1] I. Anshel, M. Anshel and D. Goldfeld. An algebraic method for public-key cryptography. *Mathematical Research Letters*, **6**: 1–5 (1999).

[2] G. M. Bergman. Some examples in PI ring theory. *Israel Journal of Mathematics*, **18**: 257–277 (1974).

[3] J.-J. Climent, P. R. Navarro and L. Tortosa. On the arithmetic of the endomorphisms ring $\mathrm{End}(\mathbb{Z}_p \times \mathbb{Z}_{p^2})$. *Applicable Algebra in Engineering, Communication and Computing*, **22(2)**: 91–108 (2011).

[4] T. ElGamal. A public key cryptosystem and a signature scheme based on discrete logarithms. *IEEE Transactions on Information Theory*, **31(4)**: 469–472 (1985).

[5] K. H. Ko, S. J. Lee, J. H. Cheon, J. W. Han, J.-s. Kang and C. Park. New public-key cryptosystem using braid groups. In M. Bellare (editor), *Advances in Cryptology – CRYPTO 2000*, volume 1880 of *Lecture Notes in Computer Science*, pages 166–183. Springer-Verlag, Berlin, 2000.

[6] A. G. Myasnikov, V. Shpilrain and A. Ushakov. *Group-based cryptography*. Birkhäuser Verlag, Basel, Switzerland, 2008.

[7] J. Proos and C. Zalka. Shor's discrete logarithm quantum algorithm for elliptic curves. arXiv:quant-ph/0301141v2.

[8] R. L. Rivest, A. Shamir and L. Adleman. A method for obtaining digital signatures and public-key cryptosystems. *Communications of the ACM*, **21(2)**: 120–126 (1978).

[9] V. Shpilrain. Cryptanalysis of Stickel's key exchange scheme. In E. A. Hirsch, A. A. Razborov, A. Semenov and A. Slissenko (editors), *Computer Science – Theory and Applications*, volume 5010 of *Lecture Notes in Computer Science*, pages 283–288. Springer-Verlag, Berlin, 2008.

[10] E. Stickel. A new method for exchanging secret keys. In *Proceedings of the Third International Conference on Information Technology and Applications (ICITA'05)*, pages 426–430. Sidney, Australia, 2005.

 ISBN: 978-84-614-6167-7

Proceedings of the 11th International Conference on Computational and Mathematical Methods in Science and Engineering, CMMSE 2011 26–30 June 2011.

Fourth and eighth-order optimal derivative-free methods for solving nonlinear equations

Alicia Cordero[1], José L. Hueso[1], Eulalia Martínez[2] and Juan R. Torregrosa[1]

[1] *Instituto de Matemática Multidisciplinar, Universitat Politècnica de València, Camino de Vera, s/n, 46022 València, Spain*

[2] *Instituto de Matemática Pura y Aplicada, Universitat Politècnica de València, Camino de Vera, s/n, 46022 València, Spain*

emails: acordero@mat.upv.es, jlhueso@mat.upv.es, eumarti@mat.upv.es, jrtorre@mat.upv.es

Abstract

Based on Steffensen-like methods and Padé approximant, a technique to obtain derivative-free methods, with optimal order in the sense of the Kung-Traub conjecture, for solving nonlinear smooth equations is described. Derivative-free methods of fourth and eighth optimal orders are theoretically discussed. Numerical examples are made to show the performance of the presented methods, on nonsmooth equations, and to compare with another ones.

Key words: Nonlinear equations, derivative-free, iterative methods, convergence order, efficiency index, Padé approximant.

1 Introduction

Finding iterative methods for solving nonlinear equations is an important area of research in numerical analysis and it has interesting applications in various branches of science and engineering. In this study, we describe new iterative methods to find a simple root α of a nonlinear equation $f(x) = 0$, where $f : I \subset \mathbb{R} \to \mathbb{R}$ is a scalar function on an open interval I. The known Newton's method for finding α uses the iterative expression

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}, \quad n = 0, 1, \ldots$$

which converges quadratically in some neighborhood of α. If the derivative $f'(x_n)$ is replaced by the forward-difference approximation

$$f'(x_n) \approx \frac{f(x_n + f(x_n)) - f(x_n)}{f(x_n)},$$

 ISBN: 978-84-614-6167-7

the Newton's method becomes

$$x_{n+1} = x_n - \frac{f(x_n)^2}{f(z_n) - f(x_n)}, \tag{1}$$

where $z_{=}x_{+}f(x_n)$, which is the known Steffensen's method (SM), (see [8]). This scheme is a tough competitor of Newton's method. Both methods are of second order, both require two functional evaluations per step, but in contrast to Newton's method, Steffensen's method is derivative-free.

The procedure of removing the derivatives usually increases the number of functional evaluations per iteration. Commonly in the literature the efficiency of an iterative method is measured by the *efficiency index* defined as $I = p^{1/d}$ (see [9]), where p is the order of convergence and d is the total number of functional evaluations per step. Kung and Traub conjectured in [6] that the order of convergence of any multipoint method cannot exceed the bound 2^{d-1}, (called *the optimal order*). Thus, the optimal order for methods with 2, 3 or 4 functional evaluations per step would be 2, 4 or 8, respectively. Newton and Steffensen's methods are optimal schemes of order 2.

Recently, Ren et al. derive in [10] a one-parameter class of fourth-order methods (RWB) with three functional evaluations per step. In these methods, an interpolation polynomial of order three is used to get a better approximation to the derivative of the given function. The iterative expression is:

$$x_{n+1} = y_n - \frac{f(y_n)}{f[x_n, y_n] + f[y_n, z_n] - f[x_n, z_n] + a(y_n - x_n)(y_n - z_n)}, \tag{2}$$

where y_n is the approximation of the Steffensen's method and $f[x, y]$ is the divided difference of order one. Other Steffensen type methods and their applications are also discussed by Liu et al. in [7] and by Cordero and Torregrosa in [4], where the authors also obtain different optimal fourth-order methods; by Zheng et al. in [11] and by Feng and He in [5]. As far as we know, there is not optimal derivative-free schemes of order greater than four.

In this paper, the technique used to improve the local order of convergence consists of the composition of two iterative methods of order p and q, respectively, to obtain a method of order pq (see [8]). In the first proposed method, we compose Steffensen and Newton's methods and we use a Padé approximant of degree one to get a good approximation to the derivative of the given function. The resulting method has order of convergence four and requires three evaluations of the function $f(x)$ per step, therefore it is an optimal method with efficiency index $4^{1/3} = 1.587$.

If we compose again the above method with Newton's method and use a Padé approximant of degree two, we can derive a new iterative scheme of eighth-order of convergence, which requires four evaluations of the function $f(x)$ per step, that is an optimal scheme with efficiency index $8^{1/4} = 1.6817$. We think that this technique can be extended in order to obtain a derivative-free optimal methods of orders 16, 32, ...

The paper is organized as follows. In Section 2 we describe the new methods and analyze its convergence order for smooth equations. In Section 3, different numerical tests confirm the theoretical results and allow us to compare these methods with other

 ISBN: 978-84-614-6167-7

known schemes mentioned in this section. We also analyze in this numerical section the behavior of the new schemes on nonsmooth equations.

2 Description of the methods and convergence analysis

We first compose the well-known Steffensen method, defined by (1), with Newton's method obtaining the fourth-order scheme

$$\begin{aligned} y_n &= x_n - \frac{f(x_n)^2}{f(z_n)-f(x_n)}, \\ x_{n+1} &= y_n - \frac{f(y_n)}{f'(y_n)}, \end{aligned} \tag{3}$$

where $z_n = x_n + f(x_n)$. Now, in order to avoid the evaluation of $f'(y_n)$, we approximate it by the derivative $m'(y_n)$ of the following Padé approximant of the first degree

$$m(t) = \frac{a_1 + a_2(t-y_n)}{1+a_3(t-y_n)}, \tag{4}$$

where a_1, a_2 and a_3 are real parameters to be determined satisfying the following conditions:

$$m(x_n) = f(x_n), \tag{5}$$

$$m(y_n) = f(y_n), \tag{6}$$

and

$$m(z_n) = f(z_n). \tag{7}$$

Directly from (6) one obtains

$$a_1 = f(y_n). \tag{8}$$

From (5) and (7), we obtain, respectively,

$$a_2 - f(x_n)a_3 = f[x_n, y_n] \tag{9}$$

and

$$a_2 - f(z_n)a_3 = f[z_n, y_n]. \tag{10}$$

After some algebraic manipulations, the following values are obtained for the parameters:

$$a_2 = f[y_n, z_n] - \frac{f(z_n)f[x_n, y_n, z_n]}{f[z_n, x_n]} \tag{11}$$

and

$$a_3 = -\frac{f[x_n, y_n, z_n]}{f[x_n, z_n]}, \tag{12}$$

where $f[x_n, y_n, z_n] = \frac{f[x_n,y_n]-f[y_n,z_n]}{x_n-z_n}$ denotes the divided difference of order 2.

 ISBN: 978-84-614-6167-7

Therefore, the derivative of the Padé approximant evaluated in y_n can be expressed as

$$m'(y_n) = \frac{f[x_n, y_n]f[y_n, z_n]}{f[x_n, z_n]}, \tag{13}$$

and substituting (13) in the last step of the iteration, we obtain a new iterative method, which we denote by $M4$, whose expression is:

$$\begin{aligned} y_n &= x_n - \frac{f(x_n)^2}{f(z_n) - f(x_n)}, \\ x_{n+1} &= y_n - \frac{f(y_n)f[x_n, z_n]}{f[x_n, y_n]f[y_n, z_n]}. \end{aligned} \tag{14}$$

Let us note that in each iteration we only evaluate $f(x_n)$, $f(y_n)$ and $f(z_n)$, so that the method will be optimal in the sense of Kung-Traub's conjecture, if we show that its convergence order is 4.

Theorem 1 *Let $\alpha \in I$ be a simple zero of a sufficiently differentiable function $f : I \subseteq \mathbb{R} \longrightarrow \mathbb{R}$ in an open interval I. If x_0 is sufficiently close to α, then the method $M4$ defined by (14) has optimal convergence order 4.*

This technique can be applied on higher order methods than Steffensen's one, in order to obtain a new high-order derivative-free method. So, let us consider a composition between $M4$ and Newton's scheme; if a Padé approximant of degree two is applied on the estimation of the derivative in the last step, the resulting method would appear as:

$$\begin{aligned} y_n &= x_n - \frac{f(x_n)^2}{f(z_n) - f(x_n)}, \\ u_n &= y_n - \frac{f(y_n)f[x_n, z_n]}{f[x_n, y_n]f[y_n, z_n]}, \\ x_{n+1} &= u_n - \frac{f(u_n)}{\bar{m}'(u_n)}, \end{aligned} \tag{15}$$

where $\bar{m}(t) = \frac{b_1+b_2(t-u_n)+b_3(t-u_n)^2}{1+b_4(t-u_n)}$, and the parameters b_1, b_2, b_3 and b_4 must satisfy the following conditions:

$$\bar{m}(x_n) = f(x_n), \tag{16}$$

$$\bar{m}(y_n) = f(y_n), \tag{17}$$

$$\bar{m}(z_n) = f(z_n), \tag{18}$$

and

$$\bar{m}(u_n) = f(u_n). \tag{19}$$

Again, from (19) one obtains

$$b_1 = f(u_n), \tag{20}$$

 ISBN: 978-84-614-6167-7

and, by solving the system described by (16), (17) and (18) in a similar way as before, the following values are obtained for the parameter b_4:

$$b_4 = \frac{f[y_n, u_n, x_n] - f[y_n, u_n, z_n]}{f[y_n, z_n] - f[y_n, x_n]} \tag{21}$$

and also b_3:

$$b_3 = f[y_n, u_n, z_n] + b_4 f[y_n, z_n] \tag{22}$$

and b_2:

$$b_2 = f[y_n, u_n] - b_3(y_n - u_n) + f(y_n)b_4. \tag{23}$$

Therefore the derivative of the second-degree Padé approximant can be expressed as

$$\bar{m}'(u_n) = b_2 - b_1 b_4, \tag{24}$$

and substituting (24) in (15), we obtain a new scheme, denoted by $M8$, whose iterative expression is:

$$\begin{aligned} y_n &= x_n - \frac{f(x_n)^2}{f(z_n) - f(x_n)}, \\ u_n &= y_n - \frac{f(y_n) f[x_n, z_n]}{f[x_n, y_n] f[y_n, z_n]}, \\ x_{n+1} &= u_n - \frac{f(u_n)}{b_2 - b_1 b_4}. \end{aligned} \tag{25}$$

Let us note that in each iteration we evaluate $f(x_n)$, $f(y_n)$, $f(z_n)$ and $f(u_n)$, so that the method would be optimal, if we show that its convergence order is 8.

Theorem 2 *Let $\alpha \in I$ be a simple zero of a sufficiently differentiable function $f : I \subseteq \mathbb{R} \longrightarrow \mathbb{R}$ in an open interval I. If x_0 is sufficiently close to α, then the method $M8$ defined by (25) has optimal convergence order 8.*

We conjecture that this technique would allow us to design new methods of increasing order of convergence (16, 32, ...) by using, respectively, Padé approximants of degree three, four, ..., which would be optimal derivative-free schemes.

3 Numerical results

In this section we check the effectiveness of the new optimal methods of order for and eight, $M4$ and $M8$, comparing them with Steffensen's and RWB methods of second and fourth order, respectively.

Nowadays, high-order methods are important because numerical applications use high precision in their computations; for this reason numerical tests have been carried out using variable precision arithmetic in MATLAB 7.1. with 2000 significant digits.

 ISBN: 978-84-614-6167-7

Tables 1 and 2 show for each initial estimation and every method, the exact absolute error at first and last iterations, the number of iterations required to obtain $incr1 = |x_{k+1} - x_k| < 10^{-150}$ or $incr2 = |f(x_{k+1})| < 10^{-150}$ and the corresponding computational order of convergence ρ (usually called ACOC), defined by Cordero et al. in [3]:

$$p \approx \rho = \frac{\ln(|x_{k+1} - x_k| \,/\, |x_k - x_{k-1}|)}{\ln(|x_k - x_{k-1}| \,/\, |x_{k-1} - x_{k-2}|)}. \tag{26}$$

The value of ρ appearing in Tables 1 and 2 is the last component of the vector defined by (26), when it is stable, in other case we will denote it by '-'.

The first test has been made on the nonsmooth function:

$$f(x) = \begin{cases} x(x+1), & \text{if } x < 0, \\ -2x(x-1), & \text{if } x \geq 0, \end{cases} \tag{27}$$

that can be found in [2]. We use three initial estimations in order to approximate the three different roots of the equation, $\{-1, 0, 1\}$. From Table 1 it can be inferred that, as every root is close to each other, the initial estimation must be quite good if convergence to the central root is looked for. In fact, for $x_0 = -20$, method M4 converges to 0, instead of the closest root, -1. Moreover, the nonsmoothness of the function (27) in 0 is the reason why the estimated order of convergence is 2 for all the methods. The stability problems do not allow the order of convergence to increase. When the initial approximation is near of 1 or -1, the behavior of the methods is stable and the ACOC is near the theoretical order of convergence. High-order methods are shown to be more efficient when the root is far enough; for $x_0 = -20$, the number of iterations needed have been reduced in a reason of 1/7 from Steffensen's method.

Now, we consider the nonsmooth function that can be found in [1]:

$$f(x) = \begin{cases} 10x(x^4 + x), & \text{if } x < 0, \\ 10x(x^3 + x), & \text{if } x \geq 0, \end{cases} \tag{28}$$

The numerical experiments made on this function are summarized in Table 2. In this case, function (28) has a unique root in $x = 0$. Then, we test the different methods with 3 initial estimations, each of them further from the zero than the previous one. Smoothness difficulties make the calculation of the ACOC unstable, so we have no information at this point. Nevertheless, the number of iterations needed is much lower in high-order methods, indeed in the optimal eighth-order method M8. Moreover, Table 2 shows that the exact error per iteration in M8 is lower than the one calculated by the fourth-order methods, in most of cases from the first iteration.

4 Conclusions

Two new high-order methods to solve nonlinear equations have been developed, both of them optimal and derivative free. The optimal eighth-order method is, as far as we know, the first optimal Steffensen-type scheme developed of this order of convergence. It seems to be quite stable and robust, as it can be faster than other known, excellent methods as RWB even on nonsmooth functions.

 ISBN: 978-84-614-6167-7

Acknowledgements

This research was supported by Ministerio de Ciencia y Tecnología MTM2010-18539 and by Vicerrectorado de Investigación, Universitat Politècnica de València PAID-06-2010-2285.

References

[1] S. Amat, S. Busquier, *On a higher order secant methods*, Applied Mathematics and Computation, **141** (2003) 321–329.

[2] S. Amat, S. Busquier, *On a Steffensen's type method and its behavior for semismooth equations*, Applied Mathematics and Computation, **177** (2006) 819–823.

[3] A. Cordero, J.R. Torregrosa, *Variants of Newton's method using 5fth-order quadrature formulas*, Applied Mathematics and Computation, **190** (2007) 686–698.

[4] A. Cordero, J.R. Torregrosa, *A class of Steffensen type methods with optimal order of convergence*, Applied Mathematics and Computation, **217** (2011) 7653–7659.

[5] X. Feng, Y. He, *High order oterative methods without derivatives for solving nonlinear equations*, Applied Mathematics and Computation, **186** (2007) 1617–1623.

[6] H.T. Kung, J.F. Traub, *Optimal order of one-point and multi-point iteration*, Applied Mathematics and Computation, **21** (1974) 643–651.

[7] Z. Liu, Q. Zheng, P. Zhao, *A variant of Steffensen's method of fourth-oder convergence and its applications*, Applied Mathematics and Computation, **216** (2010) 1978–1983.

[8] J.M. Ortega, W.G. Rheinboldt, *Iterative solutions of nonlinear equations in several variables*, Academic Press, New York, 1970.

[9] A.M. Ostrowski, *Solutions of equations and systems of equations*, Academic Press, New York-London, 1966.

[10] H. Ren, Q. Wu, W. Bi, *A class of two-step Steffensen type methods with fourth-order convergence*, Applied Mathematics and Computation, **209** (2009) 206–210.

[11] Q. Zheng, J. Wang, P. Zhao, L. Zhang, *A Steffensen-like method and its higher-order variants*, Applied Mathematics and Computation, **214** (2009) 10–16.

 ISBN: 978-84-614-6167-7

	SM		RWB		M4		M8	
$x_0 = 0.1$	iter	error	iter	error	iter	error	iter	error
	1	4.52e-2	1	1.93e-2	1	0.1	1	0.1
	2	4.60e-3	2	6.95e-4	2	1.98e-2	2	6.74e-3
$\alpha = 0$	⋮		⋮		⋮		⋮	
	8	9.98e-124	7	1.80e-92	7	6.50e-92	7	5.82e-130
	9	2.99e-246	8	6.45e-184	8	8.45e-183	8	5.08e-259
incr1	9.98e-124		1.80e-92		6.50e-92		5.82e-130	
incr2	2.99e-246		6.45e-184		8.45e-183		1.02e-258	
ρ	2.0058		2.0000		2.0000		2.0000	
$x_0 = 3$	iter	error	iter	error	iter	error	iter	error
	1	2.86e-1	1	1.75e-1	1	1.75e-1	1	3.70e-1
	2	1.53e-1	2	2.13e-3	2	4.52e-3	2	4.33e-5
$\alpha = 1$	⋮		⋮		⋮		3	4.92e-35
	9	2.72e-114	4	1.91e-43	5	2.40e-144	4	1.38e-274
	10	7.38e-228	5	1.34e-171	6	6.66e-575		
incr1	2.71e-114		1.91e-43		2.40e-144		4.92e-35	
incr2	1.48e-227		2.69e-171		0.0000		2.77e-274	
ρ	2.0000		4.0005		4.0000		7.6157	
$x_0 = -20$	iter	error	iter	error	iter	error	iter	error
	1	18.44	1	9.34e-1	1	9.13	1	4.23
	2	17.88	2	4.47e-1	2	4.19	2	5.25e-1
$\alpha = -1$	⋮		⋮		⋮		⋮	
	33	2.01e-115	6	6.19e-27	11	1	4	7.05e-36
	34	8.09e-335	7	5.61e-158	12	1	5	6.03e-422
incr1	2.0e-115		6.19e-27		1.37e-106		7.05e-36	
incr2	0.0000		5.61e-158		3.74e-212		0.0000	
ρ	-		-		2.0000		-	

Table 1: Numerical results for function (27)

 ISBN: 978-84-614-6167-7

	SM	RWB	M4	M8
$x_0 = 1$	iter error	iter error	iter error	iter error
$\alpha = 0$	1 9.96e-1 2 9.91e-1 ⋮ 100 6.61e-105 101 3.81e-311	1 4.86e-1 2 1.10e-1 ⋮ 5 1.03e-60 6 0	1 4.98e-1 2 1.37e-1 ⋮ 5 9.70e-76 6 0	1 8.97e-2 2 2.03e-6 ⋮ 3 1.05e-46 4 0
incr1	3.69e-36	1.03e-60	9.70e-76	1.05e-46
incr2	0	0	0	0
ρ	-	-	-	-
$x_0 = 16$	iter error	iter error	iter error	iter error
$\alpha = 0$	$> 10^4$	1 10.65 2 7.08 ⋮ 11 1.13e-87 12 0	1 10.65 2 7.08 ⋮ 11 4.72e-33 12 3.40e-159	1 6.76 2 2.79 ⋮ 6 1.36e-46 7 0
incr1		1.13e-87	4.72e-33	1.36e-46
incr2		0	3.40e-158	0
ρ		-	-	-
$x_0 = 32$	iter error	iter error	iter error	iter error
$\alpha = 0$	$> 10^4$	1 21.33 2 14.21 ⋮ 13 4.40e-58 14 0	1 21.33 2 14.21 ⋮ 13 1.78e-62 14 2.57e-306	1 13.56 2 5.71 ⋮ 7 2.34e-98 8 0
incr1		4.40e-58	1.78e-62	2.34e-98
incr2		0	2.57e-305	0
ρ		-	-	-

Table 2: Numerical results for function (28)

 ISBN: 978-84-614-6167-7

Proceedings of the 11th International Conference on Computational and Mathematical Methods in Science and Engineering, CMMSE 2011 26–30 June 2011.

On complex dynamics of some third-order iterative methods

Alicia Cordero[1], Juan R. Torregrosa[1] and Pura Vindel[2]

[1] *Instituto de Matemática Multidisciplinar, Universitat Politècnica de València, Camino de Vera, s/n, 46022 València, Spain*

[2] *Instituto de Matemáticas y Aplicaciones de Castellón, Universitat Jaume I, Av. de Vicent Sos Baynat s/n, 12071 Castello de la Plana, Spain*

emails: acordero@mat.upv.es, jrtorre@mat.upv.es, vindel@uji.es

Abstract

In this paper we analyze the dynamical behaviour of Potra-Pták and Midpoint methods on second degree complex polynomials. We obtain that, in both cases, the Julia set is a connected set that separates the basins of attraction of the roots of the polynomial: the Julia set of the Midpoint method is the same as for the Newton operator, but it is more complicated for the Potra-Pták operator. We explain these differences by obtaining the conjugacy function of each method.

Key words: Nonlinear equations, iterative methods, complex dynamics.

1 Introduction

Many engineering applications involve nonlinear equations $f(x) = 0$ whose solution can not be found by means of analytical methods. To approximate the solution of these equations we use iterative methods. This means that the output of the method is a sequence of images $\{x_0, R(x_0), R^2(x_0), ..., R^n(x_0), ...\}$ for the initial condition x_0, where R is a rational function that represents the fixed point operator of the iterative scheme. Therefore, it can be seen as a discrete dynamical system and we can study it from this point of view.

There is an extensive literature on the study of iteration of rational mappings R of a complex variable (see [5], [6], for example) and, moreover, the Newton's method (see [4], [7] for example) applied on polynomials is a rational function. In this case, the Riemann sphere $\hat{\mathbb{C}}$ is also considered.

To our knowledge, the study on the dynamics of Newton's method has been extended to other point-to-point iterative methods used for solving nonlinear equations,

 ISBN: 978-84-614-6167-7

with convergence order up to three (see, for example [1], [2] and, more recently, [8] and [11]).

S. Amat et al. in [3] make a brief raid into the study of the dynamics of the Potra-Pták method (see [12]) defined on the real numbers and applied on polynomials of second and third degrees. This study, although interesting by itself, does not allow to see all the richness of the dynamics of the method when it is defined on the complex numbers.

Now, let us recall some basic concepts on complex dynamics. Given a rational function $R : \hat{\mathbb{C}} \to \hat{\mathbb{C}}$. The *orbit of a point* z_0 is defined as:

$$z_0,\ R(z_0),\ R^2(z_0),...,R^n(z_0),...$$

and we are interested in the study of the asymptotic behaviour of the orbits depending on the initial condition z_0, that is, we are interested in the study of the phase plane of the map defined by the iterative method.

To obtain this phase space, the first of all is to classify the initial conditions from the asymptotic behavior of their orbits.

As α is a root of f, $f(\alpha) = 0$, *the basin of attraction* of α is defined as the set of pre-images of any order:

$$\mathcal{A}(\alpha) = \{z_0 \in \mathbb{C} | R^n(z_0) \to \alpha, n \to \infty\}.$$

A *fixed point* z_0 is a point that satisfies: $R(z_0) = z_0$. A *periodic point* z_0 of period $p > 1$ is a point such that $R^p(z_0) = z_0$ and $R^k(z_0) \neq z_0$, $k < p$. 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 vanishes, $R'(z_0) = 0$.

On the other hand, a fixed point z_0 is called *attractor* if $|R'(z_0)| < 1$, *superattractor* if $|R'(z_0)| = 0$, and *repulsor* if $|R'(z_0)| > 1$.

The set of points $z \in \hat{\mathbb{C}}$ such that their families $\{R^n(z)\}_{n\in\mathbb{N}}$ are normal in some neighborhood $U(z)$, is the *Fatou set,* $\mathcal{F}(R)$, that is, the Fatou set is composed by the set of points whose orbits tend to an attractor (fixed point, periodic orbit or infinity). Its complement in $\hat{\mathbb{C}}$ is the *Julia set,* $\mathcal{J}(R)$; therefore, the Julia set includes all repelling fixed points, periodic orbits and their pre-images. That means that the basin of attraction of any fixed point belongs to the Fatou set. On the contrary, the boundaries of the basins of attraction belong to the Julia set.

1.1 The Newton's Method

The Newton's method is the best known method to find the roots of a nonlinear function:

$$f(z) = 0$$

where $f \in C^1\left(\hat{\mathbb{C}}\right)$ is defined on the Riemann sphere $\hat{\mathbb{C}}$. The Newton's iterative operator is

$$N_f(z) = z - \frac{f(z)}{f'(z)} \tag{1}$$

 ISBN: 978-84-614-6167-7

which satisfies that $f(z) = 0$ if and only if $N_f(z) = z$. So, to find the roots of $f(z)$ is equivalent to find the fixed points of the operator $N_f(z)$. Actually, the global analysis of convergence of Newton's method of $f(z)$ is equivalent to compute individual orbits of the dynamical systems generated by the Newton map $N_f(z)$.

The equation (1) on a polynomial $p(z)$

$$N_p(z) = z - \frac{p(z)}{p'(z)} \tag{2}$$

verifies the following properties:

1. The roots of $p(z)$ correspond to the finite fixed points of N_p.
2. The point at the infinity is a repelling fixed point.
3. As the derivative of the iteration function is

$$N_p'(z) = \frac{p(z)\,p''(z)}{p'(z)^2} \tag{3}$$

the simple roots of $p(z)$ are superattracting fixed points. Multiple roots are attracting fixed points, but not superattracting.

For simplicity, we begin by studying the Newton's method on quadratic polynomials. It is known that the roots of the polynomial can be transformed by an affine map without qualitatively changing the dynamics of the correspondent Newton function. So, we can use the quadratic polynomial $p(z) = z^2 + c$.

Then, we obtain that the two basin of attraction of the roots are separated by the perpendicular bisector of the line segment from one root to the other. This bisector is the Julia set for this polynomial, that it is connected.

Moreover, it is desirable that the convergence regions of both maps be essentially the same, except for the change of coordinates, see [10] for example.

Theorem 1 *Let f be an analytic function on the Riemann sphere, and let $A(z) = \alpha z + \beta$, with $\alpha \neq 0$, be an affine map. If $g(z) = \lambda(f \circ A)(z)$, where $\lambda \in \mathbb{C} - \{0\}$, then the Newton's iteration function N_f is analytically conjugated to N_g by A:*

$$A \circ N_f \circ A^{-1}(z) = N_g(z).$$

Moreover,

Theorem 2 *[4, Th. 2] Let $p(z)$ be a quadratic polynomial with distinct roots. The Newton's method $N_p(z)$ is globally, analytically conjugate to the quadratic polynomial z^2.*

P. Blanchard, in [4], proves it by considering the conjugacy map

$$h(z) = \frac{z - i\sqrt{c}}{z + i\sqrt{c}}, \tag{4}$$

with the following properties:

 ISBN: 978-84-614-6167-7

i) $h(\infty) = 1$,

ii) $h(i\sqrt{c}) = 0$,

iii) $h(-i\sqrt{c}) = \infty$.

Then,

$$\left(h \circ N_p \circ h^{-1}\right)(z) = z^2. \tag{5}$$

So, for quadratic polynomials, the Newton operator is always conjugate to the rational map z^2, satisfying the following properties:

1. The dynamics of this operator gives the unit circle $S^1(z) = \{z \in \hat{\mathbb{C}} \ : \ |z| = 1\}$ as the invariant Julia set.

2. The Fatou set is defined by the two basins of attraction of the superattracting fixed points: 0 and ∞.

In this paper, we study the dynamics of two iterative methods of order three: the Potra-Pták's method (Section 2) and the MidPoint method (Section 3) on quadratic polynomials.

2 The Potra-Pták's method

The iterative method of Potra-Pták (see [12]) is:

$$x_{k+1} = y_k - \frac{f(y_k)}{f'(x_k)},$$

where y_k comes from the Newton's method. The fixed point operator associated to this method on the complex plane

$$T_f(z) = z - \frac{f(z)}{f'(z)} - \frac{f\left(z - \frac{f(z)}{f'(z)}\right)}{f'(z)}. \tag{6}$$

In this section, we study the dynamics of this operator on quadratic polynomials, $p(z) = z^2 + c$, $c \in \mathbb{C}$,

$$T_p(z) = \frac{3z^4 - 6cz^2 - c^2}{8z^3}. \tag{7}$$

The Scaling Theorem for this method can be found in [3].

Theorem 3 *(Amat et.al., [3]) Let f be an analytic function on the Riemann sphere, and let $A(z) = \alpha z + \beta$, with $\alpha \neq 0$, be an affine map. If $g(z) = \lambda(f \circ A)(z)$, where $\lambda \in \mathbb{C} - \{0\}$, then the Potra-Pták's iteration function T_f is analytically conjugated to T_g by A:*

$$A \circ T_f \circ A^{-1}(z) = T_g(z).$$

 ISBN: 978-84-614-6167-7

We obtain four fixed points for this operator: two of them are the roots of the polynomial. The other two are called *strange fixed points.*

$$T_p(z) = z \Rightarrow z = \pm i\sqrt{c}, \pm i\sqrt{\frac{c}{5}}$$

As each component of the Fatou set contains at least a critical point, the dynamical properties of a complex analytical functions are often determined for the dynamics of its critical points. In this case, the derivative of (7)

$$T_p'(z) = \frac{3}{8}\frac{\left(z^2+c\right)^2}{z^4},$$

allows us to deduce that the only critical points are the roots of the polynomial and these roots are also fixed points of the operator.

Moreover,

$$T_p'\left(\pm i\sqrt{c}\right) = 0$$

implies that these roots are superattractor fixed points. The other fixed points of $T_p(z)$ are repelling $\left(T_p'\left(\pm i\sqrt{\frac{c}{5}}\right) = 6\right)$; so, they are in the Julia set.

As in Newton's method, the Fatou set consists of the basins of attraction of the two roots of the polynomial. That means that this method never fails for quadratic polynomials when it is applied on open sets of the complex plane. The dynamical plane of the operator (7) is shown in the Figure 1.

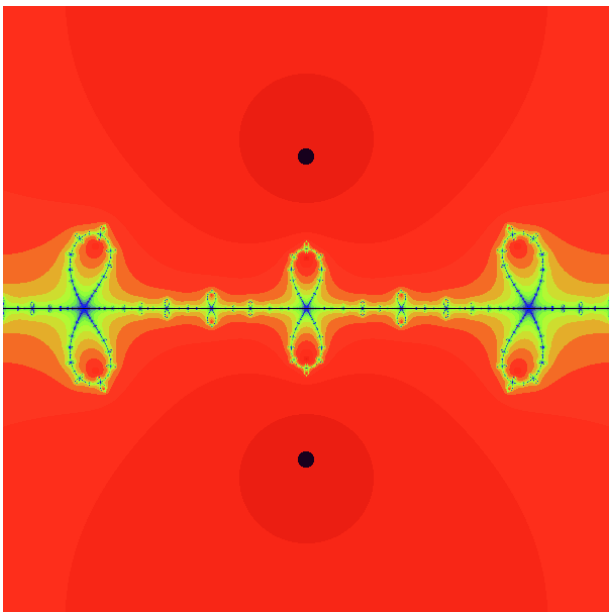


Figure 1: Dynamical plane of the Potra-Pták method on quadratic polynomials

 ISBN: 978-84-614-6167-7

From Theorem 2, we know that Newton's iteration function on any quadratic polynomial is conjugated to z^2. In this case, we prove that T_p on quadratic polynomials has a more complicated expression:

Theorem 4 *Let $p(z)$ be a quadratic polynomial with distinct roots. The fixed point operator associated to the Potra-Pták method $T_p(z)$ has the following properties:*

i) $T_p(z)$ is globally, analytically conjugate to the rational map $z^3\frac{z+2}{2z+1}$.

ii) The dynamics of this operator gives the unit circle $S^1(z) = \{z \in \hat{\mathbb{C}} \ : \ |z| = 1\}$ as the invariant Julia set. Moreover, this set is connected.

iii) The Fatou set is defined by the two basins of attraction of the superattracting fixed points: 0 and ∞.

Proof. As before, we consider the conjugacy map

$$h(z) = \frac{z - i\sqrt{c}}{z + i\sqrt{c}} \tag{8}$$

which has the same properties:

a) $h(\infty) = 1$

b) $h(i\sqrt{c}) = 0$

c) $h(-i\sqrt{c}) = \infty$

So,

$$h^{-1}(z) = i\sqrt{c}\frac{z+1}{1-z}$$

and, therefore

$$\left(h \circ T_p \circ h^{-1}\right)(z) = z^3\frac{z+2}{2z+1} \tag{9}$$

is a rational map of degree three that has superattracting fixed points at 0 and ∞. As in the previous case, this map does not depends on the parameter c, therefore the Julia set is connected for every c. ■

Because of the rational part, the Julia set is more complicated that the unit circle obtained in the Newton's method; nevertheless, all the points in the unit circle belongs to the Julia set (see Figure 2). As above, the Julia set separates the two basins of attraction of the two superattractor fixed points: 0 and ∞.

 ISBN: 978-84-614-6167-7

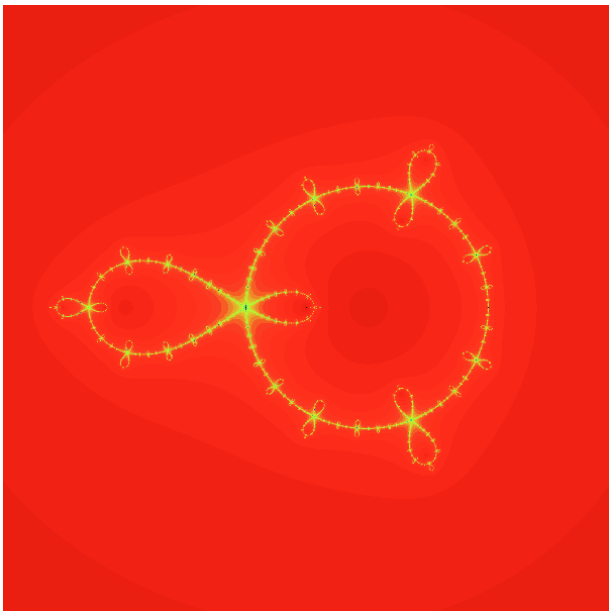


Figure 2: Dynamics of the conjugacy map for the Potra-Pták method on quadratic polynomials

3 The Midpoint Method

In this section we study the dynamics of the operator associated to the Midpoint Method (see [9]) on the complex plane

$$M_p(z) = z - \frac{f(z)}{f'\left(z - \frac{f(z)}{2f'(z)}\right)}, \tag{10}$$

for the same family of quadratic polynomials $p(z) = z^2 + c$, $c \in \mathbb{C}$.

$$M_f(z) = \frac{3cz - z^3}{c - 3z^2}. \tag{11}$$

Now, we state the Scaling Theorem for this iterative method.

Theorem 5 *Let f be an analytic function on the Riemann sphere, and let $A(z) = \alpha z + \beta$, with $\alpha \neq 0$, be an affine map. If $g(z) = \lambda(f \circ A)(z)$, where $\lambda \in \mathbb{C} - \{0\}$, then the Newton's iteration function M_f is analytically conjugated to M_g by A:*

$$A \circ M_f \circ A^{-1}(z) = M_g(z).$$

In this case, the fixed points are the roots of the polynomial:

$$M_f(z) = z \Rightarrow z = \pm i\sqrt{c},$$

 ISBN: 978-84-614-6167-7

which are also the critical points:

$$M_p'(z) = \frac{3\left(z^2+c\right)^2}{\left(c-3z^2\right)^2} = 0.$$

As in the previous cases, these roots are superattracting fixed points. As in Newton's method, the Fatou set consists of the basins of attraction of the two roots of the polynomial. That means that this method never fails for quadratic polynomials when it is applied on open sets of the complex plane. The dynamical plane of the operator (10) is the same as the corresponding of Newton's method.

This result is very well understood from the following result

Theorem 6 *Let $p(z)$ be a quadratic polynomial with distinct roots. The fixed point operator associated to the Midpoint method $M_p(z)$ has the following properties:*

i) $M_p(z)$ is globally, analytically conjugate to the cubic polynomial z^3.

ii) The dynamics of this operator gives the unit circle $S^1(z) = \{z \in \hat{\mathbb{C}} \ : \ |z| = 1\}$ as the invariant Julia set. Moreover, this set is connected.

iii) The Fatou set is defined by the two basins of attraction of the superattracting fixed points: 0 and ∞.

Proof. Similarly to the previous theorems, we consider the conjugacy map

$$h(z) = \frac{z - i\sqrt{c}}{z + i\sqrt{c}} \tag{12}$$

which have the same properties as in the previous theorem: $h(\infty) = 1$, $h(i\sqrt{c}) = 0$ and $h(-i\sqrt{c}) = \infty$. So,

$$h^{-1}(z) = i\sqrt{c}\frac{z+1}{1-z}$$

and, therefore

$$\left(h \circ M_p \circ h^{-1}\right)(z) = z^3 \tag{13}$$

is a cubic polynomial of degree three that has superattracting fixed points at 0 and ∞ separated by the unit circle. As before, this map does not depends on the parameter c, therefore the Julia set is connected for every c. ■

Moreover, (13) implies that the origin is a zero of order three.

4 Conclusions

Firstly, we would mention that we have obtained the conjugacy function of the Potra-Pták and Midpoint operators applied on quadratic polynomials.

Secondly, we note that the dynamical plane for the Midpoint operator on quadratic polynomials is the same that in the Newton case. This is explained from the fact that the two operators are conjugate to monomials of z (Theorems 2 and 6).

 ISBN: 978-84-614-6167-7

However, the Julia set in the dynamical plane for the Potra-Pták operator is more complicated. Theorem 4 explain this difference because it shows that this operator is conjugated to the product of a monomial and a rational function.

Acknowledgments: The authors thank to Professors X. Jarque and A. Garijo for their help. This research was supported by Ministerio de Ciencia y Tecnología MTM2010-18539 and by Vicerrectorado de Investigación, Universitat Politècnica de València PAID-06-2010-2285.

References

[1] S. Amat, C. Bermúdez, S. Busquier and S. Plaza, *On the dynamics of the Euler iterative function*, Applied Mathematics and Computation **197** (2008) 725–732.

[2] S. Amat, S. Busquier and S. Plaza, *A construction of attracting periodic orbits for some classical third-order iterative methods*, J. of Computational and Applied Math. **189** (2006) 22–33.

[3] S. Amat, S. Busquier and S. Plaza, *Chaotic dynamics of a third-order Newton-type method*, J. Math. Anal. Appl. **366** (2010) 24–32.

[4] P. Blanchard, *The Dynamics of Newton's Method*, Proc. of Symposia in Applied Math. **49** (1994) 139–154.

[5] J. Curry, L. Garnet and D. Sullivan, *On the iteration of a rational function: Computer experiments with Newton's method*, Comm. Math. Phys. **91** (1983) 267–277.

[6] A. Douady and J. H. Hubbard, *On the dynamics of polynomials-like mappings*, Ann. Sci. Ec. Norm. Sup. (Paris) **18** (1985) 287–343.

[7] N. Fallega. *Invariants en dinàmica complexa*, Butlletí de la Soc. Cat. de Matemàtiques **23**(1) (2008) 29–51.

[8] J. M. Gutiérrez, M. A. Hernández and N. Romero, *Dynamics of a new family of iterative processes for quadratic polynomials*, J. of Computational and Applied Math. **233** (2010) 2688–2695.

[9] A. Y. Ozban, *Some new variants of Newton's Method*, Applied Mathematics Letters, **17** (2004) 677–682.

[10] S. Plaza, *Conjugacy classes of some numerical methods*, Proyecciones (2001) 1-17.

[11] S. Plaza and N. Romero, *Attracting cycles for the relaxed Newton's method*, J. of Computational and Applied Math. **235** (2011) 3238–3244.

 ISBN: 978-84-614-6167-7

[12] F. A. Potra, V. Pták, *Nondiscrete introduction and iterative processes, Research Notes in Mathematics*, **103**, Pitman, Boston, 1984.

 ISBN: 978-84-614-6167-7

Proceedings of the 11th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2011
26–30 June 2011.

Filters method in direct search optimization, New measures to admissibility

Aldina Correia[1,2], João Matias[2], Pedro Mestre[3] and Carlos Serôdio[3]

[1] *CIICESI - Centro de Inovação e Investigação em Ciências Empresariais e Sistemas de Informação, ESTGF/IPP - Escola Superior de Tecnologia e Gestão de Felgueiras - Instituto Politécnico do Porto*

[2] *CM-UTAD - Centro de Matemática da UTAD, Universidade de Trás-os-Montes e Alto Douro*

[3] *CITAB - Centro de Investigação e de Tecnologias Agro-Ambientais e Biológicas, UTAD*

emails: aic@estgf.ipp.pt, j_matias@utad.pt, pmestre@utad.pt, cserodio@utad.pt

Abstract

Constrained nonlinear optimization problems are usually solved using penalty or barrier methods combined with unconstrained optimization methods.

Another alternative used to solve constrained nonlinear optimization problems is the filters method. Filters method, introduced by Fletcher and Leyffer in 2002, have been widely used in several areas of constrained nonlinear optimization. These methods treat optimization problem as bi-objective attempts to minimize the objective function and a continuous function that aggregates the constraint violation functions.

Audet and Dennis have presented the first filters method for derivative-free nonlinear programming, based on pattern search methods. Motivated by this work we have developed a new direct search method, based on simplex methods, for general constrained optimization, that combines the features of the simplex method and filters method.

This work presents a new variant of these methods which combines the filters method with other direct search methods and are proposed some alternatives to aggregate the constraint violation functions.

Key words: Constrained nonlinear optimization, Filters method, direct search methods

MSC 2000: 90C56; 90C30; 49M37; 65K05

 ISBN: 978-84-614-6167-7

1 Introduction

Let us consider a Constrained Nonlinear Programming Problem (NLP):

$$\begin{array}{ll} \min\limits_{x\in\mathbb{R}^n} & f(x) \\ s.t. & c_i(x) = 0, i \in \mathcal{E} \\ & c_i(x) \leq 0, i \in \mathcal{I} \end{array} \tag{1}$$

where:

- $f : \mathbb{R}^n \to \mathbb{R}$ is the objective function;
- $c_i(x) = 0,\ i \in \mathcal{E}$, with $\mathcal{E} = \{1, 2, ..., t\}$, define the Problem equality constraints;
- $c_i(x) \leq 0,\ i \in \mathcal{I}$, with $\mathcal{I} = \{t+1, t+2, ..., m\}$, represent the inequality constraints;
- $\Omega = \{x \in \mathbb{R}^n : c_i = 0,\ i \in \mathcal{E} \ \wedge\ c_i(x) \leq 0,\ i \in \mathcal{I}\}$ is the set of all feasible points, i.e., the feasible region.

In the resolution of a problem of this type we have two objectives: minimize the objective function f (Optimality) and minimize the constraints violation h, which must be 0 or tend to 0 (Viability).

If we consider the possibility of the objective function and/or the constraints functions of being non smooth, non continuous, non convex and/or with many local minimums then it is not possible to use derivative-based methods.

Derivative-based methods can be deterministic or heuristic. In this work we deal with deterministic direct search methods, i.e., methods which only need information about the objective and constraints functions values, in some points, and only use this information comparing these values to find next iteration.

The known direct search methods are unconstrained optimization methods and then we need some constrained techniques to treat the constraints information. The the most used techniques are based in penalty or barrier functions, and more recently the filter methods has proved to be effective to deal with the information given by the constraints.

Filters Method was introduced by Fletcher and Leyffer, [6]. Unlike penalty or barrier methods, the filters method considers the feasibility and optimality separately, using the concept of dominance of multiobjective optimization. A filters algorithm introduces a function that aggregates constraint violations and constructs a bi-objective problem attempts to minimize simultaneously that function (*feasibility*) and the objective function (*optimality*), giving priority to the feasibility at least until a feasible iterate is found.

First filters method for derivative-free nonlinear programming was presented by Audet and Dennis, [1]. This method is based on pattern search methods. Motivated by this work we have developed a method that combines the features of the simplex method and filters

 ISBN: 978-84-614-6167-7

method, [4, 3, 5]. The promising results that were obtained with this method encouraged the development of more features of the method, namely the combination of filters method with other direct search unconstrained optimization methods and the definition of other measures to aggregate the constraint violation functions.

2 General concepts

Filters method considers a nonnegative continuous function h which aggregates the constraint violation functions. Then h is a function such that $h(x) \geq 0$ with $h(x) = 0$ if and only if x is feasible.

This function is used in the definition of successive filters along the iterative process, because a point is accepted in a filter if and only if the point has better values of h or f than the points found so far.

A point $x \in \mathbb{R}^n$ is said to *dominate* $y \in \mathbb{R}^n$, denoted as $x \prec y$, if $f(x) \leq f(y)$ and $h(x) \leq h(y)$ or $f(x) < f(y)$ or $h(x) < h(y)$.

A *filter*, denoted by $\mathcal{F}$, is a finite set of points in the domain of f and h such that no point x in the set dominates other point y in the set, i.e., there is no pair of points x and y in the filter that have the relation $x \prec y$.

Figure 1, based in Ribeiro et. al. [7], illustrates the graphic concept of a filter with four points (a, b, c and d).

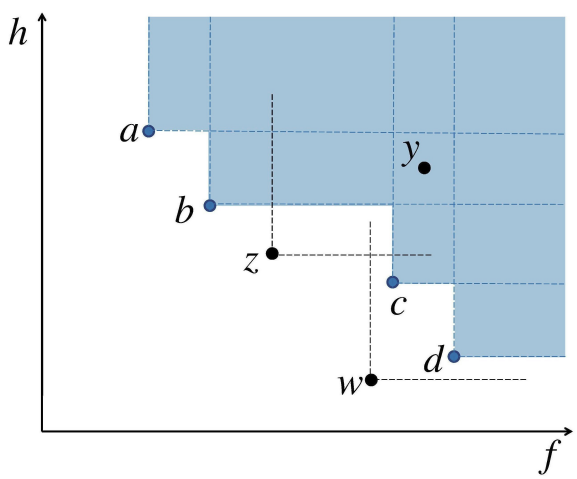


Figure 1: Filters Method - Graphic Concept

Points represented by a, b, c and d define a *forbidden region*, shaded area. Only points with better (lower) values of f and h should be in the filter, i.e. the aim is to have $h = 0$

 ISBN: 978-84-614-6167-7

and the lowest possible values for f. Therefore the point represented by y, as it is in the forbidden region, it will not be accepted in the filter. But the point represented by z is out of the forbidden region and therefore it will be included in the filter. The same applies to the point represented by w, however, in this case, the points represented by c and d would be eliminated of the filter, since they are in the forbidden region defined by w, i.e., c and d are dominated by w.

We consider that a point x is *filtered* by a filter $\mathcal{F}$ if:

- There exists a point $y \in \mathcal{F}$ such that $y \prec x$ or $y = x$;

- or $h(x) \geq h_{max}$;

- or $h(x) = 0$ and $f(x) \geq f^F$;

where f^F is the objective function value of the best feasible point found so far and h_{max} is a previously defined bound for h value, so each point $x \in \mathcal{F}$ satisfies $h(x) < h_{max}$.

3 Filters algorithm implemented

The algorithms presented in other works that use the filters method, define the procedures in a generic way, without spelling out clearly the steps of implementation, [1, 6, 8, 7]. Thus it was necessary to study the best way to do this implementation, so the work went through several phases and different versions have been implemented.

The first version tested was presented in [3] and more detailed in [4]. In this version of the filters method was implemented in combination with the Hooke and Jeeves method, a pattern search method as Audet and Dennis in [1], and in combination with Nelder-Mead method. These two combinations were compared and it was concluded that the second, which is called the Simplex Filter Algorithm (SFA), may be considered as an alternative for solving problems with nonlinear constraints and obtained satisfactory results.

In [5] some improvements were presented and a comparison was made of a New Simplex Filter Algorithm (NSFA) with the first version of the same method, SFA.

Numerical results obtained have motivated the generic implementation of filters method, i.e. so that it can be applied not only with Nelder-Mead and Hooke and Jeeves methods, in optimization of h and f, but also in other direct search type methods.

The procedure used to implement the NFSA in [5] is presented in Figure 2. Changes made in this work are adaptations and generalizations of this method, because the process is similar.

Adaptations are made in the designations. Instead of *Simplex* we have *Initial search set* and instead of *Simplex search* we use *Search set*. These designations are adapted to the generic character of the algorithm but both processes, the construction and the search, are done using the same procedure used in the previous algorithm. The generalizations are

 ISBN: 978-84-614-6167-7

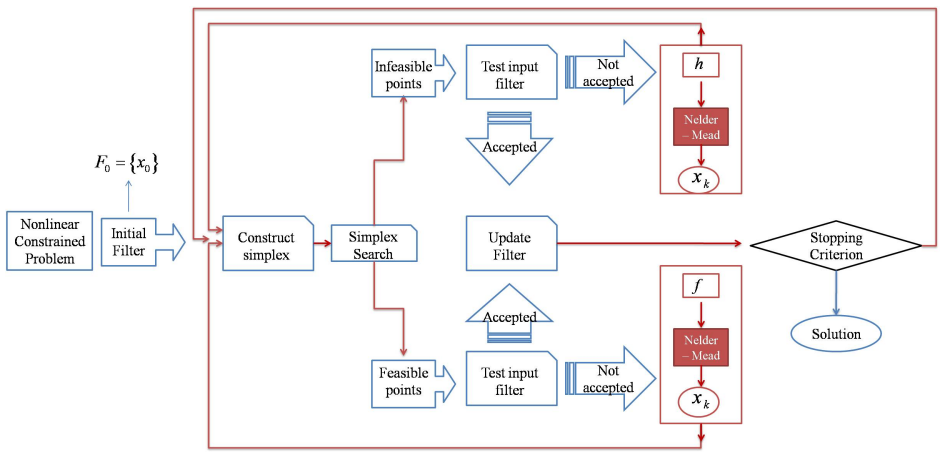


Figure 2: Implemented Algorithm in [5]

related to the internal process, i.e. the unconstrained optimization process of the previous algorithm was done exclusively using Hooke and Jeeves and Nelder-Mead algorithms and in this implementation five unconstrained direct search optimization algorithms are used.

 ISBN: 978-84-614-6167-7

Filters algorithm implemented

The procedure begins with an initial filter that contains the initial iteration, $F_0 = x_0$. *Then, it is constructed an initial set/simplex* (S_k) *containing* $n+1$ *points from that iteration* (x_k) *and:* $S_k = \{x_k\} \cup \{x_k + e_i, i = 1, ..., n\}$, *where* $e_i, i = 1, ..., n$ *represents the vectors of the canonic basis in* $\mathbb{R}^n$, *starting with the set/simplex search* $i = 0, ..., n$.

If the point, under analysis, is a feasible point its inclusion in the filter is evaluated:

- *If it is not accepted:*
  - *Nelder-Mead Method (or one of five unconstrained optimization methods) is applied to the function* $\boldsymbol{f}$;
  - *A new point is obtained,* x_k;
  - *Go back to the construction of the set/simplex:* $S_k = \{x_k\} \cup \{x_k + e_i, i = 1, ..., n\}$, *using* x_k;
- *If it is accepted:*
  - *Filter is updated with the new approximation to the solution, i.e., the new iteration;*
  - *If the stop criterion is verified, this approximation is the solution. Otherwise, go back to the set/simplex construction, using this point.*

If the vertex is an infeasible point, its inclusion in the filter is evaluated:

- *If it is not accepted:*
  - *Nelder-Mead Method (or one of five unconstrained optimization methods) is applied to the function* $\boldsymbol{h}$;
  - *A new point is obtained,* x_k;
  - *Go back to the to the construction of the set/simplex :* $S_k = \{x_k\} \cup \{x_k + e_i, i = 1, ..., n\}$, *using* x_k;
- *If it is accepted:*
  - *The filter is updated with the new approximation to the solution, i.e., the new iteration;*
  - *If the stop criterion is verified, this approximation is the solution. Otherwise, go back to the set/simplex construction, using this point.*

Thus, the method contains two distinct processes: the *external iterative process*, involving the set/simplex construction and the filter update and the *internal iterative process*, involving the optimization of f and h, where unconstrained optimization problems are solved, with objective functions f or h, using the Nelder-Mead Simplex method (or other unconstrained optimization method).

The NSFA also has a stronger component dedicated to the feasibility, when compared with SFA. It includes an unconstrained optimization process for h, when the set/simplex point being analyzed, in the set/simplex search, is infeasible. For feasible simplex points, it includes an unconstrained optimization process for f.

Furthermore, SFA only optimizes f (and not h) for infeasible points. And SFA, for feasible points, only verifies its admittance to the filter. If it is accepted, the point is added to the filter otherwise the Shrink step is applied. No optimization of h is then made in SFA, unlike NSFA.

 ISBN: 978-84-614-6167-7

With this generalization in the internal process methods availability, we can say that the method presented here is a direct search filters method for constrained optimization.

The five methods used in internal process are: Opportunistic Coordinate search method (CS); Hooke and Jeeves method (HJ); A version of Audet et. al. method (AA); Nelder-Mead method (NM) and a Convergent Simplex method (SC). The first three are Pattern Search Methods or Directional Direct-Search Methods. These methods determine possible points using fixed search directions during the iterative process, starting at an iteration x_k, the next iteration will be found in a pattern or grid of points, in the fixed directions, at a distance s_k, said step size.

The last two are Simplex Methods or Simplicial Direct-Search Methods. These methods are characterized by to construct an initial simplex and change the directions of search at each iteration, using reflection, expansion and contraction movements and shrunk steps.

4 Alternatives to aggregate the constraint violation functions

Considering equality constraints as two inequality constraints:

$$\begin{array}{rcl} c_i(x) = 0,\ \ i = 1, ..., t & \Leftrightarrow & c_i(x) \leq 0 \wedge c_i(x) \geq 0,\ \ i = 1, ..., t \\ & \Leftrightarrow & c_i(x) \leq 0 \wedge -c_i(x) \leq 0,\ \ i = 1, ..., t \end{array}$$

settle $2t + m = n$ and defining:

$$\left\{ \begin{array}{l} r_i(x) = c_i(x) \leq 0,\ \ i = 1, ..., t \\ r_j(x) = -c_i(x) \leq 0,\ \ i = 1, ..., t; j = t + 1, ..., 2t \\ r_j(x) = c_i(x) \leq 0,\ \ i = t + 1, ..., m; j = 2t + 1, ..., n \end{array} \right.$$

problem to solve is:

$$\begin{array}{ll} \min\limits_{x \in \mathbb{R}^n} & f(x) \\ s.t. & r_i(x) \leq 0, i = 1, ..., n \end{array} . \tag{2}$$

Usually norm 2 is used to define h (function that aggregate the constraint violation functions) i.e.:

$$h\left(x\right) = \left\| C_+\left(x\right) \right\|_2 = \sqrt{\sum_{i=1}^{n} \max\left(0, r_i\left(x\right)\right)^2}.$$

with

$$C_+\left(x\right) = \left\{ \begin{array}{cc} r_i\left(x\right) & \text{if} \quad r_i\left(x\right) > 0 \\ 0 & \text{if} \quad r_i\left(x\right) \leq 0 \end{array} \right. , \ \ i = 1, ..., n.$$

 ISBN: 978-84-614-6167-7

but the requirements for h only are: to be continuous and $h(x) \geq 0$ with $h(x) = 0$ if and only if x is feasible, i.e., h must be a non negative continuous function which $h(x) = 0$ if and only if x is feasible. Therefore we propose some alternatives to aggregate the constraint violation functions.

Penalty / Barrier methods penalizes (or implies the rejection) infeasible points, using penalty/barrier functions, which are measures for violation of constraints. These measures motivated the definition of some alternatives to aggregate the constraint violation functions, i.e., to define h. Some adaptations were made in the penalty or barrier functions expressions, since filters method does not require penalty/barrier parameters.

Then we can use some alternative measures to aggregate the constraint violation functions, i.e. to define h, Table 1.

Measure		h
Norm $1/\ell_1$ Penalty	**N1**	$h(x) = \|C_+(x)\|_1 = \sum_{i=1}^{n} \max[0,\ r_i(x)]$
Norm 2	**N2**	$h(x) = \|C_+(x)\|_2 = \sqrt{\sum_{i=1}^{n} \left\{\max[0,\ r_i(x)]^2\right\}}$
Extreme Barrier	**NEB**	$h(x) = \begin{cases} 0 & \text{if} \quad x \in \Omega \\ +\infty & \text{if} \quad x \notin \Omega \end{cases}$
Progressive Barrier Classic Penalty Static/Dynamic Penalty	**NP**	$h(x) = \sum_{i=1}^{n} \{\max[r_i(x),\ 0]\}^2$

Table 1: Alternatives to aggregate the constraint violation functions

5 Used parameters

In both processes, *internal* (Unconstrained Optimization - Direct Search Methods) and *external* (Constrained Optimization - Filters Method), it is necessary to choose some parameters. Parameters used are presented in Tables 2 and 3.

6 Numerical Results

Test Problems were selected from Schittkowski [9] and CUTE [2] collections. The fifteen Schittkowski problems are: S224; S225; S226; S227; S228; S231; S233; S234; S249; S264; S270; S323; S324; S325 and S326 and of Cute collection were chosen two test problems: C801 and C802. The last problem is the PA problem presented in [3].

The choice of these eighteen tests was not made in accordance with any special requirement, they are only used to illustrate the performance of the methods implemented.

 ISBN: 978-84-614-6167-7

Parameters	**Coordinate Search**	**Hooke-Jeeves**	**Audet**	**Nelder-Mead**	**Simplex Convergent**
k_{max}	100	100	100	100	100
s	1	1	*	1	1
s_m	*	*	1,5	*	*
s_p	*	*	1	*	*
s_{min}	10^{-3}	10^{-3}	10^{-3}	*	*
α	*	*	*	1	1
β	*	*	*	0,5	0,5
γ	*	*	*	2	2
T_1	10^{-5}	10^{-5}	10^{-5}	10^{-5}	10^{-5}
T_2	10^{-5}	10^{-5}	10^{-5}	10^{-5}	10^{-5}
T_{var}	*	*	*	10^{-5}	10^{-5}
T_{vol_n}	*	*	*	*	10^{-5}

k_{max} → Maximum number of iterations; s → Length of the initial step
s_m → Length of the initial mesh search step (Audet); s_p → Length of the initial poll step (Audet)
s → Length of the initial step; s_{min} → Minimum value for the step length
α → Reflexion parameter (Nelder-Mead); β → Contraction parameter (Nelder-Mead)
γ → Expansion parameter (Nelder-Mead)
$T1 = |x_k - x_{k+1}|$ → Tolerance for the distance between two consecutive iterations
or Tolerance for the distance between the last iteration and the latest iteration (Nelder-Mead)
$T2 = |f(x_k) - f(x_{k+1})|$ → Tolerance for
the distance between two values of the objective function in successive iterations
T_{var} → Tolerance to the variance of the objective function values in the vertices of the simplex (Simp. Conv.)
T_{vol_n} → Tolerance to the normalized volume of the simplex
$*$ → Parameter non used in the method

Table 2: Unconstrained Optimization - Direct Search Methods - Parameters used

$k_{max} = 40$ → Maximum number of iterations in the external process;
$\rho = 1$ → Initial search step length;
$T1 = |x_k - x_{k+1}| = 0.00001$ → tolerance for the distance between two consecutive iterations;
$T2 = |f(x_k) - f(x_{k+1})| = 0.00001$ →
Tolerance between 2 values of the objective function in two consecutive iterations;
$h_{max} = +\infty$ → Maximal valor of constraints violation.

Table 3: Constrained Optimization - Filters Method - Parameters used

 ISBN: 978-84-614-6167-7

In order to classify the solution approximations we define some criterions. Then, an approximation x_k to the problem solution is:

- a *Feasible solution approximation* if $h(x_k) = 0$, being:
  - *Good* if: $|f(x^*) - f(x_k)| \leq 0.0001$;
  - *Medium* if: $0.0001 < |f(x^*) - f(x_k)| \leq 0.01$;
  - *Bad* if: $|f(x^*) - f(x_k)| > 0.01$;
- an *Infeasible solution approximation* if $h(x_k) \neq 0$, being:
  - *Good* if: $|f(x^*) - f(x_k)| \leq 0.0001 \ \wedge \ |V(x_k)| \leq 0.0001$;
  - *Medium* if:
    - $0.0001 < |f(x^*) - f(x_k)| \leq 0.01 \ \wedge \ |V(x_k)| \leq 0.0001$;
    - ou $|f(x^*) - f(x_k)| \leq 0.0001 \ \wedge 0.0001 < |V(x_k)| \leq 0.01$;
    - ou $0.0001 < |f(x^*) - f(x_k)| \leq 0.01 \ \wedge 0.0001 < |V(x_k)| \leq 0.01$;
  - *Bad* if: $|f(x^*) - f(x_k)| > 0.01 \ \vee \ |V(x_k)| > 0.01$.

Using these criterions we classified the solution approximations obtained.

In the Table 4 are presented the numerical results obtained in the resolution of the 18 problems, using the filters method (FM) and parameters presented in Tables 2 and 3. No medium solution approximations were found in the resolutions.

We can see in Table 4 that numerical results are similar for all direct search methods. What stands out most is the HJ method with which it was possible to find 16 *Good* approximations to the admissible solution of test problems (5 with the measure N1, 5 with the measure N2, 2 with the measure NEB and 4 with the measure NP).

For these 18 test problems, the methods and combinations of measures that have proven most effective were the HJ method with measures N1, N2 and NP. In all three cases the percentage of *Bad* approaches feasible and infeasible is greater than the percentage of *Good* approximations.

With the NEB measure we cannot find any infeasible approach and the number of problems for which it was possible to find feasible approaches is reduced.

Therefore one can not specify which is the best algorithm for all problems, so that, in the presence of a problem to be solved using these methods/algorithms should be tested all the algorithms, choosing the best approach to the solution.

7 Conclusion

We can conclude that it is possible to use other methods of direct search in combination with the filter method, since the efficiency of each method is similar to the other.

 ISBN: 978-84-614-6167-7

Methods		Solution Approximation							
		Feasible				Infeasible			
IP	EP	Good	Bad	Total	%	Good	Bad	Total	%
CS	N1	2	10	12	66,7%		10	10	55,6%
	N2	2	10	12	66,7%	1	9	10	55,6%
	NEB	1	9	10	55,6%			0	0,0%
	NP	2	10	12	66,7%	1	9	10	55,6%
HJ	N1	5	8	**13**	**72,2%**	1	12	**13**	**72,2%**
	N2	5	7	**12**	**66,7%**	1	12	**13**	**72,2%**
	NEB	2	8	**10**	**55,6%**			**0**	**0,0%**
	NP	4	7	**11**	**61,1%**	1	11	**12**	**66,7%**
AA	N1	1	11	12	66,7%		12	12	66,7%
	N2	1	13	14	77,8%		11	11	61,1%
	NEB	2	8	10	55,6%			0	0,0%
	NP	1	13	14	77,8%		13	13	72,2%
NM	N1	2	9	**11**	**61,1%**		10	**10**	**55,6%**
	N2	2	10	**12**	**66,7%**		10	**10**	**55,6%**
	NEB	2	7	**9**	**50,0%**			**0**	**0,0%**
	NP	2	10	**12**	**66,7%**		10	**10**	**55,6%**
SC	N1	2	12	14	77,8%		9	9	50,0%
	N2	2	12	14	77,8%		9	9	50,0%
	NEB	2	7	9	50,0%			0	0,0%
	NP	2	12	14	77,8%		9	9	50,0%

Table 4: Numerical Results obtained with MF algorithm, with default parameters

 ISBN: 978-84-614-6167-7

We can also conclude that the proposed measures for aggregate the constraint violation functions are as efficient as the usual measure N2, therefore they represent alternatives measures for constraints violation.

Thus, the suggestions in this work are another alternative for solving constrained problems without using derivatives of the functions involved or their approximations.

Acknowledgements

This work has been partially supported by:

- CIICESI - Centro de Inovação e Investigação em Ciências Empresariais e Sistemas de Informação, ESTGF/IPP - Escola Superior de Tecnologia e Gestão de Felgueiras - Instituto Politécnico do Porto.
- CM-UTAD - Centro de Matemática da UTAD, Universidade de Trás-os-Montes e Alto Douro

References

[1] C. Audet and J. Dennis, *A pattern search filter method for nonlinear programming without derivatives*, SIAM Journal on Optimization, **(14):5** (2004) 980–1010.

[2] I. Bongartz and A. Conn and N. Gould and P. Toint, *CUTE: Constrained and Unconstrained Testing Environment*, ACM Transactions and Mathematical Software, **21** (1995) 123–160.

[3] A. Correia and J. Matias and C. Serôdio, *Derivative-free optimization and Filter Methods for solve Nonlinear Constrained Problems*, Proceedings of the 2008 International Conference on Computational and Mathematical Methods in Science and Engineering (CMMSE), ISBN: 978-84-612-1982-7 (2008) 161–172.

[4] A. Correia and J. Matias and P. Mestre and C. Serôdio, *Derivative-Free Optimization and Filter Methods to Solve Nonlinear Constrained Problems*, International Journal of Computer Mathematics, Taylor & Francis Ltd, **(86):10** (2009) 1841–1851.

[5] A. Correia and J. Matias and P. Mestre and C. Serôdio, *Derivative-free Nonlinear Optimization Filter Simplex*, International Journal of Applied Mathematics and Computer Science (AMCS), University of Zielona Góra and Lubuskie Scientific Society in Zielona Góra, Poland, **(20):4** (2010) 679–688.

[6] R. Fletcher and S. Leyffer, *Nonlinear programming without a penalty function*, Mathematical Programming, Ser. A, **(91): 2** (2002) 239–269.

 ISBN: 978-84-614-6167-7

[7] A. Ribeiro and E. Karas and C. Gonzaga, *Global Convergence of Filter Methods for Nonlinear Programming*, SIAM J. on Optimization, **(19): 3** (2008) 1231–1249.

[8] E. Karas, *Exemplos de trajetória central mal comportada em otimização convexa e um algoritmo de filtros para programação não linear*, PhD. Thesis, Universidade Federal de Santa Catarina e Universidade de Paris I - Panthéon-Sorbonne, Florianópolis, Brasil, 2002.

[9] K. Schittkowski, *More Test Examples for Nonlinear Programming Codes, Economics and Mathematical Systems*, Springer-Verlag, Berlin, 1987.

 ISBN: 978-84-614-6167-7

Proceedings of the 11th International Conference on Computational and Mathematical Methods in Science and Engineering, CMMSE 2011 26–30 June 2011.

Line graphs for directed and undirected networks: An structural and analytical comparison

Regino Criado[1], Julio Flores[1], Alejandro García del Amo[1] and Miguel Romance[1]

[1] *Departamento de Matemática Aplicada, Universidad Rey Juan Carlos, Móstoles, Spain*

emails: regino.criado@urjc.es, julio.flores@urjc.es, alejandro.garciadelamo@urjc.es, miguel.romance@urjc.es

Abstract

The centrality and efficiency measures of a network G are strongly related to the respective measures on the associated line graph $L(G)$ and bipartite graph $B(G)$ as was shown in [8]. In this note we consider different ways to obtain a line graph from a given directed or undirected network and we obtain some interesting relations. *Key words: complex networks; dual graph; line graph; line digraph.*

Many relevant properties of complex systems in the real world, such as social networks, Internet, the World Wide Web and other biological and technological systems may be described in terms of complex network properties [1, 3, 4, 5, 13, 10, 16, 17, 18, 20, 21, 22, 24]. In fact, the study of structural properties of complex networks is an attractive and fascinating branch of research in sociology (social networks, acquaintances or collaborations between individuals), science (metabolic and protein networks, neural networks, genetic regulatory networks, protein folding) and technology (Internet, computers in telecommunication networks, the World Wide Web,...).

The motivation behind this contribution is to consider the importance that edges have sometimes over nodes in the context of networks and graphs. An example of this comes from urbanism [11, 12], transport networks [25, 2] or urban traffic [19], where the line (dual) graph $L(G)$ associated to a given graph (network) G is considered.

For example, in the context of urban traffic, when the underlying (primal) graph is considered then intersections (or settlements) are seen as nodes while roads (or lines of

 ISBN: 978-84-614-6167-7

relationship) are seen as edges. In contrast when the dual (line) graph is considered roads become nodes, while intersections become links between the corresponding nodes [19].

In [8] we showed some relationships between the network's efficiency and the network's Bonacich ([6], [7]) centrality of a network G and the respective measures on the dual $L(G)$ and the bipartite $B(G)$ associated networks (see below for definitions). Some other properties and relationships between the centrality of a network G and the centrality of its dual network $L(G)$ have been studied in [9]. Note that the networks considered there were undirected.

The main goal of this note is to exhibit some relations arising from the various ways in which line graphs can be obtained from a given directed network. This in turn can be applied to obtaining estimations for several parameters that measure different properties related to the network structure and performance.

In order to investigate such properties, it is necessary to understand the main structure of the underlying network [3, 20] and also to consider other complementary topological aspects.

From a schematic point of view, a complex network is a mathematical object $G = (V, E)$ composed by a set of nodes or vertices $V = \{v_1 \ldots, v_n\}$ that are pairwise joined by links or edges $\{l_1, \ldots, l_m\}$. We consider the adjacency matrix $A(G) = (a_{ij})$ determined by the conditions

$$a_{ij} = \begin{cases} 1 & \text{if} \quad \{\mathrm{v_i}, \mathrm{v_j}\} \in \mathrm{E} \\ 0 & \text{if} \quad \{\mathrm{v_i}, \mathrm{v_j}\} \notin \mathrm{E}. \end{cases}$$

The bipartite network $B(G)$ associated to G is defined by $B(G) = (X \cup E, E(B(G)))$ whose adjacency matrix is given by

$$A(B(G)) = \left(\begin{array}{c|c} 0 & I(G) \\ \hline I(G)^t & 0 \end{array} \right)$$

where $I(G)$ is the incidence matrix of G. It is shown that

$$A(B(G))^2 = \left(\begin{array}{c|c} A(G) + gr & 0 \\ \hline 0 & A(L(G)) + 2I_n \end{array} \right)$$

where $A(G) + gr$ denotes the matrix obtained by adding to $A(G)$ the diagonal matrix (b_{ij}) and b_{ii} is the degree of the vertex v_i while $L(G)$ denotes the line (or dual) network associated to G ([14], pag. 26, [15], pag. 273).

As we showed in [8], if we know the Bonacich centrality $c(L(G))$, we can recover $c(B(G))$ and reciprocally. If, in addition, G is regular then each of the three centralities can be recovered from any of the other two. Moreover, we have a relationship between the efficiencies of the dual graph $L(G)$ and the primal graph G (see [8]):

 ISBN: 978-84-614-6167-7

If $G=(V,E)$ and $L(G)=(E,L)$, n is the number of nodes of G, m is the number of nodes $L(G)$ and p is the number of edges of $L(G)$, we have

$$\frac{n(n-1)}{8m(m-1)}E(G)+\frac{15p-2}{8m(m-1)}\leq E(L(G))\leq \max_{i\neq j}(gr_igr_j)\frac{n(n-1)}{m(m-1)}E(G)+\frac{2p}{m(m-1)}.$$

These results have potential interest in the context of urban streets networks ([11, 12]) and urban traffic ([19]).

If now $A(G)=(a_{ij})$ is the adjacency matrix of the directed network G determined by the conditions

$$a_{ij}=\left\{\begin{array}{ll} 1 & \text{if} \quad (\mathrm{v_i},\mathrm{v_j})\in \mathrm{E} \\ 0 & \text{if} \quad (\mathrm{v_i},\mathrm{v_j})\notin \mathrm{E}, \end{array}\right.$$

the bipartite network $B(G)$ associated to G is defined by $B(G)=(X\cup E, E(B(G)))$ whose adjacency matrix is given by

$$A(B(G))=\left(\begin{array}{c|c} 0 & H(G) \\ \hline T(G)^t & 0 \end{array}\right)$$

where $H=H(G)$ is the incidence matrix of heads of G defined by

$$H_{ij}=\left\{\begin{array}{ll} 1 & \text{if} \quad \mathrm{e_j}=(\mathrm{v_i},-) \\ 0 & \text{otherwise} \end{array}\right.$$

and $T=T(G)$ is the incidence matrix of tails of G defined by

$$T_{ij}=\left\{\begin{array}{ll} 1 & \text{if} \quad \mathrm{e_j}=(-,\mathrm{v_i}) \\ 0 & \text{otherwise} \end{array}\right.$$

it is shown that

$$A(B(G))^2=\left(\begin{array}{c|c} A(G) & 0 \\ \hline 0 & A(\vec{L}(G)) \end{array}\right)$$

where $\vec{L}(G)$ denotes the line (or dual) network associated to G.
Recall that the Bonacich centrality of a complex network G is the non-negative normalized eigenvector $c_G\in\mathbb{R}^n$ associated to the spectral radius of the transposed adjacency matrix of G [6, 7, 20]. The following relations between the Bonacich centralities of G, $\vec{L}(G)$ and $B(G)$ are obtained:

Theorem 1 *Let $G=(V,E)$ be a connected directed graph with n vertices and m edges. Let $c_G\in\mathbb{R}^n$, $c_{\vec{L}(G)}\in\mathbb{R}^m$ and $c_{B(G)}=(c_1,c_2)\in\mathbb{R}^n\times\mathbb{R}^m$ be the Bonacich centralities of G, $\vec{L}(G)$ and $B(G)$. Then, if $\|v\|_1=\sum_{i=1}^{n}|v_i|$ for any arbitrary $v=(v_1,\ldots,v_n)\in\mathbb{R}^n$, we have:*

 ISBN: 978-84-614-6167-7

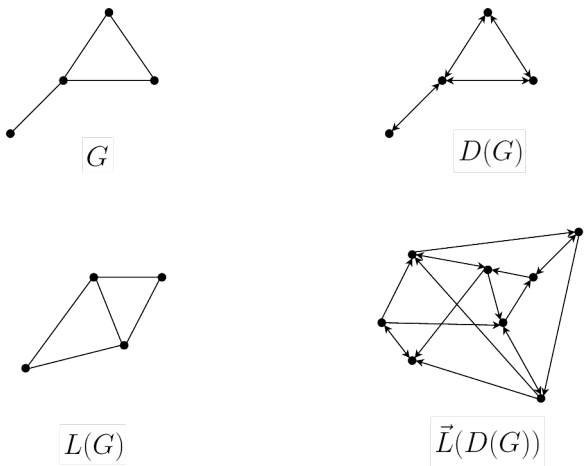


Figure 1: An example of the line-graph of a undirected network (on the left) and for a directed network (on the right).

(i) $c_G = \frac{c_1}{\|c_1\|_1}$ *and* $c_{\vec{L}(G)} = \frac{c_2}{\|c_2\|_1}$.

(ii) Reciprocally, $c_{B(G)} = \frac{1}{2}\,(c_G, c_{\vec{L}(G)})$ *and*

$$c_G = \frac{H_G\, c_{\vec{L}(G)}}{\|H_G\, c_{\vec{L}(G)}\|_1}, \quad c_{\vec{L}(G)} = \frac{T_G^t\, c_G}{\|T_G^t\, c_G\|_1}\,.$$

Let $D(G)$ denote the associated symmetric digraph obtained by replacing each edge of G by an arc pair in which the two arcs are inverse to each other. Since $A(G) = A(D(G))$, the Bonacich centralities of G and $D(G)$ coincide and, in particular, the Bonacich centralities of G and $\vec{L}(D(G))$ are closely related.

There is an alternative definition for the line graph associated with G that has received relatively little attention. We will call it the oriented line graph $L^{\rightsquigarrow}(G)$ and it will be defined as follows. If $D(G) = (V(D(G)), E(D(G)))$ denotes the associated symmetric digraph, the vertices of the oriented line graph $L^{\rightsquigarrow}(G)$ are the arcs $E(D(G))$ of $D(G)$, while (e, f) is an arc in $L^{\rightsquigarrow}(G)$ if the end of e coincides with the origin of f and f is not the inverse of e. is not the inverse of e (see [23] and the references cited therein). In the same reference [23] the oriented line graph $L^{\rightsquigarrow}(G)$ is employed to capture graph-class structure and clustering graphs.

 ISBN: 978-84-614-6167-7

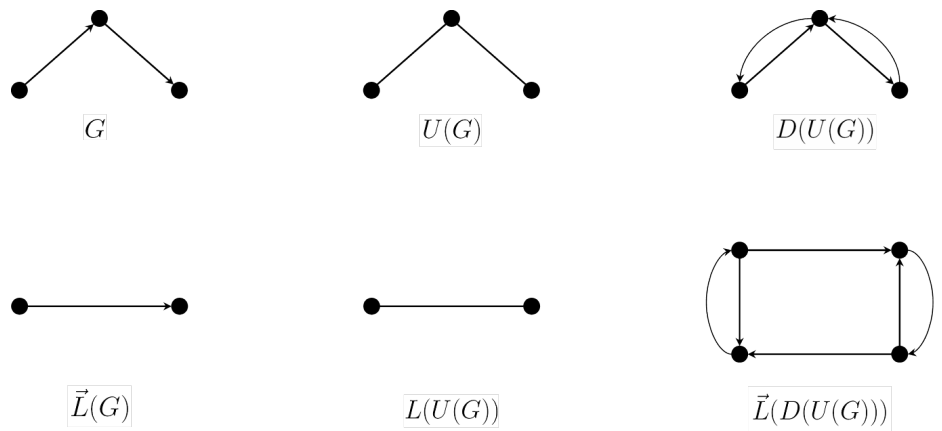


Figure 2: An example of line graphs of a directed network G, $U(G)$ and $D(U(G))$.

Acknowledgements

This work has been supported by the Spanish Government Project MTM2009-13848.

References

[1] J. Agarwal, D.I. Blockley and N.J. Woodman, "Vulnerability of structural systems". Structural Safety **25**, 263286 (2003)

[2] J. Anez, T. De La Barra and B. Perez, "Dual graph representation of transport networks", Trans. Res. B,**30**, 3, (1996) 209-216.

[3] R. Albert and A. L. Barabási, "Statistical mechanics of complex networks", *Rev. Mod. Phys.* ***74*** (2002), 47–97.

[4] Y. Bar-Yam, *Dynamics of Complex Systems*, Addison-Wesley, 1997.

[5] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, D.-U. Hwang, "Complex networks: Structure and dynamics", *Physics Reports* ***424*** (2006), 175–308.

[6] P. Bonacich, "Factoring and weighing approaches to status scores and clique information", *J. Math. Soc.* ***2*** (1972),113.

[7] P. Bonacich and P. Lloyd, "Eigenvectors-like measures of centrality for asymetric relations", *Soc. Netw.* ***23*** (2001),191.

 ISBN: 978-84-614-6167-7

[8] R. Criado, J. Flores, A. García del Amo, M. Romance, "Analytical Relationships between metric and centrality measures of a network and its dual" *JCAM* ***235***(2011)1775-1780.

[9] R. Criado, J. Flores, A. García del Amo, M. Romance, " Centrality and measure of irregularity ", preprint.

[10] P. Crucitti, V. Latora, M. Marchiori, A. Rapisarda, "Efficiency of Scale-Free Networks: Error and Attack Tolerance" *Physica A* ***320***(2003)622.

[11] P. Crucitti, V. Latora, S. Porta, "Centrality in networks of urban streets", *Chaos.* ***16***(2006)015113.

[12] P. Crucitti, V. Latora, S. Porta, "Centrality Measures in Spatial Networks of Urban Streets", *Phys. Rev. E* ***73***(2006)036125.

[13] Fontoura Costa, L. et al, "Analyzing and Modeling Real-World Phenomena with Complex Networks: A Survey of Applications". arXiv:0711.3199v3 [physics.soc-ph] (2008)

[14] CJ. L. Gross, J. Yellen (eds.), *Handbook of graph theory*, CRC Press, New Jersey, 2004.

[15] R. L. Hemminger and L. W. Beineke, *Line graphs and line digraphs, Selected Topics in Graph Theory (W. B. Lowell and R. J. Wilson, eds.)*, Academic Press, New York, 1978, pp. 271305.

[16] V. Latora, M. Marchiori, "Efficient Behavior of Small-World Networks",*Phys. Rev. Lett.* ***87***(2001)198701

[17] V. Latora, M. Marchiori, "How the science of complex networks can help developing strategies against terrorism" *Chaos, Solitons and Fractals* ***20***(2004)69

[18] V. Latora, M. Marchiori, "A measure of centrality based on the network efficiency", *New J. Phys.* ***9***(2007)188

[19] M-B. Hua, R. Jianga, R. Wang, Q-S. Wu, "Urban traffic simulated from the dual representation: Flow, crisis and congestion " *Physics Letters A* ***373***(2009)2007–2011.

[20] M. E. J. Newman, "The structure and function of complex networks", *SIAM Review* ***45*** (2003), 167–256.

[21] M.E.J. Newman *Networks: An Introduction* Oxford Univ. Press, Oxford, 2010

 ISBN: 978-84-614-6167-7

[22] M.E.J. Newman, , A.L. Barabási, D.J. Watts, , *The Structure and Dynamics of Networks*, Princeton Univ. Press, Princeton, New Jersey (2006)

[23] P. Ren, R. C. Wilson, E.R. Hancock, "Graph Characterization via Ihara Coefficients", *IEEE Transactions on Neural Networks* ***22 (2)*** (2011), 233–245.

[24] S. H. Strogatz, "Exploring complex networks", *Nature* ***410*** (2001), 268–276.

[25] Volchenkov, D. and Blanchard, Ph., "Transport Networks Revisited: Why Dual Graphs?". arXiv:0710.5494v1 [physics.soc-ph] 29 Oct 2007

 ISBN: 978-84-614-6167-7

Proceedings of the 11th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2011
26–30 June 2011.

Modeling Chagas Disease and Control Measures

Gustavo Cruz-Pacheco[1], Lourdes Esteva[2] and Cristobal Vargas[3]

[1] *Instituto de Investigaciones en Matemticas Aplicadas y Sistemas, Universidad Nacional Autónoma de México*

[2] *Facultad de Ciencias, Universidad Nacional Autónoma de México*

[3] *Departamento de Control Automático, Centro de Investigación y de Estudios Avanzados-IPN*

emails: cruz@mym.iimas.unam.mx, lesteva@lya.fciencias.unam.mx, cvargas@ctrl.cinvestav.mx

Abstract

Chagas disease, also known as American trypanosomiasis, is a potentially life-threatening illness caused by the protozoan parasite, Trypanosoma cruzi (T. cruzi). The main mode of transmission of Chagas disease in endemic areas is through an insect vector called a triatomine bug. A triatomine becomes infected with T. cruzi by feeding on the blood of an infected person or animal. Chagas disease is consider the most important vector borne infection in Latin America. It is estimated that between 16 to 15 millions of persons are infected with T. cruzi, with about 50,000 deaths each year.

In this work we formulate a model to study the transmission of this illness among the vector, humans and some mammals. Our main objective is to assess the effectiveness of some control measures for the infection. We attain this through a sensitivity analysis of the basic reproductive number R_0 with respect to the epidemiological and demographic parameters.

Key words: Chagas, Trypanosoma cruzi, triatomines, basic reproductive number, control measures

MSC 2000: AMS 92D30

1 Introduction

Chagas disease, also known as American trypanosomiasis, is a potentially life-threatening illness caused by the protozoan parasite, Trypanosoma cruzi (T. cruzi). The main mode

 ISBN: 978-84-614-6167-7

of transmission of Chagas disease in endemic areas is through the bite of an insect vector called a triatomine bug. Because they tend to feed on peoples faces, triatomine bugs are also known as *kissing bugs* [4].

Chagas disease may also be spread through blood transfusion and organ transplantation, ingestion of food contaminated with parasites, and from a mother to her fetus. The rate of trans-placental transmission from mothers with chronic T cruzi infection to their newborns is 2-10% [5].

Chagas disease is considered the most important vector borne infection in Latin America. It is estimated that between 16 to 18 millions of persons are infected with T. cruzi, with about 50,000 deaths each year [6].

In the early, stage of the infection, acute stage, the symptoms are mild and usually produce no more than local swelling at the site of infection, usually around the eyes in children. Anti parasitic treatment with benznidazol and nifurtimox in the acute phase may result in cure rates between 60 and 90 percent. There is an asymptomatic middle stage in which the infection can not be detected at all, even blood test results are negative. The length of this period is not well determined. Usually, after 48 weeks, individuals with active infections enter the chronic phase of Chagas disease that is asymptomatic for 60 to 80 percent of chronically infected individuals through their lifetime. The anti parasitic treatments also appear to delay or prevent the development of disease symptoms during the chronic phase of the disease, but 20 to 40 percent of chronically infected individuals will still eventually develop life-threatening heart and digestive system disorders [1].

Control measures include insecticides to kill the vector, screening blood donors, and treatment to patients in the acute phase. Recently, a controversial strategy, Zooprophylaxis , has been proposed for the control of vector transmitted diseases [8]. This controversial technique refers to the control of vector-borne diseases by attracting vectors to domestic animals in which the pathogen cannot amplify (a dead-end host).

2 Formulation of the model

We consider transmission by triatomine bites, and vertical transmission since these are the most common routes of infection.

We will consider the following populations:

- Humans
- Transmitters: mammals that can be infected by the Triatomine bugs, and can transmit the infection (like dogs, cats, etc.)
- Non-transmitters: animals that can be bitten by the Triatomine bugs, but can not be infected, and in consequence do not transmit the infection (like hens, birds, etc.)

 ISBN: 978-84-614-6167-7

- Vectors: Triatomine bugs

The infected human population is divided into infected humans in the acute phase, I_a, and infected humans in the chronic phase, I_c. The infected transmitters, and infected vector populations are denoted by I_t, and I_v, respectively.

The dynamics of the disease is modeled by the following system of differential equation

$$\begin{aligned}
\frac{dI_a}{dt} &= p\mu_h I_c + \frac{b_h\beta_h}{N_h+N_t+N_{nt}}(N_h - I_a - I_c)I_v - (\gamma+\mu_h)I_a \\
\frac{dI_c}{dt} &= (1-q)\gamma I_a - \mu_h I_c \\
\frac{dI_t}{dt} &= \frac{b_h r_t \beta_t}{N_h+N_t+N_{nt}}(N_t - I_t)I_v - \mu_t I_t \\
\frac{dI_v}{dt} &= \frac{b_h\alpha_a I_a + b_h\alpha_c I_c + b_h r_t \alpha_t I_t}{N_h+N_t+N_{nt}}(N_v - I_v) - \mu_v I_v
\end{aligned} \tag{1}$$

in the invariant region

$$\Omega = \{0 \le I_a + I_c \le N_h, 0 \le I_t \le N_t, 0 \le I_v \le N_v\} \subset R^4.$$

In the model N_h, N_t, N_{nt}, N_v denote the population sizes of humans, transmitters, non-transmitters, and vectors, respectively. Since the non-transmitters population do not enter in the infection process, we consider N_{nt} as a parameter rather than a dynamic variable. The other populations are constant with mortality rates given by μ_h, μ_t, and μ_v, respectively.

The parameters b_h, b_t, and b_{nt} represent the number of triatomine bites per day in humans, transmitters, and non-transmitters, respectively; β_a, and β_t the transmission probabilities from vector to susceptible humans and susceptible transmitters, respectively; and α_a, α_c, α_t the transmission probabilities from acute infective humans, chronic infective humans, and infective transmitters to susceptible vectors.

We assume that a proportion p of newborns from chronic infected humans are acute infected. Finally, we assume that acute infectious become chronic infectious at a per capita rate γ, this quantity is diminished by $q\gamma$ where q is the proportion of acute infectious that are treated and return to the susceptible class.

Triatomines can be considered as a predator of mammals and birds since they feed on these species to maintain themselves and reproduce. Then, growth of triatomine population depends upon the number of blood meals they take, their species preference, and the number of individuals of each species. We assume that the dynamics of the triatomine bugs is given by

$$\frac{dN_v}{dt} = \phi(b_h N_h + b_t N_t + b_{nt} N_{nt}) - \mu_v N_v \tag{2}$$

 ISBN: 978-84-614-6167-7

where ϕ denotes the egg-production rate from blood meals rate. It is easy to see that solutions N_v of this equation approach the equilibrium

$$\bar{N}_v = \frac{\phi(b_h N_h + b_t N_t + b_{nt} N_{nt})}{\mu_v}. \tag{3}$$

as t goes to infinity.

3 Mathematical analysis of the Model

3.1 Disease-free equilibrium

Definition. The *Basic Reproductive Number* denoted by R_0, is the average number of secondary infections caused by an infected individual in a whole susceptible population during the infection period.

The equilibrium $E_0 = (0, 0, 0, 0)$ of system (1) is called the *disease-free state*. Using the next generation operator approach ([7]), we obtain the basic reproductive number in terms of the epidemiological and demographic parameters as

$$R_0 = \sqrt{\frac{b_h^2 \beta_h N_h \bar{N}_v}{K \mu_v N^2} \left(\alpha_a + \frac{(1-q)\gamma}{\mu_h} \alpha_c \right) + \frac{b_t^2 \beta_t \alpha_t N_t \bar{N}_v}{\mu_v \mu_t N^2}}. \tag{4}$$

Hence, using Theorem 2 of [7], the following result is established.

Theorem. The disease-free equilibrium, E_0, of the model (1) is locally-asymptotically stable (LAS) if $R_0 < 1$, and unstable if $R_0 > 1$.

The above theorem shows that Chagas disease disappears if $R_0 < 1$, i.e., if the secondary cases derived from an infected individual are less than one.

R_0 can be written as

$$R_0 = \sqrt{R_h^2 \frac{N_h}{N} + R_t^2 \frac{N_t}{N}} \tag{5}$$

where

$$R_h = \sqrt{\frac{b_h^2 \beta_h \bar{N}_v}{K \mu_v N} \left(\alpha_a + \frac{(1-q)\gamma}{\mu_h} \alpha_c \right)} \tag{6}$$

is the number of secondary infections derived from an infected individual in the human-vector cycle, and

$$R_t = \sqrt{\frac{b_t^2 \beta_t \alpha_t \bar{N}_v}{\mu_v \mu_t N^2}} \tag{7}$$

 ISBN: 978-84-614-6167-7

is the number of secondary infections derived from an infected individual in the transmitters-vector cycle.

It is interesting to observe that R_0 given in (5) is the average of the basic reproductive numbers of the humans and transmitters, weighted by their corresponding population proportions with respect to the total number of the vector hosts, N. Then, the balance between the competence of humans and transmitters, and their corresponding population density will determine the evolution of the disease.

3.2 Endemic state

An endemic state is a non trivial solution of the algebraic system obtained setting the derivatives of (1) equal to zero. We have the following result.

Theorem 2. System (1) has a unique endemic state $E_1 = (\bar{I}_a, \bar{I}_c, \bar{I}_t, \bar{I}_v)$ in Ω if and only if $R_0 > 1$. Furthermore, E_1 is locally asymptotically stable.

4 Sensitivity Analysis

In this section we analyze the effect of the control measures on the spread of Chagas disease. The parameter values used in the simulations given in Table 1.

parameter	meaning	value
b_f	daily blood feeding rate per triatomine	0.26
b_h	number of triatomine bites per day in humans	0.04
b_t	number of triatomine bites per day in transmitters	0.1
b_{nt}	number of triatomine bites per day in non-transmitters	0.12
β_h	infection prob. vector-human	0.0009
β_t	infection prob. vector-transmitter	0.0009
α_a	infection prob. acute human-vector	0.04
α_c	infection prob. chronic human-vector	0.0025
α_t	infection prob. transmitter-vector	0.49
μ_h	humans mortality	0.000042
μ_t	transmitters mortality	0.00027
μ_{nt}	non-transmiters mortality	0.00039
μ_v	triatomines mortality	0.005
p	percentage of human trasplacental transmission	$2\% - 10\%$
γ	rate of acute infectious that become chronic	0.0178-0.0357

Table 1. Parameters of the model. The values are taken from [1, 3].

Measures to control Chagas disease are limited, among them are reducing the population of vectors, the early treatment of the disease, and the use of no transmitters to reduce the

 ISBN: 978-84-614-6167-7

number of bites in humans.

Periodic use of insecticides is the usual way to reduce the vector population. In terms of our model, this control translates into increasing vector mortality. If we denote by $\bar{\mu}_v = \theta\mu_v$, $\theta > 1$ the increment of the vector mortality due to insecticide application, then R_0 decreases by a factor $1/\sqrt{\theta}$.

Assuming two dogs, three chicken and five humans in a household, $N_t = 2N_h/5$, $N_{nt} = 3N_h/5$, and $N = 2N_h$. If $N_v = 100N_h$, $p = 0.06$, $q = 0.4$, $\gamma = 0.02675$, and the other parameters as in Table 1, $R_h = 2$, $R_t = 12$, and $R_0 = 6$. Disease can be eradicated if $1/\sqrt{\theta} \times 6 < 1$, or $36 < \theta$, which means than vector mortality should be increased more than 36 times. Figures 1-3 illustrate the behavior of R_0, R_h, and R_t when μ_v is incremented by the factor θ. We note that these basic reproductive numbers decrease faster for lower values of θ, suggesting that is more efficient to spray a smaller amount of insecticide at greater frequency.

Other way to control Chagas disease is by identification and treatment of the acute cases. In our model, this control is achieved increasing the proportion of acute infectious that are treated and cured, q. To assess the effect of this parameter on R_0, we assume as in the previous case that it increases by a factor $\theta > 1$, with $\theta q \leq 1$. Then, denoting $\bar{q} = \eta q$, and expanding R_0 in terms of q we obtain

$$\begin{aligned} R_0(\bar{q}) &\approx R_0(q) + \frac{\partial R_0}{\partial q}(1-\theta)q \\ &= R_0(q)\left[1 - \frac{Q}{2R_0^2(q)}\right](\theta - 1)q \end{aligned} \tag{8}$$

where

$$Q = R_h^2 \gamma \frac{(\mu_h\alpha_a + (1 + (1-q)p\gamma)\alpha_c)}{K(\mu_h\alpha_a + (1-q)\gamma\alpha_c)} \tag{9}$$

From the above expression we see that the decreasing of R_0 is more pronounced when R_h big. However, since treatment only reduce acute infected humans, the minimum value that R_0 can achieve by this control is around $R_0\left[1 - \frac{Q}{2R_0^2(q)}\right](1-q)$ when $\theta = 1/q$. In the example illustrated in Figure 1, R_0 can be reduced from 5.88 to only 5.71 given the initial values of $q = 0.04$, $\mu_v = 0.005$, and other parameters as in Figure 1, R_0 can be reduced from 5.88 to 5.71. In this particular example this small reduction is due to the fact that transmission animal-vector is more important than transmission human-vector. On the other hand, Figure 2 shows that R_h decreases form its initial value of around 1.8 to around 0.5, which indicates that human treatment is a very effective control when the transmission is mainly between humans and vectors.

Since 1982, the World Health Organization has recommended the use of animals for zooprophylaxis as a protective measure against vector borne diseases, in particular the use of cattle in the case of malaria [8]. In the following we will test the efectiveness of

 ISBN: 978-84-614-6167-7

zooprophylaxis for the case of Chagas disease. For this end, we assume that the population of non transmitters N_{nt} increases to $\bar{N}_{nt} = \theta N_{nt}$, with $\theta > 1$, and as in the previous example, we expand R_0 in Taylor series in terms of N_{nt} to obtain

$$R_0(\bar{N}_{nt}) \approx R_0(N_{nt}) \left[1 + \frac{\Phi}{2N\bar{N}_v} \left((b_{nt} - 2b_h)N_h + (b_{nt} - 2b_t)N_t - b_{nt}N_{nt}\right)\right]. \quad (10)$$

R_0 has a unique local maximum when

$$N_{nt} = \frac{(b_{nt} - 2b_h)N_h + (b_{nt} - 2b_t)N_t}{b_{nt}}, \quad (11)$$

which is positive if $(b_{nt} - 2b_h)N_h + (b_{nt} - 2b_t)N_t > 0$, and negative if the opposite inequality holds. It follows that depending on the density, and biting rate of the three populations involved, R_0 increases or decreases. For instance, if the biting rates satisfy $\frac{b_{nt}}{2} \leq \min\{b_h, b_t\}$, R_0 always decreases monotonically to zero, but if $\frac{b_{nt}}{2} > \max\{b_h, b_t\}$, R_0 increases to a maximum value and then decreases asymptotically to zero. Decreasing of R_0 can be very slow, at it is shown in the example of Figure 1, where it is necessary to increase the initial value of $N_{nt} = (3/5)N_h$ around 140 times to get $R_0 < 1$. In this particular example $b_{nt} = 0.06$, and the other parameters are as in Table 1, which implies that R_0 always decreases.

Figures 2 and 3 show the behavior of R_h, and R_t with respect to θ. It can easily show that when N_{nt} grows without bound, R_h and R_t always approach a limit value different from zero, which can be bigger or lower than the initial one. In the example illustrate in Figure 1, R_h, and R_t remain almost constant with a very small increasing from their initial values.

5 Conclusions

In this work we formulated a mathematical model to study the impact of control measures in the transmission of Chagas diseases. Here, we consider vector and vertical transmission, but we did not include blood transmission [2] Our results can be summarized as:

- Elimination of Triatomines is the control measure that has more impact on the diminishing of R_0.
- Treatment of disease in the acute phase is effective if people is isolated from other transmitters.
- Zooprophylaxis has little impact on the reduction of the infection transmission, and in some cases can even increases it.

 ISBN: 978-84-614-6167-7

- A combination of elimination of Triatomines, early treatment, and keeping the transmitters animals out of the houses can be considered the most effective control measure.

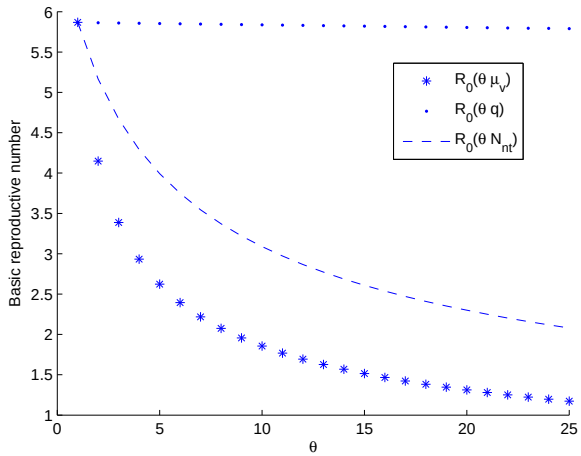


Figure 1:

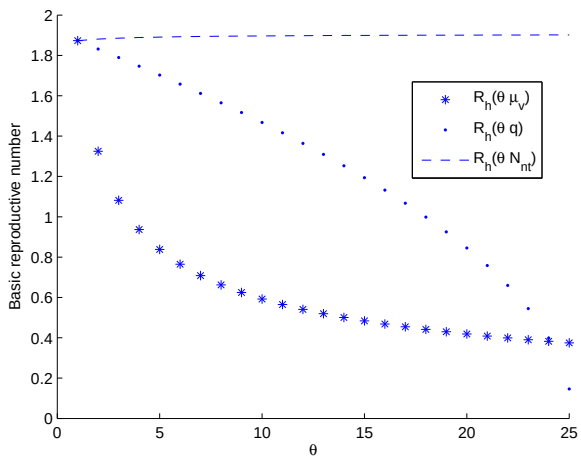


Figure 2:

Acknowledgements

This work has been partially supported by project PAPIIT IN105110-3 from Universidad Nacional Autónoma de México.

 ISBN: 978-84-614-6167-7

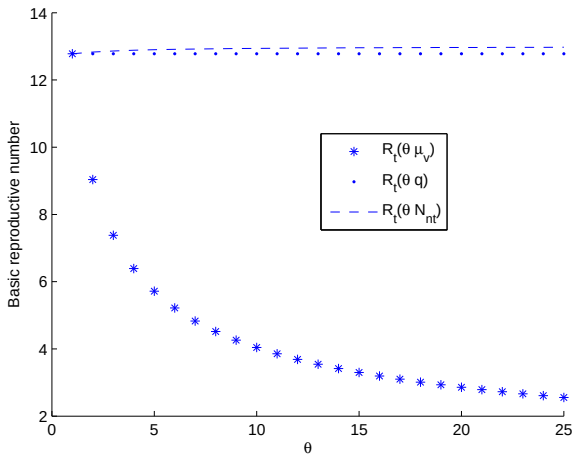


Figure 3:

References

[1] C. Bern, S. P. Montgomery, B. L. Herwaldt et al., *Evaluation and treatment of Chagas disease in the United States: a systematic review*, JAMA **298** (2007) 2171-2181.

[2] S. N. Busemberg, C. Vargas, *Modeling Chagas disease: variable population size and demographic implications* in Mathematical Population Dynamics, O. Arino, D. E. Axelrod and M. Kimmel, eds. Dekker, New York 1991 pp. 283–295.

[3] J. E. Cohen and R. E. Gurtler, *Modeling Household Transmission of American Trypanosomiasis*, Science **293** (2001) 694–698.

[4] *DPDxTrypanosomiasis*, American. Fact Sheet. Centers for Disease Control. http://www.dpd.cdc.gov/dpdx/HTML/TrypanosomiasisAmerican.htm.

[5] L. V. Kirchhoff, *Chagas Disease (American Trypanosomiasis)*, eMedicine http://emedicine.medscape.com/article/214581-overview, 2010.

[6] F. Torrico et al., *Maternal Trypanosoma cruzi infection, pregnancy outcome, morbidity and mortality of congenitally infected and non-infected newborns in Bolivia*, Am. J. of Trop. Med. and Hyg. **70** (2004) 201–209.

[7] P. van den Driessche and J. Watmough, *Reproduction numbers and sub-threshold endemic equilibria for compartmental models of disease transmission*, Math. Biosci. **180** (2002) 29–48.

[8] *WHO: Manual on environmental management for mosquito control with special emphasis on mosquito vectors.* WHO offset publication **66** Geneva, 1982.

 ISBN: 978-84-614-6167-7

Proceedings of the 11th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2011
26–30 June 2011.

Stability of numerical methods applied to families of stable linear systems

Gilner de la Hera Martínez[1], J. Vigo Aguiar[1] and M. T. de Bustos-Muñoz[1]

[1] *Departamento de Matemática Aplicada, Facultad de Ciencias, Universidad de Salamanca*

emails: gilner@usal.es , jvigo@usal.es, tbustos@usal.es

Abstract

Given a numerical method and a family of stable linear systems, we find conditions that guarantee the existence of a positive real number $0 < h^* \leq \infty$ such that for each h, $0 < h < h^*$, the numerical solutions of those systems converge to zero.

Key words: A-stability, Linear differential equations, Runge-Kutta methods, robustness, stability radius.

1 Introduction and preliminary results.

Let $A \in \mathbb{R}^{n\times n}$ be a Hurwitz stable matrix, that is to say the set of its eigenvalues $\sigma(A)$ is included in $\mathbb{C}^- = \{z \in \mathbb{C} : \ Re(z) < 0\}$ or equivalently that the solutions $x(t)$, of the system

$$\dot{x} = Ax \tag{1}$$

verifies $\lim_{t\to 0} x(t) = 0$.

Let $V \subset \mathbb{R}^{n\times n}$ be a compact balanced set with $0 \in V$, associated with system (1) we have the set of systems

$$\dot{x} = [A + M]\, x \tag{2}$$

where $M \in V^\rho$, $\rho \in \mathbb{R}_+$ and $V^\rho = \rho \cdot V = \{\rho G : \ G \in V\}$.

We define the real stability radius

$$r_{\mathbb{R}} \equiv r_{\mathbb{R}}(A, V) := inf\left\{\rho > 0 : \ \exists M \in V^\rho,\ \sigma\left(A + M\right) \nsubseteq \mathbb{C}^-\right\}$$

That is to say, if $M \in V^\rho$ and $0 < \rho < r_{\mathbb{R}}$ the solucions of the perturbed system (2) converge to zero too.

 ISBN: 978-84-614-6167-7

Example 1. *Let $A \in \mathbb{R}^{2\times 2}$ be a Hurwitz stable matrix and*

$$V = \left\{ \sum_{i,j=1,2} g_{ij} B_{ij} : \sum_{i,j=1,2} |g_{ij}| \leq 1 \right\}$$

where $\{B_{ij}\}_{i,j=1,2}$ is the canonical basis of $\mathbb{R}^{2\times 2}$. It is easy to see that V is the closed ball in $\mathbb{R}^{2\times 2}$ centred at the zero matrix and with radius $r = 1$ and considering the norm $\|G\|_m = \sum_{i,j} |g_{ij}|$, $G = (g_{ij})_{ij} \in \mathbb{R}^{2\times 2}$. Similarly V^ρ, $\rho > 0$, is the closed ball in $\mathbb{R}^{2\times 2}$ centred at the zero matrix and with radius ρ. Under these conditions we know that

$$r_{\mathbb{R}}(A, V) = min \left\{ \frac{det(A)}{\|A\|_m}, -tr(A) \right\} \tag{3}$$

Example 2. *Let $A = \begin{pmatrix} -2 & 3 \\ -4 & -1 \end{pmatrix} \in \mathbb{R}^{2\times 2}$ and V given in example (1), then we have $r_{\mathbb{R}}(A, V) = 3$. This means that each solution of all systems:*

$$\dot{x} = \begin{pmatrix} -2 + g_{11} & 3 + g_{12} \\ -4 + g_{21} & -1 + g_{22} \end{pmatrix} x$$

with $\sum_{i,j} |g_{ij}| \leq \rho < 3$, converge to zero.

Our objetive is to solve the following

Problem 1. *Given a numerical method, we want find conditions that guarantee the existence of a positive real number $0 < h^* \leq \infty$ such that for each h, $0 < h < h^*$, the numerical solutions of all the systems (2):*

$$\dot{x} = (A + M)x$$

converge to zero when $M \in V^\rho$ and $0 < \rho < r_{\mathbb{R}}$.

2 Main results.

The region of absolute stability, $\Re_A$, of numerical methods have been discussed in various papers and books, see [2]. For example, the region of absolute stability of a s-stage Runge-Kutta method is $\Re_A = \{z \in \mathbb{C} : |R(z)| \leq 1\}$ where

$$R(z) = \frac{det(I_n - z\mathcal{A} + ze^T b)}{det(I - z\mathcal{A})} \tag{4}$$

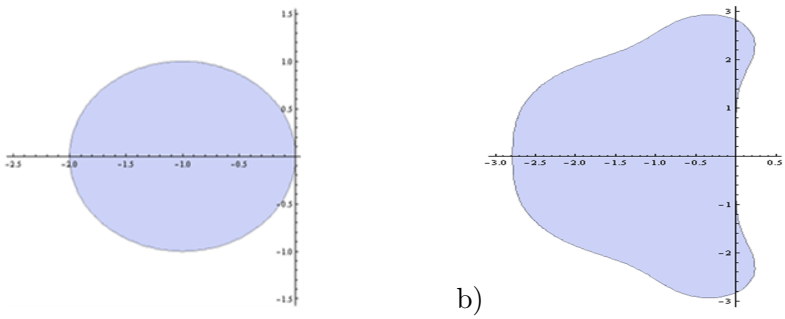
is called stability function and $\mathcal{A} \in \mathbb{R}^{s\times s}$ and $b \in \mathbb{R}^s$ determine its Butcher array, $e = (1, \ldots, 1) \in \mathbb{R}^s$.

 ISBN: 978-84-614-6167-7

The stability function $R(z)$ of a Runge-Kutta method of order p is a rational and analytic function on a neighborhood of $z = 0$ with Taylor expansion

$$R(z) = \sum_{k=0}^{p} \frac{z^k}{k!} + \sum_{k=p+1}^{\infty} \gamma_k z^k \tag{5}$$

In the case of explicit Runge-Kutta methods with s-stage, we have $p \leq s$ and $\gamma_{s+j} = 0$, $\forall j \geq 1$. Figure (1) we can see regions for explicit Euler methods and explicit Runge-Kutta fourth order and $four$-stage, here $R(z) = 1 + z$ and $R(z) = 1 + z + \frac{z^2}{2} + \frac{z^3}{6} + \frac{z^4}{24}$ respectively.



a) b)

Figure 1: Regions of absolute stability, $\Re_A$.
a) The region of absolute stability of explicit Euler method does not satisfy the property that ensures that h^* exists. ($z = 0$ is not an interior point of $\Re_A \cup \{0\}$).
b) Region of absolute stability of Runge Kutta 4th order method, for this method exists h^*. ($z = 0$ is an interior point of $\Re_A \cup \{0\}$).

Theorem 1. *Let $A \in \mathbb{R}^{n\times n}$ be a Hurwitz stable matrix and $V \subset \mathbb{R}^{n\times n}$ is a balanced compact set with $0 \in V$. Let us consider a numerical method with absolute stability region $\Re_A$. If*

(H) $z = 0$ is an interior point of $\Re_A \cup \{0\}$ in the space $\overline{\mathbb{C}^-}$ with the induced topology

then it must exist $h^* \in \overline{\mathbb{R}}$, $h^* > 0$, such that for each step size h, $0 < h < h^*$,numerical solutions of all systems (2):

$$\dot{x} = (A + M)x$$

converge to zero, where $M \in V^\rho$ and $0 < \rho < r_{\mathbb{R}}$.

Proof. Given $\rho > 0$, let $F^\rho = \sigma(A + V^\rho)$ be the set of eigenvalues of the matrices $A + M$ with $M \in V^\rho$ and let $F = \bigcup_{\rho<r_{\mathbb{R}}} F^\rho$. Since V is balanced we have $V^\rho \subset V^{\rho'}$ for $0 < \rho < \rho'$. Now, by continuity of $G \to \sigma(G)$, for each $\rho > 0$, F^ρ is a compact set with $F^\rho \subset \overline{\mathbb{C}^-}$ and $F^\rho \subset F^{r_{\mathbb{R}}}$ if $\rho < r_{\mathbb{R}}$. Then F is bounded and satisfies that $F \subset \overline{F} \subset \overline{\mathbb{C}^-}$, and so $\Re_A \cup \{0\}$ is a neighborhood of $z = 0$ in the space $\overline{\mathbb{C}^-}$, then we can ensure that there is $h > 0$ such that hF is contained entirely in $\Re_A \cup \{0\}$. □

 ISBN: 978-84-614-6167-7

Corollary 1. *If* $F = \bigcup_{\rho<r_{\mathbb{R}}} \sigma(A+V^{\rho})$ *then* $h^* = Sup\{h/hF \subseteq \Re_A \cup \{0\}\}$.

Remark 1. Theorem (1) is valid for any numerical method.

Theorem 2. *Consider a Runge-Kutta method with s-stages and order* $p \geq 1$ *and with stability funtion given by (5). Then, this method satisfies the hypothesis (H) from Theorem (1) if:*

- $\tau_m = (-1)^{(m+1)/2}\left[\frac{1}{(m+1)!} - \gamma_{m+1}\right] > 0$ *if* m *is odd.*
- $\tau_m = (-1)^{m/2}\left[\frac{m+1}{(m+2)!} + \gamma_{m+2} - \gamma_{m+1}\right] > 0$ *if* m *is even.*

where $m \in \mathbb{N}$ *is such that* $m \geq p$, $\gamma_j = \frac{1}{j!}$ *if* $p \leq j \leq m$ *and* $\gamma_{m+1} \neq \frac{1}{(m+1)!}$. *The converse is also true if we put* $\tau_m \geq 0$ *when* m *is even.*

The proof of theorem (2) makes use of the following preliminary result:

Lemma 1. Let $E_m(z)$, $m \in \mathbb{N}$, be the Taylor polynomial of degree m generated by $f(z) = e^z$ at the point $z = 0$. Consider an explicit Runge-Kutta method with s-stages and order p, then if $R(z)$ is its stability function and m as in Theorem (2) we have:

1. $R(z) = E_m(z) + \Gamma(z)$ where $\Gamma(z) = \sum_{k=m+1}^{\infty} \gamma_k z^k$.
2. If $z = iy$ then $E_m(iy) = C_m(y) + iS_m(y)$ where $C_m(y)$ and $S_m(y)$ are the Taylor polynomials, of degree m, generated by $cos(y)$ and $sin(y)$ respectively.
3. $|E_m(iy)|^2 - 1 = \alpha y^{m'} + o\left(y^{m'}\right)$ where $m' = m+2$ and $\alpha = \frac{2(-1)^{(m+2)/2}(m+1)}{(m+2)!}$ if m is even, $m' = m+1$ and $\alpha = \frac{2(-1)^{(m+1)/2}}{(m+1)!}$ if m is odd.

Proof. Statements (1) and (2) are immediate. Now, we shall prove (3). From (2) we have

$$|E_{m+2}(iy)|^2 - 1 = C_{m+2}^2(y) + S_{m+2}^2(y) - 1$$

where the right side of this equation is a polynomial, which until the $(m+2)th$ power coincides with the Taylor polynomial of the funtion $cos^2(y) + sin^2(y) - 1$ and equals zero. On the other hand

$$\begin{aligned}
|E_{m+2}(iy)|^2 - 1 =& C_{m+2}^2(y) + S_{m+2}^2(y) - 1 \\
&= \left[C_m(y) + \frac{(-1)^{\frac{m+2}{2}} y^{m+2}}{(m+2)!}\right]^2 + \left[S_m(y) + \frac{(-1)^{\frac{m}{2}} y^{m+1}}{(m+1)!}\right]^2 - 1 \\
&= \left[C_m^2(y) + S_m^2(y) - 1\right] + 2\frac{(-1)^{\frac{m+2}{2}} y^{m+2}}{(m+2)!} + \frac{(-1)^{\frac{m}{2}} y^{m+2}}{(m+1)!} + o\left(y^{m+2}\right) \\
&= \left[|E_m(iy)|^2 - 1\right] + \frac{2(-1)^{(m+2)/2}(m+1)}{(m+2)!} y^{m+2} + o\left(y^{m+2}\right)
\end{aligned}$$

 ISBN: 978-84-614-6167-7

the last equality is obtained using that $C_m(y) = 1 + ...$ y $S_m(y) = y + ...$ Then, isolate the $|E_m(iy)|^2 - 1$ term the result is obtained. A similar argument proves the statement when m is odd. □

Now, we shall prove the Theorem (2)

Proof. From Lemma (1) (1)-(2) we have

$$|R(iy)|^2 - 1 = [C_m(y) + \Gamma_1(y)]^2 + [S_m(y) + \Gamma_2(y)]^2 - 1$$

where $\Gamma(iy) = \Gamma_1(y) + i\Gamma_2(y)$, then

$$|R(iy)|^2 - 1 = \left(C_m^2(y) + S_m^2(y) - 1\right) + 2C_m(y)\Gamma_1(y) + 2S_m(y)\Gamma_2(y) + \left(\Gamma_1^2(y) + \Gamma_2^2(y)\right)$$

If p is even, then

$$\Gamma_1(y) = (-1)^{\frac{(m+2)}{2}}\gamma_{m+2}y^{m+2} + o\left(y^{m+2}\right) \qquad \Gamma_2(y) = (-1)^{\frac{m}{2}}\gamma_{p+1}y^{m+1} + o\left(y^{m+1}\right)$$

and from lemma (1)-(3) we have

$$\begin{aligned} |R(iy)|^2 - 1 &= \frac{2(-1)^{\frac{(m+2)}{2}}(m+1)}{(m+2)!}y^{m+2} + 2(-1)^{\frac{(m+2)}{2}}\gamma_{m+2}y^{m+2} \\ &+ 2(-1)^{\frac{m}{2}}\gamma_{m+1}y^{m+2} + o\left(y^{m+2}\right) = -2\tau_m y^{m+2} + o\left(y^{m+2}\right) \end{aligned} \tag{6}$$

Therefore, for y small enough if $\tau_m > 0$ then $|R(iy)|^2 - 1 < 0$ and $iy \in \Re_A$.

Now, we write $z = x + iy$ where $x \leq 0$. Using $a^k - b^k = (a-b)\sum_{j=1}^{k} a^{k-1-j}b^{j-1}$ and (5) we have

$$R(z) - R(iy) = \sum_{k=1}^{\infty}\gamma_k\left[(x+iy)^k - (iy)^k\right] = x\left[\gamma_1 + \sum_{k=2}^{\infty}\sum_{j=1}^{k}\gamma_k\,(x+iy)^{k-j}\,(iy)^{j-1}\right]$$

where $\gamma_j = \frac{1}{j!}$ if $0 \leq j \leq m$. Thus

$$R(z) = R(iy) + x\left[\gamma_1 + \sum_{k=2}^{\infty}\sum_{j=1}^{k}\gamma_k\,(x+iy)^{k-j}\,(iy)^{j-1}\right]$$

Taking the real and imaginary part, we can show that

$$|R(z)|^2 - 1 = \left(|R(iy)|^2 - 1\right) + (2\gamma_1 + f(x,y))\,x$$

 ISBN: 978-84-614-6167-7

where $f(x,y)$ is a real function with $\lim_{z\to 0} f(x,y) = 0$. Hence, for $z = x + iy \neq 0$, $x \leq 0$, with $|z|$ small enough and $\tau_m > 0$ we have

$$|R_s(z)|^2 - 1 \leq \Big(|R_s(iy)|^2 - 1\Big) < 0$$

then $z \in \Re_A$.

Conversely, suppose that $\tau_m < 0$ and y small enough, then from (6) we have $|R(iy)|^2 - 1 > 0$ and $iy \notin \Re_A$. In this case $z = 0$ is not an interior point of $\Re_A \cup \{0\}$ in the space $\overline{\mathbb{C}^-}$ and (H) is not satisfied. Therefore $\tau_m \geq 0$, but if m is odd then $\tau_m \neq 0$, because $\frac{1}{(m+1)!} \neq \gamma_{m+1}$. □

Example 3. *Let a explicit Runge-Kutta method with s-stages and order p:*

1. *If $p = 1$ or $p = 2$ with $p = s$ then $m = p = s$ and $R(z) = \sum_{k=0}^{p} \frac{z^k}{k!}$. Therefore the property (H) is not satisfied, because from (5) we have $\gamma_{p+1} = \gamma_{p+2} = 0$ and the result is obtained from Theorem (2).*

2. *If $p = 3$ or $p = 4$ with $p = s$ then the property (H) is satisfied, therefore h^* exist.*

3. *If this method is DOPRI5, see [2], we have $s = 7$, $p = 5$ and*

$$R(z) = \sum_{k=0}^{5} \frac{z^k}{k!} + \frac{z^6}{600}$$

hence, $m = p = 5$ and $\gamma_6 = \frac{1}{600}$, then from Theorem (2) $\tau_5 > 0$ and (H) is satisfied.

4. *If this method is RKF5(4), see [2], we have $p = 5$, $s = 6$, $\gamma_6 = \frac{1}{2080}$ and from Theorem (2) then (H) is not satisfied.*

5. *If this method is DOPRI8, see [2], we have $p = 8$, $s = 12$,*

$$R(z) = \sum_{k=0}^{8} \frac{z^k}{k!} + \sum_{k=9}^{12} \gamma_k z^k$$

where $\gamma_9 = 0.27521279901 \cdot 10^{-5}$, $\gamma_{10} = 0.24231996586959 \cdot 10^{-6}$, then $\tau_8 < 0$ and from Theorem (2) (H) is not satisfied.

Example 4. *Let a implicit Runge-Kutta method with s-stages and order p:*

 ISBN: 978-84-614-6167-7

1. *If this method is Lobatto IIIA we have* $s = 4$, $p = 6$ *and*

$$R(z) = \frac{1 + \frac{2z}{3} + \frac{z^2}{5} + \frac{z^3}{30} + \frac{z^4}{360}}{1 - \frac{z}{3} + \frac{z^2}{30}} = \sum_{k=0}^{6} \frac{z^k}{k!} + \frac{z^7}{5400} + \frac{z^8}{64800} + o(z^8)$$

hence, $m = p = 6$, $\gamma_7 = \frac{1}{5400}$ *and* $\gamma_8 = \frac{1}{64800}$ *then from Theorem (2)* $\tau_6 < 0$ *and* (H) *is not satisfied.*

2. *If this method is Lobatto IIIC we have* $s = 4$, $p = 6$ *and*

$$R(z) = \frac{1 + \frac{z}{3} + \frac{z^2}{30}}{1 - \frac{2z}{3} + \frac{z^2}{5} - \frac{z^3}{30} + \frac{z^4}{360}} = \sum_{k=0}^{6} \frac{z^k}{k!} + \frac{z^7}{5400} + \frac{z^8}{129600} + o(z^8)$$

hence, $m = p = 6$, $\gamma_7 = \frac{1}{5400}$ *and* $\gamma_8 = \frac{1}{129600}$ *then from Theorem (2)* $\tau_6 > 0$ *and* (H) *is satisfied.*

3. *In [3] are proposed several absolutely stable implicit methods and therefore must satisfy the condition* (H)*, consider the method for which* $p = 8$ *y* $s = 6$ *with stability function:*

$$R_6(z) = \frac{2580480 + 1290240z + 291840z^2 + 38400z^3 + 3108z^4 + 146z^5 + 3z^6}{2580480 - 1290240z + 291840z^2 - 38400z^3 + 3108z^4 - 146z^5 + 3z^6}$$

qfinding its power series expansion about $z = 0$ *we have* $\gamma_9 = \frac{2987}{1083801600}$ *and* $\gamma_{10} = \frac{299}{1083801600}$*. Hence,* $\tau_p = \left[\frac{9}{10!} + \gamma_{10} - \gamma_9\right] = 0 \geq 0$ *as stated in the converse of the Theorem (2).*

3 Numerical examples.

Example 5. *Consider the matrix* $A = \begin{pmatrix} -2 & 3 \\ -4 & -1 \end{pmatrix}$ *given in Example (2), then we have* $r_{\mathbb{R}}(A, V) = 3$.

Let $M_\mu = \begin{pmatrix} 1.5 + \mu^3 & 0 \\ 0 & 1.5 \end{pmatrix}$ *with* $-\sqrt[3]{6} < \mu < 0$ *then* $\|M_\mu\|_m < 3$ *and the matrices*

$$A + M_\mu = \begin{pmatrix} -0.5 + \mu^3 & 3 \\ -4 & 0.5 \end{pmatrix}$$

are Hurwitz stables.

The property (H) *is sufficient for the existence of* h^**, but not necessary. From Example (3)-(4) we know that RKF5(4) does not satisfy the property* (H) *and in this case we shall ensure that there is no* h^**.*

 ISBN: 978-84-614-6167-7

Figure (2)-(a) show for each μ, $-0.15 < \mu < 0$ *the value* h_u *such that for each* h, $0 < h < h_\mu$, *the numerical solutions of the systems*

$$\dot{x} = (A + M_\mu)x \tag{7}$$

converge to zero, these are just some of the systems from equation (2), hence $h^* \leq h_\mu$. *We can see that* $h_\mu \to 0$ *when* $\mu \to 0$, *therefore* h^* *does not exist.*

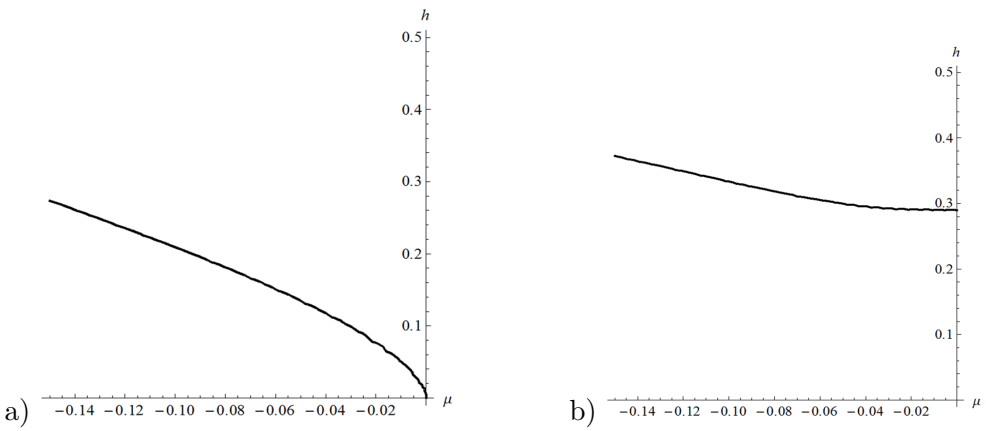


Figure 2:
a) For RKF5(4) does not exist h^*, because $h_\mu \to 0$ when $\mu \to 0$.
b) For DOPRI5 exist $h^* \simeq 0.19191$ and $h^* < h_\mu$ for each μ, $-0.15 < \mu < 0$.

Example 6. *Consider the matrices* A *and* M_μ *given above.*

From Example 3-3) we know that DOPRI5 satisfy the property (H), *then for this method* h^* *exist and in this case we have calculated* $h^* \simeq 0.19245$.

As in Example 5 we show in Figure 2-b) for each μ, $-0.15 < \mu < 0$, *the value* h_u *such that for each* h, $0 < h < h_\mu$, *the numerical solutions of the systems (7) converge to zero. We can see that* $h^* \leq h_\mu$ *for each* μ, $-0.15 < \mu < 0$.

In conclusion, note that $h_{\mu=0} \simeq 0.2909$ *is the lower of the* h_u, *see Figure 3-b), then for each* $h \leq 0.2909$, *the numerical solutions of equation (7) with* $-0.15 < \mu < 0$ *and calculated with the method DOPRI5 converges to zero, while for each* $h \leq 0.2909$, *we can find* μ *such that the corresponding solution of equation (7) calculated with the RKF5(4) method does not converge to zero. For example if we consider equation (8) and* $h = 0.25$ *is fixed then we can see in Figure (3)-a) that is sufficient to take* $\mu = -0.1$ *for that the numerical solution*

 ISBN: 978-84-614-6167-7

of equation (8) does not converge to zero:

$$\begin{aligned} \dot{x} &= (A + M_{-0.1})x \\ & x(0) = (1, 0) \end{aligned} \tag{8}$$

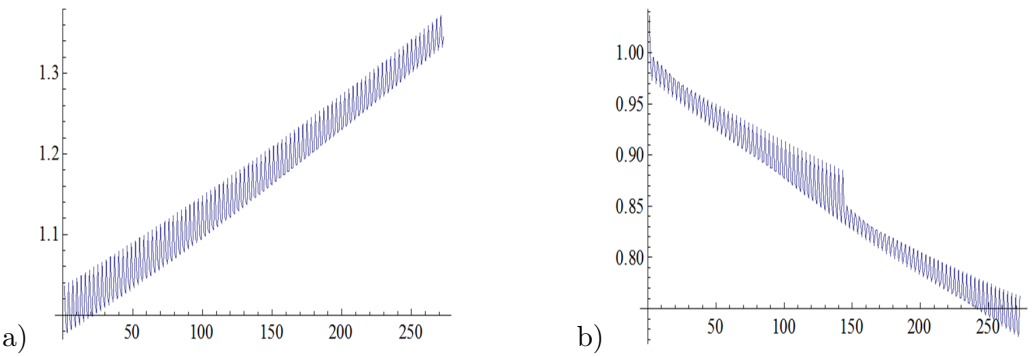


Figure 3: The figure shows euclidean norm value of the numerical solution every time it cuts the axis OX, that it to say every turn around the origin.
a) The numerical solution of the equation (8) calculated with RFK5(4) method and $h = 0.25$ does not converge to zero.
b) The numerical solution of the equation (8) calculated with DOPRI5 and $h = 0.25$ converge to zero, note that $h \leq h_{\mu=0}$, $h_{\mu=0} \simeq 0.2909$ is the lower of the h_u, see Figure 2-b).

4 Acknowledgement

The theory of this paper has been developed by the first two authors who warmly thank M. T. Bustos their dedication in checking numericaly the proposed results in several examples. Special thanks to Spanish CICYT under grant MTM2008-05489.

References

[1] Lambert J. D. Numerical methods for ordinary differential systems. Wiley, London (1991).

[2] Hairer E. Wanner G. Solving Ordinary Differential Equations II: Stiff Problems and Differential-algebraic Equations, Springer-Verlag, Berlin (1991).

[3] Jesus Vigo-Aguiar, Higinio Ramos. A family of A-stable Runge–Kutta collocation methods of higher order for initial-value problems. IMA Journal of Numerical Analysis, Volume 27 Issue 4 October 2007.

[4] G. De la Hera, E.Vazquez, H. Gonzalez. On the stability of convex symmetric polytops of matrices. Electronical Journal Differential Equations. Vol. 2000(2000), No. 09, pp. 1-17.

[7] Li Q., Bernhardsson B., Rantzer A., Davison E.J., Young P.M., Doyle J.C. A formula for computation of the Real Stability Radius. Elsevier, Automatica, Vol 31, No 6, 1995.

[8] F. Colonius, W. Kliemann. The dynamics of control. Birkhuser, Boston, 2000.

[10] Hinrichsen D., Pritchard A. J. Stability radii of linear systems, Systems and Control Letters, 7 (1986) 1-10.

[12] Hinrichsen D., Pritchard A. J. Real and complex stability radii: a survey, Progress in Systems and Control Theory. Birhuser, Boston (1990) Vol 6.

 ISBN: 978-84-614-6167-7

Volume II

 ISBN: 978-84-614-6167-7

 ISBN: 978-84-614-6167-7

Contents:

Volume I

 ISBN: 978-84-614-6167-7

 ISBN: 978-84-614-6167-7

 ISBN: 978-84-614-6167-7

Contents:

Volume II

 ISBN: 978-84-614-6167-7

 ISBN: 978-84-614-6167-7

Contents:

Volume III

 ISBN: 978-84-614-6167-7

 ISBN: 978-84-614-6167-7

 ISBN: 978-84-614-6167-7

Contents:
Volume IV

 ISBN: 978-84-614-6167-7

 ISBN: 978-84-614-6167-7

Contents:
Abstracts & Late Papers

 ISBN: 978-84-614-6167-7

 ISBN: 978-84-614-6167-7

Proceedings of the 11th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2011
26–30 June 2011.

Polynomial Chaos and Bayesian Inference in RPDE's – a biomedical application

Rob H. De Staelen[1], Karim Beddek[2] and Tineke Goessens[1]

[1] *Department of Mathematical Analysis, Faculty of Engineering and Architecture, Ghent University, Ghent, Belgium*

[2] *Laboratoire d'Électrotechnique et d'Électronique de Puissance de Lille, École Nationale Supérieure d'arts et Metiers, Lille, France*

emails: rob.destaelen@ugent.be, beddek@ensam.fr, tineke.goessens@ugent.be

Abstract

The electroencephalograph (EEG) is one of the most influential tools in the diagnosis of epilepsy and seizures. It measures electrical discharges of neurons in the human brain. The latter consists of many regions, all with a different electrical conductivity. Unfortunately one cannot measure this non invasively, e.g. preoperatively. In this paper, we investigate the uncertainty induced on the location of EEG current dipoles. A Bayesian framework is used, so as to include modeling error and noise, but combined with Polynomial Chaos expansions to represent random variables, speeding up computations. We evaluate this technique on a spherical head model with a standard clinical 27 sensor positioning.

Key words: Polynomial Chaos; Bayesian Inference; Random Partial Differential Equation (RPDE); sensitivity analysis; EEG; inverse problem

1 Introduction

In classical physics and engineering, the systems under consideration are deterministic. A *forward problem* consists of computing an output given an exact input through a known model. In practice, exact inputs or a perfect model, accounting for all external and internal processes, is rarely at hand. Instead, response quantities (e.g. hydrostatic pressure, electric potential, etc.) are more easily obtained, but with *noise*.

An *inverse problem* consists of computing the input given (noisy) output if the model is assumed to be known. Sometimes finding the model given input and output is also referred

 ISBN: 978-84-614-6167-7

to as an inverse problem, but we will not consider this here. Whereas forward problems are mostly *well-posed* in the sense of Hadamard, inverse problems are in general *ill-posed*.

For inverse problems *Bayesian Inference (BI)* [7] offers a rigorous foundation for inference from uncertain forward models and noisy data. It is a natural procedure for incorporating prior information, and a quantitative assessment of uncertainty in the results. Contrary to other methods, the output of BI is not solely a value, but a probability distribution that captures all available information about the parameters. Bayesian Inference has many benefits but a major drawback is the computational effort, see [8]. Especially when the forward model is complex, the number of computations becomes rapidly prohibitive. We will employ *Polynomial Chaos (PC)* [23] to represent uncertainties and to obtain an efficient forward model. As BI is a general and widely applicable technique, this remains the case when using PC. We will demonstrate PC in BI by means of a biomedical inverse problem.

1.1 The problem under consideration: electroencephalography

The electroencephalograph (EEG) is one of the most influential tools in the diagnosis of epilepsy and seizures, as it provides a record of ongoing electrical activity in the brain [16, 22]. The electrodes, connected to the EEG machine, measure signals produced by electrical discharge of neurons in the related areas of the brain. The quasi-static approximation of the Maxwell equations is justified by the very low frequencies (typically < 100 Hz) involved. The total electric current $\mathbf{J}$ can be partitioned [13] into two flows: a primary (driving) current $\mathbf{J}^p$ related to neural sources, and an ohmic volume (passive) current $\mathbf{J}^v$ that results from the effect of the electric field in the volume: $\mathbf{J} = \mathbf{J}^p + \mathbf{J}^v = \mathbf{J}^p + \sigma\mathbf{E} = \mathbf{J}^p - \sigma\nabla V$, where V is the electric potential. Since the total current is divergence free and no current flows outside the head, we obtain

$$\boldsymbol{\nabla} \cdot (\sigma(\mathbf{r})\nabla V(\mathbf{r})) = \boldsymbol{\nabla} \cdot \mathbf{J}^p(\mathbf{r}), \quad \text{in} \quad H \tag{1a}$$

$$\boldsymbol{\nu} \cdot \sigma(\mathbf{r})\nabla V(\mathbf{r}) = 0, \quad \text{on} \quad \partial H \tag{1b}$$

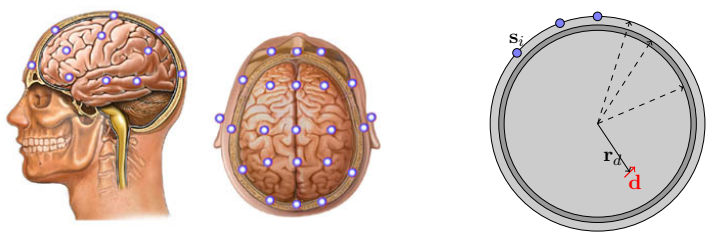
where H is the head, σ the conductivity and $\boldsymbol{\nu}$ the outward unit normal on ∂H. Usually, the head is assumed to be made up of disjoint regions (the scalp, skull, cerebrospinal fluid, grey matter and white matter) with each a constant conductivity assigned. We will use a *spherical head model* with three layers: the inner sphere (radius .87) represents the brain, the intermediate layer (radius .92) represents the skull an the outer sphere (radius 1) corresponds to the scalp, see Fig. 1(b).

A widely used approximation of the neural activity of patients suffering from epilepsy [6] is the representation of the primary current as an electric dipole with dipole moment $\mathbf{d}$ located at $\mathbf{r}_d$ inside the cortex; $\mathbf{J}^p(\mathbf{r}) = \mathbf{d}\delta(\mathbf{r} - \mathbf{r}_d)$ with $\mathbf{r}$ the position in the head measured from the center of the concentric spheres and δ the Dirac measure. The goal is to localize this dipole.

 ISBN: 978-84-614-6167-7

1.2 The intrinsic uncertainty of the conductivity and the sensitivity on source localization

It is well known in literature [3, 2, 19, 5, 21] that the conductivity of the different head layers is a key parameter in the localization of an electric dipole given an EEG measurement set (scalp potentials at the $\mathbf{s}_i$) during an epileptic seizure. Till now all these studies assumed a fixed constant conductivity value in each region and quantified the influence, on scalp potentials and source localization, of discrete perturbations. However, the conductivity is not constant in each layer and determining the conductivity values in the human head has been subject of research since many years [9]. The first attempts to measure it made use of in vitro techniques. But in fact, it is impossible to fully measure σ since tissues are inhomogeneous [18] and anisotropic, and furthermore their properties depend on physiological processes. The latter implies the conductivity changes over time; a perfect conductivity determination in advance of an EEG is not valid anymore when measuring scalp potentials during a seizure. Moreover the conductivity varies from individual to individual [14] and can only be correctly measured invasively. All this intrinsic variability will be captured by attributing the conductivity a probability distribution function, i.e. making it a random variable.



(a) (b)

Figure 1: A standard 27 electrode placement (a) and a spherical head illustration (b) with sensors $\mathbf{s}_i$.

We make a modest simplification based on the fact that the outer and inner layer are assumed to have similar electrical properties, meaning that $\sigma_{\text{scalp}} = \sigma_{\text{brain}}$. Instead of working with the two different conductivities σ_{skull} and σ_{brain}, we consider the *conductivity ratio* $X = \sigma_{\text{scull}}/\sigma_{\text{brain}}$. Based on in vivo measurements [12] we let it be uniformly distributed with mean .026 and standard deviation of .0092.

The purpose of this paper is to investigate the influence of the intrinsic uncertainty on the EEG source localization. It is clear that the source localization won't be deterministic but will be represented by a probability distribution. In Fig. 1(a) a standard 27 electrode

 ISBN: 978-84-614-6167-7

placement is shown. This setup is used in the Department of Neurology at the Ghent University Hospital for clinical practice, [4].

2 Methods

2.1 The forward model

2.1.1 The EEG lead field model

To calculate the potential at the 27 sensor locations $\mathbf{s}_i$ one needs to solve system (1) for V and evaluate $S_i = V(\mathbf{s}_i)$. Analytic solutions exist only for special cases [10]; in general the system of equations has to be solved numerically. Instead of directly solving for the electric potential induced by a current dipole (e.g. through BEM), one can solve for the electric *lead fields* [20, 17]. Invoking this method, the sensor measurements are given by $\mathbf{S} = \mathsf{L}(\mathbf{r}_d, X)\mathbf{d}$ and $\mathbf{S} = (S_i)_{i=1}^{27}$, where $\mathsf{L} \in \mathbb{R}^{27\times 3}$ is the *lead field* matrix which depends only on the dipole position, conductivity ratio, and the geometry.

2.1.2 Polynomial Chaos expansion

Consider a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ where Ω is the sample space, $\mathcal{F} \subseteq 2^\Omega$ a σ-algebra and $\mathbb{P}$ a measure. The Hilbert space of random variables $Y : \Omega \to \mathbb{R}$, for which $\mathbb{E}\left[Y^2\right] < +\infty$ is denoted $\mathcal{L}^2(\Omega, \mathcal{F}, \mathbb{P})$. Two random variables Y_1 and Y_2 are orthonormal if their inner product satisfies $\mathbb{E}\left[Y_i Y_j\right] = \delta_{ij}$. The space $\mathcal{L}^2(\Omega, \mathcal{F}, \mathbb{P})$ can be decomposed as follows:

Theorem (Wiener [23]) *Let $\{\xi_i\}_{i\in\mathbb{N}}$ be a set of orthonormal Gaussian random variables $\xi_i \in \mathcal{L}^2(\Omega, \mathcal{F}, \mathbb{P})$. Define $\hat{\Gamma}_p$ as the space of all polynomials in ξ_i of degree not exceeding p. The following decomposition holds*

$$\mathcal{L}^2(\Omega, \mathcal{F}, \mathbb{P}) = \bigoplus_{p\geq 0} \Gamma_p, \tag{2}$$

where Γ_p denotes the orthogonal complement of $\hat{\Gamma}_{p-1}$ in $\hat{\Gamma}_p$ with respect to the inner product of $\mathcal{L}^2(\Omega, \mathcal{F}, \mathbb{P})$.

The relation in (2) is called the *(homogenous) Wiener Chaos decompostion* and the space Γ_p the *(homogenous) Polynomial Chaos (PC)* of order p. Since the Polynomials Chaoses are constructed orthogonal to the normal probability measure, they are equivalent with the multidimensional Hermite polynomials. According to (2) a random variable Y can be represented in the form

$$Y(\omega) = \hat{a}_0\Gamma_0 + \sum_{i_1=1}^{\infty} \hat{a}_{i_1}\Gamma_1(\xi_{i_1}(\omega)) + \sum_{i_1=1}^{\infty}\sum_{i_2=1}^{i_1} \hat{a}_{i_1 i_2}\Gamma_2(\xi_{i_1}(\omega), \xi_{i_2}(\omega)) + \cdots . \tag{3}$$

 ISBN: 978-84-614-6167-7

In practice (3) is truncated after a finite number of terms and written conveniently as $Y(\omega) = \sum_{i=0}^{\overline{n,p}} a_i \Psi_i(\boldsymbol{\xi}(\omega))$, where $\overline{n,p} + 1 = \binom{n+p}{p}$, $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_n)$ and p is the chaos order. When $n = 1$ then $p = 3$ is usually sufficient. Since $\mathbb{E}[\Psi_i(\boldsymbol{\xi})\Psi_j(\boldsymbol{\xi})] = \delta_{ij}$, the expansion coefficients are given by $a_i = \mathbb{E}[Y\Psi_i(\boldsymbol{\xi})]$ and computed through projection or regression [11]. What distinguishes the Hermite Chaos expansion from other possible expansions is that the basis polynomials are Hermite polynomials in terms of Gaussian variables and are orthogonal with respect to the weighting function that has the form of n-dimensional independent Gaussian probability density function. The Hermite Chaos expansion proves to be effective in solving stochastic differential equations with Gaussian inputs as well as certain types of non-Gaussian inputs. However, for general non-Gaussian random inputs, the optimal exponential convergence rate will not be realized. In some cases the convergence rate is in fact severely deteriorated. In order to deal with more general random inputs the *generalized Polynomial Chaos (gPC)* [24] is used. It makes use of other well known families of orthogonal polynomials. For each type of random input variable an optimal (in the sense of convergence rate) expansion basis can be found. This expansion is in general referred to as the Wiener-Askey Polynomial Chaos since the basis polynomials are these from the *Askey-scheme* [1] of the hypergeometric functions. The original Wiener Polynomial Chaos corresponds to the Hermite Chaos and is a subset of the Wiener-Askey Polynomial Chaos.

2.1.3 Uncertainty propagation

We assume all input is stochastic. This means the conductivity ratio as well as the moment and location of the dipole are modeled by random variables. As assumed in the Introduction, the uncertainty is uniform, so we employ Legendre Chaos. The conductivity ratio is then written as $X(\xi_X) = .026 + .0159\xi_X$ where ξ_X is uniformly distributed on the interval $[-1, 1]$. Likewise the position $\mathbf{r}_d$ resp. moment $\mathbf{d}$ are written in terms of uniform variables $\boldsymbol{\xi}_{\mathbf{r}_d} = (\xi_r^{\mathbf{r}_d}, \xi_\theta^{\mathbf{r}_d}, \xi_\phi^{\mathbf{r}_d})$ resp. $\boldsymbol{\xi}_{\mathbf{d}} = (\xi_r^{\mathbf{d}}, \xi_\theta^{\mathbf{d}}, \xi_\phi^{\mathbf{d}})$ and are confined to the inner sphere. By Doob-Dynkin's lemma the solution can be expressed in terms of $\boldsymbol{\xi} = (\boldsymbol{\xi}_{\mathbf{r}_d}, \boldsymbol{\xi}_{\mathbf{d}}, \xi_X)$. A Legendre Chaos of order p of the i-the sensor reads (with λ_j the j-th 7-variate Legendre polynomial)

$$\hat{S}_i(\boldsymbol{\xi}) = \mathsf{L}(\mathbf{r}_d(\boldsymbol{\xi}_{\mathbf{r}_d}), X(\xi_X))\, \mathbf{d}(\boldsymbol{\xi}_{\mathbf{d}}) = \sum_{j=0}^{\overline{7,p}} V_{ij}\lambda_j(\boldsymbol{\xi}), \quad V_{ij} = \mathbb{E}\left[\hat{S}_i(\boldsymbol{\xi})\lambda_j(\boldsymbol{\xi})\right]. \tag{4}$$

The coefficients $V_{ij} \in \mathbb{R}$ were computed through a sparse grid or Smolyak integration scheme. We have $n = 7$ random dimensions and consider a Legendre Chaos of order $p = 5$. The effect of increasing the order of the chaos expansion may be of interest, but for our purposes we verified that 5 suffices; $\overline{7,5} = 791$.

 ISBN: 978-84-614-6167-7

2.2 The inverse model

2.2.1 Bayesian Inference

The inverse model consists of locating the dipole given a set of sensor measurements. Now this locating isn't deterministic but results in a probability density. The stochastic sensor predictions $\hat{\mathbf{S}}$ are given by (4). Sensor measurements are in fact undergoing some error – sensor noise and model error –, so $\mathbf{S} = \hat{\mathbf{S}}(\boldsymbol{\xi}) + \boldsymbol{\epsilon}$, where the components of $\boldsymbol{\epsilon}$ are i.i.d. random variables with density ϕ_ϵ. A typical assumption is ϵ being a zero mean normal variable [15], so $\boldsymbol{\epsilon} \sim \mathrm{N}(0, \sigma_\epsilon^2 \mathsf{I}_{27})$. In this case the *likelihood* becomes

$$\phi_{\mathbf{S}|\boldsymbol{\xi}}(\mathbf{x}|\boldsymbol{\xi}) = \prod_{j=1}^{27} \phi_\epsilon \left(x_j - \hat{S}_j(\boldsymbol{\xi}) \right).$$

We have a *posterior* probability density $\phi_{\boldsymbol{\xi}|\mathbf{S}}$ for the model parameters (up to a normalizing factor) defined by the product of the likelihood $\phi_{\mathbf{S}|\boldsymbol{\xi}}$ with the prior density $\phi_{\boldsymbol{\xi}}$. The latter embodies all prior knowledge on these parameters. In our case the prior is uninformative so equal to a constant. We are interested in integrals over the posterior density, especially marginalizing over X. This means that the forward model (needed in the likelihood) has to be computed many times. For Monte Carlo simulations requiring 10^3-10^5 samples, the total cost of this calculation quickly becomes prohibitive. It is therefore necessary to have an efficient way of doing so. This is exactly why we introduced Polynomial Chaos.

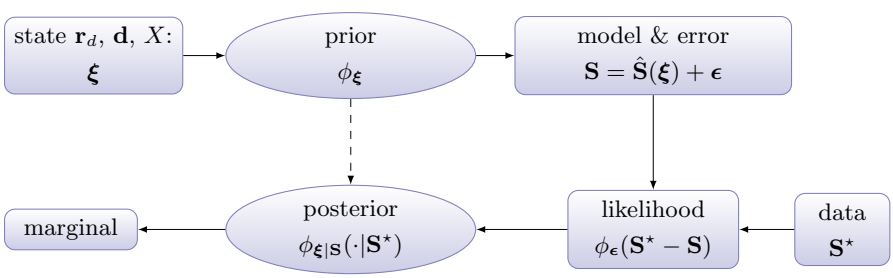


Figure 2: Different components in the Bayesian procedure.

2.2.2 Splitting and reduction of the integration

If $\mathbf{S}^\star$ is a measurement data set, one computes the posterior density $\phi_{\boldsymbol{\xi}|\mathbf{S}}(\cdot|\mathbf{S}^\star)$. But we note that given $\mathbf{r}_d$ and X we can use the EEG lead field model $\mathbf{S} = \mathsf{L}(\mathbf{r}_d, X)\mathbf{d}$ to find an *optimal* $\mathbf{d}^\star$, namely $\mathsf{L}(\mathbf{r}_d, X)^\dagger \mathbf{S}$. Using this we obtain the dipole location density $\phi_{\mathbf{r}_d}(\boldsymbol{\xi}_{\mathbf{r}_d}|\mathbf{S}^\star)$

 ISBN: 978-84-614-6167-7

proportional to

$$\int_{-1}^{1} \prod_{i=1}^{27} \phi_\epsilon \left(S_i^\star - \hat{S}_i(\boldsymbol{\xi}^\star) \right) \mathrm{d}\, \xi_X.$$

with $\boldsymbol{\xi}^\star = (\boldsymbol{\xi}_{\mathbf{r}_d}, \boldsymbol{\xi}_{\mathbf{d}}^\star(\boldsymbol{\xi}_{\mathbf{r}_d}, \xi_X), \xi_X)$ and $\boldsymbol{\xi}_{\mathbf{d}}^\star$ such that $\mathbf{d}^\star = \mathbf{d}(\boldsymbol{\xi}_{\mathbf{d}}^\star)$ is optimal.

The density $\phi_{\mathbf{r}_d}$ is not depictable so we proceed by splitting the radial and angular component of $\mathbf{r}_d$. First, we marginalize over X and the angular part to obtain the radial component density $\phi(|\mathbf{r}_d|(\xi_r^{\mathbf{r}_d})|\mathbf{S}^\star)$ as

$$\frac{1}{2^6 \nu} \int_{-1}^{1} \prod_{i=1}^{27} \phi_\epsilon \left(S_i^\star - \hat{S}_i(\boldsymbol{\xi}) \right) \mathrm{d}\, \xi_\theta^{\mathbf{r}_d}\, \mathrm{d}\, \xi_\varphi^{\mathbf{r}_d}\, \mathrm{d}\, \boldsymbol{\xi}_{\mathbf{d}}\, \mathrm{d}\, \xi_X.$$

Then, secondly, we fix the radial component at its mean $\bar{\xi}_r^{\mathbf{r}_d}$ and compute the density of the angular components of $\mathbf{r}_d$, $\phi(\angle \mathbf{r}_d(\bar{\xi}_r^{\mathbf{r}_d}, \xi_\theta^{\mathbf{r}_d}, \xi_\varphi^{\mathbf{r}_d})|\mathbf{S}^\star)$, again approximated by

$$\frac{1}{2\nu} \int_{-1}^{1} \prod_{i=1}^{27} \phi_\epsilon \left(S_i^\star - \hat{S}_i(\bar{\boldsymbol{\xi}}^\star) \right) \mathrm{d}\, \xi_X$$

with $\bar{\boldsymbol{\xi}}^\star = (\bar{\xi}_r^{\mathbf{r}_d}, \xi_\theta^{\mathbf{r}_d}, \xi_\varphi^{\mathbf{r}_d}, \boldsymbol{\xi}_{\mathbf{d}}^\star(\bar{\xi}_r^{\mathbf{r}_d}, \xi_\theta^{\mathbf{r}_d}, \xi_\varphi^{\mathbf{r}_d}, \xi_X), \xi_X)$.

3 Results and discussion

We will discuss our stochastic approach of locating an electric dipole - supposed to have evoked an epileptic seizure. However, first we construct some input to work with. We consider the stochastic position and moment vectors $\mathbf{r}_d(\boldsymbol{\xi}_{\mathbf{r}_d})$ and $\mathbf{d}(\boldsymbol{\xi}_{\mathbf{d}})$ evaluated at some fixed stochastic input vectors, namely

$$\begin{aligned} \mathbf{r}_d(-.7, .2, .43) &= (-.027, -.121, -.041), \\ \mathbf{d}(.05, -.1, .2) &= (-.365, -.265, .071), \\ X(.2) &= .0253. \end{aligned}$$

Having computed the chaos coefficients, we obtain 27 random variables S_i. We do not know its probability distribution but can regard it as function of the random variable $\boldsymbol{\xi}$. The sensor values $\mathbf{S}^\star$ corresponding to this fixed input variables will serve as our data. The inverse problem consists of recovering these inputs as described in the above.

Proceeding with the marginalizing process we obtain the radial component density $\phi(|\mathbf{r}_d|(\xi_r^{\mathbf{r}_d})|\mathbf{S}^\star)$ depicted in Fig. 3. Further we compute the mean and standard deviations of the densities; see Table 1. We find good agreement with our constructed input.

Turning to the angular components at fixed mean radius we find the level sets of the density at error level .10 in Fig. 4. The mean value is $(\bar{\xi}_\theta, \bar{\xi}_\phi) = (.222, .4385)$, which in

 ISBN: 978-84-614-6167-7

Table 1: Mean radius $\bar{\xi}_r$ and standard deviation $\check{\xi}_r$ at different σ_ϵ.

σ_ϵ	.10	.15	.20	.30
$\bar{\xi}_r$	$-.7064$	$-.7142$	$-.7213$	$-.7267$
$\check{\xi}_r$	.01334	.02245	.02772	.03287

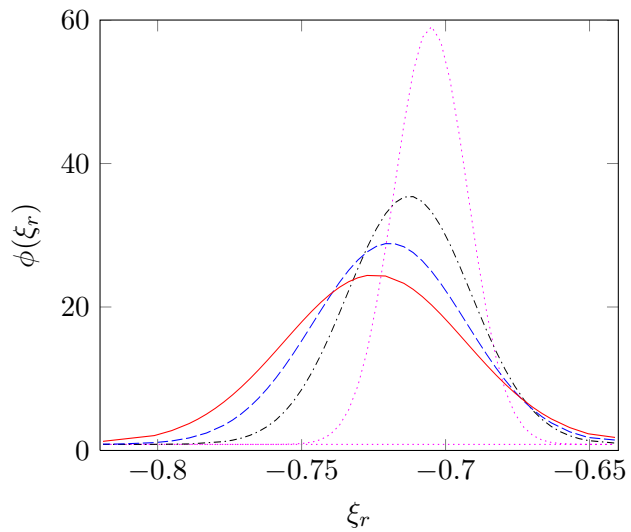


Figure 3: Density $\phi(\xi_r)$ at degree 5 with $\sigma_\epsilon = .1$ (dot), $\sigma_\epsilon = .15$ (dash-dot), $\sigma_\epsilon = .2$ (dash) and $\sigma_\epsilon = .3$ (full).

turn is good approximation of the true value $(.2, .43)$. The Bayesian approach results in a density, and here we clearly see some extra information. The oval shaped contours give insights about correlation between the angular components of the dipole location, and this correlation is positive. The angular components are thus dependent variables. We note that the approach followed lacks the difficulty of choosing physiological properties of brain tissues a priori – which we set as our goal.

Another issue which deserves attention is the error level. We performed the above analysis with different error levels and collect the results in Table 1. When we compare the error levels with the obtained standard deviations of the radial component prediction, we observe no linear dependence. We fit it to a kind of rectangular hyperbolic curve and obtain,

$$\check{\xi}_r = .037 \frac{\sigma_\epsilon^{2.458}}{.006 + \sigma_\epsilon^{2.458}}, \quad 0 \leq \sigma_\epsilon \leq .35.$$

 ISBN: 978-84-614-6167-7

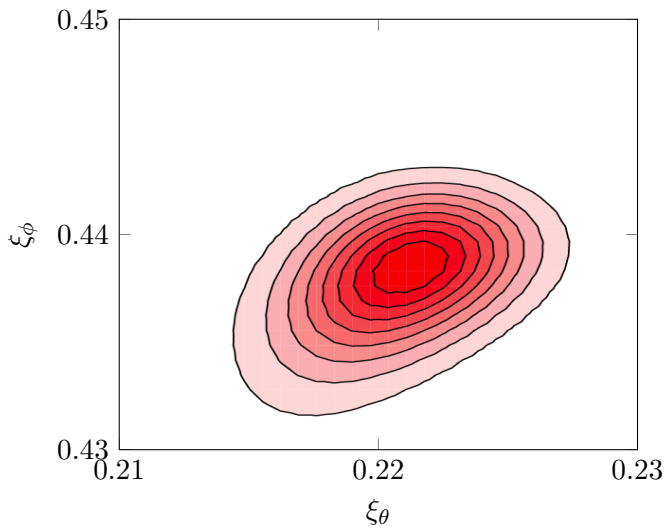


Figure 4: Contourplot of $\phi(\xi_\theta, \xi_\phi)$ with $\sigma_\epsilon = .1$ and $\bar{\xi}_r = -.7$.

Inverting the relation for a predefined $\check{\xi}_r = .01$ results in $\sigma_\epsilon = .083$. This value has to be seen as an upper bound, since $\sigma_\epsilon \to 0$ would imply $\check{\xi}_r \to \varepsilon$ with $\varepsilon > 0$ small; there will always be some intrinsic error, e.g. due to the truncation of the chaos expansion.

4 Conclusion

By employing a Polynomial Chaos expansion and using a Bayesian framework we account for all errors – model, measurement and uncertainty. With this technique we get good results and avoid the problem of blindfold guessing a patients cerebral conductivities. We proved that in general the angular components of the dipole location are dependent variables. We obtained an upper bound on the overall error level in model and measurements, given a predefined accuracy on the prediction of the radial component of the dipole location.

It will be of future research not only to relate the error level with $\check{\xi}_r$ but to include and relate the standard deviations of the angular densities, and more generally the covariance matrix, as a whole to obtain a relation or bound of the form

$$\check{\boldsymbol{\xi}}_{\mathbf{r}_d} = \mathbf{e}(\sigma_\epsilon), \quad \text{or} \quad ||\check{\boldsymbol{\xi}}_{\mathbf{r}_d}||_3 \leq e(\sigma_\epsilon),$$

with $\check{\boldsymbol{\xi}}_{\mathbf{r}_d}$ the correlation tensor between the components of the dipole position.

In this report only the results for one dipole location are stated, more locations and orientations will be published in a more elaborate paper.

 ISBN: 978-84-614-6167-7

Acknowledgement(s)

The author would like to thank the department of Electrical Energy, Systems and Automation (*EESA*) at Ghent University, and in particular Guillaume Crevecoeur, for providing some computer algorithms. His supervisors Roger Van Keer and Marián Slodička are thanked for their support.

References

[1] R. Askey and J. Wilson. *Some Basic Hypergeometric Orthogonal Polynomials that Generalize Jacobi Polynomials*. Memoirs of the American Mathematical Society, 1985.

[2] K.A. Awada, D.R. Jackson, S.B. Baumann, J.T. Williams, D.R. Wilton, P.W. Fink, and B.R. Prasky. Effect of conductivity uncertainties and modeling errors on EEG source localization using a 2-D model. *Biomedical Engineering, IEEE Transactions on*, 45(9):1135–1145, 1998.

[3] Md. Bashar, Yan Li, and Peng Wen. EEG analysis on skull conductivity perturbations using realistic head model. In Peng Wen, Yuefeng Li, Lech Polkowski, Yiyu Yao, Shusaku Tsumoto, and Guoyin Wang, editors, *Rough Sets and Knowledge Technology*, volume 5589 of *Lecture Notes in Computer Science*, pages 208–215. Springer Berlin / Heidelberg, 2009.

[4] P. Boon. Tertiary epilepsy care in Belgium: The practice at Ghent University Hospital Reference Centre for Refractory Epilepsy. *Epilepsia*, 46:318–318, 2005.

[5] Fangmin Chen, Hans Hallez, and Steven Staelens. Influence of skull conductivity perturbations on EEG dipole source analysis. *Medical Physics*, 37(8):4475–4484, 2010.

[6] J.C. de Munck, B.W. van Dijk, and H. Spekreijse. Mathematical dipoles are adequate to describe realistic generators of human brain activity. *Biomedical Engineering, IEEE Transactions on*, 35(11):960–966, 1988.

[7] A. Djafari. A Bayesian approach for data and image fusion. In Ch Williams, editor, *Bayesian Inference and Maximum Entropy Methods*, pages 386–407, Univ. of Idaho, Moscow, Idaho, USA, 2002. AIP Conference Proceedings 659.

[8] M. Evans and T. Swartz. Methods for approximating integrals in statistics with special emphasis on Bayesian integration problems. *Statistical Science*, 10:254–272, 1995.

[9] L. Geddes and L. Baker. The specific resistance of biological material–A compendium of data for the biomedical engineer and physiologist. *Medical and Biological Engineering and Computing*, 5:271–293, 1967.

 ISBN: 978-84-614-6167-7

[10] David B. Geselowitz. On bioelectric potentials in an inhomogeneous volume conductor. *Biophysical Journal*, 7(1):1–11, 1967.

[11] R. Ghanem and P. Spanos. *Stochastic Finite Elements: A Spectral Approach*. Dover Publications, 2003.

[12] S. Goncalve, J.C. de Munck, J.P.A. Verbunt, R.M. Heethaar, and F.H.L. da Silva. In vivo measurement of the brain and skull resistivities using an EIT-based method and the combined analysis of SEF/SEP data. *Biomedical Engineering, IEEE Transactions on*, 50(9):1124–1127, 2003.

[13] Matti Hämäläinen, R. Hari, R.J. Ilmoniemi, Jukka Knuutila, and Olli V. Lounasmaa. Magnetoencephalography—theory, instrumentation, and applications to noninvasive studies of the working human brain. *Rev. Mod. Phys.*, 65(2):413–497, 1993.

[14] Geertjan Huiskamp. Interindividual variability of skull conductivity: an EEG-MEG analysis. *International Journal of Bioelectromagnetism*, 10(1):25–30, 2008.

[15] C.H. Muravchik and A. Nehorai. EEG/MEC error bounds for a static dipole source with a realistic head model. *Signal Processing, IEEE Transactions on*, 49(3):470–484, 2001.

[16] MA Nolan. Memory function in childhood epilepsy syndromes. *Journal of Paediatrics and Child Health*, 40:20–27(8), 2004.

[17] G. Nolte and G. Dassios. Analytic expansion of the EEG lead field for realistic volume conductors. *Physics in Medicine and Biology*, 50(16):3807–3823, 2005.

[18] Jorma O. Ollikainen, Marko Vauhkonen, Pasi A. Karjalainen, and Jari P. Kaipio. Effects of local skull inhomogeneities on EEG source estimation. *Medical Engineering & Physics*, 21(3):143–154, 1999.

[19] Robert Pohlmeier, Helmut Buchner, Gunter Knoll, Adrian Riencker, Rainer Beckmann, and Jörg Pesch. The influence of skull-conductivity misspecification on inverse source localization in realistically shaped finite element head models. *Brain Topography*, 9:157–162, 1997.

[20] J.J. Riera and M.E. Fuentes. Electric lead field for a piecewise homogeneous volume conductor model of the head. *Biomedical Engineering, IEEE Transactions on*, 45(6):746–753, 1998.

[21] Cees J. Stok. The influence of model parameters on EEG/MEG single dipole source estimation. *Biomedical Engineering, IEEE Transactions on*, BME-34(4):289–296, 1987.

 ISBN: 978-84-614-6167-7

[22] H. Urbach. Imaging of the epilepsies. *European Radiology*, 15:494–500, 2005.

[23] Norbert Wiener. The homogeneous chaos. *American Journal of Mathematics*, 60(4):897–936, 1938.

[24] Dongbin Xiu and George Em Karniadakis. The Wiener-Askey Polynomial Chaos for stochastic differential equations. *SIAM J. Sci. Comput.*, 24(2):619–644, 2002.

 ISBN: 978-84-614-6167-7

Proceedings of the 11th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2011
26–30 June 2011.

Magnetism of platinum nanoparticles: an ab-initio point of view

C. Di Paola[1] and F. Baletto[1]

[1] *Department of Physics, School of Natural and Mathematical Science, King's College of London, Strand, WC2R 2LS London UK*

emails: cono.di_paola@kcl.ac.uk, francesca.baletto@kcl.ac.uk

Abstract

The magnetic behaviour of pure platinum nanoparticles has been studied by means of ab-initio density-functional based numerical simulations. Both unconstrained and fixed spin-polarised schemes have been used to calculate the total magnetisation as well as the local atomic polarisation of five different structural motifs of Pt_{13} nanoparticles. The dramatic role of the electronic temperature on the electronic density of states, and hence of the clusters magnetic properties, has been revealed.

1 Introduction

Nanomagnetism deals with the study of magnetic behaviour of objects which are nanoscopic at least in one dimension, including free nanoparticles (NP), nanodots, nanowires as well as thin films. Due to the broken translation symmetry and to dimensions comparable to characteristic lengths e.g., exchange length or the domain wall thickness, there are differences in the magnetic response of nano-objects as compared to their bulk counterparts, leading to novel physical properties [5, 7]. The enhancement of magnetisation in clusters of chemical species that are ferromagnetic when in the bulk has been proven thanks to SternGerlach experiments [1]. These nanoparticles can be used as magnetic nanometre devices [2] . For their practical applications, it is worthy noting that, when the size of NP becomes too small, the magnetic moment of the single-domain ferromagnet can fluctuate thermally, so that the superparamagnetic limit can be achieved. On the other hand, a magnetic behaviour has been observed quite surprisingly in NP of noble and quasi-noble metals [7]. Experimental magnetic measurements of well characterised Pt clusters, with 13±2 atoms and up to

 ISBN: 978-84-614-6167-7

8 unpaired electrons, monodisperse in NaY zeolites, have confirmed their extraordinary magnetic polarisation [3, 4].

From an atomistic point of view, the shape of the density of electronic states (DOS), due to the reduced dimensionality, depends on the geometry of the system and can influence drastically the magnetic behaviour of transition metal nanoparticles [5, 9, 8]. Generally speaking, a small cluster is expected to posses a low spin state [6] due to the average energy spacing of electronic states at the Fermi energy, the so-called Kubo gap, that scales with the width of the valence band, which is usually of the order of few eV, and the number of atoms in the cluster. Nonetheless, an opposite effect of the morphology is that quite symmetric structures may show a high degeneracy of the highest occupied molecular orbital favouring high spin states. Although the origin of ferromagnetism in non-magnetic transition metals clusters is still under debate, tailoring their magnetic properties by simply reducing the length of certain critical dimensions constitutes a really promising eld for investigation [7]. Local spin density numerical simulations represent an efficient and feasible tool for calculating magnetic properties of nano-objects, where the broadening of the DOS and the smoothening of the discontinuities at the Fermi energy have been taken into account by means of introducing a fictitious electronic temperature, T_e [10].

In this work, platinum NP of 13 atoms with different shapes have been viewed as paradigmatic example. Their energetic stability has been calculated as a function of their total magnetization (TM), using both unconstrained and fixed spin-polarised frameworks. We have demonstrated that large values of T_e wrongly reproduce the DOS, affecting directly the broadening of the *d*-band, as depicted in panel (D) of Fig. 1. This allows a non realistic overlap of *s* and *d* bands, and hence of the atomic hybridisation which has a very strong effect on the total magnetization of the cluster itself. This study may help in determination of the structural motif of Pt clusters which favour high magnetization as found in experiments [3, 4].

2 Magnetic properties of Pt_{13} clusters

In agreement with available literature [11, 12, 13] , the five most stable structures, the Mackay icosahedron (Ih), Ino decahedron (Dh) and cuboctahedron (CO), and the two bi-pyramidal geometries, indicated as the two best geometries, labelled as GM0 and GM2, are depicted in panel (A) of Fig. 1.

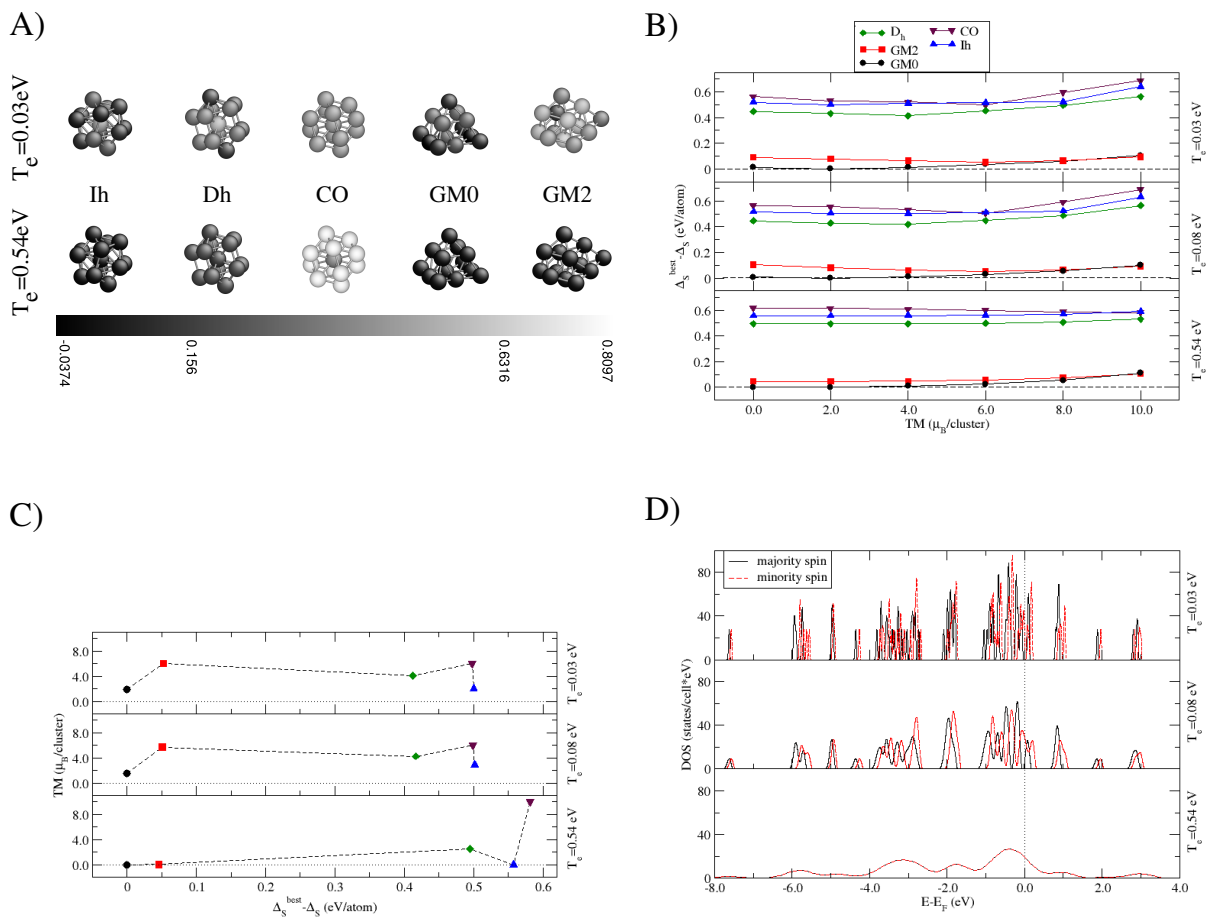
The QuantumEspresso package [15], a density functional theory based plane wave code, has been used employing the generalised gradient approximation (GGA) with the Perdew-Burke-Ernzerhof (PBE) functional [16] having an energy cut-off of 45 Ryd and a charge density cut-off of 360 Ryd. An ultrasoft pseudopotential has been included for Pt atoms with an electronic configuration $5d^9 6s^1$ and all the calculations have been done at Γ point. In all cases, a structural optimisation has been performed using the Broyden-Fletcher-

 ISBN: 978-84-614-6167-7

GoldfarbShann (BFGS) procedure. TM has been calculated for various values of T_e considered, as reported in panel (C) of Fig. 1.
It has been found that 0.03 eV is a reasonable value of T_e in order to get a better description of the DOS and thus of total magnetisation. Initially, constrained spin-polarised calculations have been preformed, in which spin multiplicity was fixed to the value of (2S+1), with the total spin S varying between 0 and 5, corresponding to a final TM between 0.00 and 10.00 μ_B, respectively.
For a better understanding of the role played by the fictitious T_e, unconstrained spin-polarised calculations have been performed as well. The energy cost to create a finite system with respect to the bulk is quantified by the excess energy $\Delta_S = \frac{E_{tot}-N\epsilon_b}{N^{2/3}}$, where E_{tot} is the total energy of NP and ϵ_b is the energy of one atom in the bulk and N is the number of atoms in the cluster. The relative stability of each structure has been given rescaling its excess energy to the value obtained for GM0, labelled as Δ_S^{best}. The energetic stability of each isomer has been plotted as a function of its TM for each T_e.
The geometrical motifs analysed are within at least 0.5 eV/atom in energy , with GM0 and GM2 being the most favourable structures, independently of the choice of the electronic temperature. However, morphologies showing degeneracy in energy only at higher spins (T_e=0.54 eV) start to show a different trend at lower T_e. Ih and CO isomers have been found to have a more complicated behaviour that ends up in a small energetic crossing between 4.0 and 6.0 μ_B as the electronic temperature is reduced. The energetic profile of Dh remains constant for $T_e \leq 0.08$ eV as well as the relative profile of GM0 and GM2, although for their energetic degeneracy starts to become relevant at lower TM.
In the case of unconstrained spin, structures like Ih, GM0 and GM2 present a TM between 2.00 and 6.00 μ_B at low T_e, while they are not magnetic at all at larger T_e. On the other hand, the CO isomer posses a TM in the range of 8.0 to 10.0 μ_B. The energy gap has been found to be lower by about 0.1 eV/atom, panel (C) of Fig. 1 and Ih and CO structures are almost degenerate in energy but different in magnetisation.
Additionally, the local atomic polarisation, which is the difference between the number of majority and minority spins per each atom, has been analysed and shown in panel (A) of Fig. 1 (where black to white gradient represents lower to higher atomic polarisations respectively). This quantity clearly reflects the effect of electronic temperature on the density of states. At low values, for example, the Ih isomer shows a slight compression along one of the 5-fold axes, reducing the bond length between one set of opposite vertexes and the centre by 2.58 $\AA$ and Dh presents a slight distortion that leaves the core atom away from the position of the centre of mass. A geometrical contraction/extension of a bond induces a local magnetic anisotropy, due to a non-homogeneous charge transfer from *s*-like orbitals and, as a consequence, a different *s*-*d* atomic overlap. These effects have been almost suppressed when T_e=0.54 eV, thus preventing the observation of the magnetic behaviour in platinum clusters.

 ISBN: 978-84-614-6167-7

In conclusion, high T_e are not able to reproduce the broadening of the *d*-band or the charge transfer providing a wrong magnetic behaviour, although the energy stability is quite independent of this parameter. It is possible that the approximation of 'superatoms', might not be valid in the limit of low T_e for Pt_{13} nanostructures. And finally, our calculations can help in understanding which structural motifs can contribute to the high magnetism for Pt_{13} NP, as found in experiments.



 ISBN: 978-84-614-6167-7

Acknowledgements

This work is supported by the UK research council EPSRC under grant number EP/GO03146/1. It made use of the HPC facilities of the School of Natural and Mathematical Science of King's College of London.

References

[1] Bucher J.P. et al. *Magnetic properties of free cobalt clusters* Phys. Rev. Lett. **66** (1991) 3052; Billas I.M.L. et al. *Magnetism from the Atom to the Bulk in Iron, Cobalt, and Nickel Clusters* Science **265** (1994) 1682

[2] Bader S.D., *Colloquium: Opportunities in nanomagnetism*, RMP **78** (2006) 1.

[3] Liu X. et al., *Structure and Magnetization of Small Monodisperse Platinum Clusters*, Phys. Rev. Lett. **97** (2006) 253401.

[4] Bartolomé J. et al., *Magnetization of Pt_{13} clusters supported in a NaY zeolite: A XANES and XMCD study* , Phys. Rev. B **80** (2009) 014404.

[5] Guimarães A.P., *Principles of Nanomagnetism*, Springer, Berlin (2009).

[6] Kubo R., *Electronic Properties of Metallic Fine Particles .1*, J. Phys. Soc. Jpn. **17** (1962) 975.

[7] Maurat E. et al., *Thermal properties of open-shell metal clusters*, New J. Phys **11** (2009) 103031.

[8] M.E. Gruner et al., *Simulating functional magnetic materials on supercomputers* J. Phys-Cond. Matter, **21** (2009) 293201.

[9] RA. Guirado-Lopez et al., Phys. Rev. Lett. , **90** (2003) 226402

[10] Marzari N. et al., *Thermal contraction and disordering of the Al(110) surface*, J. Phys. Rev. lett. **79** (1997) 1337.

[11] Hu C.H. et al., *Structural, energetic, and electronic trends in low-dimensional late-transition-metal systems*, Phys. Rev. B **79** (2009) 195416.

[12] Wang L.L. et al., *Density functional study of structural trends for late-transition-metals 13-atom clusters*, Phys. Rev. B **75** (2007) 235405.

[13] Futschek T. et al., *Stable structural and magnetic isomers of small transition-metal clusters from the Ni group: an ab initio density-functional study*, J. Phys. Condens. Matter **18** (2006) 9703.

 ISBN: 978-84-614-6167-7

[14] Liu X. et al., *A small paramagnetic platinum cluster in an NaY zeolite: Characterization and hydrogen adsorption and desorption*, J. Phys. Chem. B **110** (2006) 2013.

[15] Giannozzi P. et al., *QUANTUM ESPRESSO: a modular and open-source software project for quantum simulations of materials*, J. Phys. Condens. Matter **21** (2009) 395502.

[16] Perdew J. et al., *Generalized Gradient Approximation Made Simple*, Phys. Rev. Lett. **77** (1996) 3865.

 ISBN: 978-84-614-6167-7

Proceedings of the 11th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2011
26–30 June 2011.

A Lower Bound for Algebraic Side Channel Analysis

Thomas Eisenbarth[1]

[1] *Department of Mathematical Sciences, Florida Atlantic University*

emails: teisenba@fau.edu

Abstract

Recent results have shown significant progress by applying methods from algebraic cryptanalysis in the presence of limited side channel leakage. Unlike most related work, the methods assume a single leakage observation from an implementation of a block cipher. In this work we present two attacks targeting two block ciphers, AES and KeeLoq. The attacks perform equally well as previous work under the same leakage assumptions. However, as the attacks follow a brute-force approach, their setup assumes a less knowledgeable adversary. Hence, it is shown that equal results can be achieved with a simpler setup.

Key words: Algebraic cryptanalysis, side channel cryptanalysis, AES, KeeLoq

1 Motivation

Cryptography has become the cornerstone for building secure digital systems. A remaining threat to embedded cryptographic engines are side channel attacks. The attacks exploit information leaked via physical side channels, including power consumption [7, 8], electromagnetic emanation [10], or even the timing behavior of an implementation [6]. While their applicability is limited to cases where an adversary can sample the corresponding leakage, the attacks usually succeed easily in those cases, unless the implementer invested considerable effort to prevent these attacks.

Most work in side channel analysis focuses on improving the amount of information that can be extracted from several measurements, or on countermeasures suppressing such leakage [8]. The literature distinguishes differential attacks such as Differential Power Analysis (DPA) that exploit the leakage of several samples, and scenarios such as Simple Power Analysis (SPA), where the leakage of a single execution of the cryptographic engine is analyzed. Recently, an increasing number of publications has analyzed how to exploit a strong, well-defined leakage from a single trace, as found in SPAs. Some successful attacks [11, 9] apply advanced cryptanalytic techniques such as algebraic attacks to succeed.

 ISBN: 978-84-614-6167-7

Works in algebraic cryptanalysis attempt to recover the secret key by describing the cipher as a system of algebraic equations and solving these. Usually, equations are built around a single plaintext-ciphertext pair. However, up to now algebraic attacks have not succeeded on any relevant ciphers in use. Instead, they have been successfully applied to toy ciphers and reduced round versions of state-of-the art block ciphers.

One problem is that no intermediate information is accessible for the attacker. This problem can easily be overcome by assuming a side channel adversary, since side channel adversaries get some limited information about intermediate states. While algebraic side channel attacks, unlike classical DPA attacks, are usually not able to statistically exploit leakage of several traces, they offer the beauty of exploiting given information more efficiently. Hence, algebraic attacks seem like a good choice to enhance existing SPA approaches.

Combinations of algebraic attacks and side channel attacks have achieved some unique results. The work by Renauld et al. [11] showed that, given sufficient leakage, even the leakage of as few as three rounds of AES is sufficient to recover the full AES key. More importantly, this attack does not even require knowledge about the processed plaintext or ciphertext. While in classical encryption scenarios at least the ciphertext is assumed to be known, there are some cases, such as pseudo random number generation, where an attacker might be interested in recovering the key even though neither input nor output are known. A similar approach was taken in [9]. Besides targeting AES, an attack on KeeLoq was presented. While the attack on KeeLoq relies on either known input or output, it is, unlike the above attack, able to handle a certain amount of noise in the leakage. Since leakage information is usually derived by sampling noisy physical channels, being able to handle noise is of great advantage.

Our Approach

Attacks that focus on leakage from a single trace have received little attention in literature. This is mainly due to the fact that attacks usually exploit statistical methods which profit greatly from increasing the number of samples. Classical side channel attacks have had great successes by applying divide-and-conquer techniques to recovering the key part-by-part. In this work we apply the same approach on a single leakage trace. We analyze what an adversary can achieve with a single leakage trace and performing exhaustive key search only. In other words, unlike related work we analyze the capabilities of an adversary without applying algebraic attacks. We show that similar results can be achieved by performing techniques used in classical side channel attacks, namely by mere combination of divide-and-conquer techniques with brute-force key search. The goal of the work is to motivate further research effort and to achieve better results with more advanced methods by giving a lower bound on what an adversary is capable of without applying such advanced methods.

 ISBN: 978-84-614-6167-7

2 A Simple Side Channel Cryptanalytic Model

Classical side-channel attacks target cryptographic implementations in a special way. We describe a very general side channel adversary. As depicted in Fig.1, the adversary queries the cryptographic engine and receives an output m and a leakage λ. The output is typically a ciphertext of an unknown (random) message. Given this information, the adversary tries to recover the secret key k.

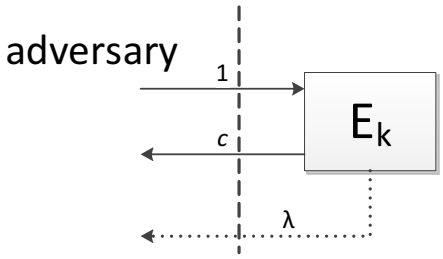


Figure 1: On query 1, the adversary receives the output c of the cryptographic operation E using key k and an unkown random input. In addition, the adversary also receives a leakage λ, which is a function of k and c.

In most scenarios, such as DPA and timing attacks, the adversary gets several queries to the target. However, in this work we will look at the case where only a single trace is available, i.e. one pair $\langle c, \lambda \rangle$ is given to the adversary.

For many block ciphers, including the ones discussed in this paper, it does not matter whether input or output, and plaintext or ciphertext is known. The attacks almost remain the same. However, the amount of leakage can significantly influence the effectiveness of an attack.

The probably most common models for power or EM side channel leakage are the Hamming weight and the closely related Hamming distance. Both have shown to be sufficient as an approximation for the actual leakage when used in correlation based DPA attacks on unprotected implementations. At the same time, it is commonly agreed that actual leakages are somewhat more complex. This is especially important when operating in an SPA scenario, where only one observation of leakage is available. It has been shown that some targets actually allow the detection of Hamming weights through side channel measurements [11]. Hence, a discussion of attacks living entirely in that model have recently attracted significant interest [11, 9].

Given the n-bit value of a number x in binary representation as $x = b_{n-1}b_{n-2}\dots b_0 = \sum_{i=0}^{n-1} b_i \cdot 2^i$, the Hamming weight (HW) is defined as

$$\mathrm{HW}(x) = \sum_{i=0}^{n-1} b_i$$

and the Hamming distance (HD) between two values x and y as $\mathrm{HD}(x, y) = \mathrm{HW}(x + y)$, where $+$ denotes binary addition, also referred to as xor-addition. These leakage functions on parts of the intermediate state are usually observed for power and EM

 ISBN: 978-84-614-6167-7

analysis attacks. Power and EM analysis attacks have received a high attention from the cryptographic community, as they are extremely difficult to completely prevent [8]. We will use the above adversarial model throughout the remainder of this work.

3 Lower-bound Attacks on Block Ciphers

In the following we present two SPA attacks on block ciphers, one on KeeLoq, one on AES. Both attacks are based on the adversarial model described in Sec. 2. The attacks are inspired by the works in [11, 9], where identical adversarial models were assumed. We show that our attacks perform comparably well, but do not require a translation of the cipher into algebraic equations. Instead, they follow a classical combination of the side-channel typical approach of divide-and-conquer with a classic exhaustive search on the partial keys. The combination of both approaches are standard methods in side channel attacks. Yet, they have not been applied in cases where only a single leakage trace is available. The approach has two advantages: The description of the cipher as equations is trivial. Furthermore, changes in algebraic representation do not result in overly complex descriptions of the cipher, which can significantly slow down the algebraic equation solver. Changes in algebraic representation are common in modern block ciphers, as a uniform algebraic structure can simplify cryptanalysis.

3.1 Simple Side-Channel Attack on KeeLoq

KeeLoq is an insecure block cipher that was designed in the 80's. It is mainly used for remote keyless entry systems, where it remains in use. The cipher as well as implementations of it have been thoroughly cryptanalyzed [1, 4, 2, 3, 5]. As shown in [3], the commercially used hardware implementation shows a strong HD leakage. The cipher consists of a 32-bit state and a 64-bit key. We follow the attack setup described in [9]. It assumes a hardware implementation of KeeLoq, showing a noisy HD leakage of the state for each round of the cipher. Such a leakage has been observed from numerous ICs that can be found in various remote keyless entry products featuring KeeLoq.

In particular, the overall leakage λ is a vector of the individual leakages of each state s_i of each of the 528 rounds of the cipher. Each state leakage λ_i gives the HD of the two consecutive states. Additionally, noise δ_i either increases or decreases the observed HD by one with a probability of $\epsilon > 0$:

$$\lambda_i = \mathrm{HD}(s_{i-1}, s_i) + \delta_i$$

Our attack focuses on the first 64 rounds, where one bit of the key is added to the state per round. These observations are very important, as they exhaust the key space entirely. Predicting a leakage in rounds 64 to 528 requires the whole 64 bit key to be guessed. Hence, we will have to use the first 64 rounds to narrow down the number of key candidates to a number that we can handle. The remaining leakages can then be used to single out the correct key from this reduced number of survivors. It is important that we do not throw away the correct key before that. Otherwise, we have to go back,

 ISBN: 978-84-614-6167-7

rerun the first step with a higher number of candidates, hoping that it contains the right key this time.

Attack Description: The followed approach is closely related to the hardware attack described in [3]. The main difference is the replacement of the correlation by a different error metric. We take advantage of the fact that we can directly count the accumulated difference between the observations and the predicted leakage for each key.

One cannot possibly compute the metric for all possible key candidates, as their number doubles with each round that is being analyzed. However, following a maximum-likelihood approach, one can reject unlikely keys throughout the attack. As with other pruning methods, it is important that the correct key is not discarded during the attack. The chosen discarding method removes all keys that (i) show more than one bit deviation from any state leakage or (ii) that result in a metric exceeding the maximum number of expected errors for the correct key. We chose the error bound in such a way that the probability of rejecting the correct key is $< 3\%$.

However, as for the original attack on KeeLoq, if the attack fails, it can be rerun with an increased error-bound, as one will notice the ever-increasing number of predicted errors. Hence, the attack is inherently error-correcting.

Results: The attack has been implemented in Python 3.1 and has been performed on a PC with 4GB memory and an Intel Core i5 processor. Depending on the error probability ϵ, the attack succeeds in comparably little time: with an $\epsilon = 1\%$ it succeeds in under 2 seconds. However, it increases quickly to about 150 seconds at $\epsilon = 6\%$ and 230 seconds at $\epsilon = 10\%$. The memory requirements also increase quickly as the number of errors increases.

3.2 Simple Side Channel Attack on AES

AES is probably the most widely used block cipher today. Having been accepted as NIST symmetric encryption standard, it has become adopted in a wide variety of products. AES implements a substitution-permutation network, with a block size of 128 bit and key sizes of 128, 192 or 256 bit. AES has a state of 16 bytes, s_0 to s_{15}. In each of the rounds, several operations are performed, i.e. `AddKey`, which xor-adds a round key byte to each state byte, `SubBytes`, which substitutes each byte value by a different one, by following a predefined permutation. `ShiftRows` rearranges the order of the state bytes. This operation comes for free on an 8-bit platform. `MixColumns` interprets 4 bytes of the state as a vector and multiplies it with a pre-defined matrix. The output are 4 updated state bytes, where each byte depends on each of the four input bytes.

For our attack we assume an 8-bit implementation that leaks the Hamming weight of each byte of the intermediate AES states. This is a common assumption for power and EM analysis attack on embedded implementations [3, 11, 9] The amount of leakage produced by each operation varies depending on the implementation. The implementation analyzed in [11] shows several leakages for the `MixColumns`, including the leakage

 ISBN: 978-84-614-6167-7

of intermediate results, while `AddKey` and `SubBytes` only leak once. Hence, the amount of leakage from each operation may vary strongly. However, for maintaining simplicity we decided to assume the following leakages: Each byte s_i of the AES state leaks its Hamming weight once per round after the `AddKey`, the `SubBytes` and the `MixColumns`. Hence, we get 3 bytewise Hamming weight leakages of the state per round, resulting in $3 \cdot 16 = 48$ leakages per round. Assuming a leakage from `ShiftRows` does not help, as it repeats the leakage observed after the `SubBytes`. Since it comes for free, it usually is not implemented as a separate operation. Hence, we assume less leakage than the model in [9]. The presented attack assumes virtually noise-free leakages, as done in [9] and [11].

Attack Description: The first leakage we observe in each round is the leakage after the `AddKey`. Each byte $s_i, 0 \leq i < 15$ of that state is the binary sum of a key byte k_i and a byte r_i of the prior state (or the plaintext in the first round), hence $s_i = k_i + r_i$. We only observe the Hamming weight of the state: $\lambda_i = \mathrm{HW}(s_i)$. The observed leakage λ_i narrows down the number of possible values for a particular byte from 256 to a value between 70 and 1, with an average remaining entropy of 5.46. The next leakage we observe is after the `SubBytes`, where the state transitioned to $s'_i = \mathrm{S}(s_i)$ and the observed leakage is $\lambda'_i = \mathrm{HW}(s'_i)$. We can use this leakage to narrow down the number of surviving states one more time. These operations are not very costly, as they can be performed on each byte individually. Hence, for each byte we store less than 70 possible values the byte can take, given the first leakage. Then, we remove all byte candidates that do not comply with the leakage observed after the s-box. This leaves approx. 8 candidates for each byte (remaining entropy: 3.04 bits per byte).

Things get a bit more complicated for the `MixColumns` operation. Each state byte s''_i of the state after the `MixColumns` operation depends on 4 input bytes, i.e. $s''_i = x \cdot s'_j + (x+1) \cdot s'_k + s'_l + s'_m$, where the indices j, k, l, m depend on the index i and $+, \cdot$ are multiplication and addition in $\mathbb{F}_{2^8}$. However, 4 bytes of one column depend on the same 4 bytes of the prior state (from the same diagonal). Hence, instead of having a list of possible bytes, we get a list of possible 4-tuples of bytes $\langle s''_{4i}, s''_{4i+1}, s''_{4i+2}, s''_{4i+3} \rangle$. The average number of surviving guesses per vector is approx. 18 (remaining entropy per byte is 1.04), resulting in approx. $2^{16.7}$ possible states per round. The experimentally observed remaining state entropy varies between 13 to 20 bits.

Since we did not use any information entering the round (neither key k_i nor the previous state r_i are assumed to be known), we can repeat the procedure for any round we need. However, gluing together the rounds and thereby further reducing the number of possible states is not possible, as the round key introduces full entropy at the beginning of each round. Nevertheless, information about two subsequent states r and s can be used to generate a list of possible round key candidates, with $k_i = r''_i + s_i$. Or, even better, three subsequent states r, s, t can produce information about 2 subsequent round keys k, l, with $k_i = r''_i + s_i$ and $l_j = s''_j + t_j$. The relation between the round keys is given by the AES key schedule. Having approx. $2^{16.7}$ candidates for each state gives a total of approx. 2^{50} possible candidates that have to be checked. This, while not undoable, requires special purpose equipment. Instead, one can follow the divide-

 ISBN: 978-84-614-6167-7

and-conquer approach: By checking the relation between k and l byte-by-byte and excluding state/key hypotheses as early as possible. It is preferable to check equations first that depend only on very few independent bytes of the three states. The strategy is as follows:

One possible state of the middle round s is checked at a time. For a given s, we can check the equations given by the key scheduling based on possible values for bytes of r and t. Instead of practically solving the system of equations given by the key schedule, we go through all possible partial values for r and t trying to find matches. This process can be done byte-by-byte, i.e., only the equation for a single byte needs to be checked over and over again. Remember that only candidates that differ in that byte have to be checked. In any of the initial equations, only candidates from at most three columns appear. Hence, the initial equation needs to be checked 2^8 times, resulting in approx. 1 valid match, in which case a second equation needs to be evaluated (including additional key hypotheses). Hence the overall attack has a complexity of approx. $2^{16.7} \cdot 2^8 = 2^{25.4}$, which is easily doable on modern PCs. Remember, these are not 2^{25} full AES encryptions, but only evaluations of simple equations, that mostly consist of binary additions of bytes.

Results: In our experiment we chose to attack the three consecutive rounds of individual encryptions that feature the lowest number of necessary guesses. The average remaining number of round candidates is approx. 2^{15} in that case. The attack performed in less than 30 minutes and recovered the correct key.

Please note that this attack possesses a unique feature. The attack neither requires plaintext nor ciphertext input. It also does not require more than 3 rounds of leakage. Hence, implementations that have a strong leakage, but feature strong countermeasures in the first 3 and last 3 rounds would still be vulnerable to this attack. While similar results were achieved in [11], we did not have to use an algebraic solver for obtaining the same results. Hence, our attack shows better performance while being much simpler to implement. It also requires less cryptanalytic knowledge of the attacker.

4 Conclusion

Two new simple side channel attacks have been presented. For both attacks a similar successful attack has been presented before. However, the presented attacks are simpler and more straightforward than prior work. Both, the attacks presented in [11] and [9] require an algebraic solver to succeed. They furthermore require the cipher to be described in equations that are understandable by the equation solver engine. Even their performance does not seem to outperform the presented results.

The presented attacks are natural extensions of classical DPA approaches. They are leakage-aware brute-force attacks and, thus, define a lower bound for attack performance in the described leakage scenarios. More complex solving methods are usually motivated by outperforming such straightforward approaches. We hope that the results may serve as a motivation for finding more advanced cryptanalytic approaches

 ISBN: 978-84-614-6167-7

that perform significantly better in the described leakage models.

References

[1] A. Bogdanov. Attacks on the KeeLoq Block Cipher and Authentication Systems. In *3rd Conference on RFID Security 2007 (RFIDSec 2007)*. `http://rfidsec07.etsit.uma.es/slides/papers/paper-22.pdf`.

[2] N. T. Courtois, G. V. Bard, and D. Wagner. Algebraic and Slide Attacks on KeeLoq. In *FSE 2008*, volume 5086 of *LNCS*, pages 97–115. Springer, 2008.

[3] T. Eisenbarth, T. Kasper, A. Moradi, C. Paar, M. Salmasizadeh, and M. T. M. Shalmani. On the power of power analysis in the real world: A complete break of the keeloqcode hopping scheme. In *CRYPTO 2008*, pages 203–220, 2008.

[4] S. Indesteege, N. Keller, O. Dunkelman, E. Biham, and B. Preneel. A Practical Attack on KeeLoq. In *EUROCRYPT 2008*, volume 4965 of *LNCS*, pages 1–18. Springer, 2008.

[5] M. Kasper, T. Kasper, A. Moradi, and C. Paar. Breaking KeeLoq in a Flash: On Extracting Keys at Lightning Speed. In *Progress in Cryptology - AFRICACRYPT 2009*, volume 5580 of *LNCS*, pages 403–420. Springer, 2009.

[6] P. C. Kocher. Timing attacks on implementations of Diffie-Hellman, RSA, DSS, and other systems. In N. I. Koblitz, editor, *Advances in Cryptology — CRYPTO '96*, volume 1109 of *Lecture Notes in Computer Science*, pages 104–113. Springer Verlag, 1996.

[7] P. C. Kocher, J. Jaffe, and B. Jun. Differential Power Analysis. In *CRYPTO '99: Proceedings of the 19th Annual International Cryptology Conference on Advances in Cryptology*, pages 388–397, London, UK, 1999. Springer-Verlag.

[8] S. Mangard, E. Oswald, and T. Popp. *Power Analysis Attacks: Revealing the Secrets of Smartcards*. Springer-Verlag, 2007.

[9] Y. Oren, M. Kirschbaum, T. Popp, and A. Wool. Algebraic side-channel analysis in the presence of errors. In S. Mangard and F.-X. Standaert, editors, *Cryptographic Hardware and Embedded Systems, CHES 2010*, volume 6225 of *Lecture Notes in Computer Science*, pages 428–442. Springer Berlin / Heidelberg, 2010.

[10] J.-J. Quisquater and D. Samyde. Electromagnetic analysis (ema): Measures and counter-measures for smart cards. In *E-SMART '01: Proceedings of the International Conference on Research in Smart Cards*, pages 200–210, London, UK, 2001. Springer-Verlag.

[11] M. Renauld, F.-X. Standaert, and N. Veyrat-Charvillon. Algebraic Side-Channel Attacks on the AES: Why Time also Matters in DPA. *Cryptographic Hardware and Embedded Systems—CHES 2009*, pages 97–111, 2009.

 ISBN: 978-84-614-6167-7

Proceedings of the 11th International Conference on Computational and Mathematical Methods in Science and Engineering, CMMSE 2011 26–30 June 2011.

A Free Boundary Problem for Polymer Crystallization in Axisymmetric Samples

Ramón Escobedo[1] and Luis A. Fernández[2]

[1] *Departamento de Matemática Aplicada y Ciencias de la Computación, Universidad de Cantabria, Av. de Los Castros s/n, Santander 39005, Spain*

[2] *Departamento de Matemáticas, Estadística y Computación, Universidad de Cantabria, Av. de Los Castros s/n, Santander 39005, Spain*

emails: escobedo@unican.es, lafernandez@unican.es

Abstract

The crystallization of a hollow cylindrical polymer sample cooled from inside with another cylindrical object of a smaller radius is described by means of a free boundary problem approximation. The cooling object is considered thermally homogeneous in space so that the applied temperature profile is constant with respect to the azimuth, thus allowing a one-dimensional formulation of the problem, provided initial and boundary conditions have axial symmetry. Numerical simulations of the crystallization process are presented for different cooling strategies. Analytical estimates are derived from the pseudo-steady state solution of the corresponding Stefan problem and are shown to be in agreement with the simulations. Analytical estimates are also provided for the crystallization time and the total amount of cold required for the (technically) complete crystallization, and are also shown to be in agreement with the simulations.

Key words: polymer crystallization, Stefan problem, axisymmetric geometry
MSC 2000: 49J20, 35R35, 35K55, 65M06, 80A22

1 Introduction

Finding the optimal cooling strategy is a core problem in polymer crystallization processes. Results in one-dimensional (1D) geometries have been recently reported in [4, 5]. The question arises as to whether the efficiency can be improved by using higher dimensional geometries taking profit of the radial and lateral heat transfer in cylindrical and rectangular samples.

In this paper our previous studies are extended to the case in which a cylindrical cooling object of radius r_c is applied to the interior of a hollow cylindrical sample of larger radius $r_a > r_c > 0$. See Fig. 1.

 ISBN: 978-84-614-6167-7

The internal cooling object is considered thermally homogeneous in space, that is, there is no variation in the temperature with respect to the z-axis and the angular coordinate ψ (the azimuth), so the time-dependent applied temperature profile is constant with respect z and ψ. This axial symmetry is also assumed for the rest of the data and parameters of the problem, in particular the boundary and initial conditions, also chosen to be independent of z and ψ.

The resulting axisymmetric geometry allows to reformulate the problem as a 1D problem for the radial spatial variable $r \in [r_c, r_a]$, often referred to as a $1\frac{1}{2}$D problem.

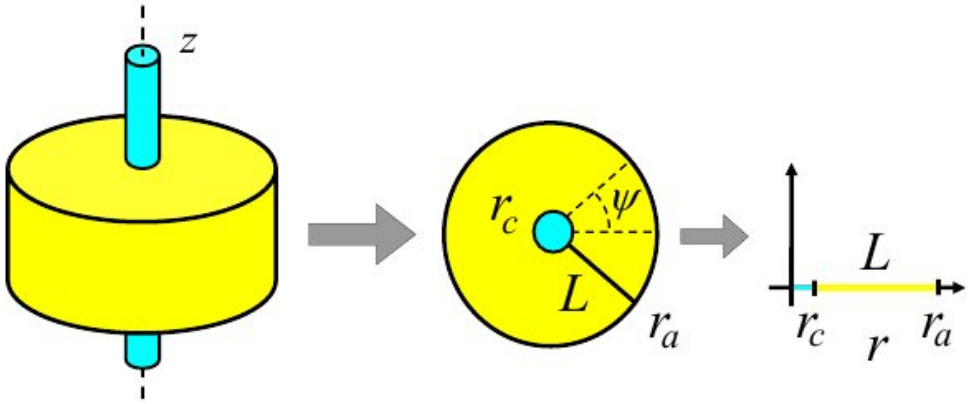


Figure 1: From a cylindrical geometry to the axisymmetric 1D formulation.

The model consists of two non-linear partial differential equations for the degree of crystallinity $y(r,t)$, defined as the mean volume fraction of the space occupied by crystals, and the temperature field $T(r,t)$, coupled by means of the rate functions of nucleation and growth $b_N(T)$ and $b_G(T)$, the function of starting of nucleation $\kappa(y) = (1-y)^2$, and the function of aggregation and saturation of nuclei $\beta(y) = y(1-y)$:

$$y_t(r,t) = \beta(y(r,t))b_G(T(r,t)) + v_0\kappa(y(r,t))b_N(T(r,t)), \tag{1}$$

$$T_t(r,t) = \sigma\left(T_{rr}(r,t) + \frac{1}{r}T_r(r,t)\right) + a_G\beta(y(r,t))b_G(T(r,t)), \tag{2}$$

for $(r,t) \in Q_\tau = (r_c, r_a) \times (0,\tau)$, where $L = r_a - r_c$ is the radial length of the sample and τ is the time at which the cooling process is stopped.

Equations (1)–(2) are solved with the following boundary and initial conditions:

$$T(r_c,t) = u_c(t), \quad T_r(r_a,t) = 0, \qquad t \in (0,\tau), \tag{3}$$

$$y(r,0) = 0, \qquad T(r,0) = T_0, \qquad r \in (r_c, r_a). \tag{4}$$

The cooling object is applied to the inner side of the sample r_c, according to a temperature profile $u_c(t) \in [0, T_f]$ during a total cooling time τ. At the outer side of the sample r_a, we consider a zero-flux condition corresponding to a thermally insulated boundary. These boundary conditions give rise to the *outward* freezing of the sample; *inward* cooling at r_a will not be taken into account in this work.

 ISBN: 978-84-614-6167-7

Let us stress that the part of Eq. (2) corresponding to the heat operator in polar coordinates does not contain the term depending on the angular coordinate ψ due to the axial symmetry of the boundary and initial data (3)-(4).

Under the widely used isokinetic assumption (see [7]), the ratio between nucleation and growth rates is constant, so we choose the nucleation and growth rate functions such that $b_G(T)/G = b_N(T)/N = \theta(T)$, where

$$\theta(T) \stackrel{def}{=} \begin{cases} \exp(-\eta T) & \text{if } T < T_f, \\ 0 & \text{if } T \geq T_f. \end{cases} \tag{5}$$

The parameters $G, v_0, N, \sigma, a_G, \eta$ and T_f are taken as positive real constants denoting the growth factor, the initial mass, the nucleation factor, the heat diffusion coefficient, the non-isothermal factor, the nucleation and growth exponent and the critical phase transition temperature (from liquid to solid), respectively. Typical values are $G = 5$ s^{-1}, $v_0 = 0.01$, $N = 20$ s^{-1}, $\sigma = 0.002$ $m^2 s^{-1}$, $a_G = 2.5 \times 10^3$ ^oC, $\eta = 0.1$ $(^oC)^{-1}$, $T_0 = 100$ ^oC and $T_f = 70$ ^oC. In the 1D case, the length of the sample was $L = 1$ m. More details of the model can be found in Refs. [2], [3] and [4].

2 Numerical simulations and PSS approximation

We have simulated numerically the crystallization process described in (1)-(5) with the above given parameter values, in a sample of inner radius $r_c = 0.1$ m and radial length $L = 1$ m to which a constant temperature $u_c = 40^oC$ is applied until the *technically complete* crystallization is reached, which happens here at $t_{\text{cryst}} = 44.35 \times 10^3$ s.

Fig. 2 depicts the resulting crystallinity and temperature distributions exhibiting the expected crystallization front. The effect of the radial geometry can be noticed in that the temperature profile in the solid phase is no longer the straight line that appeared in the 1D case [4].

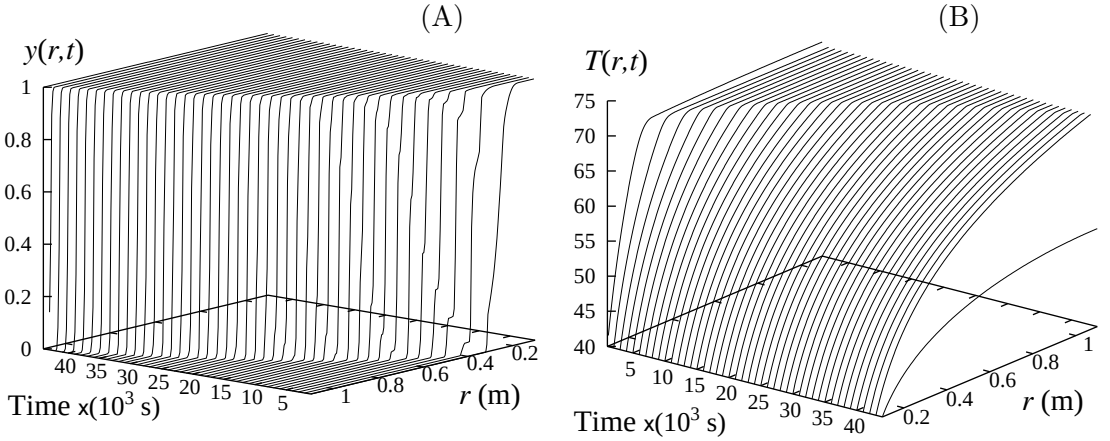


Figure 2: Numerical simulation of the system (1)-(2) in a sample of inner radius $r_c = 0.1$ m and radial length $L = 1$ m cooled with a constant temperature $u_c = 40^oC$. (A) Degree of crystallinity $y(r,t)$. (B) Temperature distribution $T(r,t)$.

 ISBN: 978-84-614-6167-7

A free boundary framework can be used, consisting in identifying the crystallization band with a free boundary $R(t)$ separating the two phases, the solid one $[r_c, R(t))$, where $y \approx 1$, and the liquid one $(R(t), r_a]$, where $y = 0$. In the liquid phase, the temperature is considered constant and equal to the freezing value, that is, $T(r,t) = T_f$ if $r \geq R(t)$.

This framework has been used successfully in the previous works in 1D geometries [4], and has been established rigorously in Ref. [6]. Then, a Stefan condition can be derived, allowing the formulation of the following one-phase Stefan problem [6]:

$$T_t(r,t) = \sigma \left(T_{rr}(r,t) + \frac{1}{r} T_r(r,t) \right), \quad r \in [r_c, R(t)),\ t > 0, \tag{6}$$

$$T(r,t) = T_f, \quad r \in (R(t), +\infty),\ t > 0, \tag{7}$$

$$T(r_c,t) = u_c(t), \quad t > 0, \tag{8}$$

$$T(R(t),t) = T_f, \quad t > 0, \tag{9}$$

$$\frac{\mathcal{L}_\delta}{c} R'(t) = \sigma T_r(R(t),t), \quad t > 0. \tag{10}$$

Here $\mathcal{L}_\delta / c = a_G K_\delta$ is the ratio of the latent heat $\mathcal{L}_\delta$ to the specific heat c, where $K_\delta = [1 + \delta(\ln \delta - 1)]/(1-\delta)^2$ and $\delta = v_0 N / G$. According to Ref. [1], a *pseudo-steady state* (PSS) exists in the limit $Ste \ll 1$, where Ste is the Stefan number, the ratio of the sensible heat $c\Delta T = c \max_t \{T_f - u(t)\}$ to the latent heat $\mathcal{L}_\delta$.

Then, $T_t = 0$ (i.e. $T_{rr} + T_r/r = 0$) and the boundary and initial conditions yield

$$T^{\mathrm{PSS}}(r,t) = \begin{cases} T_f + [u_c(t) - T_f] \dfrac{\ln (r / R^{\mathrm{PSS}}(t))}{\ln (r_c / R^{\mathrm{PSS}}(t))} & \text{if } r \leq R^{\mathrm{PSS}}(t), \\ T_f & \text{if } R^{\mathrm{PSS}}(t) \leq r, \end{cases} \tag{11}$$

where $R^{\mathrm{PSS}}(t)$ is the solution of the following transcendental equation for λ [1, p. 144],

$$2\lambda^2 \ln \left(\frac{\lambda}{r_c} \right) = \lambda^2 - r_c^2 + \frac{4\sigma c}{\mathcal{L}_\delta} Q(t), \tag{12}$$

and $Q(t)$ is the total amount of cold injected into the sample through the internal boundary r_c along the time interval $[0,t]$:

$$Q(t) \stackrel{def}{=} \int_0^t \Big(T_f - u_c(s) \Big)\, ds. \tag{13}$$

We are now interested in estimating the crystallization time t_{cryst} needed for the complete crystallization of the sample, $y(r, t_{\mathrm{cryst}}) \approx 1$, $\forall r$. Although the model recreates the well known feature that full crystallization can not be reached [6], we assume here that the state $y = 1$, $\forall r$ is reached numerically; thus, when the complete crystallization is reached, the total amount of crystallized polymer is $P(t_{\mathrm{cryst}}) = L = r_a - r_c$, where

$$P(t) \stackrel{def}{=} \int_{r_c}^{r_a} y(r,t)\, dr. \tag{14}$$

 ISBN: 978-84-614-6167-7

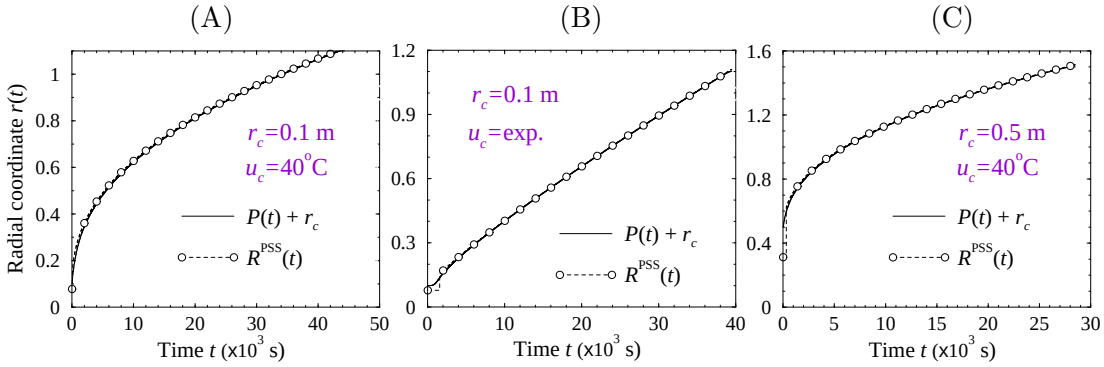


Figure 3: Free boundary $R^{\mathrm{PSS}}(t)$ compared to $P(t)+r_c$ for (A) $r_c = 0.1$ m, $u_c = 40^{\mathrm{o}}$C, (B) $r_c = 0.1$ m, $u_c(t) = T_f + (1 - e^{a_1^2\sigma t})\mathcal{L}_\delta/c$, $a_1 = 1.93 \times 10^{-2}$ m^{-1}, (C) $r_c = 0.5$ m, $u_c = 40^{\mathrm{o}}$C, in samples of radial length $L = 1$ m.

In our free boundary framework, our claim is that the amount of crystallized polymer $P(t)$ can be approximated by means of the position of the crystallization front. That is:

$$P(t) \approx \int_{r_c}^{R(t)} y(r,t)\,dr + \int_{R(t)}^{r_a} y(r,t)\,dr = R(t) - r_c. \tag{15}$$

Fig. 3 shows the excellent graphical agreement between the trajectory of the crystallization front $P(t) + r_c$ and the free boundary $R^{\mathrm{PSS}}(t)$ for three different cooling strategies with different inner radius r_c and different applied temperature $u_c(t)$.

Thus, $R(t_{\mathrm{cryst}}) \approx P(t_{\mathrm{cryst}}) + r_c = r_a$, and expression (12) provides an approximation of the total amount of cold required for the *technically* complete crystallization of the sample defined by r_c and r_a, which allows also to estimate the crystallization time t_{cryst}:

$$Q(t_{\mathrm{cryst}}) \approx \frac{\mathcal{L}_\delta}{2\sigma c}\left[r_a^2\left(\ln\left(\frac{r_a}{r_c}\right) - \frac{1}{2}\right) + \frac{r_c^2}{2}\right]. \tag{16}$$

For a constant applied temperature $u_c(t) \equiv u_c$, this yields, using $Ste = \dfrac{c(T_f - u_c)}{\mathcal{L}_\delta}$:

$$t_{\mathrm{cryst}} \approx \frac{1}{2\sigma Ste}\left[r_a^2\left(\ln\left(\frac{r_a}{r_c}\right) - \frac{1}{2}\right) + \frac{r_c^2}{2}\right]. \tag{17}$$

Note that, for the 1D case, we had $Q^{\mathrm{1D}}(t_{\mathrm{cryst}}) \approx \dfrac{\mathcal{L}_\delta}{2\sigma c}L^2$ and $t_{\mathrm{cryst}} \approx \dfrac{L^2}{2\sigma Ste}$; see [6].

3 Errors estimates

We have used the following error estimates,

$$\xi(t) = P(t) + r_c - R^{\mathrm{PSS}}(t), \tag{18}$$

$$\varepsilon(r,t) = T^{\mathrm{NUM}}(r,t) - T^{\mathrm{PSS}}(r,t), \tag{19}$$

 ISBN: 978-84-614-6167-7

where T^{NUM} is the numerical solution of (1)–(4), and the normalized L_2-norms

$$\xi_{L_2} = \frac{1}{L}\left(\frac{1}{t_2 - t_1}\int_{t_1}^{t_2} \xi^2(t)\,dt\right)^{1/2}, \tag{20}$$

$$\varepsilon_T = \frac{1}{(t_2 - t_1)T_f}\int_{t_1}^{t_2}\left(\frac{1}{L}\int_{r_c}^{r_a} \varepsilon^2(r,t)\,dr\right)^{1/2} dt, \tag{21}$$

and $[0, t_1]$ and $[t_2, t_{\mathrm{cryst}}]$ correspond to the short transient times during which the characteristic band structure is not recognizable and which are located at the beginning and the end of the crystallization process [6].

Table 1: Error estimates and crystallization and transient times.

	u_c (°C)	r_c (m)	ξ_{L_2} (10^{-4})	ε_T (10^{-3})	ϵ (10^{-2})	$t^{\mathrm{NUM}}_{\mathrm{cryst}}$ (10^3 s)	t_1 (10^3 s)	t_2 (10^3 s)	% –
A	40	0.1	1.55	0.94	2.48	44.35	3	43	90.2
B	exp.	0.1	0.5	0.61	1.18	39.47	3	38	88.67
C	40	0.5	1.44	1.5	2.97	28.5	1	27	91.22

We have also compared the numerical crystallization time $t^{\mathrm{NUM}}_{\mathrm{cryst}}$ with its estimated value $t^{\mathrm{PSS}}_{\mathrm{cryst}}$ given by (16) for arbitrary applied temperature profiles $u_c(t)$ and by (17) for constant applied temperatures; this is done by calculating the relative error

$$\epsilon = \left|1 - t^{\mathrm{NUM}}_{\mathrm{cryst}} \Big/ t^{\mathrm{PSS}}_{\mathrm{cryst}}\right|. \tag{22}$$

Error estimates ξ_{L_2}, ε_T and ϵ have been calculated for the three cooling strategies depiected in Fig. 3 for $L = 1$, where (A) $r_c = 0.1$ m, $u_c = 40$°C, (B) $r_c = 0.1$ m, $u_c(t) = T_f + (1 - e^{a_1^2\sigma t})\mathcal{L}_\delta/c$, $a_1 = 1.93 \times 10^{-2}$ m^{-1}, and (C) $r_c = 0.5$ m, $u_c = 40$°C, and are reported in Table 1.

4 Conclusion

This work is a continuation of our previous studies about the polymer crystallization in one-dimensional samples (see [4, 5, 6]) and shows that the same approach can be also used in higher dimensions.

The low values of error estimates presented in Table 1 and the excellent graphical agreement between the numerical and the pseudo-steady state approximations shown in Fig. 3 allow us to say that the polymer crystallization process with axial symmetry is satisfactorily described by our free boundary problem framework, during almost all the time (90%, approx.)

Moreover, the explicit expression for the total amount of cold injected (that can be derived from (12)) provide the key element for the resolution of the corresponding optimal control problem, similar to those considered in [4, 5].

 ISBN: 978-84-614-6167-7

Acknowledgements

This work has been partially supported by the Spanish Ministry of Science and Innovation under grant No. $MTM2008-04206$ and a "Ramón y Cajal" contract.

References

[1] V. Alexiades and A. D. Solomon, *Mathematical modeling of melting and freezing processes*, (Hemisphere Publ. Co., Washington DC, 1993).

[2] V. Capasso, H. Engl and J. Periaux, Eds., *Computational Mathematics Driven by Industrial Problems*, Springer Berlin / Heidelberg, 2000. V. Capasso, *Mathematical models for polymer crystallization processes*, pp 39–67.

[3] R. Escobedo and V. Capasso, *Moving bands and moving boundaries with decreasing speed in polymer crystallization*, Math. Mod. Meth. in Appl. Sci. (M3AS) **15**, No. 3 (2005) 325–341.

[4] R. Escobedo and L. A. Fernández, *Optimal control of chemical birth and growth processes in a deterministic model*, J. Math. Chem. **48** (2010) 118–127.

[5] R. Escobedo and L. A. Fernández, *Optimal cooling strategies in polymer crystallization*, J. Math. Chem. Online first (2011) DOI: 10.1007/s10910-011-9803-x.

[6] R. Escobedo and L. A. Fernández, *A classical one-phase Stefan problem for describing a polymer crystallization process*, (submitted 2011).

[7] A. Fasano, *Mathematical Models in Polymer Processing*, Meccanica **35** (2000) 163–198.

 ISBN: 978-84-614-6167-7

Proceedings of the 11th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2011
26–30 June 2011.

Numerical Remarks on the Preconditioned Conjugate Gradient of the Ocean Dynamics Model OPA

Raffaele Farina[1], Salvatore Cuomo[1] and Marta Chinnici[2]

[1] *Department of Mathematic and Applications "R. Caccioppoli", University of Naples Federico II*

[2] *Marta Chinnici ENEA , Ente Nazionale Energia e Ambiente*

emails: raffaele.farina@unina.it, salvatore.cuomo@unina.it, marta.chinnici@enea.it

Abstract

In this work we analyze the elliptic kernel's solver of the ocean numerical global circulation model OPA. In OPA, one of its dynamical core is rappreseted by the following Laplace's problem:

$$\begin{cases} \dfrac{\partial}{\partial x}\left[\alpha(x,y)\dfrac{\partial}{\partial x}\left(\dfrac{\partial \psi}{\partial t}\right)\right] + \dfrac{\partial}{\partial y}\left[\beta(x,y)\dfrac{\partial}{\partial y}\left(\dfrac{\partial \psi}{\partial t}\right)\right] = f(x,y,t) \quad su\ \Omega \times [t_0,t_1] \\ \dfrac{\partial \psi}{\partial t} = 0 \quad su\ \delta\Omega \times [t_0,t_1] \end{cases} \tag{0.1}$$

The problem's (0.1) discretization, by finite difference, gives N linear systems $Ex = b_m$ $m = 1,..N$ that are solved to determine the dynamic and the thermodynamic variables of ocean fluid from the time t_0 to the time t_1. The aim of this work is to analyze the existent diagonal preconditioner of the algorithm of Preconditioned Conjugate Gradient Method used by Ocean Numerical Model OPA to resolve the N linear systems. We theoretically and numerically show a decreasing of the performance of solver in terms of convergence rate when the domain grid resolution increases and the ratio between the functions α and β changes. Finally, we prove the increment in performance of the Preconditioned Conjugate Gradient Method in OPA by means of others preconditioning techniques.

Key words: Ocean Numerical Global Circulation Model OPA , Laplace's Problem, Finite Difference, Preconditioned Gradient Method.

1 Introduction

The ocean models are a component of global climate models, the climate models are increasingly being used to study not only the climate system but also ocean dynamics. Numerical

 ISBN: 978-84-614-6167-7

ocean circulation models support oceanography and climate science by providing tools to mechanistically interpret ocean observations, to experimentally investigate hypotheses for ocean phenomena, to consider future scenarios such as those associated with human-induced climate warming, and to forecast ocean conditions on weekly to decade time scales using dynamical modeling systems.
The ocean is a forced-dissipative system, with forcing largely at the boundaries and dissipation at the molecular scale. It is contained by complex land-sea boundaries with motions also constrained by rotation and stratification. Flow exhibits boundary currents, large-scale gyres and jets, boundary layers, linear and nonlinear waves, and quasi-geostrophic and three dimensional turbulence. Water mass tracer properties are preserved over thousands of mesoscale eddy turnover time scales. These characteristics of the ocean circulation pose significant difficulties for simulations. Indeed, ocean climate modeling is an application of a very different nature to those found in other areas of computational fluid dynamics (CFD). The time-scales of interest are decades to millennia, yet simulations require resolution or parameterization of phenomena whose time scales are minutes to hours. Furthermore, the most energetic spatial scales are of order 10 km-100 km (mesoscale eddies), yet the problem is fundamentally global in nature. There is no obvious place where grid resolution is unimportant, and computational costs have strongly limited the use of novel, but often more expensive, numerical methods.
In literature there are a lots of different numerical ocean models for Climate and Oceanography projections as Nemo [1], Hope [2], SEA [3], MOM [4] , POP [5] et al. They are all based on the primitive equation [6] given by:

$$\frac{\partial \mathbf{U}_h}{\partial t} = -\left[(\nabla \times \mathbf{U}) \times \mathbf{U} + \frac{1}{2}\nabla(\mathbf{U}^2)\right]_h - f\mathbf{k} \times \mathbf{U}_h - \frac{1}{\rho_0}\nabla_h(p_i + p_s) + \mathbf{D}^{\mathbf{U}} \tag{1.1}$$

$$\frac{\partial p_i}{\partial z} = -\rho g \tag{1.2}$$

$$\nabla \cdot \mathbf{U} = 0 \tag{1.3}$$

$$\frac{\partial T}{\partial t} = -\nabla \cdot (T\mathbf{U}) + D^T \tag{1.4}$$

$$\frac{\partial S}{\partial t} = -\nabla \cdot (S\mathbf{U}) + D^S \tag{1.5}$$

where U is the three-dimensional velocity field, U_h is the horizontal two-dimensional velocity field and the T, S, p_i are the temperature, the salinity and the hydrostatic pressure. Finally D^U D^T and D^S are the diffusion phenomenas and p_s is the surface pressure. The substantial differences in the ocean models [7] are:

- The representation of diffusion phenomenas D^U D^T and D^S and the surface pressure p_s as functions of the previous variables U,T, S and p_i to close the model.
- The choice of vertical coordinates as z-models, ϱ-models σ-models and the choice of horizontal grids , as A-grid B-grid, C-grid and E-grid used for the representation of the physical quantities of dynamical core.
- The use of numerical methods, as explicit, semi-implicit or implicit time stepping schemes, used to discretize (1.1), (1.4) e (1.5) in the primitive equations.

 ISBN: 978-84-614-6167-7

In this context we study the ocean modeling framework NEMO that is composed of *engines* nested in an *environment*. The engines are computational kernels that provide the numerical solutions of ocean NEMO-OPA, sea-ice NEMO-LIM, tracers and biochemistry equations NEMO-TOP and their related physics. The environment are: processing tools, interfaces to the other components of the Earth System, users interface, computer dependent functions and the documentation of the system.
In this paper we investigate one of dynamical cores of the ocean engine OPA in NEMO by numerical point of view, that is represented by Laplace's problem in (0.1). OPA adds the sistem (0.1) at (1.1), (1.2), (1.3), (1.5) e (1.4) to determinate the Surface Pressure Gradient $\nabla_h p_s$. The choice of this study is justified by the fact that the elliptic problem is almost always present in all ocean models to computing the potential in the fluid. The discretization of system (0.1) gives N linear systems that are solved to determine the dynamic and the thermodynamic variables of ocean fluid from the time t_0 to the time t_1 of the Ocean Model. In this paper we observe that the computation kernel used to solve the linear systems is the algorithm of Preconditioned Conjugate Gradient Method and study the influence of the existent diagonal preconditioner P in terms efficiency. Our aim is to characterize numerically the low convergence of the Ocean Numerical Model OPA. This behavior depends on the domain grid resolution and on the relationship of the functions α and β involved in the PDE system (0.1). Furthermore, we propose the use of a different preconditioner obtained by the incomplete Cholesky's decomposition of E that increases the performance of solver in terms of speed of convergence.

2 Preliminaries on Mathematical Method

One of OPA 's the dynamical core is described by the following Laplace's problem [8]:

$$\begin{cases} \dfrac{\partial}{\partial x}\left[\dfrac{e_2}{He_1}\dfrac{\partial}{\partial x}\left(\dfrac{\partial \psi}{\partial t}\right)\right] + \dfrac{\partial}{\partial y}\left[\dfrac{e_1}{He_2}\dfrac{\partial}{\partial y}\left(\dfrac{\partial \psi}{\partial t}\right)\right] = \\ \qquad\qquad = \dfrac{\partial}{\partial x}(e_2 \bar{M}_v) - \dfrac{\partial}{\partial y}(e_1 \bar{M}_u) \quad su\ \Omega \times [t_0\ t_1] \\ \dfrac{\partial \psi}{\partial t} = 0 \qquad\qquad su\ \delta\Omega \times [t_0\ t_1] \end{cases} \tag{2.1}$$

The calculation of the function $\frac{\partial \psi}{\partial t}$, with ψ called Volume Transport Stream function, is needed in the OPA model to determinate the horizontal velocity field (u, v). It is introduced in Nevier Stokes equations (1.1) using the following parameterization:

$$\frac{1}{\varrho_0}\nabla_h p_s = \bar{M} - \frac{1}{H}\left[\vec{k} \times \nabla\left(\frac{\partial \psi}{\partial t}\right)\right] \tag{2.2}$$

where $\nabla_h p_s$ is the Surface Pressure Gradient and ϱ_0 is the average density. The terms e_1, e_2 e e_3 in (2.1) determinate the scale factors when switching from a earth system coordinates to system curvilinear coordinates and H is the surface that parameterizes the ocean floor. Finally, the terms $\bar{M} = (\bar{M}_u, \bar{M}_v)$ represent the contributions of the Coriolis forces, of the

 ISBN: 978-84-614-6167-7

hydrostatic pressure gradient, and advection terms integrated along the z axis with direction $\vec{k}$ that points in center of the Earth [6].
The discretization of problem (2.1) is done through a finite difference scheme of second order of 5 points that gives the following linear systems:

$$Ex = B_m \quad m = 1, .., N \tag{2.3}$$

where E is a sparse matrix, symmetric and positive definite, B_m is known term at the m-th step. Finally N is the total number of system necessary to determine the place of prognostic and diagnostic variables of Ocean Model from time t_0 to the time t_1.
The software ORCA025 [9], [10] that implements OPA at configuration of $1/4^\circ$ uses a two-dimensional grid of 1440×1020 points, an order of magnitude for E of $O(10^6)$ points. Apply a direct method to resolve the systems in (2.3) is impractical and also it doesn't use the sparsity of E where all the elements vanish except those of 5 diagonals. Infact, the sparsity index of E is $SP(E) < \dfrac{n^2 - 5n}{n^2} = 1 - \dfrac{5}{n}$ with n the size of matrix E.
To solve the systems in (2.3) OPA uses the Preconditioned Conjugate Gradient Algorithm (PCG) or Successive-Over-Relaxation (SOR) Iterative Methods. We are going to analyze the (PCG) method because it is fast and easy method to use for a large number of ocean situations (variable bottom topography, complex coastal geometry, variable grid spacing, islands, open or cyclic boundaries). Furthermore it does not require the search of an optimal parameter as in a SOR method. The PCG method [12] determines, by a finite and defined for recurrence succession $\{x_m\} \quad m = 0, .., K$ an approximation of the solution x of systems (2.3) in less than a given accuracy ϵ on the residue:

$$\frac{\|Ex - Ex_K\|}{\|B_n\|} < \epsilon \tag{2.4}$$

Definition 2.1 *Let be E an invertible matrix in $R^{n\times n}$ and let be $||\ .\ ||$ a norm on a space at finite dimension. It is defined condition number of E the following quantity:*

$$\mu(E) = ||E|| \cdot ||E^{-1}|| \quad .$$

Remark 2.1 *The relationship between the speed of convergence of the sequence $x_m \quad m = 1, ..., K$ and condition number $\mu(E)$ is given by:*

$$\|x - x_m\| < 2\left(\frac{\sqrt{\mu(E)} - 1}{\sqrt{\mu(E)} + 1}\right)^{m-1} \|x - x_0\| \quad m = 0, .., K \tag{2.5}$$

By (2.5) we observe that if the condition number $\mu(E)$ increases then the speed of convergence of $\{x_m\} \quad m = 0, ..., K$ decreases while if $\mu(E)$ goes to 1 then the speed of convergence increaces .

 ISBN: 978-84-614-6167-7

As is well known [13] , the matrix E arising from a elliptic problem by finite differences has a high number condition so we need for the preconditioning techniques. In order to study how the choice of preconditioner affects on the performance of the solver in OPA recall some known results.

Remark 2.2 *Let be $\{E_m\}_{m\in\mathbb{N}}$ a succession of invertible matrices3 and E a invertible matrix belonging at $\mathbb{R}^{n\times n}$ with n finite that $E_m \to E$ for $m \to +\infty$ with $\mu(E_m) \geq \mu(E)$. Then $\mu(E_m) \to \mu(E)$ per $m \to +\infty$.*

Proof As $\mu(E_m) - \mu(E) = \|E_m\|\|E_m^{-1}\| - \|E\|\|E^{-1}\| \geq 0 \iff \|E + \Delta E_m\|\|E^{-1} + \Delta E_m^{-1}\| - \|E\|\|E^{-1}\| \geq 0$. By Minkosky's inequality we obtain that:

$$\begin{aligned} &\|E + \Delta E_m\|\|E^{-1} + \Delta E_m^{-1}\| - \|E\|\|E^{-1}\| \leq \\ &\leq \Big(\|E\| + \|\Delta E_m\|\Big)\Big(\|E^{-1}\| + \|\Delta E_m^{-1}\|\Big) - \|E\|\|E^{-1}\| = \end{aligned} \tag{2.6}$$

$$\|E\|\|\Delta E_m^{-1}\| + \|\Delta E_m\|\|E^{-1}\| + \|\Delta E_m\|\|\Delta E_m^{-1}\| \tag{2.7}$$

As $E_m \to E$ then $E_m^{-1} \to E^{-1}$ for $m \to +\infty$. So the (2.7) goes to 0 for $m \to +\infty$ and $\mu(E_m) \to \mu(E)$ for $m \to +\infty$. ■

From Remark (2.2) we deduce that the choice of the preconditioner P must be made that $P^{-1}E$ is nearly at identical matrix I to obtain a conditioning $\mu(P^{-1}E)$ next to 1 so we have to choose P that:

$$\|P^{-1}E - I\| \approx 0 \iff \|P^{-1} - E^{-1}\| \approx 0 \tag{2.8}$$

OPA uses for linear systems (2.3) the diagonal preconditioner P. Note that if E is tridiagonal the difference between the matrices $P^{-1}E$ e I is:

$$P^{-1}E - I = \begin{pmatrix} 0 & e_{1,2}\, e_{1,1}^{-1} & 0 & ... & 0 \\ e_{1,2}\, e_{2,2}^{-1} & 0 & e_{2,3}\, e_{2,2}^{-1} & ... & 0 \\ 0 & e_{2,3}\, e_{3,3}^{-1} & 0 & e_{3,4}e_{3,3}^{-1} & 0 \\ ... & ... & ... & ... & ... \\ 0 & 0 & ... & e_{n-1,n}e_{n,n}^{-1} & 0 \end{pmatrix} \tag{2.9}$$

Using the norm of Fronebius, we get that the distance between $P^{-1}E$ and I is given by:

$$\|P^{-1}E - I\| = \sqrt{\left(\frac{e_{1,2}}{e_{1,1}}\right)^2 + \left(\sum_{i=2}^{n-1}\left(\frac{e_{i,i-1}}{e_{i,i}}\right)^2 + \left(\frac{e_{i,i+1}}{e_{i,i}}\right)^2\right) + \left(\frac{e_{n,n-1}}{e_{n,n}}\right)^2} \tag{2.10}$$

It follows that if the (2.10) is close to zero then $\mu(P^{-1}E)$ is not very high so we have a good acceleration of convergence of the PCG with the diagonal preconditioner P. The (2.10) is close to zero if and only if the following condition (2.11) is true:

 ISBN: 978-84-614-6167-7

$$O(e_{i,i}) >> O(e_{i,j}) \quad i \; e \; j = 1, .., n \;\; j \neq i. \tag{2.11}$$

and the inequality (2.11) must be stronger as the dimension n is large. In the following section we are going to show some theoretical estimates on the speed of convergence of the Preconditioned Conjugate Gradient in OPA , generally valid for all elliptic problems discretized by finite difference, when use the diagonal preconditioner

3 Preconditioning and Convergence

The above considerations hold when we numerically solve the elliptic equations by the Preconditioned Conjugate Gradient Algorithm, in particular also the Laplace problem (2.1) of OPA. Let be consider an approximation of (2.1) using finite differences centered on a finite grid of $p \times q$ points $\Omega_{k,h}$ of Ω, we get the following discrete system:

$$\begin{cases} \delta_x\left[\dfrac{e_2}{He_1}\delta_x\left(\dfrac{\partial\psi}{\partial t}\right)\right] + \delta_y\left[\dfrac{e_1}{He_2}\delta_y\left(\dfrac{\partial\psi}{\partial t}\right)\right] = f(x,y,t_m) \quad on\ \Omega_{k,h}, \quad m = 1, ..., N \\ \dfrac{\partial\psi}{\partial t} = 0 \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad on\ \delta\Omega_{k,h}, \quad m = 1, ..., N \end{cases} \tag{3.1}$$

where δ is the finite difference operator of second order approximating the first derivative. Place the functions $\dfrac{e_2}{He_1} \equiv \alpha(i,j)$ e $\dfrac{e_1}{He_2} \equiv \beta(i,j)$ with $i = 1, ..., q$, $j = 1, ..., p$ and α and β belonging $C^1(\bar{\Omega})$ we get the following linear system:

$$\begin{cases} h^2\left(\alpha_{i+k,j}\dfrac{\partial\psi_{i+k,j}}{\partial t} - \left(\alpha_{i+k,j} + \alpha_{i-k,j}\right)\dfrac{\partial\psi_{i,j}}{\partial t} + \alpha_{i-k,j}\dfrac{\partial\psi_{i-k,j}}{\partial t}\right) + \\ +k^2\left(\beta_{i,j+h}\dfrac{\partial\psi_{i,j+h}}{\partial t} - \left(\beta_{i,j+h} + \beta_{i,j-h}\right)\dfrac{\partial\psi_{i,j}}{\partial t} + \beta_{i,j-h}\dfrac{\partial\psi_{i,j-h}}{\partial t}\right) \\ \qquad = (4hk)^2 f(i,j,t_m) \quad su\ \Omega_{k,h} \quad m = 1, ..., N \\ \dfrac{\partial\psi_{i,j}}{\partial t} = 0 \quad su\ \delta\Omega_{k,h} \quad m = 1, ..., N \end{cases} \tag{3.2}$$

where k and h are the discretization steps in the x and y direction. If $\Omega_{k,h}$ has $p \times q$ points then incomplete matrix E has size n, with $n = p \times q$ and is given by the following:

$$E = \begin{pmatrix} e_{11} & e_{12} & 0 & 0 & e_{1q+1} & 0 & 0 & 0 & 0 \\ e_{21} & e_{22} & e_{23} & 0 & 0 & e_{2q+2} & 0 & 0 & 0 \\ .. & .. & .. & .. & .. & .. & .. & .. & .. \\ e_{mm-q} & .. & e_{mm-1} & e_{mm} & e_{mm+1} & 0 & 0 & e_{mm+q} & 0 \\ .. & .. & .. & .. & .. & .. & .. & .. & .. \\ 0 & 0 & e_{nn-q} & 0 & 0 & 0 & 0 & e_{nn-1} & e_{nn} \end{pmatrix}$$

where its element are for each $m = 1, ..., n$ and $m = (i-1)q + j \quad i = 1, ..., q \; e \; j = 1, ..., p$:

 ISBN: 978-84-614-6167-7

$$\begin{aligned} e_{mm} &= h^2\big(\alpha_{i+1,j} + \alpha_{i-1,j}\big) + k^2\big(\beta_{i,j+1} + \beta_{i,j-1}\big) \\ e_{mm-1} &= -h^2\alpha_{i-1,j}, \quad e_{mm+1} = -h^2\alpha_{i+1,j} \\ e_{mm-q} &= -k^2\beta_{i,j-1}, \quad e_{mm+q} = -k^2\beta_{i,j+1} \end{aligned} \tag{3.3}$$

The following theorem give a result on distance between $P^{-1}E$ and I.

Theorem 3.1 *If $p = q$ and $k = h$, for the following upper and lower bounds hold:*

$$\begin{aligned} &\sqrt{\frac{1}{2}\left(\frac{\alpha_{min}}{\alpha_{max}+\beta_{max}}\right)^2(n-1) + \frac{1}{2}\left(\frac{\beta_{min}}{\alpha_{max}+\beta_{max}}\right)^2(n-\sqrt{n})} \leq \\ &\leq \|P^{-1}E - I\| \leq \sqrt{\frac{1}{2}\left(\frac{\alpha_{max}}{\alpha_{min}+\beta_{min}}\right)^2(n-1) + \frac{1}{2}\left(\frac{\beta_{max}}{\alpha_{min}+\beta_{min}}\right)^2(n-\sqrt{n})} \end{aligned} \tag{3.4}$$

Proof : If we apply the diagonal preconditioner P to the matrix E we get $P^{-1}E$ given by:

$$\begin{pmatrix} 1 & \frac{e_{12}}{e_{11}} & 0 & 0 & \frac{e_{1q+1}}{e_{11}} & 0 & 0 & 0 & 0 \\ \frac{e_{21}}{e_{22}} & 1 & \frac{e_{23}}{e_{22}} & 0 & 0 & \frac{e_{2q+2}}{e_{22}} & 0 & 0 & 0 \\ .. & .. & .. & .. & .. & .. & .. & .. & .. \\ \frac{e_{q+11}}{e_{q+1q+1}} & .. & \frac{e_{q+1q}}{e_{q+1q+1}} & 1 & \frac{e_{qq+2}}{e_{qq}} & 0 & 0 & \frac{e_{q2q}}{e_{q+1q+1}} & 0 \\ .. & .. & .. & .. & .. & .. & .. & .. & .. \\ 0 & 0 & \frac{e_{nn-q}}{e_{nn}} & 0 & 0 & 0 & 0 & \frac{e_{nn-1}}{e_{nn}} & 1 \end{pmatrix}$$

and

$$\|P^{-1}E - I\| = \left\{\left(\frac{e_{12}}{e_{11}}\right)^2 + \left(\frac{e_{1q+1}}{e_{11}}\right)^2 + \sum_{m=2}^{q}\left[\left(\frac{e_{mm-1}}{e_{mm}}\right)^2 + \left(\frac{e_{mm+1}}{e_{mm}}\right)^2 + \left(\frac{e_{mm+q}}{e_{mm}}\right)^2\right] + \tag{3.5}$$

$$+ \sum_{m=q+1}^{n-q}\left[\left(\frac{e_{mm-q}}{e_{mm}}\right)^2 + \left(\frac{e_{mm-1}}{e_{mm}}\right)^2 + \left(\frac{e_{mm+1}}{e_{mm}}\right)^2 + \left(\frac{e_{mm+q}}{e_{mm}}\right)^2\right] + \tag{3.6}$$

$$+ \sum_{m=n-q+1}^{n-1}\left[\left(\frac{e_{mm-q}}{e_{mm}}\right)^2 + \left(\frac{e_{mm-1}}{e_{mm}}\right)^2 + \left(\frac{e_{mm+1}}{e_{mm}}\right)^2\right] + \left[\left(\frac{e_{nn-q}}{e_{nn}}\right)^2 + \left(\frac{e_{nn-1}}{e_{nn}}\right)^2\right]\Bigg\}^{\frac{1}{2}} \tag{3.7}$$

Since the function α and $\beta \in \mathcal{C}^1(\bar{\Omega})$ where $\bar{\Omega}$ is a closed and limited domain, by the theorem of Weistrass exist α_{min}, α_{max}, β_{min} and β_{max} belonging $\mathbb{R}$ that $\alpha_{min} \leq \alpha \leq \alpha_{max}$ and $\beta_{min} \leq \beta \leq \beta_{max}$, furthermore being $-4\alpha\beta < 0$ then α and β have the same sign, so from the (3.3) we majorities and minorities $\|P^{-1}E - I\|$ with the (3.4).

■

The upper and lower bounds in (3.4) show that increasing the size of the problem, the distance between $P^{-1}E$ and I increases indefinitely. So the number condition $\mu(P^{-1}E)$ can be no

 ISBN: 978-84-614-6167-7

longer close to unity as stated in corollary (2.2). The result is loss of the speed of convergence of the PCG solver. Furthermore, by (3.4) we observe that the speed of convergence also depends on the functions α and β of Laplace equation:

$$\begin{cases} \alpha_{max}/\beta_{max} \approx 1 \\ \alpha_{min}/\beta_{min} \approx 1 \end{cases} \implies \tag{3.8}$$

$$\sqrt{\frac{1}{8}\left(\frac{a^2+b^2}{a^2b^2}\right)n - \frac{1}{8b^2}\sqrt{n} - \frac{1}{8a^2}} \leq \|P^{-1}E - I\| \leq \sqrt{\frac{1}{8}(a^2+b^2)n - \frac{1}{8}b^2\sqrt{n} - \frac{1}{8}a^2} \tag{3.9}$$

$$with\ a = \frac{\alpha_{max}}{\alpha_{min}}\ and\ b = \frac{\beta_{max}}{\beta_{min}}$$

instead, still by (3.4), if:

$$\alpha_{max}/\beta_{min} << 1 \implies \sqrt{\frac{1}{8b^2}(n-\sqrt{n})} \leq \|P^{-1}E - I\| \leq \sqrt{\frac{1}{2}b^2(n-\sqrt{n})} \tag{3.10}$$

$$\beta_{max}/\alpha_{min} << 1 \implies \sqrt{\frac{1}{8a^2}(n-1)} \leq \|P^{-1}E - I\| \leq \sqrt{\frac{1}{2}a^2(n-1)} \tag{3.11}$$

so we have in theory a variation of the speed of the solver also when the ratio between α e β changes. We observe from upper and lower bound determined (3.4) (3.9), (3.10) and (3.11) that increasing the size n the best case is when $\alpha_{max}/\beta_{min} << 1$ because $\|P^{-1}E - I\|$ is a $\mathcal{O}(n-\sqrt{n})$ instead the worst case is when $\beta_{max}/\alpha_{min} << 1$ because $\|P^{-1}E - I\|$ is a $\mathcal{O}(n)$. Because of the results showed by upper and lower bounds in (3.4) (3.9), (3.10) and (3.11) in order to speed up the convergence of the PCG in Opa when we increase the resolution of domain $\Omega_{k,h}$ and for all chosen of functions α and β, as E is symmetric and positive definite matrix, we replaced the preconditioner diagonal P with Cholesky's preconditioner:

$$\bar{P} = U^t \cdot U. \tag{3.12}$$

U is a upper triangular matrix with the same structure and sparsity of upper triangolar part of E, obtained by the incomplete Cholesky decomposition algorithm ([14], [15]). Since E is a sparsity matrix ($SP(E) \leq 1 - \frac{5}{n}$) then the Cholesky decomposition algorithm for the calculation of U has time and space complexity in terms of computational cost equal to $\mathcal{O}(n)$. Finally, given the sparsity of U, to solve the linear system $Pz = r_m \Leftrightarrow U^tUz = r_m, \quad m = 0, ...K$ in the PCG [12] through a backward substitution and a forward substitution algorithm, the time complexity is also a $\mathcal{O}(n)$.

In the next section we give same practical experiments of the obtained bounds and same remarks on the Cholesky precondirioner.

 ISBN: 978-84-614-6167-7

4 Numerical Experiments

We give same experiments by implementing **PCG** solvers of OPA on a `Intel Pentium IV 2.3GHz` with `Matlab R14`. To validate those theoretical observations we implements the **PCG** solver with diagonal preconditioner P for a matrix E' sparse, symmetric and positive definite, increasing the size n of E'. We investigate the ratio between α e β function in order to estimate the rate of convergence and accuracy of the PCG solver.
In the first experiment, we assign a tolerance on the residual res:

$$res = \frac{\|E'y - b\|}{\|b\|} < tol = 10^{-12} \quad y \in \mathbb{R}^n,$$

where $\|.\|$ is the Euclidean norm, we gave E', b and tol in input to the software that implements the **PCG** to resolve the linear sistem $E'x = b$. The number of iterations called $iter$, needed to get the tolerance placed on the residual, varying the size n and the ratio between α and β, are shown in the following table (1). We have reported also in the table (1) the relative accuracies between the solution x of the solver **PCG** and the real solution x' of the linear system

$$E'x = b \tag{4.1}$$

n	$iter_{\alpha<<\beta}$	$iter_{\alpha\approx\beta}$	$iter_{\alpha>>\beta}$	$acc_{\alpha<<\beta}$	$acc_{\alpha\approx\beta}$	$acc_{\alpha>>\beta}$
100	28	44	100	$2,85 \times 10^{-12}$	$1,39 \times 10^{-12}$	$6,52 \times 10^{-15}$
1000	237	180	1000	$1,40 \times 10^{-10}$	$2,78 \times 10^{-12}$	$2,45 \times 10^{-13}$
10000	250	350	7554	$1,95 \times 10^{-10}$	$7,82 \times 10^{-11}$	$1,01 \times 10^{-9}$
100000	2800	1642	56552	$1,69 \times 10^{-9}$	$1,04 \times 10^{-10}$	$8,65 \times 10^{-7}$

Table 1: the variable $iter$ is the number of iterations that the solver, with the diagonal precoditioner P, takes to get a tolerance of 10^{-12} on residual res, n is the size of matrix E' and acc is the relative accuracies between x and x' in Euclidean norm.

The numerical experiments confirm the theoretical observations made in (3.9), (3.10) and (3.11). The speed of convergence of the PCG decreases increasing the size n of E'. Moreover, numerical experiments confirm that the speed of convergence also depends on the ratio between functions α and β, in fact fixed the dimension n and the tolerance $tol = 10^{-12}$, it is greater when the magnitude of function α is less than the magnitude of function β $\left(\alpha_{max} << \beta_{min}\right)$ and when they are coincident $\left(\alpha \approx \beta\right)$ while decreases significantly when the magnitude of function α is greater than the magnitude of function β $\left(\alpha_{min} >> \beta_{max}\right)$ as also shown in figure (1).
Finally we show in table (2) and in figure (2) for $n = 10^5$ the results of experiments when PCG uses the Cholesky' preconditioner $\bar{P}$.

 ISBN: 978-84-614-6167-7

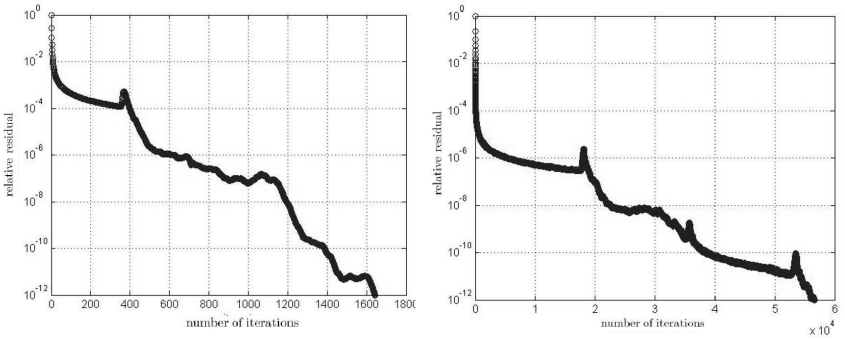


Figure 1: The figure shows the trend of the relative residual res increacing the number of iterations of the PCG with diagonal preconditioner P when ($\alpha \approx \beta$) (left) and ($\alpha_{min} >> \beta_{max}$) (right) for the dimension $n = 10^5$

n	$iter_{\alpha<<\beta}$	$iter_{\alpha\approx\beta}$	$iter_{\alpha>>\beta}$	$acc_{\alpha<<\beta}$	$acc_{\alpha\approx\beta}$	$acc_{\alpha>>\beta}$
100	4	18	6	$6,4258 \times 10^{-15}$	$9,9326 \times 10^{-13}$	$6,7649 \times 10^{-12}$
1000	7	50	15	$5,7072 \times 10^{-11}$	$1,3237 \times 10^{-12}$	$5,1436 \times 10^{-11}$
10000	10	106	61	$7,1298 \times 10^{-11}$	$5,6037 \times 10^{-11}$	$1,8441 \times 10^{-9}$
100000	27	493	429	$3,1311 \times 10^{-9}$	1.0530×10^{-10}	2.3509×10^{-7}

Table 2: the variable $iter$ is the number of iterations that the solver, with Cholesky's preconditioner $\bar{P}^{-1}$, takes to get a tolerance of 10^{-12} on residual res, n is the size of matrix E' and acc is the relative accuracies between x and x' in Euclidean norm.

If we consider the case $n = 10^5$, the tables show that using the diagonal preconditioner with the magnitude of function α greater that function β ($\alpha_{min} >> \beta_{max}$) the number of iterations need to get the tolerance $res = 10^{-12}$ is equal to about 56% of size n of the problem while using the Cholesky preconditioner need only a small number of iterations equal to about 0.5% of size n. We can observe a reduction in the number of iterations of a scaling factor equal to about 112 times.

5 Conclusions

Improve the rate of convergence and the accuracy of the numerical kernels in ocean global mathematical models is a crucial aim of modeling framework for oceanographic research, operational oceanography seasonal forecast and climate studies. In this paper we theoretically and numerically show that the convergence and accuracy depend on the function α and β of the Laplace problem and the size of the discretization resolution. These results can be used in order to construct a new Preconditioner based on the obtained bounds. We held first encouraging results on High Performance Computing ENEA Cresco Grid System in which we optimize the OPA kernel with a new numerical routine.

 ISBN: 978-84-614-6167-7

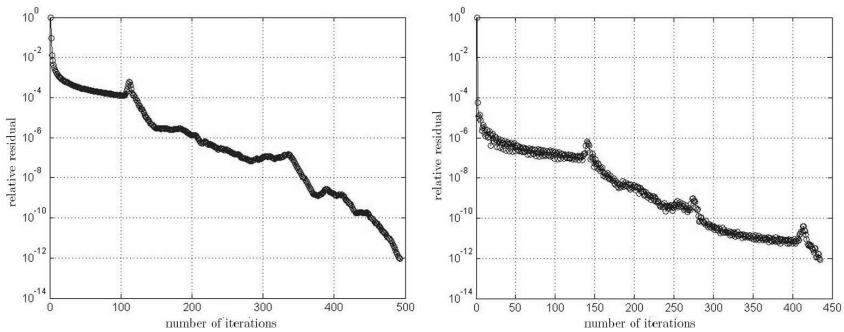


Figure 2: The figure shows the trend of the relative residual res increacing the number of iterations of the PCG when ($\alpha \approx \beta$) (left) and ($\alpha_{min} >> \beta_{max}$) (right) for the dimension $n = 10^5$

References

[1] G.Madec *2008 NEMO reference manual, ocean dynamics component: NEMO-OPA. Preliminary version. Note du Pole de modelisation Institut Pierre-Simon Laplace (IPSL), France, No 27 ISSN No 1288-1619*

[2] D. W. Pierce The Hybrid Coupled Model, Version 3: Technical Notes *.Scripps Instiution of Oceanography Ref. Series No. 96-27, 66 pp. [Available from Climate Research Division, Mail Stop 0224, Scripps Institution of Oceanography, La Jolla, CA 92093-0224, August 96.]*

[3] M.I. Beare, The Southampton - East Anglia (SEA) Model: A General Purpose Parallel Ocean Circulation Model, in: High Performance Computing. *R.J. Allan, M.F. Guest, A.D. Simpson, D.S. Henty and D.A. Nicole, eds, Plenum Publishing Company Ltd., London, 1998*

[4] S. M. Griffies, Elements of MOM4p1, GFDL Ocean Group Tech. Rep. , *Geophys. Fluid Dyn. Lab., NOAA, Princeton, N. J. (2009)*

[5] R.D. Smith and P. Gent. Reference manual for the Parallel Ocean Program (POP). *Los Alamos Unclassified Report LA-UR-02-2484.*

[6] Robert L. Higdon Numerical modelling of ocean circulation. *Acta Numerica (2006), Cambridge University Press, 2006, Pages 385470.*

[7] M. Gries, Claus Boning, F.O. Bryan et al. Developments in ocean climate modelling. *Ocean Modelling Volume 2, Issues 3-4, 2000, Pages 123-192.*

 ISBN: 978-84-614-6167-7

[8] G.MADEC, , P. DELECLUSE, M.IMBARD and M. LEVY "OPA version 8.1, Ocean General Circulation Model, reference manual *Notes du Ple de Modlisation n11, IPSL, 91p, France, 1999*

[9] G.Grieco and S.Masina Implementation of NEMO-OPA in Configuration ORCA-R025 *(November 2009). CMCC Research Paper No. 72. Available at SSRN: http://ssrn.com/abstract=1632851*

[10] I. Epicoco, S.Mocavero, E.Scoccimarro and G.Aloisio. ORCA025: Performance Analysis on Scalar Architecture *(November 1, 2008). CMCC Research Paper No. 50. Available at SSRN: http://ssrn.com/abstract=1370926*

[11] K. Bryan A numerical method for the study of the circulation of the world ocean. *Original Research Article Journal of Computational Physics, Volume 4, Issue 3, October 1969, Pages 347-376*

[12] Y.Saad *,* Iterative Methods for Sparse Linear Systems 2nd*, Society for Industrial and Applied Mathematics Philadelphia, PA, USA 2003*

[13] DL.Young, CC.Tsai, CW. Chen, CM. Fan. The method of fundamental solutions and condition number analysis for inverse problems of Laplace equation. *Computers and Mathematic swith Applications 2008 ; 55 : 11891200.*

[14] David S. Kershaw The incomplete Cholesky conjugate gradient method for the iterative solution of systems of linear equations. Journal of Computational Physics, Volume 26, Issue 1, January 1978, Pages 43-65.

[15] P. CONUS, G. H. GOLUB and D. P. O Leary, A generalized conjugate gradient method for the numerical solution of elliptic partial differential equations, in Studies in Numerical Analysis (G. H. Golub, ed.), MAA Studies in Math., Vol. 24, 1985, pp. 178-198.

 ISBN: 978-84-614-6167-7

Proceedings of the 11th International Conference on Computational and Mathematical Methods in Science and Engineering, CMMSE 2011 26–30 June 2011.

Assessment of a Hybrid Approach for Nonconvex Constrained MINLP Problems

Florbela P. Fernandes[1], M. Fernanda P. Costa[2] and Edite M.G.P. Fernandes[3]

[1] *Department of Mathematics–Polytechnic Institute of Bragança, & CMAT–University of Minho, Portugal*

[2] *Department of Mathematics and Applications, University of Minho, Portugal*

[3] *Algoritmi R&D Center, University of Minho, Portugal*

emails: fflor@ipb.pt, mfc@math.uminho.pt, emgpf@dps.uminho.pt

Abstract

A methodology to solve nonconvex constrained mixed-integer nonlinear programming (MINLP) problems is presented. A MINLP problem is one where some of the variables must have only integer values. Since in most applications of the industrial processes, some problem variables are restricted to take discrete values only, there are real practical problems that are modeled as nonconvex constrained MINLP problems. An efficient deterministic method for solving nonconvex constrained MINLP may be obtained by using a clever extension of Branch-and-Bound (B&B) method. When solving the relaxed nonconvex nonlinear programming subproblems that arise in the nodes of a tree in a B&B algorithm, using local search methods, only convergence to local optimal solutions is guaranteed. Pruning criteria cannot be used to avoid an exhaustive search in the search space. To address this issue, we propose the use of a genetic algorithm to promote convergence to a global optimum of the relaxed nonconvex NLP subproblem. We present some numerical experiments with the proposed algorithm.

Key words: mixed-integer programming, branch-and-bound, genetic algorithm.
MSC 2000: 02.60.Pn

1 Introduction

A wide range of applications in the industrial processes are modeled as nonconvex constrained mixed-integer nonlinear programming (MINLP) problems, due to the restrictions imposed on some problem variables to take only integer values. In particular, one may find applications which include gas network problems, nuclear core reloaded problems, cyclic scheduling trim-loss optimization in the paper industry, synthesis problems,

 ISBN: 978-84-614-6167-7

layout problems, thermal insulation systems. Other examples of this kind of problems appears in engineering designs, in metabolic pathway engineering or in molecular design (see for instance [1, 2, 4, 14, 19]).

In this paper we consider a mixed-integer nonlinear program in the following form:

$$\begin{array}{rl} \min & f(x,y) \\ \text{subject to} & g_j(x,y) \geq 0,\ j \in J \\ & x \in X,\ y \in Y \end{array} \tag{1}$$

where $X = \{x \in \mathbb{R}^n : l_x \leq x \leq u_x\}$ with $l_x, u_x \in \mathbb{R}^n$ and $Y = \{y \in \mathbb{Z}^p : l_y^1 \leq y \leq u_y^1\}$ with $l_y^1, u_y^1 \in \mathbb{Z}^p$. $f : \mathbb{R}^{n+p} \to \mathbb{R}$ and $g : \mathbb{R}^{n+p} \to \mathbb{R}^m$ are continuously differentiable functions, J is the index set of inequality constraints, and x and y are the continuous and discrete/integer variables, respectively. If the objective function f is convex and the constraint functions g_j are concave, the problem is known as convex, otherwise the problem is a non-convex MINLP [3].

It is known that nonconvex MINLP problems are the most difficult since they combine all the difficulties arising from the mixed-integer linear programming and nonconvex constrained nonlinear programming (NLP). Taking into account that this kind of problems appears very frequently in industrial processes, it is fundamental to develop solution techniques to efficiently solve nonconvex constrained MINLP problems [8, 18, 22].

Therefore, the goal of this study is to analyze and propose a method for nonconvex constrained MINLP problems. The proposed hybrid method combines two strategies: a Branch-and-Bound (B&B) method to find integer solutions and a genetic algorithm (GA) type method to promote convergence to global solutions of the relaxed nonconvex NLP subproblems.

The paper is organized as follows. In Section 2, the proposed hybrid method is presented and B&B and GA methods are briefly described. Section 3 presents numerical results for 21 MINLP benchmark problems from the open literature and some conclusions are drawn. Finally, Section 4 presents the major conclusions and future work. The set of test problems used in this work is listed in the appendix.

2 The proposed hybrid method

To solve nonconvex constrained MINLP problems, a B&B-type method is used. Initially developed to solve combinatorial optimization problems, the B&B strategy has evolved to a method for solving more general problems, like for example the MINLP (1). B&B computes lower and upper bounds on the optimal value of f over successively refined partitions of the search space. The generated partition elements are saved in a list. Then they are selected for further processing and partition. The partition elements are deleted when their lower bounds are no lower than the best known upper bound for the problem. While branching on a binary variable creates two subproblems with that variable fixed in both problems, branching on a continuous variable in nonlinear programming may require an infinite number of subproblems. Furthermore, the relaxed

 ISBN: 978-84-614-6167-7

NLP subproblem that appears at each node of the B&B tree search may be nonconvex and it may be difficult to solve to global optimality. Classical gradient-based or even derivative-free local search methods may fail in solving nonconvex NLP problems. Thus, the herein proposed methodology for solving nonconvex MINLP uses a heuristic to solve the relaxed nonconvex NLP subproblems of the B&B tree search.

Existing methods for global optimization can be classified into two categories [15]: stochastic methods and deterministic methods. Stochastic methods sample the objective function for a number of points, with an outcome that is random. They are particularly suited for problems that possess no known structure that can be exploited, and in general do not require derivative information. In these methods, a probabilistic convergence guarantee can be provided. The genetic algorithm is an example of a population-based stochastic method. On the other hand, deterministic methods exploit analytical properties of the problem to generate a sequence of points converging to a global solution. They typically provide a mathematical guarantee for convergence to a minimum in a finite number of steps. The B&B method is a deterministic method.

In this paper, a new methodology to solve problem (1), based on a deterministic method and on a stochastic population-based method is presented. It relies on a B&B scheme and uses a genetic algorithm to promote convergence with a high probability to a global optimum of the relaxed nonconvex NLP subproblem (that arises in each node of the tree in the B&B algorithm). A brief description of the two strategies combined in the herein proposed hybrid method is presented below.

2.1 Branch-and-Bound method

B&B, originally devised for MILP (Mixed Integer Linear Program), can be applied to mixed-integer nonlinear problems too. The reader is referred to one of the first references to nonlinear Branch-and-Bound [9] and also to MINLP problems (see [16, 17] and the references therein included).

The B&B methodology can be explained in terms of a tree search. Initially, all integer variables are relaxed and the resulting relaxed NLP subproblem is solved. If all integer variables take an integer value at the solution then this solution also solves the MINLP. Usually, some integer variables take non-integer values. In that case, a tree search is performed in the space of the integer variables. The B&B algorithm selects one of those integer variables which take non-integer value and branches on it. Branching generates new NLP problems by adding simple bounds respectively to the new relaxed NLP subproblems. After that, one of these new subproblems is selected and solved.

The solution of each subproblem provides a lower bound for the subproblems in the descent nodes of the tree. This process continues until the lower bound exceeds the best known upper bound, the NLP subproblem is infeasible, or the solution provides integer values for the integer variables. The integer solutions (at the nodes of the tree) give upper bounds on the optimal integer solution. This process stops when there are no more nodes to explore.

 ISBN: 978-84-614-6167-7

2.2 The Genetic Algorithm

To solve the nonconvex MINLP problem the herein proposed approach combines the B&B method (described above) and the genetic algorithm, which is used to solve each relaxed nonconvex NLP subproblem that appears in the nodes of the B&B tree. The relaxed NLP subproblem assumes that all variables are continuous and has the form (according to the MINLP model (1)):

$$\begin{array}{rl} \min & f(z) \\ \text{subject to} & g_j(z) \geq 0,\ j \in J \\ & l \leq z \leq u \end{array} \tag{2}$$

where $z = (x, y)$, $z \in \mathbb{R}^m$ and $m = n + p$. Further, the lower and upper bounds of the problem satisfy: $l = \left(l_x, l_y^i\right)$, $u = \left(u_x, u_y^i\right)$ with $i \geq 1$ an integer index and $[l_y^i, u_y^i] \subset [l_y^{i-1}, u_y^{i-1}]$. The used notation means that the index i refers to a problem that is in a descent node of the problem $i - 1$.

A variety of techniques have been proposed to handle inequality constraints in NLP problems. The most widely used techniques rely on penalty functions. Here we are interested in a particular case, known as Lagrangian barrier function. Using the Lagrangian approach to solve the NLP problem in the form (2), a subproblem is formulated by combining the objective function and the nonlinear constraint functions in the barrier function Θ as follows:

$$\Theta(z, \lambda, s) = f(z) - \sum_{j \in J} \lambda_j\, s_j \log(g_j(z) + s_j). \tag{3}$$

where $\lambda_j \geq 0$ represents an estimate of the Lagrange multiplier associated with the constraint $g_j(z) \geq 0$ ($j \in J$) and the components s_j of s are positive and are known as shifts. The general problem (2) is solved by a sequential minimization of the function (3) within the region defined by the simple bounds [7]. The method based on this Lagrangian barrier $\Theta(z, \lambda, s)$ proceeds by fixing λ to some estimate of the optimal Lagrange multipliers and s to some positive estimate of the initial slack variables, and then finding a value of z that approximately minimizes Θ. This new iterate z is then used to update λ, and then s, and the process is repeated. The vector of shifts s depends iteratively on the value of the multiplier vector and on a classical penalty parameter. In order to solve the problem (2), the following general algorithmic framework, as presented in Algorithm 1, is considered [7].

When the subproblem in Step 1 of Algorithm 1 is minimized to a required accuracy (verified in Step 2 of the algorithm), the Lagrangian multipliers are updated. This leads to a new function $\Theta(z, \lambda, s)$ and a new simple bound minimization problem. These steps are repeated until convergence to the optimal solution of (2) is achieved. A detailed description of the algorithm is shown in [6, 7, 13].

This paper is concerned with the use of the genetic algorithm to solve the subproblem that appears in Step 1 of Algorithm 1. The GA may be viewed as an evolutionary process wherein a population of solutions evolves over a sequence of iterations. Genetic algorithm selects individuals at random from the current population to be parents and

 ISBN: 978-84-614-6167-7

Algorithm 1 (Lagrangian barrier method)

Step 0.	Given initial estimates for z and initial estimates for λ and for penalty parameter. Set $k_o = 0$.
Step 1.	Compute shifts s. Use GA to compute z by solving $\min_z \Theta(z, \lambda, s)$ subject to $l \leq z \leq u$.
Step 2.	Test for convergence: Stop, go to Step 3 or go to Step 4.
Step 3.	Update Lagrange multiplier estimates. Maintain the penalty parameter. Increase k_o and go to Step 1.
Step 4.	Reduce the penalty parameter. Maintain the multiplier estimates. Increase k_o and go to Step 1.

uses this individuals to produce the children for the next generation. Over successive iterations and using the common operations of selection, crossover, mutation and scaling, the population "evolves" toward an optimal solution [13]. The basic structure of GA can be summarized as in Algorithm 2.

Algorithm 2 (GA algorithm)

Step 0.	Create a random initial population. Set $k_i = 1$.
Step 1.	Evaluate population using fitness Θ.
Step 2.	While stopping criteria are not satisfied: Select solutions for next population based on their fitness. Perform crossover and mutation. Accept new generation. Evaluate population using fitness Θ. Increase k_i and go to Step 2.

As far as the termination criteria are concerned, GA stops if one of the following conditions holds: the number of iterations exceeds a maximum threshold; the CPU time exceeds a maximum threshold (in seconds); or the cumulative change in the fitness Θ over stall iterations is less than or equal to a function tolerance.

3 Numerical Results

The proposed method (hereafter called BBGA) was implemented in `Matlab` and uses a B&B method combined with the GA method – *ga* function from the Optimization Tolbox of `Matlab`. The *ga* function is called inside the B&B algorithm (to solve the relaxed NLP in each node of the tree)and was run using the default options.

A collection of MINLP problems with inequality constraints and simple bounds is used to analyze the practical behaviour of BBGA. A comparison between this study, a previous work and other results in the literature is presented. The set of tested problems is displayed in the appendix. The problems are denoted by P1, P2, P2a, P3, $\cdots$, P20. For each problem the solver was run 30 times. All experiments were run on a

 ISBN: 978-84-614-6167-7

HP 2230s computer with an Intel(R) Core(TM)2 Duo CPU P7370 2.00GHz processor and 3 GB of memory.

To analyze the behaviour of the BBGA method and to perform a comparison with other results, Table 1 lists all the problems tested in this study "Prob.", the optimal solutions known in the literature,"Optimal f", and the related references, "Ref".

Table 1: Solutions obtained from the literature.

Prob.	**Ref**	**Optimal f**	**Prob.**	**Ref**	**Optimal f**	**Prob.**	**Ref**	**Optimal f**
P1	[10]	6.83325E-13	P7	[10]	-227.766	P14	[21]	-32217.4
P2	[10]	8.0839E-16	P8	[10]	0.0	P15	[12]	-17
P2a	[10]	6.66341E-11	P9	[21]	4.5796	P16	[11]	-8.5
P3	[10]	-0.35239	P10	[20]	2.00	P17	[20]	-5.68
P4	[10]	1.07746E-12	P11	[21]	2.1247	P18	[20]	2.00
P5	[10]	-0.407462	P12	[21]	1.076543	P19	[20]	3.4455
P6	[10]	-18.0587	P13	[21]	3.557473	P20	[20]	2.20

Table 2: Numerical results obtained for MINLPs with simple bounds.

				Average				
Prob.	m	p	**% success**	**time (s)**	**f eval.**	**Best f^***	**Average f^***	σ
P1	2	1	100	0.8	5779	4.930E-15	2.986E-08	2.885E-08
P2	2	1	13	1.0	7291	4.497E-10	1.86656E-08	1.754E-08
			87	1.0	7293	7.659E-11	1.7296E-08	1.751E-08
P2a	4	1	17	1.3	7965	3.984E-08	2.967E-05	3.158E-05
			50	1.1	7765	1.574E-08	1.010E-05	3.159E-05
P3	2	1	67	0.6	4642	-0.35239	-3.524E-01	5.106E-05
P4	2	1	13	1.0	7802	6.692E-13	1.323E-09	9.551E-10
			87	1.1	8403	2.700E-11	4.383E-07	2.136E-06
P5	2	1	33	1.2	4369	-0.40746	-4.075E-01	1.997E-06
			53	1.3	4492	-0.40746	-4.075E-01	4.879E-08
P6	2	1	13	1.4	5271	-18.0587	-1.806E+01	1.582E-07
			83	1.6	6289	-18.0587	-1.806E+01	9.678E-04
P7	2	1	17	2.1	8161	-227.766	-2.278E+02	8.874E-04
			70	2.0	7680	-227.766	-2.278E+02	2.249E-03
P8	2	1	100	0.8	5829	3.659E-10	5.127E-08	1.536E-07

The obtained results with BBGA method are presented in two separate tables. Table 2 contains the results with the problems with simple bounds; Table 4 contains the results from the problems with inequality constraints. In Table 2 and Table 4, the best objective function value "Best f^*" obtained over the 30 runs is reported for each test problem. In order to show more details concerning the quality of the obtained

 ISBN: 978-84-614-6167-7

solution, the average "Average f^*" and the standard deviation "σ" of the best obtained function values are also reported in both tables. Moreover, success rates of obtaining the global minimum "% success", the average numbers of CPU time "time" (in seconds) and function evaluations "f eval.", are shown in columns 4–6. It is also reported, in columns 1–3, the name of the problem, the total number of variables, "m", and the number of integer variables, "p".

From Table 2 it is possible to observe that the set of tested problems covers different situations: problems with small size, problems with or without the same number of integer/continuous variables and problems with one or more than one global minimizer.

As it can be seen, BBGA method has a success rate superior to 85% for 7 problems, out of 9 problems. It is noteworthy that the BBGA method has a success rate equal to 100% for 4 problems. The accuracy of the achieved solutions, in terms of "Best f^*" is very high when compared to the solutions reported in Table 1; the standard deviations of the f values are close to zero.

Table 3 summarizes the results obtained, for this set of problems using a strategy based on a B&B scheme and a simulated annealing heuristic pattern search – BBSAHPS method (see [10] and the references included).

Table 3: Numerical results obtained for Problems P1 – P8 from [10].

	P1	P2	P2a	P3	P4	P5	P6	P7	P8
% success	46	36	87	87	19, 18	49	49	52	42
		26	3		6, 6	40	43	41	
Av. time (s)	0.1	0.1	0.3	0.1	0.1	0.0	0.1	0.1	0.1
			0.2						
Av. f eval.	991	744	1626	966	918, 842	555	610	643	656
		733	1615		859, 865	562	593	573	

It is possible to observe, from Table 2 and Table 3, that the computational cost in terms of CPU time required by the proposed BBGA method, as well as the required number of function evaluations are greater than those of BBSAHPS method. For problem P4, BBSAHPS is able to find four global minimizers, while BBGA finds only two global minimizers. On the other hand the successful rate of BBGA is much better than the successful rate of BBSAHPS: BBGA method has a success rate greater than 67% (problem P2a and P3) and equal to 100% for 4 problems while BBSAHPS has a successful rate less than 50% for three problems.

To address the solution of constrained mixed-integer nonlinear programs, a list of well-known problems, in particular, some taken from *minlplib* [5, 20] is used. The obtained results, with BBGA method, for constrained MINLP are reported in Table 4. The dimension of these problems range between 2 and 7 avriables. The number of inequality constraints range between 1 and 9. This set of problems provides a variety of small MINLP problems.

As it can be seen, BBGA method has a success rate superior to 87% for 8 problems, out of 12 problems. It is noteworthy that the BBGA method has a success rate equal to 100% for 5 problems. Some runs did not converge to the global minimum. For instance,

 ISBN: 978-84-614-6167-7

Table 4: Numerical results obtained for constrained MINLPs.

				Average				
Prob.	m	p	**% success**	**time (s)**	f **eval.**	**Best** f^*	**Average** f^*	σ
P9	7	4	53	13.7	46678	4.5797	4.5987	2.86E-02
P10	2	1	100	2.5	10481	2.00	2.00	6.84E-05
P11	2	1	100	2.4	11527	2.1245	2.1246	2.08E-04
P12	3	1	20	3.3	13635	1.076864	1.078992	2.40E-03
P13	7	4	23	13.5	47282	3.557742	3.564265	7.63E-03
P14	5	2	100	1.3	5220	-32217.4	-32217.4	1.48E-11
P15	2	1	100	3.4	14292	-17	-17	1.80E-04
P16	2	1	97	6.0	25752	-8.5	-8.5	1.55E-04
P17	3	2	87	65.0	247055	-5.6848	-5.684E+00	1.941E-03
P18	4	2	67	3.2	12808	2.000	2.000E+00	8.920E-07
P19	2	1	100	7.8	23489	3.4455	3.446E+00	1.981E-05
P20	4	3	87	4.5	15290	2.20	2.200E+00	5.840E-05

problem P16 has 3% of the runs that have converged to a strong local minimum [11]. Exceptions are problems P12 and P13, which have a successful rate less than 50%. With all these problems the BBGA method could successfully find the global minimum, taking into account the integrality of some variables. The problem P17 requires a larger CPU time when compared with the other problems. This is related with the definition of the problem and the space of the integer variables.

The accuracy of the achieved solutions, in terms of "Best f^*", and for almost all problems, is very good and the consistency of our proposed method is high, since the standard deviations of the f values are close to zero. For problem P11, if a comparison is made in terms of "Best f^*", from Table 1, it is possible to observe that the solution of BBGA method is slightly better than those reported in the literature.

For problem P12 the objective value reported in Table 1 is slightly better than the one obtained with the BBGA. On the other hand, it is possible to observe that BBGA method needs much less functions evaluations than the algorithms used in [21]. In a general way, BBGA method needs much more function evaluations. This is due to the use of a GA method to solve the relaxed NLP problems in each node of the B&B tree search. P17 is the problem with the highest average time and average function evaluations.

Although the chosen problems are small-dimensional, it is important to emphasize that almost all the problems represent real applications of MINLPs.

 ISBN: 978-84-614-6167-7

4 Conclusions and future work

This paper presents a method, herein denoted by BBGA, which relies on a deterministic method and on a stochastic population-based method to solve nonconvex constrained MINLP problems. A Branch-and-Bound procedure is combined with a Lagrangian barrier-function-based genetic algorithm to find the minimum of relaxed nonconvex NLP problems.

The BBGA method was implemented using `Matlab` and some results are shown for 21 test problems, from which 12 are constrained MINLP problems. This method is able to find the global optimum of all problems with integer restrictions on some variables. A comparison between the results from BBGA and the results from the literature is presented. We may conclude that the performance of the BBGA method is quite satisfactory, since the results obtained with BBGA method are competitive with the results reported in the literature.

The reported CPU time is in general high. Thus, an improvement in the BBGA method in order to reduce computational requirements will be carried out in the future. Future developments will be focused, also, on solving sets of medium and large scale problems.

Appendix: Details of test problems

The collection of 21 test problems used in this study is listed below [10, 11, 12, 20, 21].

Problem 1 (P1)

$$\begin{aligned} \min \quad & (1.5 - x_1(1-x_2))^2 + (2.25 - x_1(1-x_2^2))^2 + \\ & (2.625 - x_1(1-x_2^3))^2 \\ & x_1 \in \{-5, \cdots, 5\}, \quad x_2 \in [-4.5, 4.5] \end{aligned}$$

Problem 2 (P2)

$$\begin{aligned} \min \quad & (x_1 - 1)^2 + \sum_{i=2}^{n} i(2x_i^2 - x_{i-1})^2 \\ & x_1 \in \{-10, \cdots, 10\}, \; x_i \in [-10, 10], i = 2...n, \end{aligned}$$

(P2) for $n = 2$;
(P2a) for $n = 4$

Problem 3 (P3)

$$\begin{aligned} \min \quad & 0.25x_1^4 - 0.5x_1^2 + 0.1x_1 + 0.5x_2^2 \\ & x_1 \in [-10, 10], \; x_2 \in \{-10, \cdots, 10\} \end{aligned}$$

Problem 4 (P4)

$$\begin{aligned} \min \quad & \cos^2(x_1) + \sin^2(x_2) \\ & x_1 \in [-5, 5], \quad x_2 \in \{-5, \cdots, 5\} \end{aligned}$$

Problem 5, 6, 7

$$\min \quad 10^m x_1^2 + x_2^2 - (x_1^2 + x_2^2)^2 + 10^{-m}(x_1^2 + x_2^2)^4$$

(P5): $m = 1$ and $x_2 \in [-2, 2], x_1 \in \{-2, \cdots, 2\}$;
(P6): $m = 2$ and $x_2 \in [-4, 4], x_1 \in \{-4, \cdots, 4\}$;
(P7): $m = 3$ and $x_2 \in [-8, 8], x_1 \in \{-8, \cdots, 8\}$

Problem 8 (P8)

$$\begin{aligned} \min \quad & (x_1^2 + x_2 - 11)^2 + (x_1 + x_2^2 - 7)^2 \\ & x_2 \in [-2, 4], \quad x_1 \in \{-2, \cdots, 4\} \end{aligned}$$

Problem 9 (P9)

$$\begin{aligned} \min \quad & (y_1 - 1)^2 + (y_2 - 2)^2 + (y_3 - 1)^2 - \\ & \ln(y_4 + 1) + (x_1 - 1)^2 + (x_2 - 2)^2 + \\ & (x_3 - 3)^2 \\ \text{s. t.} \quad & y_1 + y_2 + y_3 + x_1 + x_2 + x_3 \leq 5 \\ & y_3^2 + x_1^2 + x_2^2 + x_3^2 \leq 5.5 \\ & y_1 + x_1 \leq 1.2 \\ & y_2 + x_2 \leq 1.8 \\ & y_3 + x_3 \leq 2.5 \\ & y_4 + x_1 \leq 1.2 \\ & y_2^2 + x_2^2 \leq 1.64 \\ & y_3^2 + x_3^2 \leq 4.25 \\ & y_2^2 + x_3^2 \leq 4.64 \\ & x_1 \in [0, 1.2], x_2 \in [0, 1.281], \\ & x_3 \in [0, 2.062], \\ & y_1, y_2, y_3, y_4 \in \{0, 1\} \end{aligned}$$

Problem 10 (P10)

$$\begin{aligned} \min \quad & 2x + y \\ \text{s. t.} \quad & 1.25 - x^2 - y \leq 0 \\ & x + y \leq 1.6 \\ & x \in [0, 1.6], y \in \{0, 1\} \end{aligned}$$

Problem 11 (P11)

$$\begin{aligned} \min \quad & -y + 2x - \ln(x/2) \\ \text{s. t.} \quad & -x - \ln(x/2) + y \leq 0 \\ & x \in [0.5, 1.4], y \in \{0, 1\} \end{aligned}$$

Problem 12 (P12)

$$\begin{aligned} \min \quad & -0.7y + 5(x_1 - 0.5)^2 + 0.8 \\ \text{s. t.} \quad & -\mathrm{e}^{x_1 - 0.2} - x_2 \leq 0 \\ & x_2 + 1.1y \leq -1 \\ & x_1 - 1.2y \leq 0.2 \\ & x_1 \in [0.2, 1], x_2 \in [-2.22554, -1], \\ & y \in \{0, 1\} \end{aligned}$$

Problem 13 (P13)

 ISBN: 978-84-614-6167-7

$$\begin{array}{ll}\min & (y_1-1)^2+(y_2-1)^2+(y_3-1)^2-\ln(y_4+1)+\\ & (x_1-1)^2+(x_2-2)^2+(x_3-3)^2\\ \text{s. t.} & y_1+y_2+y_3+x_1+x_2+x_3\leq 5\\ & y_3^2+x_1^2+x_2^2+x_3^2\leq 5.5\\ & y_1+x_1\leq 1.2\\ & y_2+x_2\leq 1.8\\ & y_3+x_3\leq 2.5\\ & y_4+x_1\leq 1.2\\ & y_2^2+x_2^2\leq 1.64\\ & y_3^2+x_3^2\leq 4.25\\ & y_2^2+x_3^2\leq 4.64\\ & x_1\in[0,1.2],x_2\in[0,1.8],x_3\in[0,2.5],\\ & y_1,y_2,y_3,y_4\in\{0,1\}\end{array}$$

Problem 14 (P14)

$$\begin{array}{ll}\min & 5.357854x_1^2+0.835689y_1x_3+37.29329y_1-\\ & 40792.141\\ \text{s. t.} & 85.334407+0.0056858y_2x_3+0.0006262y_1x_2-\\ & 0.0022053x_1x_3\leq 92\\ & 80.51249+0.0071317y_2x_3+0.0029955y_1y_2+\\ & 0.0021813x_1^2-90\leq 20\\ & 9.300961+0.0047026x_1x_3+0.0012547y_1x_1+\\ & 0.0019085x_1x_2-20\leq 5\\ & x_1,x_2,x_3\in[27,45],y_1\in\{78,\cdots,102\},\\ & y_2\in\{33,\cdots,45\}\end{array}$$

Problem 15 (P15)

$$\begin{array}{ll}\min & 3y-5x\\ \text{s. t.} & 2y^2-2y^{0.5}-2x^{0.5}y^2+11y+8x\leq 39\\ & -y+x\leq 3\\ & 2y+3x\leq 24\\ & x\in[1,10],y\in\{1,\cdots,6\}\end{array}$$

Problem 16 (P16)

$$\begin{array}{ll}\min & -x-y\\ \text{s. t.} & xy-4\leq 0\\ & x\in[0,4],y\in\{0,\cdots,8\}\end{array}$$

Problem 17 (P17)

$$\begin{array}{ll}\min & -0.00201x_1^4x_2x_3^2\\ \text{s. t.} & -675+x_1^2x_2\leq 0\\ & -0.419+0.1x_1^2x_3\leq 0\\ & x_1,x_2\in\{1,\cdots,200\},x_3\in[0.1,0.2]\end{array}$$

Problem 18 (P18)

$$\begin{array}{ll}\min & 2+4x_3^2+2x_4+2x_1+2x_2\\ \text{s. t.} & -x_3+3x_4-5\leq 0\\ & 2x_3-x_4-5\leq 0\\ & -2x_3+x_4\leq 0\\ & x_3-3x_4\leq 0\\ & -6x_1+x_3\leq 0\\ & -5x_2+x_4\leq 0\\ & x_1,x_2\in\{0,1\},x_3\in[0,6],x_4\in[0,5]\end{array}$$

Problem 19 (P19)

$$\begin{array}{ll}\min & 1.1\left((2x_1-10)^2+(x-2-5)^2\right)+\\ & \sin\left((2x_1-10)^2+(x-2-5)^2\right)\\ \text{s. t.} & 0.7x_1+x_2\leq 7\\ & 2.5x_1+x_2\leq 19\\ & x_1\in[0,10],x_2\in\{0,\cdots,10\}\end{array}$$

Problem 20 (P20)

$$\begin{array}{ll}\min & 5x_1^2+x_2+x_3+x_4\\ \text{s. t.} & 3x_1-x_2-x_3\leq 0\\ & -x_1+0.1x_3+0.25x_4\leq 0\\ & 2-x_2-x_3-x_4\leq 0\\ & 2-x_2-x_3-2x_4\leq 0\\ & x_1\in[0.2,1],x_2,x_3,x_4\in\{0,1\}\end{array}$$

References

[1] K. Abhishek, S. Leyffer and J. Linderoth, *FilMINT: An Outer-Approximation-Based Solver for Nonlinear Mixed Integer Programs*, Technical Report ANL/MCS-P1374-0906, Mathematics and Computer Science Division, Argonne National Laboratory (2006).

[2] K. Abhishek, S. Leyffer and J. T. Linderoth, *Modeling without categorical variables: a mixed-integer nonlinear program for the optimization of thermal insulation systems*, Optimization and Engineering **11** (2010) 185–212.

[3] L. Biegler and I. Grossmann, *Retrospective on optimization*, Computers and Chemical Engineering **28** (2004) 1169–1192.

[4] P. Bonami, L. Biegler, A. Conn, G. Cornuejols, I. Grossmann, C. Laird, J. Lee,A. Lodi, F. Margot, N. Sawaya and A. Wachter, *An algorihmic framework for convex mixed integer nonlinear programs*, Discrete Optimization **5** (2008) 186–204.

[5] Michael R. Bussieck, Arne Stolbjerg Drud and Alexander Meeraus, *MINLPLib–A Collection of Test Models for Mixed-Integer Nonlinear Programming*, INFORMS Journal on Computing **15(1)** (2003) 114–119.

[6] A. R. Conn, N. I. M. Gould and Ph. L. Toint, *A Globally Convergent Augmented Lagrangian Algorithm for Optimization with General Constraints and Simple Bounds*, SIAM Journal on Numerical Analysis **28(2)** (1991) 545–572.

 ISBN: 978-84-614-6167-7

[7] A. R. Conn, N. I. M. Gould and Ph. L. Toint, *A Globally Convergent Augmented Lagrangian Barrier Algorithm for Optimization with General Inequality Constraints and Simple Bounds*, Mathematics of Computation **66(217)** (1997) 261–288.

[8] R. H. Byrd, J. Nocedal and R. A. Waltz, *KNITRO: An Integrated Package for Nonlinear Optimization in Large-Scale Nonlinear Optimization*, G. di Pillo and M. Roma (Eds.) (2006) 35–59.

[9] R.J. Dakin, *A tree search algorithm for mixed integer programming problems*, Computer Journal **8** (1965) 250–255.

[10] F. P. Fernandes, M. Fernanda P. Costa and Edite M. G. P. Fernandes, *A Deterministic-Stochastic Method for Non-convex MINLP Problems*, Proceedings of 2nd International Conference on Engineering Optimization, H. Rodrigues et al. (Eds.), (2010) CD-ROM (1198) .

[11] C.A. Floudas, *Nonlinear and Mixed-Integer Optimization Fundamentals and Applications*, Oxford University Press, New York, 1995.

[12] C.A. Floudas, P. M. Pardalos, C. S. Adjiman, W. R. Esposito, Z. H. Gumus, S. T. Harding, J. L. Klepeis, C. A. Meyer and C. A. Schweiger, *Handbook of Test Problems in Local and Global Optimization*, Kluwer Academic Publishers, London, 1999.

[13] D. E. Goldberg, *Genetic Algorithms in Search, Optimization & Machine Learning*, Addison-Wesley, 1989.

[14] I. Grossmann, *Review of Nonlinear Mixed-Integer and Disjunctive Programming Techniques*, Optimization and Engineering **3** (2002) 227–252.

[15] J. Lee, *A novel three-phase trajectory informed search methodology for global optimization*, Journal of Global Optimization **38** (2007) 61–77.

[16] S. Leyffer, *Deterministic Methods for Mixed Integer Nonlinear Programming*, PhD Thesis, University of Dundee, 1993.

[17] S. Leyffer, *Integrating SQP and branch and bound for mixed integer nonlinear programming*, Computational Optimization and Applications **18** (2001) 295–309.

[18] S. Leyffer, A. Sartenaer and E. Wanufelle, *Branch-and-Refine for Mixed Integer Nonconvex Global Optimization*, Preprint ANL/MCS-P1547-0908, Mathematics and Computer Science Division, Argonne National Laboratory, 2008.

[19] S. Leyffer, J. Linderoth, J. Luedtke, A. Miller and T. Munson, *Applications and Algorithms for Mixed Integer Nonlinear Programming*, J. Phys.: Conf. Ser. **180** (2009) 1–5.

 ISBN: 978-84-614-6167-7

[20] *MINLPLib–A Collection of Test Models for Mixed-Integer Nonlinear Programming*, `http://www.gamsworld.org/minlp/minlplib/minlpstat.htm` (2011).

[21] T. W. Liao, *Two hybrid differencial evolution algorithms for engineering design optimization*, Applied Soft Computing **10** (2010) 1188–1199.

[22] I. Nowak and S. Vigerske, *LaGO: a (heuristic) branch and cut algorithm for nonconvex MINLPs*, Central European Journal of Operations Research **16** (2008) 127–138.

 ISBN: 978-84-614-6167-7

Proceedings of the 11th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2011
26–30 June 2011.

A mathematical kit for simulating drug delivery through polymeric membranes

J.A. Ferreira[1], P. de Oliveira[1] and P.M. da Silva[2]

[1] *Department of Mathematics, University of Coimbra*

[2] *Department of Physics and Mathematics, Coimbra Institute of Engineering*

emails: ferreira@mat.uc.pt, poliveir@mat.uc.pt, pascals@isec.pt

Abstract

The controlled release of drug through polymeric membranes provides a mechanism with a wide range of applications. A review of the pharmaceutical literature has been carried on and two main classes of controlled drug release devices using polymeric matrices have been identified: ophthalmic therapeutic lenses that deliver drug to the cornea and transdermal drug delivery systems (TDSS) that release therapeutic agents to the systemic circulation or directly to target tissues. The aim of the paper is to present mathematical models that simulate drug release from such polymeric devices. The kinetics of the release is controlled by the diffusion of drug, the binding to immobilizing sites, the degradation of the polymer and the transference mechanisms between drug encapsulations and the matrix. Numerical simulations are included and compared with laboratorial results. A good agreement between both shows the effectiveness of the approach.

Key words: Controlled drug delivery, polymeric matrix, differential equations.
MSC 2000: 35BOJ; 35B30.

1 Introduction

Controlled drug delivery occurs when a polymer is combined with a drug in a such a way that the release profile is predefined. Conventional forms of drug delivery were based on tablets, eye drops, ointments and intravenous solutions. These delivery systems were characterized by an immediate and non controlled kinetics depending essentially on the properties of tissues to absorb drugs. In the last decades drug delivery devices have moved from simple pills to complex controlled systems. Advances in polymer science have led to the development of several novel drug-delivery systems which purpose is to maintain drug concentration in the blood or in target tissues at a desired value and during an extended period of time. The identification of accurate delivery mechanisms, the roles of material properties and the establishment of mathematical models are

 ISBN: 978-84-614-6167-7

essential for attaining the goals of providing scientific guidelines to researchers and product manufacturers. The drug can be dispersed in devices that are made of a single polymeric layer, multiple layers, multiple layers linked by void spaces where drug is entrapped ([1]), and single or multiple layers with particles where drug is also encapsulated ([2], [4], [5]). Some examples are showed in Figure 1.

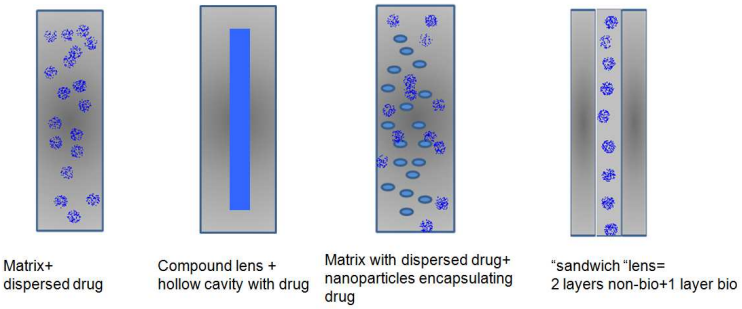


Figure 1: Examples of Single and Multilayer drug delivery systems.

The behavior of the concentration of the free drug in a monolayer device, modeled as a unidimensional platform of width ℓ , $u(x,t)$, can be described by a coupled system of partial differential equations

$$\begin{cases} u_t = (D(u,x,t)u_x)_x + f(u,u^b),\ x \in (0,\ell),\ t > 0 \\ u_t^b = (D(u^b,x,t)u_x^b)_x + g(u,u^b),\ x \in (0,\ell),\ t > 0 \\ u(x,0) = u_0,\ x \in (0,\ell) \\ u^b(x,0) = u_0^b,\ x \in (0,\ell) \end{cases}, \tag{1}$$

where $u^b(x,t)$ represents the concentration of the bound drug, $D(u,x,t)$ stands for the diffusion coefficient of the drug in the polymer, f and g represent reactions terms and u_0, v_0 are the initial free and bound concentrations, respectively.

The basic models presented here can be viewed as a tool kit to build more complex models describing drug release devices. In fact they can be combined to simulate drug delivery from existing devices making it possible to quickly investigate influences of a large number of factors on the efficacy of drug delivery. We mention among other factors influencing the release drug diffusion coefficient, polymer's rate of degradation, number and thicknesses of layers or chemical affinities. Furthermore the models can be also used as a tool kit to help devise administration strategies and reduce the number of pharmaceutical experiments in new devices. For example layers with dispersed drug can be combined with layers containing particles encapsulating drug or layers linked by void spaces where drug is encapsulated. We note that exterior enhancement factors as heat sources, used in heat-aided TDSS ([8]), or iontophoresis ([9]) have not been took into account.

System (1) is completed with boundary conditions. In order to simulate *in vitro*

 ISBN: 978-84-614-6167-7

experiments or *in vivo* release two types of boundary conditions can be considered

$$\begin{cases} D(u,0,t)u_x(0,t) = \alpha(u(0,t) - u_{ext}),\ t > 0 \\ -D(u,\ell,t)u_x(\ell,t) = \alpha(u(\ell,t) - u_{ext}),\ t > 0 \end{cases}, \tag{2}$$

or

$$\begin{cases} D(u,0,t)u_x(0,t) = 0,\ t > 0 \\ -D(u,\ell,t)u_x(\ell,t) = \alpha(u(\ell,t) - u_{ext}),\ t > 0 \end{cases}, \tag{3}$$

where u_{ext} represents the exterior drug concentration and α stands for a transfer coefficient.

Conditions (2) simulate *in vitro* experiments where devices are placed in a vial and the drug is released through both extremities. Conditions (3) simulate *in vivo* experiments as one of the extremities is insulated. For TDDS the impermeable layer is a polymer film farthest from the skin; for therapeutic ophthalmic lens the impermeable layer is a polymer film farthest from the cornea.

In Section 2 different models of drug delivery from therapeutic contact lens are established. We begin by presenting models that simulate drug delivery from simple layer devices: a non biodegradable layer and a biodegradable film. In order to introduce some delay in the delivery process we present a model which describes the delivery from a polymeric lens with dispersed drug and containing particles encapsulating drug. This lens has been first presented in [4] and a modified version has been recently described in [2], [3] and [7]. A delay in the delivery can be also induced by multilayer type lens. A new model that simulates the behavior of a multilayer prototype recently described in [1] is also proposed. Numerical simulations obtained using real data in our models show a good agreement with laboratory experiments. Combining the platforms described in the previous section we present in Section 3 models for Transdermal Drug Delivery Systems. The models simulate the behavior of commercialized systems ([6]). In Section 4 some conclusions are addressed.

2 Mathematical models for therapeutical lenses

In this section several models to simulate drug delivery from therapeutic ophthalmic lenses are presented. In the case of monolayers lenses the models are obtained as special cases of (1). The simplest model is a polymeric film-which can be biodegradable or non biodegradable- with dispersed drug. To induce a delay in the delivery the release mechanisms are modified leading to more complex systems. Two main mechanisms are described in the literature: the inclusion in the polymeric platform of particles encapsulating drug ([2], [3], [4], [5]) and the coupling of three layers in a sandwich form [1]. Multilayer models are obtained by combining models of type (1),(2) or (1),(3) completed with interface conditions between the layers. The numerical simulations presented in this paper are obtained following a method of lines approach, where spatial derivatives are discretized with finite differences and time integration with appropriate multistep methods.

 ISBN: 978-84-614-6167-7

A simple polymeric membrane with dispersed drug.

In the case of a non biodegradable platform, where no binding occurs, $u^b = 0$, $f = 0$, $g = 0$, a simple diffusion model is obtained from (1). If there is chemical affinity between the drug and the polymer, the following system (Model I) is obtained

$$\begin{cases} u_t = Du_{xx} - k_1 u + k_2 u^b, \ x \in (0,\ell), \ t > 0 \\ {u^b}_t = k_1 u - k_2 u^b, \ x \in (0,\ell), \ t > 0 \end{cases} . \tag{4}$$

where u represents the free drug concentration, u^b the bound drug concentration and the parameters k_1 and k_2 stand for binding and unbinding rates respectively. In (4) the reaction term has been defined following [10] and [11]. Model (4) is completed with initial conditions

$$\{ \ u(x,0) = u_0, \ u^b(x,0) = u_0^b, \ x \in (0,\ell) \quad , \tag{5}$$

and boundary conditions

$$\begin{cases} Du_x(0,t) = \alpha(u(0,t) - u_{ext}), \ t > 0 \\ -Du_x(\ell,t) = \alpha(u(\ell,t) - u_{ext}), \ t > 0 \end{cases} , \tag{6}$$

which simulate *in vitro* delivery, where the drug leaves the platform through the two extremities.

If a biodegradable platform is used and no binding is considered (Model II), the model is obtained from (1) with $u^b = 0$, $f(u,x,t) = c_0\gamma e^{-\gamma t}$, $g = 0$, giving

$$\begin{cases} u_t = (D(t)u_x)_x + c_0\gamma e^{-\gamma t}, \quad D(t) = D_2 e^{-\beta_2 e^{-\gamma t}}, \ x \in (0,\ell), \ t > 0 \\ u(x,0) = u_0, \ x \in (0,\ell) \end{cases} , \tag{7}$$

where c_0 represent the concentration of drug initially bound to the polymer and β, γ are positive parameters.

System (7) is completed with boundary conditions analogous to (6).

Platform with drug and dispersed particles encapsulating drug.

As cornea tissues have a limited absorption capacity it is crucial to delay the drug delivery process. A delay mechanism has been proposed in [2]. Here particles encapsulating drug are dispersed in the polymeric platform that contains itself drug (Model III). We note that this lens whether conceptually analogous to the lens presented in [4], [5] was prepared with different polymers and techniques. Also the mathematical models simulating the behavior of the two lenses are different.

In (1) let u represent the concentration of the drug in the platform and u^b the concentration in the particles. Then *in vitro* release is described by

$$\begin{cases} u_t = Du_{xx} - \lambda u + \lambda u^b, \ x \in (0,\ell), \ t > 0 \\ {u^b}_t = \lambda u - \lambda u^b, \ x \in (0,\ell), \ t > 0 \\ u(x,0) = u_0, \ u^b(x,0) = u_0^b, \ x \in (0,\ell) \\ Du_x(0,t) = \alpha(u(0,t) - u_{ext}), \ t > 0 \\ -Du_x(\ell,t) = \alpha(u(\ell,t) - u_{ext}), \ t > 0 \end{cases} , \tag{8}$$

where λ stands for the transfer coefficient between the particles and the platform and α represents a partition coefficient. We note that this model can be obtained from (4)-(6) with $k_1 = k_2 = \lambda$.

 ISBN: 978-84-614-6167-7

Multi-layer models.

A second mechanism to induce delay in drug delivery from ophthalmic lenses has been presented in [1]. The idea lies in creating sandwich type structures composed by three polymeric layers as represented in Figure 1: two non biodegradable layers (HEMA) coating a biodegradable PLGA film containing drug (Model IV). The behavior of the drug release is modeled by the coupled diffusion-reaction system of partial differential equations

HEMA layer

$$\begin{cases} u_t = (D_1 u_x)_x, \quad D_1(u) = D_{1e} e^{\beta_1 (1-u/v_0)}, \ x \in (0, \ell_1), \ t > 0 \\ u(x,0) = 0, \ x \in (0, \ell_1) \\ D_1 u_x(0,t) = \alpha(u(0,t) - u_{ext}), \ t > 0 \\ D_1 u_x(\ell_1, t) = D_2 v_x(\ell_1, t), \ t > 0 \end{cases}, \tag{9}$$

PLGA film

$$\begin{cases} v_t = (D_2 v_x)_x + c_0 \gamma e^{-\gamma t}, \quad D_2(t) = D_{2e} e^{-\beta_2 e^{-\gamma t}}, \ x \in (\ell_1, \ell_2), \ t > 0 \\ v(x,0) = v_0, \ x \in (\ell_1, \ell_2) \\ v(\ell_1, t) = \beta u(\ell_1, t), \ t > 0 \\ D_2 v_x(\ell_2, t) = D_1 w_x(\ell_2, t), \ t > 0 \end{cases}, \tag{10}$$

HEMA layer

$$\begin{cases} w_t = (D_1 w_x)_x, \quad D_1(w) = D_{1e} e^{\beta_1 (1-w/v_0)}, \ x \in (\ell_2, \ell_3), \ t > 0 \\ w(x,0) = 0, \ x \in (\ell_2, \ell_3) \\ w(\ell_2, t) = \beta v(\ell_2, t), \ t > 0 \\ -D_1 w_x(\ell_3, t) = \alpha(w(\ell_3, t) - u_{ext}), \ t > 0 \end{cases}. \tag{11}$$

In (9)-(11) u and w represent the drug concentration in the non biodegradable layers, v represent the drug concentration in the biodegradable PLGA film, D_{1e} and D_{2e} stand for the initial diffusion coefficients in HEMA and PLGA, respectively. We note that ℓ_i are the thicknesses of the different layers, v_0 and c_0 are the free and bound initial concentrations in PLGA, respectively. Parameters α and β are related with the flux conditions at the boundary and at the interfaces, respectively; β_1 and β_2 are positive parameters. We note that apart from initial and boundary conditions, interface conditions are also included in (9)-(11). We assume that binding is not significant in HEMA layers.

The authors in [1] also report experiments carried with a different type of sandwich structure: two HEMA layers linked by a void space containing drug (Model V). The kinetics of the release can be described by equations (9), (11) and an evolution equation in the void space of type

$$\begin{cases} v_t = -D_1 u_x(\ell_1, t) - D_1 w_x(\ell_1 + \epsilon, t), \ t > 0 \\ v(0) = v_0 \end{cases}. \tag{12}$$

In Table I we present a synthesis of the five models previously described.

 ISBN: 978-84-614-6167-7

Table I: Description of the models.

Models	Definition
Model I	Drug delivery from a non biodegradable film, where binding can occur
Model II	Drug delivery from a biodegradable film
Model III	Drug delivery from a polymeric matrix with particles encapsulating drug
Model IV	Drug delivery from the lens of "sandwich" type
Model V	Delivery of drug entrapped between two polymeric layers

We present in Figure 2 the plots of the total released masses, corresponding to models I, II, III, IV and V with boundary conditions of type (2) that simulate *in vitro* results.

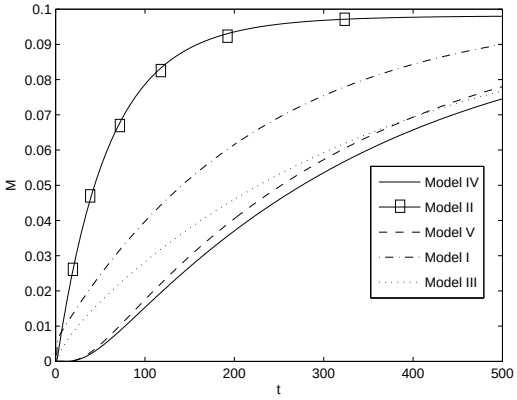


Figure 2: Comparison of Models I, II, III and IV.

We considered $u_{ext} = 0$, $\alpha = 0.01$, in all simulations and the values of the parameters exhibited in Table II.

We note that if the drug is entrapped in a single non biodegradable layer where binding can occur (model I) the release is faster than in models IV and V, but slower than in model II. We conclude that "sandwich platforms" - models IV and V - lead to a slower drug release than "non sandwich platforms" - models I, II and III.

Table II: Parameters of simulation in Figure 2.

Models	Parameters
Model I	$D = 0.005$, $k_1 = 0.8$, $k2 = 0.6$, $u_0^b = 0.01$, $u_0 = 0.04$, $\ell = 2$
Model II	$D_2 = 0.03$, $\beta_2 = 0.001$, $\gamma = 0.01$, $c_0 = 0.01$, $u_0 = 0.09$, $\ell = 1$
Model III	$D = 0.005$, $\lambda = 0.05$, $u_0^b = 0.01$, $u_0 = 0.04$, $\ell = 2$
Model IV	$D_{1e} = 0.005$, $D_2e = 0.03$, $\beta_1 = 0.002$, $\beta_2 = 0.001$, $\gamma = 0.01$, $c_0 = 0.01$, $v_0 = 0.09$, $\ell_1 = \ell_2 = \ell_3 = 1$
Model V	$D_{1e} = 0.005$, $\beta_1 = 0.002$, $v_0 = 0.1, \ell_1 = \ell_3 = 1$

In Figure 3-left- we compare the total released mass of model IV for two different degradation coefficients, and in Figure 3-right- we illustrate the behaviour of released

 ISBN: 978-84-614-6167-7

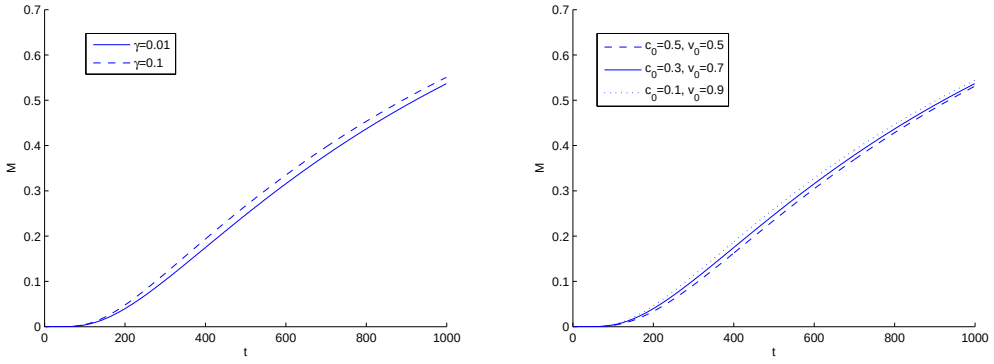


Figure 3: A comparison of released mass from model IV for two different degradation coefficients γ (with $c_0 = 0.3$, $v_0 = 0.7$) -left- and different free and bound initial concentration -right- (with $\gamma = 0.1$).

mass for different free and bound initial concentration. In these simulations the following values were used: $u_{ext} = 0$, $D_{1e} = 0.001$, $D_{2e} = 0.02$, $\alpha = 0.01$, $\beta_1 = 0.02$, $\beta_2 = 0.02, \ell_1 = \ell_2 = \ell_3 = 1$. From the figure in the left we conclude that the delivered mass is an increasing function of γ. In fact as the polymer erodes the bound drug is free to diffuse through the HEMA layers and the largest is the degradation rate the fastest is the release. The influence of initial concentration is also illustrated in the right of Figure 3: for each t the total released mass is a decreasing function of the initial bound mass.

We observe that the values used for the parameters do not correspond to physical values.

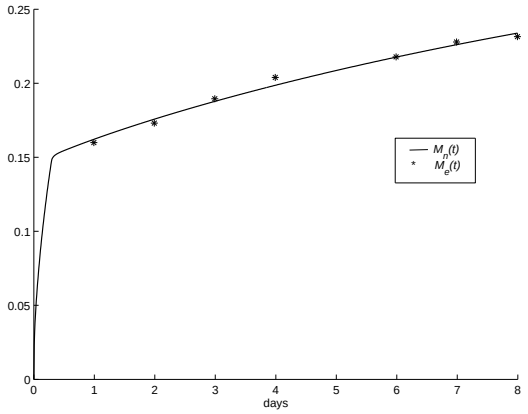


Figure 4: Comparison of delivered numerical ($M_{1,n}$) and experimental ($M_{1,e}$) masses for model III.

 ISBN: 978-84-614-6167-7

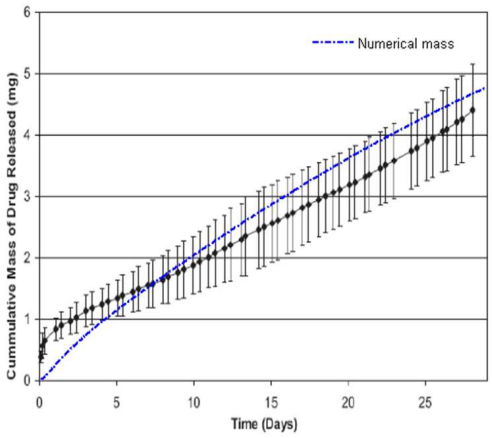


Figure 5: Comparison of delivered numerical and experimental masses for model IV.

We compare now experimental results with the corresponding numerical simulations obtained with models III and IV. In Figure 4 we plot the results obtained with model III using experimental values for the parameters: $u_0^b = 0.05102$, $u_0 = 0.28$, $\alpha = 0.01, \lambda = 0.02$, and

$$D(t) = \begin{cases} 0.1996 \times 10^{-3},\ t \in [0, 420] \\ \\ 0.9 \times 10^{-5},\ t \in (420, 11520] \end{cases} . \tag{13}$$

To represent more realistically the exterior concentration we defined $u_{ext}(t) = \gamma_1 u(-\ell, t)$ with $\gamma_1 = 0.5$. This plot shows a good agreement with the experimental results presented in [2].

In Figure 5 numerical simulations of model IV are compared with laboratorial results in [1]. We consider $u_{ext} = 0$, $D_{1e} = 0.8554$, $D_{2e} = 4.2336 \times 10^{-7}$, $\alpha = 0.5$, $\beta_1 = 1.5$, $\beta_2 = 0.1$, $\gamma = 0.0714$, $c_0 = 0.03475$, $v_0 = 0.1$, $\ell_1 = 0.02$, $\ell_2 = 0.01$, $\ell_3 = 0.02$. As referred in [1] after 30 days the lens is still releasing drug. The qualitative behaviour of the numerical prediction shows a good agreement after day 5. We note that the experimental results exhibit an initial burst that is not present in the numerical solution. This is a point deserving some attention. In fact if there was no drug at all in the HEMA layers ([1]) this initial burst is not expectable. This argument suggests that the non biodegradable layers are not completely drug free.

3 Transdermal Drug Delivery Systems

The aim of this section is to present a multi layer system to deliver drug through the skin. The device contains two layers, with dispersed drug and containing particles also containing drug, and a void membrane to be put in contact with the skin.

We begin by comparing the delivery *in vitro* of one single layer described by equation (8) and its behavior *in vivo* described by

 ISBN: 978-84-614-6167-7

$$\begin{cases} u_t = Du_{xx} - \lambda u + \lambda u^b, \ x \in (0,\ell), \ t > 0 \\ {u^b}_t = \lambda u - \lambda u^b, \ x \in (0,\ell), \ t > 0 \\ u(x,0) = u_0, \ u^b(x,0) = u_0^b, \ x \in (0,\ell) \\ Du_x(0,t) = 0, \ t > 0 \\ -Du_x(\ell,t) = \alpha(u(\ell,t) - u_{ext}), \ t > 0 \end{cases}, \tag{14}$$

where u represents the free drug concentration and u^b stands for the drug concentration in particles. We recall that D is the diffusion coefficient, λ is the transfer coefficient, α stands for the distribution coefficient and u_{ext} the exterior concentration.

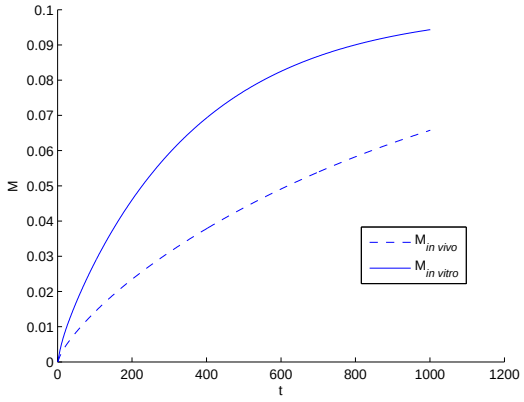


Figure 6: Comparison of the delivery masses obtained from system (8) -*in vitro* release- and system (14) -*in vivo* release-.

In Figure 6 we compare the plots obtained from systems (8) and (14), with $D = 0.005$, $u_{ext} = 0$, $\alpha = 0.01$, $\lambda = 0.05, \ell = 2$, $u_0^b = 0.01$, $u_0 = 0.09$. As expected the delivery process is slower *in vivo* than *in vitro*, where the release occurs through both extremities.

Let us consider now the multilayer system built from two models of type III, represented by (14), linked by an interface condition as described in

$$\begin{cases} u_t = Du_{xx} - \lambda u + \lambda u^b, \ x \in (0,\ell_1), \ t > 0 \\ {u^b}_t = \lambda u - \lambda u^b, \ x \in (0,\ell_1), \ t > 0 \\ u(x,0) = u_0, \ u^b(x,0) = u_0^b, \ x \in (0,\ell_1) \\ Du_x(0,t) = 0, \ t > 0 \\ Du_x(\ell_1,t) = Dv_x(\ell_1,t), \ t > 0 \\ v_t = Dv_{xx} - \lambda v + \lambda v^b, \ x \in (\ell_1,\ell_2), \ t > 0 \\ v_t^b = \lambda v - \lambda v^b, \ x \in (\ell_1,\ell_2), \ t > 0 \\ v(x,0) = v_0, \ v^b(x,0) = v_0^b, \ x \in (\ell_1,\ell_2) \\ -Dv_x(\ell_2,t) = \alpha(v(\ell_2,t) - u_{ext}), \ t > 0 \end{cases}. \tag{15}$$

In (15) u represents the drug concentration in the first layer, $[0,\ell_1]$, and v the drug concentration in the second layer, $[\ell_1,\ell_2]$; u^b, v^b stand for the drug concentrations in

 ISBN: 978-84-614-6167-7

the particles in each one of the layers.

The interest of these type of platforms lies in the possibility of using different initial conditions in the two layers. This leads to a customization of the delivery which can be tailored to a specific treatment. In Figure 7 we illustrate the effect of the use of two different concentrations. We considered $D = 0.5$, $u_{ext} = 0$, $\alpha = 0.5$, $\lambda = 0.1$, $\ell = 0.5$ in both layers. The simulations resulting from two set of values for the concentration are compared. The first set is defined by $u_0^b = 0.01$, $u_0 = 0.09$ and $v_0^b = 0.1$, $v_0 = 0.9$; in the second set $u_0^b = 0.1$, $u_0 = 0.9$ and $v_0^b = 0.01$, $v_0 = 0.09$.

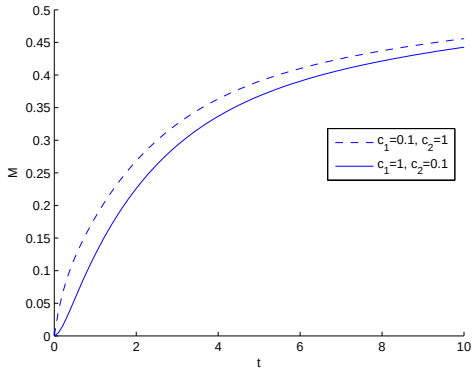


Figure 7: Mass for two different initial concentrations, obtained from system (15), where $c_1 = u_0 + u_0^b$ and $c_2 = v_0 + v_0^b$.

We note that the global concentration is the same in the two simulation. However these concentrations are differently distributed. As expected the release is faster when the layer closest to the skin is the most loaded.

Some commercial platforms [6] contain a membrane in contact with the skin- that introduces a further delay in the drug release. This membrane is initially void. The model that simulates the device can be described from models of type IV,

$$\begin{cases} u_t = Du_{xx} - \lambda u + \lambda u^b,\ x \in (0,\ell_1),\ t > 0 \\ u_t^b = \lambda u - \lambda u^b,\ x \in (0,\ell_1),\ t > 0 \\ u(x,0) = u_0,\ u^b(x,0) = u_0^b,\ x \in (0,\ell_1) \\ Du_x(0,t) = 0 \\ Du_x(\ell_1,t) = Dv_x(\ell_1,t) \\ v_t = Dv_{xx} - \lambda v + \lambda v^b,\ x \in (\ell_1,\ell_2),\ t > 0 \\ v_t^b = \lambda v - \lambda v^b,\ x \in (\ell_1,\ell_2),\ t > 0 \\ v(x,0) = v_0,\ v^b(x,0) = v_0^b,\ x \in (\ell_1,\ell_2),\ t > 0 \\ -Dv_x(\ell_2,t) = D_m C_x^m(\ell_2,t),\ t > 0 \\ C_t^m = D_m C_{xx}^m,\ x \in (\ell_2,\ell_3),\ t > 0 \\ C^m(x,0) = 0,\ x \in (\ell_2,\ell_3) \\ -D_m C_x^m(\ell_3,t) = \alpha(C^m(\ell_3,t) - u_{ext}),\ t > 0 \end{cases}, \tag{16}$$

where C^m represents the drug concentration in the membrane and D_m stands for its

 ISBN: 978-84-614-6167-7

diffusion coefficient.

The delay effect of the initially void membrane can be observed in Figure 8 where numerical simulations of systems (15) and (16) are plotted, with $D = D_m = 0.5$, $u_{ext} = 0$, $\alpha = 0.5$, $\lambda = 0.1, \ell = 0.5$, $u_0^b = v_0^b = 0.1$, $u_0 = v_0 = 0.9$.

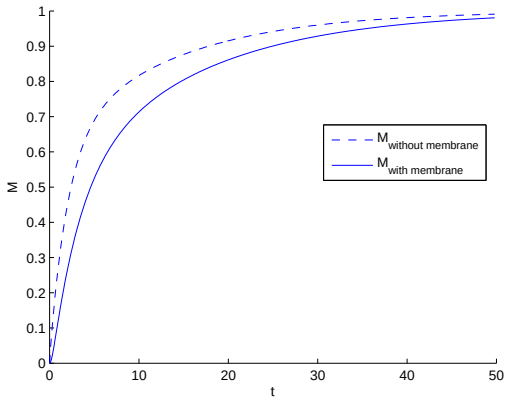


Figure 8: Numerical simulations of systems (15) and (16).

4 Conclusion

A mathematical kit was proposed to simulate drug delivery trough polymeric membranes. Two main class of devices have been considered: ophthalmic therapeutic lenses and Transdermal Drug Delivery Systems. The developed procedure allows researchers to investigate influences of a large of number of factors on the efficacy of drug delivery.

As far as therapeutic lenses are concerned we conclude that the delay induced by sandwich type structures is more efficient that the encapsulation of drug in particles. In the case of TDDS multilayer platforms are analyzed. Namely the effect of using different concentrations in the layers and the delay effect of a void membrane, closest to the skin, are addressed.

Acknowledgements

This work has been partially supported by the Center of Mathematics of the University of Coimbra (CMUC).

References

[1] J. B. Ciolino, T. R. Hoare, N. G. Iwata, I. Behlau,C. H. Dohlman, R. Langer, D. S. Kohane, *A Drug-Eluting Contact Lens*, Investigative Ophthalmology & Visual Science, **50** (2009) 7.

 ISBN: 978-84-614-6167-7

[2] J.A. Ferreira, P. Oliveira, P. M. Silva, A. Carreira, H. Gil, J. N. Murta, *Sustained drug release from contact lens*, Computer Modeling in Engineering and Science, **60**, (2010), 152–179.

[3] J.A. Ferreira, P. Oliveira, P. M. Silva, J. N. Murta, *Drug delivery: from a ophthalmic lens to the anterior chamber*,Computer Modeling in Engineering and Science, (2011).

[4] D. Gulsen, A. Chauhan, *Ophthalmic drug delivery from contact lenses*, Investigative Ophthalmology and Visual Science, **45**, (2004), 2342–2347.

[5] G. Gulsen, A. Chauhan, *Dispersion of microemulsion drops in HEMA hydrogel: a potential ophthalmic drug delivery vehicle*, International Journal of Pharmaceutics, **292**, (2005), 95–117.

[6] E. B. Nauman, K. Patel, P. Karande, *On the design and optimization of diffusion-controlled planar delivery devices*, Chemical Engineering Science, **65**, (2010), 23–930.

[7] P. M. Silva, *Controlled Drug Delivery: Analytical and Numerical Study* , Phd-Thesis, University of Coimbra, Portugal, (2010).

[8] L. Simon,*Analysis of heat-aided membrane controlled drug released from a process control prespective*, Int. J. of Heat and Mass Transfer, **50**, (2007), 2425–2433.

[9] L. Simon, *Timely drug delivery from controlled-release devices: Dynamic analysis and novel design concepts*, Mathematical Bioscience, **217**, 2009, 151–158.

[10] R.A. Siegel, E.L. Cussler, *Reactive barrier membranes: some theoretical observations regarding the time lag and breakthrough curves*, J. Memb. Sci., **229**, (2004), 33–41.

[11] R.A. Siegel, *Characterization of relaxation to steady state in membranes with binding and reaction*, J. Memb. Sci., **251**, (2004), 91–99.

 ISBN: 978-84-614-6167-7

Proceedings of the 11th International Conference on Computational and Mathematical Methods in Science and Engineering, CMMSE 2011 26–30 June 2011.

A Non Fickian single phase flow model

José A. Ferreira[1] **and Luís Pinto**[1]

[1] *CMUC, Department of Matehmatics, University of Coimbra, Portugal*

emails: ferreira@mat.uc.pt, luisp@mat.uc.pt

Abstract

A single phase incompressible flow problem is usually modeled by a system of three equations:a differential equation for the velocity, an algebraic equation linking the velocity and the pressure and a parabolic equation for the concentration depending on the velocity. Some limitations have been pointed out in the literature on the use of a parabolic equations to describe the the concentration evolution, namely related with the use of Fick's law to describe the mass flux. To avoid the pathologic behavior of the classical diffusion equation, non Fickian corrections have been proposed in the literature. In this paper we introduce a new model to describe a single phase incompressible flow problem and its stability will be studied. A numerical method that mimics the continuous model is also studied and some numerical experiments are included.

Key words: Fikian model, Non Fickian model, Diffusion, pressure, concentration, numerical method, stability.

1 Introduction

A single phase incompressible flow problem is usually modeled by a system of three equations: a differential equation for the pressure, an algebraic equation linking the velocity and the pressure and a parabolic equation for the concentration depending on the velocity (see [1], [2], [4], [5], [9] and [10]). This system can be rewritten as a system of an elliptic equation for the pressure and a parabolic equation for the concentration that depends on the gradient of the pressure.

Traditionally, the a diffusion process in a porous medium is described by the parabolic equation

$$\frac{\partial u}{\partial t} + \nabla J = q_2, \tag{1}$$

where u denotes the concentration, J represents the mass flux and q_2 denotes the reaction term. In (1) J can be expressed as

$$J = J_{adv} + J_{dif} + J_{dis}, \tag{2}$$

 ISBN: 978-84-614-6167-7

where

$$J_{adv} = uv \tag{3}$$

represents the advection mass due to the the fluid velocity v,

$$J_{dif} = -D_m \nabla u \tag{4}$$

denotes the mass flux due to molecular diffusion, being D_m the effective molecular diffusion coefficient, and J_{dis} satisfies the so called Fick's law $J_{dis} = -D_d \nabla u$ and represents the dispersive mass flux associated with random deviations of fluid velocities within the porous space from their macroscopic value v. In the definition of J_{dis}, D_d denotes the dispersive tensor $D_d = \alpha_t \|v\| I + (\alpha_\ell - \alpha_t)\frac{1}{\|v\|} vv^t$ being α_ℓ and α_t the longitudinal and transversal dispersivities.

Combining (1) with (2) we obtain the parabolic equation

$$\frac{\partial u}{\partial t} + \nabla(uv) = \nabla((D_m I + D_d)\nabla u) + q_2, \tag{5}$$

where I is the identity tensor.

Some limitations have been pointed out in the literature on the use of a parabolic equation (5) to describe the concentration evolution (see, for instance, [3], [6], [8]): equation (5) prescribes an infinite speed of propagation for the concentration; it is based on Fick's law for the mass flux which establishes a linear relation between the concentration and dispersive mass flux; the mass flux J is independent of the history of dispersion; in the dispersive tensor the dispersivities coefficients are medium constant and invariant with time and space (often they increase with the distance and/or with time).

To avoid the pathologic behavior of the classical diffusion equation, hyperbolic or non Fickian corrections have been proposed in the literature (see [6], [8]). One possible approach is to consider that the dispersive mass flux satisfies the following differential equation

$$\tau \frac{\partial J_{dis}}{\partial t}(x,t) + J_{dis}(x,t) = -D_d \nabla u(x,t), \tag{6}$$

where τ is a delay parameter ([7]). We remark that the left hand side of (6) is a first order approximation of the left hand side of $J_{dis}(x, t+\tau) = -D_d \nabla u(x,t)$, which means that the dispersion mass flux at the point x and time $t+\tau$ depends on the gradient of the concentration at the same point but at a delayed time. Equation (6) leads to

$$J_{dis}(t) = -\frac{1}{\tau}\int_0^t e^{-\frac{t-s}{\tau}} D_d \nabla u(s)\, ds, \tag{7}$$

provided that $J_{dis}(0) = 0$. Combining the partition (2), where J_{adv}, J_{dif} and J_{dis} are given by (3), (4) and (7), respectively, with (1) we obtain the integro-differential equation

$$\frac{\partial u}{\partial t} + \nabla(uv) - \nabla(D_m \nabla u) = \frac{1}{\tau}\int_0^t e^{-\frac{t-s}{\tau}} \nabla(D_d \nabla u)(s)\, ds + f \tag{8}$$

which replaces (5).

 ISBN: 978-84-614-6167-7

In this paper we consider the following system of equations:

$$-\nabla(A(u)\nabla p) = q_1 \text{ in } (a,b)\times(0,T], \tag{9}$$

$$\begin{aligned}\frac{\partial u}{\partial t} &+ \nabla(B(u,\nabla p)u) + \nabla(D_m(u,\nabla p)\nabla u)\\ &= \int_0^t k_{er}(t-s)\nabla(D_d(u,\nabla p)\nabla u)(s)\,ds + q_2 \text{ in } (a,b)\times(0,T],\end{aligned} \tag{10}$$

where q_1 and q_2 are source terms, A, D_m and D_d are smooth enough satisfying the following assumptions:

$$0 < A_0 \le A(x,y), (x,y) \in \mathbb{R}\times\mathbb{R}, \tag{11}$$

$$|B(x,y)| \le C_B|y|, (x,y) \in \mathbb{R}\times\mathbb{R}, \tag{12}$$

$$0 < D_{m,0} \le D_m(x,y), (x,y) \in \mathbb{R}\times\mathbb{R}, \tag{13}$$

$$0 < D_{d,0} \le D_d(x,y) \le D_{d,1}|y|, (x,y) \in \mathbb{R}\times\mathbb{R}. \tag{14}$$

In (10) $K_{er}(s)$ denotes a kernel that satisfies some assumptions that will be specified later but it can be in particular defined by $K_{er}(s) = \frac{1}{\tau}e^{-\frac{s}{\tau}}$. System (9), (10) is complemented by Dirichelet boundary conditions

$$p(a,t) = p_a(t), p(b,t) = p_b(t), u(a,t) = u_a(t), u(b,t) = u_b(t),\ t\times]0,T], \tag{15}$$

and initial conditions

$$p(x,0) = p_0(x), u(x,0) = u_0(x), x \in (a,b). \tag{16}$$

The paper is organized as follows. In Section 2 we study the stability of the initial boundary value problem (9), (10), (15), (16). A numerical method that mimics the initial boundary value problem (9), (10), (15), (16) will be presented in Section 3 and its stability will be analized. In Section 4 we include some numerical experiments illustrating the behavior of the pressure and concentration when we replace (5) by (10).

2 Stability analysis

By $H^1(a,b)$ and $H_0^1(a,b)$ we denote the usual Sobolev spaces with the usual norm $\|.\|_1$. By $(.,.)$ we represent the usual inner product defined in $L^2(a,b)$ and $\|.\|$ denotes the norm induced by such inner product. By $L^2(0,T;H^1(a,b))$ we denote the space of functions $v : (0,T) \to H^1(a,b)$ such that $\displaystyle\int_0^T \|v(s)\|_1^2\,ds < \infty$. We also consider the space $\mathcal{W}(0,T) = \{v \in L^2(0,T;H^1(a,b)) : \frac{dv}{dt} \in L^2(0,T;L^2(a,b))\}$, where $L^2(0,T;L^2(a,b))$ is the space of functions $v : (0,T) \to L^2(a,b)$ such that $\displaystyle\int_0^T \|v(s)\|^2\,ds < \infty$.

 ISBN: 978-84-614-6167-7

We study in what follows the stability of the solution (u,p), $u \in \mathcal{W}(0,T)$, $p \in L^2(0,T;H^1(a,b))$, that satisfies the variational equations

$$(A(u(t))\nabla p(t), \nabla w) = (q_1, w),\ \forall w \in H_0^1(a,b), \tag{17}$$

$$\begin{aligned}(\frac{du}{dt}(t), w) - (B(u(t), \nabla p(t))u(t), \nabla w) + (D_m(u(t), \nabla p(t))\nabla u(t), \nabla w)\\ = -(\int_0^t K_{er}(t-s)D_d(u(s), \nabla p(s))\nabla u(s)\, ds, \nabla w) + (q_2(t), w),\ \forall w \in H_0^1(a,b),\end{aligned} \tag{18}$$

almost everywhere in $(0,T]$, the boundary conditions (15) and the initial conditions (16) with $p_0, u_0 \in L^2(a,b)$.

As this section focuses in the stability analysis of initial boundary value problem (9), (10), (15), (16) we assume that $p_a(t) = p_b(t) = u_a(t) = u_b(t) = 0$, $t \in]0,T]$. We also introduce the space $L^2(0,T;H_0^1(a,b))$ which is obtained from $L^2(0,T;H^1(a,b))$ replacing $H^1(a,b)$ by $H_0^1(a,b)$. By $\mathcal{W}_0(0,T)$ we denote the subspace of $\mathcal{W}(0,T)$ that is obtained replacing $H^1(a,b)$ by $H_0^1(a,b)$.

Theorem 1 *Let us suppose that $p_0, u_0 \in L^2(a,b)$, A, B, D_m and D_d satisfy the conditions (11), (12), (13) and (14), respectively. If the solution (u,p) of (17), (18) with initial conditions (16) is in $\mathcal{W}_0(0,T) \times L^2(0,T;H_0^1(a,b))$, then*

$$\|\nabla p(t)\|^2 \leq \frac{(b-a)^2}{2A_0^2}\|q_1\|^2, \tag{19}$$

$$\|u(t)\|^2 + \int_0^t \|\nabla u(s)\|^2\, ds \leq e^{-\tilde{C}t}(\|u_0\|^2 + \int_0^t e^{\tilde{C}s} g(s)\, ds) \tag{20}$$

where $g(s) = \dfrac{1}{\min\{1, 2(D_{m,0} - \sigma^2 - \gamma^2)\}}\Big(\|u_0\|^2 + \dfrac{1}{2\eta^2}\displaystyle\int_0^s \|q_2(\mu)\|^2\, d\mu,\Big)$, $\eta \neq 0$ *is an arbitrary constant,* $\sigma \neq 0, \gamma \neq 0$ *satisfy*

$$D_{m,0} - \sigma^2 - \gamma^2 > 0, \tag{21}$$

K_{er} and C_{q_1} are such that

$$\int_0^t k_{er}(t-s)^2 \|q_1(s)\|^2\, ds \leq C_{q_1},\ t \in (0,T], \tag{22}$$

holds, and

$$\tilde{C} = \frac{\max\Big\{D_{d,1}C_{q_1}\frac{(b-a)^2}{2A_0^2}\frac{1}{2\sigma^2}, 2\Big(\eta^2 + (C_B\frac{(b-a)^2}{2A_0^2})^2\frac{1}{4\gamma^2}\max_{t\in[0,T]}\|q_1(t)\|^2\Big)\Big\}}{\min\Big\{1, 2\Big(D_{m,0} - \sigma^2 - \gamma^2\Big)\Big\}}. \tag{23}$$

Proof: Considering in (17) $w = p(t)$ and using the assumption (11)) we deduce

$$A_0\|\nabla p(t)\|^2 \leq \frac{1}{4\epsilon^2}\|q_1(t)\|^2 + \epsilon^2\|p(t)\|^2, \tag{24}$$

 ISBN: 978-84-614-6167-7

where $\epsilon \neq 0$ is an arbitrary constant.
As the Friedrichs-Poincaré inequality $\|p(t)\|^2 \leq \frac{(b-a)^2}{2}\|\nabla p(t)\|^2$ holds, from (24) we obtain

$$\frac{A_0}{2}\|\nabla p(t)\|^2 + (\frac{A_0}{2} - \epsilon^2\frac{(b-a)^2}{2})\|\nabla p(t)\|^2 \leq \frac{1}{4\epsilon^2}\|q_1\|^2. \tag{25}$$

Then, for ϵ such that $\frac{A_0}{2} - \epsilon^2\frac{(b-a)^2}{2} = 0$ we conclude (19).

Taking in (18) $w = u(t)$ and using (12), (13) and (14) we deduce

$$\begin{aligned}\frac{1}{2}\frac{d}{dt}\|u(t)\|^2 - C_B\|\nabla p(t)\|\|u(t)\|\|\nabla u(t)\| + D_{m,0}\|\nabla u(t)\|^2 \\ \leq \int_0^t |K_{er}(t-s)|D_{d,1}\|\nabla p(s)\|\|\nabla u(s)\|ds\|\nabla u(t)\| + \frac{1}{4\eta^2}\|q_2(t)\|^2 + \eta^2\|u(t)\|^2,\end{aligned} \tag{26}$$

where $\eta \neq 0$ is an arbitrary constant.
It can be shown that

$$\begin{aligned}\int_0^t |K_{er}(t-s)|D_{d,1}\|\nabla p(s)\|\|\nabla u(s)\|ds\|\nabla u(t)\| \\ \leq D_{d,1}^2\frac{(b-a)^2}{2A_0^2}C_{q_1}\frac{1}{4\sigma^2}\int_0^t \|\nabla u(s)\|^2\,ds + \sigma^2\|\nabla u(t)\|^2,\end{aligned} \tag{27}$$

where $\sigma \neq 0$ is an arbitrary constant and C_{q_1} is fixed by (22).

Considering (19) and (27) in (26) we get

$$\begin{aligned}\frac{1}{2}\frac{d}{dt}\|u(t)\|^2 - C_B\frac{(b-a)^2}{2A_0^2}\|q_1(t)\|\|u(t)\|\|\nabla u(t)\| + \Big(D_{m,0} - \sigma^2\Big)\|\nabla u(t)\|^2 \\ \leq D_{d,1}C_{q_1}\frac{(b-a)^2}{2A_0^2}\frac{1}{4\sigma^2}\int_0^t \|\nabla u(s)\|^2 ds + \frac{1}{4\eta^2}\|q_2(t)\|^2 + \eta^2\|u(t)\|^2.\end{aligned} \tag{28}$$

Furthermore, as

$$-C_B\frac{(b-a)^2}{2A_0^2}\|q_1(t)\|\|u(t)\|\|\nabla u(t)\| \geq -(C_B\frac{(b-a)^2}{2A_0^2})^2\frac{1}{4\gamma^2}\|q_1(t)\|^2\|u(t)\|^2 - \gamma^2\|\nabla u(t)\|^2,$$

where $\gamma \neq 0$ is an arbitrary constant, from (28) we obtain

$$\begin{aligned}\frac{1}{2}\frac{d}{dt}\|u(t)\|^2 + \Big(D_{m,0} - \sigma^2 - \gamma^2\Big)\|\nabla u(t)\|^2 \leq \frac{1}{4\eta^2}\|q_2(t)\|^2 \\ + D_{d,1}C_{q_1}\frac{(b-a)^2}{2A_0^2}\frac{1}{4\sigma^2}\int_0^t \|\nabla u(s)\|^2 ds + \Big(\eta^2 + (C_B\frac{(b-a)^2}{2A_0^2})^2\frac{1}{4\gamma^2}\|q_1(t)\|^2\Big)\|u(t)\|^2,\end{aligned}$$

that leads to

$$\begin{aligned}\|u(t)\|^2 + \int_0^t \|\nabla u(s)\|^2\,ds \leq +\tilde{C}\int_0^t \Big(\int_0^s \|\nabla u(\mu)\|^2 d\mu + \|u(s)\|^2\Big)\,ds, \\ + \frac{1}{\min\{1, 2\Big(D_{m,0} - \sigma^2 - \gamma^2\Big)\}}\Big(\|u_0\|^2 + \frac{1}{2\eta^2}\int_0^t \|q_2(s)\|^2\,ds\Big)\end{aligned} \tag{29}$$

 ISBN: 978-84-614-6167-7

provided that σ and γ are fixed by (21) and $\tilde{C}$ is defined by (23). The proof of (20) is then concluded applying Gronwall Lemma to (29).

■

Theorem 1 can be easily generalized to analyze the stability of the two dimensional or three dimensional versions of the initial boundary value problem (15), (16), (17), (18).

Let us consider the particular non Fickian coupled diffusion model: equation (8),

$$\nabla v = q_1 \text{ in } (a,b) \times (0,T], \tag{30}$$

where v is given by Darcy's law

$$v = -\frac{K}{\mu(u)} \nabla p \text{ in } (a,b) \times (0,T]. \tag{31}$$

In (31) K denotes the permeability tensor and $\mu(u)$ represents the viscosity. System (9), (10) with $K_{er}(s) = \frac{1}{\tau} e^{-\frac{s}{\tau}}$, $A(x,y) = 1, B(x,y) = -\frac{K}{\mu(x)} y$, $D_m(x,y) = D_m(const)$, $D_d(x,y) = D_d|y|$, is a non Fickian version of (5), (30) and (31). The coefficient functions satisfy the conditions (11), (12), (13) and (14) provided that K is bounded and $\mu(x) \geq \mu_0$.

3 The semi-discrete approximation

In this section we introduce the semi-discrete approximation for the variational problem (9), (10) (15), (16). Let $\mathbb{I}_h = \{x_i, i = 0, \ldots, N, x_0 = a, x_N = b, x_i - x_{i-1} = h, i = 1, \ldots, N\}$ be a uniform partition of $[a,b]$. By $\mathbb{P}_h u_h$ we represent the piecewise linear interpolator of a grid function u_h defined in $\mathbb{I}_h$. By $\mathbb{W}_h$ we represent the space of all grid function defined on $\mathbb{I}_h$ and by $\mathbb{W}_{h,0}$ its subspace of all grid function null on the boundary points. The space of piecewise linear functions induced by the partition $\mathbb{I}_h$ is denoted by $S_h = \{\mathbb{P}_h u_h, u_h \in \mathbb{W}_h\}$.

By $L^2(0,T;S_h)$ we denote the subspaces of $L^2(0,T;H^1(a,b))$ that is obtained replacing $H^1(a,b)$ by S_h. We introduce now the piecewise linear approximations for the pressure p and for the concentration u, respectively, $\mathbb{P}_h p_h \in L^2(0,T;S_h)$ and $\mathbb{P}_h u_h \in \{v \in L^2(0,T;S_h) : \frac{dv}{dt} \in L^2(0,T;S_h)\}$ such that

$$p_h(x_0,t) = p_a(t), p_h(x_N,t) = p_b(t), u_h(x_0,t) = u_a(t), u_h(x_N,t) = u_b(t), t \in (0,T], \tag{32}$$

$$p_h(x_i,0) = p_0(x_i), u_h(x_i,0) = u_0(x_i), i = 1, \ldots, N-1, \tag{33}$$

and

$$\Big(A(\mathbb{P}_h u_h(t)) \nabla(\mathbb{P}_h p_h)(t), \nabla(\mathbb{P}_h w_h)\Big) = (q_1(t), \mathbb{P}_h w_h),\ w_h \in \mathbb{W}_{h,0}, \tag{34}$$

 ISBN: 978-84-614-6167-7

$$\begin{aligned}\Big(\frac{\partial}{\partial t}(\mathbb{P}_h u_h)(t),\mathbb{P}_h w_h\Big) &- \Big(B(\mathbb{P}_h u_h(t),\nabla(\mathbb{P}_h p_h)(t))\mathbb{P}_h u_h(t),\nabla(\mathbb{P}_h w_h)\Big)\\ &+\Big(D_m(\mathbb{P}_h u_h(t),\nabla(\mathbb{P}_h p_h)(t))\nabla(\mathbb{P}_h u_h)(t),\nabla(\mathbb{P}_h w_h)\Big)\\ =&-\int_0^t K_{er}(t-s)\Big(D_d(\mathbb{P}_h u_h(s),\nabla(\mathbb{P}_h p_h)(s))\nabla(\mathbb{P}_h u_h)(s),\nabla(\mathbb{P}_h w_h)\Big)ds\\ &+\Big(q_2(t),\mathbb{P}_h w_h\Big), w_h\in\mathbb{W}_{h,0}.\end{aligned} \tag{35}$$

In the space $\mathbb{W}_h$ we consider the norm $\|c_h\|_{1,h}^2=\|c_h\|_h^2+\|\nabla(\mathbb{P}_h c_h)\|^2$, where $\|.\|_h$ denotes the norm induced by the inner product

$$(w_h,v_h)_h=\sum_{i=1}^{N}\frac{h}{2}\Big(w_h(x_{i-1})v_h(x_{i-1})+w_h(x_i)v_h(x_i)\Big),\ w_h,v_h\in\mathbb{W}_h.$$

Let $L^2(0,T;\mathbb{W}_h)$ be a discrete version of $L^2(0,T;H^1(a,b))$ which is the space of grid functions $v_h:[0,T]\to\mathbb{W}_h$ such that $\displaystyle\int_0^T\|v_h(t)\|_1^2\,dt<\infty$.

We introduce now the fully discrete approximations (in space) $p_h\in L^2(0,T;\mathbb{W}_h)$ and $u_h\in\mathcal{W}_h(0,T)=\{v_h\in L^2(0,T;\mathbb{W}_h):\frac{dv_h}{dt}\in L^2(0,T;\mathbb{W}_h)\}$ as the grid functions that satisfy the conditions (32), (33) and the discrete variational equations

$$(A_h(t)\nabla(\mathbb{P}_h p_h)(t),\nabla(\mathbb{P}_h w_h))=(q_{1,h},w_h)_h,\ w_h\in\mathbb{W}_{h,0}, \tag{36}$$

$$\begin{aligned}&\Big(\frac{\partial u_h}{\partial t}(t),w_h\Big)_h-\Big(M_h(B_h(t)u_h(t)),D_{-x}w_h\Big)_{h,+}+\Big(D_{m,h}(t)\nabla(\mathbb{P}_h u_h(t)),\nabla(\mathbb{P}_h w_h)\Big)\\ &=-\int_0^t K_{er}(t-s)\Big(D_{d,h}(s)\nabla(\mathbb{P}_h u_h(s)),\nabla(\mathbb{P}_h w_h)\Big)ds+\Big(q_{2,h}(t),w_h\Big)_h, w_h\in\mathbb{W}_{h,0},\end{aligned} \tag{37}$$

where $M_h(v_h)(x_i)=\frac{1}{2}(v_{i-1}+v_i), i=1,\dots,N.$

In the previous equations the following notations were used: $D_{-x}w_h(x_i)=\dfrac{w_i-w_{i-1}}{h}$, $i=1,\dots,N$, $w_j=w_h(x_j)$, $(v_h,w_h)_{h,+}=\displaystyle\sum_{j=1}^{N}hv_jw_j$,

$$q_{\ell,h}(x_i,t)=\frac{1}{h}\int_{x_{i-1/2}}^{x_{i+1/2}}q_\ell(x,t)\,dx, i=1,\dots,N-1,\ell=1,2, \tag{38}$$

$A_h(x,t)$ and $D_{m,h}(x,t)$ are x piecewise constant functions defined by

$$A_h(x,t)=A(\frac{1}{2}(u_h(x_i,t)+u_h(x_{i+1},t))), \tag{39}$$

$$D_{m,h}(x,t)=D_m(\frac{1}{2}(u_h(x_i,t)+u_h(x_{i+1},t)),D_{-x}p_h(x_{i+1},t)), \tag{40}$$

for $x\in[x_i,x_{i+1})$, and the grid function $B_h(t)$ is given by

$$B_h(x_i,t)=\begin{cases}B(u_h(x_0,t),D_{-x}p_h(x_1,t)), i=0,\\ B(u_h(x_i,t),D_c p_h(x_i,t)), i=1,\dots,N-1,\\ B(u_h(x_N,t),D_{-x}p_h(x_N,t)), i=N,\end{cases} \tag{41}$$

 ISBN: 978-84-614-6167-7

and D_c will be defined bellow. The definition of the piecewise constant function $D_{d,h}$ is analogous to the definition of $D_{m,h}$.

In what follows we establish an ordinary differential system equivalent to the fully discrete (in space) variational problem (32), (33) (36), (37). In order to do that we introduce the following finite difference operators

$$D_c w_h(x_i) = \frac{w_{i+1} - w_{i-1}}{2h},\ D_x w_h(x_{i+1/2}) = \frac{w_{i+1} - w_i}{h}, D_x^{1/2} w_h(x_i) = \frac{w_{i+1/2} - w_{i-1/2}}{h},$$

where and $w_{j\pm 1/2}$ is used as far it makes sense.

It can be shown that the approximations $p_h(t)$ and $u_h(t)$ are solutions of the discrete problem

$$-D_x^{1/2}(A_h(t)D_x p_h(t)) = q_{1,h}(t) \text{ in } \mathbb{I}_h - \{a, b\}, \tag{42}$$

$$\begin{aligned} &\frac{du_h}{dt}(t) + D_c(B_h(t)u_h(t)) - D_x^{1/2}(D_{m,h}(t)D_x p_h(t)) \\ &\qquad = \int_0^t K_{er}(t-s)D_x^{1/2}(D_{d,h}(s)D_x p_h(s))ds + q_{2,h}(t) \text{ in } \mathbb{I}_h - \{a, b\} \end{aligned} \tag{43}$$

with the conditions (32), (33).

4 Stability of concentration and pressure

We establish now the stability of the coupled variational problem (36), (37) or equivalently the stability of the coupled finite difference problems (42), (43) under Dirichlet boundary conditions, that is $p_a(t) = p_b(t) = u_a(t) = u_b(t) = 0$. Let $C^1([0,T]; W_{h,0})$ be the space of grid functions $u_h : [0,T] \to W_{h,0}$ such that $\frac{du_h}{dt} : [0,T] \to W_{h,0}$ is continuous when in $W_{h,0}$ we consider the norm $\|.\|_h$.

Theorem 2 *If $u_h \in C^1([0,T]; W_{h,0})$ then, under the conditions of Theorem 1,*

$$\|p_h(t)\|_1 \leq \frac{b-a}{A_0}\|q_{1,h}(t)\|_h,\ t \in [0,T]. \tag{44}$$

and

$$\|u_h(t)\|_h^2 + \int_0^t \|\nabla(\mathbb{P}_h u_h)(s)|^2\, ds \leq e^{\tilde{C}t}\Big(\|u_h(0)\|_h^2 + \int_0^t e^{-\tilde{C}s} g_h(s)\ ds\Big), t \in [0,T], \tag{45}$$

provided that

$$\int_0^t k_{er}(t-s)^2 \|q_{1,h}\|_h^2\, ds \leq C_{q_1},\ t \in [0,T]. \tag{46}$$

In (45) $g_h(s)$ is given by

$$g_h(s) = \frac{1}{\min\{1, 2(D_{m,0} - \sigma^2 - \eta^2)\}}\Big(\|u_h(0)\|_h^2 + \frac{1}{2\eta^2}\int_0^s \|q_{2,h}(s)\|_h^2\, ds\Big),$$

 ISBN: 978-84-614-6167-7

$\eta \neq 0$ is an arbitrary constant, $\tilde{C}$ is defined by

$$\tilde{C} = \frac{\max\left\{C_{q_1} D_{d,1}^2 \frac{(b-a)^2}{A_0^2} \frac{1}{4\sigma^2}, \left(\eta^2 + \frac{C_p^2 C_B^2}{4\gamma^2} \max_{t\in[0,T]} \|q_{1,h}(t)\|_h^2\right\}\right.}{\min\{1, 2(D_{m,0} - \sigma^2 - \gamma^2)\}} \tag{47}$$

and $\sigma \neq 0, \gamma \neq 0$ are such that

$$D_{m,0} - \sigma^2 - \gamma^2 > 0. \tag{48}$$

Proof: As Friedrich's-Poincaré inequality $(b-a)^2 \|\nabla(\mathbb{P}_h w_h)\|^2 \geq \|w_h\|_h^2$ holds, the proof of (44) follows the proof of the correspondent continuous inequality (19).
Taking in (37) w_h replaced by $u_h(t)$ we easily deduce that

$$\begin{aligned}\frac{1}{2}\frac{d}{dt}\|u_h(t)\|_h^2 - (M_h(B_h(t)u_h(t)), D_{-x}u_h(t))_{h,+} + D_{m,0}\|\nabla(\mathbb{P}_h u_h)(t)\|^2 \\ \leq \int_0^t K_{er}(t-s)(D_{d,h}(s)\nabla(\mathbb{P}_h u_h)(s), \nabla(\mathbb{P}_h u_h)(t))ds + \frac{1}{4\eta^2}\|q_{2,h}(t)\|_h^2 + \eta^2\|u_h(t)\|_h^2,\end{aligned} \tag{49}$$

where $\eta \neq 0$ is an arbitrary constant.
As from (12) we have

$$|(M_h(B_h(t)u_h(t)), D_{-x}u_h(t))_{h,+}| \leq 2C_B\|\nabla(\mathbb{P}_h p_h)(t)\|\|u_h(t)\|_h\|\nabla(\mathbb{P}_h u_h)(t)\|,$$

considering (44) we obtain

$$|(M_h(B_h(t)u_h(t)), D_{-x}u_h(t))_{h,+}| \leq C_B^2 \frac{(b-a)^2}{A_0^2}\frac{1}{\gamma^2}\|q_{1,h}(t)\|_h^2\|u_h(t)\|_h^2 + \gamma^2\|\nabla(\mathbb{P}_h u_h)(t)\|^2, \tag{50}$$

where $\gamma \neq 0$ is an arbitrary constant.
As in the proof of Theorem 1, it can be shown that

$$\begin{aligned}\int_0^t K_{er}(t-s)|(D_{d,h}(s)\nabla(\mathbb{P}_h u_h(s)), \nabla(\mathbb{P}_h u_h)(t))ds| \\ \leq D_{d,1}^2 \frac{(b-a)^2}{A_0^2} C_{q_1} \frac{1}{4\sigma^2}\int_0^t \|\nabla(\mathbb{P}_h u_h)(t)\|^2 ds + \sigma^2\|\nabla(\mathbb{P}_h u_h)(s)\|^2,\end{aligned} \tag{51}$$

where $\sigma \neq 0$ is an arbitrary constant and C_{q_1} is fixed by (46).
Finally, using (50) and (51) in (49) we obtain

$$\begin{aligned}\frac{1}{2}\frac{d}{dt}\|u_h(t)\|^2 + \Big(D_{m,0} - \sigma^2 - \gamma^2\Big)\|\nabla(\mathbb{P}_h u_h)(t)\|^2 \leq \frac{1}{4\eta^2}\|q_{2,h}(t)\|_h^2 \\ + C_{q_1} D_{d,1}^2 \frac{(b-a)^2}{A_0^2}\frac{1}{4\sigma^2}\int_0^t \|\nabla(\mathbb{P}_h u_h)(s)\|^2 ds + \Big(\eta^2 + \frac{C_p^2 C_B^2}{4\gamma^2}\|q_{1,h}(t)\|_h^2\Big)\|u_h(t)\|_h^2.\end{aligned} \tag{52}$$

Inequality (52) implies that

$$\begin{aligned}\|u_h(t)\|^2 + \int_0^t \|\nabla(\mathbb{P}_h u_h)(t)\|^2 \leq \tilde{C}\int_0^t \Big(\int_0^s \|\nabla(\mathbb{P}_h u_h)(\mu)\|^2 d\mu + \|u_h(s)\|_h^2\Big) ds \\ + \frac{1}{\min\{1, 2(D_{m,0} - \sigma^2 - \gamma^2)\}}\Big(\|u_h(0)\|_h^2 + \frac{1}{2\eta^2}\|q_{2,h}(t)\|_h^2\Big),\end{aligned} \tag{53}$$

 ISBN: 978-84-614-6167-7

where σ and γ are fixed by (48) and $\tilde{C}$ is given by (47). Finally, the inequality (53) leads to (45).

■

Stability results similar to Theorems 1 and 2 hold when q_2 depends on u. In fact we need only to assume that $|q_2(y)| \leq C_{q_2}|y|$.

5 Numerical results

In this section, as an example, we apply the proposed non Fickian model to a single phase incompressible flow problem in a porous media with a point source and sink term. This fluid flow process involves fully miscible displacement of one incompressible fluid by another. In this case p is the pressure of the fluid mixture, u the volumetric concentration of the injected fluid, ϕ the porosity of the medium, K the permeability of the medium, D_m the molecular diffusion coefficient and $D_d(u, \nabla p)$ the mechanical dispersion $D_d(u, \nabla p) = D_d|v|$, where D_d denotes the dispersion coefficient and v represents Darcy's velocity of the fluid mixture which is given by (31). In (31) the viscosity of the mixture $\mu(u)$ is determined by the commonly used rule $\mu(u) = \mu_0((1-u) + M^{\frac{1}{4}}u)^{-4}$, where M is the mobility ratio and μ_0 the viscosity of resident fluid. In (5) the function q_2 is given by $q_2 = u^* q_1$, where u^* is the injected concentration at sources or the concentration u at sinks. The function q_1 is the source and sink terms.

To closed the system (8), (30), (31) we assume natural boundary conditions $v = 0$ on $\{a, b\} \times (0, T]$, $D_m \nabla u + \dfrac{1}{\tau}\displaystyle\int_0^t e^{-\frac{t-s}{\tau}}(D_d \nabla u)\, ds = 0$ on $\{a, b\} \times (0, T]$. In order to compare the Fickian and non Fickian models, we integrate in time the ordinary differential system (36), (37) using the implicit Euler method with a very small step size and discretizing the integral term using the right rectangular rule.

Let $0 = t^0 < t^1 < \ldots < t^N = T$ be a partition of the time interval $[0, T]$ with $\Delta t = t^{n+1} - t^n$ and N the number of time steps. Denote by p_h^n, v_h^n and u_h^n the approximations of p, v and u, respectively, at time level t^n. To compute the numerical approximations at time level t^{n+1} we use the following algorithm:

Step 1: Given u_h^n, solve the finite difference equation

$$-D_x^{1/2}\Big(\frac{K}{\mu(u_h^n)} D_x p_h^{n+1}\Big) = q_1^{n+1}$$

to compute p_h^{n+1};

Step 2: With p_h^{n+1}, compute the velocity of the convective term using the discretization in (41) and v_h^{n+1} in $D_d(v_h^{n+1})$ by

$$v_h^{n+1} = -\tfrac{K}{\mu(u_h^n)} D_{-x} p_h^{n+1};$$

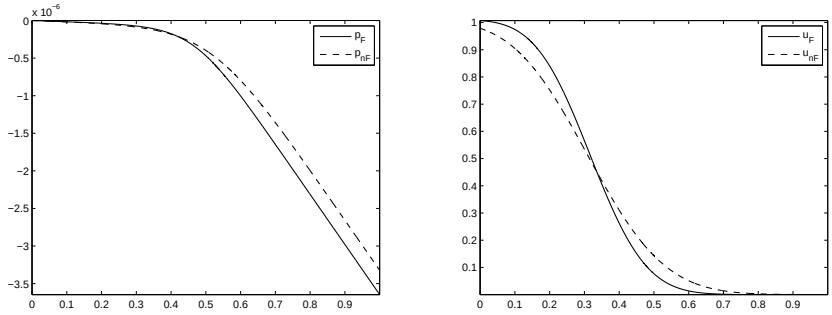
Step 3: Compute u_h^{n+1} using

$$\phi \frac{u_h^{n+1} - u_h^n}{\Delta t} + D_c(u_h^n v_h^{n+1}) - D_x^{1/2}(D_m D_x u_h^n) = \frac{\Delta t}{\tau} \sum_{j=1}^{t_{n+1}} e^{-\frac{t_{n+1}-t_j}{\tau}} D_x^{1/2}(D_d(v_h^j) D_x u_h^j) + q_2^n.$$

 ISBN: 978-84-614-6167-7

The simulation was performed considering $[a, b] = [0, 1]$ and the following parameters: $T = 800$, $\phi = 1$, $K = 60$, $M = 41$, $\mu_0 = 1$, $u^* = 5$, $D_m = \phi 10^{-5}$, $D_d = \phi 10^{-3}$, $\tau = 0.001$, $h = 10^{-4}$, $\Delta t = 0.025$ The injection well cover one cell at the left extreme of the interval $[a, b]$ and has a constant injection rate equal to 5. The production well also cover one cell which is located at the right extreme with the production rate equal to -5.

In Figure 1 we plot the numerical approximation for Fickian and non Fickian pressures and concentrations at $t = 800$. From the numerical experiments we observe for the Fickian and non Fickian pressure, p_F and p_{nF} respectively, a similar behavior. However for Fickian and non Fickian concentrations u_f, u_{nf}, respectively we observe that $u_F > u_{nF}$ near the injection point and $u_F < u_{nF}$ near the sink point.



Acknowledgements

The authors gratefully acknowledge the support of this work by Centro de Matemática da Universidade de Coimbra and by the project UTAustin/MAT/0066/2008.

References

[1] R. E. Ewing, *Problems arising in the modeling of hydrocarbon recorvery*, in *The Mathematics of Reservoir Simulation*, SIAM, Philadelphia, 1983, 3–34.

[2] T. F. Russel and M. F; Wheeler, *Finite element and finite difference methods for continuous flows in porous media*, in *The Mathematics of Reservoir Simulation*, SIAM, Philadelphia, 1983, 35–106.

[3] H-T. Chen and K-C. Liu *Analysis of non-Fickian diffusion problems in a composite medium*, Comp, Phys. Comm. **150** (2003) 31–42.

[4] J. Douglas Jr *Superconvergence in the pressure in the simulation of miscible displacement*, SIAM J. Numer. Anal. **22** (1985) 962–969.

[5] R. E. Ewing and M. F. Wheeler, *Galerkin methods for miscible displacement problems in porous media*, SIAM J. Numer. Anal. **17** (1980) 351–365.

 ISBN: 978-84-614-6167-7

[6] S. M. Hassahizadeh, *On the transient non-Fickian dispersion theory*, Transp. in Porous Media **23** (1996) 107–124.

[7] C. Maas, *A hyperbolic dispersion equation to model the bounds of a contaminated groundwater body*, J. of Hydrol. **226** (1999) 234–241.

[8] S. P. Neuman, and D. M. Tartakovski, *Perspectives on theories of non-Fickian transport in heterogeneous medias*, Adv. in Water Resources **32** (2008) 670–680.

[9] D. W. Peaceman, *Improved tratment of dispersion in numerical calculation of multidimensional miscible displacement*, Soc. of Petro. Eng. **6** (1966) 213–216.

[10] A. Settari, H. S. Price and T. Dupont, *Development and application of variational methods for simulation of miscible displacement in porous media*, Soc. of Petrol. Eng. **17** (1977) 228–246.

 ISBN: 978-84-614-6167-7

Proceedings of the 11th International Conference on Computational and Mathematical Methods in Science and Engineering, CMMSE 2011 26–30 June 2011.

Development of an unified FDTD-FEM library for electromagnetic analysis with CPU and GPU computing

Jorge Francés[1], Sergio Bleda[1,2], Sergi Gallego[1,2], Cristian Neipp[1,2], Andrés Marquez[1,2], Inmaculada Pascual[2,3]and Augusto Beléndez[1,2]

[1] *Department of Physics, Systems Engineering and Signal Theory, Universidad de Alicante, E-03080, Alicante (Spain)*

[2] *University Institute of Physics to Sciences and Technologies, University of Alicante, E-03080, Alicante (Spain)*

[3] *Department of Optics, Pharmacology and Anatomy, University of Alicante, E-03080, Alicante (Spain)*

emails: jfmonllor@ua.es, sergio.bleda@ua.es, sergi.gallego@ua.es, cristian@dfists.ua.es, andres.marquez@ua.es, pascual@ua.es, a.belendez@ua.es

Abstract

We describe a C++ library for electromagnetics based on the Finite-Difference Time-Domain method for transient analysis, and the Finite Element Method for modal analysis. Both methods share the same core and also both methods are optimized for CPU and GPU computing. The FEM method is applied for solving Laplace's equation and analyzes the relation between surface curvature and electrostatic potential of a long cylindric conductor. The FDTD method is applied for analyzing Thin Film Filters in optical wavelengths. Furthermore, the performance of both CPU and GPU versions are analyzed as a function of the grid size simulation. This approach allows to analyze a wide range of electromagnetic situations taking advantage of the benefits of each numerical method and also of the modern graphics processing units.

Key words: Electromagnetic analysis, Finite-Difference Time-Domain, Finite Element Method, electrostatic potential, Thin Film Filters, Optical wavelengths, Graphics Processing Units

MSC 2000: 78A97, 81V80, 80M50

1 Introduction

Graphics Processing Units (GPU) are of considerable importance in such areas as electromagnetics, optics, and acoustics. The architecture of the new Fermi family [1] has

 ISBN: 978-84-614-6167-7

been fully designed for numerical computing and many researchers have been used recently the GPUs in many different fields such as astronomy, soft tissue simulation, image computing, and much more. The adaption of many computational methods is still in progress and unfortunately, not all algorithms can be ported efficiently onto a GPU architecture. Using a recent consumer graphics card, we accelerated the Finite Element Method (FEM) and the Finite-Difference Time-Domain (FDTD) method for electromagnetics. A highly optimized sequential implementation for Central Processor Unit (CPU) was also developed in order to compare and contrast the performance between the GPU version and the CPU code. The optimization in the sequential code is performed by programming strategies that benefits the auto-vectorization provided by modern compilers, which are mainly focused on parallelize the main loops in the algorithms [2]. Both implementations, CPU and GPU, requiere a previous analysis of the operations and a rearrangement of the instructions in order to take advantage of the CPU and GPU architectures. Due to the fact that the programming paradigm is changed in great manner for GPU computing, we are focused on analyzing the degree of improvement as a function of the size simulation between single CPU and GPU approaches.

On the other hand, the library implemented provides the possibility of using both FEM and FDTD method for solving a wide range of electromagnetic experiences. Specifically, the FEM method is applied to analyze the relation between surface curvature and electrostatic potential [3], whereas the FDTD method is used for analyzing optical diffractive elements such as Thin Film Filters (TFF). The FEM method has been applied in many areas and nowadays is a reference in numerical methods. In this work, it has been used for solving the Laplace's equation in electrostatics. The aim of the numerical analysis performed by the FEM is to compare the analytical expressions developed in [3] for analyzing the relation between the the surface curvature of an isolated charged conductor with uniform cross section and the resultant electrostatic potential. An illustration showing this scheme is shown in Fig. 1a. Due to the fact that FEM method is useful for static analysis, a well-known alternative for transient analysis is the FDTD method [4, 5]. Specifically, it was applied at optical wavelengths for analyzing the reflectance of High-Reflecting Coatings (HRCs') [6]. HRC is a basic type of TFF and is composed by a stack of alternate high- and low-index films, all one quarter wavelength thick as it has been illustrated in Fig. 1b. Light reflected within the high-index layers not suffers any phase shift while that a change of 180 in the low-index layers is produced. It is straightforward to see that the light produced by reflection at successive boundaries throughout the assembly reappear at the front surface all in phase so that they recombine constructively. Because of this behavior, HRC's have many applications such as photovoltaic cells and Micro-Electro-Mechanical Systems (MEMS).

 ISBN: 978-84-614-6167-7

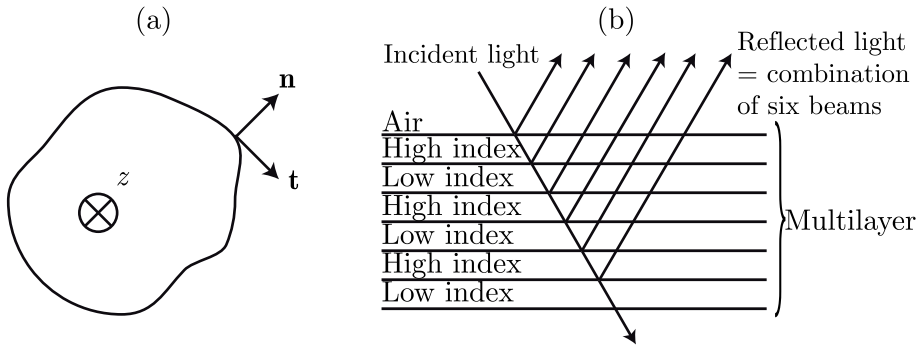


Figure 1: (a) Long conductor with uniform cross section. (b) Scheme of a high-reflectance coating [6].

2 Theory

This section gives a brief summary of the basis of the numerical methods implemented. First, the theoretical basis of FEM method for solving Poisson's equation is shown. Second, the theory related with the FDTD method and its add-ons needed for our specific applications are introduced.

2.1 FEM analysis of Laplace's equation

In this section, we will apply the FEM to electroestatic problems. More specifically to a infinite cylinder perturbed by a small cosine function,

$$r(\theta) = a\left(1 + \varepsilon\cos(n\theta)\right), \tag{1}$$

where a is the radius of the cylinder, n is an integer parameter and ϵ is an small number [3]. The analytical expression developed by Neipp *et al* in [3] relates the curvature of an infinite conductor and the electrostatic potential.

$$\phi(r,\theta) = \phi_0 + b_0 \ln\frac{r}{a} - \varepsilon b_0 \left(\frac{a}{r}\right)^n \cos(n\theta), \tag{2}$$

where r and θ are the cylindrical coordinates, ϕ_0 is the potential at the surface of the conductor and b_0 is a constant coefficient defined in [3]. Therefore, the Laplace's equation $\nabla^2\phi = 0$ is considered. In the FEM, the two-dimensional region in which the potential distribution solution $\phi(x,y)$ is defined, is divided into a number of finite elements as illustrated in Fig 2. The subdivision of the solution region into elements is done by an automatic scheme able to provide uniform and nonuniform meshes by a few parameters related with the geometry [7]. The approximate solution of the whole region is

$$\phi(x,y) \approx \sum_{e=1}^{N} \phi_e(x,y), \tag{3}$$

 ISBN: 978-84-614-6167-7

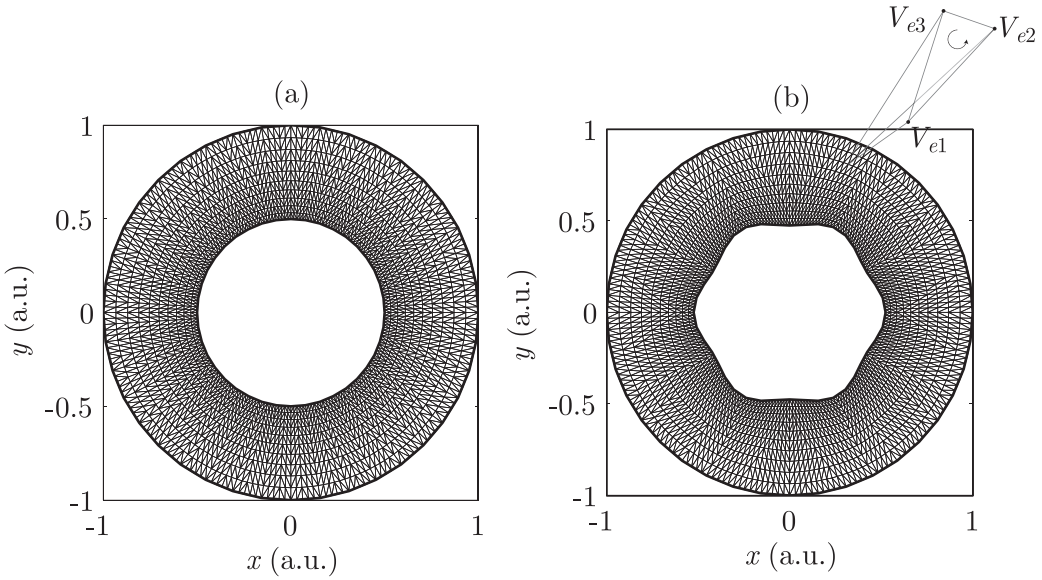


Figure 2: Finite element discretization for the cross section of cylindrical conductor perturbed by a cosine term. (a)$\varepsilon = 0$ and $a = 1$. (b)$\varepsilon = 0.05$, $n = 6$, $a = 1$.

where N is the number of elements into which the solution region is divided. Here the approximation of ϕ_e within an element is the polynomial approximation for the triangular element shown in Fig. 2b,

$$\phi_e(x, y) = a + bx + cy. \tag{4}$$

The constants a, b and c are to be determined. The potential V_{e1}, V_{e2} and V_{e3} at nodes 1, 2 and 3 are obtained from Eq. (4) as:

$$\begin{bmatrix} \phi_{e1} \\ \phi_{e2} \\ \phi_{e3} \end{bmatrix} = \begin{bmatrix} 1 & x_1 & y_1 \\ 1 & x_2 & y_2 \\ 1 & x_3 & y_3 \end{bmatrix} = \begin{bmatrix} a \\ b \\ c \end{bmatrix} \Rightarrow \begin{bmatrix} a \\ b \\ c \end{bmatrix} = \begin{bmatrix} 1 & x_1 & y_1 \\ 1 & x_2 & y_2 \\ 1 & x_3 & y_3 \end{bmatrix}^{-1} \begin{bmatrix} \phi_{e1} \\ \phi_{e2} \\ \phi_{e3} \end{bmatrix}. \tag{5}$$

Substituting this into Eq. (4) gives

$$\phi_e = \sum_{i=1}^{3} \alpha_i(x, y)\phi_{ei}, \tag{6}$$

where,

$$\alpha_1 = \frac{1}{2A}\left[(x_2y_3 - x_3y_2) + (y_2 - y_3)x + (x_3 - x_2)y\right], \tag{7}$$

$$\alpha_2 = \frac{1}{2A}\left[(x_3y_1 - x_1y_3) + (y_3 - y_1)x + (x_1 - x_3)y\right], \tag{8}$$

$$\alpha_3 = \frac{1}{2A}\left[(x_1y_2 - x_2y_1) + (y_1 - y_2)x + (x_2 - x_1)y\right], \tag{9}$$

 ISBN: 978-84-614-6167-7

with A being the area of the element e [7], and α_i are linear interpolation functions and are called the *element shape functions*.
The functional corresponding to Laplace's equation is given by:

$$W_e = \frac{1}{2}\int \epsilon |\nabla \phi_e|^2 dS, \tag{10}$$

Physically, the functional W_e is the energy per unit length associated with the element e. Notice that the region is homogenous, thus ϵ is a constant related with the electric permittivity. From Eq. (6):

$$\nabla \phi_e = \sum_{i=1}^{3} \phi_{ei} \nabla \alpha_i. \tag{11}$$

Substituting Eq. (11) into (10) gives

$$W_e = \frac{1}{2}\sum_{i=1}^{3}\sum_{j=1}^{3} \epsilon \phi_{ei} \left[\int \nabla \alpha_i \cdot \nabla \alpha_j dS\right] \phi_{ej}. \tag{12}$$

If the term in brackets is defined as $C_{ij}^{(e)}$, Eq. (12) can be rewritten in matrix form as

$$We = \frac{1}{2}\epsilon \boldsymbol{\phi_e}^t \mathbf{C^{(e)}} \boldsymbol{\phi_e}, \tag{13}$$

where the t denotes the transpose of the matrix. All terms of $\mathbf{C^{(e)}}$ are fully detailed in [7].
The next step is to assemble all such elements in the solution region.

$$W = \sum_{e=1}^{N} W_e = \frac{1}{2}\epsilon \boldsymbol{\phi}^t \mathbf{C} \boldsymbol{\phi}, \tag{14}$$

where $\boldsymbol{\phi}$ is a vector with n values related with the n nodes of the system, and N is the number of elements. The matrix $\mathbf{C}$ is the *global coefficient matrix*, which is the assemblage of individual element coefficient matrices and is fully defined in [7].
The Eq. (14) can be easily solved if the free nodes are numbered first and the fixed nodes last

$$W = \frac{1}{2}\epsilon \begin{bmatrix} \boldsymbol{\phi}_f & \boldsymbol{\phi}_p \end{bmatrix} \begin{bmatrix} \mathbf{C}_{ff} & \mathbf{C}_{fp} \\ \mathbf{C}_{pf} & \mathbf{C}_{pp} \end{bmatrix} \begin{bmatrix} \boldsymbol{\phi}_f \\ \boldsymbol{\phi}_p \end{bmatrix}, \tag{15}$$

where subscripts f and p, respectively, refer to nodes with free and fixed potentials. For solving this system of equations the partial derivatives of W with respect to each nodal value of the potential be zero ($\partial W/\partial \phi_1 = \partial W/\partial \phi_2 = \cdots = \partial W/\partial \phi_n = 0$. Since $\boldsymbol{\phi_p}$ is constant, we only differentiate with respect $\boldsymbol{\phi}_f$.

$$\begin{bmatrix} \mathbf{C}_{ff} & \mathbf{C}_{fp} \end{bmatrix} \begin{bmatrix} \boldsymbol{\phi}_f \\ \boldsymbol{\phi}_p \end{bmatrix} = 0 \Rightarrow \mathbf{C}_{ff}\boldsymbol{\phi}_f = -\mathbf{C}_{fp}\boldsymbol{\phi}_p. \tag{16}$$

This equation can be rewritten as $\mathbf{A} \cdot \mathbf{x} = \mathbf{y}$ where $\mathbf{x}$ is the unknown. Here, this equation system is solved applying the Conjugate Gradient Method (CGM) [8]. The CGM is an iterative approach that basically uses matrix vector multiplications and inner products ($\mathbf{v}^T\mathbf{v}$). Therefore, these operators are critical because they are the main core of the method and the objective of our optimization.

 ISBN: 978-84-614-6167-7

2.2 FDTD analysis at optical wavelengths

Light propagation is described by means of Maxwell's time-dependent curl equations:

$$\frac{\partial \widetilde{\mathbf{D}}}{\partial t} = \frac{1}{\sqrt{\epsilon_0 \mu_0}} \left(\nabla \times \mathbf{H} - \sigma \widetilde{\mathbf{E}} \right), \tag{17}$$

$$\widetilde{\mathbf{D}}(\omega) = \epsilon_r^*(\omega) \cdot \widetilde{\mathbf{E}}, \tag{18}$$

$$\frac{\partial \mathbf{H}}{\partial t} = -\frac{1}{\sqrt{\epsilon_0 \mu_0}} \nabla \times \widetilde{\mathbf{E}} - \frac{\sigma_m}{\mu_0} \mathbf{H}, \tag{19}$$

where ϵ_0 is the electrical permittivity in vacuum in farads per meter, ϵ_r is the medium's relative complex permittivity constant, that has been assumed real, μ_0 is the magnetic permeability in vacuum in henrys per meter. The flux density is denoted by $\widetilde{\mathbf{D}}$ and both, $\widetilde{\mathbf{D}}$ and $\widetilde{\mathbf{E}}$ are normalized respect to the vacuum impedance $\eta_0 = \sqrt{\mu_0/\epsilon_0}$. The FDTD algorithm used here is based on the Yee lattice [10]. The electrical field components $\mathbf{E}$ and the magnetic field component $\mathbf{H}$ are defined in a bidimensional cell [4, 5, 10]. As a result, the Maxwell's curl equations can be discretized and solved by using the central-difference expressions, for both the time and space derivatives. So, considering TM^z polarization and bidimensional analysis, the Eq. (17) can be reformulated as follows:

$$\widetilde{D}_z|_{i,j}^{n+1/2} = \widetilde{D}_z|_{i,j}^{n-1/2} + \frac{\Delta t}{\sqrt{\epsilon_0 \mu_0}} \left[\frac{H_y|_{i+1/2,j}^n - H_y|_{i-1/2,j}^n}{\Delta x} - \frac{H_x|_{i,j+1/2}^n - H_x|_{i,j+1/2}^n}{\Delta y} \right] \tag{20}$$

where Δx and Δy are the spatial and time resolution respectively. In order to simulate unbounded free space, it must be included a formalism in order to avoid the interferences produced by outgoing waves reaching the grid simulation limits. For this reason a simplified version of the Perfectly Matched Layers (PML) developed by Berenger has been implemented in this work [11].

The PML's are a good technique for the absorption of electromagnetic waves by means of a nonphysical absorbing medium adjacent to the outer FDTD mesh boundary. The basic idea of this formalism is to create a medium that is lossy and minimize the amount of reflection between vacuum and the PML region. Related with the illumination method, note that the incidence is assumed to be from air to medium. In connection with the propagation in the FDTD region, it must be said that the source is introduced along the connecting boundary by using a Total Field/Scattered Field (TF/SF) algorithm [4], where the linearity of Maxwells's equations and their decomposition of the electromagnetic field are assumed: $(\mathbf{E};\mathbf{H})_{\text{Total}} = (\mathbf{E};\mathbf{H})_{\text{inc}} + (\mathbf{E};\mathbf{H})_{\text{scat}}$. Where $(\mathbf{E};\mathbf{H})_{\text{inc}}$ are the values of the incident field, which are assumed to be known at all space points in the FDTD grid and also at all time steps. $(\mathbf{E};\mathbf{H})_{\text{scat}}$ are the values of the scattered wave fields, which are unknown and produced by the optical device in our particular case. This method avoids the computation of the incident wave in the whole bidimensional grid and only two one-dimensional arrays are needed.

 ISBN: 978-84-614-6167-7

3 Computational optimization

In this section the approach followed for implementing both methods in an unified C++ library are shown. The rearrangement of the instructions and memory alignment strategies are explained for both numerical methods, the FEM and the FDTD methods. The implementation and the application for GPU computing is also detailed. Whole library is implemented in C++ object oriented language. This type of language provides several characteristics such as class definition, overloading or inheritance, that benefit the development of a complex and big project. Nevertheless, here only classes and overloading were considered, since advanced inheritance directives are not already allowed in Compute Unified Device Architecture (CUDA) [12] and in some cases can reduce the performance of the application. Fig. 3 shows the class diagram of the library developed. As can be seen the class Array is common to both numerical methods. This class implements an one-dimensional array that can store a matrix of dimensions $n \times m \times p$ in columns priority. This approach is more convenient because it makes easier physical memory alignment, since all data is stored in a single column vector. This class also implements several methods related with arithmetic operations such as matrix vector multiplication, inner products, dot product, etc. Therefore, these methods can be used in the FEM and the FDTD.

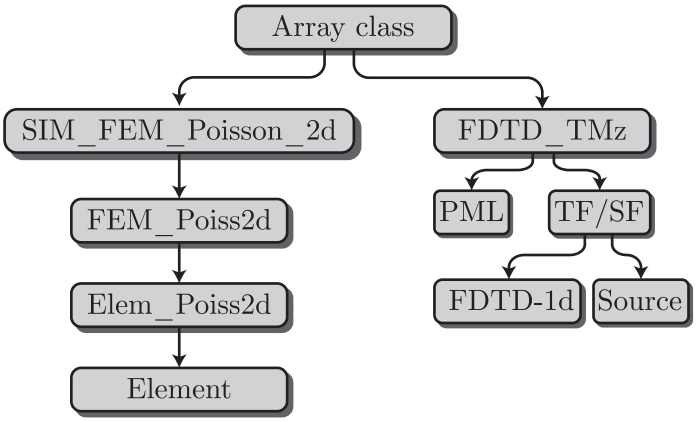


Figure 3: Class diagram of the C++ library implemented. As can be seen the class Array is common to FEM and FDTD simulations

3.1 Instruction rearrangement for auto-vectorization in CPU

In this work the software runs under a Unix based platform with an Intel Core i7-950 Processor with 8MB of cache, a clock speed of 3.06 GHz and 6 GB of global DDRAM3. The auto-vectorization provided by modern compilers (with the flags O1, O2 and O3) are based on predict which loops can be automatically vectorized, or converted into Streaming SIMD (*Singe Instruction Multiple Data*) Extensions (SSE) [2].

 ISBN: 978-84-614-6167-7

This vectorization is sensitive to the layout of the loop and the data structures used, and the dependencies among the data accesses in each iteration and across iterations. Once the compiler has made such a determination, it can generate vectorized code for the loop, allowing the application to use the SIMD instructions. With this approach, applications can theoretically achieve a full 4x performance gain. Although, 2x is the realized gain on real applications in part because of latency during I/O instructions and general issues related with the microarchitecture [9].
Therefore, the following strategies have been followed in order to take advantage of this capability: memory alignment, use of arrays to make data contiguous and padding for avoiding misalignment.
Due to the fact that matrix vector multiplication $\mathbf{A} \cdot \mathbf{b} = \mathbf{c}$ is one of the most common operator in FEM, the optimization of this operation would improve the performance of the method. For that purpose the matrix A can be redefined as $\mathbf{A} = \left[\mathbf{a}^1 \dots \mathbf{a}^n\right]$, where $\mathbf{a}^i$ is the i-th column of the matrix $\mathbf{A}$. Therefore, the matrix vector multiplication instructions can be rearranged as $\mathbf{c} = b_1\mathbf{a}^1 + \cdots + b_n\mathbf{a}^n$. This rearrangement improves the access to contiguous data, reducing the cache misses.
Due to the fact that the maximum number of nonzero elements per rows is 7, the global coefficient matrix $\mathbf{C}$ is sparse. It means that the number of zeros is greater than the number of nonzeros. For instance, for a matrix with $n = 450$ the matrix is 98.97 % sparse. Clearly, it makes little sense to store these zero entries. Therefore, the Compressed Sparse Column (CSC) format was used for storing the values of the nonzero matrix $\mathbf{C}$ by columns. This scheme stores all nonzero values in a single column vector and uses a couple of pointers for indexing these values in $\mathbf{C}$. Here, the number of nonzero elements per column was padded for reducing misalignment in memory accesses. This padding ensures that the number of nonzero elements per column is multiple of 4.
Regarding FDTD optimization under CPU, it must be said that the same techniques related with FEM method can be applied. Although, in this case to solve the FDTD equations the *leapfrog* algorithm is used [5]. This method is an iterative approach for solving the electromagnetic fields along each field component, thus a double loop is needed for compute each field as a function of space. This procedure is repeated as time simulations are defined. Therefore, each field component can be redefined as follows:

$$\widetilde{\mathbf{D}}_z^{n+1/2} = \widetilde{D_z}|_{i,j}^{n+1/2} = \left[\begin{array}{cccc} \widetilde{\mathbf{d}}_z^0 & \widetilde{\mathbf{d}}_z^1 & \cdots & \widetilde{\mathbf{d}}_z^{m-1} \end{array} \right]^{n+1/2}, \tag{21}$$

where $\mathbf{d}_z^j = [d_z[0] \quad d_z[1] \quad \cdots \quad d_z[n-1]]^T$ with T denoting the hermitian transpose matrix. If Eq. (20) is evaluated firstly by columns the performance of the method is improved due to the fact that the cache misses are minimized and also the spatial proximity of the data is ensured.

 ISBN: 978-84-614-6167-7

3.2 GPU implementation

The Fermi architecture, released in the spring of 2010, is NVIDIAs next-generation GPU. It is the successor of the GT200 architecture, and is the first in which NVIDIA focussed on general-purpose computation performance. The memory hierarchy is one of the most distinguish features of the NVIDIA GPUs. The addition of a cache hierarchy, consisting of a global L2 cache, as well as a per-Streaming Multiprocessors (SM) L1 cache gives more flexibility in non-uniform memory accesses. The Fermi GTX470 features 448 Scalar Processors (SPs) organized in 14 SMs. Each SM has 32 Single Precision (SPs), 16 Double Precision (DPs), and 4 Specieal Functions Units (SFUs). Each SM also has a block of local memory called shared memory visible to all threads within a thread block, and a scheduling unit used to schedule warps. The basic computing unit (called warp) consists of 32 threads. The GPU is capable of swapping warps into and out of context without any performance overhead. This functionality provides an important method of hiding memory and instruction latency on the GPU hardware.
To ease the mapping of data to threads, the threads identifiers may be multidimensional and, since a very high number of threads run in parallel, CUDA groups threads in blocks and grids. One of the crucial requirements to achieve a good performance on the NVIDIA GPU is to hide the high latency of the global memory ensuring coalesced memory accesses.
Therefore, to be concerned about the architecture of the GPU is mandatory to be successful in GPU computing. For the FEM method the operator matrix vector multiplication has been implemented by means the CUSPARSE Library [13]. The inner product needed for the product of two vectors is completed by means of the reduction technique developed in [12].
Regarding FDTD implementation in the GPU it must be said that, a number of blocks related with the number of columns is invoked by means of the kernels functions and an array of 128×2 threads are launched per block. Each column of threads works along one column of the simulation grid as many times as necessary to evaluate Eq. (21) and those related with the magnetic field. Besides the potential of the CUDA kernel, it is necessary to divide the whole computation process in several kernels focused on compute each component of the electromagnetic field. This segmentation improves the efficient use of the shared memory in the device and also the correct usage of the cache. This effect is maximized in the new Fermi architecture, where each SM has 64 KB of on-chip memory that can be configured as 48 KB of Shared memory with 16 KB of L1 cache or as 16 KB of Shared memory with 48 KB of L1 cache. Nonetheless, applications that do not use Shared memory automatically benefit from the L1 cache, allowing high performance CUDA programs to be built with minimum time and effort.

4 Results

Our first group of results shown in Fig. 4a-b are the comparison of the electrostatic potential ϕ obtained by Eq. (2) and the FEM method. In Fig. 4b the charge density σ

 ISBN: 978-84-614-6167-7

is shown and it has been obtained taking into account that $\sigma \approx |\nabla\phi|$. As can be seen a good agreement between numerical an theoretical values is achieved, thus validating our approach.

Fig. 4c-d illustrates the results for the analysis of the reflectance ($\propto |E|^2$) for normal incidence of alternating $\lambda_0/4$ layers of high-($n_H = 2.3$) and low-index (n_L=1.38) dielectric materials on a transparent substrate (n_s=1.52), which scheme is illustrated in Fig. 1, as a function of the wavelength ratio $\lambda_0\lambda$. Although, the FDTD is defined for transient analysis, a singleton simulation in time domain can provide information that easily can be transformed into the spectral domain. The parameters of the simulation are: $\lambda_0 = 633$ nm, $\Delta x = 0.32$ m, $\Delta t = 4.74 \cdot 10^{-6}$ns, and they were carefully chosen for ensuring stability of the method and convergency of the solution. The Effective Medium Theory EMT is related with the characteristic matrix of a layer detailed in [6]. Also in this case a good agreement between the FDTD method and the matrix approach is achieved.

The computational performance is given by means of the SpeedUp that is defined as the ratio between the simulation times of the sequential and parallel codes. Therefore, Fig 4e-f shows the improvments ratio of using the single CPU auto-vectorized by the compiler and the GPU implementation. As can be seen in Fig. 4e, the benefit of using GPU computing grows as the number of elements is increased, but there is a region in which a single CPU with an auto-vectorized code is a better option. Regarding the FDTD method, the SpeedUP of the GPU version is always greater than the single CPU in all cases, because of FDTD algorithm is more suitable to take advantage of the GPU programming model.

5 Conclusions

We have implemented an unified library for electromagnetic analysis based on the FEM and FDTD method. The FEM method was used for compare the analytical expressions obtained for the analysis of the surface curvature of an infinite cylinder in electrostatics, whereas the FDTD method has been applied in optical wavelengths for analyzing the reflectance of high-reflecting coatings. In both cases, the analytical and numerical results are quite similar, thus validating our implementation.

Moreover, both methods have been developed following a set of rules that benefits the auto-vectorization of modern compilers in order to take advantage of the SSE registers in the CPU. This optimization has revealed that an improvement near of four is achieved with this auto-vectorization in both cases. In addition, FEM and FDTD has been also implemented in a GPU. The benefits of GPU computing in both methods are quite different, since for FEM analysis the SpeeduUp increases with the number of elements, whereas for FDTD computation behaves more constant and in all cases is higher than the CPU auto-vectorized version.

 ISBN: 978-84-614-6167-7

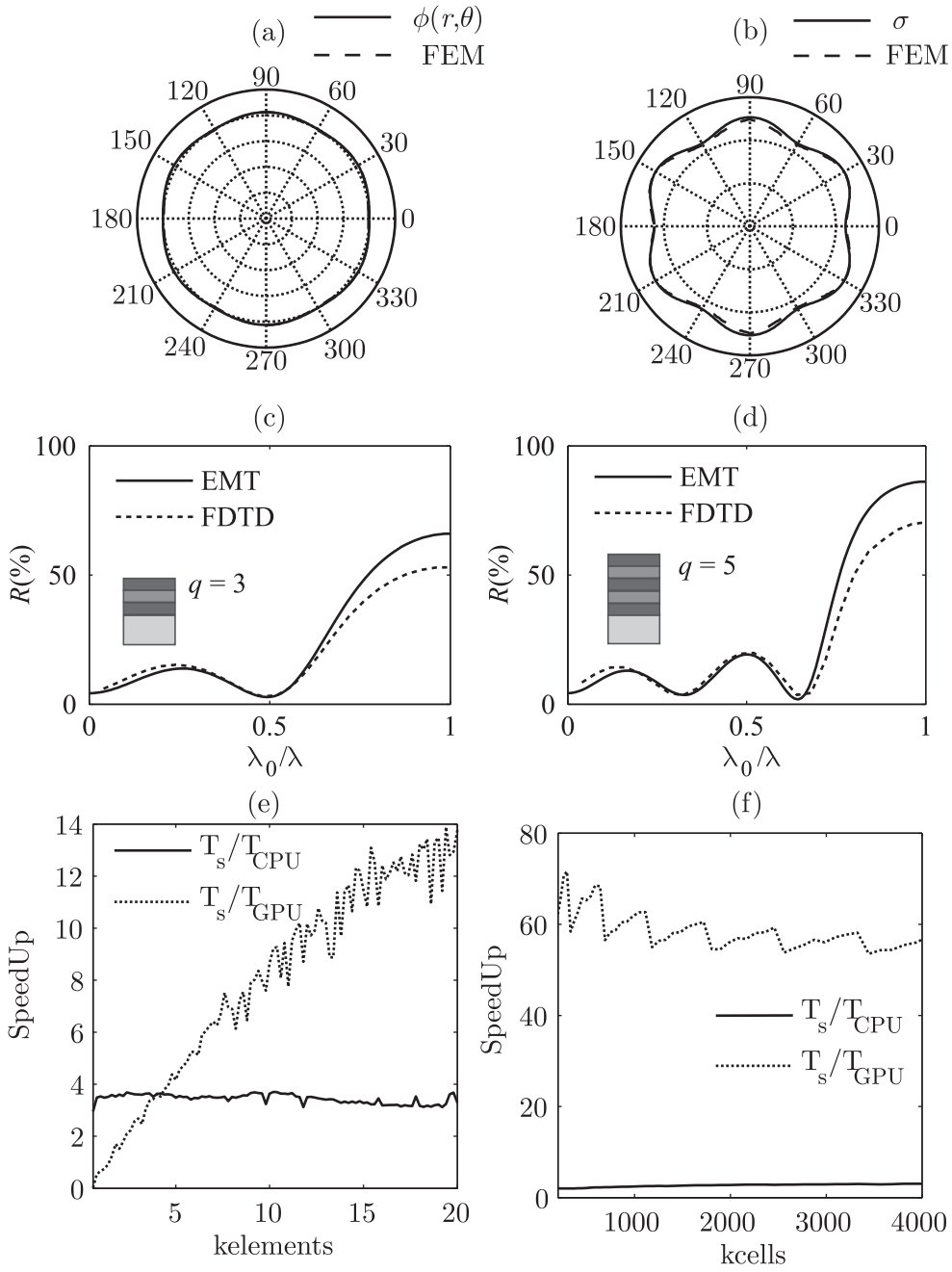


Figure 4: (a) Analytical and numerical potential for $\varepsilon = 0.02$ and $n = 6$.(b) Analytical and numerical charge density for $\varepsilon = 0.02$ and $n = 6$. (c) and (d) reflectance of the stack of homogeneous dielectric layers. (e) Comparison of sequential versus CPU auto-vectorized and GPU codes for FEM method as a function of the number of triangular elements. (f) Comparison of sequential versus CPU auto-vectorized and GPU codes for the FDTD method as a function of the number cells considering square grids.

 ISBN: 978-84-614-6167-7

Acknowledgements

This work was supported by the "Ministerio de Ciencia e Innovación" of Spain under projects FIS2008-05856-C02-01 and FIS2008-05856-C02-02, and by the "Generalitat Valenciana" of Spain under project PROMETEO/2011/021.

References

[1] Nvidia Corporation, *Whitepaper NVIDA's Next Generation CUDA Compute Architecture*, Version 1.1, 2009.

[2] Intel Corporation, *Intel 64 and IA-32 Architectures: Optimization Reference Manual*, April, 2011.

[3] C. Neipp, J. C. Moreno, J. J. Rodes, J. Francés, M. Pérez-Molina, S. Gallego, A. Beléndez, *Relation Between Surface Curvature and Electrostatic Potential of a Long Charged Isolated Conductor*, J. of Electromagn. Waves and Appl., **24** (2010) 1647–1659.

[4] D. M. Sullivan, *Electromagnetic Simulation using the FDTD Method*, IEEE Press Editorial Board, 2000.

[5] A. Taflove, *COMPUTATIONAL ELECTRODYNAMICS: The Finite-Difference Time-Domain Method*, Artech House Publishers, 1995.

[6] H. A. Macleod, *Thin-Film Optical Filters*, Institute of Physics Publishing, 2001.

[7] M. N. O. Sadiku, *Numerical Techniques in Electromagnetics* , CRC Press, New York, 2001.

[8] W. H. Press, S. A. Teukolsky, W. T. Vetterling and B. P. Flannery, *NUMERICAL RECIPES: The Art of Scientific Computing*, Cambridge University Press, 2007.

[9] S. Thakkar, *The Internet Streaming SIMD Extensions*, Intel Technology Journal **Q2** (1999) 1–8.

[10] K. S. Yee, *Numerical solution of initial boundary value problems involving Maxwell's equations in isotropic media*, IEEE Trans. On Antennas and Propagation **17** (1996) 585–589.

[11] S. M. Sullivan, *A simplified PML for use with the FDTD method*, IEEE Microwave and Guided Wave letters **6** (1996) 97–99.

[12] J. Sanders and E. Kandrot, *CUDA by Example: An introduction to general-purpose GPU programming*, Addison-Wesley, Upper Saddle River, NJ, 2011.

[13] NVIDIA CORPORATION, *CUDA: CUSPARSE Library*, NVIDIA, August, 2010.

 ISBN: 978-84-614-6167-7

Proceedings of the 11th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2011
26–30 June 2011.

Integrating dense and sparse data partitioning

Javier Fresno[1], Arturo González-Escribano[1] and Diego R. Llanos[1]

[1] *Dept. Informática, Universidad de Valladolid*

emails: javfres@gmail.com, arturo@infor.uva.es, diego@infor.uva.es

Abstract

Layout methods for dense and sparse data are often seen as two separate problems with its own particular techniques. However, they are based on the same basic concepts. This paper studies how to integrate automatic data-layout and partition techniques for both dense and sparse data structures. In particular, we show how to include support for sparse matrices or graphs in Hitmap, a library for hierarchical-tiling and automatic mapping of arrays. The paper shows that is possible to offer a unique interface to work with both dense and sparse data structures, without losing significant performance. Thus, the programmer can use a single and homogeneous programing style, reducing the development effort and simplifying the use of sparse data structures in parallel computations.

Key words: data partition, layouts, distributed computing, sparse.

1 Introduction

In parallel applications the data distribution and layout is a key issue that can determine the performance and scalability. Regarding the type data structures, dense or sparse, the data distribution problem is treated differently. On the one hand, there are many languages with primitives and tools to deal with data locality and/or distribution, such as HPF [11], OpenMP [4], or UPC [2]. On the other hand, sparse structure support are not usually integrated in the programming languages. However, a wide range of important problems are based in unstructured graph structures instead of dense arrays. The common approach to manage sparse data is using a library. We can find a high number of libraries for partitioning graphs, meshes and other sparse structures, such as Metis [10], Scotch [12], or Jostle [13].

 ISBN: 978-84-614-6167-7

There are some proposals that use an unique representation for different domains. Chapel, a new parallel language, uses a representation of indexes-set called a domain [3]. The Chapel's proposal aims to support domains for dense, sparse, strided, associative, and unstructured data aggregates. However, there is not yet a full nor efficient implementation.

Another previous proposal [9] have evolved into the Trasgo system and the Hitmap library. Hitmap is a basic tool in the back-end of the Trasgo compilation system [8]. Trasgo proposes the use of a global-view approach, with flexible and explicit mechanisms to deal with locality. The code generated by Trasgo uses the Hitmap library to perform a highly efficient data distribution and aggregated communications.

In this paper, we present a new approach to hide the internal details of dense and sparse data structure, using a common interface to deal with both types of data. Combining the dense and sparse manipulation under a common interface has great advantages. It simplifies programming by hiding the partition details in reusable and flexible plug-ins. The programmer focus on the algorithm parallel implementation without thinking in terms of the underlying data structure. Coding a parallel application follows the same basic pattern, using the same API for the same functionalities. The implementaton of Hitmap abstractions is focused on further native compiler optimizations. Our experimental results show that using Hitmap simplifies programing with a negligible impact on performance.

This paper is organized as follows: Section 2 provides a brief overview of the Hitmap library. Section 3 introduces the benchmark that illustrates our proposal and describes its different implementations. Section 4 contains the experimental work. Finally, the paper ends with the conclusions at Section 5.

2 Hitmap library

Hitmap is a highly-efficient library for hierarchical tiling and mapping of arrays [6, 7]. It aims to simplify parallel programming, providing functionalities to create, manipulate, distribute, and communicate tiles and hierarchies of tiles. In this section we will present the basic ideas of the Hitmap library needed for the further discussion.

Hitmap library supports functionalities to: (1) Generate a virtual topology structure; (2) mapping the data to the different processor with chosen load-balancing techniques; (3) automatically determine inactive processors at any stage of the computation; (4) identify the neighbor processors to use in communications; and (5) build communication patterns to be reused across algorithm iterations.

Hitmap is designed with an object-oriented approach, although it is implemented in C language. Fig. 1 shows a class diagram of the library architecture. The classes are implemented as C structures with associated functions. A *Signature*, represented by a HitSig object, is a selection of array indexes in a one-dimensional domain. Hitmap uses a *Shape* object to represents a domain of data. In the previous Hitmap version, this object

 ISBN: 978-84-614-6167-7

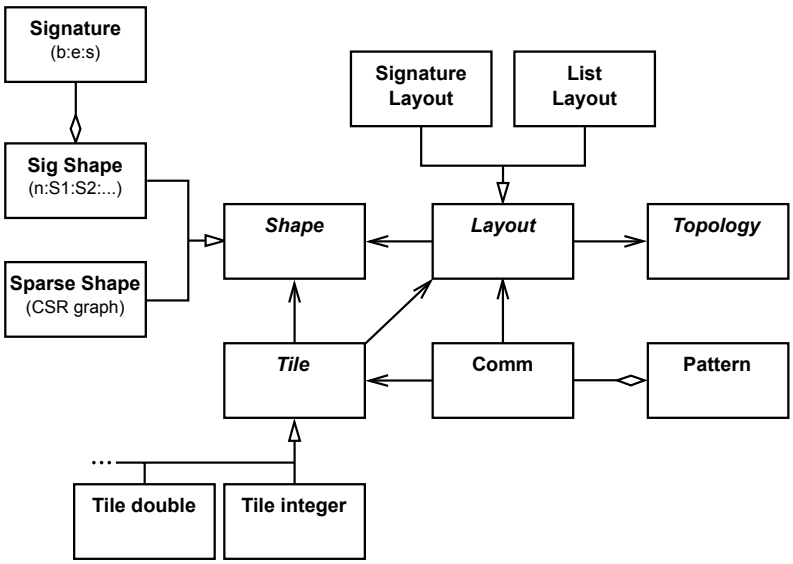


Figure 1: Hitmap new architecture.

was composed by HitSig objects, and it represented a contiguous or stride subset of indexes in a multidimensional domain. The new Hitmap version uses inheritance to integrate shape objects that represent sparse data domains. A *Tile* is an array which domain is defined by a shape. Hitmap has functionalities to dynamically declare selections of tiles to construct tile hierarchies. A tile may be defined without allocated memory allowing to declare and partition arrays before assigning memory to them, finally allocating only the parts mapped to a given processing unit.

Hitmap provides programming tools to apply different data-partition and layout techniques over automatically generated virtual topologies, hidding the details of the physical processors, topology, and mapping. Both the virtual topology generation and the partition techniques are integrated in the library as plug-in modules extending abstract classes. Programmers may include their own new techniques.

The virtual topology techniques are invoked by name with no extra parameters. They use the internal information of the target system. The result is a HitTopology object which can be queried and it is used as a parameter for data partition and layout. The layout plug-in modules allow to compute a partition of a shape domain over a virtual topology. The result of a layout plug-in is a HitLayout object that contains information about the local part of the domain mapped to the current processor, the neighbor relations, and methods to return or compute on the fly all the information needed to exchange data. The plug-ins encapsulate the computations needed to deal with physical topology and data location at

 ISBN: 978-84-614-6167-7

all the mapping stages; details which are usually hardwired in the code by the programmer. The combination of these plug-in systems allows the programmer to easily create abstract codes which also simplifies debugging operations with any kind of data structures.

Finally, an information exchange operation can be specified creating a HitCom object. The constructor receives a HitShape to specify the data to be moved and a HitLayout object with the mapping and topology information. The result is a HitCom object with all the information needed to execute the data exchange as many times as required. Moreover, several HitCom objects can be composed in reusable communications patterns represented by HitPattern objects. The library is built on top of the MPI communication library, for portable communication and synchronization on different architectures. Hitmap internally exploits several MPI techniques that increase performance.

2.1 Hitmap Sparse support

The previous version of the Hitmap library was oriented to manipulate and communicate tiles of data with contiguous domains, or non-contiguous but regular index selections. In this new version of the library, we have extended the Hitmap functionalities to support sparse data structures. This allows the programmer to work with graph or sparse matrices using a similar API, and a homogeneous programming methodology. To support the manipulation, mapping, and communication of these new data structures, we need to make several structural changes in the library.

The first change is related to the data domains represented by the HitShape class. In the previous version of the library, a HitShape object was composed by several HitSig objects, to represent a multidimensional selection of indexes. We have transformed the HitShape class into an interface, with two different implementations, one for the old dense domain and another for the new sparse domain (see Fig. 1).

The HitSparseShape class encapsulate a sparse matrix format to represent the sparse data domain. The first sparse matrix format that we have implemented is Compressed Sparse Row (CSR). It is a well-known and widely used format for sparse data. CSR is simple; it does not make any assumptions about the matrix structure and it has minimal storage requirements [1]. There are other formats that can offer a better performance in some particular applications. It is possible to support any of these formats with new implementations of the proposed interface.

To illustrate how to develop new layout plug-ins that make specific partitions and mapping techniques for sparse domains, we have implemented an example HitLayout plug-in. We have integrated one of the graph partition techniques of the Metis library [10]. The plug-in receives a HitShape with the sparse domain, and it calls the Metis library to compute the local part. Metis also uses the CSR format. Thus, the plug-in needs to apply a minimal data-format transformation. With the result returned by Metis, the plug-in creates a new HitShape with the local graph part. The resulting HitLayout object can optionally

 ISBN: 978-84-614-6167-7

contain lists with vertices belonging to other processors. This information can be used to find the owner of a given vertex and to exchange data between processors.

To implement simple FDM (Finite Differences Methods) applications, we have included code to automatically compute the list of vertices in other processors which are adjacent to the local vertices. This example plug-in can be further extended to include more neighbors information useful for more complex FEM (Finite Elements Methods) applications.

The tiles constructor receives a HitShape object. There is a method to allocate the memory for the associated data. The tile constructor internally checks the type of shape. For sparse shapes we allocate memory for the vertices. Extending the current implementation to store weight information for the edges is straightforward. We have also added macros and functions to easily access data and iterate across the vertices values.

In some applications, a processor needs the values of neighbor vertices mapped to other processors. The new version of the HitCom object supports a new global communication type. It is designed to exchange the data of adjacent vertices assigned to different processors. The HitCom object uses the internal information calculated in the layout to determine which vertices should be sent to any other processor, and which vertices should the current processor receive from any other processor.

Finally, other functionalities have been added to the library to facilitate the sparse data management. For example, functions for input of sparse data, like reading the Harwell-Boeing format or plain CSR format.

3 Neighbor-vertices synchronization benchmark

In this section we discuss the methodology followed to create a benchmark to test the efficiency of our implementation. We select a simple problem that involves a computation over a sparse data structure.

We extend the idea of neighbor synchronization in FDM applications on dense matrices to civil engineering structural graphs, see e.g. [14]. The application performs several iterations of a graph update operation. It traverses the graph nodes, updating each node value with the result of a function on the neighbor nodes values. To simulate the load of a real scientific application, we write a dummy loop, which issues 10 times a mathematical library operation (sin). We use as benchmarks different graphs from the Pothen group of the *University of Florida Sparse Matrix Collection* [5].

The codes have been run on Geopar, an Intel S7000FC4URE server, equipped with four quad-core Intel Xeon MPE7310 processors at 1.6GHz and 32GB of RAM. The MPI implementation used is MPICH2, compiled with a backend that exploits shared memory for communications if available in the target system.

 ISBN: 978-84-614-6167-7

3.1 C implementation

We have first developed a serial C implementation of the benchmark, to use as the reference to measure speedups and to verify the results of the parallel versions.

Our first parallel version includes the code to compute the data distribution manually. This highly-tuned and efficient implementation is based on Metis library to calculate the partition of the graph, and MPI to communicate the data between processors. The Hitmap library implementation is also based on the same tools. Thus, the comparison between performance results of a Hitmap parallel version and the manual one will show any potential inefficiency introduced in the design or implementation of Hitmap.

The manual parallel version has the following stages: (1) A sparse graph file is read; (2) the data partition is calculated using the Metis library; (3) each processor initializes the values of the local vertices using a random function; (4) a loop performs 10,000 iterations of the main computation, including updating the values of each local vertex, and the communication of the values for neighbor vertices to other processors; (6) the final result is checked with the help of a hash function.

3.2 Hitmap implementation

Using the manual C implementation as a starting point, we have developed a Hitmap version of the program. The Hitmap implementation uses the main computation and other sequential parts of the previous one, adapting them to work with Hitmap functions for accessing data structures. A new layout plug-in module has been developed to apply the Metis data partition to the Hitmap internal sparse-shape structures. Data-layout and communications have been generated using Hitmap functionalities. In this section we discuss the Hitmap techniques needed to automatically compute the data-layout, allocate the proper part of the graph and communicate the neighbor vertices values.

In Fig. 2 we show the main function of the Hitmap code. The first line initializes the Hitmap environment. Line 5 uses a function to read a graph stored in the file system, and returns a shape object. Then, a virtual topology of processors that uses the internal information available about the real topology is created transparently to the programmer with a single call.

In line 8, the data-layout is generated with a single Hitmap call. The layout parameters are: (a) the layout plug-in name, (b) the virtual topology of processors generated previously, and (c) the shape with the domain to distribute. The result is a HitLayout object, containing the shape assigned to the local processor and information about the neighbors.

In line 11, we obtain the shape of the local part of the graph with only the local vertices. On the following line, we use the layout to obtain an extended shape with local vertices plus the neighbor vertices from other processors. This is the equivalent to the shape of a

 ISBN: 978-84-614-6167-7

```
1     // Read the global graph.
2     HitShape shape_global = hit_shapeHBRead("graph_file.rb");
3
4     // Create the topologoy object.
5     HitTopology topo = hit_topology(plug_topPlain);
6
7     // Distribute the graph among the processors.
8     HitLayout lay = hit_layout(plug_layMetis,topo,&shape_global);
9
10    // Get the shapes for the local and local extended graphs.
11    HitShape local_shape = hit_layShape(lay);
12    HitShape ext_shape = hit_layExtendedShape(lay);
13
14    // Allocate memory for the graph.
15    HitTile_double graph;
16    hit_tileDomainShapeAlloc(&graph,sizeof(double),HIT_NONHIERARCHICAL,ext_shape);
17
18    // Init the local graph.
19    init_graph(graph, shape_global);
20
21    // Create the communicator and send the initial values of the neighbor vertices.
22    HitCom com = hit_comSparseUpdate(lay, &graph, HIT_DOUBLE);
23    hit_comDo(&com);
24
25    int i;
26    // Update loop.
27    for(i=0; i<ITERATIONS; i++){
28       // Update the graph.
29       synchronization_iteration(local_shape,ext_shape,&graph);
30       // Communication.
31       hit_comDo(&com);
32    }
```

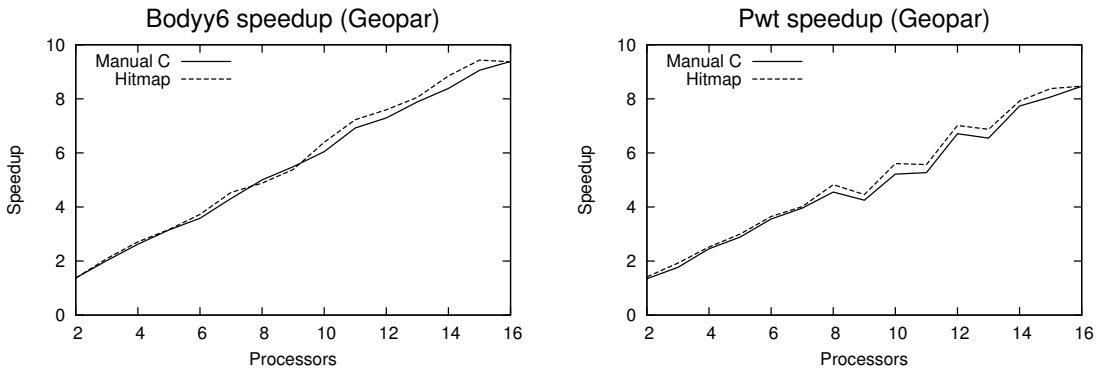
Figure 2: Kernel code of the Hitmap version.

 ISBN: 978-84-614-6167-7

```
int vertex, edge;

// Iterate through all the vertices.
hit_sparseShapeVertexIterator(vertex,local_shape){

   // Set new value to 0.
   double value = 0;

   hit_sparseShapeEdgeIterator(edge,ext_shape,vertex){

      // Get the neighbor.
      int neighbor = hit_sparseShapeEdgeTarget(ext_shape,edge);
      // Add its contribution.
      value +=  hit_tileElemAt(1,graph,neighbor);
   }

   // Dummy workload = 10.
   int i;
   for(i=0; i<WORKLOAD; i++){
      value = sin(value+1);
   }

   int nedge = hit_sparseShapeNumberEdgesFromVertix(ext_shape,vertex);

   // Update the value of the vertex.
   hit_tileElemAt(1,graph_aux,vertex) =  (value / nedge);
}
```

Figure 3: Function that serially updates the local part of the graph.

 ISBN: 978-84-614-6167-7



(a) Speedup for the bodyy6 graph.

(b) Speedup for the pwt graph.

Figure 4: Speedup in Geopar.

tile with a shadow region in a FDM solver for dense matrices. This shape is used to declare and allocate the local tile with double elements (line 16).

In line 22, a HitCom object is created to contain the information needed to issue the communications that will update the neighbor vertices values. Data marshaling and unmarshaling is automatized by the communications objects when the communication is invocated (see lines 23 and 31).

Lines 25 to 32 contain the main iteration loop to update local nodes values and reissue communications with a single Hitmap call thar reuses the HitCom object previously defined.

Fig. 3 show the code of the function that updates the local values of the graph. It uses two iterators defined in the library. The first one traverses the local vertices (line 4), and the second one iterates over the edges to get their neighbor vertices (line 9). Each neighbor-vertex value contributes to the new value of the local vertex that is set in line 20.

4 Experimental results

In this section we compare the performance obtained with the benchmark described in the previous section. We have tested the benchmark with different graphs from the Pothen group of the *University of Florida Sparse Matrix Collection* [5]. In this section, we discuss the results from two representative cases in the group collection: The bodyy6 graph with 19366 vertices and pwt graph with 36519 vertices. Fig. 4 shows the speedup for both manual C and Hitmap benchmark implementations. There is no significant difference between the two implementations in terms of performance. Therefore, the abstractions introduced by Hitmap (such as the common interface for dense and sparse data structures, or the adapta-

 ISBN: 978-84-614-6167-7

tion of the partition technique in the plug-in module system), do not lead to performance reduction comparing with the manual version.

Fig. 5 shows a comparison of the Hitmap version with the C version in terms of lines of code. We distinguish lines devoted to sequential computation, declarations, parallelism (data layouts and communications), and other non-essential lines (input-output, etc). Taking into account only essential lines, our results show that the use of Hitmap library leads to a 72% reduction on the total number of code lines with respect to C version. Regarding lines devoted specifically to parallelism, the percentage of reduction is 78.7%. We have also use the cyclomatic complexity metric to compare the codes. The total cyclomatic complexity of the manual version is 74 whereas the Hitmap versions has a total value of 17. The reason for the reduction of complexity in the Hitmap version is that Hitmap greatly simplifies the programmer effort for data distribution and communication, compared with the equivalent code to manually calculate the information needed in the MPI routines.

5 Conclusions

This paper studies how to integrate the support for dense and sparse data structures in an automatic data partitioning parallel library. We have add sparse structure support in Hitmap, a highly-efficient modular library for hierarchical tiling and mapping of arrays. We have illustrated how to use the library to implement a simple graph algorithm. We have also measured the efficiency of the library in terms of performance comparing with a manual implementation. The results show that the abstraction introduced by the library does not reduce performance. We also measure the code complexity in terms of lines of code and cyclomatic complexity. Our results show that it is possible to use a common interface for both dense and sparse data structures with a homogeneous coding style, and reducing the associated development cost comparing with manually coding the data structure management, its partition, and the communication of locally mapped subdomains when needed. As it is shown by the experimental results, this can be done without sacrificing performance.

Our ongoing work includes the integration of new partition techniques in the Hitmap framework. For example, there are other libraries that could be used instead of the Metis library, with different partitioning properties. We are also working on alternative implementations for the communication classes, that are currently built on top of the MPI communication library, to better exploit different low-level parallel tools and models.

Acknowledgements

This research is partly supported by the Ministerio de Industria, Spain (CENIT MARTA, CENIT OASIS, CENIT OCEANLIDER), Ministerio de Ciencia y Tecnología (CAPAP-H3 network, TIN2010-12011-E), and the HPC-EUROPA2 project (project number: 228398)

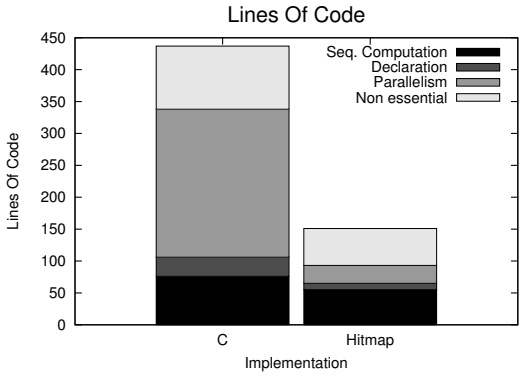


Figure 5: Comparison of the lines of code.

with the support of the European Commission - Capacities Area - Research Infrastructures Initiative.

References

[1] Richard Barrett, Michael Berry, Tony. F. Chan, James Demmel, June M. Donato, Jack Dongarra, Victor Eijkhout, Roldan Pozo, Charles Romine, and Henk Van der Vorst. *Templates for the Solution of Linear Systems: Building Blocks for Iterative Methods, 2nd Edition*, volume 64. SIAM, July 1995.

[2] William W. Carlson, Jesse M. Draper, David E. Culler, Kathy Yelick, Eugene Brooks, and Karen Warren. Introduction to UPC and Language Specification, 1999.

[3] Bradford L Chamberlain, Steven J Deitz, David Iten, and Sung-Eun Choi. User-defined distributions and layouts in chapel: philosophy and framework. In *Proceedings of the 2nd USENIX conference on Hot topics in parallelism*, HotPar'10, page 12, Berkeley, CA, USA, 2010. USENIX Association.

[4] Robit Chandra, Leonardo Dagum, Dave Kohr, Dror Maydan, Jeff McDonald, and Ramesh Menon. *Parallel programming in OpenMP*. Morgan Kaufmann, 1 edition, 2001.

[5] Timothy A. Davis and Yifan Hu. The University of Florida Sparse Matrix Collection. *To appear in ACM Transactions on Mathematical Software.*

[6] Carlos de Blas Cartón, Arturo González-Escribano, and Diego R. Llanos. Effortless and Efficient Distributed Data-Partitioning in Linear Algebra. In *2010 IEEE 12th Inter-*

national Conference on High Performance Computing and Communications (HPCC), pages 89–97. IEEE, September 2010.

[7] Javier Fresno, Arturo González-Escribano, and Diego R. Llanos. Automatic Data Partitioning Applied to Multigrid PDE Solvers. In *2011 19th International Euromicro Conference on Parallel, Distributed and Network-Based Processing*, pages 239–246. IEEE, February 2011.

[8] Arturo González-Escribano and Diego R. Llanos. Trasgo: a nested-parallel programming system. *The Journal of Supercomputing*, December 2009.

[9] Arturo González-Escribano, Arjan J.C. van Gemund, Valentín Cardeñoso Payo, Raúl Portales-Fernández, and Jose A. Caminero-Granja. A preliminary nested-parallel framework to efficiently implement scientific applications. *High Performance Computing for Computational Science-VECPAR 2004*, pages 541–555, 2005.

[10] George Karypis and Vipin Kumar. MeTiS–A Software Package for Partitioning Unstructured Graphs, Partitioning Meshes, and Computing Fill-Reducing Orderings of Sparse Matrices–Version 4.0, 1998.

[11] Ken Kennedy, Charles Koelbel, and Hans Zima. The rise and fall of High Performance Fortran. In *Proceedings of the third ACM SIGPLAN conference on History of programming languages - HOPL III*, pages 7.1–7.22, New York, New York, USA, 2007. ACM Press.

[12] François Pellegrini. PT-Scotch and libScotch 5.1 User's Guide, 2010.

[13] Chris Walshaw. The serial JOSTLE library user guide : Version 3.0, 2002.

[14] Barry Wilkinson and Michael Allen. *Parallel Programming: Techniques and Applications Using Networked Workstations and Parallel Computers*. Prentice Hall, 2004.

 ISBN: 978-84-614-6167-7

Proceedings of the 11th International Conference on Computational and Mathematical Methods in Science and Engineering, CMMSE 2011 26–30 June 2011.

Improving the discrete wavelet transform computation from multicore to GPU-based algorithms

V. Galiano[1], O. López[1], M.P. Malumbres[1] and H. Migallón[1]

[1] *Physics and Computer Architecture Department, Miguel Hernández University, 03202 Elche, Spain*

emails: vgaliano@umh.es, otoniel@umh.es, mels@umh.es, hmigallon@umh.es

Abstract

In this work we analyze the behavior of some parallel algorithms when computing the two dimensional discrete wavelet transform (2D-DWT) using both OpenMP over a multicore platform and CUDA (Compute Unified Device Architecture) over a GPU (Graphics Processing Unit). The proposed algorithms are based on both regular filter-bank convolution and lifting transform. Finally we will also compare our algorithms against other recently proposed algorithms.

Key words: CUDA, OpenMP, wavelet transform, image coding, parallel algorithms

Introduction

During the last decade, several image compression schemes emerged in order to overcome the known limitations of block-based algorithms that use the Discrete Cosine Transform (DCT) [1], the most widely used compression technique at that moment. Some of these alternative proposals were based on more complex techniques, like vector quantization and fractal image coding, while others simply proposed the use of a different and more suitable mathematical transform, the Discrete Wavelet Transform (DWT). Wavelet transforms have proven to be very powerful tools for image compression and many state-of-the-art image codecs, including the JPEG2000 image coding standard, employ a wavelet transform in their algorithms (see for example [2, 3])

Unfortunately, despite the benefits that the wavelet transform entails, some other problems are introduced. Wavelet-based image processing systems are typically implemented by memory-intensive algorithms, with higher execution time than other transforms. In the usual DWT implementation [4], the image decomposition is computed by means of convolution filtering process and so, its complexity rises as the filter length increases. Moreover, in the regular DWT computation, the image is transformed at

 ISBN: 978-84-614-6167-7

every decomposition level first row by row and then column by column, and hence it must be kept entirely in memory. These problems are not as noticeable in other image transforms as in the DWT. For example, when the DCT is used for image compression, it is applied in small block sizes, and thus a large amount of memory is not specifically needed for the transformation process.

The lifting scheme [5, 6] is probably the best-known algorithm to calculate the wavelet transform in a more efficient way. Since it uses less filter coefficients than the equivalent convolution filter, it provides a faster implementation of the DWT. This scheme also provides memory reduction through in-place computation of wavelet coefficients. However, if in-place computation is applied, the low-frequency coefficients are interleaved with the wavelet coefficients, and the subsequent wavelet processing can be non-optimal (specially in cache-based systems), requiring a more careful process. We can overcome this problem with coefficient reordering, at the cost of increasing the complexity of the algorithm.

Other fast wavelet transform algorithms has been proposed in order to reduce both memory requirements and complexity, like line-based [7] and block-based [8] wavelet transform approaches that performs wavelet transformation at image line or block level. These approaches increases flexibility when applying wavelet transform and significantly reduce the memory requirements. At the other hand, in [9], authors present a novel way of computing the wavelet transform called Symmetric Mask-based Discrete Wavelet Transform (SMDWT). This new wavelet transform is computed as a matrix convolution, using four matrix masks, one for each subband type, that are built in order to reduce the repetitive computations found in the classical convolution approach. In this scheme, the 2D-DWT is performed in only one pass, avoiding multiple-layer transpose decomposition operations. One of the most interesting advantages of this method is that the computation of each wavelet subband is completely independent.

When designing fast wavelet-based image/video encoders, one of the most computational intensive tasks is the 2D-DWT, which in some cases may take up between 30% and 50% of the overall encoding time (depending of image size and the number of decompositions levels). So, it is very important to reduce 2D-DWT computation time to develope fast real-time image/video encoders. To do that, we will take profit of the available hardware resources that are present in current off-the-shelf computers, in particular multicore processing and GPU co-processing units.

In this paper, we perform optimized parallel algorithms based on the methods introduced in [4] and [5]. The main goals of the performed optimizations are to obtain low memory requirements as well as good computational behavior, exploiting multicore architectures, i.e. shared memory platform. After that, we will apply the same scheme introduced in the multicore algorithm to develop CUDA-based DWT algorithms on GPU. Algorithms developed on Graphics Processing Units (GPU) require an efficient use of memory to exploit the GPU architecture in an efficient way. The developed algorithms are focused in the use of the GPU shared memory. We have also compared the CUDA based algorithms developed with the algorithms proposed in [10], in both computation performance and memory requirements.

 ISBN: 978-84-614-6167-7

1 Discrete Wavelet Transform (DWT)

DWT is a multiresolution decomposition scheme for input signals, see detailed description in [4]. The original signals are firstly decomposed into two frecuency subbands, low-frequency (low-pass) subband and high-frequency (high-pass) subband. For the classical DWT, the forward decomposition of a signal is implemented by a low-pass digital filter H and a high-pass digital filter G. Both of digital filters are derived using the scaling function $\Phi(t)$ and the corresponding wavelets $\Psi(t)$. The system downsamples the signal to half of the filtered results in decomposition process. If the four-tap and non-recursive FIR filters with length L are considered, the transfer functions of H and G can be represented as follows:

$$H(z) = h_0 + h_1 z^{-1} + h_2 z^{-2} + h_3 z^{-3} \tag{1}$$

$$G(z) = g_0 + g_1 z^{-1} + g_2 z^{-2} + g_3 z^{-3} \tag{2}$$

1.1 Lifting-based Wavelet Transform (LDWT)

One of the drawbacks of the DWT is that it doubles the memory requirements because it is implemented as a filter. A proposal that reduces the amount of memory needed for the computation of the 1D DWT is the lifting scheme [5]. Despite this disadvantage, the main benefit of this scheme is the reduction in the number of operations needed to perform the wavelet transform if compared with the usual filtering algorithm (also known as convolution algorithm). The order of this reduction depends on the type of wavelet transform, as shown in [11].

The lifting scheme implements the DWT decomposition as an alternative algorithm to the classical filtering algorithm introduced in the previous section. In the filtering algorithm, in-place processing is not possible because each input sample is required as incoming data for the computation of its neighbor coefficients. Therefore, an extra array is needed to store the resulting coefficients, doubling the memory requirements. On the other hand, the lifting scheme provides in-place computation of the wavelet coefficients and hence, it does not need extra memory to store the resulting coefficients. In addition, the lifting scheme can be computed on an odd set of samples, while the regular transform requires the number of input samples to be even.

The Euclidean algorithm can be used to factorize the poly-phase matrix of a DWT filter into a sequence of alternating upper and lower triangular matrices. In 3, the variables $h(z)$ and $g(z)$ denote the low-pass and high-pass inverse filters, respectively, which can be divided into even and odd parts to form a poly-phase matrix $P(z)$ as in 4.

$$g(z) = g_e\left(z^2\right) + z^{-1} g_o\left(z^2\right), \;\; h(z) = h_e\left(z^2\right) + z^{-1} g_o\left(z^2\right) \tag{3}$$

$$P(z) = \begin{pmatrix} h_e(z) & g_e(z) \\ h_o(z) & g_o(z) \end{pmatrix} \tag{4}$$

Using the Euclidean algorithm, it recursively finds the greatest common divisors of the even and odd parts of the original filters. Since $h(z)$ and $g(z)$ form a complementary

 ISBN: 978-84-614-6167-7

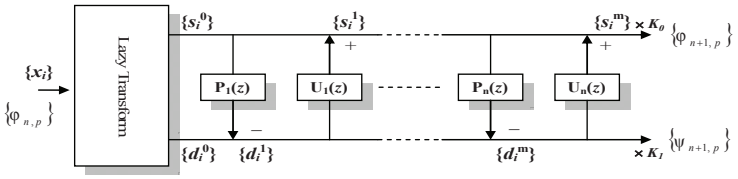


Figure 1: General diagram for a wavelet decomposition using the lifting scheme.

filter pair, $P(z)$ can be factorized into three lifting steps as below.

$$P(z) = \prod_{i=1}^{m} \begin{pmatrix} 1 & s_i(z) \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ t_i(z) & 1 \end{pmatrix} \begin{pmatrix} k & 0 \\ 0 & 1/k \end{pmatrix} \tag{5}$$

where $s_i(z)$ and $t_i(z)$ denote the Laurent polynomials corresponding to prediction and update steps, respectively, and k is a nonzero constant.

The whole process consists of a first lazy transform, one or several prediction and update steps, and coefficient normalization. In the lazy transform, the input samples are split into two data sets, one with the even samples and the other one with the odd ones. Thus, if we consider $\{X_i\} = \{\Phi_{n,p}\}$ the input samples at a level n, we define:

$$\left\{s_i^0\right\} = \{X_{2i}\} \tag{6}$$

$$\left\{d_i^0\right\} = \{X_{2i+1}\} \tag{7}$$

Then, in a prediction step (sometimes called dual lifting), each sample in $\left\{d_i^0\right\}$ is replaced by the error committed in the prediction of that sample from the samples in $\left\{s_i^0\right\}$:

$$d_i^1 = d_i^0 - P\left(\left\{s_i^0\right\}\right) \tag{8}$$

while in an update step (also known as primal lifting), each sample in the set $\left\{s_i^0\right\}$ is updated by $\left\{d_i^1\right\}$ as:

$$s_i^1 = s_i^0 + U\left(\left\{d_i^1\right\}\right) \tag{9}$$

After m successive prediction and update steps, the final scaling and wavelet coefficients are achieved as follows:

$$\{\Phi_{n+1,p}\} = K_0 \times \{s_i^m\} \tag{10}$$

$$\{\Psi_{n+1,p}\} = K_1 \times \{d_i^m\} \tag{11}$$

A special case of wavelet filter is the Daubechies 9/7 filter. This filter has been widely used in image compression [3, 12], and it has been included in the JPEG2000 standard [2]. In this paper, all the DWT algorithms will be focused on this filter because

 ISBN: 978-84-614-6167-7

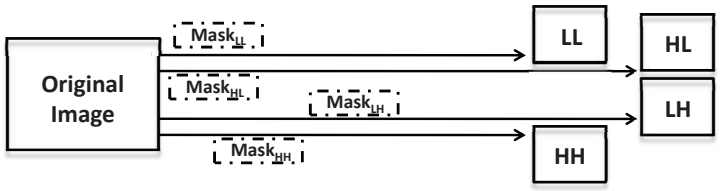


Figure 2: 2D SMDWT structure.

of its goodness behavior. The coefficients of the Daubechies 9/7 decomposition filters, $h[n]$ and $g[n]$ are:

$$\begin{aligned} h[n] &= 0.026749, -0.016864, -0.078223, 0.266864, 0.602949, \\ &\quad 0.266864, -0.078223, -0.016864, 0.026749 \\ g[n] &= 0.091272, -0.057544, -0.591272, 1.115087, \\ &\quad -0.591272, -0.057544, 0.091272 \end{aligned}$$

while the result of the lifting-based decomposition is:

$$P(z) = \begin{pmatrix} 1 & \alpha\left(1+z^{-1}\right) \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ \beta(1+z) & 1 \end{pmatrix} \begin{pmatrix} 1 & \gamma\left(1+z^{-1}\right) \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ \delta(1+z) & 1 \end{pmatrix} \begin{pmatrix} \zeta & 0 \\ 0 & 1/\zeta \end{pmatrix} \tag{12}$$

where $\alpha = -1.586134342$,$\beta = -0.052980118$,$\gamma = 0.882911075$,$\delta = 0.443506852$ and $\zeta = 1.230174105$.

1.2 Symmetric Mask-based Wavelet Transform (SMDWT)

In [9], authors present a novel way of computing the wavelet transform trying to reduce the computational complexity for the wavelet filtering process. The Symmetric Mask-based Discrete Wavelet Transform (SMDWT) is performed as a matrix convolution, using four matrix derived from the 2D DWT of 9/7 floating point lifting-based coefficients (see Figure 2). The 2D LDWT lifting scheme require vertical and horizontal 1D LDWT calculations, and each of the 1D LDWT requires four steps: splitting, prediction, updating, and scaling. Conversely, the four subband 2D SMDWT can be yielded using four independent matrices of size 7×7, 7×9, 9×7 and 9×9 for the 9/7 filter.

2 Multicore Wavelet Transform

We have used the regular filter-bank convolution based on Daubechies 9/7 filter, in order to develop the optimized parallel 2D discrete wavelet transform (DWT), proposed in [4]. On the other hand, we have used the lifting scheme proposed by Sweldens in [5], in order to develop the optimized parallel 2D lifting wavelet transform (LWT). As we have previously mentioned, we require the image size memory space to store the computed

Image Size	Cores	Extra memory size		Image Size	Cores	Extra memory size	
		Conv.	Lifting			Conv.	Lifting
512 x 512	1	520	1024	2048 x 2560	1	2568	4608
	2	1040	2048		2	5136	9216
	4	2080	4096		4	10272	18432
2048 x 2048	1	2056	4096	4096 x 4096	1	4104	8192
	2	4112	8192		2	8208	16384
	4	8224	16384		4	16416	32768

Table 1: Amount of extra memory size using four-tap filter (pixel size).

wavelet coefficients. In the convolution based wavelet transform, an extra memory space to store the current image row/column is required. On the other hand, in lifting wavelet based transform, we need the memory space to store a copy of both one row and one column. Remark that, the SMDWT algorithm requires twice the image size space to perform the four mask filtering.

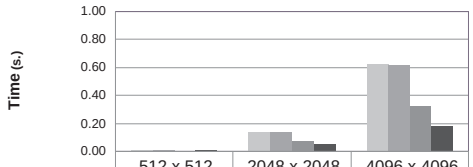
We have used OpenMP [13] paradigm in order to develop the parallel algorithms. The multicore platform used is an Intel Core 2 Quad Q6600 2.4 GHz, with 4 cores, where a block of rows and a block of columns has been assigned to one process in each core to compute the wavelet transform, therefore each process (or core) requires the above mentioned amount of extra memory. Remark that the objective of this buffer is to compute the wavelet transform, so we could store the final wavelet coefficients in the same memory space occupied by the image, avoiding in this manner to double the memory requirements. Table 1 shows the amount of extra memory in pixels (i.e. floats) used by each algorithm depending on the number of cores used. As it can be seen, the worst case is for the smallest image, requiring less than 2% of extra memory overhead, being for the rest of the images less than 1%. As mentioned, the extra memory size needed by the SMDWT algorithm is the size of the image. Note that we work with grayscale images where a pixel is represented by a float, therefore the data shown in Table 1 are pixels or floats.

The operating system used by the multicore platform is Ubuntu 9.04 (Jaunty Jackalope) for 64 bit systems. We have used the GNU compiler `gcc` included in `gcc 4.3.3`. Compiler flags used to exploit the multicore architecture are: "-O3 -m64 -fopenmp", while the ones used to avoid multicore architecture are: "-O3 -m64".

We have considered two scenarios for the parallel algorithms. In the first one, we assign a set of consecutive rows/columns to each processor, while in the second scenario the compiler perform the distribution of computational load. We will not present different results for both scenarios because the computational times obtained are quite similar.

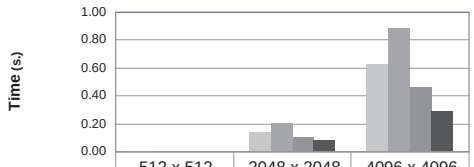
We have tuned the algorithms to obtain the best performance on multicore architectures, taking into account that these algorithms are characterized by an intensive use of memory. In figure 3 we show the computational times obtained for both convolution-based and lifting wavelet transform, for different images sizes: 512×512, 2048×2048, and 4096×4096 pixels. Although the memory access bottleneck is the major obstacle

 ISBN: 978-84-614-6167-7



	512 x 512	2048 x 2048	4096 x 4096
CPU	0.0051	0.1387	0.6278
CPU-1Core	0.0050	0.1384	0.6205
CPU-2Core	0.0029	0.0696	0.3240
CPU-4Core	0.0039	0.0508	0.1839

(a) Convolution



	512 x 512	2048 x 2048	4096 x 4096
CPU	0.0051	0.1387	0.6278
CPU-1Core	0.0037	0.2066	0.8870
CPU-2Core	0.0020	0.1055	0.4626
CPU-4Core	0.0033	0.0871	0.2865

(b) Lifting

Figure 3: Computational times for multicore fast wavelet transform algorithms.

to obtain ideal efficiencies, in Figure 3 we can observe that the computational time decreases, except for small images, as we increase the number of processes. Note that each core executes only one process. Working with small pictures do not achieve good performance due to the relationship between the computational load and the memory accesses degrading the inherent parallelism. Note that each column and/or row has few elements, hence the work performed over each row or column stored in the buffer is not significant. If we calculate efficiency between multicores algorithms, we obtain an efficiency closely ideal one using 2 cores, while we obtain a good efficiency using 4 cores. Note that the memory access bottleneck get worse as the number of cores increase because the number of entities that use the memory is greater.

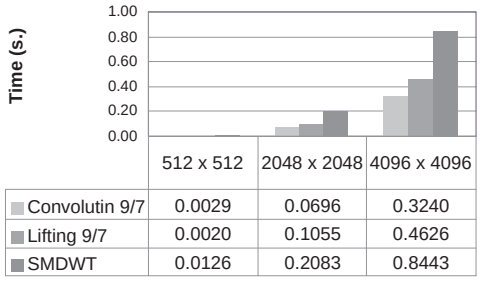
Finally, we have compared our algorithms with a recent and not classical implementation of the fast DWT called "symmetric mask-based DWT" (SMDWT) [9]. We have developed the method introduced in [9] and also, we have parallelized its reference algorithm. In Figure 4 we present a comparison between convolution, lifting and the SMDWT algorithm, using two and four cores. As it can be seen, our convolution and lifting implementations are 2.5 times as fast as the SMDWT algorithm. Note that the authors in [9] propose the SMDWT algorithm to improve the computational complexity of the lifting scheme and also for the ability of the SMDWT algorithm to compute the four subbands (LL, LH, HL and HH) independently.

Some applications only require computing the LL subband, in Figure 5 we present the same comparison as the one in Figure 4, only computing the LL subband when SMDWT algorithm is used, and computing all subbands in our algorithms. Note that the behavior of our algorithms computing the four subbands is similar to the SMDWT behavior only computing the LL subband.

3 CUDA GPU-based Wavelet Transform Algorithm

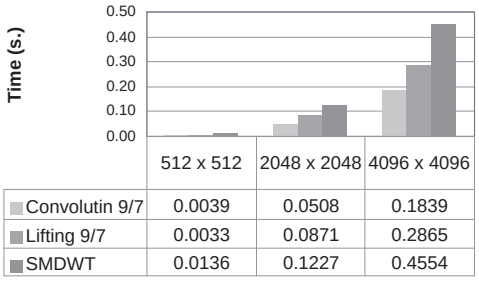
In the previous section, we have confirmed that our shared memory parallel algorithm for computing the 2D DWT presents a good behavior. Moreover, we question in this section if better performance can be achieved with other architecture. The Graphical Processor

 ISBN: 978-84-614-6167-7



	512 x 512	2048 x 2048	4096 x 4096
Convolutin 9/7	0.0029	0.0696	0.3240
Lifting 9/7	0.0020	0.1055	0.4626
SMDWT	0.0126	0.2083	0.8443

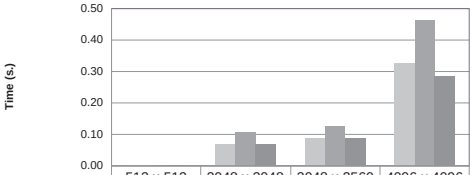
(a) 2 Cores



	512 x 512	2048 x 2048	4096 x 4096
Convolutin 9/7	0.0039	0.0508	0.1839
Lifting 9/7	0.0033	0.0871	0.2865
SMDWT	0.0136	0.1227	0.4554

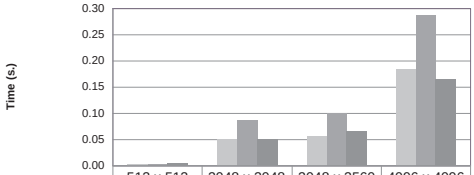
(b) 4 Cores

Figure 4: Comparison between Convolution, Lifting and SMDWT algorithms.



	512 x 512	2048 x 2048	2048 x 2560	4096 x 4096
Convolutin 9/7	0.0029	0.0696	0.0898	0.3240
Lifting 9/7	0.0020	0.1055	0.1245	0.4626
SMDWT	0.0037	0.0698	0.0879	0.2851

(a) 2 Cores



	512 x 512	2048 x 2048	2048 x 2560	4096 x 4096
Convolutin 9/7	0.0039	0.0508	0.0567	0.1839
Lifting 9/7	0.0033	0.0871	0.1007	0.2865
SMDWT	0.0048	0.0501	0.0660	0.1649

(b) 4 Cores

Figure 5: Comparison between Convolution, Lifting and LL subband computation using SMDWT algorithm.

Unit (GPU) architecture is based on a set of multiprocessor units called streaming multiprocessors (SM), containing each one a set of processor cores called streaming processors (SP). CUDA is a heterogeneous computing model that involves both the CPU and the GPU. In the CUDA parallel programming model [14, 15], an application consists of a sequential host program, that may execute parallel programs known as kernels on a parallel device, i.e. a GPU. Note that the CPU could be a multicore processor running an OpenMP model program, but in this case only one of the available cores can call a kernel, i.e. kernel calls must be serialized, therefore we do not use both models in an single algorithm. A kernel is an Single Program Multiple Data (SPMD) computation that is executed using a potentially large number of parallel threads. Each thread runs the same scalar sequential program. The programmer organizes the threads of a kernel into a grid of thread blocks. The threads of a given block can cooperate among themselves using a barrier synchronization. The main kind of memory units available in GPUs are: the global memory, which has the highest latency; the constant and the texture memory units, which are read only units and, the shared memory and the registers, which both are on-chip memory units. The shared memory is owned by a block while the registers are owned by a thread.

So, in order to implement a GPU-based algorithm with the same scheme that the

 ISBN: 978-84-614-6167-7

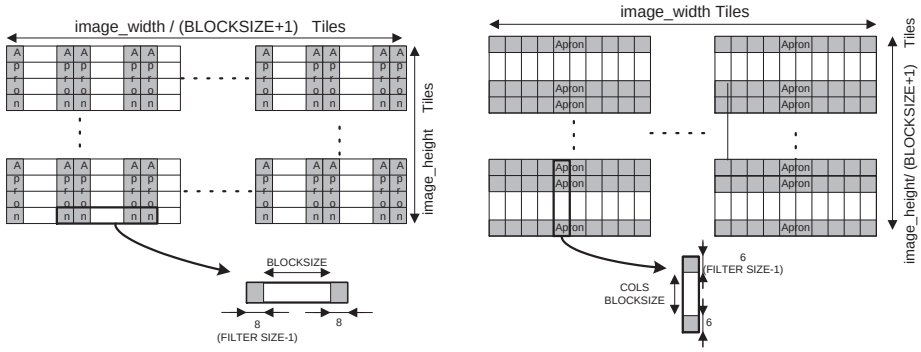
one presented in Section 2, the key element is the use of shared memory to store the buffer that contains a copy of the working data row, and the constant memory to store the filter taps $h[n]$ and $g[n]$. We call each CUDA kernel with a one-dimensional number of blocks NBLOCKS and a one-dimensional number of threads NTHREADS. The number of blocks (NBLOCKS) must be equal to or greater than the maximum size of a row or a column. Each block computes a single row or a single column, the copy of the row or column to be computed is stored in the GPU shared memory. Remark that the available size of shared memory in a GTX 280 is 16KB.

Note that, one of the main goals, in the proposed CUDA based methods is to minimize memory requirements, so we will store the resulting wavelet coefficients in the image memory space. On the other hand, the methods included in the CUDA SDK [10] use three times the image size. These methods perform two steps; in the first step, compute and store, the convolution of rows in GPU global memory, and, in the second step, compute and store the convolution of columns. Remark that, the memory requirements of these methods can be easily reduced using the image memory space to store the wavelet coefficients of the second step. Nevertheless, the memory requirements, using this improvement, is twice the size of the image. We have developed two methods based on the naive implementation described in the SDK (see [10, 16]), the first one using global memory (*CUDA-Mem 9/7*), and the second one using texture memory (*CUDA-Text 9/7*).

As proposed in [10], the behavior of these methods computed over a GPU can be improved optimizing the memory coalescence. In order to optimize the memory coalescence, separable filters must be used. Using a separable filter allows the convolution of rows and convolution of columns to be computed separately. Based on the properties of separable filters, we have developed the method *CUDA-Sep 9/7*, which uses the Daubechies 9/7 filter. The expected improvement should be based on (a) the reduction of the times the pixels are read, (2) on coalescing access to memory, (3) high memory throughput, and (4) the reduction of the number of idle threads (see [10]). As we have said, the convolution is separated in two stages, 1) the rows stage and 2) the columns stage; each stage is separated into two sub-stages, a) the first sub-stage loads the data from global memory into shared memory, and b) the second sub-stage processes the data and stores the results in the global memory. In the computation stage, as it can be seen in Figure 6, each thread loops over a width of twice the filter radius plus 1 (8 in rows and 6 in columns for Daubechies 9/7 filter), multiplying each pixel by the corresponding filter tap stored in the constant memory. Each thread in a half-warp (a warp is composed by 32 CUDA threads) accesses to the same constant memory address and hence there is no penalty due to constant memory bank conflicts. Also, consecutive threads always access consecutive shared memory addresses so no shared memory bank conflicts occur as well, see [10] for a detailed description.

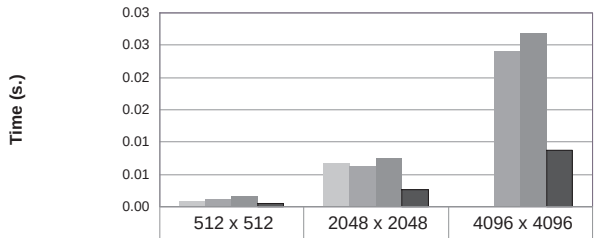
In Figure 7, we compare execution times to obtain the 2D DWT using the four proposed CUDA based algorithms. The four algorithms considered are: the algorithm based on convolution described in Section 1 (labeled *CUDA-Conv 9/7*); the naive aforementioned algorithms described in the SDK, the first one using global memory (labeled as *CUDA-Mem 9/7*) and the second one using texture memory (labeled as *CUDA-*

 ISBN: 978-84-614-6167-7



(a) Shared memory for row filter (b) Shared memory for column filter

Figure 6: Shared Memory for separable filter 9/7



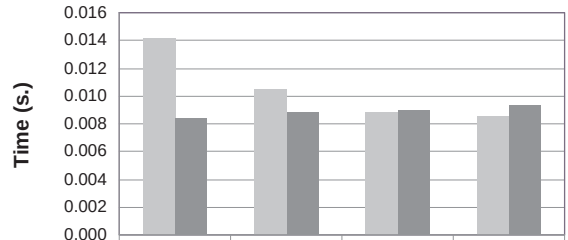
	512 x 512	2048 x 2048	4096 x 4096
CUDA-Conv 9/7	0.0008	0.0068	
CUDA-Mem 9/7	0.0011	0.0063	0.0240
CUDA-Text 9/7	0.0016	0.0077	0.0269
CUDA-Sep 9/7	0.0005	0.0026	0.0088

Figure 7: 2D DWT computation over GPUs with CUDA

Text 9/7); and the algorithm developed to exploit the characteristics of the convolution based on a separable filter (labeled as *CUDA-Sep 9/7*). Daubechies 9/7 filter is used in all experiments performed. In Figure 7 we can observe that the results obtained by the proposed algorithm *CUDA-Conv 9/7* are similar to results obtained by algorithms *CUDA-Mem 9/7* and *CUDA-Text 9/7*, note that the memory requirement of algorithm *CUDA-Conv 9/7* is the lowest one, because the image is overwritten with wavelet coefficients. On the other hand, the best results are obtained using the algorithm *CUDA-Sep 9/7*, note that in this algorithm we optimize the memory coalescence using a separable filter. We want to point out that the speed-up obtained is up to 2.7 for 4096×4096 image size.

In algorithm *CUDA-Sep 9/7* the shared memory stores a block of pixels of one row or a block of one column, the block data stored in shared memory will be computed by a CUDA block. Due to each block of threads computes a block data, the number of threads by block must be selected according the row block size and column block size. Figure 6 shows this structure for both rows and columns. In Figure 8 we present results varying the row block size and column block size for 4096×4096 image size. The best

 ISBN: 978-84-614-6167-7



	4	8	16	32
By Rows	0.0141	0.0105	0.0088	0.0086
By Cols	0.0084	0.0088	0.0089	0.0094

Figure 8: Algorithm *CUDA-Sep 9/7* varying column and row block size

results are obtained with the row block size equal to 16 or 32 and with the column block size equal to 4, 8 or 16. Since shared memory is limited (16 Kbytes in GTX 280), the smaller optimal values of row block size and column block size must be used.

4 Conclusions

We have presented both multicore-based (convolution and lifting) and CUDA-based algorithms (convolution) that perform the two dimensional discrete wavelet transform. We have analyzed the behavior of the proposed algorithms over a shared-memory multiprocessor and a GPU architecture. Furthermore, we have compared our multicore-based proposals against a recent algorithm called SMDWT. The multicore-based algorithms obtain a speed-up above 1.9 using two processors and above 2.4 and up to 3.4 using four processors. Since the best results over a multicore platform have been obtained by the convolution algorithm which also requires a smaller buffer size, we have developed the corresponding GPU-based algorithm using CUDA and implemented the row/column buffer in the GPU shared memory. The speed-up achieved by the GPU-based algorithm is up to 20. We have also compared several CUDA-based algorithms, based on both the proposed multicore-based algorithms and the CUDA SDK proposals. In conclusion, we would like to point out that (1) the use of a multicore platform obtains good performance, and (2) we obtain a good speed-up in a GPU respect to the results obtained in the multicore platform. The CUDA based algorithm to be chosen depends on the parameters to optimize, which can be either the computation time or the memory requirements.

Acknowledgements

This research was supported by the Spanish Ministry of Education and Science under grant DPI2007-66796-C03-03 and the Spanish Ministry of Science and Innovation under grant number TIN2008-06570-C04-04.

References

[1] K. Rao and P. Yip. Discrete cosine transform: Algorithms, advantages, applications. In *Academic Press, USA*, 1990.

[2] ISO/IEC 15444-1. JPEG2000 image coding system, 2000.

[3] A. Said and A. Pearlman. A new, fast and efficient image codec based on set partitioning in hierarchical trees. *IEEE Transactions on Circuits, Systems and Video Technology*, 6(3):243–250, 1996.

[4] S. G. Mallat. A theory for multi-resolution signal decomposition: The wavelet representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(7):674–693, July 1989.

[5] W. Sweldens. The lifting scheme: a custom-design construction of biorthogonal wavelets. *Applied and Computational Harmonic Analysis*, 3(2):186–200, April 1996.

[6] W. Sweldens. The lifting scheme: a construction of second generation wavelets. *SIAM Journal on Mathematical Analysis*, 29(2):511–546, March 1998.

[7] C. Chrysafis and A. Ortega. Line-based, reduced memory, wavelet image compression. *IEEE Transactions on Image Processing*, 9(3):378–389, March 2000.

[8] Y. Bao and C.C. Jay Kuo. Design of wavelet-based image codec in memory-constrined environment. *IEEE Trans. on Circuits and Systems for Video Technology*, 11(5):642–650, May 2001.

[9] Chih-Hsien Hsia, Jing-Ming Guo, Jen-Shiun Chiang, and Chia-Hui Lin. A novel fast algorithm based on smdwt for visual processing applications. In *Circuits and Systems, 2009. ISCAS 2009. IEEE International Symposium on*, pages 762 –765, May 2009.

[10] V. Podlozhnyuk. Image convolution with cuda, June 2007.

[11] I. Daubechies and W. Sweldens. Factoring wavelet transforms into lifting steps. *Fourier Analysis and Applications*, 4(3):247–269, 1998.

[12] J.M. Shapiro. Embedded image coding using zerotrees of wavelet coefficients. *IEEE Transactions on Signal Processing*, 41(12), December 1993.

[13] OpenMP Architecture Review Board. Openmp c and c++ application program interface, version 2.0. March 2002.

[14] J. Nickolls, I. Buck, M. Garland, and K. Skadron. Scalable parallel programming with cuda. In *Queue*, volume 6, pages 40–53, 2008.

[15] NVIDIA Corporation. Nvidia cuda c programming guide. version 3.2.

[16] Ian Buck. Gpu computing with nvidia cuda. In *ACM SIGGRAPH 2007 courses*, SIGGRAPH '07, New York, NY, USA, 2007. ACM.

 ISBN: 978-84-614-6167-7

Proceedings of the 11th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2011
26–30 June 2011.

Extension of the Babuška–Brezzi theory on mixed variational formulations to reflexive spaces

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

The Hilbertian Babuška–Brezzi theory guarantees existence of a unique solution for the mixed variational formulation of a wide class of linear elliptic variational problems, as well as its numerical approximation by the corresponding Galerkin scheme. In this work we generalize the classical theory to the framework of reflexive Banach spaces, providing new examples of application.

Key words: Mixed variational formulations, Babuška–Brezzi theory, discrete mixed formulations, Galerkin methods, linear elliptic boundary value problems.
MSC 2000: 49J40, 65N30, 65M12, 46N40.

1 Motivation

The Babuška–Brezzi theory provides a satisfactory and systematic study of the existence of solution for the mixed variational formulation of many elliptic boundary value problems (see [5, 6, 9, 4]). That kind of variational formulation leads to consider the following abstract problem: suppose that E and F are real Hilbert spaces, $a : E \times E \longrightarrow \mathbb{R}$, $b : E \times F \longrightarrow \mathbb{R}$ and $c : F \times F \longrightarrow \mathbb{R}$ are continuous bilinear forms (in most cases $c = 0$) and $x_0^* : E \longrightarrow \mathbb{R}$ and $y_0^* : F \longrightarrow \mathbb{R}$ continuous linear bilinear functionals. Can we find $(x_0, y_0) \in E \times F$ with

$$\left\{ \begin{array}{ll} x \in E \Rightarrow & x_0^*(x) = a(x_0, x) + b(x, y_0) \\ y \in F \Rightarrow & y_0^*(y) = b(x_0, y) + c(y_0, y) \end{array} \right. .$$

Because of the strongly Hilbertian nature of the Babuška–Brezzi theory, it does not apply to a large class of linear elliptic boundary value problems, those whose function data belong to reflexive Banach spaces (see Example 3). For that very reason, we consider the following generalization of the problem above: given E, F, G and H reflexive Banach spaces, $a : E \times F \longrightarrow \mathbb{R}$, $b : F \times G \longrightarrow \mathbb{R}$, $c : E \times H \longrightarrow \mathbb{R}$ and

 ISBN: 978-84-614-6167-7

$d : G \times H \longrightarrow \mathbb{R}$ continuous bilinear forms and $y_0^* : F \longrightarrow \mathbb{R}$ and $w_0^* : H \longrightarrow \mathbb{R}$ continuous linear functionals, is there exists $(x_0, z_0) \in E \times G$ such that

$$\left\{ \begin{array}{rrl} y \in F \Rightarrow & y_0^*(y) = & a(x_0, y) + b(y, z_0) \\ w \in H \Rightarrow & w_0^*(w) = & c(x_0, w) + d(z_0, w) \end{array} \right. ? \tag{1}$$

2 A reflexive Babuška–Brezzi's theory

As a consequence of the non Hilbert Lax–Milgram's type result appeared in [10], we characterize the existence of a solution of the mixed variational problem (1). First a bit of notation. We write E^* for the dual space of a Banach space E and given real vector spaces E and F, a bilinear form $a : E \times F \longrightarrow \mathbb{R}$ and $(x_0, y_0) \in E \times F$, $a(\cdot, y_0)$ denotes the linear functional on E

$$x \in E \longmapsto a(x, y_0) \in \mathbb{R}$$

and $a(x_0, \cdot)$ is the analogous linear functional on F.

Theorem 1. *Assume that E, F, G and H are real reflexive Banach spaces, $y_0^* \in F^*$, $w_0^* \in H^*$ and that $a : E \times F \longrightarrow \mathbb{R}$, $b : F \times G \longrightarrow \mathbb{R}$, $c : E \times H \longrightarrow \mathbb{R}$ and $d : G \times H \longrightarrow \mathbb{R}$ are continuous bilinear forms. Then there exists $(x_0, z_0) \in E \times G$ solution of the mixed variational problem (1) if, and only if, there exists $\rho \geq 0$ such that the inequality*

$$y_0^*(y) + w_0^*(w) \leq \rho(\|a(\cdot, y) + c(\cdot, w)\| + \|b(y, \cdot) + d(\cdot, w)\|)$$

holds for any $(y, w) \in F \times H$. Moreover, if these equivalent statements are satisfied and there exists $(y, w) \in F \times H$ with $\|a(\cdot, y) + c(\cdot, w)\| + \|b(y, \cdot) + d(\cdot, w)\| \neq 0$, then

$$\min\{\max\{\|x_0\|, \|z_0\|\} : (x_0, z_0) \in E \times G \text{ is a solution of problem (1)}\}$$

$$= \sup_{(y,w) \in F \times H,\ \|a(\cdot,y)+c(\cdot,w)\|+\|b(y,\cdot)+d(\cdot,w)\| \neq 0} \frac{y_0^*(y) + w_0^*(w)}{\|a(\cdot, y) + c(\cdot, w)\| + \|b(y, \cdot) + d(\cdot, w)\|}.$$

Although this control of the norm of the solution seems to give little explicit information, it is the key step in order to derive an estimation of that norm in Theorem 2 below.

Let us recall that a closed subspace K of a Banach space F is said to be *complemented* if there exists a closed subspace M of F such that each $y \in F$ admits a unique representation $y = k + m$, with $k \in K$ and $m \in M$. Equivalently, there exists a continuous linear operator $P : F \longrightarrow F$ such that $P^2 = P$, $P(F) = K$ and $(I - P)(F) = M$. We also say that K is *complemented in F by means of P*. Each closed subspace of a Hilbert space is complemented and for any Banach space any finite–dimensional subspace is complemented, as well as each cofinite–dimensional one.

Now we show how the main classical results on the mixed variational formulation of an elliptic boundary value problem, in the setting of Hilbert spaces, follow from

 ISBN: 978-84-614-6167-7

Theorem 1. Moreover, in Theorem 2 below we state a global version of Theorem 1 (deduced from it), in the sense that we obtain necessary and sufficient conditions, not only for some fixed $y_0^* \in F^*$ and $w_0^* \in H^*$, but also for all continuous and linear functionals on F and H, in order that the mixed variational problem admits a solution. In addition we introduce the ingredient of uniqueness. For the sake of exposition, and taking into account that the most common case is $d = 0$, we restrict ourselves to that concrete situation, firstly stated in [11]. Let us note that even for that specific case, the main result in the Babuška–Brezzi theory for Hilbert spaces ($E = F$ and $G = H$ Hilbert spaces, a coercive, $b = c$ and $d = 0$, see [9, Theorem 10.1] or [4, §1 Theorem 5.2]) is a straightforward consequence of the next result, since K_b is complemented in F by means of its orthogonal projection P, and $\|I - P\| = 1$.

Theorem 2. *If E, F, G and H are real reflexive Banach spaces and $a : E \times F \longrightarrow \mathbb{R}$, $b : F \times G \longrightarrow \mathbb{R}$ and $c : E \times H \longrightarrow \mathbb{R}$ are continuous bilinear forms, we write*

$$K_b := \{y \in F : b(y, \cdot) = 0\}$$

and

$$K_c := \{x \in E : c(x, \cdot) = 0\}$$

and in addition we suppose that K_b is complemented in F by means of a projection P, then, for all $y_0^* \in F^*$, $w_0^* \in H^*$ there exists a unique $(x_0, z_0) \in E \times G$ solution of problem (1) with $d = 0$ if, and only if,

$$x \in K_c \text{ and } a(x, \cdot)_{|K_b} = 0 \Rightarrow x = 0$$

and there exist $\lambda,\ \beta,\ \gamma > 0$ such that

$$y \in K_b \Rightarrow \lambda\|y\| \leq \|a(\cdot, y)_{|K_c}\|,$$

$$z \in G \Rightarrow \beta\|z\| \leq \|b(\cdot, z)\|$$

and

$$w \in H \Rightarrow \gamma\|w\| \leq \|c(\cdot, w)\|.$$

Furthermore, if one of these equivalent conditions is satisfied, then

$$\|x_0\| \leq \frac{\|y_0^*\|}{\lambda} + \frac{1}{\gamma}\left(1 + \frac{\|a\|}{\lambda}\right)\|w_0^*\|$$

and

$$\|z_0\| \leq \frac{\|I - P\|}{\beta}\left(1 + \frac{\|a\|}{\lambda}\right)\left(\|y_0^*\| + \frac{\|a\|}{\gamma}\|w_0^*\|\right).$$

Example 3. As a model problem, consider $\Omega := (0, 1)$, $h \in L^p(\Omega)$ with $1 < p < \infty$ and the Poisson problem with homogeneous Dirichlet boundary conditions

$$\begin{cases} -z'' = h & \text{in } \Omega \\ z(0) = z(1) = 0 & \end{cases},$$

 ISBN: 978-84-614-6167-7

which is equivalently given by

$$\left\{\begin{array}{rcl} x = & z' & \text{in } \Omega \\ -x' = & h & \text{in } \Omega \\ z(0) = z(1) = & 0 & \end{array}\right. .$$

Now, as in the Babuška–Brezzi's treatment, which is not possible here unless $p = 2$, if $E := W^{1,p}(\Omega)$, $F := W^{1,q}(\Omega)$, $G := L^p(\Omega)$ and $H := L^q(\Omega)$, $a : E \times F \longrightarrow \mathbb{R}$, $b : F \times G \longrightarrow \mathbb{R}$ and $c : E \times H \longrightarrow \mathbb{R}$ are the continuous bilinear forms given for all $x \in E,\ y \in F,\ z \in G$ and $w \in H$ by

$$a(x, y) := \int_\Omega xy, \quad b(y, z) := \int_\Omega y'z, \quad c(x, w) := \int_\Omega x'w,$$

and $y_0^* : F \longrightarrow \mathbb{R}$ and $w_0^* : H \longrightarrow \mathbb{R}$ are the continuous linear functionals defined for each $(y, w) \in F \times H$ as

$$y_0^*(y) := 0, \quad w_0^*(w) := -\int_\Omega hw,$$

then we arrive at the following mixed variational problem: find $(x_0, z_0) \in E \times G$ solution of problem (1). It is not difficult to check the assumptions in Theorem 2 (K_b and K_c are one–dimensional subspace, and make use of the description of the dual spaces of the preceding integrable and Sobolev spaces (see for instance [1])) and deduce that

$$\max\{\|x_0\|_{W^{1,p}(\Omega)}, \|z_0\|_{L^p(\Omega)}\} \leq 4\|h\|_{L^p(\Omega)}.$$

□

It is possible to consider the corresponding discrete problem, obtain its Galerkin scheme and derive some stability conditions that generalize the classical results in the framework of Hilbert spaces (see [9, Theorem 10.4] or [4, §1 Theorem 5.3]). To this is end, we note that Fortin's lemma ([7]) also holds for reflexive spaces. A process for generating finite–dimensional subspaces is to fix Schauder bases in the respective reflexive spaces. This kind of biorthogonal system has been successfully used in the numerical treatment of integral, integro–differential or differential equations (see [2, 3, 8]).

Acknowledgements

Research partially supported by Junta de Andalucía Grant FQM 359.

References

[1] R.A. Adams and J.J.F. Fournier, *Sobolev spaces*, 2nd Edition, Elsevier, 2003.

[2] M.I. Berenguer, D. Gámez, A.I. Garralda–Guillem, M. Ruiz Galán and M.C. Serrano Pérez, *Biorthogonal systems for solving Volterra integral equation systems of the second kind*, J. Comput. Appl. Math. **235** (2011), 1875–1883.

[3] M.I. Berenguer, A.I. Garralda–Guillem and M. Ruiz Galán, *Biorthogonal systems approximating the solution of the nonlinear Volterra integro–differential equation*, Fixed Point Theory Appl., Volume **2010**, Article ID 470149, (2010), 9 pp.

[4] D. Boffi et al., *Mixed finite elements, compatibility conditions and applications*, Lecture Notes in Mathematics **1939**, Springer–Verlag, Berlin, 2008.

[5] F. Brezzi, *On the existence, uniqueness and approximation of saddle–point problems arising from Lagrangian multipliers*, RAIRO Model. Math. Anal. Numer. **21** (1974), 129–151.

[6] F. Brezzi and M. Fortin, *Mixed and hybrid finite element methods*, Springer Series in Computational Mathematics **15**, Springer–Verlag, New York, 1991.

[7] M. Fortin, *An analysis of the convergence of mixed finite element methods*, RAIRO Anal. Numér. **11** (1977), 341–354.

[8] D. Gámez, A.I. Garralda Guillem and M. Ruiz Galán, *High order nonlinear initial–value problems countably determined*, J. Comput. Appl. Math. **228** (2009), 77–82.

[9] J.E. Roberts and J.M. Thomas, *Mixed and hybrid methods*, in: *Handbook of numerical analysis, Vol. II* (P.G. Ciarlet and J.L. Lions, eds.), Finite Element Methods (Part 1), North Holland, 1989.

[10] M. Ruiz Galán, *A version of the Lax–Milgram Theorem for locally convex spaces*, J. Convex Anal. **16** (2009), 993–1002.

[11] M. Ruiz Galán, *Generalizad mixed variational formulations and their Galerkin schemes*, submitted for publication.

 ISBN: 978-84-614-6167-7

Proceedings of the 11th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2011
26–30 June 2011.

A note on the dynamic analysis using Generalized Finite Difference Method.

L. Gavete[1], F. Ureña[2], J.J. Benito[3], E. Salete[3] and María Lucía Gavete[4]

[1] *Departamento de Matemática Aplicada a los Recursos Naturales, Universidad Politécnica de Madrid*

[2] *Departamento de Matemática Aplicada, Universidad de Castilla-La Mancha*

[3] *Departamento de Construcción y Fabricación, Universidad Nacional de Educación a Distancia*

[4] *Departamento Bioquímica, Fisiología y Genética Molecular, Universidad Rey Juan Carlos, Madrid, Spain*

emails: lu.gavete@upm.es, francisco.urena@uclm.es, jbenito@ind.uned.es, esalete@ind.uned.es, lucia.gavete@urjc.es

Abstract

This paper shows the application of generalized finite difference method (GFDM) to the problem of dynamic analysis of plates. The stability condition for a fully explicit algorithm is given.

Key words: meshless methods, generalized finite difference method, moving least squares, plates, stability.
MSC 2000: 65M06, 65M12, 74S20, 80M20

1 Introduction

The Generalized finite difference method (GFDM) is evolved from classical finite difference method (FDM). Benito, Ureña and Gavete have made interesting contributions to the development of this method [1, 5, 6, 7, 8].
This paper decribes how the GFDM can be applied for solving dynamic analysis problems of plates [3, 4].
The paper is organized as follows. Section 1 is the introduction. Section 2 describes the

 ISBN: 978-84-614-6167-7

explicit generalized finite difference schemes. In section 3 is studied the consistency and the von Neumann stability. In Section 4 some applications of the GFDM for solving problems of dynamic analysis are included. Finally, in Section 5 some conclusions are given.

2 Explicit generalized finite difference schemes

Let us consider the problems governed by the following partial differential equations (pde)

$$\frac{\partial^2 U(\{x\},t)}{\partial t^2} + A^2\Delta^2 U(\{x\},t) = F_1(\{x\},t) \quad \{x\} \in \Omega \quad (\Omega = (0,L)\, or\, (0,L)\times(0,L), \quad t > 0 \tag{1}$$

with boundary conditions at the ends of the beam of length L or at the edges of plate $[0,L]\times[0,L]$ for each particular case and initial conditions

$$U(\{x\},0) = 0; \quad \frac{\partial U(\{x\},t)}{\partial t}|_{(\{x\},0)} = F_2(\{x\}) \tag{2}$$

where F_1 and F_2 are two known smooth functions, the constant A depends of the material and geometry of the beam.
Firstly, we use the explicit difference formulae for the values of partial derivatives in the space variable. The intention is to obtain explicit linear expressions for the approximation of partial derivatives in the points of the domain.
First of all, an irregular grid or cloud of points is generated in the domain. On defining the composition central node with a set of N points surrounding it (henceforth referred as nodes), the star then refers to the group of established nodes in relation to a central node. Each node in the domain have an associated star assigned [1, 5, 6, 7].
If u_0 is an approximation of fourth-order for the value of the function at the central node (U_0) of the star, with coordinate $\{x_0\}$ and u_j is an approximation of fourth-order for the value of the function at the rest of nodes, of coordinates $\{x_j\}$ with $j = 1, \cdots, N$, then, according to the Taylor series expansion in 1-D and 2-D respectively

$$U_j = U_0 + h_j\frac{\partial U_0}{\partial x} + \frac{h_j^2}{2}\frac{\partial^2 U_0}{\partial x^2} + \frac{h_j^3}{6}\frac{\partial^3 U_0}{\partial x^3} + \frac{h_j^4}{24}\frac{\partial^4 U_0}{\partial x^4} + \cdots \tag{3}$$

$$\begin{aligned} U_j = U_0 &+ h_j\frac{\partial U_0}{\partial x} + k_j\frac{\partial U_0}{\partial y} + \frac{h_j^2}{2}\frac{\partial^2 U_0}{\partial x^2} + \frac{k_j^2}{2}\frac{\partial^2 U_0}{\partial y^2} + h_jk_j\frac{\partial^2 U_0}{\partial x\partial y} + \\ &+ \frac{h_j^3}{6}\frac{\partial^3 U_0}{\partial x^3} + \frac{k_j^3}{6}\frac{\partial^3 U_0}{\partial y^3} + \frac{h_j^2k_j}{2}\frac{\partial^3 U_0}{\partial x^2\partial y} + \frac{h_jk_j^2}{2}\frac{\partial^3 U_0}{\partial x\partial y^2} + \frac{h_j^4}{24}\frac{\partial^4 U_0}{\partial x^4} + \frac{k_j^4}{24}\frac{\partial^4 U_0}{\partial y^4} + \\ &+ \frac{h_j^3k_j}{6}\frac{\partial^4 U_0}{\partial x^3\partial y} + \frac{h_j^2k_j^2}{4}\frac{\partial^4 U_0}{\partial x^2\partial y^2} + \frac{h_jk_j^3}{6}\frac{\partial^4 U_0}{\partial x\partial y^3} + \cdots \end{aligned} \tag{4}$$

 ISBN: 978-84-614-6167-7

where $h_j = x_j - x_0$ and $k_j = y_j - y_0$.
If in equations 3 or 4 the terms over fourth order are ignored. It is then possible to define the function $B_4(u)$ in 1-D or $B_{14}(u)$ in 2-D as in [1, 3, 8, 10, 11, 13, 14]

$$B_4(u) = \sum_{j=1}^{N} [(u_0 - u_j + h_j \frac{\partial u_0}{\partial x} + \frac{h_j^2}{2} \frac{\partial^2 u_0}{\partial x^2} + \frac{h_j^3}{6} \frac{\partial^3 u_0}{\partial x^3} + \frac{h_j^4}{24} \frac{\partial^4 u_0}{\partial x^4}) w(h_j)]^2 \tag{5}$$

$$\begin{aligned} B_{14}(u) = \sum_{j=1}^{N} [(u_0 - u_j + h_j \frac{\partial u_0}{\partial x} + k_j \frac{\partial u_0}{\partial y} + \frac{h_j^2}{2} \frac{\partial^2 u_0}{\partial x^2} + \frac{k_j^2}{2} \frac{\partial^2 u_0}{\partial y^2} + h_j k_j \frac{\partial^2 u_0}{\partial x \partial y} + \\ + \frac{h_j^3}{6} \frac{\partial^3 u_0}{\partial x^3} + \frac{k_j^3}{6} \frac{\partial^3 u_0}{\partial y^3} + \frac{h_j^2 k_j}{2} \frac{\partial^3 u_0}{\partial x^2 \partial y} + \frac{h_j k_j^2}{2} \frac{\partial^3 u_0}{\partial x \partial y^2} + \frac{h_j^4}{24} \frac{\partial^4 u_0}{\partial x^4} + \frac{k_j^4}{24} \frac{\partial^4 u_0}{\partial y^4} + \\ + \frac{h_j^3 k_j}{6} \frac{\partial^4 u_0}{\partial x^3 \partial y} + \frac{h_j^2 k_j^2}{4} \frac{\partial^4 u_0}{\partial x^2 \partial y^2} + \frac{h_j k_j^3}{6} \frac{\partial^4 u_0}{\partial x \partial y^3}) w(h_j, k_j)]^2 \end{aligned} \tag{6}$$

where $w(h_j)$ and $w(h_j, k_j)$ are the denominated weighting function in 1-D or 2-D respectively.
If the norms 5 or 6 are minimized with respect to the partial derivatives the following linear equation systems are obtained

$$\boldsymbol{A_4 D_{u_4}} = \boldsymbol{b_4} \tag{7}$$

$$\boldsymbol{A_{14} D_{u_{14}}} = \boldsymbol{b_{14}} \tag{8}$$

where The matrices $\boldsymbol{A_4}$ and $\boldsymbol{A_{14}}$ are of 5×5 and 14×14, respectively, and the vectors

$$\boldsymbol{D_{u_4}} = \left\{ \begin{array}{cccc} \frac{\partial u_0}{\partial x} & \frac{\partial^2 u_0}{\partial x^2} & \frac{\partial^3 u_0}{\partial x^3} & \frac{\partial^4 u_0}{\partial x^4} \end{array} \right\}^T \tag{9}$$

$$\boldsymbol{D_{u_{14}}} = \left\{ \begin{array}{ccccccccc} \frac{\partial u_0}{\partial x} & \frac{\partial u_0}{\partial y} & \frac{\partial^2 u_0}{\partial x^2} & \frac{\partial^2 u_0}{\partial y^2} & \frac{\partial^2 u_0}{\partial x \partial y} & \frac{\partial^3 u_0}{\partial x^3} & \cdots & \frac{\partial^4 u_0}{\partial x \partial y^3} & \frac{\partial^4 u_0}{\partial x^2 \partial y^2} \end{array} \right\}^T \tag{10}$$

On solving systems 7 and 8, the following explicit difference formulae are obtained as in [1, 5, 6, 7]. On including the explicit expressions for the values of the partial derivatives the star equation is obtained

$$\Delta^2 U(\{x\}, t)|_{(\{x_0\}, n\triangle t)} = m_0 u_0 + \sum_{j=1}^{N} m_j u_j \tag{11}$$

with

$$m_0 + \sum_{j=1}^{N} m_j = 0 \tag{12}$$

 ISBN: 978-84-614-6167-7

Secondly, we shall use an explicit formula for the part of the equation 1 that depends on time. This explicit formula can be used to solve the Cauchy initial value problem. This method involves only one grid point at the advanced time level. The second derivative with respect to time is approached by

$$\frac{\partial^2 U}{\partial t^2}|_{(\{x_0\},n\triangle t)} = \frac{u_0^{n+1} - 2u_0^n + u_0^{n-1}}{(\triangle t)^2} \tag{13}$$

If the equations 11 and 13 are substituted in equation 1 the following recursive relationship is obtained

$$u_0^{n+1} = 2u_0^n - u_0^{n-1} - A^2(\triangle t)^2[m_0 u_0^n + \sum_{j=1}^{N} m_j u_j^n] + F_1(\{x_0\}, n\triangle t) \tag{14}$$

The first derivative with respect to the time is approached by the central difference formula

$$\frac{\partial U}{\partial t}|_{(\{x_0\},0)} = \frac{u_0^1 - u_0^{-1}}{2\triangle t} = F_2(\{x_0\}) \Rightarrow u_0^{-1} = u_0^1 - 2\triangle t F_2(\{x_0\}) \tag{15}$$

If equation 15 is substituted in equation 14 and taking into account initials conditions (2), the following equation is obtained

$$u_0^1 = \triangle t F_2(\{x_0\}) + \frac{F_1(\{x_0\}, 0)}{2} \tag{16}$$

The equation 16 relates the value of the function at the central node of the star, at time $n = 1$, with the values $F_1(\{x_0\}, 0)$ and the initial conditions $F_2(\{x_0\})$.

3 Convergence

According to Lax's equivalence theorem, if the consistency condition is satisfied, stability is the necessary and sufficient condition for convergence. In this section we study firstly the truncation error of the equations 12, and secondly consistency and stability.

3.1 Truncation error

As it is well known, the truncation errors for second order time derivative (TEt) is given as follows:

$$\frac{\partial^2 U(\boldsymbol{x}, t)}{\partial t^2} = \frac{u_0^{t+\triangle t} - 2u_0^t + u_0^{t-\triangle t}}{(\triangle t)^2} - \frac{(\triangle t)^2}{12}\frac{\partial^4 U(\boldsymbol{x}, t_1)}{\partial t^4} + \Theta((\triangle t)^4), t < t_1 < t + \triangle t \tag{17}$$

 ISBN: 978-84-614-6167-7

$$(TE_t) = -\frac{(\triangle t)^2}{12}\frac{\partial^4 U(\boldsymbol{x}, t_1)}{\partial t^4} + \Theta((\triangle t)^4), t < t_1 < t + \triangle t \tag{18}$$

In order to obtain the truncation error for space derivatives, Taylor's series expansion including higher order derivatives is used and then higher order functions $B_p^*[u]$, $p = 4, 14$ are obtained. The expressions of $B_p^*[u]$, $p = 4, 14$ are similar to the ones given in Eq. 5 and Eq. 6, but incorporating now higher order derivatives. If the new norms $B_p^*[u]$, $p = 4, 14$ are minimized with respect to the partial derivatives until the fourth order, the following linear equation systems are obtained:

$$\boldsymbol{A_p D_{u_p}} = \boldsymbol{b_p^*} \tag{19}$$

where $\boldsymbol{A_p}$, $\boldsymbol{D_{u_p}}$ and $\boldsymbol{b_p}$ with $(p = 4, 14)$ are as previously calculated in previous section and $\boldsymbol{b_p^*}$ can be split in two parts as follows

$$\boldsymbol{b_p^*} = \boldsymbol{b_p} + \boldsymbol{b_p^{**}} \tag{20}$$

where the news terms $\boldsymbol{b_p^{**}}$ correspond to the new higher order derivatives incorporated in the Taylor's series expansion to extend the functions from $B_p[u]$, $p = 4, 14$ to $B_p^*[u]$, $p = 4, 14$. Then a better approximation of the partial derivatives can be obtained using the inverse matrix $\boldsymbol{A_p^{-1}}$

$$\boldsymbol{D_{u_p}} = \boldsymbol{A_p^{-1} b_p} + \boldsymbol{A_p^{-1} b_p^{**}} \tag{21}$$

In the Eq. 21 the expression $\boldsymbol{A_p^{-1} b_p}$ is the approximation used in the GFDM (see [10, 14]) and then the truncation errors for spatial derivatives are given by

$$TE_{x_p} = \boldsymbol{A_p^{-1} b_p^{**}} \tag{22}$$

We develop only the truncation error corresponding to $p = 14$ case. The other truncation error for $p = 4$ case can be obtained in a similar way that the one used in $p = 14$ case.

$$\begin{aligned} B^*(u) = \sum_{j=1}^{N} [(u_0 - u_j + h_j\frac{\partial u_0}{\partial x} + k_j\frac{\partial u_0}{\partial y} + \frac{1}{2}(h_j\frac{\partial u_0}{\partial x} + k_j\frac{\partial u_0}{\partial y})^{2)} \\ + \frac{1}{3!}(h_j\frac{\partial u_0}{\partial x} + k_j\frac{\partial u_0}{\partial y})^{3)} + \frac{1}{4!}(h_j\frac{\partial u_0}{\partial x} + k_j\frac{\partial u_0}{\partial y})^{4)} \\ + \frac{1}{5!}(h_j\frac{\partial u_0}{\partial x} + k_j\frac{\partial u_0}{\partial y})^{5)} + \frac{1}{6!}(h_j\frac{\partial u_0}{\partial x} + k_j\frac{\partial u_0}{\partial y})^{6)} + \cdots) w(h_j, k_j)]^2 \end{aligned} \tag{23}$$

If the Eq. 23 is minimized with respect partial derivatives up to the fourth order, the following linear equations system is defined

$$\boldsymbol{A D_u} = \left(\sum_{j=1}^{N} \Xi h_j \quad \sum_{j=1}^{N} \Xi k_j \quad \sum_{j=1}^{N} \Xi \frac{h_j^2}{2!} \quad \cdots \sum_{j=1}^{N} \Xi \frac{h_j^3}{3!} \quad \sum_{j=1}^{N} \Xi \frac{h_j^4}{4!} \quad \cdots \sum_{j=1}^{N} \Xi \frac{h_j^2 k_j^2}{4} \right)^T \tag{24}$$

 ISBN: 978-84-614-6167-7

where

$$\Xi = [-U_0 + U_j - \frac{1}{5!}(h_j\frac{\partial u_0}{\partial x} + k_j\frac{\partial u_0}{\partial y})^{5)} - \frac{1}{6!}(h_j\frac{\partial u_0}{\partial x} + k_j\frac{\partial u_0}{\partial y})^{6)} - \cdots)]w^2(h_j, k_j) \tag{25}$$

with $N \geq 14$, and then

$$TE_{(x,y)} = -A_2^2\boldsymbol{A^{-1}} \times \left(\sum_{j=1}^{N}\Upsilon h_j \quad \sum_{j=1}^{N}\Upsilon k_j \quad \sum_{j=1}^{N}\Upsilon\frac{h_j^2}{2!} \quad \ldots\sum_{j=1}^{N}\Upsilon\frac{h_j^3}{3!} \quad \sum_{j=1}^{N}\Upsilon\frac{h_j^4}{4!} \quad \ldots\sum_{j=1}^{N}\Upsilon\frac{h_j^2k_j^2}{4} \right)^T \tag{26}$$

where

$$\Upsilon = -[\frac{1}{5!}(h_j\frac{\partial u_0}{\partial x} + k_j\frac{\partial u_0}{\partial y})^{5)} + \frac{1}{6!}(h_j\frac{\partial u_0}{\partial x} + k_j\frac{\partial u_0}{\partial y})^{6)} + \cdots)]w^2(h_j, k_j) \tag{27}$$

and operating

$$TE_{(x,y)} = A_2^2[\sum_{j=1}^{N}\Psi_{1,j}\frac{\partial^5 U}{\partial x^5} + \cdots + \Psi_{i,j}\frac{\partial^6 U}{\partial x^6} + ...] + \Theta(h_j, k_j) \tag{28}$$

where $\Psi_{i,j}(h_j)$ are homogeneous rational functions of order two and $\Theta(h_j, k_j)$ is a series of third- and higher-order functions.
The Eq. 28 is the truncation error for spatial derivatives.
Taking into account that the total truncation errors (TTE) is given by

$$TTE = TE_t + TE_{(x,y)} \tag{29}$$

where TE_t and $TE_{(x,y)}$ are given by Eqs. 18 and 28 respectively.

3.2 Consistency

By considering bounded derivatives in Eq. 29

$$\lim_{(\triangle t, h_j, k_j)\to(0,0)} TTE \to 0 \tag{30}$$

Then, the truncation error condition given in Eq. 30 shows the consistency of the approximation.

3.3 Stability criterion

For the difference schemes, the von Neumann condition is sufficient as well as necessary for stability [2]. "Boundary conditions are neglected by the von Neumann method which applies in theory only to pure initial value problems with periodic initial data. It does however provide necessary conditions for stability of constant coefficient problems regardless of the type of boundary condition".

 ISBN: 978-84-614-6167-7

For the stability analysis the first idea is to make a harmonic decomposition of the approximated solution at grid points and at a given time level n. Then we can write the finite difference approximation in the nodes of the star at time n, as

$$u_0^n = \xi^n e^{i\boldsymbol{\nu}^T \boldsymbol{x_0}}; \quad u_j^n = \xi^n e^{i\boldsymbol{\nu}^T \boldsymbol{x_j}} \tag{31}$$

where ξ is the amplification factor,

$$\boldsymbol{x_j} = \boldsymbol{x_0} + \boldsymbol{h_j}; \qquad \xi = e^{-iw\triangle t}$$

$\boldsymbol{\nu}$ is the column vector of the wave numbers

$$\boldsymbol{\nu} = \left\{ \begin{array}{c} \nu_x \\ \nu_y \end{array} \right\}$$

then we can write the stability condition as: $\|\xi\| \leq 1$.
Including the equation 31 into the equation 14, cancelation of $\xi^n e^{i\boldsymbol{\nu}^T \boldsymbol{x_0}}$, leads to

$$\xi = 2 + \frac{1}{\xi} - (\triangle t)^2 A^2 (m_0 + \sum_1^N m_j e^{i\boldsymbol{\nu}^T \boldsymbol{h_j}}) \tag{32}$$

Using the equations 12 and after some calculus we obtain the quadratic equation

$$\xi^2 - \xi[2 + A^2(\triangle t)^2 (\sum_1^N m_j (1 - \cos \boldsymbol{\nu}^T \boldsymbol{h_j}) - i \sum_1^N m_j \sin \boldsymbol{\nu}^T \boldsymbol{h_j})] + 1 = 0 \tag{33}$$

Hence the values of ξ are

$$\xi = b \pm \sqrt{b^2 - 1} \tag{34}$$

where

$$b = 1 + \frac{A^2(\triangle t)^2}{2} \sum_1^N m_j (1 - \cos \boldsymbol{\nu}^T \boldsymbol{h_j}) - i \frac{A^2(\triangle t)^2}{2} \sum_1^N m_j \sin \boldsymbol{\nu}^T \boldsymbol{h_j} \tag{35}$$

If we consider now the condition for stability, we obtain

$$\|b \pm \sqrt{b^2 - 1}\| \leq 1 \tag{36}$$

Operating with the Eqs. 35 and 36, canceling with conservative criteria, the condition for stability of star is obtained as

$$\triangle t \leq \frac{1}{4A\sqrt{|m_0|}} \tag{37}$$

4 Numerical Results

In this section we present different numerical results.

 ISBN: 978-84-614-6167-7

4.1 Transverse vibrations of a beam with one end fixed and other end free

In this section, the weighting function used is

$$\Omega(h_j) = \frac{1}{(\sqrt{h_j^2})^3} \tag{38}$$

The global exact error can be calculated as

$$Global \quad exact \quad error = \sqrt{\frac{\sum_{i=1}^{NT} e_i^2}{NT}} \tag{39}$$

where NT is the number of nodes in the domain and e_i is the exact error in the node i.
The pde is

$$\frac{\partial^2 U(x,t)}{\partial t^2} + \frac{1}{1.875^4}\frac{\partial^4 U(x,t)}{\partial x^4} = 0 \quad x \in (0,1), \quad t > 0 \tag{40}$$

with boundary conditions

$$\begin{cases} U(0,t) = 0 \\ \frac{\partial U(x,t)}{\partial x}|_{(0,t)} = \frac{\partial^2 U(x,t)}{\partial x^2}|_{(1,t)} = \frac{\partial^3 U(x,t)}{\partial x^3}|_{(1,t)} = 0, \end{cases} \tag{41}$$

and initial conditions

$$U(x,0) = 0;\ \frac{\partial U(x,t)}{\partial t}|_{(x,0)} = \cos(1.875x) - \cosh(1.875x) - 0.7340327[\sin(1.875x) - \sinh(1.875x)] \tag{42}$$

The exact solution is given by

$$U(x,t) = (\cos(1.875x) - \cosh(1.875x) - 0.7340327[\sin(1.875x) - \sinh(1.875x)])\sin t \tag{43}$$

Figure 1 shows the approximated solution of the equation 40, 41 and 42 in the last time step ($n = 1000$) with $\triangle t = 0.001$.

4.2 Free vibrations of a simply supported plate

In this section, the weighting function used is

$$\Omega(h_j, k_j) = \frac{1}{(\sqrt{h_j^2 + k_j^2})^3} \tag{44}$$

and the global exact error can be calculated by 39
The pde is

 ISBN: 978-84-614-6167-7

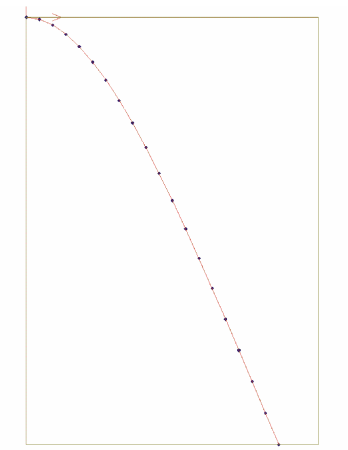


Figure 1: Approximated solution in last time step

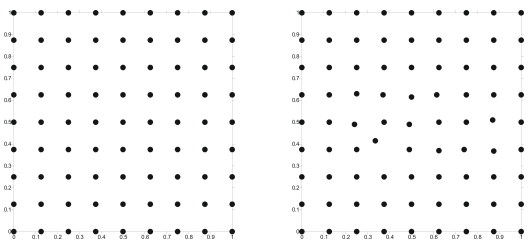


Figure 2: Regular and irregular mesh

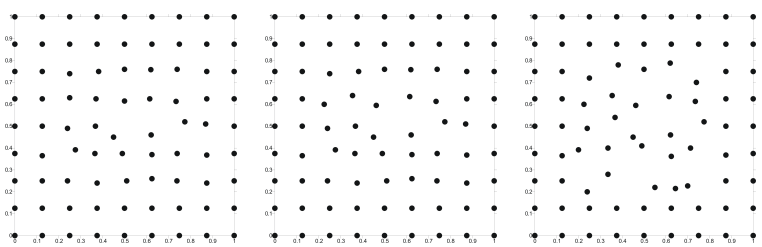


Figure 3: Three irregular meshes

 ISBN: 978-84-614-6167-7

$$\frac{\partial^2 U(x,y,t)}{\partial t^2} + \frac{1}{4\pi^4}[\frac{\partial^4 U(x,y,t)}{\partial x^4} + 2\frac{\partial^4 U(x,y,t)}{\partial x^2 \partial^2} + \frac{\partial^4 U(x,y,t)}{\partial y^4}] = 15 \sin t \sin(2\pi x)\sin(2\pi y) \quad (x,y) \in (0,1)\times(0,1), \quad t > 0 \quad (45)$$

with boundary conditions

$$\begin{cases} U(x,y,t)|_\Gamma = 0 \\ \frac{\partial^2 U(x,y,t)}{\partial y^2}|_{(0,y,t)} = \frac{\partial^2 U(x,y,t)}{\partial y^2}|_{(1,y,t)} = 0, \forall y \in [0,1] \\ \frac{\partial^2 U(x,y,t)}{\partial x^2}|_{(x,0,t)} = \frac{\partial^2 U(x,y,t)}{\partial x^2}|_{(x,1,t)} = 0, \forall x \in [0,1] \end{cases} \quad (46)$$

where Γ is the boundary of the domain $[0,1]\times[0,1]$, and initial conditions

$$U(x,y,0) = 0; \quad \frac{\partial U(x,y,t)}{\partial t}|_{(x,y,0)} = \sin(\pi x)\sin(\pi y) \quad (47)$$

The exact solution is given by

$$U(x,y,t) = \sin(\pi x)\sin(\pi y)\sin t \quad (48)$$

Table 1 shows the results of the global error, using a regular mesh of 81 nodes (fig. 10), for several values of $\triangle t$.

Table 2 shows the results of global error with $\triangle t = 0.001$ for several irregular meshes of

$\triangle t$	Global error
0.01	0.08254
0.005	0.03513
0.002	0.01339
0.001	0.00212

Table 1: Influence of $\triangle t$ in the global error.

IIC	Global error
0.92	0.00224
0.83	0.00224
0.76	0.00231
0.58	0.00251

Table 2: Influence of irregularity of mesh in the global error.

81 nodes (figures 2 and 3).

Figure 4 shows the approximated solution of the equation 45 in the last time step ($n = 1000$).

As new initial conditions let us assume that due to impact an initial velocity is given to a point ($x = y = 0.5$) of the plate, which give the conditions

$$U(x,y,0) = 0; \quad \begin{cases} \frac{\partial U(x,y,t)}{\partial t}|_{(x,y,0)} = 1 & if \quad x = y = 0.5 \\ \frac{\partial U(x,y,t)}{\partial t}|_{(x,y,0)} = 0 & if \quad (x,y) \neq (0.5, 0.5) \end{cases} \quad (49)$$

The exact solution is given by

$$U(x,y,t) = 2[\sin(\pi x)\sin(\pi y)\sin(t) - \frac{1}{9}\sin(3\pi x)\sin(3\pi y)\sin(9t) + \frac{1}{25}\sin(5\pi x)\sin(5\pi y)\sin(25t) - \cdots] \quad (50)$$

 ISBN: 978-84-614-6167-7

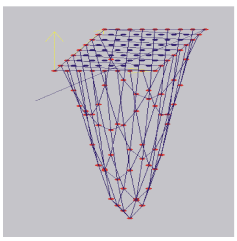


Figure 4: Approximated solution in the last time step

n	Global error
100	0.01122
200	0.01858
600	0.02690
1200	0.03363

Table 3: Variation of global error versus the number of time steps

Table 3 shows the results of the global error, using a regular mesh of 81 nodes (figure 2) and $\triangle t = 0.001$, versus the number of time steps (n).
Figures 5 shows the approximated solution of the equation 45 with the initial conditions 49 in the last time steps for the cases $n = 100$, $n = 200$, $n = 600$ and $n = 1200$ time steps respectively.

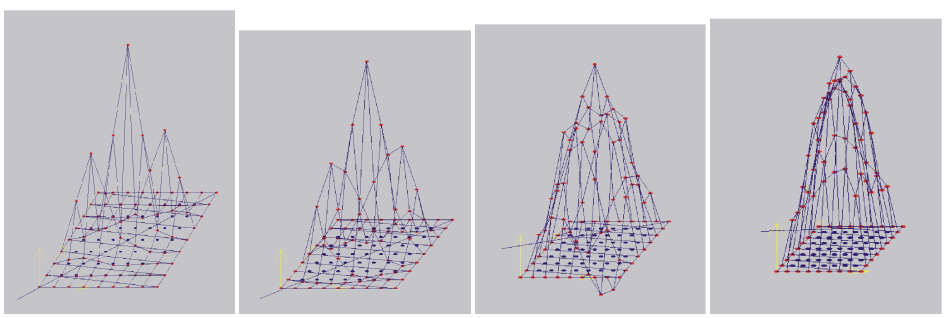


Figure 5: Approximated solution with: n=100, n=200, n=600 and n=1200

 ISBN: 978-84-614-6167-7

4.3 Forced vibrations of a simply supported plate

In this section, the weighting function used is 44 and the global error is calculated by 39 The pde is

$$\frac{\partial^2 U(x,y,t)}{\partial t^2} + \frac{1}{4\pi^4}[\frac{\partial^4 U(x,y,t)}{\partial x^4} + 2\frac{\partial^4 U(x,y,t)}{\partial x^2 \partial^2} + \frac{\partial^4 U(x,y,t)}{\partial y^4}] = 15\sin t \sin(2\pi x)\sin(2\pi y)) \quad (x,y) \in (0,1)\times(0,1), \quad t>0 \quad (51)$$

with boundary conditions

$$\begin{cases} U(x,y,t)|_\Gamma = 0 \\ \frac{\partial^2 U(x,y,t)}{\partial y^2}|_{(0,y,t)} = \frac{\partial^2 U(x,y,t)}{\partial y^2}|_{(1,y,t)} = 0, \forall y \in [0,1] \\ \frac{\partial^2 U(x,y,t)}{\partial x^2}|_{(x,0,t)} = \frac{\partial^2 U(x,y,t)}{\partial x^2}|_{(x,1,t)} = 0, \forall x \in [0,1] \end{cases} \quad (52)$$

and initial conditions

$$U(x,y,0) = 0; \quad \frac{\partial U(x,y,t)}{\partial t}|_{(x,y,0)} = \sin(\pi x)\sin(\pi y) + \sin(2\pi x)\sin(2\pi y) \quad (53)$$

The exact solution is given by

$$U(x,y,t) = (\sin(\pi x)\sin(\pi y) + \sin(2\pi x)\sin(2\pi y))\sin t \quad (54)$$

Table 4 shows the results of the global error, using regular mesh of 81 nodes (figure 2), for several values of $\triangle t$. Table 5 shows the results of global error with $\triangle t = 0.001$ for several

$\triangle t$	Global error
0.01	0.53070
0.005	0.14640
0.002	0.07837
0.001	0.01444

Table 4: Influence of $\triangle t$ in the global error.

IIC	Global error
0.92	0.01412
0.83	0.01437
0.76	0.01442
0.58	0.01447

Table 5: Influence of irregularity of mesh in the global error.

irregular meshes of 81 nodes (figures 2 and 3).
Figure 6 shows the approximated solution of the equation 51 in the last time step ($n = 1000$).

5 Conclusions

The extension of the generalized finite difference to the explicit solution of some dynamic analysis problems has been developed.

 ISBN: 978-84-614-6167-7

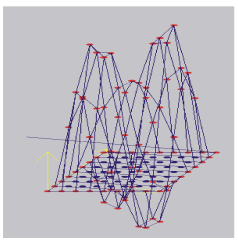


Figure 6: Approximated solution in the last time step

The von Neumann stability criterion has been expressed in function of the coefficients of the star equation for irregular cloud of nodes.
As it is shown in the numerical results, a decrease in the value of the time step, always below the stability limits, leads to a decrease of the global error.

Acknowledgments.

The authors acknowledge the support from Ministerio de Ciencia e Innovación of Spain, project CGL2008 – 01757/CLI.

References

[1] J. J. Benito, F. Ureña and L. Gavete , *Leading-Edge Applied Mathematical Modelling Research (chapter 7)*, Nova Science Publishers, New York, (2008).

[2] A. R. Mitchell, D. F. Griffiths, *The Finite Difference Method in Partial Differential Equations*, John Wiley & Sons, New York, 1980.

[3] W. T.Thomson, *Vibration Theory and Applications*, Prentice Hall Publishers, (1965)

[4] W. Weaver Jr., S. P. Timoshenko, D. H. Young, *Vibration Problems in Engineering*, Jhon Wiley & Sons, Inc, New York (1990)

[5] J. J. Benito, F. Ureña, L. Gavete, B. Alonso, *Solving parabolic and hyperbolic equations by Generalized Finite Difference Method*, Journal of Computational and Applied Mathematics **209 Issue 2** (2007) 208–233.

[6] J. J. Benito, F. Ureña, L. Gavete, B. Alonso, *Application of the Generalized Finite Difference Method to improve the approximated solution of pdes*, Computer Modelling in Engineering & Sciences **38** (2009) 39–58.

 ISBN: 978-84-614-6167-7

[7] 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.

 ISBN: 978-84-614-6167-7

Proceedings of the 11th International Conference on Computational and Mathematical Methods in Science and Engineering, CMMSE 2011 26–30 June 2011.

Special Functions in Engineering: Why and How to Compute Them

Amparo Gil[1], Javier Segura[2] and Nico M. Temme[3]

[1] *Departamento de Matemática Aplicada y CC. de la Computación , Universidad de Cantabria. 39005-Santander, Spain.*

[2] *Departamento de Matemáticas, Estadística y Computación , Universidad de Cantabria. 39005-Santander, Spain.*

[3] *Centrum voor Wiskunde & Informatica (CWI), Science Park 123, 1098 XG Amsterdam, The Netherlands.*

emails: amparo.gil@unican.es, javier.segura@unican.es, nico.temme@cwi.nl

Abstract

Special functions appear in a enormous variety of problems in Engineering. In spite of their importance, there is a lack of validated software for computing important members of this big community of functions, especially for large and/or complex parameters. In this paper we review methods of computation and point out possible failures of routines available in commercial software packages for computing special functions.

Key words: Mathematical functions; Bessel, Legendre and parabolic cylinder functions; stability analysis of recurrence relations; numerical quadrature; asymptotic expansions; computational algorithms; software.

1 Introduction

The so-called special functions of mathematical physics [21] play a key role in many applications in science and technical applications. For example, Bessel or Legendre are familiar names for everyone involved in any area of electromagnetism. This is not surprising because these functions appear in the solution of partial differential equations in cylindrically symmetric domains (such as optical fibers) or in the Fourier transform of radially symmetric functions, to mention just a couple of applications. On the other hand, Legendre functions appear in the solution of electromagnetic problems involving spherical or spheroidal geometries. A couple of papers on the topic are [25] and [4], for example.

 ISBN: 978-84-614-6167-7

But the world of special functions doesn't end in the area of Bessel or Legendre functions, and there are many more functions under the term "special functions", such as cumulative distribution functions [11, Ch. 10], which need to be evaluated in many problems in Engineering. The question is: how to compute these functions? Few engineers have the time (and the will) to develop their own algorithms for computing special functions and most people rely on the black boxes of routines provided by commercial software (like Matlab or Mathematica) to compute them. In some cases, these routines lack of a rigorous testing and validation and fail in providing a uniform accuracy, specially for large parameter cases and/or complex arguments.

In this paper, we review some of the methods for computing special functions and provide some hints about possible failures of routines included in commercial software.

2 Ideas about how to compute special functions

The methods used for the evaluation of special functions are varied, depending on the function under consideration as well as on the efficiency and the accuracy demanded. Usual tools for evaluating special functions are the evaluation of convergent and divergent series, the computation of continued fractions, the use of Chebyshev approximations, the computation of the function using integral representations (numerical quadrature) and the numerical integration of ODEs. Usually, several of these methods are needed in order to build an algorithm able to compute a given function for a large range of values of parameters and argument.

2.1 Some examples: Airy functions, PCFs and toroidal functions

As a first example we consider the Airy functions [20], which are solutions of the differential equation

$$y''(z) - zy(z) = 0\,. \tag{1}$$

Different methods have to be used in order to build an efficient and accurate algorithm for computing the Airy functions in the complex plane ([10],[9]): Maclaurin series for small $|z|$, asymptotic expansions for large $|z|$ and numerical quadrature for intermediate values of $|z|$. Airy functions appear in a large number of applications in physics and engineering (for few examples see, for instance, [24],[18]); these functions are also useful because they can be used as approximants to differential equations with turning point (the Airy equation being the simplest possible equation of this type).

Of course, the degree of complexity of the algorithm for computing a function increases when the function depends on additional parameters, particularly for large values of the parameters. This happens, for instance, in the case of the parabolic cylinder functions $U(a,x)$, $V(a,x)$ [15], where several methods (series, recurrences, quadrature and several types of asymptotic expansions) had to be considered. Something similar happens for the parabolic cylinder functions $W(a,\pm x)$ ([12],[13]). A second example

 ISBN: 978-84-614-6167-7

of the complexity of the analysis is the evaluation of toroidal functions, which are Legendre functions of half-integer degrees: $P^m_{n-1/2}(x)$, $Q^m_{n-1/2}(x)$; for a recent application in telecom engineering see [1]. These functions depend on an argument and two parameters and, in this case, key ingredients in the algorithms [7, 8] were the study of the stability of recurrence relations and the analysis the convergence of the associated continued fraction [7], together with the development of an asymptotic expansion for large orders m [14] (uniformly valid with respect to the argument).

The difficulty in the analysis of the methods in examples like these, could explain the absence of validated software for relevant functions depending on two or more parameters.

These examples illustrate the importance of the asymptotic analysis in the development of efficient methods of computation. Specifically, when the function depends on several parameters, the asymptotic expansions for large values of one or more parameters are powerful tools of computation.

2.2 Some techniques: recurrence relations and numerical quadrature

We briefly describe two important techniques for computing special functions which, together with the evaluation of series (convergent or divergent), appear ubiquitously in algorithms for special function evaluation.

2.2.1 Recurrence relations

In many important cases there exist recurrence relations relating different values of the function for different values of its variables; in particular, one can usually find three term recurrence relations ([11, Ch. 4]). In these cases, the efficient computation of special functions uses at some stage the recurrence relations satisfied by such families of functions. In fact, it is difficult to find a computational task which does not rely on recursive techniques: the great advantage of having recursive relations is that they can be implemented with ease. However, the application of recurrence relations can be risky: each step of a recursive process generates not only its own rounding errors but also accumulates the errors of the previous steps. An important aspect is then the study of the numerical condition of the recurrence relations, depending on the initial values for starting recursion. Let's briefly explain the procedure.

We write the three-term recurrence satisfied by the function y_n as

$$C_n y_{n+1} + B_n y_n + A_n y_{n-1} = 0. \tag{2}$$

If a solution $y_n^{(m)}$ of (2) exists that satisfies

$$\lim_{n\to+\infty} \frac{y_n^{(m)}}{y_n^{(D)}} = 0$$

for all solutions $y_n^{(D)}$ that are linearly independent of $y_n^{(m)}$, then we call $y_n^{(m)}$ the *minimal solution*. The solution $y_n^{(D)}$ is said to be a *dominant solution* of the three-term

 ISBN: 978-84-614-6167-7

recurrence relation. From a computational point of view, the crucial point is the identification of the character of the function to be evaluated (either minimal or dominant) because the stable direction of application of the recurrence relation is different for evaluating the minimal or a dominant solution of (2).

Let N be a (possibly large) positive number. Then we have the following scheme of recursions:

$$\begin{array}{lcll} y_0^{(D)}\,,\, y_1^{(D)} & \longrightarrow & y_N^{(D)} & \text{well conditioned;} \\ y_0^{(m)}\,,\, y_1^{(m)} & \longrightarrow & y_N^{(m)} & \text{ill conditioned;} \\ y_0^{(m)} & \longleftarrow & y_N^{(m)}\,,\, y_{N+1}^{(m)} & \text{well conditioned.} \end{array}$$

For analyzing whether a function is minimal or not, it is needed that analytical information is needed regarding its behavior as $n \to +\infty$; also some numerical experiments can be considered for elucidating this fact (by the method of "see what happens").

2.2.2 Numerical quadrature

Another example where the study of numerical stability is of concern is the computation of special functions via integral representations. It is tempting, but usually wrong, to believe that once an integral representation is given, the computational problem is solved. One has to choose a stable quadrature rule and this choice depends on the integral under consideration. Particularly problematic is the integration of strongly oscillating integrals (Bessel and Airy functions, for instance); in these cases an alternative approach consists in finding non-oscillatory representations by properly deforming the integration path in the complex plane. Let's explain this in a simple example (so simple that it is, in fact, computable in closed form).

Consider the numerical computation of

$$I(\lambda) = \int_0^{+\infty} \cos(2\lambda t) e^{-t^2}\, dt.$$

Needless to say that it is not necessary to evaluate this integral numerically because the exact result is known, but we take this as an illustration on how to build stable representations starting with unstable representations. Straightforward computation of this integral by using any quadrature rule (for instance the trapezoidal rule) is very unstable when λ is large due to the fast oscillations of the integrand.

By symmetry of the integrand, we can write the integral as $I(\lambda) = \frac{1}{2} G(\lambda)$ where

$$G(\lambda) = \int_{-\infty}^{\infty} e^{-t^2+2i\lambda t}\, dt, \quad \lambda > 0.$$

Now, for building a stable integral representation we shift the path of integration upwards in the complex $t-$plane to make it run through the point $t = i\lambda$, or write

$$-t^2 + 2i\lambda t = -(t - i\lambda)^2 - \lambda^2.$$

 ISBN: 978-84-614-6167-7

This gives

$$G(\lambda) = e^{-\lambda^2} \int_{-\infty}^{\infty} e^{-(t-i\lambda)^2}\, dt$$

or, by writing $t = i\lambda + s$,

$$G(\lambda) = e^{-\lambda^2} \int_{-\infty}^{\infty} e^{-s^2}\, ds.$$

In this simple example we deform the original contour of integration to let it run through the saddle point. The resulting integral has no oscillation in the integrand and it is suited for a computation by the trapezoidal rule (which in fact is very efficient for these type of analytic and fast decaying integrands). Of course, in this case it is not really needed that we use a numerical approximation for the integral because it is obvious that

$$G(\lambda) = \sqrt{\pi} e^{-\lambda^2}.$$

This method for building stable integral representations by deforming the contour of integration can be used for many special functions defined by real or contour integrals (see [11, Ch. 5]).

3 A brief guide to what is available for computing special functions

In [19] an overview is given of available software for special functions. We mention a few examples and additions.

3.1 Repositories of software packages

Let's start with two important repositories of software packages that include routines for computing special functions: ACM Algorithms and CPC Programs.

Software associated with papers published in the Transactions on Mathematical Software, as well as other ACM journals are included in CALGO (http://calgo.acm.org/). The software is refereed, among other aspects, for accuracy, robustness and portability. On the other hand, the routines are usually programmed in Fortran (Fortran 77, Fortran 90) or C but there are also routines for Matlab and Maple. There are routines for computing a large number of special functions.

The CPC Program Library (http://cpc.cs.qub.ac.uk) contains a large number of programs in computational physics and chemistry, including several routines for the computation of special functions. The programs are available for subscribers of the journal.

Other sources are software repositories as Netlib, which contains a great number of routines for special functions (and also contains, as a subset, the algorithms published in ACM journals). See [19] for a comprehensive survey.

 ISBN: 978-84-614-6167-7

3.2 Commercial software

The two previous software repositories contain routines (most of them programmed in Fortran or C) which have to be compiled and linked before they can be executed. A very popular alternative to the use of this kind of routines are the commercial interactive systems, such as

- Matlab (http://www.mathworks.com/),
- Maple (http://www.maplesoft.com/), and
- Mathematica (http://www.wolfram.com/).

An interactive system provides a set of commands which the user can enter at the keyboard. The response to each command can be seen immediately on the screen.

Matlab supports Bessel functions of real order and complex argument. The algorithms use a routine by D.E. Amos [2] and there is a warning in the online help system that the functions may produce inaccurate results for large order and argument. Built-in special functions for real arguments and parameters include error and inverse error functions; gamma function; incomplete gamma and beta functions; Bessel functions I, J, K and Y; complete elliptic integrals of first and second kind; Jacobi's elliptic functions; exponential integral and the psi function.

Maple and Mathematica are also examples of general-purpose commercial mathematics software packages. Differently from Matlab, both are examples of Computer Algebra Systems which are software programs that facilitate symbolic mathematics. The core functionality of these packages is the manipulation of mathematical expressions in symbolic form. The symbolic manipulations supported typically include: simplification to the smallest possible expression or some standard form, including automatic simplification with assumptions and simplification with constraints; change of form of expressions: expanding products and powers, rewriting as partial fractions, constraint satisfaction, rewriting trigonometric functions as exponentials, and so on.

4 A few examples of inexact computations with commercial software

Next we give some examples of special function evaluation which give incorrect or misleading results with Matlab, Maple or Mathematica. The examples include complex Bessel functions, prabolic cylinder functions, conical functions and toroidal functions.

4.1 Computation of Bessel functions in Matlab and Mathematica

In [6] du Toit pointed out failures in the routines used in early versions of Matlab (and Mathematica) for computing the Bessel functions of second kind $Y_\nu(z)$ for integer orders and complex arguments. Although these particular failures seem to be corrected in later versions of Matlab and Mathematica, still the routines give inexact answers

 ISBN: 978-84-614-6167-7

	$z = 10i$	
ν	Matlab value (version 7.6.0)	Correct value
57	$2.0909 \times 10^{+34} i$	$-2.6307 \times 10^{-37} + 2.0909 \times 10^{+34} i$
58	$2.4021 \times 10^{+35} - 2.2515 \times 10^{-38} i$	$2.4021 \times 10^{+35} - 2.2515 \times 10^{-38} i$
59	$-2.8073 \times 10^{+36} i$	$1.8947 \times 10^{-39} - 2.8073 \times 10^{+36} i$
60	$-3.3367 \times 10^{+37} + 1.5683 \times 10^{-40} i$	$-3.3367 \times 10^{+37} + 1.5683 \times 10^{-40} i$
	$z = 100i$	
ν	Matlab value (version 7.6.0)	Correct value
57	$-1.3173 \times 10^{+35} + 8.0660 \times 10^{+18} i$	$-1.3173 \times 10^{+35} + 2.0993 \times 10^{-38} i$
58	$3.6210 \times 10^{-38} - 7.6043 \times 10^{+34} i$	$3.6210 \times 10^{-38} - 7.6043 \times 10^{+34} i$
59	$4.3518 \times 10^{+34} - 2.6647 \times 10^{+18} i$	$4.3518 \times 10^{+34} - 6.2996 \times 10^{-38} i$
60	$-1.1055 \times 10^{-37} + 2.4691 \times 10^{+34} i$	$-1.1055 \times 10^{-37} + 2.4691 \times 10^{+34} i$

Table 1: Computation of the Bessel function of the second kind $Y_\nu(z)$, ν integer and z purely imaginary in Matlab (version 7.7.0).

for certain values of argument and order (although there is no warning about this). Table 1 shows an example of computation of $Y_\nu(z)$ using Matlab (version 7.6.0) for two purely imaginary arguments ($z = 10i, 100i$) and four consecutive integer orders ($\nu = 57, 58, 59, 60$). As can be seen, a correct answer is obtained for even values of the order ν (both for the real and imaginary parts of the function) but for odd values of the order, the computed imaginary part of the function seems to be completely wrong.

In addition, the Bessel function $J_\nu(ix)$ has a small real part (relative to the imaginary part) when ν is and odd integer, which is incorrect. For instance, Matlab gives $J_{57}(10i) = 1.610810^{-53} + 2.630710^{-37} i$, but the real part is really zero. The Hankel functions ($H_\nu^{(1)}(z) = J_\nu(z) + iY_\nu(z)$, $H_\nu^{(2)}(z) = J_\nu(z) - iY_\nu(z)$) and for odd integer ν also have problems, and the imaginary part is incorrect; for even ν only the first Hankel function has problems (an small real part appears which should not be there).

For the values of Table 1, it appears that Mathematica does a good job and no problems appear. However, serious problems appear for other parameter ranges. For example, when we try to compute $H_9^{(1)}(18i)$ by introducing the command **HankelH1[9., 18. I]** we get the almost completely wrong answer $-2.57604\,10^{-8} - 1.97906\,10^{-9}$ (the correct number being 2.449910^{-8}). Only by adding 18 zeros after the decimal point we get a correct answer and with less zeros it is totally wrong.

These appear to be examples of errors due to some special properties that the functions satisfy, like the fact that the real part or imaginary parts of $H_\nu^{(1)}(ix)$ vanish when ν is an integer number. However, this problems are carried for other values and also, for example, in Mathematica **HankelH1[9.5, 18. I]** gives an incorrect answer, both for the real and imaginary parts.

When computing special functions one should be aware of some properties the functions satisfy and not take for granted that an algorithm, no matter how powerful it seems, will be necessarily able to compute the function for the whole range of pa-

 ISBN: 978-84-614-6167-7

Digits	Maple 12 value
20	-9.2548×10^{45}
40	2.7764×10^{-54}
60	-3.7019×10^{-153}
80	3.0957×10^{-159}

Table 2: Computation of the parabolic cylinder function $U(-10, 40)$ using Maple 9. The correct answer is 3.0957×10^{-159}.

Digits	Maple 12 value
20	$1.3048 \times 10^{+174}$
40	0
60	1.3048×10^{-25}
80	0
100	6.5241×10^{-225}
200	3.5205×10^{-358}

Table 3: Computation of the parabolic cylinder function $U(50, 50)$ using Maple 9. The correct answer is 3.5205×10^{-358}.

rameters. Caution must be taken, even if our software package gives no warning (as is the case in the previous examples).

4.2 A Maple example: computation of parabolic cylinder functions

Parabolic cylinder functions (PCFs) ineluctably appear in electromagnetic problems in parabolic cylinder geometries (see, for instance, [5] or [3] for a couple of problems involving these functions). Maple allows the computation of the PCFs $U(a, x)$ and $V(a, x)$, solutions of the differential equation

$$y''(x) - \left(\frac{x^2}{4} + a\right) y(x) = 0$$

by means of the commands **CylinderU** and **CylinderV**, respectively. When the parameters a and x are large (or not so large), Maple needs a very large number of digits in order to compute the correct values of these functions. Tables 2 and 3 show examples of the variety of answers given by Maple in the computation of $U(-10, 40)$ and $U(50, 50)$, respectively, when varying the number of digits used by Maple in the calculations.

The problem with these Maple computation seem to be of a different nature to the problems found in the previous section. In this case, the most probable source of error is the method of computation. $U(a, x)$ is a function which decays exponentially for large x (and also for large positive a). In [26, §12.4], $U(a, x)$ is given as a combination of two power series, both of which are exponentially large for large x. $U(a, x)$ is a recessive

 ISBN: 978-84-614-6167-7

τ	Mathematica 6 value
0.1	$1.04237 \times 10^{+155} + 1.90166 \times 10^{+140} i$
1	$7.90237 \times 10^{+155} + 3.97168 \times 10^{+141} i$
10	$3.59416 \times 10^{+161} + 4.33418 \times 10^{+147} i$
30	$-1.64285 \times 10^{+173} - 5.94165 \times 10^{+159} i$

Table 4: Computation of the conical function $P^{100}_{-1/2+i\tau}(100)$ using Mathematica 6. These functions should be real valued.

solution of the differential equation, which is written in terms of two dominant solutions (i.e., two solutions which are much larger for large x). From a numerical point of view, this is an ill conditioned computation; it is not a wise decision to compute a recessive solution from a combination of two dominant solutions, because more and more digits will be needed as the dominant solutions become larger and the recessive solution becomes smaller. Just as for the case of recurrence relations mentioned before, it is crucial to identify recessive solutions. Parabolic cylinder functions also satisfy three-term recurrence relations with respect to the parameter a, and $U(a, x)$ is also recessive (or minimal) with respect to this parameter when a becomes large and positive.

4.3 A Mathematica example: computation of the conical function $P^m_{-1/2+i\tau}(x)$ and toroidal harmonics

The conical functions $P^m_{-1/2+i\tau}(x)$ appear in a large number of applications in electrical engineering (see, for instance, [23], [22]). The only existing (refereed) code for computing these functions (for $m = 0, 1$) is [17], although an algorithm for computing these functions for m variable has been recently proposed in [16].

These functions, which are real valued, can be computed in Mathematica using the command **LegendreP**. Table 4 shows few examples of computed values of the conical function $P^{100}_{-1/2+i\tau}(100)$ with Mathematica 6. As can be seen, Mathematica gives a non-zero imaginary part for these functions. This is not an effect of the large arguments and this imaginary part appears also for lower values, and, for example computing $P^5_{-1/2+i}(5)$ with the command **N[LegendreP[-1/2 + 1 I, 10, 3, 5.], 16]** we get $127387. + 2.39325\,10^{-9} i$. Only after adding around 18 zeros after the digital comma in all the real parameters, we get as output $127387.2890841 + 0.10^{-8} i$. The results are in any case correct when the imaginary part is discarded, but for this we need to be aware of the fact that the function is real. But when a function is real, it is preferable to consider specific algorithms for that function and not to compute the real function in terms of complex functions.

For the case of conical functions, we could suspect that something wrong may happen because it has a complex parameter, though it is real. However, there are cases which are not so apparently problematic. Take for instance the case of toroidal functions $P^m_{n-1/2}(z)$, $Q^m_{n-1/2}(z)$, $z > 1$ and n and m integer numbers. Several definitions for this functions are possible, but, in particular, it is useful to consider real representations

 ISBN: 978-84-614-6167-7

for this functions, for instance by taking the branch cut in the complex plane to be the interval $(-\infty, 1)$. These are the Legendre functions of type 3, available in Mathematica. With this definition, both $P^m_{n-1/2}(z)$ and $Q^m_{n-1/2}(z)$ are real functions. However, when computing $Q^{50}_{51/2}(4.4)$ with the Mathematica command **N[LegendreQ[51/2, 50, 3, 4.4], 16]** (16 digits): $4.28053\,10^{59} + 4.20431\,10^{44}i$; the imaginary part is meaningless; the situation only improves after adding 18 zeros after the decimal point in each of the parameters. Again, the reason for this is that a real function is being computed from complex functions and, either the imaginary part should be dropped or specific algorithms should be considered for this function.

In Maple, this problem does not exist and $Q^m_{n-1/2}(z)$ is real when the branch cut is chosen to be $(-\infty, 1)$. Similar problems, however, appear with Maple when computing $Q^{m-1/2}_{n-1/2}(z)$ with m integer; in this case, $Q^{m-1/2}_{n-1/2}(z)$ is imaginary but a small real part appears. For instance, with 15 digits in Maple 12, we get $Q^{50.5}_{51/2}(4.4) \approx 1.87921\,10^{39} + 3.82844248154908\,10^{60}i$, with a meaningless real part. These cases, however do not appear to be so useful as the toroidal function cases $P^m_{n-1/2}(z)$ and $Q^m_{n-1/2}(z)$.

Acknowledgements

The authors acknowledge financial support from *Ministerio de Ciencia e Innovación*, project MTM2009-11686. NMT acknowledges financial support from *Gobierno of Navarra*, Res. 07/05/2008.

References

[1] I. Al Falujah and V.K. Prabhu. Error performance of dqpsk with egc diversity reception over fading channels. *IEEE Trans. Wireless Commun*, 7(4):1190–1194, 2008.

[2] D. E. Amos. Algorithm 644: A portable package for Bessel functions of a complex argument and nonnegative order. *ACM Trans. Math. Softw.*, 12(3):265–273, 1986.

[3] R. Borghi. Analytical solution for the eigenmodes of closed waveguide resonators with small curvature mirrors. *IEEE Journal of Quantum Electronics*, 36(3):363–365, 2000.

[4] I.R. Capoglu and G.S. Smith. The input admittance of a prolate-spheroidal monopole antenna fed by a magnetic frill. *IEEE Trans. Antennas and Propagation*, 54(2):572–585, 2006.

[5] P. K. Choudhury. On the propagation of electromagnetic waves through parabolic cylindrical chiroguides with small flare angles. *Microwave and Optical Technology Letters*, 33(6):414–419, 2002.

 ISBN: 978-84-614-6167-7

[6] C. F. du Toit. Evaluation of some algorithms and programs for the computation of integer-order Bessel functions of the first and second kind with complex arguments. *IEEE Antennas and Propagation Magazine*, 35(3):19–25, 1993.

[7] A. Gil and J. Segura. Evaluation of toroidal harmonics. *Comput. Phys. Comm.*, 124:104–122, 2000.

[8] A. Gil and J. Segura. DTORH3 2. 0: A new version of a computer program for the evaluation of toroidal harmonics. *Comput. Phys. Commun.*, 139(2):186–191, 2001.

[9] A. Gil, J. Segura, and N. M. Temme. Algorithm 819: AIZ, BIZ: two Fortran 77 routines for the computation of complex Airy functions. *ACM Trans. Math. Softw.*, 28(3):325–336, 2002.

[10] A. Gil, J. Segura, and N. M. Temme. Computing complex Airy functions by numerical quadrature. *Numer. Algorithms*, 30(1):11–23, 2002.

[11] A. Gil, J. Segura, and N. M. Temme. *Numerical methods for special functions*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2007.

[12] A. Gil, J. Segura, and N.M. Temme. Fast and accurate computation of the Weber parabolic cylinder function $W(a, x)$. *IMA J. Numer. Anal, to appear*.

[13] A. Gil, J. Segura, and N.M. Temme. Parabolic cylinder function $W(a, x)$ and its derivative (algorithm). *ACM Trans. Math. Softw., to appear*.

[14] A. Gil, J. Segura, and N.M. Temme. Computing toroidal functions for wide ranges of the parameters. *J. Comput. Phys.*, 161(1):204–217, 2000.

[15] A. Gil, J. Segura, and N.M. Temme. Computing the real parabolic cylinder functions $U(a, x)$, $V(a, x)$. *ACM Trans. Math. Softw.*, 32(1):70–101, 2006.

[16] A. Gil, J. Segura, and N.M. Temme. Computing the conical function $P^m_{-1/2+i\tau}(x)$. *SIAM J. Sci. Comput.*, 31:1716–1741, 2009.

[17] K.S. Kölbig. A program for computing the conical functions of the first kind $P^m_{-1/2+i\tau}(x)$ for $m = 0$ and $m = 1$. *Comput. Phys. Commun.*, 23:51–61, 1981.

[18] M.F. Levy. Horizontal parabolic equation solution of radiowave propagation problems on large domains. *IEEE Trans. Antennas and Propagation*, 43(2):137–144, 1995.

[19] D. W. Lozier and F. W. J. Olver. Numerical evaluation of special functions. In *Mathematics of Computation 1943–1993: a half-century of computational mathematics (Vancouver, BC, 1993)*, volume 48 of *Proc. Sympos. Appl. Math.*, pages 79–125. Amer. Math. Soc., Providence, RI, 1994. Updates are available at http://math.nist.gov/mcsd/Reports/2001/nesf/.

 ISBN: 978-84-614-6167-7

[20] F. W. J. Olver. Airy and related functions. In *NIST handbook of mathematical functions*, pages 193–213. U.S. Dept. Commerce, Washington, DC, 2010.

[21] F.W. Olver, D.W. Lozier, R.F. Boisvert, and C.W. Clark. *NIST Handbook of Mathematical Functions*. Cambridge University Press, 2010. http://dlfm.nist.gov.

[22] A. Passian, R.H. Ritchie, A.L. Lereu, T. Thundat, and T.L. Ferrell. Curvature effects in surface plasmon dispersion and coupling. *Phys. Rev. B*, 71:115425, 2005.

[23] A. Passian, A. Wig, F. Meriaudeau, M. Buncick, T. Thundat, and T.L. Ferrell. Electrostatic force density for a scanned probe above a charged surface. *J. Appl. Phys.*, 90(2):1011–1016, 2001.

[24] M.P. Shatz and G.H. Polychronopoulos. An algorithm for the evaluation of radar propagation in the spherical earth diffraction region. *IEEE Transactions on Antennas and Propagation*, 38(8):1249–1252, 1990.

[25] T Simpson and J Cahill. The electrically small elliptical loop with an oblate spheroidal core. *IEEE Antennas and Propagation Magazine*, 49(5):83–92, 2007.

[26] N. M. Temme. Parabolic cylinder functions. In *NIST handbook of mathematical functions*, pages 303–319. U.S. Dept. Commerce, Washington, DC, 2010.

 ISBN: 978-84-614-6167-7

Proceedings of the 11th International Conference on Computational and Mathematical Methods in Science and Engineering, CMMSE 2011 26–30 June 2011.

Lane mark detection using statistical measures over compressed domain video data

Juan Giralt[1], Luis Rodriguez-Benitez[1], Cayetano Solana-Cipres[1], Juan Moreno-Garcia[2] and Luis Jimenez-Linares[1]

[1] *Escuela Superior de Informatica, Universidad de Castilla-La Mancha, Paseo de la Universidad. 4, 13071 Ciudad Real, Spain*

[2] *Escuela de Ingenieria Industrial, Universidad de Castilla-La Mancha, Avenida Carlos III s/n, 45071 Toledo, Spain*

emails: juan.giralt@uclm.es, luis.rodriguez@uclm.es, cayetanoj.solana@uclm.es, juan.moreno@uclm.es, luis.jimenez@uclm.es

Abstract

An approximation to the detection of lane marks from video sequences is described. Detecting and localizing this kind of marks is relevant in many applications of driving assistance. The novelty of the proposed technique is that it works directly in video compressed domain and no pixel information is needed. The proposed method applies a set of filters and removes progressively blobs obtained from segmentation depending on their position in the scene, their size and their shape. The major contributions are the simplicity of the technique because its operation is based on simple statistical filters and the shape analysis based on the position and the distribution in the scene of macroblocks belonging to segmented regions. Finally, it must be remarked the proposed method shows encouraging results in road traffic video sequences.

Key words: H264/AVC compressed video domain, statistical filtering, driving assistance

1 Introduction

Detecting and localizing lanes from a road image is an important component of many intelligent-transportation-system applications and there has been active research on lane detection in recent years because of its relevance in the prevention of traffic accidents. A wide variety of algorithms of various representations, detection and tracking techniques have been proposed and they are usually divided in two classes: feature-based techniques and model-based techniques. The first class algorithms localizes the lanes by using low-level features like line edges identified with traditional edge-based

 ISBN: 978-84-614-6167-7

segmentation algorithms [1], [8]. These techniques work at pixel level and usually have problems with occlusions and noise. The second class approaches use a few parameters to represent the lines by assuming straight lines [4] or parabolic curves [11]. These algorithms try to calculate the model parameters by using probabilistic or fuzzy inference. The model-based techniques are more robust against noise and missing date and could not need pixel processing, but high level processing. Furthermore, it is important to know the road curvature to distinguish objects on the sidewalk and avoid false alarms.

Some relevant works are described now. Wang et al. [11] propose a lane detection and tracking algorithm based on B-snakes without any camera parameters. This method allows to describe a lane through a wide range of lane structures since B-splines can form any arbitrary shape by a set of control points. Jung and Kelber [3] develop a lane departure warning system based on a linear-parabolic lane model. They divide the road into a near field and a far field; a linear function fits the near vision field and a quadratic function is used to the far field. On the other side, Kim [5] presents a fast and robust lane detection and tracking algorithm. It is based on random-sample consensus and particle filtering to generate hypotheses in real time about the correct lines. After that, a framework combines a likelihood-based object-recognition algorithm with a Markov-style process to track identified lines. A recent approach is proposed by Wang et al. [10]. They combine a self-clustering algorithm, fuzzy C-means and fuzzy rules to process the spatial information and Canny algorithms to get good edge detection. Their lane departure warning uses instantaneous information from the lane detection to calculate angle relations of the boundaries.

This paper presents a technique to identify lane marks in road scenes by using an in-car camera system where input video is coded in H264/AVC. The paper is organized as follows. In Section 2 an introduction to H264/AVC is done. Later, in Section 3 the proposed method is described. In Section 4 an example of the behavior of the method is shown and Section 5 shows the experimental results. Finally, conclusions are described in Section 6.

2 Motion compensation in H.264 Advanced Video Coding

H.264 [2], also known as MPEG-4 Part 10, is a standard for video compression developed jointly between the *Motion Picture Expert Group* (MPEG) and the *Video Coding Experts Group* (VCEG). This standard provides mechanisms for video coding that are optimized for a better compression efficiency and aims to meet the multimedia communication applications. Richardson explains deeply in [6] the features of the H.264 compressed domain, but in this section only the motion compensation in H.264 is described. That is because input data for the proposed method is obtained from motion prediction and for a better comprehension of the shape analysis process the understanding of the macroblock and the motion vector concepts are necessary.

 ISBN: 978-84-614-6167-7

2.1 Macroblocks and motion vectors

The basic unit in which an image is divided into is the **macroblock**. It contains the information of a 16x16 pixels region and there are two types depending on the encoding: **Intra macroblock**, in which Intra-prediction algorithms are applied directly to exploit the spatial redundancy according to the H.264 standard, and **Inter macroblock**, in which motion compensation is used to exploit temporal redundancy from a reference macroblock (earlier, later or a combination of both).

H.264 uses *block-based motion compensation*, the same principle adopted by every major coding standard since H.261. This motion compensation is done through the redundant information between consecutive frames looking for a pattern that captures the kind of movement between pictures. This pattern is represented as a **motion vector**, which defines a distance and a direction and has two dimensions: *right_x* and *down_x*. Important differences from earlier standards include the support for a range of block sizes and the use of multiple reference frames to improve the performance of the coding.

H.264 supports motion compensation block sizes ranging from 16x16 to 4x4 samples. Each macroblock may be split up in 4 ways: 16x16, 16x8, 8x16 or 8x8. Each of the sub-divided regions is a **macroblock partition**. If the 8x8 mode is chosen, each of the four 8x8 macroblock partitions within the macroblock may be further split in 4 ways: 8x8, 8x4, 4x8 or 4x4 (known as **sub-macroblock partitions**). These partitions and sub-partitions give rise to a large number of possible combinations within each macroblock (Fig. 1). This method of partitioning macroblocks into motion compensated sub-blocks of varying size is known as *tree structured motion compensation*.

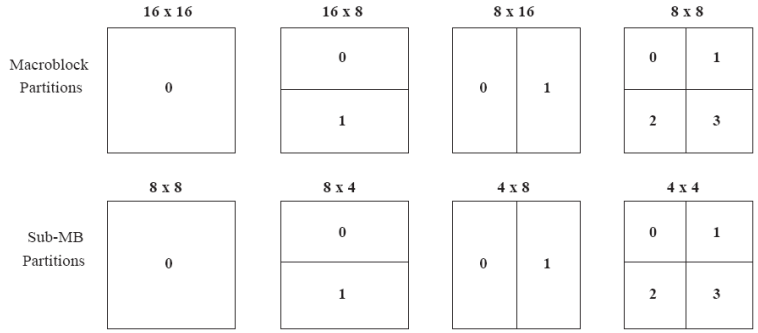


Figure 1: Macroblock types.

Since each macroblock and sub-macroblock partition has a motion vector associated, a macroblock has from 0 to 16 pairs of *motion vectors*. For example, an Intra macroblock has not any motion vector, an Inter macroblock partitioned into two 16x8 blocks has two pairs of motion vectors and an Inter macroblock partitioned into four 8x8 blocks -each of them partitioned into 4x4 sub-macroblocks- has 16 pairs of motion vectors ($4 * 4 = 16$).

 ISBN: 978-84-614-6167-7

3 Lane marking identification

In this section, a technique to identify lane markings in road scenes is described. A scheme of the process is shown in Figure 2.

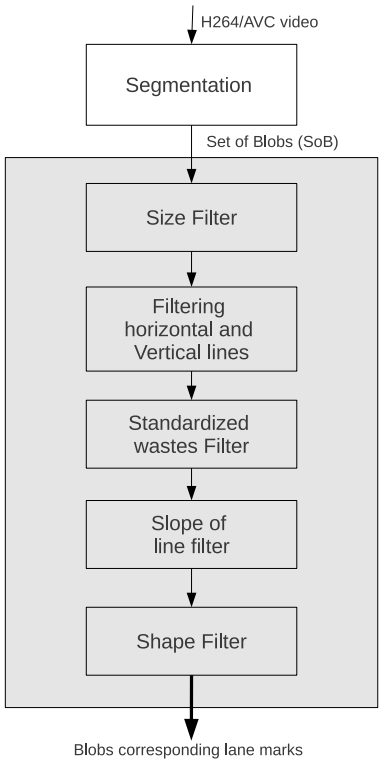


Figure 2: Operation scheme of the proposed method.

Initially, the objects of interest in the scene are detected using a segmentation method based in approximate reasoning [9]. From this algorithm a set of blobs (SoB) (Equation 1) is obtained.

$$SoB = \{B_1, B_2, ..., B_n\} \tag{1}$$

A blob is defined as a group of connected pixels but in the context of video compressed domain, a blob is defined as a group of similar macroblocks. Similarity between macroblocks is considered when motion vectors associated to them are similar in magnitude and direction. So, we can conclude each blob represents a region of the image with similar motion characteristics. The formal representation of a blob is:

$$B =< FN, Size, ML > \tag{2}$$

where FN is the frame number in which is located the blob, $Size$ is the number of motion vectors grouped in the blob and ML is the list of macroblocks belonging to

 ISBN: 978-84-614-6167-7

the blob as it is shown in Equation 3:

$$ML = \{(x_1, y_1), (x_2, y_2)..., (x_N, y_N)\} \tag{3}$$

where the components (x_i, y_i) represents the position in the picture of each macroblock.

As the list of the macroblocks is saved in each blob, the objects can be made up to obtain the visual representation as viewed in Figure 3. An example of blob could be the next: $< 325, 5, \{(22, 11), (23, 11), (13, 1), (14, 2), (12, 22)\} >$.

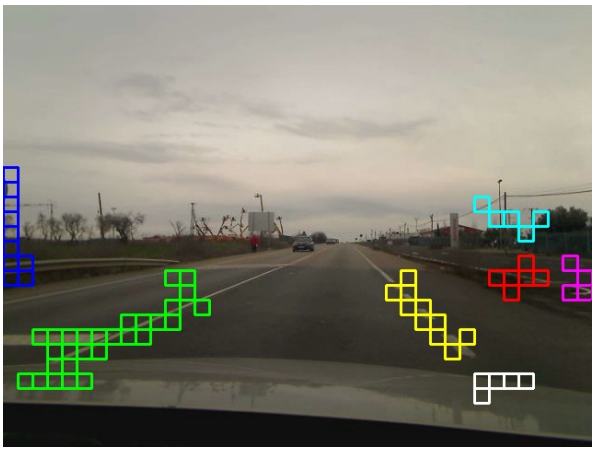


Figure 3: Example of segmented frame.

Once obtained the set of blobs (SoB) in a frame from the segmentation stage a set of filters that removes progressively blobs depending of their position in the scene, their size and their shape are applied. Finally, the remaining blobs from the input data set will be considered the road lane markings in the scene.

3.1 Filtering of blob position and size

Now, the filters that discard blobs as lane marks depending of the position and size are described. In addition, a mask is applied to remove spurious macroblocks of blobs that may come from merge errors (shadows, brightness,...) of the segmentation algorithm. First, those blobs with $Size$ (number of macroblocks in ML) too small or too big are discarded. Only the blobs satisfying Equation 4 are considered for the following steps. In the tested videos it was observed that if $Size$ was smaller than 5, significant statistical conclusions can not be obtained and also that lines rarely were bigger than 80 elements.

$$(Size(B) \geq M) \; or \; (Size(B) \leq M') \tag{4}$$

where $Size(B)$ selects the attribute size from a Blob B.

Later, those blobs whose macroblocks in ML have same x, $x_i = x_j \forall i, j$ or same y, $y_i = y_j \forall i, j$ are rejected because they represent vertical and horizontal lines respectively, which knowing the camera position can not be lane markings.

 ISBN: 978-84-614-6167-7

Now, we describe the only filter that does not remove a complete blob but modifies the set of macroblocks in ML. More concretely, it tries to remove noise caused by shadows, merge of regions, etc., and represented in the linguistic blob as a subset of spurious macroblocks. To do this, using standardized wastes, those macroblocks represented by pairs (x_i, y_i) that have a measure (expressed in absolute value) greater than or equal to 2 are removed.

Once removed that noise, linear regression [7] is used to discard blobs representing too horizontal or too vertical lines since they can not be considered road lane marks because of the position of the camera. Linear regression [7] tries to fit a straight line to a "cloud of points". Straight line of regression of Y on X is represented by Equation 5, where $\bar{y}$ represents the arithmetic mean of the marginal variable Y, S_x^2 is the variance of X and S_{xy} is the covariance.

$$y - \bar{y} = m(x - \bar{x}) \tag{5}$$

The straight line $y = (m \times x) + b$ is the one that better adjusts to the cloud of points. The values of m and b are obtained using Equations 7 and 6 respectively. N represents the number of elements in ML.

$$m = \frac{\sum_{i=1}^{N} x_i y_i - N\bar{x}\bar{y}}{\sum_{i=1}^{N} x_i^2 - N\bar{x}^2} \tag{6}$$

$$b = \bar{y} - m\bar{x} \tag{7}$$

Once the slope m is calculated, the blobs whose set of macroblocks ML do not take values in the range: $0.1 < |m| < 0.9$ are removed.

Now, the position of the blobs in the picture is studied. Blobs on the left of the scene and having a negative slope or blobs on the right having positive slope are removed. This step is useful because it is well known that left side lane markers always slope upwards from left to right, while right side lane markers slope downwards from left to right. Then, if the resolution of the video is $widthxheight$ we eliminate a blob if it verifies the condition describe in Equation 8.

$$((max(x_i) < width/2) \; and \; m < 0) \; or \; ((min(x_i) > width/2) \; and \; m > 0) \tag{8}$$

where $max(x_i)$ and $min(x_i)$ represent maximum and minimum of x_i in ML.

Finally, blobs in upper areas of the pictures (power lines, clouds, ...) are deleted. By means of Equation 9 blobs detected in the upper third of the scene are removed.

$$max(y_i) \geq (2/3) * height \tag{9}$$

3.2 Filtering based on shape

In this section, the shape of the blobs is determined by using a set of statistical measures. Once selected the candidate blobs to be road lane marks, correlation of values in ML

 ISBN: 978-84-614-6167-7

is studied. That is because the coefficient correlation r (Equation 10) allows to check the linearity of the cloud points of ML. This linearity allows to recognize lane marks because this is the shape (distribution of the macroblocks of ML in the picture) of the blobs corresponding with these lane marks.

$$r = \frac{SC_r}{S_{yy}} \tag{10}$$

Now, operations needed to solve Equation 10 are described (Equations 11 to 12) where $\bar{x}$ and $\bar{y}$ are the arithmetic mean of x_i and y_i respectively.

$$SC_r = \sum_{i=1}^{N} (y_i - ((m \times x_i) + b))^2 \tag{11}$$

$$S_{yy} = (\sum_{i=1}^{N} y_i^2) - N\bar{y}^2 \tag{12}$$

Those blobs with r less than or equals to $2/\sqrt{N}$ are refused. This value is a simplification of the t-Student (Equation 13).

$$t = \sqrt{\frac{N-2}{1-r^2}} \tag{13}$$

In a first set of experiments, two additional shape filters were considered. Dependence between variables was studied through the $test\beta$. In addition, this test was corroborated with the $F - Snedecor$ test. As no improvement of the results was obtained, both tests were finally not considered in this work.

4 Example of application of filters

Now, the filtering process described in Sections 3.1 and 3.2 is detailed by means of an example where the blobs detected in a frame are analysed. In Figure 4, the blobs detected in the segmentation stage are shown. Each blob is identified by a number. Let note that the same macroblock can be assigned to different blobs. That is because H.264/AVC allows the generation of more than one motion vector for each macroblock. For example, blob number 3 is overlapped by blob 15, and so, interpretation of data in this figure can be confusing.

In Table 1, values that are used to apply the different filtering stages are shown. More concretely, N_b is the number of detected blobs, N is the number of elements in ML; columns X_i and Y_i take the value $TRUE$ if every x_i, y_i are equals $\forall i$; m is the slope of the regression line; $Mx(x)$ is the maximum value of x_i; $Mn(x)$ is the minimum value of x_i; $Mx(y)$ is the maximum value of y_i; r is the correlation coefficient; $2/R(N)$ is the value of contrast to r; L_i is the value that takes $TRUE$ if the blob is detected as a lane mark and takes $FALSE$ in other case; L represents the right value for this blob.

 ISBN: 978-84-614-6167-7

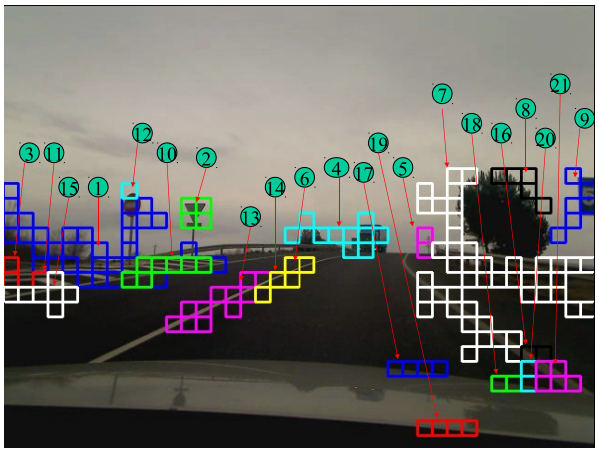


Figure 4: Example of segmented frame with blobs number.

In Figure 4 it can be observed that blobs number 1 and 7 are discarded because their size is bigger than 80 points (N column). Blobs number 3, 5, 6, 11, 12, 16, 17, 18, and 19 are not considered because have the same values of x_i or y_i (X_i and Y_i columns). Blobs 2, 4 and 15 are removed for having a very small slope (m column). Blobs number 9 and 20 have positive slope and they are on a right position in the scene (m, $Mx(x)$,$Mn(x)$ columns). There is no blob is detected in areas corresponding to the sky. Comparing the correlation coefficient r with the value $2/R(N)$, those blobs in which $r < 2/R(N)$ are discarded (blobs number 14 and 21). It can be observed that in both cases the difference is very small and the lines could be considered. Once applied the set of filters, blobs 8, 10, and 13 are not removed and they are considered as road lane marks. Actually, blob 8 is a wrong result.

5 Experimental results

The proposed technique was tested over a dynamic traffic video scene. The resolution of the video was 640x480 pixels and 28 frames have been selected from a total number of 2471. The selection criteria of this subset of frames is the exploration of different conditions of operation. In the segmentation stage, 358 blobs have been detected. In Table 2, experimental results are shown. As can be observed 69 of the total number of blobs correspond with lane markings in the scene, 50 of them are detected by our technique (72.46%). From a total number of 289 blobs not corresponding to lane marks, the system determines that 283 are effectively not lane markings (97.92%).

In Figure 5, it can be observed the number of blobs removed in each step, and from those blobs deleted which of them are correctly removed and those that are not.

 ISBN: 978-84-614-6167-7

Table 1: Values extracted from individual blobs.

N_b	N	X_i	Y_i	m	$Mx(x)$	$Mn(x)$	$Mx(y)$	r	$2/R(N)$	L_i	L
1	171									F	F
2	9	F	F	-610.35	13	12	17	0.0	0.6666	F	F
3	10	T	F							F	F
4	18	F	F	-0.0147	25	19	16	0.0547	0.4714	F	F
5	6	T	F							F	F
6	6	F	T							F	F
7	123									F	F
8	14	F	F	-0.8381	36	33	19	0.7903	0.5345	T	F
9	28	F	F	0.9244	39	37	19	0.5242	0.3779	F	F
10	10	F	F	0.2266	13	8	13	0.7399	0.6325	T	T
11	5	F	T							F	F
12	5	T	T							F	F
13	29	F	F	0.4532	19	11	12	0.9163	0.3714	T	T
14	6	F	F	0.4000	19	17	12	0.8000	0.8165	F	F
15	14	F	F	0.0						F	F
16	7	F	T							F	F
17	5	F	T							F	F
18	5	F	T							F	F
19	7	F	T							F	F
20	5	F	F	0.4999	36	35	6	0.6124	0.8944	F	F
21	13	F	F	-0.3333	38	36	6	0.5375	0.5547	F	F

Table 2: Experimental results extracted from 358 blobs.

	Lane Mark	**No Lane Mark**	**Total**
Lane Marks Detected	50	5	55
Lane Marks not Detected	19	284	303
Total	69	289	358

6 Conclusions

A novel technique that works only with information obtained from compressed domain is presented. The proposed algorithm can deduce if one blob corresponds with a road lane mark or not by means of filtering process of data obtained from a segmentation algorithm and using information about the size, the position and the shape of these elements. Although input data for the tests is obtained from a generic algorithm not oriented to traffic video sequences, encouraging results in this first approximation to the problem are obtained.

As future work, the authors consider the improvement of computational efficiency using the described filters directly in the segmentation stage and not once this phase is finished. More precisely, it can be said that it must be studied if during the segmentation process (grouping similar macroblocks to compose blobs) is possible to detect what is a lane mark and what is not.

 ISBN: 978-84-614-6167-7

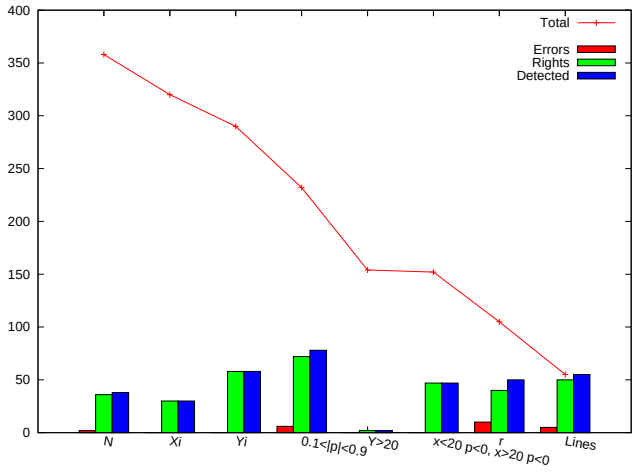


Figure 5: Filtering capacity of each test.

Acknowledgements

This work has been funded by the Regional Government of Castilla-La Mancha under the Research Project PII1C09-0137-6488 and by the Spanish Ministry of Science and Technology under the Research Project TIN2009-14538-C02-02.

References

[1] S. G. Jeong, C. S. Kim, K. S. Yoon, J. N. Lee, J. I. Bae and M. H. Lee, *Real-time lane detection for autonomous navigation*, In: Proceedings of 4th IEEE International Conference on Intelligent Transportation Systems, Oakland, California (2001) 508–513.

[2] Joint Video Team (JVT), *Draft ITU-T Recommendation and Final Draft International Standard of Joint Video Specification, ITUT Recommendation H.264 and ISO/IEC 14496/10 AVC* (2003)

[3] C. R. Jung and C. R. Kelber, *A lane departure warning system based on a linear-parabolic lane model*, in: IEEE Intelligent Vehicles Symposium, Parma. Italy (2004) 891–895.

[4] K. Kaliyaperumal, S. Lakshmanan and K. Kluge, *An algorithm for detecting roads and obstacles in radar images*, IEEE Transactions on Vehicle Technology **50 (1)** (2001) 170–182.

[5] Z. Kim, *Robust Lane Detection and Tracking in Challenging Scenarios*, IEEE Trans. on Intelligent Transportation Systems **9 (1)** (2008) 16–26.

[6] I. E. G. Richardson, *H.264 and MPEG-4 Video Compression*, John Willey & Sons Ltd, (2003).

[7] G. A. F. Seber and A. J. Lee, *Linear Regression Analysis*, Wiley-interscience, 2003.

[8] B. Serge and B. Michle, *Road segmentation and obstacle detection by a fast watershed transform*, In: Proc. Intelligent Vehicles Symposium, Paris (1994) 296–301.

[9] C. Solana-Cipres, G. Fernandez-Escribano, L. Rodriguez-Benitez, J. Moreno-Garcia and L. Jimenez-Linares, *Real-time moving object segmentation in H.264 compressed domain based on approximate reasoning*, Int. J. Approx. Reasoning **51 (1)** (2009) 99–114.

[10] J. G. Wang, C. J. Lin and S.M. Chen, *Applying fuzzy method to vision-based lane detection and departure warning system*, Expert Syst. Appl. **37 (1)** (2010) 113–126.

[11] Y. Wang, E.K. Teoh and D. Shen, *Lane detection and tracking using B-Snake*, Image Vision Comput. **22** (2004) 269–280.

 ISBN: 978-84-614-6167-7

Proceedings of the 11th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2011
26–30 June 2011.

A predictive estimator of the proportion with missing data

Silvia González Aguilera[1] and M del Mar Rueda García[2]

[1] *Department of Statistics & OR, University of Jaén*

[2] *Department of Statistics & OR, University of Granada*

emails: sgonza@ujaen.es, mrueda@ugr.es

Abstract

This paper considers the problem of estimating a population proportion when there are missing values. The prediction approach is used to define a new estimator that presents desirable efficiency properties. Simulation studies are considered to evaluate the performance of the proposed estimator via the empirical relative bias and the empirical relative efficiency, and favourable results are achieved.

Key words: superpopulation models, missing data, auxiliary information

1 Introduction

The use of the auxiliary population information, provided by one or several auxiliary variables, at the estimation stage is a very usual technique with many advantages. Powerful auxiliary information can produce a successful reduction of the bias and the sampling error. Techniques as ratio, difference o calibration produce new estimators that are generally more efficient than other methods which do not make use of auxiliary information.

The auxiliary information can be also used in order to treatment of missing values. The problem of missing data, is a common aspect in almost all investigations. Indeed, often some sampling units does not participate in the study or refuse to answer all the questionnaire, the interviewer is not able to contact with all sampling units, or they are accidental loss of information caused by unknown factors. For it, some sample data are not complete for the variables included in the study. The treatment of missing data in survey research is not a simple job. A variety of methods have been developed to attempt to compensate for missing data in a general purpose way that enables the survey's data file to be analyzed without regard for the missing data.

When some observations in the sample are missing, the simplest solution is to eliminate the incomplete observations and to restrict the study to complete observations

 ISBN: 978-84-614-6167-7

for all variables to apply the calibration techniques with the sample of complete data. One obvious consequence, is that the actual sample size is less than the planned one. This can produce biases in the estimations and greater sampling variance.
Another solution is to employ some imputation techniques to replace the missing observations. see e.g *Little* and *Rubin* (1987), *Rao* and *Toutenburg* (1995), *Saŕndal* (1992) and treating these imputed values as true observations, one may apply the indirect methods using the standard procedures without any missing observation. Such a practice may tend to invalidate the inferences and may often have serious consequences.

Considering that the deleted observations may contain valuable information, a third option will be to try to improve the precision of the estimators by including all cases available for their calculation. Some authors have defined ratio estimators for the population mean when the sample is drawn according to the procedure of simple random sampling without replacement when some observations are missing, see, e.g., *Tracy* and *Osahan* (1994) and *Toutenburg* and *Srivastava* (1998, 1999, 2000). However, there appears to be no investigation reported in the literature when the population proportion is investigated.

Therefore, our pretension in this work, is to build some modifications of the prediction estimator that employ the information in the sample for the study and auxiliary variables, for the estimation of the population proportion, on the basis of a logistic regression superpopulation model.

2 Proportion estimators in presence of missing values

Let $U = \{1, 2, ..., N\}$ be a population of N identifiable elements. We consider the problem of estimating the population proportion $P_A = N^{-1} \sum_{i \in U} y_i$, where y_i is an attribute indicator for unit i, i.e., $y_i = 1$ if unit i has the attribute of interest Y, and $y_i = 0$ otherwise. P_A is the parameter of interest, which needs to be estimated. For this purpose, a random sample s, of size n, is selected from U according to a given sampling design. The first- and the second-order inclusion probabilities associated to the sampling design are denoted, respectively, as π_i and π_{ij}, and which are assumed to be strictly positive. The design weight associated to the unit i is given by $d_i = \pi_i^{-1}$.

The population proportion P_A can be estimated by using the well-known Horvitz-Thompson estimator, which is given by

$$\widehat{p}_A = \frac{1}{N} \sum_{i \in s} d_i y_i, \tag{1}$$

Estimator (1) makes no use of auxiliary information. However, it is common to assume that there exists an auxiliary variable which can be used at the estimation stage to improve the estimation of the parameter of interest. For this reason, we assume an auxiliary attribute X, whose values are known from a previous census. This assumption is commonly used in the survey sampling context. On the other hand, a problem of

 ISBN: 978-84-614-6167-7

missing data can occur in the sample s and the estimator (1) can not be calculated in this situation.

Throughout this paper, we assume missing data on the sample s, which can be divided into the disjoint sets

$$s_1 = \{i \in s/y_i, x_i \text{ are non-missing}\}$$
$$s_2 = \{i \in s/y_i \text{ are missing, and } x_i \text{ are non-missing}\},$$

with s_1 of size $n-c$ and s_2 of size c. We assume that c is an integer numbers verifying $0 < c < n/2$.

Prediction theory for sampling surveys can be considered as a general framework for statistical inference on the characteristics of finite populations. The general prediction theory is based on superpopulation models, that assumes that the population under study $\mathbf{y} = (y_1, ..., y_N)'$ is considered to be a realization of super-population random variables $\mathbf{Y} = (Y_1, ..., Y_N)'$ having a superpopulation model ξ. The value of the variable of interest, associated with the i-th unit of the population is comprised of a deterministic element η_i (known) and a random element:

$$Y_i = m(x_i) + e_i \tag{2}$$

$i = 1, ..., N$. The random vector $e = (e_1, ..., e_N)$ is assumed to have zero mean and a positive definite covariance matrix which is diagonal (Y_i are mutually independent).

The superpopulation model defines a class of distributions ξ, which supposes that all sample values are known and they are not missing values.

A good deal of inference in survey sampling emerges from the postulation of a suitable distribution ξ for Y. Traditionally, parametric methods utilize regression models to incorporate auxiliary information: $m(x_i) = \beta x_i$. Also one can use nonparametric models as Breidt and Opsomer (2000). Since the variable Y is dichotomous when we estimate a proportion, it may be more natural to consider a logistic model for the population, where it is assumed that .. k U.. is known. For a given x_k, the model is given by:

$$Pr(Y_k = 1) = \frac{exp(x_k'\beta)}{1 + exp(x_k'\beta)} + e_i$$

and $Pr(Y_k) = 0) = 1 - Pr(Y_k = 1)$.

We denote by $\mu_k = E(Y_k|x_k, \beta) = Pr(Y_k = 1|x_k, \beta)$ and $\overline{\mu} = \frac{1}{N}\sum \mu_k$. This model was used by Duchesne (2003).

In our working context, we consider the problem of the prediction or population proportion. For any given $s \in S_d$ of size n we can write:

$$P_A = f_{s1}P_{As1} + (1 - f_{s1})P_{A\widetilde{s_1}} \tag{3}$$

where $f_{s1} = \dfrac{n-c}{N}$, P_{As_1} is the proportion for units in the sample s_1, and $P_{A\widetilde{s_1}}$ is the proportion for the units with not known values of y .

In this representation of the proportion, sample proportion P_{As1} is known, and then we attempt a post survey prediction of the proportion $P_{A\widetilde{s_1}}$ of the unknown units.

 ISBN: 978-84-614-6167-7

We denote by E_ξ the expected value under the model ξ and E_d the expected value under the design d.

Consider any predictor T of P_A; it can be represented, for any given sample s as:

$$T = f_{s1}P_{As1} + (1 - f_{s1})U \tag{4}$$

where U is considered as predictor of $P_{A\widetilde{s_1}}$.

The minimum $E_\xi MSE_d$ criterium will be considered. If T is design-model unbiased, we have:

$$E_\xi MSE_d(T) = E_d E_\xi (T - \overline{\mu})^2 - V_\xi(P_A)$$

Hence, if we can find T_B to minimize $E_\xi(T - \overline{\mu})^2$ for any $s \in S_d$, and d is noninformative, then T_B also has the property of minimizing $E_\xi MSE_d(T)$ for any given design. We only consider the linear and model unbiased predictors.

Now, T is model-unbiased if

$E_\xi(U) = \mu_{\widetilde{s_1}} = \frac{1}{N-n+c} E_\xi(Y_i)$, $\forall s | p(s) > 0$

$$E_\xi MSE_d(T) = E(V_\xi(U) + V_\xi(P_{A\widetilde{s_1}}) - 2cov(U, P_{A\widetilde{s_1}})$$

and if Y_i are independents $cov(U, P_{A\widetilde{s_1}}) = 0$.

Minimize $E_\xi MSE_d(T)$ in T is equivalent to minimize $V_\xi(U)$ in U assuming that $E_\xi(U) = \mu_{\widetilde{s_1}}$. Hence, in this case, for a given s, the optimal model unbiased predictor of T is (Royall, 1970)

$$T_B = f_{s1}P_{As1} + (1 - f_{s1})\hat{U} \tag{5}$$

where $E_\xi(\hat{U}) = \mu_{\widetilde{s_1}}$ and $V_\xi(\hat{U}) \leq V_\xi(U)$.

The best linear unbiased predictor for U under this model is, therefore,

$$\hat{U} = \frac{1}{N - (n - c)} \sum_{\widetilde{s_1}} \frac{exp(x_k'\widehat{\beta})}{1 + exp(x_k'\widehat{\beta})}$$

being $\widehat{\beta}$ is the BLUP estimator of β under the regression logistic model.

Then we define the predictor of P_A as

$$T_B = f_{s1}P_{As1} + \frac{1}{N - (n - c)} \sum_{\widetilde{s_1}} \frac{exp(x_k'\widehat{\beta})}{1 + exp(x_k'\widehat{\beta})}. \tag{6}$$

3 Simulation study

We now compare empirically the performance of the proposed estimator with the simple available cases estimators ($\hat{P}_{As1}$), the ratio complete cases estimator ($\hat{P}^r_{As1}$) and the regression complete cases estimator ($\hat{P}^{reg}_{As1}$).

Estimators are evaluated by using a total of 4 simulated populations with population size $N = 1000$. These populations were generated as a random sample of 1000

 ISBN: 978-84-614-6167-7

units from a Bernoulli distribution with parameter $P_A = \{0.5, 0.75\}$, and the attributes of interest were thus achieved with the aforementioned population proportions. It is known that the performance of indirect estimators depends on the relationship between the auxiliary and interest variables. In the context of binary variables, this relationship is measured by the Cramer's V coefficient, which we denote as ϕ. Auxiliary attributes were also generated by using the same distribution, but we randomly change a given proportion of values in order to the Cramer's V coefficient between the attribute of interest and the auxiliary attribute takes the values 0.5 and 0.7.

For each simulation, 1000 samples, with size $n = 100$, were selected under SRS to compute the empirical relative efficiency (RE) of the estimators $\hat{P}^r_{As1}$, $\hat{P}^{reg}_{As1}$ and T_B with respect to $\hat{P}_{As1}$,

$$RE(\hat{P}^{reg}_{As1}) = \frac{MSE[\hat{P}_{As1}]}{MSE[\hat{P}^{reg}_{As1}]}$$

$$RE(\hat{P}^{reg}_{As1}) = \frac{MSE[\hat{P}_{As1}]}{MSE[\hat{P}^{reg}_{As1}]}$$

$$RE(T_B) = \frac{MSE[\hat{P}_{As1}]}{MSE[T_B]},$$

where $MSE[\cdot]$ denote the empirical mean square error.

For some values of c, we generated missing values for the attribute y, following 3 different missing mechanism: uniform response, unconfounded mechanism increasing x and confounded mechanism increasing y.

The obtained results are shown in table 1.

As you can see in the former table, in the sense of efficiency, the proposed estimator is better than estimator $\hat{P}^{reg}_{As1}$ in 66.67% of the cases, better than estimator $\hat{P}^r_{As1}$ in 63.89% of the cases, and better than estimator $\hat{P}_{As1}$ in 77.78 % of the cases.

Anyway, the estimator T_B is better than the rest in 47.22% of the cases. $\hat{P}^r_{As1}$ and $\hat{P}^{reg}_{As1}$ are better than the rest in 33.33% and 19.44% of the cases respectively. Finally, $\hat{P}_{As1}$ is equal or less efficient than the rest in all cases.

Results derived from this simulation study indicate that the proposed estimator can provide desirable estimates in the presence of missing data, as it makes a good use of the available information at the estimation stage.

References

[1] Breidt, F.J. and Opsomer, J.D. (2000) *Local polynomial regression estimators in survey sampling.* The Annals of Statistics 28,4, 1026-1053.

[2] Duchesne, P. (2003) *Estimation of a Proportion with Survey Data* Journal of Statistics Education 11, 3. John Wiley, New York.

[3] Little, R. J. A. and Rubin, D. B. (1987) *Statistical analysis with missing data.* John Wiley, New York.

 ISBN: 978-84-614-6167-7

Table 1: Empirical relative eficiency (RE) of the estimators $\hat{P}^{r}_{As1}$, $\hat{P}^{reg}_{As1}$ and T_B with respect to $\hat{P}_{As1}$

Uniform response										
V	p		c=10			c=20			c=30	
		T_B	$\hat{P}^{reg}_{As1}$	$\hat{P}^{r}_{As1}$	T_B	$\hat{P}^{reg}_{As1}$	$\hat{P}^{r}_{As1}$	T_B	$\hat{P}^{reg}_{As1}$	$\hat{P}^{r}_{As1}$
0.5	0.5	0.9108	1.0757	0.7819	1.1435	1.0702	0.7370	1.5049	1.0188	0.6580
	0.75	1.3054	1.0653	0.6173	1.1583	1.1120	0.6376	0.9936	1.0759	0.5002
0.9	0.5	1.0447	1.0753	2.2055	1.3476	1.0086	1.4857	1.7790	0.9651	1.1971
	0.75	1.5971	1.0855	2.2001	1.5067	1.0483	1.4011	1.4342	1.0757	1.1809
Unconfounded mechanism Increasing x										
V	p		c=10			c=20			c=30	
		T_B	$\hat{P}^{reg}_{As1}$	$\hat{P}^{r}_{As1}$	T_B	$\hat{P}^{reg}_{As1}$	$\hat{P}^{r}_{As1}$	T_B	$\hat{P}^{reg}_{As1}$	$\hat{P}^{r}_{As1}$
0.5	0.5	0.8240	1.1304	0.8271	0.9204	1.0715	0.8538	1.1452	1.1841	1.1148
	0.75	1.3061	1.0834	0.6661	1.2505	1.0300	0.6324	1.7188	1.0759	0.6009
0.9	0.5	0.9919	1.0634	2.3760	1.0053	1.0198	1.9860	1.3492	1.0292	3.9238
	0.75	1.4383	0.9712	2.0902	1.7050	1.0160	1.3853	2.2957	1.0275	1.7039
Confounded mechanism Increasing y										
V	p		c=10			c=20			c=30	
		T_B	$\hat{P}^{reg}_{As1}$	$\hat{P}^{r}_{As1}$	T_B	$\hat{P}^{reg}_{As1}$	$\hat{P}^{r}_{As1}$	T_B	$\hat{P}^{reg}_{As1}$	traz
0.5	0.5	0.7377	1.0503	0.7610	0.8320	1.1037	0.7789	1.3416	1.1352	1.5930
	0.75	1.4121	1.1159	0.6777	1.4330	1.0396	0.5454	2.8746	1.1055	0.7259
0.9	0.5	0.9790	1.0722	2.2471	1.0748	1.1266	2.1710	1.4027	1.0735	3.7032
	0.75	1.5930	1.0791	2.2544	1.7032	1.0148	1.4987	2.3685	1.0415	1.6627

 ISBN: 978-84-614-6167-7

[4] SRao C.R., Toutenburg, H. (1995), *Linear Models - Least Squares and Alternatives*. Series in statistics

[5] Särndal, C.E. (1992), *Methods for estimating the precision of survey estimates when imputation has been used*. Survey Methodology, 18, 241-252.

[6] Toutenburg, H. and Srivastava, V. K. (1998) *Estimation of ratio of population means in survey sampling when some observations are missing*. Metrika 48, 177–187.

[7] Toutenburg, H., Srivastava. V.K. (1999) *Amputation versus imputation of missing values through ratio method in sample surveys*, Unpublished document.

[8] Toutenburg, H. and Srivastava, V. K. (2000) *Efficient estimation of population mean using incomplete survey data on study and auxiliary characteristic.* Unpublished document.

[9] Tracy, D. S. and Osahan, S. S. (1994) Random non-response on study variable versis on study as well as auxiliary variables. Statistica 54, 163–168.

 ISBN: 978-84-614-6167-7

Proceedings of the 11th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2011
26–30 June 2011.

SparseBLAS Products in UPC: an Evaluation of Storage Formats

Jorge González-Domínguez[1], Óscar García-López[1], Guillermo L. Taboada[1], María J. Martín[1] and Juan Touriño[1]

[1] *Computer Architecture Group, Department of Electronics and Systems, University of A Coruña, Spain*

emails: jgonzalezd@udc.es, oscar.garcia@udc.es, taboada@udc.es, mariam@udc.es, juan@udc.es

Abstract

The performance of a significant number of applications in High Performance Computing (HPC) is determined by the efficiency of the sparse matrix-vector and matrix-matrix products. These computational kernels generally present poor scalability due to the lack of memory locality exploitation. Selecting the most appropriate storage format, which generally depends on the specific application scenario, can significantly improve their efficiency. This paper presents an evaluation of the most common sparse storage formats using Unified Parallel C (UPC). UPC is a Partitioned Global Address Space (PGAS) language that provides high programmability and performance through an efficient exploitation of data locality, especially on hierarchical architectures such as multicore clusters. Different combinations of storage formats and data distributions for the SparseBLAS matrix products are analyzed. Experimental results on an HP supercomputer using representative sparse matrices show that a suitable combination of storage formats and parallel algorithms has a great influence on the performance of the sparse products.

Key words: Sparse Matrices, Storage Formats, SparseBLAS, PGAS, UPC

1 Introduction

Sparse matrices are pervasive in many scientific and engineering areas, and the efficiency in their processing is critical for the performance of many applications. Many storage formats have been proposed to represent them. The minimization of the storage requirements is

 ISBN: 978-84-614-6167-7

not the only goal of these formats, but also the computational efficiency in sparse matrices operations. Thus, sparse numerical libraries must take into account the most suitable combination of storage formats and algorithms, especially when it comes to their processing in parallel on hybrid shared/distributed memory architectures.

The Partitioned Global Address Space (PGAS) programming model provides significant productivity advantages over traditional parallel programming paradigms. In this model all threads share a global address space, just as in the shared memory model. However, this space is logically partitioned among threads, just as in the distributed memory model. Thus, the data locality exploitation increases performance, whereas the shared memory space facilitates the development of parallel codes. As a consequence, the PGAS model has been gaining rising attention. A number of PGAS languages are now ubiquitous, being Unified Parallel C (UPC) [1] a representative example.

UPC is an extension of ANSI C for parallel computing. In [2] El-Ghazawi and Cantonnet established, through an extensive evaluation of experimental results, that UPC can potentially perform at similar levels to those of MPI. Besides, the one-sided communications present in languages such as UPC were demonstrated to be able to obtain even better performance than the traditional two-sided communications [3]. Barton et al. [4] further demonstrated that UPC codes can scale up to thousands of processors with the right support from the compiler and the run-time system. More up-to-date evaluations [5, 6] have confirmed this analysis.

This paper presents an evaluation of the most suitable combinations of representative sparse storage formats (Coordinate, Compressed Sparse Row, Block Sparse Row, Compressed Sparse Column, Diagonal and Skyline) and parallel algorithms for the implementation of the SparseBLAS matrix-vector and matrix-matrix products using UPC. SparseBLAS products are core routines for most iterative solvers and matrix factorizations and thus their performance has a great influence on a wide variety of scientific and engineering applications.

The rest of this paper is organized as follows. Section 2 summarizes the related work. Section 3 describes the sparse storage formats evaluated. Sections 4 and 5 outline the different algorithms used to perform the sparse matrix-vector and matrix-matrix products, respectively, depending on the storage format. Section 6 presents the analysis of the experimental results obtained on an HP supercomputer (Finis Terrae). Finally, conclusions are discussed in Section 7.

2 Related Work

Due to the significant presence and impact in science and engineering of the sparse products, several optimization techniques have been proposed for their parallel implementation. Williams et al. [7] provide an efficient implementation of the matrix-vector product for multicore systems using the Compressed Sparse Row format by applying thread blocking

 ISBN: 978-84-614-6167-7

together with sequential optimizations such as cache blocking, loop optimizations or software memory prefetching. Liu et al. [8] provide another implementation for the Block Sparse Row format using OpenMP. This work also evaluates three different types of load balancing, determining that the non-zero scheduling presented in [9] usually obtains the best performance. A new method for load balancing for the sparse matrix-vector product in heterogeneous systems was presented in [10].

Regarding the sparse matrix-matrix product, Buluc and Gilbert [11] compare different algorithms and data distributions for the multiplication of two sparse matrices. However, this routine is not the same as the one in the SparseBLAS library [12], which multiplies a sparse matrix by a dense one. Our paper will analyze this latter one together with the sparse matrix-vector product.

The selection of the most suitable storage format is one of the main decisions in order to perform efficient products, and this decision can be influenced by the size of the problem, the sparsity pattern of the matrix, the programming language or the architecture of the system. In [13] Luján et al. presented a performance evaluation of different storage formats for the sparse matrix-vector product in Java. This study was complemented in [14] with a similar evaluation using Fortran. Regarding parallel computing, Shahnaz et al. provide in [15] and [16] a comparison of the performance of the sparse matrix-vector product with seven different formats in a small cluster using MPI. Similar studies for GPUs are presented in [17] and [18].

Nevertheless, none of these works take advantage of the use of PGAS languages. Bell and Nishtala [19] deal with sparse matrices in UPC but restricted to the sparse triangular solver and the Compressed Sparse Row format. Therefore, the novelty of our work in the PGAS programming model area is twofold: it is the first approach to the parallel sparse multiplications and it provides the first performance comparison among different sparse storage formats.

3 Sparse Matrix Storage Formats

The sparse storage formats define the structure to keep the data of the sparse matrices, and thus play an important role in achieving a good efficiency in sparse routines. This paper studies the formats described by Dongarra in [20], each of them tailored to particular sparsity patterns:

- Coordinate Format: It is the most intuitive, simple and flexible scheme to represent sparse matrices. It consists of three arrays, *values*, *rows* and *columns*, which store the values, row indices and column indices of the non-zero entries, respectively. In most occasions (and always in this work) the non-zero elements of the same row are assumed to be stored contiguously.

 ISBN: 978-84-614-6167-7

- Compressed Sparse Row (CSR) Format: This format is probably the most popular sparse representation. It explicitly stores subsequent non-zero values of the matrix rows in array *values*. Array *columns* keeps the column indices. A third array *rowPtr* stores, for each row, the index of the entry in the array *columns* which is the first non-zero element of the given row. It has an additional entry with the total number of non-zero elements in the matrix. Therefore, CSR presents a compressed view of Coordinate as the length of *rowPtr* is the total number of rows plus one instead of the total number of non-zero values.

- Block Sparse Row (BSR) Format: It is a variant of CSR, very useful for sparse matrices where the non-zero elements are grouped in blocks. It consists of dividing the matrix in a grid of blocks and keeping, for each block with non-zero entries, its values (including zeros) and the information of the position of the block within the grid according to the CSR scheme. The values are stored consecutively by blocks and, inside them, by rows.

- Compressed Sparse Column (CSC) Format: It is similar to CSR, but storing consecutively in array *values* the non-zero elements by columns, using *rows* for the row indices and *columnPtr* for keeping for each column the index of the entry in the array *rows* which is the first non-zero element of the given column.

- Diagonal Format: Many sparse matrices in scientific computing present their non-zero entries restricted to a small number of diagonals. In order to take advantage of this pattern the Diagonal scheme has been defined. In this case *values* stores consecutively all the elements of the diagonals with any non-zero element. Another array, *distance*, represents, for each stored diagonal, its offset from the main diagonal. Diagonals above and below the main one have positive and negative distance, respectively.

- Skyline Format: This format has been specifically designed for sparse triangular matrices, which frequently arise when solving linear systems. The concrete storage of the elements depends on whether the matrix is lower or upper triangular. The values of all the entries from the first non-zero element to the diagonal in each row/column are consecutively stored in *values* in the lower/upper case. Besides, an additional array *ptr* is necessary. In lower/upper matrices, it keeps for each row/column the index of the entry of *values* with the first element of this row/column. In both cases an additional entry with the total number of non-zero elements is needed.

4 Matrix-Vector Product

This section analyzes the SparseBLAS matrix-vector product: $\alpha * A * x + y = y$, where α is a scalar, A a sparse matrix and x and y dense vectors.

 ISBN: 978-84-614-6167-7

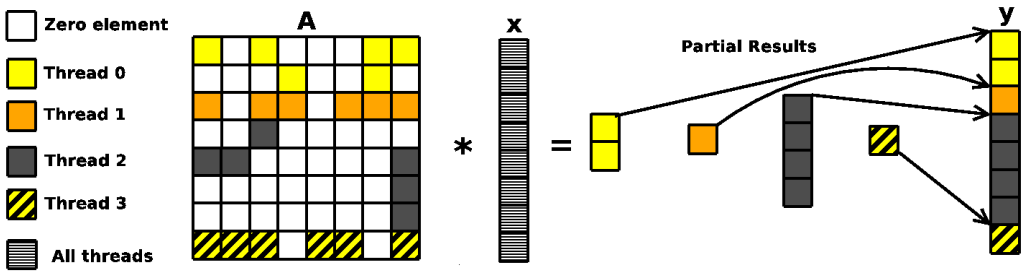


Figure 1: Sparse matrix-vector product using a row distribution for the matrix

All PGAS languages, and thus UPC, expose a global shared address space to the user which is logically divided among threads, so each thread is associated or presents affinity to a part of the shared memory. Moreover, UPC also provides a private memory space per thread for local computations. Therefore, each thread has access to both its private memory and to the whole global space. However, the accesses to remote data will be much more expensive than the accesses to data in local memory (private memory or shared memory with affinity to the thread). Thus, data distributions will have a serious impact on the performance of the parallel codes. For the parallel implementation of the matrix-vector product three different data distributions will be considered (by rows, columns and diagonals).

Figure 1 illustrates the behavior of the sparse matrix-vector product distributing the matrix by rows. This distribution relies on the consecutive storage by rows of the data in the `values` array so it can be used in the Coordinate, CSR, BSR and Skyline (with lower matrices) formats. Previous works have pointed out that a key aspect in the performance of the sparse matrix-vector product is the computational load balance [7]. In order to achieve a good load balance the matrix is distributed by blocks of rows of different size, trying to evenly distribute the number of non-zeros per thread (in the example, six non-zero elements per thread). In order to exploit data locality as much as possible each thread only accesses the rows of the matrix that correspond to it. Then, by applying a sequential partial sparse matrix-vector product with these rows and all the elements of x, each thread calculates a partial result that corresponds with its rows of A. Thus, the distribution of y must match the distribution of the matrix so that the partial sums can be performed without remote accesses. Besides, in the common case that the result vector is needed completely stored in an array in the local memory of one thread, this distribution by blocks only requires one bulk copy of remote data per thread. This bulk copy is performed in one go with the `upc_memget` function which is much more efficient than copying all the elements one-by-one (the UPC default access).

For the CSC and Skyline (with upper matrices) formats, where the data in the `values` array are consecutively stored by columns, the use of this row distribution would lead to

 ISBN: 978-84-614-6167-7

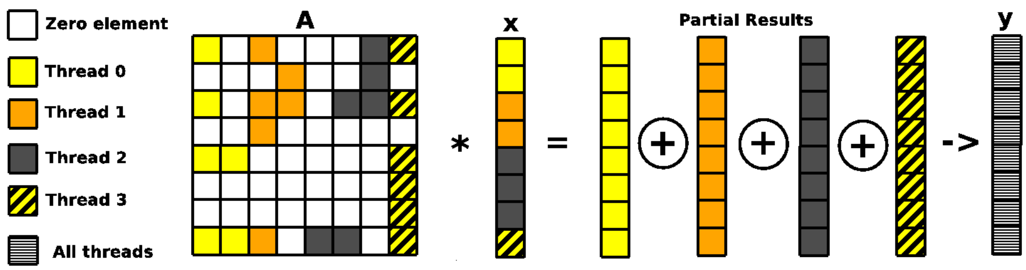


Figure 2: Sparse matrix-vector product using a column distribution for the matrix

several data movements, which can represent an important performance overhead. The natural distribution for these formats is by blocks of columns with variable block sizes in order to achieve a good computational load balance. Figure 2 shows this distribution for the same sparse matrix used in the row example. In this case the source vector x must be always distributed according to the size of the blocks in the matrix. In order to compute the i^{th} element of the result, the i^{th} values of all partial results should be added. These additions need reduction operations involving all UPC threads, so their performance is usually poor.

Regarding the Diagonal format, as the elements in the `values` array are consecutively stored neither by rows nor by columns, none of the presented distributions is eligible. In this case, the sparse matrix is distributed by diagonals as shown in Figure 3 and the final reductions are also mandatory. Furthermore, as the number of non-zero elements per diagonal is unknown, the computational load might be unbalanced (in the example, seven non-zero elements for threads 0 and 1 and five non-zero elements for threads 2 and 3). Nevertheless, the impact of this drawback is alleviated by using a cyclic distribution which achieves a balanced load distribution in most sparse matrices.

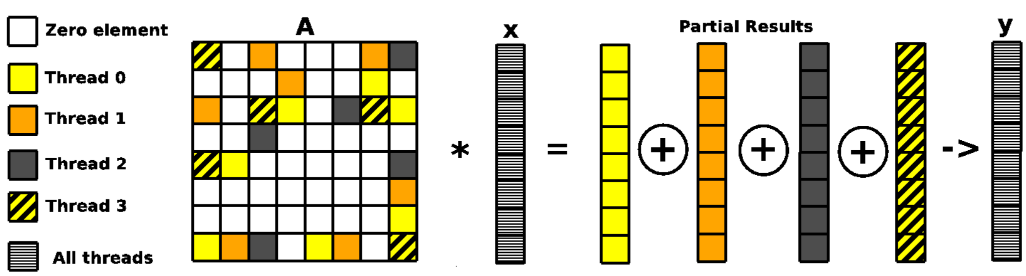


Figure 3: Sparse matrix-vector product using a diagonal distribution for the matrix

In UPC shared arrays can not be distributed with a variable block size. Thus, for all storage formats the sparse matrices are explicitly distributed into the memories of the threads using private arrays. Vectors in the matrix-vector product and dense matrices in

 ISBN: 978-84-614-6167-7

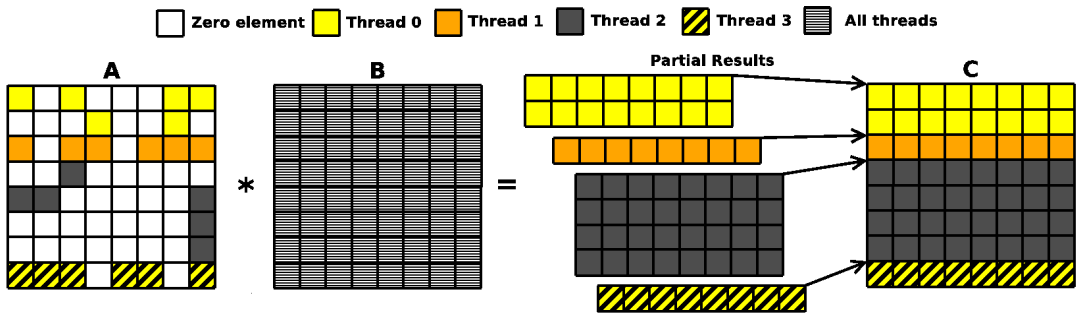


Figure 4: Sparse matrix-matrix product using a row distribution for the sparse matrix

the matrix-matrix product are instead stored in shared memory.

UPC provides functionality to access memory through pointers. A pointer to shared memory contains 3 fields: thread, block and phase. When performing pointer arithmetic on a pointer-to-shared all three fields must be updated, making the operations slower than with private pointer arithmetic. Thus, in all the implemented sparse products, when dealing with shared data with affinity to the local thread, the access is performed through standard C pointers instead of using UPC pointers to shared memory.

5 Matrix-Matrix Product

This section focuses on the SparseBLAS matrix-matrix routine: $\alpha * A * B + C = C$, where α is a scalar value, A a sparse matrix, and B and C dense matrices.

The first approach to parallelize this kernel consists of adapting the matrix-vector distributions and algorithms to this problem. For instance, Figure 4 shows the adaptation of the row distribution for matrices with eight columns. This data distribution will be applied to the Coordinate, CSR, BSR and Skyline (with lower matrices) sparse formats. Each thread needs the whole matrix B and the same rows of C as in A. As in the matrix-vector product, only one bulk copy per thread is required in case all the elements of the result matrix need to be consecutively stored in one local memory.

Nevertheless, the adaptation of the distribution by columns and diagonals employed in the matrix-vector multiplication would eventually involve a significant number of final reductions, leading to a very poor performance. Therefore, the approach illustrated in Figure 5, which avoids all the reductions, has been developed for CSC, Diagonal and Skyline (with upper matrices) formats. Each thread needs to access the whole sparse matrix but only the same columns of B and C. The block distribution is used because it allows to aggregate the copies of all elements of the same row of C using only one call to `upc_memget` per thread and row in a scenario where the output elements must be in one array in local

 ISBN: 978-84-614-6167-7

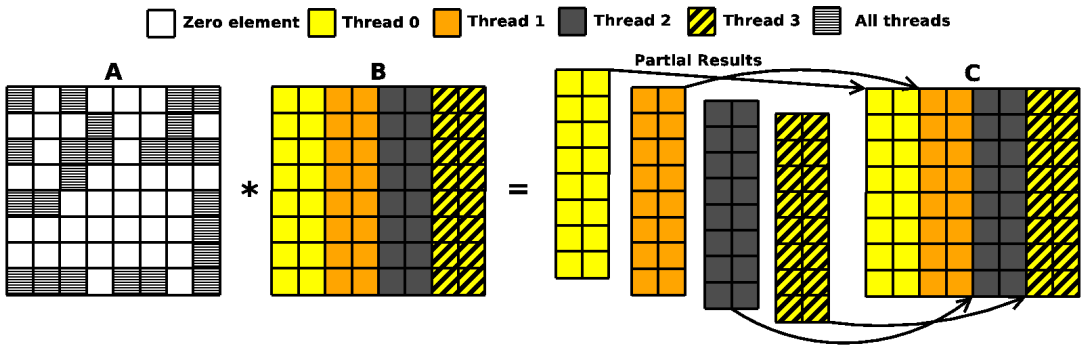


Figure 5: Sparse matrix-matrix product using a column distribution for the dense matrices

memory. This approach could also be used for the other sparse formats but it has been discarded because, in that scenario, it would lead to a greater number of copies (one bulk copy per row and thread) than in the row distribution of A (only one bulk copy per thread).

6 Performance Evaluation

The performance evaluation of the storage formats for SparseBLAS products in UPC has been conducted on the Finis Terrae supercomputer [21] at the Galicia Supercomputing Center (CESGA). This system consists of 142 HP RX7640 nodes, each of them with 16 IA64 Itanium2 Montvale cores at 1.6 Ghz, 128 GB of memory and a dual 4X InfiniBand port (16 Gbps of theoretical effective bandwidth). The cores of each node are distributed in two cells, each of them with 4 dual-core processors, grouped in pairs that share the memory bus (8 cores and 64 GB of shared memory per cell). As for software, the code was compiled using Berkeley UPC 2.12.1 [22]. The intra-node and inter-node communications are performed through shared memory and GASNet over InfiniBand, respectively.

In this evaluation four representative matrices, with different sparsity characteristics, have been selected from the University of Florida Matrix Collection [23]. Their characteristics are shown in Table 1. Larger versions (labeled with "large") have been obtained by replicating the original matrices, which preserves the sparsity and the pattern of the original ones. The larger versions have been used in the matrix-vector product whereas the original matrices have been used for the matrix-matrix product.

Figures 6 and 7 show the speedups of the double precision matrix-vector and matrix-matrix products, respectively, using up to 64 threads. These results have been obtained discarding the overhead of the initial data distribution (for many applications several consecutive products are performed with the same input data distributions). For clarity purposes, results with less than 8 threads are not shown as there are no significant differences among

 ISBN: 978-84-614-6167-7

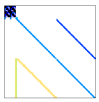
Plot	Name	Dimensions	Non-zeros	% sparsity
	nemeth26	9506x9506	760,633	0.842
	nemeth26_large	85554x85554	61,630,079	0.842
	TSOPF	18696x18696	4,396,289	1.258
	TSOPF_large	56088x56088	39,574,965	1.258
	gupta3	16783x16783	4,670,105	1.658
	gupta3_large	67132x67132	74,721,175	1.658
	exdata	6001x6001	1,137,751	3.159
	exdata_large	84014x84014	222,973,345	3.159

Table 1: Overview of the sparse matrices used in the evaluation

the analyzed formats. Some storage formats are not appropriate for storing some matrices due to the significant number of zeros that the format would require to store them, namely `gupta3` with Diagonal, and `TSOPF` and `exdata` matrices with Skyline. Thus, these combinations have not been considered. In the matrix-matrix product, `TSOPF` and `gupta3` matrices are not multiplied by square dense matrices but by matrices with 2000 and 3000 columns, respectively, because of memory constraints and to show the behavior of the function with different settings.

As expected, in the matrix-vector product row-based storage formats outperform significantly column and diagonal-based ones due to the avoidance of the final reduction operations, as shown in Section 4.

The analysis of the matrix-matrix results confirms that the differences between the two approaches presented in Section 5 is mainly due to the workload balance of the row distribution. Thus, when the workload is balanced with the row distribution approach, as in the case of `nemeth` and `gupta3` matrices, the formats that use this distribution (Coordinate, CSR, BSR and Skyline with lower matrices) are the best choice because they only need one data copy at the end of the algorithm (see Figures 4 and 5). However, for `TSOPF` and `exdata` matrices, workload is not completely balanced when relying on a row distribution because there are square blocks with a high number of non-zero elements. Therefore, formats that distribute the dense matrices by columns show higher scalability. Finally, the poor speedups of the Diagonal format for this routine is due to the fact that the sequential times are lower than using other formats thanks to the efficient exploitation of the cache hierarchy provided by this format on the evaluated matrices. Nevertheless, when the data

 ISBN: 978-84-614-6167-7

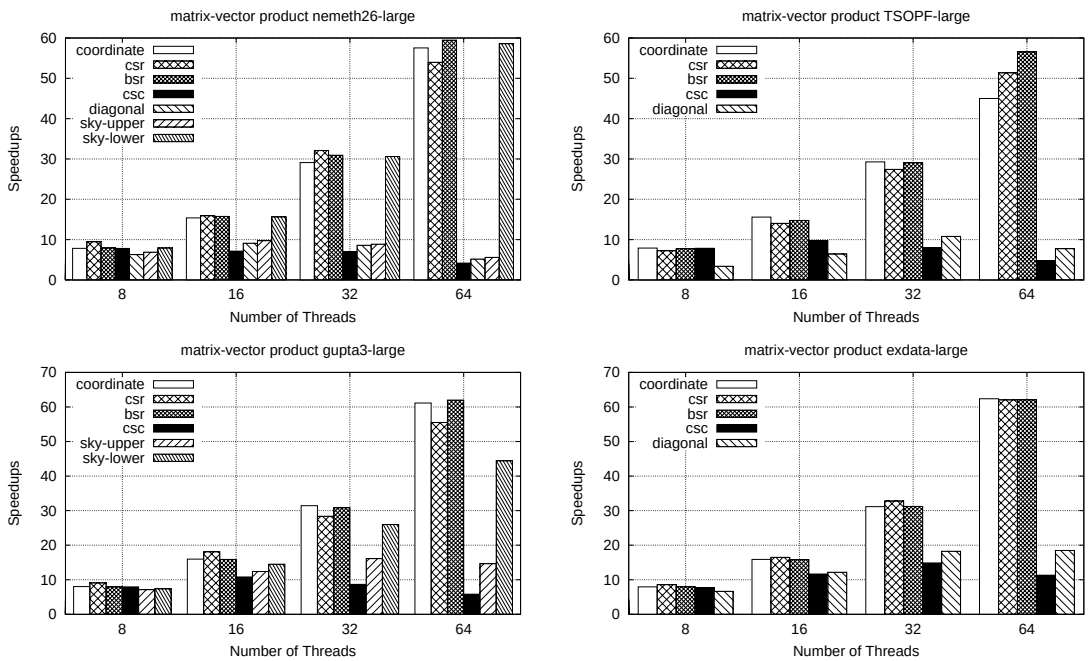


Figure 6: Speedups of the matrix-vector product

 ISBN: 978-84-614-6167-7

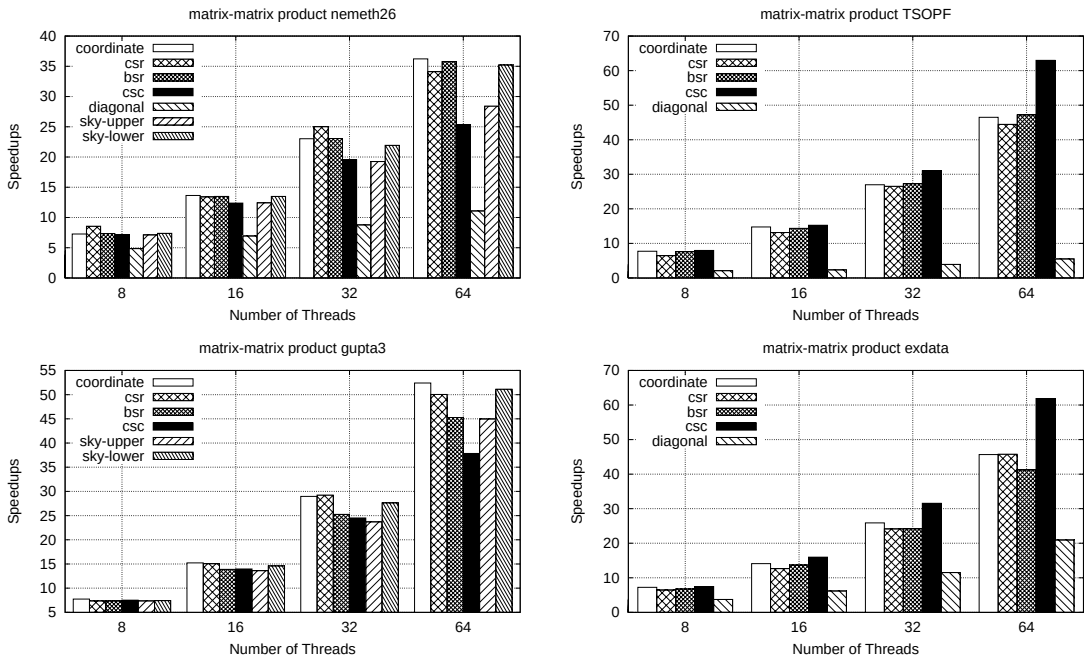


Figure 7: Speedups of the matrix-matrix product

in the Diagonal format is distributed among several threads this cache efficiency decreases, showing significantly poorer scalability.

7 Conclusions

The efficiency of the sparse products is critical for the performance of many applications. In this paper, a PGAS language, UPC, has been used for the parallelization of these computational kernels, as it provides productivity advantages and good data locality explotation, especially on hierarchical architectures such as multicore clusters. The parallel algorithms proposed take into account both the most suitable storage format of the sparse matrix and its influence on the data distributions in order to obtain a good efficiency. The performance evaluation of the routines on a supercomputer has shown the efficiency of the algorithms implemented, achieving speedups of up to 62 for the sparse matrix-vector product and 63 for the sparse matrix-matrix one on 64 cores. Furthermore, it has been assessed the suitability of the combination of different storage formats and workload distributions, depending on the matrix sparsity pattern.

The routines implemented will be included in a UPC sparse BLAS library to extend

 ISBN: 978-84-614-6167-7

the dense counterpart described in [24].

Acknowledgements

This work was funded by Hewlett-Packard (Project "Improving UPC Usability and Performance in Constellation Systems: Implementation/Extensions of UPC Libraries"), the Ministry of Science and Innovation of Spain (Project TIN2010-16735), the Ministry of Education (FPU grant AP2008-01578), and the Spanish network CAPAP-H3 (Project TIN2010-12011-E). We gratefully thank CESGA (Galicia Supercomputing Center) for providing access to the Finis Terrae supercomputer.

References

[1] UPC Consortium. UPC Language Specifications, v1.2, 2005. `http://upc.lbl.gov/docs/user/upc_spec_1.2.pdf`, Last visit: May 2011.

[2] T. El-Ghazawi and F. Cantonnet. UPC Performance and Potential: a NPB Experimental Study. In *Proc. 15th ACM/IEEE Conf. on Supercomputing (SC'02)*, pages 1–26, Baltimore, MD, USA, 2002.

[3] C. Bell, D. Bonachea, R. Nishtala, and K. Yelick. Optimizing Bandwidth Limited Problems using One-Sided Communication and Overlap . In *Proc. 20th Intl. Parallel and Distributed Processing Symp. (IPDPS'06)*, Rhodes Island, Greece, 2006.

[4] C. Barton, C. Casçaval, G. Almási, Y. Zheng, M. Farreras, S. Chatterjee, and J. N. Amaral. Shared Memory Programming for Large Scale Machines. In *Proc. ACM SIGPLAN Conf. on Programming Language Design and Implementation (PLDI'06)*, pages 108–117, Ottawa, Canada, 2006.

[5] D. A. Mallón, G. L. Taboada, C. Teijeiro, J. Touriño, B. B. Fraguela, A. Gómez, R. Doallo, and J. C. Mouriño. Performance Evaluation of MPI, UPC and OpenMP on Multicore Architectures. In *Proc. 16th European PVM/MPI Users' Group Meeting (EuroPVM/MPI'09)*, pages 174–184, Espoo, Finland, 2009.

[6] H. Shan, F. Blagojević, S.-J. Min, P. Hargrove, H. Jin, K. Fuerlinger, A. Koniges, and N. J. Wright. A Programming Model Performance Study using the NAS Parallel Benchmarks. *Scientific Programming*, 18(3-4):153–167, 2010.

[7] S. Williams, L. Oliker, R. W. Vuduc, J. Shalf, K. Yelick, and J. Demmel. Optimization of Sparse Matrix-Vector Multiplication on Emerging Multicore Platforms. In *Proc. 20th ACM/IEEE Conf. on Supercomputing (SC'07)*, Reno, NV, USA, 2007.

 ISBN: 978-84-614-6167-7

[8] S. Liu, Y. Zhang, X. Sun, and R. Qiu. Performance Evaluation of Multithreaded Sparse Matrix-Vector Multiplication using OpenMP. In *Proc. 11th IEEE Intl. Conf. on High Performance Computing and Communications (HPCC'09)*, pages 659–665, Seoul, Korea, 2009.

[9] K. Kourtis, G. I. Goumas, and N. Koziris. Improving the Performance of Multithreaded Sparse Matrix-Vector Multiplication using Index and Value Compression. In *Proc. 37th Intl. Conf. on Parallel Processing (ICPP'08)*, pages 511–519, Portland, OR, USA, 2008.

[10] C. D. Jiogo, P. Manneback, and P. Kuonen. Well Balanced Sparse Matrix-Vector Multiplication on a Parallel Heterogeneous System. In *Proc. 8th IEEE Intl. Conf. on Cluster Computing (CLUSTER'06)*, Barcelona, Spain, 2006.

[11] A. Buluç and J. R. Gilbert. Challenges and Advances in Parallel Sparse Matrix-Matrix Multiplication. In *Proc. 37th Intl. Conf. on Parallel Processing (ICPP'08)*, pages 503–510, Portland, OR, USA, 2008.

[12] Sparse Basic Linear Algebra Subprograms (SparseBLAS) Library. `http://math.nist.gov/spblas`, Last visit: May 2011.

[13] M. Luján, A. Usman, T. L. Freeman, and John R. Gurd. Storage Formats for Sparse Matrices in Java. In *Proc. 5th Intl. Conf. on Computational Science (ICCS'05)*, pages 364–371, Atlanta, GA, USA, 2005.

[14] A. Usman, M. Luján, L. Freeman, and J. R. Gurd. Performance Evaluation of Storage Formats for Sparse Matrices in Fortran. In *Proc. 8th IEEE Intl. Conf. on High Performance Computing and Communications (HPCC'06)*, pages 160–169, Munich, Germany, 2006.

[15] R. Shahnaz, A. Usman, and I. R. Chughtai. Implementation and Evaluation of Parallel Sparse Matrix-Vector Products on Distributed Memory Parallel Computers. In *Proc. 8th IEEE Intl. Conf. on Cluster Computing (CLUSTER'06)*, Barcelona, Spain, 2006.

[16] R. Shahnaz and A. Usman. Blocked-Based Sparse Matrix-Vector Multiplication on Distributed Memory Parallel Computers. *The International Arab Journal of Information Technology*, 8(2):130–136, 2011.

[17] N. Bell and M. Garland. Implementing Sparse Matrix-Vector Multiplication on Throughput-Oriented Processors. In *Proc. 22nd Intl. Conf. on Supercomputing (SC'09)*, Portland, OR, USA, 2009.

[18] M. R. Hugues and S. G. Petiton. Sparse Matrix Formats Evaluation and Optimization on a GPU. In *Proc. 12th IEEE Intl. Conf. on High Performance Computing and Communications (HPCC'10)*, pages 122–129, Melbourne, Australia, 2010.

 ISBN: 978-84-614-6167-7

[19] C. Bell and R. Nishtala. UPC Implementation of the Sparse Triangular Solve and NAS FT, 2004. http://www.cs.berkeley.edu/~rajeshn/pubs/bell_nishtala_spts_ft.pdf, Last visit: May 2011.

[20] J. Dongarra. *Templates for the Solution of Algebraic Eigenvalue Problems: A Practical Guide*, chapter 10. SIAM, 2000.

[21] Finis Terrae Supercomputer. http://www.top500.org/system/9500, Last visit: May 2011.

[22] Berkeley UPC Project. http://upc.lbl.gov, Last visit: May 2011.

[23] The University of Florida Sparse Matrix Collection. http://www.cise.ufl.edu/research/sparse/matrices/, Last visit: May 2011.

[24] J. González-Domínguez, M. J. Martín, G. L. Taboada, J. Touriño, R. Doallo, and A. Gómez. A Parallel Numerical Library for UPC. In *Proc. 15th Intl. European Conf. on Parallel and Distributed Computing (Euro-Par 2009)*, pages 630–641, Delft, The Netherlands, 2009.

 ISBN: 978-84-614-6167-7

Proceedings of the 11th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2011
26–30 June 2011.

Chameleon Hashes in the Forward-Secure ID-Based Setting

Madeline González Muñiz[1] **and Peeter Laud**[2]

[1] *Cybernetica AS, Akadeemia tee 21, Tallinn, Estonia*

[2] *Cybernetica AS, Ülikooli 2, Tartu, Estonia*

emails: madeline@research.cyber.ee, peeter@cyber.ee

Abstract

Due to the possibility of key exposure, forward security in digital signature schemes has been well studied. In the identity-based setting where an entity's public key is that entity's name, our aim is to allow sanitizations in digital signature schemes which provide forward security. After introducing the notion of key-evolving chameleon hash schemes, we present a construction that provides forward-secure collision-resistance and uses non-interactive proofs of knowledge. The public key of the key generation center and all private keys are updated during each time period.

Key words: identity-based cryptography, chameleon hashes, bilinear pairings, sanitizable signatures
MSC 2000: 68P01, 94A60

1 Introduction

Chameleon hashes are collision-resistant functions with a trapdoor for finding collisions [11]. They are particularly useful in sanitizable signature schemes, a variant of digital signature schemes that allow a certain degree of modification to the original message [5, 10]. The sanitizer can create collisions to modify the message using chameleon hashes [1]. An identity-based chameleon hash based on bilinear pairings was proposed by Zhang et al. [15] but suffers from a key exposure problem as shown by Ateniese and de Medeiros [2]. Recently, a new ID-based sanitizable signature scheme in the standard model was proposed by Ming et al. [12]. Our contribution is to adapt the definitions and construction of work by Chen et al. [7] to include forward security. That is, once the intended time period has passed, honest recipients will not be able to create collisions in the chameleon hash.

 ISBN: 978-84-614-6167-7

2 Preliminaries and Definitions

We write $x \xleftarrow{\$} X$ to denote that the value x is uniformly chosen from the set X. Our construction builds on bilinear pairings commonly used in identity-based schemes (such as [4]).

Definition 1 (Bilinear Map) *A pairing is a bilinear map $e : \mathbb{G}_1 \times \mathbb{G}_1 \to \mathbb{G}_2$ where $\mathbb{G}_1$ and $\mathbb{G}_2$ have prime order q and satisfy the following properties:*

1. Bilinearity: $\forall P_1, P_2 \in \mathbb{G}_1$, $\forall \alpha, \beta \in \mathbb{Z}_q^*$, $e(\alpha P_1, \beta P_2) = e(P_1, P_2)^{\alpha\beta}$.
2. Non-Degeneracy: *For any* $P_1 \in \mathbb{G}_1$, $e(P_1, P_2) = 1$ *for all* $P_2 \in \mathbb{G}_1$ *iff* $P_1 = \mathcal{O}$.
3. Computability: *There exists an efficient algorithm to compute* $e(P_1, P_2)$ *from* $P_1, P_2 \in \mathbb{G}_1$.

The security proofs of our constructions depend on the hardness of certain computational problems involving bilinear pairings. Namely, the *bilinear decisional Diffie-Hellman* (BDDH) problem states that it is infeasible to distinguish tuples of the form $(P, \alpha P, \beta P, \gamma P, e(P, P)^{\alpha\beta\gamma})$ from tuples $(P, \alpha P, \beta P, \gamma P, z)$. Here, P is a generator of $\mathbb{G}_1$, α, β, γ are uniformly chosen from $\mathbb{Z}_q^*$, and z is uniformly chosen from $\mathbb{G}_2$. The hardness of BDDH problem implies that the *bilinear computational Diffie-Hellman* (BCDH) problem — given $P, \alpha P, \beta P, \gamma P \in \mathbb{G}_1$, find $e(P, P)^{\alpha\beta\gamma}$ — is hard, too.

In identity-based cryptography, the names of the parties serve as their public keys. There is a trusted party, the key generation center (KGC), that is capable of computing the secret key corresponding to any public key and uses this capability to give to each party its secret key. Secret keys are bound to public keys through the *master public key* known to everybody. The master public key, together with the *master secret key* are generated by KGC as well as the system parameters params.

We will construct a non-interactive key-evolving identity-based chameleon hash scheme with a similar setup to the one modeled by Steinwandt and Suárez Corona [14] in which users update their states during every time period (including the KGC). All users, including the adversary, are assumed to be stateful; we assume that the old state of each user is cleanly erased and thus unrecoverable. After defining these notions, we explore the security model in the next section. For time t, we denote the master public key by P_t and the master secret key by S_t.

Definition 2 (Key-Evolving Identity-Based Chameleon Hash Scheme) *A key-evolving identity-based chameleon hash (KE-ID-CH) scheme is a 6-tuple of probabilistic polynomial-time algorithms $(\mathcal{S}, \mathcal{M}, \mathcal{E}, \mathcal{U}, \mathcal{H}, \mathcal{F})$ such that*

$\mathcal{S}$ *The setup algorithm is run by the* KGC *which upon input security parameter* 1^k *and total number of time periods* T *generates* params, S_0, *and* P_0.

 ISBN: 978-84-614-6167-7

$\mathcal{M}$ *The master key update algorithm is run by the* KGC *which upon input* params *and* S_{t-1} *where* $t-1 < T$ *outputs* S_t*. Then,* S_{t-1} *is erased.*

$\mathcal{E}$ *The extraction algorithm is run by the* KGC *which on input* ID *and* P_t *outputs* D_t^{ID}*, the secret key corresponding to* ID *during time period* t[1]*.*

$\mathcal{U}$ *The user key update algorithm is run by user* ID *which upon input* params *and* D_{t-1}^{ID} *where* $t-1 < T$ *outputs* D_t^{ID}*. Then,* D_{t-1}^{ID} *is erased.*

$\mathcal{H}$ *The hashing algorithm is run by the sender which on input* P_t*, receiving identity* ID*, message* m*, random element* r*, and transaction label* L *outputs the hash value* $h = \mathcal{H}(P_t, \mathsf{ID}, L, m, r)$*.*

$\mathcal{F}$ *The forging algorithm is run by the receiving identity* ID *which on input* $(P_t, \mathsf{ID}, L, m, r, m')$ *and* D_{ID} *outputs* r' *such that* $h = \mathcal{H}(P_t, \mathsf{ID}, L, m, r) = \mathcal{H}(P_t, \mathsf{ID}, L, m', r')$*.*

For algorithms $\mathcal{E}$*,* $\mathcal{H}$ *and* $\mathcal{F}$*, we require that* $t \leq T$*. Furthermore, if* r *is uniformly distributed in finite space* $\mathcal{R}$*, then the distribution of* r' *is computationally indistinguishable from uniform in* $\mathcal{R}$*.*

In the construction that we will propose, a secure pseudorandom bit generator (PRBG) will be used in the KE-ID-CH scheme by the KGC in algorithm $\mathcal{M}$ and by the users in algorithm $\mathcal{U}$ to update the secret keys. A PRBG is a deterministic algorithm which, given a truly random binary sequence of length l, outputs a binary sequence of length $k \gg l$ such that no efficient algorithm can tell it apart from a sequence of uniformly random strings of the same length. A well-known result by Håstad et al. shows that PRBGs may be constructed from any one-way function [9].

In terms of notation (as modeled by Bellare and Yee in [3]), we say that $\mathcal{G} = (\mathcal{G}_{\mathsf{k}}, \mathcal{G}_{\mathsf{n}}, k, T)$ is a (stateful) PRBG which consists of a pair of algorithms and a pair of natural numbers. The probabilistic key generation algorithm $\mathcal{G}_{\mathsf{k}}$ takes no inputs and outputs an initial seed $Seed_0$. The deterministic next step algorithm $\mathcal{G}_{\mathsf{n}}$ runs a maximum of T times. It takes as input $Seed_{t-1}$ and outputs $(Out_t, Seed_t)$ where Out_t is a k-bit block and $Seed_t$ is the seed for the next period.

3 Security Model

A forward-secure ID-based chameleon hash must have collision resistance and indistinguishability as defined below.

[1]We assume that time period t can be extracted from P_t in the scheme.

 ISBN: 978-84-614-6167-7

Definition 3 (Forward-Secure Collision Resistance) *Let $(\mathcal{S}, \mathcal{M}, \mathcal{E}, \mathcal{U}, \mathcal{H}, \mathcal{F})$ be a KE-ID-CH scheme and $\mathcal{A}^{\text{col}}$ a probabilistic polynomial time algorithm. Consider the following attack scenario:*

1. *Run the setup algorithm $\mathcal{S}$ and hand* params *as input to $\mathcal{A}^{\text{col}}$.*
2. *Starting with $t = 1$, the adversary $\mathcal{A}^{\text{col}}$ is successively given access to P_t for each time period t.*
3. *Eventually, $\mathcal{A}^{\text{col}}$ outputs* breakin *and* ID *for time period t or $t = T$ at which point it receives the secret key D_t^{ID}.*
4. *$\mathcal{A}^{\text{col}}$ outputs $(P_{t'}, \mathsf{ID}', L, m, r)$ and $(P_{t'}, \mathsf{ID}', L, m', r')$.*

We say the KE-ID-CH scheme is forward-secure collision resistant *if the success probability $\Pr[\mathcal{H}(P_{t'}, \mathsf{ID}', L, m, r) = \mathcal{H}(P_{t'}, \mathsf{ID}', L, m', r')$ and $t' < t]$ is negligible for all probabilistic polynomial time adversaries $\mathcal{A}^{\text{col}}$.*

The adversary $\mathcal{A}^{\text{col}}$ receives the system parameters and the public key of the system for each time period in order; that is, $\mathcal{A}^{\text{col}}$ is an outsider, not a legitimate user of the system at this point. Then, $\mathcal{A}^{\text{col}}$ selects a time to break into the secret key of a particular user (becoming an insider in the system with the information that an honest user would have stored) and must create a collision in the past. The identity queried during breakin need not be the same as the one used in the forgery. The label L is used in the scheme in order to ensure *key-exposure freeness*. We will briefly describe what this means and how this works in our proposed construction after the proof of Theorem 1.

Definition 4 (Indistinguishability) *Let $(\mathcal{S}, \mathcal{M}, \mathcal{E}, \mathcal{U}, \mathcal{H}, \mathcal{F})$ be a KE-ID-CH scheme and $\mathcal{A}^{\text{ind}}$ a probabilistic polynomial time algorithm. Consider the following attack scenario:*

1. *Run the setup algorithm $\mathcal{S}$ and hand* params *as input to $\mathcal{A}^{\text{ind}}$.*
2. *The adversary $\mathcal{A}^{\text{ind}}$ is given unrestricted access to an extraction oracle $\mathcal{O}^{\mathcal{E}}$ to run $\mathcal{E}(\cdot, \cdot)$.*
3. *Eventually, $\mathcal{A}^{\text{ind}}$ outputs P_t, ID, L, and m.*
4. *Select $b \xleftarrow{\$} \{0, 1\}$ and random element r and do as follows:*
   (a) *If $b = 0$, the adversary $\mathcal{A}^{\text{ind}}$ is given $\mathcal{H}(P_t, \mathsf{ID}, L, m^*, r)$ for randomly chosen m^* of the same length as m.*
   (b) *If $b = 1$, the adversary $\mathcal{A}^{\text{ind}}$ is given $\mathcal{H}(P_t, \mathsf{ID}, L, m, r)$.*

 ISBN: 978-84-614-6167-7

5. $\mathcal{A}^{\text{ind}}$ *outputs bit* b'.

The KE-ID-CH scheme is indistinguishable *if the success probability* $\left|\Pr[b = b'] - \frac{1}{2}\right|$ *is negligible for all probabilistic polynomial time adversaries* $\mathcal{A}^{\text{ind}}$.

The adversary $\mathcal{A}^{\text{ind}}$ receives the system parameters, the public key of the system and an extraction oracle for each time period. After selecting the values that it wants in the hash, a random value is chosen and either the original message is hashed or a random message of the same length. $\mathcal{A}^{\text{ind}}$ wins if it can determine whether the message it selected was the one hashed. In Section 5, we propose a construction that achieves both forward-secure collision resistance and indistinguishability, but first we discuss the shortcomings of the original scheme.

4 Cryptanalysis of Collision Resistance in the Original Scheme

We recall in Figure 1 the scheme by Chen et al. [7] that we will adapt in the next section to include forward security.

$\mathcal{S}$ The KGC takes as input the security parameter 1^k. Then it chooses two groups $\mathbb{G}_1$ and $\mathbb{G}_2$ of prime order q, a bilinear map $e : \mathbb{G}_1 \times \mathbb{G}_1 \to \mathbb{G}_2$, a generator P of $\mathbb{G}_1$, master secret key $S_0 = s$, and sets the public key to $P_0 = sP$. The hash functions in the random oracle is $H_1 : \{0,1\}^* \to \mathbb{G}_1$ (which is used to compute the public keys of the identities).

$\mathcal{E}$ The KGC takes as input ID and S_0 and returns the secret key $D_{\mathsf{ID}} = sQ_{\mathsf{ID}}$ where $Q_{\mathsf{ID}} = H_1(\mathsf{ID})$ is the public key of ID.

$\mathcal{H}$ The sender selects $a \xleftarrow{\$} \mathbb{Z}_q^*$ then returns $r = (aP, e(a \cdot P_0, Q_{\mathsf{ID}}))$ and $h = aP + mH_1(L)$.

$\mathcal{F}$ The receiver selects message m' then sets $a'P = aP + (m - m')H_1(L)$ and $r' = (a'P, e(a'P, D_{\mathsf{ID}}))$.

Figure 1: Chen et al. ID-based chameleon hash scheme without key exposure

Given $(\mathsf{ID}, L, m, r, m')$, the security of the scheme in Figure 1 relies on the difficulty of coming up with r' such that the hash values are equal. The first component of r' is $a'P = aP + (m - m')H_1(L)$ which is easy to compute, but computing the second component $e(a' \cdot sP, Q_{\mathsf{ID}})$ without having D_{ID} implies solving the BCDH problem. Thus, we have that $r' = (a'P, e(a' \cdot sP, Q_{\mathsf{ID}}))$. If a user selects $u \xleftarrow{\$} \mathbb{G}_2$ and returns $r'' = (a'P, u)$, we still have that $h = a'P + m'H_1(L) = (aP + (m - m')H_1(L)) + m'H_1(L) = aP + mH_1(L)$. How can

 ISBN: 978-84-614-6167-7

users other than the sender and receiver be sure that r and r' were correctly formed? A user that can distinguish between r' and r'' can solve the BDDH problem.

If solving the BDDH problem is hard, the issue is solved by attaching a non-interactive zero knowledge (NIZK) proof that the pairing is correctly formed. NIZK protocols enable a prover to non-interactively convince a verifier about the truth of a statement without leaking any information except the fact that the statement is true. Authors Chen et al. address this problem in the resulting chameleon signature scheme that they propose by having a "Deny" protocol. However, when viewed independently from the signature scheme, it is clear that the chameleon hash scheme in Figure 1 is secure if the BDDH problem is easy and the BCDH problem is hard, and this is not what the authors proved. The second component of r and r' does not function as a trapdoor for the receiving identity as intended. As an open problem, one could explore the possibility of constructing an ID-based chameleon hash without key exposure that does not require a NIZK.

5 Proposed Construction

In Figure 2, we use a NIZK proof π that $e(aP_t, Q_{\mathsf{ID}})$ is the solution to the BCDH problem on input $(aP, P_t, Q_{\mathsf{ID}})$. Respectively, π' is the NIZK proof that $e(a'P, s_tQ_{\mathsf{ID}})$ is the solution to the BCDH problem on input $(a'P, P_t, Q_{\mathsf{ID}})$. These NIZK proofs may be done efficiently as shown by Groth and Sahai [8]. Although NIZK proofs are not present in the original scheme that we modified [7], we showed the need for them in Section 4.

The correctness of the forgery follows from the fact that $h = a'P + m'H_1(L) = (aP + (m - m')H_1(L)) + m'H_1(L) = aP + mH_1(L)$ and $r' = e(a' \cdot P_t, Q_{\mathsf{ID}}) = e(a' \cdot s_tP, Q_{\mathsf{ID}}) = e(a'P, s_tQ_{\mathsf{ID}})$. The security of the scheme follows from Theorems 1 and 2. Clearly, one can run $\mathcal{G}$ repeatedly to compute the public key necessary to hash values in future time periods.

Theorem 1 *If the BCDH assumption holds, then the KE-ID-CH scheme in Figure 2 is secure in the sense of forward-secure collision resistance.*

Proof Let $\mathcal{A}^{\text{col}}$ be an adversary in the KE-ID-CH scheme that can forge with non-negligible advantage p according to Definition 3. Adversary $\mathcal{A}^{\text{bcdh}}$ receives challenge $(P, \alpha P, \beta P, \gamma P)$ and must compute $e(P, P)^{\alpha\beta\gamma}$. $\mathcal{A}^{\text{bcdh}}$ simulates the KGC and all random oracle queries and uses $\mathcal{A}^{\text{col}}$ as a subroutine to solve the BCDH problem. To win, $\mathcal{A}^{\text{col}}$ must output $(P_{t'}, \mathsf{ID}, L, m, r)$ and $(P_{t'}, \mathsf{ID}, L, m', r')$ which hash to the same h with the restrictions outlined in Definition 3. In order to introduce the group elements from the challenge above, $\mathcal{A}^{\text{bcdh}}$ will guess when $\mathcal{A}^{\text{col}}$ will query H_1 with the values used in the forgery and set $P_{t'} = \alpha P$, $Q_{\mathsf{ID}} = \beta P$ and $H(L) = \gamma P$. We now present the details of the simulation and how this allows $\mathcal{A}^{\text{bcdh}}$ to solve the BCDH problem.

 ISBN: 978-84-614-6167-7

$\mathcal{S}$ The KGC takes as input the security parameter 1^k and T. Then it chooses two groups $\mathbb{G}_1$ and $\mathbb{G}_2$ of prime order q, a bilinear map $e : \mathbb{G}_1 \times \mathbb{G}_1 \to \mathbb{G}_2$, a generator P of $\mathbb{G}_1$, secret key s_0, and sets the public key to $P_0 = s_0 P$. The hash functions in the random oracle are $H_1 : \{0,1\}^* \to \mathbb{G}_1$ and $H_2 : \{0,1\}^k \to \mathbb{Z}_q^*$. The KGC also chooses a secure PRBG $\mathcal{G} = (\mathcal{G}_\mathsf{k}, \mathcal{G}_\mathsf{n}, k, T)$, runs $\mathcal{G}_\mathsf{k}$ to compute seed $Seed_0$ and sets $S_0 = (s_0, Seed_0)$.

$\mathcal{M}$ The KGC takes as input the master key $S_{t-1} = (s_{t-1}, Seed_{t-1})$, computes $\mathcal{G}_\mathsf{n}(Seed_{t-1}) = (Out_t, Seed_t)$, sets $s_t = H_2(Out_t)s_{t-1} \mod q$ and outputs $S_t = (s_t, Seed_t)$.

$\mathcal{E}$ The KGC takes as input ID and $S_t = (s_t, Seed_t)$ and returns the secret key $D_t^\mathsf{ID} = (s_t Q_\mathsf{ID}, Seed_t)$ and $P_t = s_t P$ to the user where $Q_\mathsf{ID} = H_1(\mathsf{ID})$ is the public key of ID and P_t is the public key of the KGC at time t.

$\mathcal{U}$ The user takes as input the secret key $D_{t-1}^\mathsf{ID} = s_{t-1}Q_\mathsf{ID}$ and $P_{t-1} = s_{t-1}P$, then computes $\mathcal{G}_\mathsf{n}(Seed_{t-1}) = (Out_t, Seed_t)$, and outputs $D_t^\mathsf{ID} = (H_2(Out_t)s_{t-1}Q_\mathsf{ID}, Seed_t) = (s_t Q_\mathsf{ID}, Seed_t)$ and $P_t = H_2(Out_t)P_{t-1} = H_2(Out_t)s_{t-1}P = s_t P$.

$\mathcal{H}$ The sender selects $a \xleftarrow{\$} \mathbb{Z}_q^*$ then returns $r = (aP, e(a \cdot P_t, Q_\mathsf{ID}))$, NIZK proof π and $h = aP + mH_1(L)$.

$\mathcal{F}$ The receiver selects message m' then sets $a'P = aP + (m - m')H_1(L)$, $r' = (a'P, e(a'P, s_t Q_\mathsf{ID}))$ and constructs NIZK proof π'.

Figure 2: Chameleon hash scheme in the forward-secure ID-based setting

PRBG Since $\mathcal{A}^{\text{col}}$ has access to the public keys, $\mathcal{A}^{\text{bcdh}}$ selects $t' \in \{1, \ldots, T'\}$ when it will return $P_{t'} = \alpha P$ where T' is the polynomial number of time periods queried by $\mathcal{A}^{\text{col}}$. $\mathcal{A}^{\text{bcdh}}$ selects a PRBG $\mathcal{G} = (\mathcal{G}_\mathsf{k}, \mathcal{G}_\mathsf{n}, k, T)$ and creates an empty list R. First, $\mathcal{A}^{\text{bcdh}}$ runs $\mathcal{G}_\mathsf{k}$ and stores $Seed_0$. Then, $\mathcal{A}^{\text{bcdh}}$ runs $\mathcal{G}_\mathsf{n}(Seed_{t-1})$ and stores $(Out_t, Seed_t)$ in R; list R is used to compute P_t for $t \neq t'$. If $\mathcal{A}^{\text{col}}$ can efficiently distinguish between a faithful simulation using $\mathcal{G}$ versus a simulation in which we insert αP in place of the public key $P_{t'}$, then $\mathcal{A}^{\text{col}}$ distinguishes between the PRBG chosen and a function that outputs uniformly at random bits of the same length.

Random Oracle Queries $\mathcal{A}^{\text{bcdh}}$ begins with empty lists R_1 and R_2 where it stores the random oracle queries of H_1 and H_2, respectively. $\mathcal{A}^{\text{col}}$ makes a polynomial number of queries of at most q_{H_1} to H_1 and q_{H_2} to H_2. To guess when $\mathcal{A}^{\text{col}}$ will query the values that

 ISBN: 978-84-614-6167-7

it will use to create a forgery, $\mathcal{A}^{\text{bcdh}}$ selects $i \in \{1, \ldots, q_{H_1}\}$ and then $j \in \{1, \ldots, q_{H_1}\} \setminus \{i\}$ and will return βP upon the i^{th} query x_i and γP upon the j^{th} query x_j to H_1.

- H_1 Queries: Except for the i^{th} and j^{th} queries, in simulating H_1 for new query $x \in \{0,1\}^*$, $\mathcal{A}^{\text{bcdh}}$ selects $r_x \xleftarrow{\$} \mathbb{Z}_q$, stores $(x, r_x, r_x P)$, and returns $r_x P$; otherwise, $\mathcal{A}^{\text{bcdh}}$ returns the value $r_x P$ already stored in R_1.

- H_2 Queries: When simulating H_2 for new query $x^* \in \{0,1\}^k$, $\mathcal{A}^{\text{bcdh}}$ selects $r_{x^*} \xleftarrow{\$} \mathbb{Z}_q^*$, stores (x^*, r_{x^*}), and returns r_{x^*}; otherwise, $\mathcal{A}^{\text{bcdh}}$ returns the value r_{x^*} already stored in R_2.

Collisions Let Collision be the event that during the simulation, $\mathcal{A}^{\text{bcdh}}$ stores pairs $(x, r_x, r_x P)$ and $(\widehat{x}, r_{\widehat{x}}, r_{\widehat{x}} P)$ in R_1 (and analogously for R_2) where $x \neq \widehat{x}$ and $r_x = r_{\widehat{x}}$. Whenever the event Collision occurs, $\mathcal{A}^{\text{bcdh}}$ will restart the simulation. As $\mathcal{A}^{\text{bcdh}}$ is polynomially bounded, Collision occurs with negligible probability only, and subsequently we may assume that the event Collision does not occur. By similar reasoning, one may assume that $\mathcal{A}^{\text{col}}$ has queried all random oracle values which make the forgery valid; otherwise, the probability that $\mathcal{A}^{\text{bcdh}}$ selects the same values during verification is also negligible.

Break-in and Forgery $\mathcal{A}^{\text{col}}$ will request breakin for some time period $\tilde{t} > t'$ and will receive $D_{\tilde{t}}^{\mathsf{ID}}$ of which $s_{\tilde{t}} Q_{\mathsf{ID}}$ is computed by $\mathcal{A}^{\text{bcdh}}$ using lists R and R_1 and $Seed_{\tilde{t}}$ comes from $\mathcal{G}$; if $P_{\tilde{t}} = \alpha P$ and $Q_{\mathsf{ID}} = \beta P$, $\mathcal{A}^{\text{bcdh}}$ restarts the simulation.

Let $(P_{t'}, \mathsf{ID}, L, m, r)$ and $(P_{t'}, \mathsf{ID}, L, m', r')$ be the forgery returned by $\mathcal{A}^{\text{col}}$. When $\mathcal{A}^{\text{col}}$ uses $P_{t'} = \alpha P$, $Q_{\mathsf{ID}} = \beta P$ and $H(L) = \gamma P$ to create its forgery, $\mathcal{A}^{\text{bcdh}}$ can solve the BCDH problem. Since $aP + mH_1(L) = a'P + m'H_1(L)$ implies that $(m - m')H_1(L) = a'P - aP$, we can compute the pairing $e\left((m - m')H_1(L), \alpha Q_{\mathsf{ID}}\right) = e(a'P - aP, \alpha Q_{\mathsf{ID}}) = e(a'P, \alpha Q_{\mathsf{ID}})/e(aP, \alpha Q_{\mathsf{ID}})$ because $e(aP, \alpha Q_{\mathsf{ID}})$ and $e(a'P, \alpha Q_{\mathsf{ID}})$ are the second components of r and r', respectively. By raising both sides to $(m - m')^{-1}$ (computed modulus q), one can solve for $e(H_1(L), \alpha Q_{\mathsf{ID}}) = e(\gamma P, \alpha\beta P) = e(P, P)^{\alpha\beta\gamma}$. $\mathcal{A}^{\text{bcdh}}$'s probability of success $\dfrac{p}{(q_{H_1})^2 T'}$ is non-negligible, thus proving the theorem. ∎

Key exposure freeness In some proposed chameleon hash schemes, creating a collision would result in exposing the recipient's secret key (such as Krawczyk and Rabin's scheme [11]) and this is known as the key exposure problem. As shown in the proof, when a collision occurs in the scheme, the pairing $e(H_1(L), \alpha Q_{\mathsf{ID}})$ may be computed for that label L. However, this does not reveal αQ_{ID}. Furthermore, computing $e(H_1(L'), \alpha Q_{\mathsf{ID}})$ for $L' \neq L$ involves solving the BCDH problem when L' has not been used before by ID to create a collision.

 ISBN: 978-84-614-6167-7

Theorem 2 *The KE-ID-CH scheme in Figure 2 is secure in the sense of indistinguishability.*

In order to distinguish between $h = aP + mH_1(L)$ and $h' = aP + m'H_1(L)$, $\mathcal{A}^{\text{ind}}$ would have to correctly guess the first component of $r = (aP, e(a \cdot s_t P, Q_{\mathsf{ID}}))$ which is not given and chosen uniformly at random. As this occurs with negligible probability $1/q$, the scheme has the indistinguishability property.

6 Conclusion and Open Problems

In this work, we constructed a provably secure forward-secure ID-based chameleon hash scheme based on a construction that is free of key exposure. Chameleon hashes have often been used in sanitizable signature schemes, but we are the first to propose a scheme in a setting which involves forward security, as far as we know. Our forward-secure construction is non-interactive but requires that the master public key be updated.

Naturally, one wonders if it is possible to have a forward-secure key-evolving chameleon hash function that is non-interactive and has a fixed master public key. In many forward-secure schemes (for example Canetti et al.'s [6] forward-secure encryption), the user updates its internal state non-deterministically. Does this imply that the sender who is hashing needs to interact with the receiving identity?

On another direction, an interesting problem is whether it is possible to construct an attribute-based chameleon hash scheme with the conditions that follow. The hash would be a function of the message, attribute authority's public key and attributes. A user with the appropriate threshold number of attribute secret keys should be able to compute a trapdoor to the chameleon hash. Also, one would like to limit the interaction between each user and the attribute authority to an initial exchange as is typically done (for example in fuzzy ID-based encryption by Sahai and Waters [13]).

Acknowledgements

This research was supported by the European Regional Development Fund through the Estonian Center of Excellence in Computer Science, EXCS, and by the Estonian Science Foundation, grant #8124.

References

[1] Giuseppe Ateniese, Daniel H. Chou, Breno de Medeiros, and Gene Tsudik. Sanitizable signatures. In Sabrina de Capitani di Vimercati, Paul Syverson, and Dieter Gollmann,

editors, *Computer Security ESORICS 2005*, volume 3679 of *Lecture Notes in Computer Science*, pages 159–177. Springer Berlin / Heidelberg, 2005.

[2] Giuseppe Ateniese and Breno de Medeiros. Identity-based chameleon hash and applications. In Ari Juels, editor, *Financial Cryptography*, volume 3110 of *Lecture Notes in Computer Science*, pages 164–180. Springer Berlin / Heidelberg, 2004.

[3] Mihir Bellare and Bennet Yee. Forward-security in private-key cryptography. In Marc Joye, editor, *Topics in Cryptology CT-RSA 2003*, volume 2612 of *Lecture Notes in Computer Science*, pages 1–18. Springer Berlin / Heidelberg, 2003.

[4] Dan Boneh and Matt Franklin. Identity-based encryption from the weil pairing. In Joe Kilian, editor, *Advances in Cryptology CRYPTO 2001*, volume 2139 of *Lecture Notes in Computer Science*, pages 213–229. Springer Berlin / Heidelberg, 2001.

[5] Christina Brzuska, Marc Fischlin, Tobias Freudenreich, Anja Lehmann, Marcus Page, Jakob Schelbert, Dominique Schröder, and Florian Volk. Security of sanitizable signatures revisited. In Stanislaw Jarecki and Gene Tsudik, editors, *Public Key Cryptography PKC 2009*, volume 5443 of *Lecture Notes in Computer Science*, pages 317–336. Springer Berlin / Heidelberg, 2009.

[6] Ran Canetti, Shai Halevi, and Jonathan Katz. A forward-secure public-key encryption scheme. In Eli Biham, editor, *Advances in Cryptology EUROCRYPT 2003*, volume 2656 of *Lecture Notes in Computer Science*, pages 646–646. Springer Berlin / Heidelberg, 2003.

[7] Xiaofeng Chen, Fangguo Zhang, Willy Susilo, Haibo Tian, Jin Li, and Kwangjo Kim. Identity-based chameleon hash scheme without key exposure. In Ron Steinfeld and Philip Hawkes, editors, *Information Security and Privacy*, volume 6168 of *Lecture Notes in Computer Science*, pages 200–215. Springer Berlin / Heidelberg, 2010.

[8] Jens Groth and Amit Sahai. Efficient non-interactive proof systems for bilinear groups. In Nigel Smart, editor, *Advances in Cryptology - EUROCRYPT 2008*, volume 4965 of *Lecture Notes in Computer Science*, pages 415–432. Springer Berlin / Heidelberg, 2008.

[9] Johan Håstad, Russell Impagliazzo, Leonid A. Levin, and Michael Luby. A pseudorandom generator from any one-way function. *SIAM Journal on Computing*, 28(4):1364–1396, 1999.

[10] Marek Klonowski and Anna Lauks. Extended sanitizable signatures. In Min Rhee and Byoungcheon Lee, editors, *Information Security and Cryptology ICISC 2006*, volume 4296 of *Lecture Notes in Computer Science*, pages 343–355. Springer Berlin / Heidelberg, 2006.

 ISBN: 978-84-614-6167-7

[11] Hugo Krawczyk and Tal Rabin. Chameleon signatures. In *Proceedings of the Network and Distributed System Security Symposium, NDSS 2000, San Diego, California, USA*. The Internet Society, 2000.

[12] Yang Ming, Xiaoqin Shen, and Yamian Peng. Identity-based sanitizable signature scheme in the standard model. In Rongbo Zhu, Yanchun Zhang, Baoxiang Liu, and Chunfeng Liu, editors, *Information Computing and Applications*, volume 105 of *Communications in Computer and Information Science*, pages 9–16. Springer Berlin Heidelberg, 2011.

[13] Amit Sahai and Brent Waters. Fuzzy identity-based encryption. In Ronald Cramer, editor, *Advances in Cryptology EUROCRYPT 2005*, volume 3494 of *Lecture Notes in Computer Science*, pages 557–557. Springer Berlin / Heidelberg, 2005.

[14] Rainer Steinwandt and Adriana Suárez Corona. Identity-based non-interactive key distribution with forward security. *Designs, Codes and Cryptography*, pages 1–14, 2011.

[15] Fangguo Zhang, Reihaneh Safavi-Naini, and Willy Susilo. Id-based chameleon hashes from bilinear pairings. Cryptology ePrint Archive, Report 2003/208, 2003. `http://eprint.iacr.org/`.

 ISBN: 978-84-614-6167-7

Proceedings of the 11th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2011
26–30 June 2011.

A Numerical Study of Viscoelastic Strings Using a Discrete Model

G. González-Santos[1] and C. Vargas-Jarillo[2]

[1] *Departamento de Matemáticas, Escuela Superior de Física y Matemáticas del IPN, México., D.F. 07738, México.*

[2] *Departamento de Control Automático, , CINVESTAV-IPN, A. P. 14-740, México., D. F. 07000, México.*

emails: gsantos@esfm.ipn.mx, cvargas@ctrl.cinvestav.mx

Abstract

The vibrating string is a problem that has been considered in many areas of science and engineering. We propose a quasimolecular simulation of a nonlinear viscoelastic string by means of a molecular type approach. The discrete model is an array of masses connected by Kelvin units with nearest-neighbor interactions.

Key words: Discrete nonlinear string, Nonlinear vibrations, Viscoelasticity, Discrete mechanics
MSC 2000: AMS 65L04, 65L05, 70F10

1 Introduction

In this work we propose the simulation of a nonlinear viscoelastic string by means of a molecular type approach [6]-[8]. The string is composed of a finite array of particles joined by massless Kelvin units. We study the string in two dimensions for several initial and boundary conditions. The interest of vibrations in strings has hold attention since the seventeen century [4]. Other studies have been done in [1], [2], and [12]. We study the string, allowing the particles to move in two dimensions; therefore the nonlinearity is introduced in our case by the geometry of the problem.

 ISBN: 978-84-614-6167-7

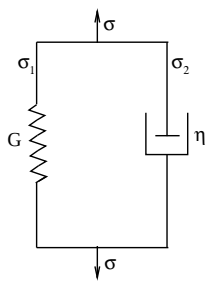


Figure 1: The Kelvin's unit

2 Model formulation

The physical model of a viscoelastic string consists of N particles, $P_1, P_2, ..., P_N$ with masses $m_1, m_2, ..., m_N$ respectively. The number of Kelvin units, M, depends on the boundary conditions; when both ends are fixed there are $M = N + 1$ and $M = N$ if the string has only a free end. The Kelvin model is a two-element model consisting of a linear spring element and a linear viscous dashpot connected in parallel as shown in Fig. 1. The spring and dashpot have the following stress-strain relation.

$$\sigma_1 = G\epsilon, \tag{1}$$
$$\sigma_2 = \eta\dot{\epsilon}, \tag{2}$$

where σ is the stress (force per unit area) applied to the Kelvin unit, ϵ is the strain (deformation per unit length), G is the Young's modulus and η is the viscosity coefficient. Since both elements are connected in parallel, the total stress stress is given by

$$\sigma = \sigma_1 + \sigma_2. \tag{3}$$

From (1)-(3) follows the equation between the stress σ and strain ϵ:

$$\dot{\epsilon} + \frac{G}{\eta}\epsilon = \frac{\sigma}{\eta} \tag{4}$$

The initial lengths of the Kelvin's units are $l_1, l_2, ..., l_M$ and $G_1, G_2, ..., G_M$, $\eta_1, \eta_2, ..., \eta_M$ are the corresponding Young's and viscosity coefficients. Fig. 2 shows four consecutive masses of a viscoelastic string in two dimensions. The variable at time t are:

 ISBN: 978-84-614-6167-7

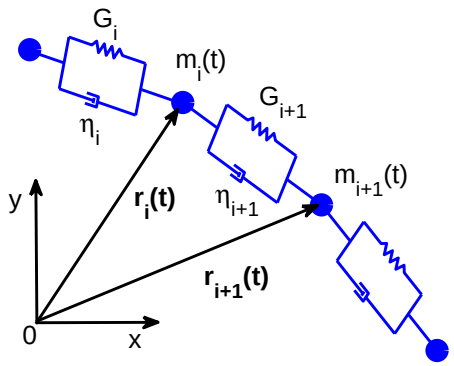


Figure 2: Small segment of a viscoelastic string

$\mathbf{r}_i$ Coordinate of the particle P_i,
$\mathbf{r}_{i,j}$ Vector $\mathbf{r}_i - \mathbf{r}_j$,
$\dot{\mathbf{r}}_i$ Velocity of the particle P_i,
$\ddot{\mathbf{r}}_i$ Aceleration of the particle P_i,
$\mathbf{F}_i^*$ Force acting on P_i due to nearest neighbor particles,
$\mathbf{f}_i$ long range force acting on particle P_i (like gravity), and
$\mathbf{F}_i$ Total force acting on P_i
for $i = 1, 2, ..., N$.

The force $\mathbf{F}_i^*$ exerted on the particle P_i by the Kelvin's units i and $i+1$ is given by:

$$\mathbf{F}_i^* = -\left[k_i(r_{i,i-1} - l_i) + \eta_i \dot{r}_{i,i-1}\right] \frac{\mathbf{r}_{i,i-1}}{r_{i,i-1}} + \left[k_{i+1}(r_{i,i+1} - l_i) + \eta_{i+1}\dot{r}_{i,i+1}\right] \frac{\mathbf{r}_{i,i+1}}{r_{i,i+1}},$$

where $k_i = l_i G_i$ and $r_{i,j} = \sqrt{(x_j - x_i)^2 + (y_j - y_i)^2}$ is the Euclidian distance between the particles P_i and P_j. This introduces the nonlinearity in our problem since we are allowing vibrations in both directions. In contrast to the usual model where the particles are constrained to move only along the vertical axes, therefore, the total force acting upon the particle P_i is

$$\mathbf{F}_i = \mathbf{F}_i^* + \mathbf{f}_i.$$

The acceleration of P_i at time t is related to the force by a discrete Newton's Law:

$$\mathbf{F}_i = m_i \ddot{\mathbf{r}}_i \tag{5}$$

We recall that the acceleration is the second derivative respect to time, thus (5) is a nonlinear system of 2N second order differential equations. In general this system can not be

 ISBN: 978-84-614-6167-7

solved analytically from initial positions and velocities, therefore it must be solved numerically.
Since the velocity, at time t of the particle P_i is $\dot{\mathbf{r}}_i$ we can determine the kinetic energy of the string by

$$T(t) = \frac{1}{2}\sum_{i=1}^{N} m_i \parallel \dot{\mathbf{r}}_i \parallel^2 . \tag{6}$$

The potential energy $V(t)$ is found by considering the increase of length of each Kelvin's unit. The $i-th$ unit has increased or decreased its length from l_i to $r_{i,i-1}$. Therefore, we have done an amount of work $k_i(r_{i,i-1} - l_i)$. Summing up for all the Kelvin's units of the string, we obtain the potential energy of the string, at time t:

$$V_k = \frac{1}{2}\sum_{i=1}^{M} k_i r_{i,i-1}^2. \tag{7}$$

When a long range (gravity) and viscous force are present two additional terms must be added to (7):

$$\sum_{i=1}^{M} m_i y_i g + \sum_{i=1}^{M} \eta_i \dot{r}_{i,i-1} r_{i,i-1}^2.$$

3 Transversal vibrations of a string

We use the discrete model to analyze two viscoelastic strings. In this section we analyze the transversal vibrations of a horizontal string with both ends are fixed.
The system (5) was solved numerically by using the fortran subroutines DDRIV2 [9]. The relative accuracy in the all solution components was taken equal to $1E-6$.

We consider a viscoelastic string with fixed ends at the same height. The mass and the length of the string are $L_c = M_c = 1$. The discrete model consists of $N = 64$ particles of mass $m_i = M_c/N$ for $i = 1, 2, ..., N$ and $M = N + 1$ Kelvin's units of initial length $l_i = L_c/M, i = 1, 2, ..., M$. The gravity is equal to zero, $g = 0$. We consider that the initial position of the string is a single sine wave; the positions of the particles are given by:

$$\mathbf{r}_i(0) = [L_i, a\sin(\pi L_i/L_c)], i = 1, 2, ..., N, \tag{8}$$

and the initial velocity is zero, $\dot{\mathbf{r}}_i = 0$, for $i = 1, 2, ..., N$. The x-position of the i-th particle is given by:

$$L_i = \sum_{j=1}^{i} l_j.$$

 ISBN: 978-84-614-6167-7

The string behavior depends on the quotient $\alpha = a/L_c$; when this values is small, $\alpha < 1.5E-2$, the string oscillates in this mode indefinitely and its amplitude varies periodically in time. In this case the string practically shows a linear behavior. When α increases the nonlinearity of the string appears; higher modes of oscillation are present in the string evolution. In this case the energy of the system is shared with higher modes. For example, when the Young's coefficients are equal for the all Kelvin's units, $G_i = 1/l_i$, and the viscosity coefficients are $\eta_i = 0, i = 1, 2, ..., N$ the evolution of the string can be seen in Fig. 3.

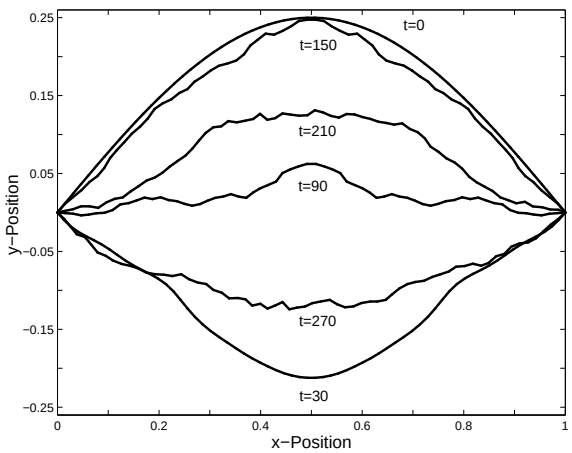


Figure 3: Vertical displacement of the string at several times. $L_c = 1, M_c = 1, \alpha = 0.25, G_i = 1/l_i, \eta_i = 0$, for $i = 1, 2, ..., N$.

The kinetic and potential energy for this string are shown in Fig. 4. Both energies are almost periodic; initially the kinetic energy is zero and it increases as the string moves to the middle position. After this time the kinetic energy decreases as the string moves to the bottom position. The kinetic energy shows a similar behavior when the string moves from the bottom to the top extreme. Both energies show high frequency oscillations of small amplitude due mainly to the oscillations in the x-component. The total energy of the string increases slowly after $t = 100$ due to the numerical method.

Fermi, Pasta and Ulam in 1955 [5], proposed a model for a nonlinear string, where the nonlinearity was given as quadratic power of the displacement in the force term. They analyzed the string behavior with the initial condition given by (8). They found that the total energy of the string was concentrated at only the first modes. In our case we introduced a nonlinearity due to the geometry, and we observe energy in the modes 1, 3, 5 and 7 and there is not energy for higher odd modes. When the viscosity is nonzero the evolution of

 ISBN: 978-84-614-6167-7

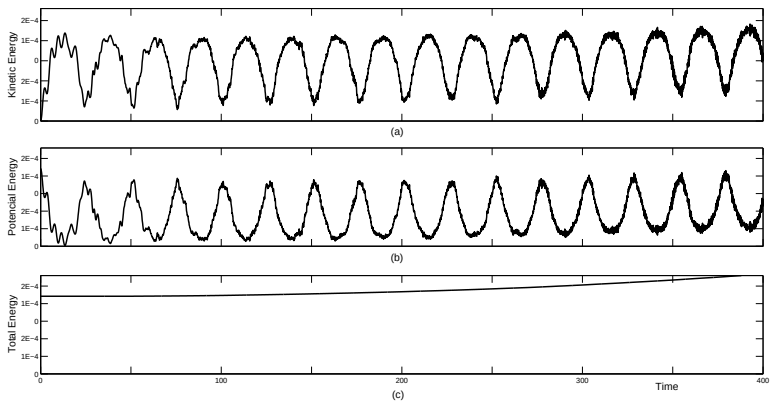


Figure 4: Kinetic, potential an total energy of the string during the first 400 units of time.

the string still shows higher modes but the high frequency oscillations disappear. Figure 5 shows the y-displacement of the viscoelastic at the same times as in Fig. 3.

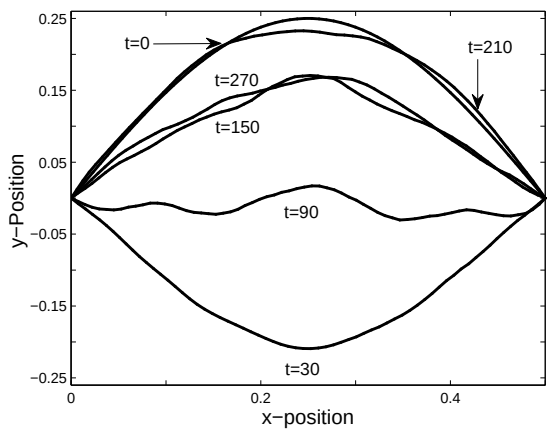


Figure 5: Vertical displacement of the viscoelastic string at several times. $L_c = 1, M_c = 1, \alpha = 0.25, \eta_i = 0.1$, for $i = 1, 2, ..., N$.

The kinetic, potential and total energy lessen to zero and the speed of convergence depends on the viscosity; bigger values of η produces faster convergence. Figure 6 shows the energies when the viscosity coefficients η_i are equal to 0.1 for all the Kelvin's units.

 ISBN: 978-84-614-6167-7

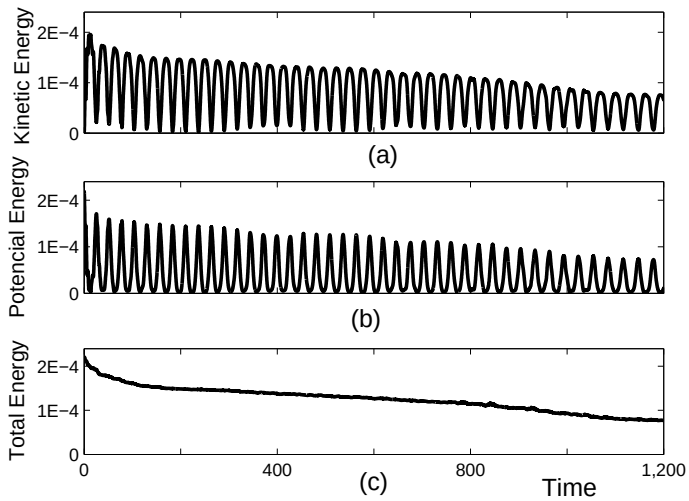


Figure 6: Kinetic, potential an total energy of the string during the first 1200 units of time.

4 The vertical viscoelastic string

Now we consider a vertical string with the upper end fixed and the other end free. The mass and the length of the string are $L_c = M_c = 1$. The discrete model of the hanging string consists of $N = 64$ particles with masses $m_i = M_c/N$ for $i = 1, 2, ..., N$. The Kelvin's units have the same length, $l_i = L_c/N$ for $i = 1, 2, ..., N$, initially. The equilibrium position of the vertical string is obtained by solving the linear system:

$$\mathbf{K}\mathbf{y}^{eq} = -(\mathbf{T}\mathbf{I} + g\mathbf{M}\mathbf{e}),$$

where

$$\mathbf{K} = \begin{pmatrix} -(k_1+k_2) & k_2 & & \\ k_2 & -(k_2+k_3) & k_3 & \\ & \ddots & \ddots & k_N \\ & & k_N & -k_N \end{pmatrix},$$

$$\mathbf{T} = \begin{pmatrix} k_1 & -k_2 & & \\ & k_2 & -k_3 & \\ & & \ddots & -k_N \\ & & & k_N \end{pmatrix},$$

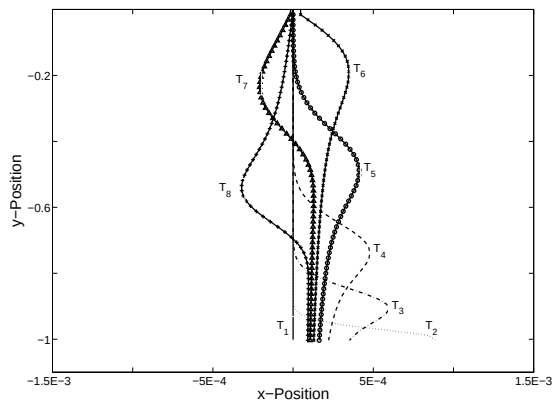
y $\mathbf{M} = diag(m_1, m_2, ..., m_N), \mathbf{l} = (l_1, l_2, ..., l_N)$ y $\mathbf{e} = (1, 1, ..., 1)^t$.

 ISBN: 978-84-614-6167-7

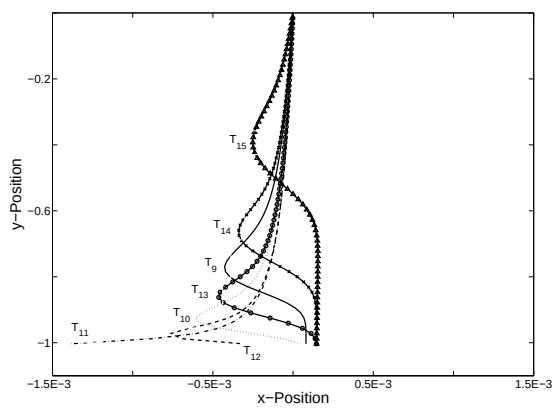
4.1 Vertical string with initial velocity

In this numerical experiment we consider $\eta_i = 0$ for $i = 1, 2, ..., N$, and the Young's coefficients are $G_i = 1/l_i$. The string is at equilibrium initially and the free end, particle N, receives an horizontal impulse of magnitude $\beta > 0$. Thus the initial conditions are:

$$\begin{aligned} \mathbf{r}_i(0) &= (0, y_i^{eq}), \\ \dot{\mathbf{r}}_i(0) &= (0,0),\ for\ i = 1, 2, ..., N-1,\ and \\ \dot{\mathbf{r}}_N(0) &= (\beta, 0),\ \beta = 2.5E-3. \end{aligned}$$

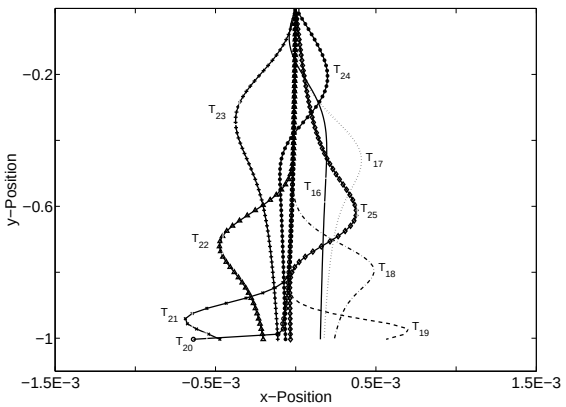


(a)



(b)

 ISBN: 978-84-614-6167-7



(c)

Figure 7: Position of the vertical string at $t_k = (k-1)\Delta t$. (a) A-B: $k = 1-8, \Delta t = 1.25$, (b)B-C: $k = 9-15, \Delta t$ [illegible]

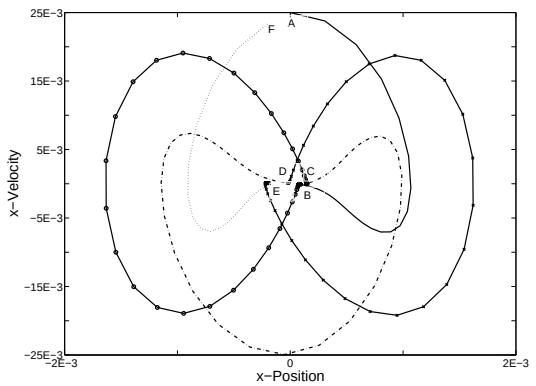


Figure 8: The phase portrait of the x-component of the free end of the string.

The evolution of the vertical string and the phase portrait of the x-component of the free end are shown in Fig. 7 and 8 respectively.
The initial impulse produces a sudden displacement to the right of the free end and then it returns quickly near to equilibrium position, without swinging to the left, and remains there for a little while. This behavior is shown in figure 7(a). This stage, in the phase portrait, corresponds from point A to B along the curve in Fig. 8.
Next to the point B we observe a sudden jump to the left of the free end return again near to the equilibrium position as is shown in Fig. 7(b). This stage corresponds from the point B to C along the curve in Fig. 8.
Immediately after we observe a whip of the free end; it oscillates quickly from to left to the

 ISBN: 978-84-614-6167-7

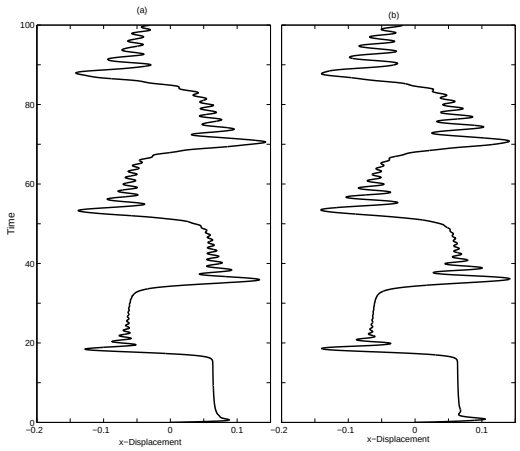


Figure 9: Evolution of free end (x-coordinate) of the viscoelastic string. (a) $\eta = 0$, (b) $\eta = 1000$. $N = 32, L_c = 1, M_c = 1, k_i = 1$, for $i = 1, 2, ..., N$.

right and finish again near the equilibrium position, see Fig. 7(c). This stage starts at the point C of the curve in Fig. 8 and finish at point D. After reaching this point the free end suddenly jump to the right and return again near the equilibrium position. Finally the free end arrives near to its initial condition (point F of the phase portrait); zero displacement and velocity β.
In our case the time between two consecutive whips is not constant in comparison with Bailey [2] who shows that for a inextensible string time between two consecutive whips is constant and it has been accurately predicted by the solution of an ordinary differential equation

When the viscosity is nonzero the evolution of the free end is similar to the elastic case ($\eta = 0$); the evolution of the x-component in both case is practically the same, see Fig. 9. A different behavior shows the y-component; while the amplitude of the oscillations of the y-component, in the elastic case, does not show any tendency as the time increases, the amplitude of y-component, in the viscoelastic case, tends to zero. Figure 10 shows this behavior.

References

[1] S. S. Antman., *The Equations for Large Vibration Strings*, The American Mathematical Montly **87(5)** (1980) 359–370.

 ISBN: 978-84-614-6167-7

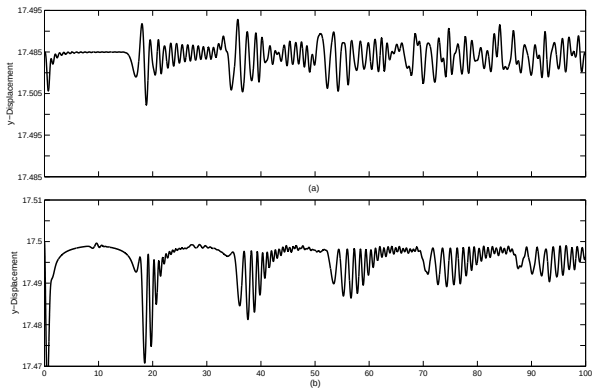


Figure 10: Evolution of free end (y-coordinate) of the viscoelastic string. (a) $\eta = 0$, (b) $\eta = 1000.N = 32, L_c = 1, M_c = 1, k_i = 1$, for $i = 1, 2, ..., N$.

[2] H. Bailey., *Motion of a Hanging Chain After the Free End is Given an Initial Velocity,* Am. J. Phys. **68(8)** (2000)

[3] U. Bottazzini., *The Higher Calculus: A History of Real and Complex Analysis from Euler to Weierstrass.*, Springer-Verlag, New York, 1986.

[4] J. T. Cannon and S. Dostrovsky., *The Evolution of Dynamics Vibration Theory from 1687 to 1742.*, Springer-Verlag , New York, 1981.

[5] E. Fermi, J. Pasta and S. Ulam., *Studies of nonlinear problems*, In Alan C. Newell, editor, Proceedings of the Summer Seminar, sponsored by the America Mathematical Society and Society for Industrial and Applied Mathematics, held at Clarkson College of Technology, Potsdam, N. Y., 1972, pages, 143-156. American Mathematical Society, Providence , R. I., 1974. Los Alamos Sci. Lab. Rep. No. LA-1940 , 1955.

[6] D. Greenspan., *Discrete, Nonlinear String Vibrations*, The Computer Journal, **13(2)** (1970) 195–201.

[7] D. Greenspan., *Computer Simulation of Transverse String Vibrations*, BIT **11** (1971) 399-408.

[8] D. Greenspan., *Numerical Solution of Ordinary Differential Equations: For classical, relativistic and nano systems*, Wiley-VCH 2006.

[9] D. Kahaner, C. Moler and S. Nash., *Numerical Methods and Software*, Prentice Hall, 1989.

[10] H. Levine., *Unidirectional Wave Motions*, North-Holland Publishing Company, New York, 1978.

 ISBN: 978-84-614-6167-7

[11] A. J. Roberts., *A One-Dimensional Introduction to Continuum Mechanics*, World Scientific, Singapore, 1994.

[12] D. Yong.,*Strings, Chains and Ropes*, SIAM REVIEW **48(4)** (2006) 771–781.

 ISBN: 978-84-614-6167-7

Proceedings of the 11th International Conference on Computational and Mathematical Methods in Science and Engineering, CMMSE 2011 26–30 June 2011.

On parallelizing a bi-blend optimization algorithm

J.F.R. Herrera[1], L.G. Casado[1], I. García[2] and E.M.T. Hendrix[2]

[1] *Department of Computer Architecture and Electronics, University of Almería*

[2] *Department of Computer Architecture, University of Málaga*

emails: juanfrh@ual.es, leo@ual.es, igarciaf@uma.es, eligius.hendrix@wur.nl

Abstract

Blending algorithms aim for a solution to the problem of determining the mixture of raw materials in order to obtain a cheap and feasible recipe. The mixture has linear and quadratic constraints to establish the correctness of the final product. The aim of the algorithm is to provide a Pareto solution consisting of minimizing the cost and the number of raw materials involved in the mixture. Manufacturers produce several products from a given set of raw materials. Scarcity in the availability of materials can happen. This scarcity introduces capacity constraints that change the Pareto solution. The authors have developed branch-and-bound (B&B) algorithms to solve this type of blending problems in which the computational complexity increases with the dimension of the problem. Due to this complexity, the authors only addressed the bi-blending problem, where two products are designed.

The bi-blending problem is more difficult than the blending problem because apart from the fact that each product must satisfy its design constraints, it also extends the Pareto front to two products and takes into account the availability of materials. A final combination between all B&B solutions of product one and product two has to be performed to remove combinations of recipes that are shown not to be feasible. The set of left over combinations can be used as the input data for a second execution when more accurate results are requested. The computational cost of the combination phase will depend on the number of elements of the final feasible set for each product.

Here, we study the computational cost of the different phases of the sequential bi-blending algorithm and provide threaded versions for the most time consuming ones. We have carried out experiments on an eight-core shared memory machine, using a small-medium size problem to avoid very large execution times. Experiments show that parallel computation will be a necessary tool to do an exhaustive search for larger dimensional and n-blending problems.

Key words: Shared memory, parallel processors, multi-threaded, branch-and-bound, global optimization.

MSC 2000: 68W10, 65Y05, 90C26, 65G20.

1 Introduction

Finding a cheap robust recipe for a blending problem that satisfies quadratic design requirements is a hard problem. In practice, companies are also dealing with so-called multi-blending problems where the same raw materials are used to produce several products. Descriptions from practical cases can, among others, be found in [1], and [3]. This complicates the search process for feasible and optimal robust solutions if we intend to guarantee the optimality and robustness of the final solutions.

Exhaustive search for a blending algorithm and its components are described in [5, 6, 10], while a bi-blending approach appears in [11].

Based on previous articles, Section 1.1 describes the blending problem and Section 1.2 defines the blending problem to obtain two mixture designs (bi-blending). Section 2 describes the sequential version of the bi-blending algorithm and Section 3 describes its parallel model. Section 4 shows the computational results and Section 5 summarizes the conclusions and future work.

1.1 Blending problem

The blending problem we researched is described in [10] as a semi-continuous quadratic mixture design problem (SQMDP). Here we summarize the main characteristics of this problem.

Variables x_i represent the fraction of material i in a mixture x. The set of possible mixtures is mathematically defined by the unit simplex

$$S = \left\{ x \in \mathbb{R}^n \ : \ \sum_{i=1}^{n} x_i = 1.0; \ x_i \geq 0 \right\}. \tag{1}$$

where n denotes the number of raw materials.

The objective is to find a recipe x that minimizes the cost of the material, $f(x) = c^T x$, where vector c gives the cost of the raw materials. Not only should the cost of the material be minimised, but also the number of raw materials in the mixture x given by $\sum_{i=1}^{n} \delta_i(x)$, where

$$\delta_i(x) = \begin{cases} 1 & \textbf{if} \quad x_i > 0, \\ 0 & \textbf{if} \quad x_i = 0. \end{cases}$$

The semi-continuity of the variables is due to a minimum acceptable dose (md) that the practical problems reveal, i.e. either $x_i = 0$ or $x_i \geq md$. The number of resulting sub-simplices (faces) is

$$\sum_{t=1}^{n} \binom{n}{t} = 2^n - 1,$$

where t denotes the number of raw materials in each sub-simplex. All points x in an initial simplex $P_u, \ u = 1, \ldots, 2^n - 1$, are mixtures of the same group of raw materials. The index u represents the group of raw materials corresponding to initial simplex P_u:

$$u = \sum_{i=1}^{n} 2^{i-1} \delta_i(x), \quad \forall x \in P_u.$$

 ISBN: 978-84-614-6167-7

Recipes have to satisfy certain requirements. For relatively simple blending problems, the bounds or linear inequality constraints (see [1], [3], and [16])

$$h_i(x) \leq 0; \quad i = 1, \ldots, l,$$

define the design space $X \subset S$.

In practice, however, quadratic requirements appear [6, 10]. Quadratic constraints are written as

$$g_i(x) = x^T A_i x + b_i^T x + d_i \leq 0; \quad i = 1, \ldots, m,$$

in which A_i is a symmetric n by n matrix, b_i is an n-vector and d_i is a scalar. Feasible space, according to quadratic constraints, is defined as

$$Q = \{x \in S \,:\, g_i(x) \leq 0;\ i = 1, \ldots, m\}.$$

Moreover, the design to have an ε-robustness with respect to the quadratic requirements in order to maintain the feasibility of the result when small variations in the mixture appear. One can define robustness $R(x)$ of a design $x \in Q$ with respect to Q as

$$R(x) = \max\{R \in \mathbb{R}^+ : (x + r) \in Q,\ \forall r \in \mathbb{R}^n,\ \|r\| \leq R\}.$$

[6] describes tests based on so-called infeasibility spheres that identify areas where an ε-robust solution cannot be located. [10] describes a B&B algorithm to solve SQMDP using rejection tests based on linear, quadratic and robustness constraints.

The problem of finding the best robust recipe becomes a Global Optimization (GO) problem for which a guaranteed optimal solution is hard to obtain, because it can have several local optima and the feasible area may be nonconvex and even consist of several compartments.

A threaded version of the B&B blending (SQMDP) algorithm was presented in [4], with an Asynchronous Multiple Pool scheme [9], following a similar strategy to the one used in a parallel Interval Global Optimization algorithm (Local-PAMIGO), which is published in [7].

1.2 Bi-blending problem

In this paper, our focus is on parallelizing the algorithm that finds the best recipes if one wants to design two mixture products that share raw materials. Due to capacity constraints and availability of raw materials, a bigger dosage of other ingredients could be needed. The recipe that looks for the best for one product is not necessarily the best when designing both products simultaneously.

As described in [11], each product has its own demand and quality requirements consisting of design constraints. Here we summarize the main characteristics of the problem.

We identify an index j for each product with demand D_j. The amount of available raw material i is given by B_i. Now, the main decision variable $x_{i,j}$ is the fraction of raw material i in recipe of product j.

 ISBN: 978-84-614-6167-7

Let $x_{*,j}$ represent column j of matrix $\mathbf{x}$ of decision variables. Then we define linear restrictions per product: $x_{*,j} \in X_j$; and quadratic requirements per product: $x_{*,j} \in Q_j = \{y \in S : g_i(y) \leq 0;\ i = 1, \ldots, m_j\}$.

In principle, all final products can make use of all n raw materials; $x_{*,j} \in \mathbb{R}^n,\ j = 1, 2$. This means that $x_{i,1}$ and $x_{i,2}$ denote fractions of the same ingredient for products 1 and 2. The main restrictions that give the bi-blending problem the "bi" character are the capacity constraints:

$$\sum_{j=1}^{2} D_j x_{i,j} \leq B_i; \quad i = 1, \ldots, n.$$

Therefore, the cost function of the bi-blending problem can be written as:

$$f_{bi}(x_{*,1}, x_{*,2}) = \sum_{j=1}^{2} D_j f(x_{*,j}).$$

Redefining the other optimization criterion on the number of distinct raw materials having two mixtures $x_{*,1}$ and $x_{*,2}$ sharing ingredients, the function to minimize is

$$\omega(x_{*,1}, x_{*,2}) = \sum_{i=1}^{n} \delta_i(x_{*,1}) \vee \delta_i(x_{*,2}),$$

where $\vee$ denotes the bitwise *or* operation.

The Quadratic Bi-Blending problem (QBB) is defined as follows:

$$\begin{array}{ll} \min & f_{bi}(x_{*,1}, x_{*,2}),\ \omega(x_{*,1}, x_{*,2}) \\ \text{s.t.} & x_{*,1} \in X_1 \cap Q_1,\ x_{*,2} \in X_2 \cap Q_2 \\ & R(x_{*,j}) \geq \varepsilon;\ j = 1, 2 \\ & \sum_{j=1}^{2} D_j x_{i,j} \leq B_i;\ i = 1, \ldots, n \end{array}$$

2 B&B algorithm for the QBB

Solving the QBB problem in an exhaustive way (the method obtains all global solutions with the established precision) requires the design of a specific Branch-and-Bound algorithm. The concept of Branch-and-Bound is not to generate all the points, but to partition the search region to avoid visiting those regions (partition sets) which are known not to contain an optimal ε-robust solution. B&B methods can be characterized by four rules: *Branching, Selection, Bounding,* and *Elimination* [12, 14]. A *Termination rule* can be incorporated, for instance, based on the smallest sampling precision.

In the branch-and-bound method, the search region is subsequently partitioned in more and more refined parts (branching) over which bounds of an objective function value and bounds on the constraint functions can be determined (bounding). A global upper bound f^U is defined as the objective function value of the best ε-robust (robust) solution found so far. Subsets C_k with a lower bound f_k^L of the objective function

 ISBN: 978-84-614-6167-7

Algorithm 1 QBB algorithm

1: Set $ns := 2 \times (2^n - 1)$ *Number of simplices*
2: Set the working list $\Lambda_1 := \{C_1, \ldots, C_{2^n-1}\}$
3: Set the working list $\Lambda_2 := \{C_{2^n}, \ldots, C_{ns}\}$
4: Set the final lists $Q_1 := \{\}$ and $Q_2 := \{\}$
5: **while** $\Lambda_1, \Lambda_2 \neq \{\}$ **do**
6: Select a simplex $C = C_k$ from Λ_j *Selection rule*
7: Evaluate C
8: Compute $f^L(C)$ and $b_i^L(C),\ i = 1, \ldots, t_k$ *Bounding rule*
9: **if** C cannot be eliminated **then** *Rejection rule*
10: **if** C satisfies the termination criterion **then** *Termination rule*
11: Store C in Q_j
12: **else**
13: Divide C into C_{ns+1}, C_{ns+2} *Branching rule*
14: $C := \arg\min\{f^L(C_{ns+1}), f^L(C_{ns+2})\}$ *Select the cheapest simplex*
15: Store $\{C_{ns+1}, C_{ns+2}\} \setminus C$ in Λ_j
16: $ns := ns + 2$
17: Go to 7
18: **end if**
19: **end if**
20: $j := (j \mod 2) + 1$ *Alternate product*
21: **end while**
22: **return** Q_1 and Q_2

which is larger than the upper bound can be discarded, because they cannot contain an optimal solution.

A detailed description of Algorithm 1 can be found in [11]. Here we summarize the important characteristics of the parallel version.

Every simplex is a region satisfying (1) and its vertices are possible recipes or mixtures. In Algorithm 1, some of the B&B rules can be applied to each product individually and others must be done taking into account both products. The per product rules are:

Division: Divide the longest edge or that edge with the cheapest and the most expensive vertices.

Selection: A hybrid Best-Depth search is done. The cheapest simplex, based on the sum of the cost of its vertices, is selected and a Depth-first is done until no further subdivision is allowed (see Algorithm 1, lines 6, 14 and 17). This is done to reduce the memory requirement of the algorithm.

Rejection: Several tests based on linear, quadratic and robustness constraints are applied to simplices of one product (see [10]).

Termination: Non-rejected simplices that reach the requested size α are stored in Q_j belonging to the solution area.

Bounding: Two bound values have to be calculated for each simplex:

 ISBN: 978-84-614-6167-7

Cost: $f^L(C)$ is a lower bound of the cost of a simplex and it is equal to the minimum cost of the vertices of the simplex, because the simplices are convex and the cost function is linear.

Amount of each raw material: $b_i^L(C)$ is a lower bound of the raw material i in the simplex C. It is obtained in an analogous way than the lower bound of the cost.

The following rules need to take into account both products:

Rejection: Several rejection tests can be applied:

Capacity test: We define

$$\beta_{i,j}^L = D_j \times \min_{x \in C \in \Lambda_j \cup Q_j} x_i. \tag{2}$$

as a lower bound of the demand of a material i in the current search space of product j. Then, a simplex C_k in product j does not satisfy the Capacity test if

$$D_j \times b_i^L(C_k) + \beta_{i,j'}^L > B_i, \tag{3}$$

Pareto test: If a pair of mixtures $x \in C \in \Lambda_1 \cup Q_1$ and $y \in C \in \Lambda_2 \cup Q_2$ has been found with $f(x) + f(y) < f_{\omega(x,y)}^U$, the value of f_p^U is updated for $p = \omega(x,y), \ldots, n$ and the pair (x,y) is stored as a valid solution. Algorithm 1 returns the Pareto vector f^U and the corresponding mixtures. We define

$$\varphi_{u,j}^L = \min\{f(v) : v \in C \subset P_{u,j},\ C \in \Lambda_j \cup Q_j\} \tag{4}$$

as a vector containing the cost value of the cheapest non-rejected mixture for initial simplex $P_{u,j},\ u = 1, \ldots, 2^n - 1$. Then, a simplex C_k in product j does not satisfy the Pareto test if:

$$f^L(C_k) + \varphi_{u,j'}^L > f_{\omega(x,y)}^U; \quad x \in C_k,\ y \in P_{u,j'}. \tag{5}$$

The result of Algorithm 1 is a set of (α-guaranteed) Pareto bi-blending recipes and lists $Q_j,\ j = 1, 2$, that besides the recipes contain mixtures that have not been thrown out. During the execution of the algorithm, lower bounds $\beta_{i,j}^L$ and $\varphi_{u,j}^L$ are updated based on non-rejected vertices. Therefore, a final combination of simplices must be done in order to reject those that cannot contain a Pareto solution:

$$f^L(C) + f^L(C') \leq f_{\omega(x,y)}^U; \quad x \in C,\ y \in C', \tag{6}$$

or do not satisfy the Capacity constraint:

$$b_i^L(C) + b_i^L(C') \leq B_i; \quad i = 1, \ldots, n. \tag{7}$$

This is done by Algorithm 2. When $j = 1$, in line 1 of Algorithm 2, a simplex $C' \in Q_2$ satisfying (6) and (7) with C is tagged as valid and it will not be processed in line 2, when $j = 2$.

 ISBN: 978-84-614-6167-7

Algorithm 2 Comb algorithm

```
1: for j = 1, 2 do
2:   for all C ∈ Q_j not tagged as valid do
3:     if ∃ C' ∈ Q_j' that satisfies (6) and (7) then
4:       Tag C' as valid
5:       Continue with the next C          Remaining C' ∈ Q_j' are not visited
6:     else
7:       Remove C
8:     end if
9:   end for
10: end for
```

3 Parallel strategy

Algorithm 1 obtains all final simplices for both products, before executing Algorithm 2. Therefore, the parallelization of QBB and Comb algorithms can be done independently.

The number of final simplices of the QBB algorithm will depend on several factors: the dimension, the accuracy α of the termination rule and the feasible region of the instances to solve. Preliminary experimentation shows that this number can be relatively large. Consequently, the Comb algorithm is computational much more expensive than QBB algorithm. Therefore, we first study the parallelization of the Comb algorithm.

The Comb algorithm uses a nested loop, and two lists Q_1 and Q_2. For each simplex $C \in Q_j$, a simplex $C' \in Q_{j'}$ must be found that satisfies (6) and (7) to retain it on the list. In the worst case (when the simplex can be removed), list $Q_{j'}$ will be explored completely (all nodes on the other list will be visited).

A number of threads is assigned to the inner loop to perform one iteration of the outer loop. The following notation is needed:

- Pos(C, Q_j): position of the simplex C in Q_j.
- NTh: number of threads.
- Id(Th): identification of thread Th. Identification numbers are consecutive and start at zero.

Each thread Th checks simplices C satisfying $\text{modulus}\,(\text{Pos}(C, Q_j)/NTh) = \text{Id}(Th)$. To avoid contention, simplices are not deleted but tagged to be removed. Otherwise, the list can be modified by several threads, when simplices are removed. Deletion of nodes (tagged to be removed) is done after all simplices are checked and it is done before performing the next iteration ($j = 2$).

Parallelization of the QBB algorithm is more difficult because the pending computational work is not known beforehand. A study on the prediction of the pending work in B&B Interval Global Optimization algorithms can be found in [2]. Although authors show their experience in B&B parallel algorithms in [7, 8, 13, 15], these papers tackle only one B&B algorithm. However, QBB uses two B&B algorithms, one for each product, sharing $\beta^L_{i,j}$, $\varphi^L_{u,j}$ and $f^U_{\omega(x,y)}$ (see Eqs. 2, 3, 4 and 5). The problem is

 ISBN: 978-84-614-6167-7

to determine how many threads are dedicated to each product. This will be addressed in a future study. Here, we will use just one static thread per product to show the difficulty.

4 Experimental results

To evaluate the performance of our parallel algorithm, we have used a pair of five-dimensional products, called UniSpec1-5 and UniSpec5b-5. Both of them are adapted from seven-dimensional ones (UniSpec1 and UniSpec5b, respectively) taken from [10] removing elements $\{a_{i,j} \in A : i = 6, 7;\ j = 6, 7\}$ and $\{b_i \in b : i = 6, 7\}$. The problem was solved with a robustness $\varepsilon = \sqrt{2}/100$, an accuracy $\alpha = \varepsilon$, and a minimal dose $md = 0.03$. The demand of each product is $D^T = (1, 1)$. The availability of raw material one (RM1) and RM3 is restricted to 0.62 and 0.6, respectively; while the others are not limited.

The algorithms were coded in C and run on a Dell PowerEdge R810 with one eight-core Intel Xeon 1.87 GHz processor, 16 GB of RAM, and Linux operating system with 2.6 kernel. POSIX Threads API was used to create and manipulate threads. The LAPACK library is also used by the algorithm.

Table 1 provides information about the test problem described above. The following notation has been used:

- NEvalS: Number of evaluated simplices.
- NEvalV: Number of evaluated vertices.
- QLR: Number of simplices rejected by linear infeasibility, quadratic infeasibility or lack of robustness.
- Pareto: Number of simplices rejected by Pareto test.
- Capacity: Number of simplices rejected by Capacity test.
- $|Q_S|$: Number of simplices in final lists $Q_j,\ j = 1, 2$.
- $|Q_V|$: Number of vertices associated to simplices in $Q_j,\ j = 1, 2$.
- NTh: Number of created threads.
- Th_1 time: The running time of Th_1 in seconds.
- Th_2 time: The running time of Th_2 in seconds.
- Time: The running time in seconds.
- Speedup: Speedup obtained.

The speedup with regard to execution time of a parallel algorithm with p process units is defined as $S(p) = t(1)/t(p)$, where $t(p)$ is the execution time when p processors are used.

Table 1 shows the numerical results obtained from running the sequential algorithm (BiBlendSeq), and the parallel version (BiBlendPar) for $NTh = 2$ in B&B phase, and $NTh = 1$, 2, 4 and 8 threads in Comb phase. The data shown in Table 1 is the average value of five executions.

BiBlendPar exhibits a good scalability and speedup when compared to BiBlend-Seq. For the B&B phase, a slight acceleration is obtained in the execution time. A

 ISBN: 978-84-614-6167-7

Table 1: Computational effort

B&B Phase

	BiBlendSeq	BiBlendPar			
NEvalS	2,536,862	2,537,430	=	=	=
NEvalV	168,186	168,299	=	=	=
QLR	887,609	888,004	=	=	=
Pareto	54,050	54,050	=	=	=
Capacity	18,277	18,211	=	=	=
$\|Q_S\|$	308,443	308,465	=	=	=
$\|Q_V\|$	49,317	49,324	=	=	=
NTh	–	2	2	2	2
Th_1 time	–	0.83	0.81	0.76	0.81
Th_2 time	–	7.03	7.03	6.96	7.00
Time	7.23	7.03	7.03	6.96	7.00
Speedup	–	1.03	1.03	1.04	1.03

Comb. Phase

	BiBlendSeq	BiBlendPar			
Pareto	27,284	27,284	=	=	=
Capacity	105,499	105,521	=	=	=
$\|Q_S\|$	175,660	175,660	=	=	=
$\|Q_V\|$	24,861	24,861	=	=	=
NTh	–	1	2	4	8
Time	1,991.46	1,966.51	956.01	465.02	228.63
Speedup	–	1.01	2.08	4.28	8.71

Total

Time	1,998.70	1,973.54	936.05	471.98	235.63
Speedup	–	1.01	2.08	4.23	8.48

linear speedup is not reached due to the difference of complexity between two products: UniSpec1-5 has a simpler quadratic requirement compared to UniSpec5b-5. For the Comb phase, a linear speedup is obtained compared to the sequential one. Notice that Comb phase is able to filter out almost half of the final simplices obtained in B&B phase.

Two solutions have been found for UniSpec1-5 & UniSpec5b-5 with a different number of raw materials involved: $(x_{*,1}^{\star[1]}, x_{*,2}^{\star[1]})$ and $(x_{*,1}^{\star[2]}, x_{*,2}^{\star[2]})$. The first one uses four raw materials (RM1, RM3, RM4 and RM5):

$$\mathbf{x}^{\star[1]} = \begin{array}{c} \\ \text{UniSpec1-5} \\ \text{UniSpec5b-5} \end{array} \begin{array}{c} \begin{array}{ccccc} \text{RM1} & \text{RM2} & \text{RM3} & \text{RM4} & \text{RM5} \end{array} \\ \begin{pmatrix} 0.428125 & 0.0 & 0.4352344 & 0.0 & 0.1366406 \\ 0.146875 & 0.0 & 0.1640625 & 0.2328125 & 0.4562500 \end{pmatrix} \end{array}^T$$

 ISBN: 978-84-614-6167-7

Its cost value is $f(x^{\star[1]}_{*,1}, x^{\star[1]}_{*,2}) = 111.09 + 116.334375 = 227.424375$. The second one involves five raw materials:

$$\mathbf{x}^{\star[2]} = \begin{array}{c} \\ \text{UniSpec1-5} \\ \text{UniSpec5b-5} \end{array} \begin{array}{c} \begin{array}{ccccc} \text{RM1} & \text{RM2} & \text{RM3} & \text{RM4} & \text{RM5} \end{array} \\ \begin{pmatrix} 0.428125 & 0.0 & 0.442344 & 0.0 & 0.129531 \\ 0.156172 & 0.03 & 0.152852 & 0.212617 & 0.448359 \end{pmatrix}^T \end{array}$$

Its cost value is $f(x^{\star[2]}_{*,1}, x^{\star[2]}_{*,2}) = 111.033125 + 116.172422 = 227.205547$.

5 Conclusions and future work

A parallelization of an algorithm to solve the bi-blending problem has been studied for a small-medium size instance of the problem. This single case shows the difficulties of this type of algorithm. Bi-blending increases the challenges of the parallelization of a B&B algorithm for single blending because it actually runs two B&B algorithms that share information. Additionally, in bi-blending algorithms, a combination of final simplices has to be done after the B&B phase to discard infeasible regions. This combination phase is computationally several orders of magnitude larger than the B&B phase. Here we use just one thread for each product in the B&B phase and several threads for the combination phase. Linear speedup is obtained on a shared memory machine with eight cores.

Our intention is to experiment with larger dimensional problems for the parallel bi-blending algorithm, trying to reduce the computational cost. Another future research question is to develop the n-blending algorithm and its parallel version, which is the problem of interest to the industry.

Acknowledgements

This work has been funded by grants from the Spanish Ministry of Science and Innovation (TIN2008-01117), and Junta de Andalucía (P08-TIC-3518), in part financed by the European Regional Development Fund (ERDF). Eligius M. T. Hendrix is a fellow of the Spanish "Ramon y Cajal" contract program, co-financed by the European Social Fund.

References

[1] J. Ashayeri, A.G.M. van Eijs, and P. Nederstigt. Blending modelling in a process manufacturing: A case study. *European Journal of Operational Research*, 72(3):460–468, 1994.

[2] J.L. Berenguel, L.G. Casado, I. García, and E.M.T. Hendrix. On estimating workload in interval branch-and-bound global optimization algorithms. *Journal of Global Optimization*. Submitted.

[3] J.W.M. Bertrand and W.G.M.M. Rutten. Evaluation of three production planning procedures for the use of recipe flexibility. *European Journal of Operational Research*, 115(1):179–194, 1999.

[4] L.G. Casado, I. García, J.A. Martínez, and E.M.T. Hendrix. Shared memory parallel exhaustive search of epsilon-robust mixture design solutions. In *Volume of Abstracts of 22nd European Conference on Operational Research (EURO XXII)*, page 178, 2007.

[5] L.G. Casado, I. García, B.G. Tóth, and E.M.T Hendrix. On determining the cover of a simplex by spheres centered at its vertices. *Journal of Global Optimization*, pages 1–11, 2010. DOI: 10.1007/s10898-010-9524-x.

[6] L.G. Casado, E.M.T Hendrix, and I. García. Infeasibility spheres for finding robust solutions of blending problems with quadratic constraints. *Journal of Global Optimization*, 39:577–593, 2007. DOI: 10.1007/s10898-007-9157-x.

[7] L.G. Casado, J.A. Martínez, I. García, and E.M.T. Hendrix. Branch-and-bound interval global optimization on shared memory multiprocessors. *Optimization Methods Software*, 23:689–701, October 2008. DOI: 10.1080/10556780802086300.

[8] J.F.S. Estrada, L.G. Casado, and I. García. Adaptive parallel interval global optimization algorithms based on their performance for non-dedicated multicore architectures. In *Parallel, Distributed and Network-Based Processing (PDP), 2011 19th Euromicro International Conference on*, pages 252–256, February 2011. DOI: 10.1109/PDP.2011.54.

[9] B. Gendron and T.G. Crainic. Parallel branch-and-bound algorithms: Survey and synthesis. *Operations Research*, 42(6):1042–1066, 1994.

[10] E.M.T. Hendrix, L.G. Casado, and I. García. The semi-continuous quadratic mixture design problem: Description and branch-and-bound approach. *European Journal of Operational Research*, 191(3):803–815, 2008.

[11] J.F.R. Herrera, L.G. Casado, E.M.T. Hendrix, and I. García. Pareto optimality and robustness in bi-blending problems. *Top*. Submitted.

[12] T. Ibaraki. Theoretical comparisons of search strategies in branch and bound algorithms. *International Journal of Computing and Information Sciences*, 5(4):315–344, 1976.

[13] J. Martínez, L.G Casado, J.A. Alvarez, and I. García. Interval parallel global optimization with Charm++. In Jack Dongarra, Kaj Madsen, and Jerzy Wasniewski, editors, *Applied Parallel Computing. State of the Art in Scientific Computing*, volume 3732 of *Lecture Notes in Computer Science*, pages 161–168. 2006. DOI: 10.1007/11558958_18.

 ISBN: 978-84-614-6167-7

[14] L.G. Mitten. Branch and bound methods: general formulation and properties. *Operations Research*, 18(1):24–34, 1970.

[15] J.F. Sanjuan-Estrada, L.G. Casado, and I. García. Adaptive parallel interval branch and bound algorithms based on their performance for multicore architectures. *The Journal of Supercomputing*, pages 1–9, 2011. DOI: 10.1007/s11227-011-0594-4.

[16] H.P. Williams. *Model Building in Mathematical Programming.* Wiley & Sons, Chichester, 1993.

 ISBN: 978-84-614-6167-7

Proceedings of the 11th International Conference on Computational and Mathematical Methods in Science and Engineering, CMMSE 2011 26–30 June 2011.

On construction of second order schemes for Maxwell's equations with discontinuous dielectric permittivity

Timur Z. Ismagilov[1]

[1] *Department of Information Technologies, Novosibirsk State University*

emails: ismagilov@academ.org

Abstract

We suggest an approach to construction of second order finite volume schemes for Maxwell's equations with discontinuous dielectric permittivity. The key idea is to use stencils that are comprised of cells with the same dielectric permittivity for calculating derivatives used to increase the order of approximation. The idea was implemented in Godunov scheme built with Law-Wendroff and Van Leer approaches. Results of test computations for problems with linear and curvilinear discontinuities show second order of approximation.

Key words: Maxwell's equations, Godunov scheme, discontinuous permittivity, second order, finite volume

1 Introduction

Maxwell's equations describe propagation of electromagnetic waves. For many practical problems analytical solutions do not exist and as a result various numerical methods were developed [1, 2]. Probably the most popular method for numerical solution of Maxwell's equations today is the method of finite differences. It can provide satisfactory results for a number of problems.

One of the first works devoted to construction of finite difference schemes for numerical solution of Maxwell's equations was paper [3]. In this work Yee suggested a scheme with second order of approximation in space and time for the case of constant dielectric permittivity, based on staggered cartesian grids. Later on various finite difference methods were successfully applied to solution of many problems [4, 5]. The main advantage of finite difference schemes is simplicity of constructed algorithms.

The main disadvantage of finite difference schemes is poor precision for problems with curvilinear boundaries of computational region and subregions with different dielectric permittivity. For this reason for problems with complicated geometry finite

 ISBN: 978-84-614-6167-7

volume schemes [6] can be preferable. These schemes allow to perform approximation using unstructured meshes and in this way more precisely approximate computational region boundaries and boundaries between subregions with different dielectric permittivity. One of the first finite volume schemes for numerical solution of Maxwell's equations was suggested in [7]. In this paper Shankar suggested a finite volume scheme based on structured non-cartesian meshes. In the following many finite volume schemes for solution of Maxwell's equations using unstructured meshes were suggested [8, 9, 10].

One of the main problems for construction of finite volume and finite difference schemes for Maxwell's equations is the case of discontinuous dielectric permittivity. For finite difference schemes in [2] various ways of smoothing dielectric permittivity were considered. For finite volume schemes in [10] use of continuous variables was suggested. Both approaches allowed to achieve better precision for some test cases but not the second order of convergence.

We suggest a finite volume scheme for numerical solution of Maxwell's equations with discontinuous dielectric permittivity on unstructured meshes. The scheme is second order accurate in space and time even for curvilinear discontinuities of dielectric permittivity. Godunov scheme is used as a basis [11]. In order to achieve second order accuracy in space and time approaches of Lax-Wendroff and Van Leer are used respectively [12, 13] similar to [10]. The key difference of the scheme built is gradient approximation that is first order accurate even for cells adjacent to dielectric permittivity discontinuity. This makes proposed scheme second order accurate in space and time. Computational results are presented that confirm second order approximation of the scheme built for linear as well as curvilinear discontinuities.

It is important to emphasize that our results do not contradict Godunov theorem since solution discontinuities are fixed, coincide with discontinuities of dielectric permittivity and the scheme depends on this information.

2 Maxwell's equations

The system of Maxwell's equations in the absence of charges and currents in dimensionless variables in two-dimensions can be written in conservative form as

$$\frac{\partial}{\partial t}\mathbf{U} + \frac{\partial}{\partial x_1}\mathbf{F}_1 + \frac{\partial}{\partial x_2}\mathbf{F}_2 = 0, \tag{1}$$

where $\mathbf{U}$ — conservative variables vector, $\mathbf{F}_1$ and $\mathbf{F}_2$ — flux vectors. For TM case they can be written as

$$\mathbf{U} = \begin{pmatrix} D_3 \\ B_1 \\ B_2 \end{pmatrix}, \ \mathbf{F}_1 = \begin{pmatrix} -H_2 \\ 0 \\ -E_3 \end{pmatrix}, \quad \mathbf{F}_2 = \begin{pmatrix} H_1 \\ E_3 \\ 0 \end{pmatrix}, \tag{2}$$

and for TE case as

$$\mathbf{U} = \begin{pmatrix} D_1 \\ D_2 \\ B_3 \end{pmatrix}, \ \mathbf{F}_1 = \begin{pmatrix} 0 \\ H_3 \\ E_2 \end{pmatrix}, \ \mathbf{F}_2 = \begin{pmatrix} -H_3 \\ 0 \\ -E_1 \end{pmatrix}. \tag{3}$$

 ISBN: 978-84-614-6167-7

In the above formulas E_1, E_2, E_3 — electric field, H_1,H_2, H_3 — magnetic field, $D_1 = \varepsilon E_1$, $D_2 = \varepsilon E_2$, $D_3 = \varepsilon E_3$ — electric induction, $B_1 = \mu H_1$, $B_2 = \mu H_2$, $B_3 = \mu H_3$ — magnetic induction, ε — dielectric permittivity, μ — magnetic permeability assumed to be zero in this paper.

In the regions with constant dielectric permittivity an equivalent non-divergent form of initial system (1) can be written

$$\frac{\partial}{\partial t}\mathbf{U} + A_1\frac{\partial}{\partial x_1}\mathbf{U} + A_2\frac{\partial}{\partial x_2}\mathbf{U} = 0, \tag{4}$$

where for the TM case matrices A_1 and A_2 are defined as

$$A_1 = \begin{pmatrix} 0 & 0 & -1 \\ 0 & 0 & 0 \\ -\frac{1}{\varepsilon} & 0 & 0 \end{pmatrix}, \quad A_2 = \begin{pmatrix} 0 & 1 & 0 \\ \frac{1}{\varepsilon} & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \tag{5}$$

and for the TE case as

$$A_1 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & \frac{1}{\varepsilon} & 0 \end{pmatrix}, \quad A_2 = \begin{pmatrix} 0 & 0 & -1 \\ 0 & 0 & 0 \\ -\frac{1}{\varepsilon} & 0 & 0 \end{pmatrix}. \tag{6}$$

On the dielectric permittivity discontinuity a vector of continuous variables $\mathbf{W}$ can be considered. Vector $\mathbf{W}$ is related to conservative variable vector $\mathbf{U}$ by the transition matrix Φ: $\mathbf{W} = \Phi\mathbf{U}$. For TM case continuous variables vector and transition matrix can be written as

$$\mathbf{W} = \begin{pmatrix} E_3 \\ B_1 \\ B_1 \end{pmatrix}, \quad \Phi = \begin{pmatrix} \frac{1}{\varepsilon} & 0 & 1 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \tag{7}$$

and for TE case as

$$\mathbf{W} = \begin{pmatrix} D_n \\ E_\tau \\ H_3 \end{pmatrix}, \quad \Phi(\phi,\varepsilon) = \begin{pmatrix} \varepsilon\cos(\phi) & \varepsilon\sin(\phi) & 0 \\ -\sin(\phi) & \cos(\phi) & 0 \\ 0 & 0 & 1 \end{pmatrix}, \tag{8}$$

where D_n - is a normal component of electric induction, E_τ - tangential component of electric field, ϕ - angle defining normal to dielectric permittivity discontinuity.

3 Finite volume scheme

Consider a computational region in two-dimensional space. Assume that an unstructured mesh composed of triangles Δ was constructed. By integrating the system of equations (1) over the i-th mesh cell Δ_i we can obtain integral conservation law

$$\frac{\partial}{\partial t}\int\limits_{\Delta_i} \mathbf{U} d\Omega + \sum_{k=1}^{3}\int\limits_{\Gamma_k} \left(n_1\mathbf{F}_1 + n_2\mathbf{F}_2\right) d\Gamma = 0, \tag{9}$$

 ISBN: 978-84-614-6167-7

where Γ_k — k-th cell edge, $\mathbf{n} = (n_1, n_2)$ — outward normal. For numerical approximation of integral equation (9) consider a finite volume scheme

$$\Omega_{\Delta_i} \frac{\mathbf{U}_i^{n+1} - \mathbf{U}_i^n}{\tau} + \sum_{k=1}^{3} s_{\Delta_i}^k \mathbf{F}_i^k = 0, \tag{10}$$

where vector $\mathbf{U}_i^n$ — denotes the value of $\mathbf{U}$ at the i–th cell center $\mathbf{X}_{\Delta_i}^B$ at the time $t_n = n\tau$, τ — time step, $s_{\Delta_i}^k$ — length of the k-th edge, Ω_{Δ_i} — area of the i-th cell, $\mathbf{F}_i^k$— flux through the k-th edge approximation. For approximation of the flux through the edge Riemann problem solution is used [11]

$$\mathbf{F} = A^+ \mathbf{U}_L \left(\mathbf{X}^C\right) + A^- \mathbf{U}_R \left(\mathbf{X}^C\right), \tag{11}$$

where initial states are taken as interpolations $\mathbf{U}_L$ and $\mathbf{U}_R$ from cell centers Δ_L and Δ_R approximating values at the edge center at time half step $t_{n+1/2} = n\tau + \tau/2$ with second order of approximation in time and space

$$\mathbf{U}_L(\mathbf{X}^C) = \mathbf{U}(\mathbf{X}_L^B) + \frac{\partial \mathbf{U}}{\partial \mathbf{x}}(\mathbf{X}_L^B)(\mathbf{X}^C - \mathbf{X}_L^B) - \frac{\tau}{2}\left(A_1 \frac{\partial \mathbf{U}}{\partial x_1}(\mathbf{X}_L^B) + A_2 \frac{\partial \mathbf{U}}{\partial x_2}(\mathbf{X}_L^B)\right). \tag{12}$$

In this case scheme (10) will approximate initial conservation law (9) with second order in space and time. Approaches used to increase order of approximation in time and space were suggested by Lax Wendroff [12] and Van Leer [13], respectively and used in [10]. For TM case A_1 and A_2 are given by

$$A^+ = \frac{1}{\sqrt{\varepsilon_L} + \sqrt{\varepsilon_R}} \begin{pmatrix} \sqrt{\frac{\varepsilon_R}{\varepsilon_L}} & \sqrt{\varepsilon_R} n_2 & -\sqrt{\varepsilon_R} n_1 \\ \frac{1}{\sqrt{\varepsilon_L}} n_2 & n_2^2 & -n_1 n_2 \\ -\frac{1}{\sqrt{\varepsilon_L}} n_1 & -n_1 n_2 & n_1^2 \end{pmatrix}, \tag{13}$$

$$A^- = \frac{1}{\sqrt{\varepsilon_L} + \sqrt{\varepsilon_R}} \begin{pmatrix} -\sqrt{\frac{\varepsilon_L}{\varepsilon_R}} & \sqrt{\varepsilon_L} n_2 & -\sqrt{\varepsilon_L} n_1 \\ \frac{1}{\sqrt{\varepsilon_R}} n_2 & -n_2^2 & n_1 n_2 \\ -\frac{1}{\sqrt{\varepsilon_R}} n_1 & n_1 n_2 & -n_1^2 \end{pmatrix}, \tag{14}$$

and for TE case by

$$A^+ = \frac{1}{\sqrt{\varepsilon_L} + \sqrt{\varepsilon_R}} \begin{pmatrix} \sqrt{\frac{\varepsilon_R}{\varepsilon_L}} n_2^2 & -\sqrt{\frac{\varepsilon_R}{\varepsilon_L}} n_1 n_2 & -\sqrt{\varepsilon_R} n_2 \\ -\sqrt{\frac{\varepsilon_R}{\varepsilon_L}} n_1 n_2 & \sqrt{\frac{\varepsilon_R}{\varepsilon_L}} n_1^2 & \sqrt{\varepsilon_R} n_1 \\ -\frac{1}{\sqrt{\varepsilon_L}} n_2 & \frac{1}{\sqrt{\varepsilon_L}} n_1 & 1 \end{pmatrix}, \tag{15}$$

$$A^- = \frac{1}{\sqrt{\varepsilon_L} + \sqrt{\varepsilon_R}} \begin{pmatrix} -\sqrt{\frac{\varepsilon_L}{\varepsilon_R}} n_2^2 & \sqrt{\frac{\varepsilon_L}{\varepsilon_R}} n_1 n_2 & -\sqrt{\varepsilon_L} n_2 \\ \sqrt{\frac{\varepsilon_L}{\varepsilon_R}} n_1 n_2 & -\sqrt{\frac{\varepsilon_L}{\varepsilon_R}} n_1^2 & \sqrt{\varepsilon_L} n_1 \\ -\frac{1}{\sqrt{\varepsilon_R}} n_2 & \frac{1}{\sqrt{\varepsilon_R}} n_1 & -1 \end{pmatrix}. \tag{16}$$

 ISBN: 978-84-614-6167-7

To compute gradients of conservative variables $\mathbf{U}$ at cell barycenters with the first order of approximation in the presence of dielectric permittivity discontinuity we will use a three stage process.

During the first stage we will compute preliminary gradients with the first order of approximation using values in the cell and values in neighboring cells with the same dielectric permittivity with the help of least square method [14]

$$\frac{\partial \mathbf{U}}{\partial \mathbf{x}} = (\mathbf{U}_1, ..., \mathbf{U}_n) X^T \left(X X^T\right)^{-1}, \tag{17}$$

where $\mathbf{U}_1, ..., \mathbf{U}_n$ are values at cell barycenters used to calculate preliminary gradients, $X = \{x_{ij}\} = \{x_i^j - \sum_{k=1}^n x_i^k\}$, x_i^j - i-th coordinate of j-th barycenter. We specifically emphasize that if a cell is adjacent to the dielectric permittivity discontinuity the neighboring cell on the other side of discontinuity will not be used.

During the second stage for each cell we will calculate values for each of its vertices $\mathbf{X}^P$ with second order of approximation using values and preliminary gradients approximations at cell barycenter $\mathbf{X}^B$ with the help of

$$\mathbf{U}(\mathbf{X}^P) = \mathbf{U}(\mathbf{X}^B) + \frac{\partial \mathbf{U}}{\partial \mathbf{x}}(\mathbf{X}^P - \mathbf{X}^B). \tag{18}$$

After obtaining in each vertex a set of interpolated values from all the adjacent cells we partition them into groups depending on the value of dielectric permittivity in a cell from which interpolation was carried out. For each group we will calculate arithmetic average. As a result for each vertex for each value of dielectric permittivity present in adjacent cells we will have a separate value $\mathbf{U}$. The second version of this stage is to carry out averaging of all the interpolated values with the help of continuous variables. This can be accomplished by changing variables from $\mathbf{U}$ to $\mathbf{W}$ before averaging. In this case at a vertex only one average value is obtained $\mathbf{W}$.

During the third stage we will calculate final gradients in each cell using values in adjacent vertices. If the first version of the second stage was used we will take values in vertices corresponding to the dielectric permittivity in the cell. If the second version of the second stage was used we will switch from continuous variables back to conservative variables before calculation.

4 Test computations

The schemes suggested were tested using several test problems for which analytical solutions are available. In our computations we used unstructured triangular meshes built with the help of Gmsh software [15]. For each problem we performed calculations using a sequence of five meshes. Every next mesh had characteristic size two times smaller than the previous one. Time step also was two times smaller. To evaluate approximation properties of the scheme we compared computational results with analytic solution. Error of numerical solution was calculated as deviation from analytic solution

 ISBN: 978-84-614-6167-7

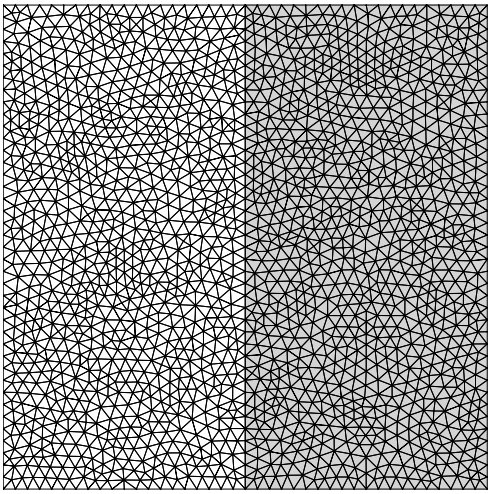


Figure 1:

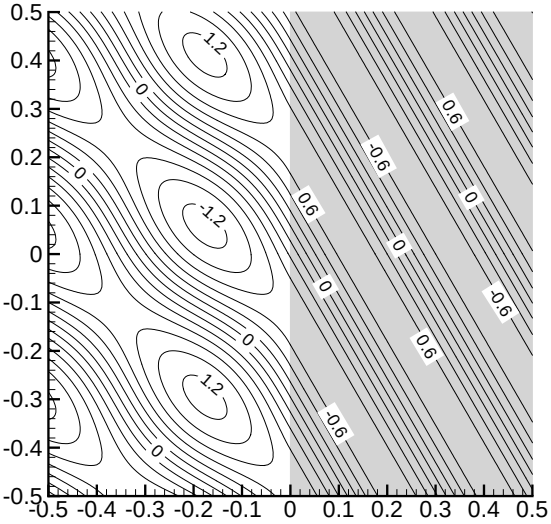


Figure 2:

at time $t^n = n\tau$ in the L_2 norm and was calculated using

$$\frac{\left\|\mathbf{U}^n\left(\mathbf{X}^B\right) - \mathbf{U}^{\mathbf{exact}}\left(\mathbf{X}^B, t^n\right)\right\|_{L_2}}{\left\|\mathbf{U}^{\mathbf{exact}}\left(\mathbf{X}^B, t^n\right)\right\|_{L_2}} = \sqrt{\frac{\sum\limits_{i=1}^{T}\left[\sum\limits_{k=1}^{3}\left(\mathbf{U}_k^n\left(\mathbf{X}^{B_i}\right) - \mathbf{U}_k^{\mathbf{exact}}\left(\mathbf{X}^{B_i}, t^n\right)\right)^2\right]\cdot S_{\Delta_i}}{\sum\limits_{i=1}^{T}\left[\sum\limits_{k=1}^{3}\left(\mathbf{U}_k^{\mathbf{exact}}\left(\mathbf{X}^{B_i}, t^n\right)\right)^2\right]\cdot S_{\Delta_i}}}, \tag{19}$$

where T — total number of cells in the computational region, $\mathbf{U}_k^n\left(\mathbf{X}^{B_i}\right)$ and $\mathbf{U}_k^{\mathbf{exact}}\left(\mathbf{X}^{B_i}, t^n\right)$ — calculated and exact values of electromagnetic fields at the cell i barycenter respectively. We present results using first and second versions of graduate calculations for TE and TM cases respectively.

4.1 Test 1

Consider a problem of interaction of plane electromagnetic wave with a linear dielectric boundary. Dielectric permittivity of half plane $x_1 > 0$ is ε. Wave numbers inside and outside of half plane β and β_1 are related by the equation $\beta_1 = \beta\sqrt{\varepsilon}$. Incident wave propagates in the direction $\mathbf{n}_0 = (\cos(\theta_0), \sin(\theta_0))$. The solution to this problem also includes a plane wave reflected from the dielectric border propagating in the direction $\mathbf{n}_2 = (-\cos(\theta_0), \sin(\theta_0))$ and transmitted plane wave inside of dielectric propagating in the direction $\mathbf{n}_1 = (\cos(\theta_1), \sin(\theta_1))$. The angle of incidence and the angle of refraction θ_0 and θ_1 are related by $\sin(\theta_1)\sqrt{\varepsilon} = \sin(\theta_0)$.

For the case of TM wave electric field of incident wave is $E_z^i = E_0 e^{j\beta - \mathbf{n}_0\cdot\mathbf{x} + j\omega t}$, electric field of reflected wave $E_z^s = E_1 e^{-j\beta\mathbf{n}_1\cdot\mathbf{x} + j\omega t}$, and electric field for transmitted

 ISBN: 978-84-614-6167-7

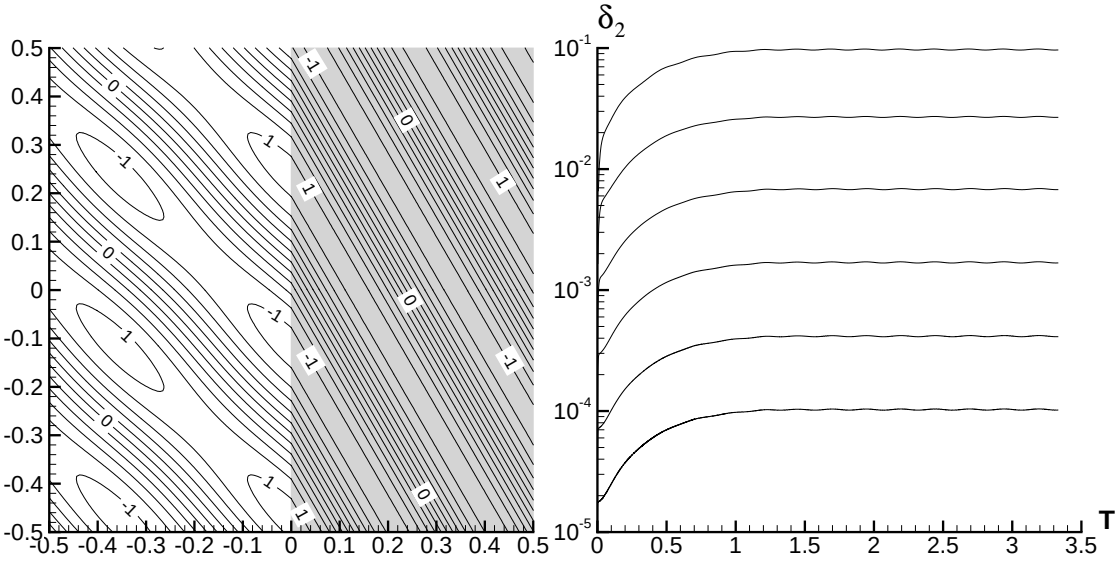


Figure 3:

Figure 4:

wave inside dielectric is $E_z^d = E_2 e^{-j\beta \mathbf{n}_2 \cdot \mathbf{x} + j\omega t}$, where E_1 and E_2 are defined as

$$E_1 = \frac{2\cos(\theta_0)\sin(\theta_1)}{\sin(\theta_0+\theta_1)}, \quad E_2 = \frac{\sin(\theta_1-\theta_0)}{\sin(\theta_0+\theta_1)}. \tag{20}$$

For the case of TE wave magnetic field for incident wave is $H_z^i = H_0 e^{-j\beta \mathbf{n}_0 \cdot \mathbf{x} + j\omega t}$, magnetic field for reflected wave is $H_z^s = H_1 e^{-j\beta \mathbf{n}_1 \cdot \mathbf{x} + j\omega t}$, and magnetic field for transmitted wave inside dielectric is $H_z^d = H_2 e^{-j\beta \mathbf{n}_2 \cdot \mathbf{x} + j\omega t}$, where H_1 and H_2 are defined as

$$H_1 = \frac{\sin(2\theta_0)}{\sin(\theta_0+\theta_1)\cos(\theta_0-\theta_1)}, \quad H_2 = \frac{\tan(\theta_0-\theta_1)}{\tan(\theta_0+\theta_1)} \tag{21}$$

The angle of incidence was chosen as $\pi/4$, incident wavelength 0.5, dielectric permittivity of half plane was 2.0. Computational region was a square with side length 1.0. Square center coordinates were $(0.0, 0.0)$. We used meshes composed of 3634, 14446, 58488, 236880 and 958556 triangles. A mesh of 3634 triangles is shown on Fig. 1. Distribution of E_z at time 2.5 for the case of TM wave obtained using a mesh of 58488 triangles is shown on Fig. 2. Distribution of H_z at time 2.5 for the case of TE wave obtained using a mesh of 58488 triangles is shown on Fig. 2. Error evolution in L_2 norm for different meshes for the cases of TM and TE waves is shown on Fig. 5 and Fig. 6. Values of maximum errors for different meshes for TM and TE waves are presented in Table 1 and in Table 2. Error behavior demonstrates second order of convergence of numerical solution to analytic solution and confirms second order of approximation of the scheme considered for TM and TE waves for the case of linear discontinuity of dielectric permittivity.

 ISBN: 978-84-614-6167-7

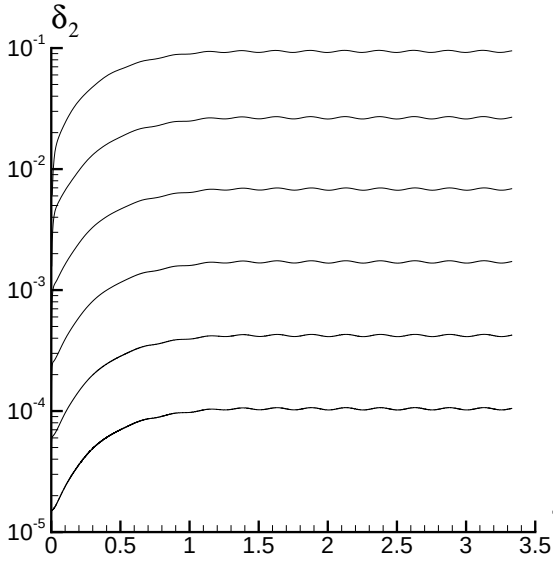


Figure 5:

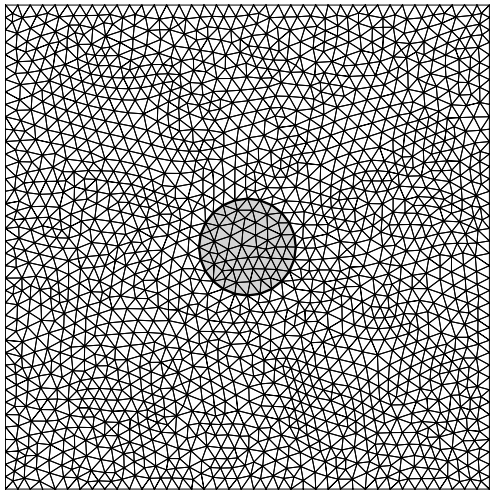


Figure 6:

Table 1:

Cells	δ_2	Order
3634	0.02724913	—
14446	0.00690299	1.98
58488	0.00171342	2.00
236880	0.00042076	2.01
958556	0.00010414	2.01

Table 2:

Cells	δ_2	Order
3634	0.02714944	—
14446	0.00698887	1.96
58488	0.00174735	1.98
236880	0.00043037	1.99
958556	0.00010677	2.00

4.2 Test 2

As a second test consider a problem of plane electromagnetic wave interaction with a dielectric cylinder. Incident wave propagates in the direction (1,0), cylinder radius is a, cylinder dielectric permittivity is ε. The solution to this problem includes incident plane wave, wave reflected from the boundary of the cylinder and wave inside the

 ISBN: 978-84-614-6167-7

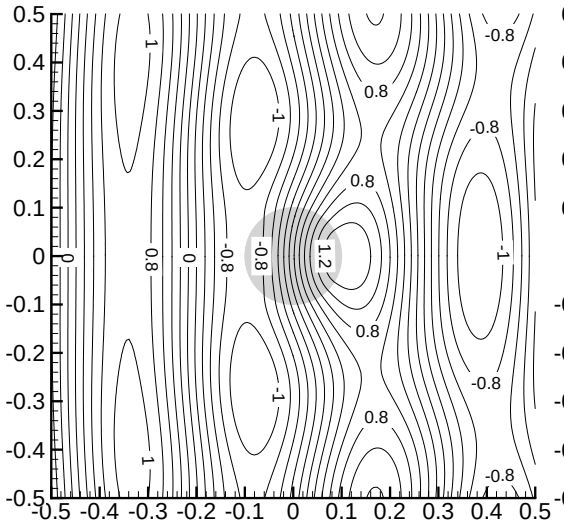


Figure 7:

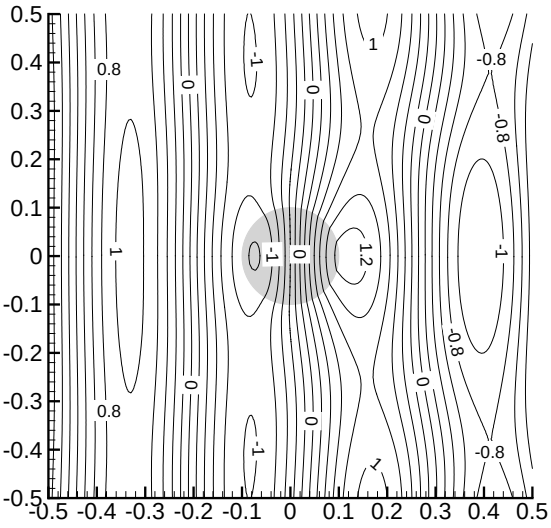


Figure 8:

cylinder. It can be written in polar coordinates radius ρ and angle φ with the help of Bessel functions J_n and Hankel functions of the second kind $H_n^{(2)}$.

For the TM case electric field for the incident wave E_z^i, electric field for the reflected wave E_z^s and electric field inside the cylinder E_z^d can be written as a series

$$E_z^i = H_0 \sum_{n=-\infty}^{\infty} j^{-n} \left[J_n(\beta\rho) \right] e^{jn\varphi + j\omega t} \tag{22}$$

$$E_z^s = H_0 \sum_{n=-\infty}^{\infty} j^{-n} \frac{J_n'(\beta a) J_n(\beta_1 a) - \sqrt{\varepsilon} J_n(\beta a) J_n'(\beta_1 a)}{\sqrt{\varepsilon} J_n'(\beta_1 a) H_n^{(2)}(\beta a) - J_n(\beta_1 a) H_n^{(2)'}(\beta a)} H_n^{(2)}(\beta\rho) e^{jn\varphi + j\omega t} \tag{23}$$

$$E_z^d = H_0 \sum_{n=-\infty}^{\infty} j^{-n} \frac{J_n(\beta a) H_n^{(2)'}(\beta a) - J_n'(\beta a) H_n^{(2)}(\beta a)}{J_n(\beta_1 a) H_n^{(2)'}(\beta a) - \sqrt{\varepsilon} J_n'(\beta_1 a) H_n^{(2)}(\beta a)} J_n(\beta_1\rho) e^{jn\varphi + j\omega t} \tag{24}$$

For the TE case magnetic field for the incident wave H_z^i, magnetic field for the reflected wave H_z^s and magnetic field inside the cylinder H_z^d can be written as a series

$$H_z^i = H_0 \sum_{n=-\infty}^{\infty} j^{-n} \left[J_n(\beta\rho) \right] e^{jn\varphi + j\omega t} \tag{25}$$

$$H_z^s = H_0 \sum_{n=-\infty}^{\infty} j^{-n} \frac{J_n'(\beta a) J_n(\beta_1 a) - \sqrt{1/\varepsilon} J_n(\beta a) J_n'(\beta_1 a)}{\sqrt{1/\varepsilon} J_n'(\beta_1 a) H_n^{(2)}(\beta a) - J_n(\beta_1 a) H_n^{(2)'}(\beta a)} H_n^{(2)}(\beta\rho) e^{jn\varphi + j\omega t} \tag{26}$$

$$H_z^d = H_0 \sum_{n=-\infty}^{\infty} j^{-n} \frac{J_n(\beta a) H_n^{(2)'}(\beta a) - J_n'(\beta a) H_n^{(2)}(\beta a)}{J_n(\beta_1 a) H_n^{(2)'}(\beta a) - \sqrt{1/\varepsilon} J_n'(\beta_1 a) H_n^{(2)}(\beta a)} J_n(\beta_1\rho) e^{jn\varphi + j\omega t} \tag{27}$$

 ISBN: 978-84-614-6167-7

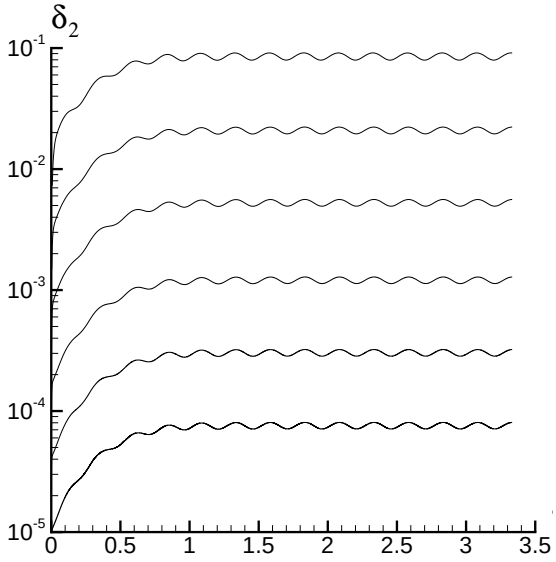


Figure 9:

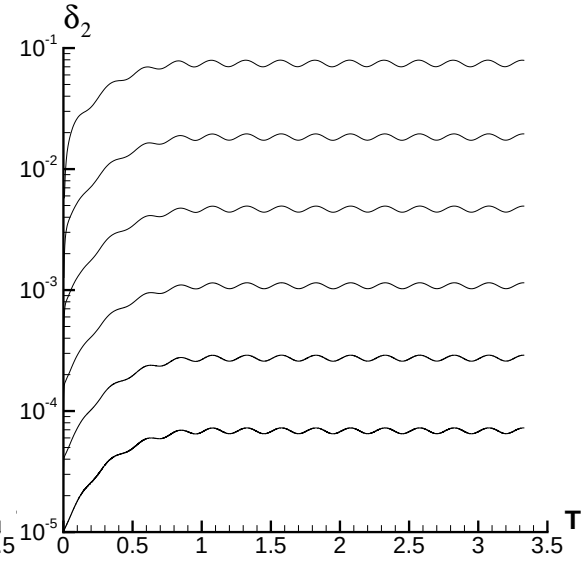


Figure 10:

Cylinder radius was chosen as 0.1, incident wave length was 0.5, cylinder center coordinates were $(0.0, 0.0)$, cylinder dielectric permittivity was 2.0. Computational region constituted a square with side length 1.0 and center coordinates $(0.0, 0.0)$. We used meshes composed of 3464, 14398, 59888, 240654 and 968090 triangles. A mesh of 3464 triangles is shown on Fig. 6. Distribution of E_z at time 2.5 for the case of TM wave obtained using a mesh of 59888 triangles is shown on Fig. 7. Distribution of H_z at time 2.5 for the case of TE wave obtained using a mesh of 59888 triangles is shown on Fig. 8. Error evolution in L_2 norm for different meshes for the cases of TM and TE waves is shown on Fig. 9 and Fig. 10. Values of maximum errors for different meshes for TM and TE waves are presented in Table 3 and in Table 4. Error behavior demonstrates second order of convergence of numerical solution to analytic solution and confirms second order of approximation of the scheme considered for TM and TE waves for the case of curvilinear discontinuity of dielectric permittivity.

Table 3:

Cells	δ_2	Order
3464	0.02225090	—
14398	0.00560872	1.99
59888	0.00128325	2.06
240654	0.00032258	2.04
968090	0.00008057	2.03

 ISBN: 978-84-614-6167-7

Table 4:

Cells	δ_2	Order
3464	0.01954593	—
14398	0.00494726	1.98
59888	0.00114908	2.04
240654	0.00028971	2.03
968090	0.00007259	2.02

5 Conclusion

A finite volume scheme for numerical solution of Maxwell's equations with dielectric permittivity discontinuity was suggested. The scheme is second order accurate in space and time. Scheme was tested using a number of problems. Test problems included linear and curvilinear dielectric permittivity discontinuities for TE and TM cases. To analyze rate of convergence calculations were performed using a sequence of five meshes. Calculation results demonstrate second order of convergence and confirm second order of approximation of the proposed scheme.

A Order of approximation of Lebedev's scheme

Here we will prove that scheme suggested in [10] has only first order of approximation. To accomplish this we will show that calculation of derivatives using continuous variables without choosing stencil with constant dielectric permittivity will have zero order of approximation for some cells. As a result approximation of values at edge centers and approximation of the resulting scheme will be first order.

Assume that derivative calculation using continuous variables suggested in [10] has order of approximation higher than zero. Then for any continuous function and a sequence of meshes with cell sizes tending to zero calculated derivative values should tend to exact values. Consider a computational region in the form of a regular hexagon with two opposite vertices lying on the x_2 axis. In this region consider a sequence of triangular meshes. As a first mesh we will choose six equilateral triangles of the regular hexagon. Every following mesh will be obtained from the previous one by dividing every equilateral triangle into four equilateral triangles. As a continuous function we choose

$$f(x_1, x_2) = \begin{cases} 0, & x_1 < 0; \\ x_1, & x_1 \geq 0\ . \end{cases} \tag{28}$$

Consider cells that are not adjacent to the border of computational region. For cells adjacent to the x_2 axis on the left derivative with respect to x_1 will tend to $1/9$, and for cells adjacent to the x_2 axis on the right derivative with respect to x_1 will tend to $66/81$. Contradiction. The order of approximation for derivative calculation in [10] is zero.

 ISBN: 978-84-614-6167-7

References

[1] S. M. Rao, *Time Domain Electromagnetics*, Academic Press, San Diego, 1999.

[2] A. Taflove, *Advances in Computational Electrodynamics: the Finite-Difference Time-Domain Method*, Artech House, Boston, 1998.

[3] K. S. Yee, *Numerical solution of initial boundary value problems involving Maxwell's equations in isotropic media*, IEEE Trans. Antennas Propagat. **17** (1966) 585–589.

[4] D. M. Sullivan, *Electromagnetic Simulation Using the Finite-Difference Time-Domain Method*, IEEE, New York, 2000.

[5] A. Taflove and S. C. Hagness, *Computational Electrodynamics: the Finite-Difference Time-Domain Method*, Artech House, Boston, 2000.

[6] R. J. LeVeque, *Numerical methods for conservation laws*, Birkhäuser, Basel, 1990.

[7] V. Shankar, W. F. Hall, A. H. Mohammadian, *A CFD-based finite-volume procedure for computational electromagnetics - interdisciplinary applications of CFD methods*, AIAA **A89-41776 18-02** (1989) 551–564.

[8] F. Hermeline, *Two coupled particle-finite volume methods using Dalaunay-Voronoi meshes for applorimation of Vlasov—Poisson and Vlasov — Maxwell equations*, J. Comput. Phys. **106** (1993) 1–18.

[9] J.-P. Cioni, L. Fezoui, D. Issautier, *Higher order upwind schemes for solving time domain Maxwell equations*, La Recherche Aĕrospatiale **5** (1994) 319–328.

[10] A. S. Lebedev, M. P. Fedoruk, and O. V. Shtyrina, *Finite volume algorithm for solving the time-dependent Maxwell equations on unstructured meshes*, Journal of Computational Mathematics and Mathematical Physics, **47** (2006) 1219–1233.

[11] S. K. Godunov, *A Difference Scheme for Numerical Solution of Discontinuos Solution of Hydrodynamic Equations*, Math. Sbornik **47** (1959) 271–306.

[12] P. D. Lax and B. Wendroff, *Systems of conservation laws*, Commun. Pure Appl. Math. **13** (1960) 217–237.

[13] B. Van Leer, *Towards the ultimate conservative difference scheme. A second order sequel to Godunov's method*, J. Comput. Phys. **32** (1979) 101–136.

[14] A. J. Laub, *Matrix Analysis for Scientists and Engineers*, SIAM, Philadelphia, 2004.

[15] C. Geuzaine and J.-F. Remacle, *Gmsh: a three-dimensional finite element mesh generator with built-in pre- and post-processing facilities*, Int. J. Numer. Meth. Engng. **3** (2009) 1–24.

 ISBN: 978-84-614-6167-7

Proceedings of the 11th International Conference on Computational and Mathematical Methods in Science and Engineering, CMMSE 2011 26–30 June 2011.

First steps in the mathematical modeling of a bioreactor behavior

Riccardo Jadanza[1], Luisa Testa[1], Shinnosuke Oharu[2] and Ezio Venturino[1]

[1] *Dipartimento di Matematica "Giuseppe Peano", Università di Torino, via Carlo Alberto 10, 10123 Torino, Italy*

[2] *Department of Mathematics, Chuo University, Tokyo, Japan*

emails: riccardo.jadanza@gmail.com, testa_luisa@libero.it, oharu@gug.math.chuo-u.ac.jp, ezio.venturino@unito.it

Abstract

In this paper we model the functioning of reactor for the treatment of waste, which makes use of effective microorganisms. The functioning of the system is described by a mathematical model, whose space-independent form is then theoretically analysed. For the more complex model including diffusion and tranport terms, we use simulations. The ultimate goal of the investigation is the control of the release of the maleodorant volatile acid fats produced by the reactions.

Key words: bioreactor, interacting population models, effective microorganisms, dynamical systems, diffusion, reaction, transport
MSC 2000: AMS codes (92D25)

1 Introduction

The starting point of this investigation is the plenary lecture given by one of the authors (S.O.) at the CMMSE 2010 conference, [1]. Our aim here is to model the functioning of a biological reactor, used for sewage processing. Two bacteria populations are present, one which is facultative anaerobic, which gets advantage from the presence of oxygen, and the other one that tolerates only low concentrations of oxygen. Nutrients for these populations are represented by elements like calcium, phosphorous, natrium, potassium and magnesium. Nutrient supply is provided regularly by the waste dumped into the reactor. The aerotolerant bacteria produce adenosin-triphosphate (ATP) due to fermentation. During this process they generate volatile acid fats, which are then reabsorbed by the other facultative anaerobic bacteria. These volatile fats are maleodorant, and therefore their release needs to be controlled. The model we develop and analyse may provide insights on how to achieve this control.

 ISBN: 978-84-614-6167-7

2 The model

The mathematical model is based on the following system of partial differential equations describing an advection-reaction-diffusion system

$$\begin{cases} u_t = \Delta u + \gamma_{11} O u + \gamma_{12} N u - \gamma_{13} u, \\ v_t = \Delta v + \gamma_{21}(O_{\text{sat}} - O)u + \gamma_{22} N v - \gamma_{23} v, \\ N_t = \gamma_{41} \boldsymbol{V} \cdot \nabla N - \gamma_{41} u - \gamma_{42} v + N_{\text{supply}}, \\ O_t = \Delta O - \gamma_{51} u - \gamma_{53} \boldsymbol{p} \cdot \nabla O, \\ F_t = \Delta F + \gamma_{71} N v - \gamma_{72}(O_{\text{sat}} - O)u + \boldsymbol{V} \cdot \nabla F. \end{cases} \tag{1}$$

All the variables are functions of space $\boldsymbol{x} \in \mathbb{R}^n$, $n \leqslant 3$, and time $t \in \mathbb{R}_+$. Here $u(\boldsymbol{x}, t)$ denotes the population density of facultative anaerobic bacteria; $v(\boldsymbol{x}, t)$: is the density of the aerotolerant bacteria; $N(\boldsymbol{x}, t)$ is the concentration of nutrients (Ca, P, Na, K, Mg); $O(\boldsymbol{x}, t)$ is the oxygen concentration in the control system; $F(\boldsymbol{x}, t)$ represents the fermentation level, i.e. the concentration of the volatile fat acids.

The vectors $\boldsymbol{V}$ and $\boldsymbol{p}$, assumed to be uniform and stationary, i.e. constant vectors in space and time, represent the tranport directions of N, F and O, respectively. All other parameters are positive. In addition to the transport terms, the equations contain obviously also diffusion terms.

The first equation models the evolution of the facultative anaerobic bacteria, which have different reproduction terms, $\gamma_{11} O u$ and $\gamma_{12} N u$ since they reproduce both in absence as well as in presence of oxygen. When oxygen is present, their reproduction is enhanced. The last term in the first equation represents their mortality.

The second equation describes the aerotolerant bacteria v with a similar behavior, but with the difference that in place of O we find the difference $O_{\text{sat}} - O$, to model the fact that if the oxygen concentration falls below a certain saturation threshold O_{sat}, which represents the maximum level of oxygen that can be tolerated by these bacteria, the population thrives by feeding on the wastes produced by the u population. If instead $O \geqslant O_{\text{sat}}$, the corresponding term $\gamma_{21}(O_{\text{sat}} - O)u$ becomes an additional mortality, independent of the population size v. Natural mortality instead appears again as the last term in the equation.

The first term in the third equation denotes nutrients N transport. Their temporal evolution is further regulared by consumption by the two kinds of bacteria, and by their regular supply N_{supply}.

Oxygen dynamics comes next. It diffuses, and it is used only by the bacteria u for respiration. It also moves within the reactor.

The anaerobic bacteria v produce energy by the fermentation process, in this way they contribute to the growth of F. This explains the second term in the last equation, the first one clearly accounting for diffusion. Instead the third term describes the fact that the facultative anaerobic bacteria u, use fermentation only if the oxygen level is below the threshold O_{sat}; otherwise, if the oxygen suffices, they produce ATP by cellular respiration, and therefore contribute to a decrease in the level of F. This behavior of

 ISBN: 978-84-614-6167-7

the population U is highly desirable, since in ideal conditions it allows to control the volatile acid fats.

In this model formulation however a small incongruence appears, which will be resolved further on. A consistent interpretation of this behavior should entail a positive sign of the term $\gamma_{72}(O_{\text{sat}} - O)u$, in the last equation of (1).

3 The zero-dimensional model

We analyse at first the local dynamics, ignoring therefore the space dependence in (1). The simplified model reads then

$$\begin{cases} \dfrac{du}{dt} = \gamma_{11}Ou + \gamma_{12}Nu - \gamma_{13}u, \\ \dfrac{dv}{dt} = \gamma_{21}(O_{\text{sat}} - O)u + \gamma_{22}Nv - \gamma_{23}v, \\ \dfrac{dN}{dt} = -\gamma_{41}u - \gamma_{42}v + N_{\text{supply}}, \\ \dfrac{dO}{dt} = -\gamma_{51}u, \\ \dfrac{dF}{dt} = \gamma_{71}Nv - \gamma_{72}(O_{\text{sat}} - O)u. \end{cases} . \tag{2}$$

Looking for equilibria, we find the following region

$$E_1 := \big\{(u, v, N, O, F) \in \mathbb{R}^5_+ : u = 0, v = 0 \text{ per ogni } t \geqslant 0\big\},$$

which is diffeomorphic to the first octant of $\mathbb{R}^3$, but this happens if and only if N_{supply} vanishes. This equilibrium is not unique, therefore, and moreover not realistic, as when the reactor functions, in fact wastes are continuously inputted into it. If instead the system includes nutrient supply, $N_{\text{supply}} \neq 0$ and the following condition holds,

$$\frac{\gamma_{13}\gamma_{22} - \gamma_{12}\gamma_{23}}{\gamma_{11}\gamma_{22}} = \frac{\gamma_{13}}{\gamma_{11}},$$

which is equivalent to

$$\gamma_{12}\gamma_{23} = 0,$$

another region of equilibria exists, namely

$$E_2 := \left\{(u, v, N, O, F) \in \mathbb{R}^5_+ : u = 0, v = \frac{N_{\text{supply}}}{\gamma_{42}}, N = \frac{\gamma_{13}}{\gamma_{12}} - \frac{\gamma_{13}}{\gamma_{22}}, O = \frac{\gamma_{13}}{\gamma_{11}}\right\}.$$

In spite of the fact that the above conditions on the parameters are very unlikely to be verified, so that we could say that the probability of finding this equilibrium region is almost zero, it turns out that also this set of equilibria is generally unstable, as the Jacobian matrix possesses vanishing eigenvalues. In conclusion, this system does not admit equilibria, when the reactor works, i.e. with a nonzero supply of nutrients.

 ISBN: 978-84-614-6167-7

4 Implementation

Due to the lack of much information gathered from the analytic approach, we have then used a numerical approach to try to better understand the behavior of (2). At first, we use the Matlab routine `ode45`, with initial condition the constant vector $(1, 1, 1, 1, 1)$, with all parameters initially at the value 1. But this leads to some problems, as some variables become negative. This occurs also by modifying the parameter values. We attribute this bad behavior to the oversimplification made in (2).

So at this point we consider the full model (1) in which however we still ignore the transport terms. With a finite difference method in one space dimension, [3], with suitable time stepsize so as to satisfy the von Neumann stability condition, with the same above initial condition and parameter values, the Dirichlet boundary condition keeps all the variables at a constant value for all times. Here the solutions are meaningful for small time intervals, as they are nonnegative, although for longer time simulations they still become negative. Moreover, the boundary condition does not seem to be realistic.

We then include a different initial condition and above all a Neumann boundary condition, modeling the absence of flux toward the exterior. For the previous parameter values however, the results are once again not realistic. Therefore we add also the transport terms but also this modification does not change much the model's behavior. We impose then a restriction on the model, by setting and keeping to the value zero the variables that eventually become negative, thus considering only the positive part of the solutions, which is necessary from the biological viewpoint.

Finally we try to investigate the influence of each and every parameter on the solutions behavior, to get an optimal parameter configuration. This is related in the next Section.

5 Simulations

5.1 Parameter settings

We use the following measurement units: bacteria are counted in pure numbers:

- $\star$ $\mathbf{m}^n$ for the space, where n denotes the dimension $\mathbb{R}^n$, $n \leqslant 3$;
- $\star$ $\mathbf{h}$ for time;
- $\star$ $\mathbf{m}^{-n}$ for the bacteria density;
- $\star$ $\dfrac{\mathbf{kg}}{\mathbf{m}^n}$ for the nutrient, oxygen and volatile fat acids densities.

Searching the literature for getting reasonable parameter values, [2], we estimate the bacteria population density in the reactor between $4 \cdot 10^{13}$ and $8 \cdot 10^{13}$ bacteria per m^3, assuming a soil density between 1000 and 2000 kg/m^3 and an average dimension of the bacteria of around tens of micrometers. For the reproduction rate we have assumed an average splitting rate between 0.17 and 6 fissions per hour. For their mortality we estimate a value of about $1 \cdot 10^{13}$ batteria per hour. The density for oxygen in air is 0.28

 ISBN: 978-84-614-6167-7

Parameter	Value	Interpretation
γ_{11}	0.5000	Reproduction rate of u due to oxygen
γ_{12}	2.5000	Reproduction rate of u due to nutrients
γ_{13}	1.0000	Mortality of u
γ_{21}	0.0010	Reproduction rate of v due to $(O_{\text{sat}} - O)u$
γ_{22}	2.5000	Reproduction rate of v due to nutrients
γ_{23}	1.0000	Mortality of v
γ_{41}	0.0001	Nutrient consumption rate by u
γ_{42}	0.0001	Nutrient consumption rate by v
γ_{51}	0.0010	Oxygen consumption rate by u
γ_{53}	0.1000	Oxygen transport
γ_{71}	0.0100	F production rate by v
γ_{72}	0.0100	F consumption rate by u
N_{supply}	0.0510	Nutrient input rate
O_{sat}	0.0500	Oxygen saturation threshold

Table 1: Parameter values used in the simulations

kg/m^3. The density of the fermentation gases is assumed equal to the one of the air, at environmental pressure and temperature, i.e. 1.2 kg/m^3. The one of the nutrients in the wastes is assumed to be around one gram per cubic meter.

In Table 1 we report all the parameter values used in the numerical simulation. Note that the saturation value of oxygen, O_{sat}, has been set to about 20% of the oxygen density in air, a consistent assumption with the aerotolerant nature of the bacteria v. The vectors $\boldsymbol{p}$ and $\boldsymbol{V}$ are set equal to 1 in the monodimensional case and equal to the vector $(1, 1)$ in the bidimensional case.

5.2 The monodimensional case

We consider a bioreactor length $L = 15$ m and we simulate the temporal evolution of the system for one week, i.e. $T = 168$ hours. We also assume that the nutrients input N_{supply} occurs only once a day. The initial conditions are given by a Gaussian function, centered at $x = L/2$, for all variabiles but for oxygen, which is taken to be constant, uniformly distributed in the reactor at time $t = 0$. Figure 1 reports the graph of the first two system's variables. There are several equispaced peaks, one per day. The first one is higher because it is still part of the transient period. The remaining ones are slightly decreasing in time. Analog configurations arise for the remaining variables, N and F. For the facultative anaerobic bacteria u however, these maxima slightly decrease in time, for N they are constant, for v and F they slightly increase. Our heavy experimental evidence shows that the frequency of such oscillations is regulated by the parameter N_{supply}, their amplitude is instead directly proportional to O_{sat}, while the growth is directly proportional to γ_{21}. However, if the latter parameter becomes larger than $\mathcal{O}(10^{-2})$, F exhibits a higher slope, Figure 2 (right). Fermentation is kept

 ISBN: 978-84-614-6167-7

low also by the fact that the u population is about 10 times larger than that of the aerotolerant bacteria.

A particular case is given by $\gamma_{21} = 0$: here the peaks for u, v and F are stationary. In this way, after the transient regime of just one day, the periodic values of F attain a very reasonable range, usable also for longer time runs. From Figure 2 (center) oxygen decays irreversibly to zero after about a day, i.e. after the transient phase. This appears not to be very realistic. In fact, during the waste input also some oxygen should penetrate into the system. This suggests a possible improvement in the model formulation, to be discussed later on in Section 6.

All these remarks hold true also for much longer simulation times. In fact we have checked them by simulating the system evolution over a time period up to 10 weeks.

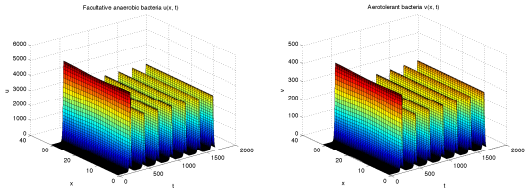


Figure 1: Model (1). Temporal evolution of $u(x,t)$ (left), $v(x,t)$ (right).

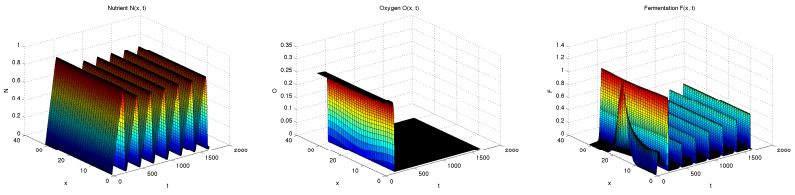


Figure 2: Model (1). Temporal evolution of $N(x,t)$ (left), $O(x,t)$ (center), $F(x,t)$ (right).

5.3 The bidimensional case

We now add one space dimension, leaving out at first for simplicity the diffusion and transport terms, since ultimately we are interested in the steady state behavior and these terms influence mainly the transient behavior. The bioreactor is now represented by a rectangle with an area of 20×30 m^2. We reuse the previous parameter values and reproduce all the behaviors observed in the one dimensional case, Fig.s 3 and 4. Obvious changes are clearly the fact that now we have circular waves, in view of the radial simmetry of the initial conditions.

However, all the configurations here found are highly unstable. In the one-dimensional case it is indeed very difficult to tune the parameters to find satisfactory results. In the

 ISBN: 978-84-614-6167-7

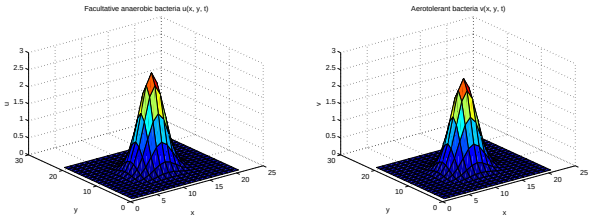


Figure 3: Model (3). Bidimensional simulations: $u(x,t)$ (left), $v(x,t)$ (right).

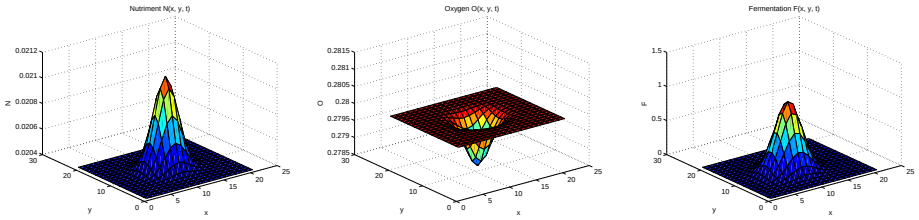


Figure 4: Model (3). Bidimensional simulations: $N(x,t)$(left), $O(x,t)$ (center), $F(x,t)$ (right).

two-dimensional case, a bad tuning of the parameters leads to integration errors that take over.

6 Possible model improvements

The original model has some shortcomings and inaccuracies. We now examine some changes in the equations and investigate their influence on the system's evolution. The simulations are here run up to two weeks time.

6.1 The role of nutrients

First of all, it seems reasonable to substitute $\gamma_{11}Ou$ with $\gamma_{11}ONu$, since, in absence of nutrient, bacteria are bound to die out. The reproduction term in this way is tied to the presence of nutrient and if the latter is absent, the first equation contains only mortality in addition to diffusion. The first modified system is then

$$\begin{cases} u_t = \Delta u + \gamma_{11}ONu + \gamma_{12}Nu - \gamma_{13}u, \\ v_t = \Delta v + \gamma_{21}(O_{\text{sat}} - O)u + \gamma_{22}Nv - \gamma_{23}v, \\ N_t = \gamma_{41}\boldsymbol{V}\cdot\nabla N - \gamma_{41}u - \gamma_{42}v + N_{\text{supply}}, \\ O_t = \Delta O - \gamma_{51}u - \gamma_{53}\boldsymbol{p}\cdot\nabla O, \\ F_t = \Delta F + \gamma_{71}Nv - \gamma_{72}(O_{\text{sat}} - O)u + \boldsymbol{V}\cdot\nabla F. \end{cases} \tag{3}$$

 ISBN: 978-84-614-6167-7

The same modification is not applied to the term $\gamma_{21}(O_{\text{sat}} - O)u$ in the second equation since this can be interpreted as a second nutrient source for the bacteria v. With the same parameter values used in the original model, the fermentation level rapidly increases, reaching unacceptable values for an optimal functioning of the reactor. We thus diminish the reproduction rate of v due to nutrient N uptake, by slightly changing the value of γ_{22} from 2.5 to 2.1. In this way, the peaks reduce sensibly, up almost to disappearing. Consequently F drops to zero after a brief transient phase, Figure 6 (right). We will use this modification also in all subsequent changes.

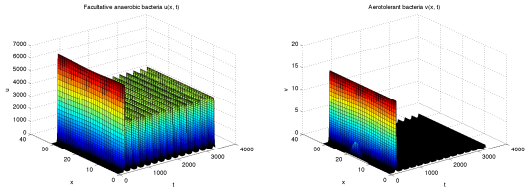


Figure 5: Model (3). Temporal evolution of $u(x,t)$ (left), $v(x,t)$ (right).

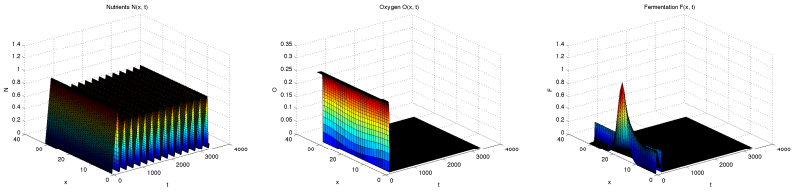


Figure 6: Model (3). Temporal evolution of $N(x,t)$(left), $O(x,t)$ (center), $F(x,t)$ (right).

6.2 The role of facultative anaerobic bacteria

We rewrite (3) by changing the sign of $\gamma_{72}(O_{\text{sat}} - O)u$, to agree with the definition of facultative anaerobic bacteria. We need to insert a term $O_{\text{supply}} \approx \mathcal{O}(10^{-1})$ in the oxygen equation, since fermentation is too quick, to induce a periodic raise in the bacteria that feed on fermentation products. The equilibrium value of F settles to a level inversely proportional to the new parameter O_{supply}. The model thus reads

$$\begin{cases} u_t = \Delta u + \gamma_{11} ONu + \gamma_{12} Nu - \gamma_{13} u, \\ v_t = \Delta v + \gamma_{21}(O_{\text{sat}} - O)u + \gamma_{22} Nv - \gamma_{23} v, \\ N_t = \gamma_{41} \boldsymbol{V} \cdot \nabla N - \gamma_{41} u - \gamma_{42} v + N_{\text{supply}}, \\ O_t = \Delta O - \gamma_{51} u - \gamma_{53} \boldsymbol{p} \cdot \nabla O + O_{\text{supply}}, \\ F_t = \Delta F + \gamma_{71} Nv + \gamma_{72}(O_{\text{sat}} - O)u + \boldsymbol{V} \cdot \nabla F. \end{cases} \tag{4}$$

 ISBN: 978-84-614-6167-7

The oxygen density becomes periodic, Fig. 8 (center), even though introducing O_{supply} increases the number of peaks, which in this way are no longer tied to the feeding once a day of waste material. Fermentation becomes periodic within acceptable levels, Fig. 8 (right). In the following models, except the last one, the third term in the last equation will always have a positive sign.

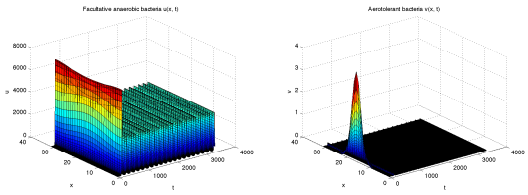


Figure 7: Model (4). Temporal evolution of $u(x,t)$ (left), $v(x,t)$ (right).

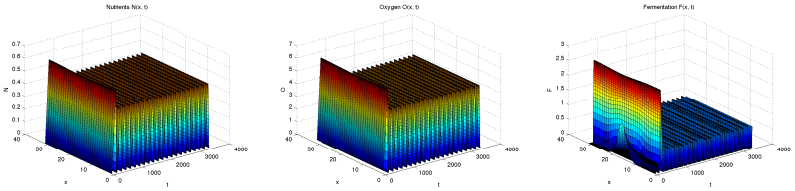


Figure 8: Model (4). Temporal evolution of $N(x,t)$ (left), $O(x,t)$ (center), $F(x,t)$ (right).

6.3 The role of nutrient reproduction

Here we substitute $\gamma_{21}(O_{\text{sat}} - O)u$ with $\gamma_{21}(O_{\text{sat}} - O)Nv$, changing its interpretation now either as a reproduction term due to nutrients N, if $O < O_{\text{sat}}$, or mortality, if $O > O_{\text{sat}}$. Thus the aerotolerant character of v becomes clearer for low oxygen concentrations. Again with the original parameter values we find an excessive increase of F. To avoid it, we add again O_{supply}. The system becomes

$$\begin{cases} u_t = \Delta u + \gamma_{11} ONu + \gamma_{12} Nu - \gamma_{13} u, \\ v_t = \Delta v + \gamma_{21}(O_{\text{sat}} - O)Nv + \gamma_{22} Nv - \gamma_{23} v, \\ N_t = \gamma_{41} \boldsymbol{V} \cdot \nabla N - \gamma_{41} u - \gamma_{42} v + N_{\text{supply}}, \\ O_t = \Delta O - \gamma_{51} u - \gamma_{53} \boldsymbol{p} \cdot \nabla O + O_{\text{supply}}, \\ F_t = \Delta F + \gamma_{71} Nv + \gamma_{72}(O_{\text{sat}} - O)u + \boldsymbol{V} \cdot \nabla F. \end{cases} \tag{5}$$

Further, in this way oxygen has still a periodic behavior, Fig. 10 (center), and v attains lower values than those reached in the model (1).

 ISBN: 978-84-614-6167-7

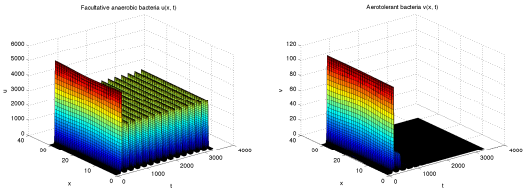


Figure 9: Model (5). Temporal evolution of $u(x,t)$ (left), $v(x,t)$ (right).

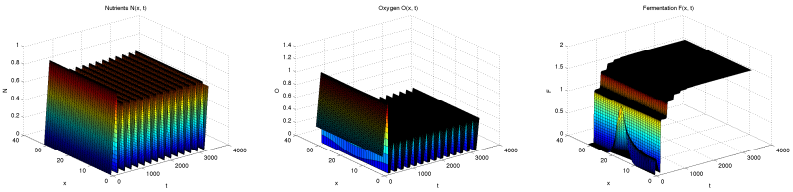


Figure 10: Model (5). Temporal evolution of $N(x,t)$ (left), $O(x,t)$ (center), $F(x,t)$ (right).

6.4 A generalization

We try now to generalize the previous cases. The term $\gamma_{21}(O_{\text{sat}} - O)u$ is changed into $\gamma_{21}(O_{\text{sat}} - O)Nuv$ and O_{supply} is again used

$$\begin{cases} u_t = \Delta u + \gamma_{11} ONu + \gamma_{12} Nu - \gamma_{13} u, \\ v_t = \Delta v + \gamma_{21}(O_{\text{sat}} - O)Nuv + \gamma_{22} Nv - \gamma_{23} v, \\ N_t = \gamma_{41} \boldsymbol{V} \cdot \nabla N - \gamma_{41} u - \gamma_{42} v + N_{\text{supply}}, \\ O_t = \Delta O - \gamma_{51} u - \gamma_{53} \boldsymbol{p} \cdot \nabla O + O_{\text{supply}}, \\ F_t = \Delta F + \gamma_{71} Nv + \gamma_{72}(O_{\text{sat}} - O)u + \boldsymbol{V} \cdot \nabla F. \end{cases} \tag{6}$$

As in the previous case, with the values of Table 1, we find a quick decrease to zero of v, while F stabilizes at acceptable levels, Fig. 12 (right). All other variables are almost periodic.

 ISBN: 978-84-614-6167-7

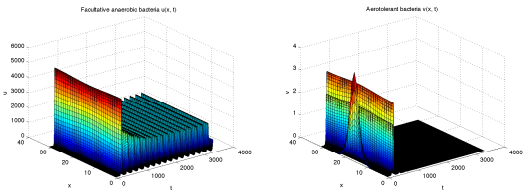


Figure 11: Model (6). Temporal evolution of $u(x,t)$ (left), $v(x,t)$ (right).

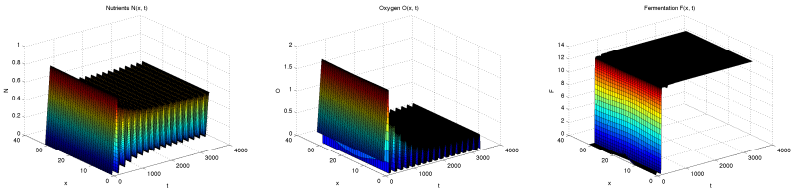


Figure 12: Model (6). Temporal evolution of $N(x,t)$ (left), $O(x,t)$ (center), $F(x,t)$ (right).

6.5 The role of obligated aerobic bacteria

In the last case we assume that the bacteria u are obligated aerobic instead of being anaerobic facultative. Thus we need to set γ_{12} in (3) to zero. In this way reproduction of u is tied to presence of oxygen. Next we replace $\gamma_{72}(O_{\text{sat}}-O)u$ by $\gamma_{72}Ou$, to indicate consumption of F by u independently of the oxygen level. Once again we introduce the parameter $O_{\text{supply}} = 0.3$ to guarantee the presence of the obligated aerobic bacteria. The system becomes

$$\begin{cases} u_t = \Delta u + \gamma_{11} ONu - \gamma_{13} u, \\ v_t = \Delta v + \gamma_{21}(O_{\text{sat}} - O)u + \gamma_{22} Nv - \gamma_{23} v, \\ N_t = \gamma_{41} \boldsymbol{V} \cdot \nabla N - \gamma_{41} u - \gamma_{42} v + N_{\text{supply}}, \\ O_t = \Delta O - \gamma_{51} u - \gamma_{53} \boldsymbol{p} \cdot \nabla O + O_{\text{supply}}, \\ F_t = \Delta F + \gamma_{71} Nv - \gamma_{72} Ou + \boldsymbol{V} \cdot \nabla F. \end{cases} \tag{7}$$

All variables now show a periodic behavior, Fig.s 13 and 14. The height of the peaks for F is inversely proportional to O_{supply}.

References

[1] S. Oharu, Y. Matsuura, T. Arima, *Flow analysis around structures in slow fluids and its applications to the environmental fluid phenomena*, Proceedings of

 ISBN: 978-84-614-6167-7

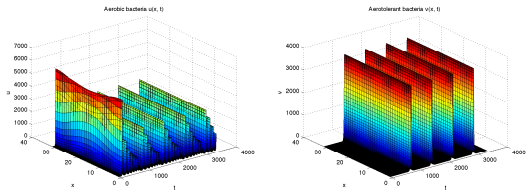


Figure 13: Model (7). Temporal evolution of $u(x,t)$ (left), $v(x,t)$ (right).

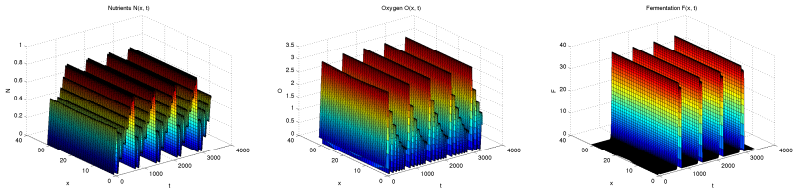


Figure 14: Model (7). Temporal evolution of $N(x,t)$ (left), $O(x,t)$ (center), $F(x,t)$ (right).

the 10th International Conference on Computational and Mathematical Methods in Science and Engineering, CMMSE, **1** (2010) 718–729.

[2] U. Schöttler, *On the Anaerobic Metabolism of Three Species of Nereis (Annelida)*, Marine Ecology Progress Series **1** (1979) 249–254.

[3] G.D. Smith, *Numerical solution of partial differential equations: finite difference methods*, Oxford Univ. Press, 1978.

 ISBN: 978-84-614-6167-7

Proceedings of the 11th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2011
26–30 June 2011.

A Sound Semantics for Bousi∼Prolog

Pascual Julian Iranzo[1] **and Clemente Rubio Manzano**[1]

[1] *Department of Information Technologies and Systems, University of Castilla-La Mancha*

emails: Pascual.Julian@uclm.es, Clemente.Rubio@uclm.es

Abstract

Bousi∼Prolog is an extension of the standard Prolog language aiming at to make more flexible the query answering process and to deal with vagueness applying declarative techniques. In this paper we precise a model-theoretic semantics for a pure subset of this language, we recall both the WSLD-resolution principle and a similarity-based unification algorithm which is the basis of its operational mechanism and then we prove the soundness of WSLD-resolution.

Key words: Fuzzy Logic Programming, (Least) Fuzzy Herbrand Model, Fixpoint Semantics, Weak Unification, Weak SLD-Resolution, Proximity/Similarity Relations.

1 Introduction

In recent years there has been a renewed interest in amalgamating *Logic Programming* [11] with concepts coming from *Fuzzy Logic* [13] or akin to this field. As tokens of this interest we mention the works on *Fuzzy Logic Programming* [3, 7, 12], *Qualified Logic Programming* [8, 1] (which is a derivation of the van Emden's *Quantitative Logic Programming* [10]) or *Similarity-Based Logic Programming* [2, 9]. Bousi∼Prolog is a representative of the last class of fuzzy logic programming languages. It replaces the syntactic unification mechanism of the classical **S**election-function driven **L**inear resolution for **D**efinite clauses (SLD–resolution) by a fuzzy unification algorithm based on fuzzy binary relations on a syntactic domain. The result is an operational mechanism, called Weak SLD-resolution, which differs in some aspects w.r.t. the one of [9], based exclusively on similarity relations.

This work can be seen as a continuation of the investigation started in [4]. In this paper after introducing some refinements to the model-theoretic and fix-point semantics of Bousi∼Prolog defined in [4] for definite programs, we present, for the first time in our framework, the concept of a *correct answer* providing a declarative description for

 ISBN: 978-84-614-6167-7

the output of a program and a goal. It is noteworthy that, although the refinements introduced in the declarative semantics do not dramatically alter the original definitions, given in [4], they are important in order to establish the soundness of our proposal. Afterwards, we recall the operational semantics of Bousi∼Prolog and we prove, among other results, its soundness. The soundness theorem is established following a proof strategy comparable with the one appeared in [5]. It is important to remark that, the soundness in our framework will be proven under certain conditions. To be precise, we only consider programs without negation and we restrict ourselves to similarity relations on syntactic domains. Finally, it is worthy to say that, along this paper we also clarify some of the existing differences between our framework and the related proposal introduced by [9].

2 Preliminaries.

2.1 Fuzzy relations, proximity and similarity relations

A *binary fuzzy relation* on a set U is a fuzzy subset on $U \times U$ (that is, a mapping $U \times U \longrightarrow [0,1]$). There are some important properties that fuzzy relations may have: i) (Reflexivity) $\mathcal{R}(x,x) = 1$ for any $x \in U$; i) (Symmetry) $\mathcal{R}(x,y) = \mathcal{R}(y,x)$ for any $x, y \in U$; i) (Transitivity) $\mathcal{R}(x,z) \geq \mathcal{R}(x,y) \wedge \mathcal{R}(y,z)$ for any $x, y, z \in U$; where the operator '$\wedge$' is the minimum t-norm. A *proximity relation* is a binary fuzzy relation which is reflexive and symmetric. A proximity relation is characterized by a set $\Lambda = \{\lambda_1, ..., \lambda_n\}$ of approximation levels. We say that a value $\lambda \in \Lambda$ is a *cut value*. A special, and well-known, kind of proximity relations are *similarity* relations, which are nothing but transitive proximity relations.

In classical logic programming different syntactic symbols represent distinct information. Following [9], this restriction can be relaxed by introducing a proximity or similarity relation $\mathcal{R}$ on the alphabet of a first order language. This makes possible to treat as indistinguishable two syntactic symbols which are related by the proximity or similarity relation $\mathcal{R}$ with a certain degree greater than zero. A similarity relation $\mathcal{R}$ on the alphabet of a first order language can be extended to terms by structural induction in the usual way:

1. $\mathcal{R}(x,x) = 1$;
2. Let f and g be two n-ary function symbols and let $t_1, \ldots, t_n, s_1, \ldots, s_n$ be terms. $\mathcal{R}(f(t_1, \ldots, t_n), g(s_1, \ldots, s_n)) = \mathcal{R}(f,g) \wedge (\bigwedge_{i=1}^{n} \mathcal{R}(t_i, s_i))$;

Otherwise, the approximation degree of two expressions is zero. The extension for atomic formulas and compound formulas can be done in an analogous form. See [9] for a precise characterization of this problem. The extension of a proximity relation is more cumbersome[1] and it is not addressed in this paper.

[1] We know that a naive approach may cause the incompleteness of the weak SLD resolution procedure. Moreover, a indispensable property of an inference relation such as the cut rule is not fulfilled.

 ISBN: 978-84-614-6167-7

2.2 Formulas, interpretations and truth in the context of a proximity relation

In this section we discuss the notions of interpretation and truth for a first order theory in the context of a proximity relation. A *fuzzy Interpretation* $\mathcal{I}$ of a first order language $\mathcal{L}$ is a pair $\langle \mathcal{D}, \mathcal{J} \rangle$ where $\mathcal{D}$ is the domain of the interpretation and $\mathcal{J}$ is a mapping which assigns meaning to the symbols of $\mathcal{L}$: specifically n-ary relation symbols are interpreted as mappings $\mathcal{D}^n \longrightarrow [0,1]$. In order to evaluate open formulas we need to introduce the notion of variable assignment. A *variable assignment*, ϑ, w.r.t. $\mathcal{I} = \langle \mathcal{D}_\mathcal{I}, \mathcal{J} \rangle$, is a mapping $\vartheta : \mathcal{V} \longrightarrow \mathcal{D}_\mathcal{I}$, which can be extended to the set of the terms of $\mathcal{L}$ by structural induction. Given a fuzzy interpretation $\mathcal{I} = \langle \mathcal{D}, \mathcal{J} \rangle$ and a variable assignment ϑ in $\mathcal{I}$, the *valuation* of a formula w.r.t. $\mathcal{I}$ and ϑ is[2]:

$$\mathcal{I}(p(t_1, \ldots, t_n))[\vartheta] = \bar{p}(t_1\vartheta, \ldots, t_n\vartheta), \text{ where } \mathcal{J}(p) = \bar{p}$$
$$\mathcal{I}(\mathcal{A} \wedge \mathcal{B}))[\vartheta] = \inf\{\mathcal{I}(\mathcal{A})[\vartheta], \mathcal{I}(\mathcal{B})[\vartheta]\}$$
$$\mathcal{I}(\mathcal{A} \leftarrow \mathcal{B})[\vartheta] = \text{ if } \mathcal{I}(\mathcal{Q})[\vartheta] \leq \mathcal{I}(\mathcal{A})[\vartheta] \text{ then 1 else } \mathcal{I}(\mathcal{A})[\vartheta].$$
$$\mathcal{I}((\forall x)\mathcal{A})[\vartheta] = \inf\{\mathcal{I}(\mathcal{A})[\vartheta'] \mid \vartheta' \ \ x\text{–equivalent to } \ \vartheta\}$$

where p is a predicate symbol, and $\mathcal{A}$ and $\mathcal{B}$ are formulas. An assignment ϑ' is x–equivalent to ϑ when $z\vartheta' = z\vartheta$ for all variable $z \neq x$ in $\mathcal{V}$. When the assignment would not be relevant, we shall omit it during the valuation of a formula. In the context of a first order theory equipped with a proximity relation $\mathcal{R}$, characterized by a set $\Lambda = \{\lambda_1, ..., \lambda_n\}$ of approximation levels, it makes sense that the notion of truth be linked to a certain approximation level $\lambda \in \Lambda$. For a fixed value λ and a formula $\mathcal{A}$ of $\mathcal{L}$:

$\mathcal{A}$ is λ-*true* in $\mathcal{I}$ iff for every assignment ϑ in $\mathcal{I}$, $\mathcal{I}(\mathcal{A})[\vartheta] \geq \lambda$
$\mathcal{A}$ is λ-*false* in $\mathcal{I}$ iff for every assignment ϑ in $\mathcal{I}$, $\mathcal{I}(\mathcal{A})[\vartheta] < \lambda$
$\mathcal{A}$ is λ-*valid* iff $\mathcal{A}$ is λ-true for all interpretation $\mathcal{I}$.
$\mathcal{A}$ is λ-*unsatisfiable* iff $\mathcal{A}$ is λ-false for all $\mathcal{I}$.
$\mathcal{A}$ is λ-*satisfiable* iff there exists an $\mathcal{I}$ and a ϑ in $\mathcal{I}$ such that $\mathcal{I}(\mathcal{A})[\vartheta] \geq \lambda$.

Intuitively, a cut value λ is delimiting truth degrees equal or greater than λ as true. Since the valuation of a closed formula is completely determined by an interpretation, independently of a variable assignment, we say that an interpretation $\mathcal{I}$ of $\mathcal{L}$ is λ-*model* for $\mathcal{A}$ if and only if $\mathcal{I}(\mathcal{A}) \geq \lambda$.

2.3 Closed conditional formulas and models

In this section we elucidate the notion of model for a set of closed conditional formulas in the context of a similarity relation. By conditional formula we mean a formula of the form $\mathcal{C} \equiv A \leftarrow \mathcal{Q}$, where A (called the head) is an atom, $\mathcal{Q}$ a formula (called the body) and all variables are assumed universally quantified. When $\mathcal{Q} \equiv B_1 \wedge \ldots \wedge B_n$ is

[2]Note that, $t_i\vartheta$ is the usual notation for the application of a substitution ϑ to a expression t_i. That is, it is equivalent to $\vartheta(t_i)$.

 ISBN: 978-84-614-6167-7

a conjunction of atoms, the formula $\mathcal{C}$ is called a *Horn clause* or *definite clause*. As it is well known, this kind of formulas play a special role in logic programming where a set of definite clauses is called a *program* and a *goal* is any conjunctive body. A direct naive translation to our context of the classical concept of model for a set of formulas does not work. We need a new definition supported by the notion of what we called an annotated set of formulas of level λ. In particular, if we are working with the set $\Gamma \equiv \{p(a)\}$ and the similarity defined by the entry $\mathcal{R}(a,b) = 0.8$ then the intended meaning of Γ and $\mathcal{R}$ is that we believe in $p(a)$ with truth degree 1 but also, because b is similar to a, we believe in $p(b)$ with truth degree 0.8. That is $\mathcal{R}$ induces meaning into Γ and we can reflect this fact by means of an annotated set of formulas $\{\langle p(a), 1\rangle, \langle p(b), 0.8\rangle\}$ (see [4] to obtain more intuitive insides on this idea).

We want to formalize this concept, but before doing that we need some technical definitions introduced to cope with some problems that appear when conditional formulas have non-linear atoms on their heads[3]. Given a non-linear atom A, the linearization of A (as defined in [1]) is a process by which it is computed the structure $\langle A_l, C_l\rangle$, where: A_l is a linear atom built from A by replacing each one of the n multiple occurrences of the same variable X_i by new fresh variables $Y_k (1 \leq k \leq n_i)$; and C_l is a set of proximity constrains $X_i \sim Y_k$ (with $1 \leq k \leq n_i$). The operator "$s \sim t$" is asserting the proximity of two terms s and t and when interpreted, $\mathcal{I}(s \sim t) = \mathcal{R}(s,t)$, whatever the interpretation $\mathcal{I}$ of $\mathcal{L}$. Now, let $\mathcal{C} \equiv A \leftarrow \mathcal{Q}$ be a conditional formula and $C_l = \{X_1 \sim Y_1, \ldots, X_n \sim Y_n\}$, $lin(\mathcal{C}) = A_l \leftarrow X_1 \sim Y_1 \wedge \ldots \wedge X_n \sim Y_n \wedge \mathcal{Q}$. For a set Γ of conditional formulas, $lin(\Gamma) = \{lin(\mathcal{C}) \mid \mathcal{C} \in \Gamma\}$. The following algorithm, which is a reformulation of the one that appears in [4] to cope with the linearization process, gives a precise procedure for the construction of the set of annotated formulas of level λ.

Algorithm 1

Input: *A set of conditional formulas* Γ *and a similarity relation* $\mathcal{R}$ *with a set of levels* Λ *and a cut value* $\lambda \in \Lambda$.
Output: *A set* Γ^λ *of annotated formulas of level* λ.
Initialization: $\Gamma_l := lin(\Gamma)$ *and* $\Gamma^\lambda := \{\langle \mathcal{C}, 1\rangle \mid \mathcal{C} \in \Gamma_l\}$
For each *conditional formula* $\mathcal{C} \equiv A \leftarrow \mathcal{Q} \in \Gamma_l$ **do**
  $\mathcal{K}_\lambda(\mathcal{C}) = \{\langle \mathcal{C}' \equiv A' \leftarrow \mathcal{Q}, \alpha\rangle \mid \mathcal{R}(A, A') = \alpha \geq \lambda\}$
  For each *element* $\langle \mathcal{C}', \alpha\rangle$ *in* $\mathcal{K}_\lambda(\mathcal{C})$ **do**
    If $\langle \mathcal{C}', L\rangle \in \Gamma^\lambda$ **then** $\Gamma^\lambda = (\Gamma^\lambda \setminus \{\langle \mathcal{C}', L\rangle\}) \cup \langle \mathcal{C}', L \wedge \alpha\rangle$
    else $\Gamma^\lambda = (\Gamma^\lambda \cup \{\langle \mathcal{C}', \alpha\rangle\})$
Return Γ^λ

The general idea behind this algorithm is to start annotating each formula in the set Γ_l with a truth degree equal to 1. On the other hand, the rest of the formulas generated by proximity, starting from formulas of the original set Γ_l, are annotated with its corresponding approximation degree (regarding the original formula). Afterward, if several formulas of the set generate the same approximate formula, with different

[3]The apparition of this problem in our framework was pointed out by R. Caballero, M. Rodríguez and C. Romero in a private communication. So we want to express them our gratitude.

 ISBN: 978-84-614-6167-7

approximations degrees, we take the least degree as annotation. Now we are ready to define the core concepts of model and logical consequence of level λ for a set of closed conditional formulas w.r.t. a similarity relation. Let Γ be a set of closed conditional formulas of a first order language $\mathcal{L}$, $\mathcal{R}$ be a similarity relation which is characterised by a set Λ of approximation levels with cut value $\lambda \in \Lambda$ and $\mathcal{I}$ be a fuzzy interpretation of $\mathcal{L}$.

1) $\mathcal{I}$ is λ-model for $\{\Gamma, \mathcal{R}\}$ iff for all annotated formula $\langle \mathcal{A}, \lambda' \rangle \in \Gamma^{\lambda}$, $\mathcal{I}(\mathcal{A}) \geq \lambda'$;

2) $\mathcal{A}$ is a λ-logical consequence of $\{\Gamma, \mathcal{R}\}$ if and only if for each fuzzy interpretation $\mathcal{I}$ of $\mathcal{L}$, $\mathcal{I}$ is a λ-model for $\{\Gamma, \mathcal{R}\}$ implies that $\mathcal{I}$ is a λ-model for $\mathcal{A}$.

3 Declarative Semantics.

In this section we recall the declarative semantics of Bousi∼Prolog. Roughly speaking, BPL programs are sequences of (normal) clauses plus a proximity relation. However, in the rest of the paper we restrict ourselves to definite clauses. Also observe that, since the notion of approximation and the fuzzy unification algorithm we are going to work with is only well-defined for similarity relations, in this and the following sections we solely deal with this kind of relations.

3.1 Fuzzy Herbrand interpretations and models

When we use a logic programming language whose instructions are clauses and employs a refutation procedure, it is well-known that it suffices to pay attention only on (fuzzy) Herbrand interpretations in order to determine the unsatisfiability of a set of clauses. Herbrand interpretations are defined on a syntactic domain, called the Herbrand universe. For a first order language $\mathcal{L}$, the *Herbrand universe* $\mathcal{U}_{\mathcal{L}}$ for $\mathcal{L}$, is the set of all ground terms in $\mathcal{L}$. Roughly speaking, in a Herbrand interpretation, constant and function symbols are interpreted as themselves in a fixed way while n-ary relation symbols are freely interpreted as n-ary (fuzzy) relations on $\mathcal{U}_{\mathcal{L}}$, i.e. (fuzzy) subsets on $\mathcal{U}_{\mathcal{L}}^n$ (or equivalently, mappings from $\mathcal{U}_{\mathcal{L}}^n$ into the $[0,1]$ interval). On the order hand, the *Herbrand base* $\mathcal{B}_{\mathcal{L}}$ for $\mathcal{L}$ is the set of all ground atoms which can be formed by using the predicate symbols of $\mathcal{L}$ jointly with the ground terms from the Herbrand universe taken as arguments. As in the classical case, it is possible to identify a Herbrand interpretation with a fuzzy subset of the Herbrand base. That is, a *fuzzy Herbrand interpretation* for $\mathcal{L}$ can be considered as a mapping $\mathcal{I} : \mathcal{B}_{\mathcal{L}} \longrightarrow [0,1]$. The ordering $\leq$ in the lattice $[0,1]$ can be easily extended to the set of Herbrand interpretations $\mathcal{H}$, as follows: $\mathcal{I}_1 \sqsubseteq \mathcal{I}_2$ iff $\mathcal{I}_1(A) \leq \mathcal{I}_2(A)$ for all ground atom $A \in \mathcal{B}_{\mathcal{L}}$. It is important to note that the pair $\langle \mathcal{H}, \sqsubseteq \rangle$ is a complete lattice.

In the following, we focus our attention on Herbrand λ-models. For this special kind of λ-models we proved in [4] an analogous property to the *model intersection property* and we defined the *least Herbrand model of level* λ, for a program Π and a similarity relation $\mathcal{R}$, as the mapping $\mathcal{M}_{\Pi}^{\lambda} : \mathcal{B}_{\mathcal{L}} \longrightarrow [0,1]$ such that,

$$\mathcal{M}_{\Pi}^{\lambda}(A) = \inf\{\mathcal{I}(A) \mid \mathcal{I} \text{ is a } \lambda\text{-model for } \Pi \text{ and } \mathcal{R}\},$$

 ISBN: 978-84-614-6167-7

for each $A \in \mathcal{B}_{\mathcal{L}}$. The interpretation $\mathcal{M}_{\Pi}^{\lambda}$ is the natural interpretation for a program Π and a similarity relation $\mathcal{R}$, since, as it was proved in [4], for each $A \in \mathcal{B}_{\mathcal{L}}$ such that $\mathcal{M}_{\Pi}^{\lambda}(A) \neq 0$, A is a logical consequence of level λ for Π and $\mathcal{R}$.

3.2 Fixpoint semantics

In this section we recall an alternative characterization of the least Herbrand model of level λ for a definite program Π and a similarity relation $\mathcal{R}$, using fixpoint concepts, given in [4]. The idea is to provide a constructive vision of the meaning of a program by defining an immediate consequences operator which allows to construct the least Herbrand model of level λ, by means of successive applications.

Definition 3.1 (Immediate consequences operator of level λ) *Let Π be a definite program and $\mathcal{R}$ be a similarity relation. We define the immediate consequences operator of level λ, $\mathcal{T}_{\Pi}^{\lambda}$, as a mapping $\mathcal{T}_{\Pi}^{\lambda} : \mathcal{H} \longrightarrow \mathcal{H}$ such that, for all $A \in \mathcal{B}_{\mathcal{L}}$,*

$$\mathcal{T}_{\Pi}^{\lambda}(\mathcal{I})(A) = \inf\{\mathcal{PT}_{\Pi}^{\lambda}(\mathcal{I})(A)\}$$

where $\mathcal{PT}_{\Pi}^{\lambda}$ is a non deterministic operator such that $\mathcal{PT}_{\Pi}^{\lambda}(\mathcal{I}) : \mathcal{B}_{\mathcal{L}} \longrightarrow \wp([0,1])$ and it is defined as follows: Let $\Pi_l = lin(\Pi)$,

1. *For each fact $H \in \Pi_l$, let $K_{\lambda}(H) = \{\langle H', \lambda' \rangle \mid \mathcal{R}(H, H') = \lambda' \geq \lambda\}$ be the set of approximate atoms of level λ for H. Then $\mathcal{PT}_{\Pi}^{\lambda}(\mathcal{I})(H'\vartheta) \ni \lambda'$, for all H' and assignment ϑ.*

2. *For each clause $\mathcal{C} \equiv (A \leftarrow \mathcal{Q}) \in \Pi_l$. Let $K_{\lambda}(\mathcal{C}) = \{\langle \mathcal{C}' \equiv A' \leftarrow \mathcal{Q}, \lambda' \rangle \mid \mathcal{R}(A, A') = \lambda' \geq \lambda\}$ be the set of approximate clauses of level λ for $\mathcal{C}$. Then $\mathcal{PT}_{\Pi}^{\lambda}(\mathcal{I})(A'\vartheta) \ni \lambda' \wedge \mathcal{I}(\mathcal{Q}\vartheta)$, for all $\mathcal{C}'$ and assignment ϑ.*

In [4], we proved that the immediate consequences operator (of level λ) is monotonous and continuous and the least fuzzy Herbrand model (of level λ) coincides with its least fixpoint.

3.3 Correct Answer

In this section we define the concept of a correct answer, which provide a declarative description of the desired output from a program, a similarity relation, and a goal. This is a central concept for the later theoretical developments.

Definition 3.2 (Answer of level λ.) *Let Π be a definite program, $\mathcal{R}$ be a proximity relation, which is characterised by a set Λ of approximation levels with cut value $\lambda \in \Lambda$, and $\mathcal{G}$ be a goal. An answer of level λ for $\{\Pi, \mathcal{R}\}$ and $\mathcal{G}$ is a pair $\langle \theta, \beta \rangle$ where θ is any substitution for variables of $\mathcal{G}$ and β an approximation degree such that $\lambda \leq \beta \leq 1$.*

Definition 3.3 (Correct Answer of level λ.) *Let Π be a definite program and $\mathcal{R}$ be a similarity relation, which is characterised by a set Λ of approximation levels with cut value $\lambda \in \Lambda$. Let $\mathcal{G} \equiv \leftarrow \mathcal{A}_1, ..., \mathcal{A}_k$ be a goal and $\langle \theta, \beta \rangle$ an answer of level λ for $\{\Pi, \mathcal{R}\}$ and $\mathcal{G}$. We say that $\langle \theta, \beta \rangle$ is a correct answer of level λ of level λ for $\{\Pi, \mathcal{R}\}$ and $\mathcal{G}$ if:*

 ISBN: 978-84-614-6167-7

1. $\forall(A_1, ..., A_k)\theta$ *is a* λ*-logical consequence of* $\{\Pi, \mathcal{R}\}$.
2. $\mathcal{M}^{\lambda}_{\Pi}(\forall(A_1, ..., A_k)\theta) \leq \beta$.

4 Operational Semantics.

The operational semantics of Bousi∼Prolog is an adaptation of the SLD resolution principle, where classical unification has been replaced by a fuzzy unification algorithm. In this section we recall the features of both the fuzzy unification algorithm and the resolution procedure.

4.1 Weak Unification based on similarity relations.

Bousi∼Prolog uses a weak unification algorithm that, when we work with similarity relations, coincides with the one defined by M. Sessa [9]. However, there exists some remarkable differences between our proposal and Sessa's proposal that we shall treat to put in evidence along this section. In presence of similarity relations on syntactic domains, it is possible to define an extended notion of a unifier and a more general unifier of two expressions[4].

Definition 4.1 *Let* $\mathcal{R}$ *be a similarity relation,* λ *be a cut value and* $\mathcal{E}_1$ *and* $\mathcal{E}_2$ *be two expressions. The substitution* θ *is a* weak unifier *of level* λ *for* $\mathcal{E}_1$ *and* $\mathcal{E}_2$ *w.r.t* $\mathcal{R}$ *(or* λ*-unifier) if its* unification degree*,* $\mathcal{D}eg_{\mathcal{R}}(\mathcal{E}_1\theta, \mathcal{E}_2\theta)$*, defined as* $\mathcal{D}eg_{\mathcal{R}}(\mathcal{E}_1\theta, \mathcal{E}_2\theta) = \mathcal{R}(\mathcal{E}_1\theta, \mathcal{E}_2\theta)$*, is greater than* λ*.*

Note that in Sessa's proposal the idea of "cut value" is missed. Also in order that a substitution θ be a weak unifier for $\mathcal{E}_1$ and $\mathcal{E}_2$ she put a strong constrain: the unification degree of $\mathcal{E}_1$ and $\mathcal{E}_2$ w.r.t. θ must be the maximum of the unification degrees of $\mathcal{D}eg_{\mathcal{R}}(\mathcal{E}_1\varphi, \mathcal{E}_2\varphi)$ for whatever substitution φ. Therefore, some substitution that we consider as a weak unifier, are disregarded by her proposal.

Definition 4.2 *Let* $\mathcal{R}$ *be a similarity relation and* λ *be a cut value. The substitution* θ *is* more general *than the substitution* σ with level λ*, denoted by* $\theta \leq_{\mathcal{R},\lambda} \sigma$*, if there exist a substitution* δ *such that, for any variable* x *in the domain of* θ *or* σ*,* $\mathcal{R}(x\sigma, x\sigma\delta) \geq \lambda$*.*

Definition 4.3 *Let* $\mathcal{R}$ *be a similarity relation and* $\mathcal{E}_1$ *and* $\mathcal{E}_2$ *be two expressions. The substitution* θ *is a* weak most general unifier *(w.m.g.u.) of* $\mathcal{E}_1$ *and* $\mathcal{E}_2$ *w.r.t* $\mathcal{R}$*, denoted by* $wmgu(\mathcal{E}_1, \mathcal{E}_2)$*, if: (1)* θ *is a* λ*-unifier of* $\mathcal{E}_1$ *and* $\mathcal{E}_2$*; and (2)* $\theta \leq_{\mathcal{R},\lambda} \sigma$*, for any* λ*-unifier* σ *of* $\mathcal{E}_1$ *and* $\mathcal{E}_2$*.*

The weak unification algorithm we are using is a reformulation of the one appeared in [9], which, in turn, is an extension of Martelli and Montanari's unification algorithm for syntactic unification [6]. The main difference is regarding the so called *decomposition rule* [5]: Given the unification problem $\langle\{f(t_1, \ldots, t_n) \approx g(s_1, \ldots, s_n)\} \cup E, \sigma, \alpha\rangle$, if

[4] We mean by "expression" a first order term or an atomic formula.

[5] Here, the symbol "$\mathcal{E}_1 \approx \mathcal{E}_2$" represents the potential possibility that two expressions $\mathcal{E}_1$ and $\mathcal{E}_2$ be close.

 ISBN: 978-84-614-6167-7

$\mathcal{R}(f,g)=\beta>\lambda$, it is not a failure but it is equivalent to solve the new configuration $\langle\{t_1\approx s_1,\ldots,t_n\approx s_n\}\cup E,\sigma,\alpha\wedge\beta\rangle$, where the approximation degree α has been compounded with the degree β. It is important to note that, differently to [9], the resulting approximation degree is casted by a cut value λ.

The weak unification algorithm allows us to check if a set of expressions $S=\{\mathcal{E}_1\approx\mathcal{E}'_1,\ldots,\mathcal{E}_n\approx\mathcal{E}'_n\}$ is weakly unifiable. The w.m.g.u. of the set S is denoted by $wmgu(S)$. In general, a w.m.g.u. of two expressions $\mathcal{E}_1$ and $\mathcal{E}_2$ is not unique [9]. Therefore, the weak unification algorithm computes a representative of a w.m.g.u. class.

4.2 Weak SLD-Resolution.

Let Π be a set of Horn clauses and $\mathcal{R}$ a similarity relation on the alphabet of a first order language $\mathcal{L}$. Let Λ be the set of approximation levels of $\mathcal{R}$. We define *Weak SLD* (WSLD) *resolution* as a labeled transition system $\langle\mathcal{G}oals,\mathcal{L}abels,\Longrightarrow_{\text{WSLD}}\rangle$, where $\mathcal{G}oals$ is the set of goals of $\mathcal{L}$, $\mathcal{L}abels$ is a set of triples $\langle\mathcal{C},\theta,\alpha\rangle$ (Clause, substitution, approximation degree) and whose transition relation $\Longrightarrow_{\text{WSLD}}\subseteq(\mathcal{G}oals\times\mathcal{L}abels\times\mathcal{G}oals)$ is the smallest relation that satisfies:

$$\frac{\mathcal{C}=(\mathcal{A}\leftarrow\mathcal{Q})\ll\Pi,\sigma=wmgu(\mathcal{A},\mathcal{A}')\neq fail,\ \beta=\mathcal{R}(\mathcal{A}\sigma,\mathcal{A}'\sigma))\geq\lambda}{\leftarrow\mathcal{A}',\mathcal{Q}'\ \overset{[\mathcal{C},\sigma,\beta]}{\Longrightarrow}_{\text{WSLD}}\leftarrow(\mathcal{Q},\mathcal{Q}')\sigma}$$

where $\mathcal{Q}$, $\mathcal{Q}'$ are conjunctions of atoms, the notation "$\mathcal{C}\ll\Pi$" is representing that $\mathcal{C}$ is a standardized apart clause in Π, and the value λ is a cut value in Λ, which imposes a limit to the expansion of the search space in a computation. We say that the performed step is a *step of level* λ because the computed approximation degree is greater or equal than λ. A WSLD *derivation of level* λ for $\Pi\cup\{\mathcal{G}_0\}$ and $\mathcal{R}$ is a sequence of steps of level λ: $\mathcal{G}_0\overset{[\mathcal{C}_1,\theta_1,\beta_1]}{\Longrightarrow}_{\text{WSLD}}\ldots\overset{[\mathcal{C}_n,\theta_n,\beta_n]}{\Longrightarrow}_{\text{WSLD}}\mathcal{G}_n$. That is, each $\beta_i\geq\lambda$. And a WSLD *refutation of level* λ for $\Pi\cup\{\mathcal{G}_0\}$ and $\mathcal{R}$ is a WSLD derivation of level λ for $\Pi\cup\{\mathcal{G}_0\}$ and $\mathcal{R}$: $\mathcal{G}_0\overset{\theta,\beta}{\Longrightarrow}{}^*_{\text{WSLD}}\square$, where the symbol "$\square$" stands for the empty clause, $\theta=\theta_1\theta_2\ldots\theta_n$ is the *computed substitution* and $\beta=\bigwedge_{i=1}^n\beta_i$ is its *approximation degree*. The output of a WSLD refutation is the pair $\langle\theta_{\restriction(\mathcal{V}ar(\mathcal{G})},\beta\rangle$, which is said to be the *computed answer*. Certainly, a WSLD refutation computes a family of answers, in the sense that, if $\theta=\{x_1/t_1,\ldots,x_n/t_n\}$ then, by definition, whatever substitution $\theta'=\{x_1/s_1,\ldots,x_n/s_n\}$, holding that $\mathcal{R}(s_i,t_i)\geq\lambda$, for any $1\leq i\leq n$, is also a computed substitution with approximation degree $\beta\wedge(\bigwedge_1^n\mathcal{R}(s_i,t_i))$.

Observe that our definition of similarity based SLD resolution is parameterized by a cut value $\lambda\in\Lambda$. This introduces an important conceptual distinction between our approach an the similarity based SLD resolution presented in [9]. Moreover, we differ in the way we obtain a family of computed answers (see [4] for details). This may have a determinant impact in the correctness of the overall proposal.

5 Soundness of WSLD-Resolution

In this section we establish the soundness of WSLD-Resolution, but before proving the main result of the paper we need to introduce some important intermediate lemmas.

 ISBN: 978-84-614-6167-7

Lemma 5.1 *Let Π be a definite program, $\mathcal{R}$ be a similarity relation and λ be a cut value. Given $(\mathcal{A} \leftarrow \mathcal{Q}) \in \Pi$ and $\mathcal{A}'$ an atom such that $\mathcal{R}(\mathcal{A}, \mathcal{A}') = \alpha \geq \lambda$. If $(\forall \mathcal{Q})$ is a λ-logical consequence of $\{\Pi, \mathcal{R}\}$ then $(\forall \mathcal{A}')$ is a λ-logical consequence of $\{\Pi, \mathcal{R}\}$.*
Proof. *If $\forall \mathcal{Q}$ is an λ-logical consequence of $\{\Pi, \mathcal{R}\}$, then for all, $\mathcal{I}$, λ-model for $\{\Pi, \mathcal{R}\}$, $\mathcal{I}$ is a λ-model for $\forall \mathcal{Q}$. That is, $\mathcal{I}(\forall \mathcal{Q}) \geq \lambda$ and hence $\mathcal{I}(\mathcal{Q}\vartheta) \geq \lambda$ for every assignment ϑ.*

On the other hand, if $(\mathcal{A} \leftarrow \mathcal{Q}) \in \Pi$ and $\mathcal{R}(\mathcal{A}, \mathcal{A}') = \alpha \geq \lambda$, by definition of annotated program, Π^λ, there exists an annotated clause $\langle \mathcal{A}' \leftarrow \mathcal{Q}, \alpha \rangle \in \Pi^\lambda$. Moreover, by definition of λ-model for $\{\Pi, \mathcal{R}\}$, $\mathcal{I}(\forall(\mathcal{A}' \leftarrow \mathcal{Q})) \geq \alpha \geq \lambda$ and hence $\mathcal{I}(\mathcal{A}'\vartheta \leftarrow \mathcal{Q}\vartheta) \geq \alpha \geq \lambda$ for every assignment ϑ. Now, there are two cases by definition of valuation:

1. $\mathcal{I}(\mathcal{A}'\vartheta \leftarrow \mathcal{Q}\vartheta) = 1$ because $\mathcal{I}(\mathcal{A}'\vartheta) \geq \mathcal{I}(\mathcal{Q}\vartheta) \geq \lambda$.
2. $\mathcal{I}(\mathcal{A}'\vartheta \leftarrow \mathcal{Q}\vartheta) = \mathcal{I}(\mathcal{A}'\vartheta)$ because $\mathcal{I}(\mathcal{Q}\vartheta) > \mathcal{I}(\mathcal{A}'\vartheta)$. But, as we just mentioned, $\mathcal{I}(\mathcal{A}'\vartheta \leftarrow \mathcal{Q}\vartheta) \geq \alpha \geq \lambda$ and therefore $\mathcal{I}(\mathcal{A}'\vartheta) \geq \lambda$.

So, in both cases $\mathcal{I}(\mathcal{A}'\vartheta) \geq \lambda$ for every assignment ϑ. That is, $\mathcal{I}(\forall \mathcal{A}') = inf\{\mathcal{I}(\mathcal{A}'\vartheta) | \vartheta$ assignment $\} \geq \lambda$. Therefore $(\forall \mathcal{A}')$ is a λ-logical consequence of $\{\Pi, \mathcal{R}\}$. □

Lemma 5.2 *Let $\mathcal{A}$ and $\mathcal{B}$ be two atoms such that $\mathcal{A} \leq \mathcal{B}$. Then, $\mathcal{I}(\forall \mathcal{A}) \leq \mathcal{I}(\forall \mathcal{B})$.*
Proof. *Immediate because as $\mathcal{A} \leq \mathcal{B}$ then it implies that there exists a substitution ξ, such that $\mathcal{B} = \mathcal{A}\xi$ and moreover: $\mathcal{I}(\forall \mathcal{A}) = \inf\{\mathcal{I}(\mathcal{A}\vartheta) \mid \vartheta$ assignment$\} \leq \inf\{\mathcal{I}(\mathcal{A}\xi\vartheta') \mid \vartheta'$ assignment$\} = \mathcal{I}((\forall \mathcal{A}\xi)) = \mathcal{I}(\forall \mathcal{B})$.* □

Corollary 5.3 *Let Π be a definite program, $\mathcal{R}$ be a similarity relation and λ be a cut value. Given $(\mathcal{A} \leftarrow \mathcal{Q}) \in \Pi$ and $\mathcal{A}'$ an atom such that $\mathcal{R}(\mathcal{A}, \mathcal{A}') = \alpha \geq \lambda$. If $(\forall \mathcal{Q}\theta)$ is a λ-logical consequence of $\{\Pi, \mathcal{R}\}$ then $(\forall \mathcal{A}'\theta)$ is a λ-logical consequence of $\{\Pi, \mathcal{R}\}$, whatever the substitution θ is.*
Proof. *Immediate by Lemma 5.1 and Lemma 5.2.* □

Corollary 5.4 *Let Π be a program, $\mathcal{R}$ be a similarity relation and λ be a level of cut. Given $\mathcal{A} \leftarrow\in \Pi$ and $\mathcal{A}'$ an atom such that $\mathcal{R}(\mathcal{A}, \mathcal{A}') = \alpha \geq \lambda$. Then $(\forall \mathcal{A}'\theta)$ is an λ-logical consequence of Π for every substitution θ.*
Proof. *Trivial, since this is a specific case of Corollary 5.3, when the clause is a fact. Then, if $\mathcal{A}$ is a program fact, the atoms $\mathcal{A}'$ close to $\mathcal{A}$ and their instances are λ-logical consequence of $\{\Pi, \mathcal{R}\}$.* □

In order to prove Lemma formally, we need to introduce the notion of position of a expression and some notations. *Positions* of a expression t (also called *occurrences*) are represented by sequences of natural numbers used to address subterms of t. The concatenation of the sequences p and w is denoted by $p.w$. $\mathcal{P}os(t)$ denotes the set of positions of the expression t. A expression t can be see as a mapping between the set $\mathcal{P}os(t)$ and the set of symbols compounding it. If $p \in \mathcal{P}os(t)$, $t[p]$ denotes the symbol of t at position p.

 ISBN: 978-84-614-6167-7

Lemma 5.5 *Let $\mathcal{A}$ and $\mathcal{B}$ be two atoms, $\mathcal{R}$ be a similarity relation, with cut level λ, and θ be a λ-unifier for $\mathcal{A}$ and $\mathcal{B}$ with degree α. Then, there exists an atom $\mathcal{A}'$ such that, $\mathcal{R}(\mathcal{A},\mathcal{A}') = \alpha$ and $\mathcal{A}'\theta = \mathcal{B}\theta$ (That is there exists $\mathcal{A}'$ which is close to $\mathcal{A}$, with degree α which unifies syntactically with $\mathcal{B}$, through the unifier θ)*

Proof. *[Sketch] If θ is a λ-unifier for $\mathcal{A}$ and $\mathcal{B}$ with degree α then $\mathcal{R}(\mathcal{A}\theta,\mathcal{B}\theta) = \alpha \geq \lambda$. Moreover, $\mathcal{A}\theta$ and $\mathcal{B}\theta$ share the same positions (i.e., $\mathcal{P}os(\mathcal{A}\theta) = \mathcal{P}os(\mathcal{B}\theta)$) and for all position $u_i \in \mathcal{P}os(\mathcal{A}\theta)$, $\mathcal{R}(\mathcal{A}\theta[u_i],\mathcal{B}\theta[u_i]) = \alpha_i$ being $\alpha = \bigwedge_{i=1}^{n}\alpha_i$. Note that, only positions $w_i \in \mathcal{P}os(\mathcal{A}) \cap \mathcal{P}os(\mathcal{B})$ contribute to the computation of the degree α (for the rest of positions w_i', $\mathcal{R}(\mathcal{A}\theta[w_i'],\mathcal{B}\theta[w_i']) = 1$).*

Now, we take $\mathcal{A}$ and build an atom $\mathcal{A}'$ of the following form: for each $u_i \in \mathcal{P}os(\mathcal{A}) \cap \mathcal{P}os(\mathcal{B})$, we replace the symbol of $\mathcal{A}$ at position w_i (i.e., $\mathcal{A}[w_i]$) by the corresponding symbol of $\mathcal{B}$ (i.e., $\mathcal{B}[w_i]$) in $\mathcal{A}$. That is, we build an atom $\mathcal{A}'$ which is exactly equal than $\mathcal{A}$ except that the symbols at non-variable positions (shared with $\mathcal{B}$) has been replaced with symbols of $\mathcal{B}$. Therefore, $\mathcal{R}(\mathcal{A},\mathcal{A}') = \alpha$ and $\mathcal{A}'\theta = \mathcal{B}\theta$ by construction of $\mathcal{A}'$. □

Theorem 5.6 (Soundness of the WSLD–Resolution) *Let Π be a definite program, $\mathcal{R}$ a similarity relation, λ a cut value and $\mathcal{G}$ a definite goal. Then every computed answer $\langle\theta,\beta\rangle$ of level λ for $\{\Pi,\mathcal{R}\}$ and $\mathcal{G}$ is a correct answer of level λ for $\{\Pi,\mathcal{R}\}$ and $\mathcal{G}$.*

Proof. *Assume $\mathcal{G} \equiv \leftarrow \mathcal{A}_1,\ldots,\mathcal{A}_k$ and $\theta_1,\ldots,\theta_n$ be the sequence of w.m.g.u.'s in a WSLD-refutation of level λ for $\{\Pi \cup \{\mathcal{G}\},\mathcal{R}\}$: $\mathcal{G} \overset{[\theta_1,\alpha_1]}{\Longrightarrow}_{WSLD} \overset{[\theta_2,\alpha_2]}{\Longrightarrow}_{WSLD} \ldots \overset{[\theta_n,\alpha_n]}{\Longrightarrow}_{WSLD} \square$, leading to the computed answer $\langle\theta,\alpha\rangle$ where $\alpha = \bigwedge_{i=1}^{n}\alpha_i$ is the approximation degree computed in the refutation. We have to prove that: (1) $\mathcal{I}(\forall(A_1,\ldots,A_k)\theta)$ is λ-logical consequence of $\{\Pi,\mathcal{R}\}$; and (2) $\mathcal{M}_\Pi^\lambda(\forall(A_1,\ldots,A_k)\theta) \leq \alpha$. The result is proven by induction on the number of steps of the WSLD-refutation.*

Base case ($n = 1$)*: First, we consider refutations of length one. This means that $\mathcal{G}$ must be a goal of the form $\mathcal{G} \equiv \leftarrow \mathcal{A}_1$ and the program Π contains a unit clause (a fact) $\mathcal{C}_1 \equiv \mathcal{A} \leftarrow$ which weakly unifies with $\mathcal{A}_1$. That is, there exists a w.m.g.u. θ_1 of $\mathcal{A}_1$ and $\mathcal{A}$ such that $\mathcal{R}(\mathcal{A}\theta_1,\mathcal{A}_1\theta_1) = \alpha \geq \lambda$ (i.e., its approximation degree is $\alpha_1 = \alpha \geq \lambda$). On the other hand, it is easy to prove that $\mathcal{A}_1\theta_1$ is an instance of a clause of the annotated program Π^λ. More precisely, there exist an annotated clause $\langle\mathcal{A}',\alpha_1\rangle \in \Pi^\lambda$, with $\mathcal{R}(\mathcal{A},\mathcal{A}_1) = \alpha_1 = \alpha$, such that $\mathcal{A}' \leq \mathcal{A}_1\theta_1$. Therefore, by Corollary 5.4, $\forall(A_1\theta_1)$ is λ-logical consequence for $\{\Pi,\mathcal{R}\}$ and item 1 is proved.*

For proving item 2, remember that if $\mathcal{R}(A,A) = \alpha$, by definition of the immediate consequences operator of level λ (Deninition 3.1), $\alpha \in \mathcal{PT}_\Pi^\lambda(\mathcal{M}_\Pi^\lambda)(\mathcal{A}'\vartheta')$ for every assignment ϑ'. So, for the least Herbrand model $\mathcal{M}_\Pi^\lambda(\mathcal{A}'\vartheta') \leq \alpha$ whatever be the assignment ϑ'. On the other hand, if $\mathcal{A}' \leq \mathcal{A}_1\theta_1$, then there exists a γ such that $\mathcal{A}_1\theta_1 = \mathcal{A}'\gamma$. Hence, for every ϑ_1, $\mathcal{M}_\Pi^\lambda(\mathcal{A}_1\theta_1\vartheta_1) \leq \mathcal{M}_\Pi^\lambda(\mathcal{A}'\gamma\vartheta_1) \leq \alpha$. Therefore, $\mathcal{M}_\Pi^\lambda(\forall(\mathcal{A}_1\theta_1)) = \inf\{\mathcal{M}_\Pi^\lambda(\mathcal{A}_1\theta_1\vartheta_1) \mid \vartheta_1 \ \ assignment\} \leq \alpha$.

Inductive case ($n > 1$)*: Next, assume that the theorem holds for WSLD-refutations with $n > 1$ steps. Also suppose that $\mathcal{C} \equiv \mathcal{A} \leftarrow \mathcal{Q} \in \Pi$ is the input clause and A_l is the selected atom of $\mathcal{G}$ in the first step of the WSLD-refutation:*

 ISBN: 978-84-614-6167-7

$$\leftarrow \mathcal{A}_1,\ldots,\mathcal{A}_l,\ldots,\mathcal{A}_k \overset{\theta_1,\alpha_1}{\Longrightarrow}_{WSLD} \leftarrow (\mathcal{A}_1,\ldots,\mathcal{Q},\ldots,\mathcal{A}_k)\theta_1 \Longrightarrow_{WSLD}{}^{*} \square$$

Now, by the induction hypothesis, we suppose that the result fulfills for computed answers of refutations with length less than n. Then: (a) $\forall(\mathcal{A}_1,\ldots,\mathcal{Q},\ldots,\mathcal{A}_k)\theta_1,\theta_2,\ldots,\theta_n$ *is* λ*-logical consequence of* $\{\Pi,\mathcal{R}\}$*; (b)* $\mathcal{M}^{\lambda}_{\Pi}(\forall(\mathcal{A}_1,\ldots,\mathcal{Q}_1,\ldots,\mathcal{A}_k)\theta_1\theta_2\ldots\theta_n) \leq \beta = \bigwedge_{i=2}^{n}\alpha_i$*. For proving item 1, first note that, from (a), we can immediately infer that* $\forall(\mathcal{Q}\theta_1\theta_2,\ldots,\theta_n)$ *is* λ*-logical consequence of* $\{\Pi,\mathcal{R}\}$*. On the other hand, if the first resolution step, with the clause* $\mathcal{C}$ *and the selected atom* $\mathcal{A}_l$*, is possible, there must exist a wmg* θ_1 *with* $\mathcal{R}(\mathcal{A}_l\theta_1,\mathcal{A}\theta_1) = \alpha_1 \geq \lambda$*. In a similar way to the base case, we can claim that there exists an annotated clause* $\langle \mathcal{A}' \leftarrow \mathcal{Q},\alpha_1\rangle \in \Pi^{\lambda}$ *with* $\mathcal{R}(\mathcal{A},\mathcal{A}') = \alpha_1$ *and such that* $\mathcal{A}' \leq \mathcal{A}_1\theta_1$*. Therefore, by the Corollary 5.3,* $\forall(\mathcal{A}_l\theta_1)$ *is a* λ*-logical consequence of* $\{\Pi,\mathcal{R}\}$*. Moreover, by Lemma 5.2, all of the instances of* $\mathcal{A}_l\theta$ *are* λ*-logical consequences of* $\{\Pi,\mathcal{R}\}$*. Notably,* $\forall(\mathcal{A}_l\theta_1\theta_2,\ldots,\theta_n$ *is a* λ*-logical consequence of* $\{\Pi,\mathcal{R}\}$*. Finally, it is immediate to prove that* $\forall((\mathcal{A}_1,\ldots,\mathcal{A}_k)\theta_1\theta_2,\ldots,\theta_n$ *is an* λ*-logical consequence of* $\{\Pi,\mathcal{R}\}$*.*

For proving the item 2, note that, since the selected atom $\mathcal{A}_l$ *an the head* $\mathcal{A}$ *of the input clause* $\mathcal{C}$ *weakly unify with approximation degree* α_1*, by Lemma 5.5, there exists an atom* $\mathcal{A}'$ *such that* $\mathcal{R}(A,A') = \alpha_1$ *and* $\mathcal{A}'\theta_1 \leq \mathcal{A}_l\theta_1$*. Now, we can build the ground instance:* $\mathcal{A}_l\theta_1\theta_2\ldots\theta_n\vartheta_1 = \mathcal{A}'\theta_1\theta_2\ldots\theta_n\vartheta_1$ *being the head of an instance of an annotated clause* $\langle \mathcal{C}' \equiv \mathcal{A}' \leftarrow \mathcal{Q},\alpha_1\rangle$ *belonging to the proximity class of* $\mathcal{C}$*,* $\mathcal{K}_{\lambda}(\mathcal{C})$ *(see Definition 3.1). Now by definition of the immediate consequences operator of level* λ*:* $\mathcal{PT}^{\lambda}_{\Pi}(\mathcal{M}^{\lambda}_{\Pi}) \ni \alpha_1 \wedge \mathcal{M}^{\lambda}_{\Pi}((\mathcal{Q}\theta_1\ldots\theta_n)\vartheta_1)$ *Consequently, for every assignment* ϑ_1*: So,*

$$\mathcal{M}^{\lambda}_{\Pi}(\forall(\mathcal{A}_l\theta_1\ldots\theta_n)) = \inf\{\mathcal{M}^{\lambda}_{\Pi}(\mathcal{A}_l\theta_1\ldots\theta_n\vartheta_1) \mid \vartheta_1 \quad \textit{assignment}\} \leq$$
$$\alpha_1 \wedge \inf\{\mathcal{M}^{\lambda}_{\Pi}(\mathcal{Q}\theta_1\ldots\theta_n\vartheta_1) \mid \vartheta_1 \quad \textit{assignment}\} = \alpha_1 \wedge \mathcal{M}^{\lambda}_{\Pi}(\forall(\mathcal{Q}\theta_1\ldots\theta_n)).$$

Now, by the distributivity of the universal quantifier with respect to the conjunction:

$$\mathcal{M}^{\lambda}_{\Pi}(\forall(\mathcal{A}_1,\ldots,\mathcal{A}_l,\ldots,\mathcal{A}_k)\theta_1\ldots\theta_n) =$$
$$\textstyle\bigwedge_{i=1}^{l-1}\mathcal{M}^{\lambda}_{\Pi}(\forall(\mathcal{A}_i\theta_1\ldots\theta_n) \wedge \mathcal{M}^{\lambda}_{\Pi}(\forall(\mathcal{A}_l\theta_1\ldots\theta_n) \wedge \bigwedge_{j=l+1}^{k}\mathcal{M}^{\lambda}_{\Pi}(\forall(\mathcal{A}_j\theta_1\ldots\theta_n) \leq$$
$$\textstyle\bigwedge_{i=1}^{l-1}\mathcal{M}^{\lambda}_{\Pi}(\forall(\mathcal{A}_i)\theta_1\ldots\theta_n) \wedge \alpha_1 \wedge \mathcal{M}^{\lambda}_{\Pi}(\forall(\mathcal{Q})\theta_1\ldots\theta_n) \wedge \bigwedge_{j=l+1}^{k}\mathcal{M}^{\lambda}_{\Pi}(\forall(\mathcal{A}_j\theta_1\ldots\theta_n) =$$
$$\alpha_1 \wedge \mathcal{M}^{\lambda}_{\Pi}(\forall(\mathcal{A}_1,\ldots,\mathcal{Q},\ldots,\mathcal{A}_k)\theta_1\ldots\theta_n) \leq \alpha_1 \wedge (\textstyle\bigwedge_{n}^{i=1}\alpha_i) = \alpha$$

□

6 Conclusions and Future Work

In this paper we revisited the declarative semantics of Bousi∼Prolog which were defined for a pure subset and presented in [4]. We have given more accurate definitions for the semantic concepts and thereby solved some problems that may arise when we work with non-linear programs. Moreover, we introduce for the first time a notion of correct answer inside our framework. Then, after recalling both the WSLD-resolution principle and a similarity-based unification algorithm, which is the basis of the Bousi∼Prolog operational mechanism for definite programs, we prove the soundness of WSLD-resolution as well as other auxiliary results. Finally, it is worthy to say that, along this paper we

 ISBN: 978-84-614-6167-7

have clarified some of the existing differences between our framework and the related proposal introduced by [9].

As a matter of future work we want to go ahead proving the completeness theorem for this restricted subset of Bousi∼Prolog. On the other hand, at the present time we know that a naive extension of Sessa's unification algorith to proximity relations does not work, because correctness problems may arise. Therefore, it is necessary to define a complete new algorithm able to deal with proximity relations properly and to lift some of the current results to the new framework.

Acknowledgements This work has been partially supported by FEDER and the Spanish Science and Innovation Ministry under grants TIN2007-65749 and TIN2011-25846 and by the Castilla-La Mancha Regional Administration under grant PII1I09-0117-4481.

References

[1] R. Caballero, M. Rodríguez, and C.A. Romero. Similarity-based reasoning in qualified logic programming. In *Proc. PPDP '08*, pages 185–194, ACM. 2008.

[2] F. Arcelli and F. Formato. A similarity-based resolution rule. *Int. J. Intell. Syst.*, 17(9):853–872, 2002.

[3] S. Guadarrama, S. Muñoz, and C. Vaucheret. Fuzzy Prolog: A new approach using soft constraints propagation. *Fuzzy Sets and Systems, Elsevier*, 144(1):127–150, 2004.

[4] P. Julián and C. Rubio. A declarative semantics for Bousi∼Prolog. In *PPDP*, pages 149–160, ACM. 2009.

[5] J.W. Lloyd. *Foundations of Logic Programming.* Springer-Verlag, Berlin, 1987.

[6] A. Martelli and U. Montanari. An Efficient Unification Algorithm. *ACM Transactions on Programming Languages and Systems*, 4:258–282, 1982.

[7] J. Medina, M. Ojeda and P. Vojtáš. Similarity-based unification: a multi-adjoint approach. *Fuzzy Sets and Systems*, 146(1):43–62, 2004.

[8] M. Rodríguez and C.A. Romero. Quantitative logic programming revisited. In *Proc. FLOPS'08*, pages 272–288, Springer. 2008.

[9] M.I. Sessa. Approximate reasoning by similarity-based sld resolution. *Theoretical Computer Science*, 275(1-2):389–426, 2002.

[10] M.H. van Emden. Quantitative deduction and its fixpoint theory. *Journal of Logic Programming*, 3(1):37–53, 1986.

[11] M.H. van Emden and R.A. Kowalski. The semantics of predicate logic as a programming language. *Journal of the ACM*, 23(4):733–742, 1976.

[12] P. Vojtáš. Fuzzy Logic Programming. *Fuzzy Sets and Systems*, 124(1):361–370, 2001.

[13] L. Zadeh. Fuzzy sets. *Information and Control*, 8:338–353, 1965.

Proceedings of the 11th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2011
26–30 June 2011.

A Stochastic Game Analysis of a Multi-Power Diversity Binary Exponential Backoff Algorithm

Abdelilah Karouit[1], Essaid Sabir[2], Fernando Ramirez-Mireles[3], Luis Orozco Barbosa[4] and Abdelkrim Haqiq[1]

[1] *L-IR2M, Faculty of Sciences and Techniques, Hassan 1st University, Settat, Morocco*

[2] *LIA/CERI, University of Avignon, Agroparc, BP 1228, 84911, Avignon, France*

[3] *Engineering Division, ITAM, Mexico City*

[4] *I3A, University of Castilla La Mancha (UCLM), 02071 Albacete, Spain*

emails: akarouit@gmail.com, essaid.sabir@univ-avignon.fr, ramirezm@ieee.org, luis.orozco@uclm.es, ahaqiq@gmail.com

Abstract

In the past few years, there has been an increasing interest on the benefits of applying the principles of game theory to better understand and plan the operation of telecommunications systems. For example, game theory has been used to analyze network congestion control, network routing, transmission power control, topology control, etc. In particular, in this work we deal with access to channel resources. The scenario is a distributed network with mobile stations (MS) competing for the channel bandwidth to communicate with a centralized entity. To solve the conflicts arising from MS trying to access simultaneously, the network relies on a protocol named the Binary Exponential Backoff (BEB), a popular bandwidth allocation mechanism used by a large number of wireless technologies. Our game scenario is made up of three components: 1) The decision makers are the MSs, 2) The individual action is either to "transmit" or to "wait", the retransmission probability is then the strategy. The resulting strategy profile determines the amount of congestion in the network (outcome of the game), and 3) The utility function of each Ms is its own throughput (or alternatively minus its delay). Clearly, the payoff of each station depends not only on its own decision, but also depends on the actions of other adversarial stations. Although our scenario does have a centralized entity, this entity does not have the full picture of the network conditions created by the individual decisions of the MS. Our intention is to predict what might or should happen when aggressive MSs competing with other MS act selfishly using a

 ISBN: 978-84-614-6167-7

high retransmission probability to increase its own throughput, and what would be the effect of using multiple power levels (MPL) during the transmissions. Our results show that the capture effect introduced by the MPL-BEB mechanism proves effective on the individual/global performance. Via price of anarchy, our results identify a behavior similar to the well-know prisoners dilemma. A non-efficiency of Nash equilibrium is observed for both schemes (BEB and MPL-BEB) under heavy traffic with a notable outperformance of MPL-BEB.

Key words: Random access, Capture effect, Nash equilibrium, Price of anarchy.

1 Introduction

With its potential synergy in analyzing actors/players behaviors and predicting the outcome of a conflict situation (called "game"), game theory is the ultimate tool to adopt when studying decentralized systems. Under rationality of actors/players, the most common solution concept is called the "Nash Equilibrium". A Nash equilibrium point is a strategy profile where no player has any incentive to deviate unilaterally. Recently, the selfish behavior of mobile stations in NET, MAC or PHY layers has been widely analyzed using game theory. For instance many works on power control, random access, routing games have been driven, see [1, 2, 4, 7, 11, 12]. A performance collapse is then predicted for collision-channel systems using aloha or binary exponential backoff algorithm (BEB), see [5, 9]. In order to overcome the limitations of the standard BEB, we develop a stochastic game-based framework to model and evaluate BEB while exploiting power diversity. Indeed, it has been verified that the system's performance can be improved when: 1) The Mobile Station (MS) use power diversity, i.e., before starting transmission each MS picks randomly a power level among the N available levels. This power diversity produces a "capture effect" in which even when two or more packets collide, one of them can be decoded successfully with a certain probability [4, 5], and 2) The value of the initial window W_0 (directly related with the retransmission probability) is optimized using team theory (in which the MSs cooperate together) [6], or using game theory (in which the MS act selfishly) [7]. Previous works combine team theory and power diversity for Aloha [5, 9] and for BEB [8], and use game theory with power diversity for Aloha [9]. To the best of our knowledge, this work proposes for the first time in the literature the implementation of power diversity in a system operating using the BEB to increase throughput and reduce delay in a decentralized context (selfish MSs).

In the traditional BEB, although the terminals compete, they do not do so on base of a decision-making process. The MSs do not have an individual choice of retransmission probability, they use a predefined parameters or a parameter given by a central entity. In this case finding the retransmission probability that maximizes throughput is an optimization problem carried out by a single decision maker: the central entity.

The rest of the paper is organized as follows. We describe the problem in Section 2. In Section 3, we build the induced stochastic game for both standard BEB and MPL-BEB. Later, we derive performance measures of interest in Section 5. We perform a detailed nu-

 ISBN: 978-84-614-6167-7

merical investigation and present many related results in Section 6. Finally, our concluding remarks are drawn in Section 7.

2 A MAC/PHY cross-layer design

We consider a wireless multiple access system composed of one central receiver (base station BS) and m geographically dispersed mobile stations communicating with the BS. There is no central control and consider that MSs communicate using BEB as a random access method. Time is divided into multiple equal and synchronized slots. Transmission feedback (success or collision) is received at the end of the current slot.

As mentioned before, power diversity produces a capture effect. Due to this capture effect, even if a collision occurs the receiver is able to decode the message transmitted with the highest power among concurrent transmissions. In fact, the unsuccessful concurrent messages are lost and are treated as interference. In this MAC/PHY cross-layer design an MS i contending for a message transmission, randomly chooses a power level T_i among N available levels $T = \{T_1, T_2, ..., T_N\}$. The power levels random selection follows the probability vector $X = [x_1, x_2, ..., x_N]$, where the j-th entry x_j is the probability to select the power level T_j. We consider a general capture model where a message transmitted by an MS i is received successfully when and only when its Signal to Interference plus Noise ratio (SINR) is higher than some given threshold Θ_{th}. The received power on the BS can be related to the transmitted power by the propagation relation $h_i \cdot T_i$, where h_i is the channel gain experienced by the base station on that link. Denoting by σ^2 the power of the thermal noise, the instantaneous SINR of MS i is then given by:

$$\Theta_i = \frac{h_i \cdot T_i}{\sum\limits_{k=1,k\neq i}^{m} h_k \cdot T_k \cdot \mathbf{1}_k + \sigma^2}, \tag{1}$$

where $\mathbf{1}_k$ is an indicator function of the event that at the current slot, MS k transmits its message.
We denote by A_s , $s \geq 2$, the probability of a successful transmission among s simultaneous attempts, i.e., transmitted during the same slot. Let us denote by a_i^s the probability that transmission of some tagged MS i is successful while $s-1$ other MSs attempt simultaneous transmission. It can be derived using the following events decomposition $a_i^s = \sum\limits_{t=2}^{N} P(\mathrm{A}_{i,s}^t \bigcap \mathrm{B}_{i,s}^t \bigcap \mathrm{C}_{i,s}^t)$ where

- $\mathrm{A}_{i,s}^t$ is the event "*Mobile station i attempts transmission with power level* T_t".

- $\mathrm{B}_{i,s}^t$ is the event "*Other* $s-1$ *mobile stations transmit with powers less than* T_t".

- $\mathrm{C}_{i,s}^t$ is the event "*Instantaneous SINR of MS i is higher than the target SINR* Θ_{th}".

 ISBN: 978-84-614-6167-7

Since all MSs are symmetric and are assumed to experience the same channel gain, i.e., $h_i = h$, $i = 1, \ldots, m$. Then $A_s = s.a_i^s$, it follows that

$$A_s = s \sum_{l=0}^{N-2} \sum_{s_1=0}^{s-1} \cdots \sum_{s_{N-l-1}=0}^{s-1} x_1^{s_1} \cdot x_2^{s_2} \cdots x_{N-l}^{s_{N-l}} \cdot u \left(\frac{T_{N-l}}{\sum_{r=1}^{N-l-1} T_r s_r + \frac{\sigma^2}{h}} - \Theta_{th} \right) \cdot \delta \left(s - 1 - \sum_{r=1}^{N-l-1} s_r \right), \tag{2}$$

with, $A_0 = 0$ and $A_1 = 1$. In order to get a successful transmission, we need to take $s_{N-l} = 1$. Moreover, T_{N-l} is the power level chosen by the MS whose transmission maybe potentially succeed, i.e., the one corresponding to the highest power selected in the current slot. Whereas s_r denotes the number of MSs that have chosen the power level T_r in the current slot. $\delta(t)$ is the Dirac distribution and $u(t)$ is the Heaviside function (unit step function) are given by the following expressions

$$\delta(t) = \begin{cases} 1, & t = 1, \\ 0, & else. \end{cases} \qquad \text{and} \qquad u(t) = \begin{cases} 1, & t \geq 0, \\ 0, & else. \end{cases} \tag{3}$$

Computing the success probability is a challenging task. The difficulty of formula (2) is to consider one single transmitting MS at the highest power level and list all the cases where the $s-1$ remaining MSs transmit at lower power levels. This corresponds exactly to the set of partitions [1] of the positive integer $k-1$ considering all possible permutations. Generating all the partitions of an integer has been widely studied in the literature and several algorithms have been proposed, e.g., see [10]. The computational complexity of such algorithms is very expensive and may take long time to compute the set of all partitions and their permutations. Fortunately, in our model the success probability depends on none of the following: the instantaneous state of the system n; the arrival probability λ; and, the retransmission probability q. Henceforth, the success probability vector $\mathbf{A} = (A_s)$, $s = 0 \cdots m$ can be computed once and reused to derive the transition matrix.

3 A stochastic game formulation

In this section we introduce the basic definitions and notation needed to build our model. All mobile users share the same channel to transmit their respective messages. Clearly, the payoff of MS i depends not only on its own decision of i but also depends on the actions of other adversarial MSs. A typical conflict situation is then induced, and game theory tools seems to be suitable to analyze these kind of situations. The primary focus of this work is to build a non-cooperative framework for BEB, in which each MS optimizes her or his payoff. We start by presenting the markovian model for both BEB and MPL-BEB and then analyze the non-cooperative game.

We build a markovian model for both the standard BEB as well as for the MPL-BEB scheme making use of power diversity. We take the case of a system consisting of a finite

[1] A partition of a positive integer η is a way of writing η as a sum of positive integers.

 ISBN: 978-84-614-6167-7

number of m mobile stations that intend to transmit a message. Each MS i handles a buffer sufficient to store exactly one message, thus no new message is generated by MS i till success/drop of current message. An MS i can be in one of two distinguished states: 'I' (idle) or 'T' (transmitting). At the beginning of each slot and being in state I', an MS i has no message to transmit and does generate a new message with some probability λ (for lack of simplicity we restrict to the case where λ is the same for all MSs). MSs at state 'I' that generate a new message switch to state 'T' at the beginning of the next slot. Being at state 'T', the tagged MS i attempts to transmit with probability q_i, until its message is successfully transmitted. If two or more MSs at state 'T' attempt the channel simultaneously, then messages collide. In the case that the messages could not be properly decoded, then the corresponding MSs immediately return to state 'T'. All corrupted messages get backlogged and are retransmitted after some random time. Whereas, if exactly one mobile station attempts a transmission from state 'T' (BEB) or if the SINR of the received signal is higher than the target (MPL-BEB), then the transmission is successful, and the corresponding mobile station jumps to state 'I'. The determination of the above random time can be considered as a stochastic control problem. The information structure, however, is not a classical one: the MSs do not have knowledge of the full state information. They do not know the number of backlogged MSs, nor do they know the number of messages involved in a collision. Moreover, implementation of random access-based procedure such as aloha or BEB as a centralized system is a real issue since it needs a high signaling protocol and full information on the activity of the mobile stations and their instantaneous QoS (Quality of Service) requirements. Then, a high amount of bandwidth should be reserved for signaling. Furthermore, mobile users are not forced to cooperate and may present a selfish behavior. This is why we convert this problem and study it from game theory perspective as a non cooperative game. Now, each mobile station wishes to optimize its own objective function. We shall then formulate a distributed model using game theory tools with all its powerful and robust equilibrium concepts. For instance, we will be interested in studying Nash equilibrium of the power diversity-enabled BEB. Now, at equilibrium and assuming rationality of mobile users, no MS has incentive to deviate unilaterally. The elements of our contention game are as follows

- We consider a set $\mathcal{N}$ of bufferless mobile users with cardinality $|\mathcal{N}| = m$. Each MS is labeled by an integer from 1 to m.
- Packets arrive from higher layers of each mobile station i following a Bernoulli process with parameter λ (non-saturated regime).
- The state of the system is given by the number of MSs in state 'T', i.e., backlogged messages and messages that will be transmitted for the first time.
- Mobile station i transmits its messages/packets with probability q_i in every slot.
- Each mobile station has two actions (transmit or not) and its retransmission probability q_i is considered to be its strategy.
- For simplicity we assume that transmissions are cost free.

 ISBN: 978-84-614-6167-7

- The objective function of each MS i is denoted by U_i. It can be either individual throughput or alternatively minus expected delay at MS i.

Let $\mathbf{q} = (q_1, q_2, \cdots, q_m)$ be the policy vector of retransmission probabilities for all MSs. Under rationality, each MS i seeks to maximize its own function utility. We shall use as

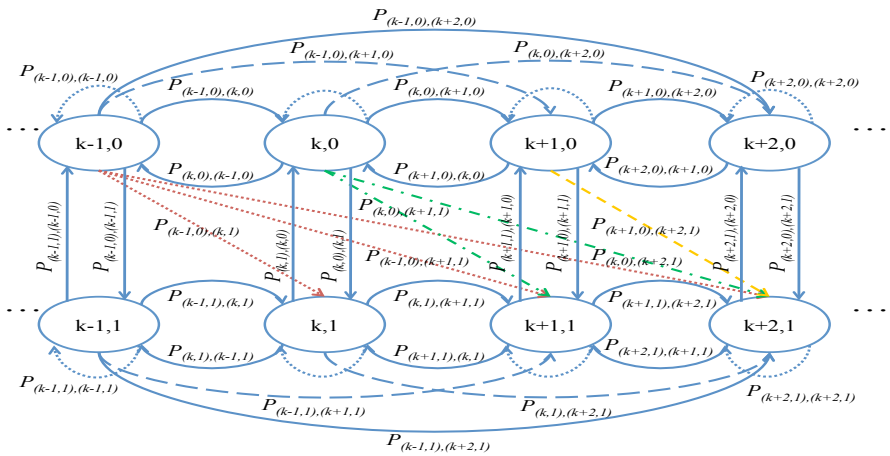


Figure 1: The state of the system is the backlog vector; the first component corresponds to the number of message in transmission (backlogged plus first time transmissions) among the mobile users $\mathcal{N} \setminus \{i\}$, whereas the second component indicates the number of messages (either backlogged or first time transmissions) of the tagged mobile i (either 0 or 1). In order to stay inside the state space (for this illustrative example) we need to have $1 \leq k \leq m-2$ and $m \geq 2$.

the state of the system (n, a) the number n of MSs in state 'T' among the set $\mathcal{N} \setminus \{i\}$ (denoted by n and takes value in $0, 1, \cdots, m-1$) and the number a of messages (either backlogged or first time transmission) of tagged MS i (denoted by a and takes value in $0, 1$) at the beginning of a slot. For any choice of values $q_i \in (0, 1]$, the state process is a Markov chain that contains a single ergodic sub-chain (and possibly transient states as well). Indeed, it is easy to check that, conditioning on the actual state of the system, the future and the past are mutually independent (Markov property), see [8] for a proof of other properties. We denote by $P_{(n,a)(n',b)}(\mathbf{q})$ the probability that the system jumps from state (n, a) to state (n', b). The transition probability diagram is depicted in Figure 1. Let $\bar{\pi}(\mathbf{q})$ be the corresponding vector of the steady-state probabilities where its n-th entry $\pi_{n,a}(\mathbf{q})$ denotes the probability that the state of the system is (n, a). The only point where the Markov chain P does not have a single stationary distribution is at $\mathbf{q} = (0, 0, \cdots, 0)$, where it has the absorbing states: $(n = m-1, a = 1)$. All remaining states are transient (for any $\lambda > 0$), and the probability to end in one of the absorbing states, depends on the initial distribution of the Markov chain. When this state is reached, then the throughput equals 0, which means that it is a deadlock state. For $\lambda > 0$ and $q = 0$, the deadlock state is reached with positive probability from any initial state other than the absorbing state. We shall

therefore exclude the case of $q = 0$ and optimize only on the range $\epsilon < q_i \leq 1 \; \forall i = 1, \cdots, m$. Naturally, the outcome of any instance of the game is a Nash equilibrium (if it exists).

Definition 1 *The strategy profile* $\boldsymbol{q}^* = (q_1^*, q_2^*, \cdots, q_i^*, \cdots, q_m^*)$ *is a Nash equilibrium if no mobile station can improve its utility by unilateral deviation, namely*

$$U_i\left(\boldsymbol{q}^*\right) \geq U_i\left(q_1^*, q_2^*, \cdots, q_i, \cdots, q_m^*\right), \quad \forall i = 1, \cdots, m \quad \textit{and} \quad \forall q_i \in [\epsilon, 1], \; \epsilon > 0. \tag{4}$$

For computation tractability and without any loss of generality, we restrict throughout this paper to find a symmetric policy where all MSs are payoff-balanced. Assume that there are n actual backlogged MSs/messages among the set $\mathcal{N} \setminus \{i\}$, and all use the same value q as retransmission probability. Let $Q_r(j, n)$ be the probability that j out of the n $(n = 0, 1, ..., m-1)$ backlogged messages are retransmitted in the current slot. Then

$$Q_r(j, n) = \binom{n}{j} (1-q)^{n-j} (q)^j . \tag{5}$$

Similarly, let $Q_a(j, n)$ denote the probability that j unbacklogged MSs among the set $\mathcal{N} \setminus \{i\}$ generate new messages in the current slot. Thus

$$Q_a(j, n) = \binom{m-n-1}{j} (1-\lambda)^{m-n-j-1} (\lambda)^j . \tag{6}$$

4 Equilibrium analysis

Define $(\mathbf{q}_{-i}, q_i)$ to be a retransmission policy where each MS $j \in \mathcal{N} \setminus \{i\}$ retransmits at any slot with probability q_j for all $j \neq i$ and where tagged MS i retransmits with probability q_i. Since we restrict to symmetric policies $\mathbf{q}_{-i}$ where all mobiles are balanced-payoff, then we shall also identify it (with some abuse of notation) with the actual transmission probability which is the same for all users in $\mathcal{N} \setminus \{i\}$. Now, we shall assume that all MSs in $\mathcal{N} \setminus \{i\}$ retransmit with a given probability $\mathbf{q}_{-i} = (q, q, \cdots, q)$ and the tagged MS i retransmits with probability q_i.

Define the set $\mathcal{BR}_i(\mathbf{q})$ as the set of best response strategies of tagged mobile station i, it can be written as

$$\mathcal{BR}_i(\mathbf{q}) := \underset{q_i}{\operatorname{argmax}} \, U_i\left(\mathbf{q}_{-i}, q_i\right), \tag{7}$$

where $\mathbf{q}$ denotes the policy where all MSs in $\mathcal{N} \setminus \{i\}$ retransmit with probability q, and the maximization is taken with respect to q_i.
Therefore, the strategy profile $\mathbf{q}^*$ is a symmetric Nash equilibrium if and only if

$$\mathbf{q}^* \in \mathcal{BR}_i(\mathbf{q}^*). \tag{8}$$

As mentioned in [5] and [9], the Nash equilibrium in such a game may be inefficient and may provide bad performance. Indeed, one note that mobile stations become more and more aggressive as the arrival probability increases which results in a dramatic decrease in the

 ISBN: 978-84-614-6167-7

system's aggregate average throughput. Moreover, the equilibrium retransmission quickly goes to 1 when the number of mobile users increases. We note that a similar aggressive behavior at equilibrium has been observed in [11] in the context of flow control by several competing mobile users that share a common drop tail buffer. However in that context, the most aggressive behavior (of transmission at maximum rate) is the "equilibrium" solution for *any arrival rate*, and not just at high rates as in our case. We may thus wonder why retransmission probabilities of 1 are not an equilibrium in our BEB game (in the case of light traffic), see Section 6. An intuitive reason could be that if a mobile station deviates and retransmits with probability 1 (while other continue to retransmit with the equilibrium probability $q^* < 1$), the congestion level in the system (i.e., the number of backlogged messages) increases; This induces more retransmissions from other MSs which then results on more collisions of messages from the deviating mobile and a degradation on its own payoff.

5 Steady state and performance evaluation

In order to evaluate and quantify the performance of MPL-BEB taking as benchmark the standard BEB, we introduce a Markov chain with a two-dimensional state, see Figure 1. Transition probabilities of both schemes BEB and MPL-BEB are given in Appendix A and Appendix B, respectively. Based on the steady state of the system, one can estimate several performance measures. For instance, we are particularly interested to derive the average throughput, the expected delay and the failure probability of transmitted messages. We first discuss the procedure to obtain the steady state probabilities. Then we derive the expressions of the performance metrics of interest functions of the steady state equations. Denote by $\pi_{n,a}\left(\mathbf{q}_{-i}, q_i\right)$ the steady state of the Markov chain where n is the number of backlogged messages among the $\mathcal{N} \setminus \{i\}$ MSs, and a is the binary number of the backlogged messages of tagged MS i.

The steady state of the Markovian process is given by the following system

$$\begin{cases} \bar{\pi}\left(\mathbf{q}_{-i}, q_i\right) = \bar{\pi}\left(\mathbf{q}_{-i}, q_i\right) P\left(\mathbf{q}_{-i}, q_i\right), \\ \sum\limits_{n=0}^{m-1} \sum\limits_{a=0}^{1} \pi_{n,a}\left(\mathbf{q}_{-i}, q_i\right) = 1, \\ \pi_{n,a}\left(\mathbf{q}_{-i}, q_i\right) \geq 0, \quad n = 0, ..., m-1 \text{ and } a = 0, 1. \end{cases} \tag{9}$$

Using a simple iterative method, one may compute the stationary distribution from system (9). Hence, the mean number of the bandwidth requests in the system is

$$B\left(\mathbf{q}_{-i}, q_i\right) = \sum_{n=0}^{m-1} \sum_{n'=0}^{m-1} P_{(n,0),(n',1)} \pi_{n,0}(\mathbf{q}_{-i}, q_i). \tag{10}$$

Average throughput: We define the average throughput of the tagged MS i as the sample average of the number of messages that are successfully transmitted by this MS. Using the

 ISBN: 978-84-614-6167-7

rate balance equation at steady state, We can easily derive the expression of the throughput of tagged MS i as follows :

$$\Gamma_i\left(\mathbf{q}_{-i}, q_i\right) = \lambda \sum_{n=0}^{m-1} \pi_{n,0}(\mathbf{q}_{-i}, q_i). \tag{11}$$

Mean delay: The expected delay at tagged MS i can be easily obtained using Little's result [13], namely

$$D(\mathbf{q}_{-i}, q_i) = 1 + \frac{\Gamma_i\left(\mathbf{q}_{-i}, q_i\right)}{B\left(\mathbf{q}_{-i}, q_i\right)}. \tag{12}$$

Failure probability of transmitted messages: The central receiver is unable to decode correctly the received message transmitted at the highest power level (messages transmitted at lower power levels are systematically corrupted) due to the accumulative noise plus interference. The failure probability encountered by a transmitted message is given by

$$\gamma\left(q\right) = \begin{cases} 1-(1-q)^{m-1}, & \text{BEB}, \\ 1-\sum\limits_{j=0}^{m-1} Q_r(j, m-1)A_{j+1}, & \text{MPL-BEB}. \end{cases} \tag{13}$$

Initial contention window : Let K denotes the maximum backoff stage. Since the behavior of BEB and MPL-BEB is the same either for a failure due to a collision of the message itself or a failure due to a lost ACK. It follows that from earlier alternative studies (e.g., see [3]), the retransmission probability can be written as the following Bianchi's fixed point equation :

$$q = \frac{2(1-2\gamma)}{(1-2\gamma)(W_0+1) + \gamma W_0\left(1-(2\gamma)^K\right)}. \tag{14}$$

Thus, one can easily estimate the initial contention window W_0 function of transmission probability q and conditional failure probability γ. Yet

$$W_0 = \frac{2-q}{q\left[1+\gamma\sum\limits_{k=0}^{K-1}(2\gamma)^k\right]}. \tag{15}$$

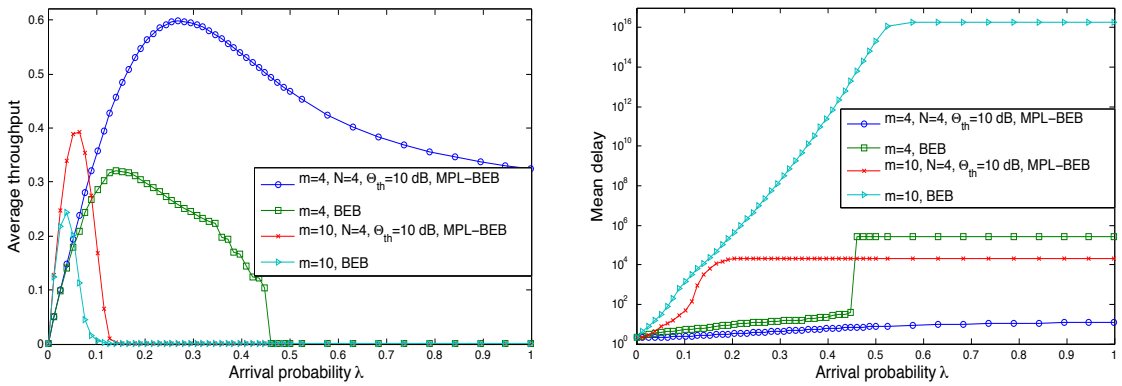
Remark 1 *It is important to highlight that, when K=0 i.e. no exponential backoff is considered, the probability q results to be independent of γ and expression (14) becomes:*

$$q = \frac{2}{W_0+1}. \tag{16}$$

6 Numerical investigation

In this section, we undertake the numerical investigation of the game problem for the standard scheme, and the mechanism introduced herein, MLP-BEB. Throughout this section, we consider the average throughput as the utility function. Similar trends are obtained when minimizing expected delay.

 ISBN: 978-84-614-6167-7



(a) (b)

Figure 2: Average throughput and mean delay for m=4 MSs and m=10 MSs, at Nash equilibrium when taking the individual throughput as the payoff function.

MPL-BEB Vs BEB: We depict in Figures 2(a) and 2(b) the total throughput and the expected delay obtained at Nash equilibrium as function of the arrival probability. We see that equilibrium throughput is unimodal function of arrival probability, at low load it presents an increasing behavior until achieving a maximum throughput of $\Gamma_{max} \simeq 0.6$ at $\lambda \simeq 0.27$ and $\Gamma_{max} \simeq 0.4$ at $\lambda \simeq 0.06$ and $\Gamma_{max} \simeq 0.32$ at $\lambda \simeq 0.12$ and $\Gamma_{max} \simeq 0.5$ at $\lambda \simeq 0.24$ for respectively MPL-BEB (m=4, 10) and BEB (m=4, 10). Under delay minimization, we obtain nearly the same profile as the one obtained when maximizing individual throughput, which means that optimal retransmission probability that maximizes the throughput is very close to the delay minimizer. Figures 3 show the retransmission probability (Nash equilibrium strategy) and the initial window backoff as function of arrival probability λ. We see that the mobile stations become more and more aggressive as the arrival probability increases or the number of MSs increases which explains the dramatic decrease in the system's throughput and the respective huge delay. Moreover, the equilibrium retransmission drastically increases to 1 at heavy load.

From Figures 2 we observe that MPL-BEB achieves higher throughput and lower delay than BEB. For a low number of terminals ($m = 4$), results in Figure 3 (retransmission probabilities) show that the use of power diversity helps to "break" the effects of the selfish behavior of the MS, i.e. the MS that transmits with higher power plays the role of "dominant station" during that slot. Since the power level is chosen randomly, this role is fairly shared by all the stations in the subsequent slots. In fact, it is observed that MSs operating in the power diversity-enabled scheme are quite aggressive, so the capture effect produces successful transmissions even when all MSs use a high retransmission probability (i.e., a smaller initial contention window in figure 3). For a larger number of terminals ($m = 10$), results in Figure 3 show that the degree of power diversity is not enough to suppress the effects of the selfish behavior for large values of the arrival probability, e.g., it is more probable that two or more stations transmits using the higher power value.

 ISBN: 978-84-614-6167-7

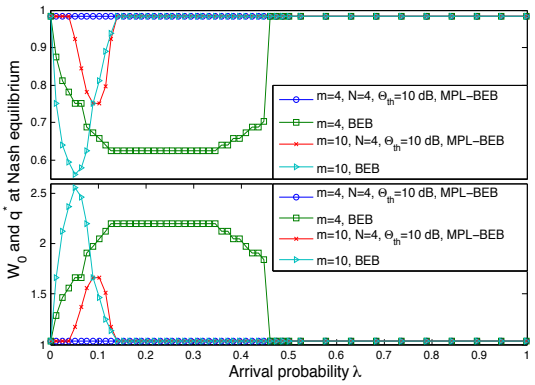


Figure 3: Retransmission probability and initial contention window (at Nash equilibrium) when maximizing the individual throughput for m=4 MSs and m=10 MSs.

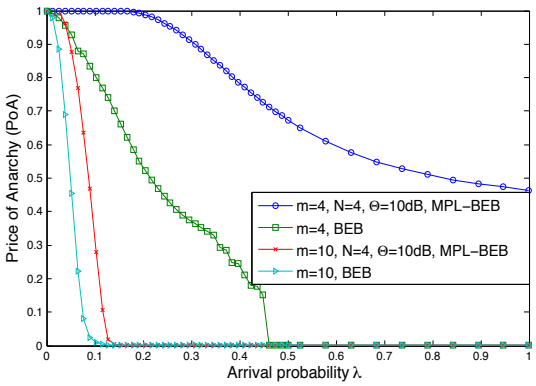


Figure 4: Price of anarchy of both BEB and MPL-BEB, for m=4 MSs and m=10 MSs.

Towards an efficient Nash equilibrium and anarchy removal: It has been proved that distributed channel access methods dealt successfully with coordination, synchronization, signaling complexity and fault-tolerance issues. Unfortunately, the related players (mobile stations in our case) don't care whether their decisions improve or hurt the overall system. They only care about maximizing their own payoff. In order to measure how the loss of efficiency generated by BEB and MPL-BEB due to selfish behavior of mobile stations, we proceed here to analyze the Price of Anarchy (PoA). We define is as the the ratio of the Nash throughput to the social optimum throughput. Figure 4 depicts the price of anarchy shows that for small-sized network and under low workload, MPL-BEB achieves the social optimum. Under average and high workload, MPL-BEB seems to keep a very good tradeoff performance-decentralization, whereas standard BEB experiences a performance collapse. For large-sized networks, the loss of efficiency is considerable and the global performance tends to 0 except for low workload. From these statements, one argue that an admission control could be a promising solution to take advantage from power diversity.

7 Concluding remarks

In this paper, we undertake a stochastic game analysis of the binary exponential backoff mechanism making use of multiple power levels. Our analysis shows that the capture effect improves the rate of successful transmissions. Under this scheme, since the power level is chosen randomly and independently by each station, the MS transmitting at the highest power level, playing the role of "dominant station", is fairly shared by all the stations.
Our results show that for large-sized networks, the price of anarchy decreases rapidly as a function of the arrival probability, which means a huge gap between centralized setup and decentralized setup. In order to keep the load of the system at reasonable levels and therefore take advantage from power diversity, an admission control seems to be a good solution. Another possible extension is to implement jointly to the power diversity enabled BEB a suitable pricing policy to heal the aggressive behavior of the MSs.

 ISBN: 978-84-614-6167-7

Acknowledgements

This work was supported in part by the Ministry of Science and Technology of Mexico (Conacyt) under Project 000000000139822, and by the Ministry of Science and Technology of Spain under CONSOLIDER Project CSD2006-46, and by ITAM.

References

[1] D. Bertsekas, R. Gallager, "Data Networks". Prentice Hall, Englewood Cliffs, New Jersey, 1987.

[2] J. Goodman, A. Greenberg, N. Madras, P. March, "Stability of binary exponential backoff". Journal of the ACM, Volume 35(3), Pages 579-602, 1988.

[3] G. Bianchi "Performance Analysis of the IEEE 802.11 Distributed Coordination Function". IEEE Journal on Selected Areas in Communications, Volume 18(3), Pages 535-548, 2000.

[4] J. H. Sarker, M. Hassan, S. Halme, "Power level selection schemes to improve throughput and stability of Slotted Aloha under heavy load". Computer Communication 25, 2002.

[5] E. Altman, D. Barman, A. Benslimane and R. El-Azouzi, "Slotted Aloha with priorities and random power". In Proceedings of IFIP Networking, Ontario, pages 610-622, 2005.

[6] A. Karouit, L. Orozco Barbosa and A. Haqiq,"Team study of the IEEE 802.16 Collision Resolution Protocol". In Proceeding of IFIP Wireless days, pages 1-6, Venezia, Italy, 2010.

[7] J. Lee, A. Tang, J. Huang, M. Chiang, and A. Calderbank,"Reverse-Engineering MAC: A Non-Cooperative Game Model".IEEE Journal on Selected Areas in Communications, pages 1135-1145, Volume 25(6), August 2007.

[8] A. Karouit, E. Sabir, L. Orozco Babrobsa, F. Ramirez-Mireles and A. Haqiq, "A Cross-layered Binary Exponential Backoff Algorithm for Initial and Bandwidth Request Ranging in IEEE 802.16". Unpublished manuscript.

[9] R. El-Azouzi, E. Sabir, T. Jiminez and E. H. Bouyakhf, "Modeling Slotted Aloha as a Stochastic Game with Random Discrete Power Selection Algorithms". International Journal of Computer Systems, Networks, and Communications, 2009.

[10] G. E. Andrews, *The theory of partitions*, Addison-Wesley Pub. Co., Advanced Book Program (Reading, Mass), 1976.

[11] D. Dutta, A. Goel and J. Heidemann, "Oblivious AQM and Nash equilibria". In proceedings of IEEE Infocom, 2003.

[12] E. Sabir, R. El-Azouzi and Y. Hayel, "Hierarchy Sustains Partial Cooperation and Induces a Braess-like Paradox in Slotted Aloha-based Networks". To appear in Computer Communications.

[13] R. Nelson, "Probability, stochastic process, and queueing theory". The Mathematics of Computer Performance Modelling. Springer-Verlag, third printing, 2000.

A Transition probabilities of standard BEB

$P_{(n,a),(n+k,b)}(q_{-i}, q_i) =$

$$\begin{cases} \left.\begin{array}{ll} Q_a(k,n)[(1-q_i)(1-Q_r(1,n)) + q_i(1-Q_r(0,n)] & \\ +(1-q_i)Q_a(k+1,n)Q_r(1,n)], & a=1,b=1 \\ (1-\lambda)[Q_a(k,n)(1-Q_r(1,n)) + Q_a(k+1,n)Q_r(1,n)], & a=0,b=0 \\ \lambda[Q_a(k,n)(1-Q_r(1,n)) + Q_a(k+1,n)Q_r(1,n)], & a=0,b=1 \\ q_i Q_a(k,n)Q_r(0,n), & a=1,b=0 \end{array}\right\} 0 \le k \le M-n+1 \\ \\ \left.\begin{array}{ll} (1-q_i)Q_a(0,n))Q_r(1,n) & a=1,b=1 \\ (1-\lambda)Q_a(0,n))Q_r(1,n) & a=0,b=0 \\ \lambda Q_a(0,n))Q_r(1,n) & a=0,b=1 \end{array}\right\} k=-1 \\ \\ 0 \quad \text{otherwise.} \end{cases}$$

B Transition probabilities of MPL-BEB

$P_{(n,a),(n+k,b)}(q_{-i}, q_i) =$

$$\begin{cases} \left.\begin{array}{ll} Q_a(k,n)[(q_i \sum_{j=0}^{n} Q_r(j,n)(1-A_{j+1}) + (1-q_i)\sum_{j=0}^{n} Q_r(j,n)(1-A_j)] + & \\ (1-q_i)Q_a(k+1,n)\sum_{j=0}^{n} Q_r(j,n)A_j + q_i Q_a(k+1,n)\sum_{j=0}^{n} \frac{j}{j+1}Q_r(j,n)A_{j+1}, & a=1,b=1 \\ (1-\lambda)[Q_a(k,n)\sum_{j=0,j\neq 1}^{n} Q_r(j,n)(1-A_j) + Q_a(k+1,n)\sum_{j=1}^{n} Q_r(j,n)A_j], & a=0,b=0 \\ \lambda[Q_a(k,n)\sum_{j=0,j\neq 1}^{n} Q_r(j,n)(1-A_j) + Q_a(k+1,n)\sum_{j=1}^{n} Q_r(j,n)A_j], & a=0,b=1 \\ q_i Q_a(k,n)\sum_{j=0}^{n} Q_r(j,n)\frac{A_{j+1}}{j+1}, & a=1,b=0 \end{array}\right\} 0 \le k \le M-n \\ \\ \left.\begin{array}{ll} Q_a(0,n)[(1-q_i)\sum_{j=1}^{n} Q_r(j,n)A_j + q_i\sum_{j=1}^{n}\frac{j}{j+1}Q_r(j,n)A_{j+1}], & a=1,b=1 \\ (1-\lambda)Q_a(0,n)\sum_{j=1}^{n} Q_r(j,n)A_j, & a=0,b=0 \\ \lambda Q_a(0,n)\sum_{j=1}^{n} Q_r(j,n)A_j, & a=0,b=1 \end{array}\right\} k=-1 \\ 0 \quad \text{otherwise.} \end{cases}$$

 ISBN: 978-84-614-6167-7

Proceedings of the 11th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2011
26–30 June 2011.

Interactions and Focusing of Nonlinear Water Waves

Harihar Khanal[1], Stefan C. Mancas[1] and Shahrdad G. Sajjadi[2]

[1] *Department of Mathematics, Embry-Riddle Aeronautical University*

[2] *Center for Geophysics and Planetary Physics, Embry-Riddle Aeronautical University*

emails: harihar.khanal@erau.edu, mancass@erau.edu, sajja8b5@erau.edu

Abstract

A theoretical and computational study is undertaken for the modulational instabilities of a pair of nonlinearly interacting two-dimensional waves in deep water. It has been shown that the full dynamics of these interacting waves gives rise to localized large-amplitude wavepackets (wave focusing). The coupled cubic nonlinear Schrödinger (CNLS) equations are used to derive a nonlinear dispersion equation which give rise to new class of modulational instabilities and demonstrates the dependence of obliqueness of the interacting waves. The computations, due to nonlinear wave-wave interactions, waves that are separately modulationally stable can give rise to the formation of large-amplitude coherent wave packets with amplitudes several times that of the initial waves. In the case for the original Benjamin-Feir instability, in contrast, waves disintegrate into a wide spectrum.

Key words: water waves, coupled nonlinear Schrödinger, fast spectral algorithm

1 Introduction

Extremely large size waves (commonly known as freak, rogue or giant waves) are very common in the open sea or ocean and they pose major hazard to mariners. As early as 1976, Peregrine [11] suggested that in the region of oceans where there is a strong current present, freak waves can form when action is conncentrated by reflection into a caustic region. A variable current acts analogously to filamentation instability in laser-plasma interactions [8, 9]. Freak waves are very steep and is a nonlinear phenomena, hence they cannot be represented and described by a linear water wave theory. Zakharov [19] has noted that in the last stage of their evolution, their steepness becomes 'infinite', thereby forming a 'wall of water', such as that shown in Fig. 1. However, before such an instant in time, the

 ISBN: 978-84-614-6167-7

steepness is higher than one for the limiting Stokes wave and before breaking the wave crest reaches three to four (sometimes even more) times higher than the crests of neighboring waves. The freak wave is preceded by a deep trough appearing as a 'hole in the sea'. On the other hand, a characteristic life time of a freak wave is short, typically ten of wave periods or so. For example, if the wave period is fifteen seconds, then their life time is just few minutes. Freak wave appears almost instantly from a relatively calm sea. It is, therefore, easy to appriciate that such peculiar features of freak waves cannot be explained by means of a linear theory. Even the focusing of ocean waves is a preconditions for formation of such waves.

Figure 1: A photograph of a rouge wave, depicting the enormous height of the wave and its nonlinear character.

It is now quite common to associate appearance of freak waves with the modulation instability of Stokes waves. This instability (known as the Benjamin–Feir instability) was first discovered by Lighthill [7] and the detail of theory was developed independently by Benjamin and Feir [2] and by Zakharov [15]. Zakharov showed slowly modulated weakly nonlinear Stokes wave can be described by nonlinear Shrödinger equation (NLSE) and that this equation is integrable [16] and is just the first term in the hierarchy of envelope equations describing packets of surface gravity waves. The second term in this hierarchy was calculated by Dysthe [4].

Since the pioneer work of Smith [13], many researchers attempted (both theoretically or numerically) to explain the freak wave formation by NLSE. Among diversed results obtained by them there is one important common observation which has been made by

 ISBN: 978-84-614-6167-7

all, and that is, nonlinear development of modulational instability leads to concentration of wave energy in a small spatial region. This marks the possibility for formation of freak wave. Modulation instability leads to decomposition of initially homogeneous Stokes wave into a system of envelope solitons, or more stricktly quasi-solitons [17, 18].

This state can be called "solitonic turbulence", or "quasisolitonic turbulence".

In this paper, we consider the problem of a single solition in a homogeneous media, being subjected to modulational instability which eventually leads to formation of a system of soliton. We will show that the supercritical instability leads to maximum formation of soliton, concentrated in a small region. Moreover, in going through subcritical instability the solitons coagulate to early stages of supercritical instability.

Moreover, we investigate the full dynamics of nonlinearly interacting deep water waves subjected to modulational/filamentation instabilities. It is found that random perturbations can grow to form inherently nonlinear water wave structures, the so called freak waves, through the nonlinear interaction between two coupled water waves. The latter should be of interest for explaining recent observations in water wave dynamics.

2 Formulation of the problem

In a pioneering work, a theory for the modulational instability of a pair of two-dimensional nonlinearly coupled water waves in deep water, as well as the formation and dynamics of localized freak wave packets was presented [12]. Likewise we follow suite in derivation of CNLS equations. Thus, we use the CNLS equations derived by Onorato *et al.* [10], which are valid for a system of obliquely propagating waves. As in [10] we define the $x-$axis as the middle between the two directions of propagation. Thus we define wavenumbers as $\mathbf{k}_A = (k_{A,x}, k_{A,y}) \equiv (k, \ell)$ and $\mathbf{k}_B = (k_{B,x}, k_{B,y}) \equiv (k, -\ell)$ with the undestanding that both k and ℓ are positive. The frequencies ω_j of the two carrier waves ($j = A,\ B$) are then related to the wavevectors $\mathbf{k}_j$ by the dispersion relation for deep water waves [6] $\omega_j = \sqrt{g|\mathbf{k}_j|}$, where g is acceleration due to gravity. Accordingly, we may define $\omega_A = \omega_B = \sqrt{g\kappa}$, where κ is the wavenumber norm given by $\kappa \equiv \sqrt{k^2 + \ell^2}$.

Multiplying the system of two-dimensional CNLS equations given in [10] by i, we obtain

$$i\left(\frac{\partial A}{\partial t} + C_x\frac{\partial A}{\partial x} + C_y\frac{\partial A}{\partial y}\right) + \alpha\frac{\partial^2 A}{\partial x^2} + \beta\frac{\partial^2 A}{\partial y^2} + \gamma\frac{\partial^2 A}{\partial x\partial y} - \xi\,|A^2|A - 2\zeta\,|B|^2 A = 0\,, \qquad (1)$$

and

$$i\left(\frac{\partial B}{\partial t} + C_x\frac{\partial B}{\partial x} - C_y\frac{\partial B}{\partial y}\right) + \alpha\frac{\partial^2 B}{\partial x^2} + \beta\frac{\partial^2 B}{\partial y^2} - \gamma\frac{\partial^2 B}{\partial x\partial y} - \xi\,|B^2|B - 2\zeta\,|A|^2 B = 0, \qquad (2)$$

where A and B are the amplitudes of the slowly varying wave envelopes and the corresponding surface elevations are given by

$$\{\eta_A, \eta_B\} = \tfrac{1}{2}\{A(\mathbf{r}, t), B(\mathbf{r}, t)\}\exp(ikx + i\ell y - i\omega t) + \text{c.c.}$$

 ISBN: 978-84-614-6167-7

where c.c. denotes complex conjugate. The x and y components of the group velocity are given respectively by

$$C_x = \omega k/2\kappa^2 \quad \text{and} \quad C_y = \omega \ell/2\kappa^2$$

and the group velocity dispersion coefficients are

$$\alpha = \omega(2\ell^2 - k^2)/8\kappa^4, \quad \beta = \omega(2k^2 - \ell^2)/8\kappa^4 \quad \text{and} \quad \gamma = -3\omega\ell k/4\kappa^4.$$

Also, the nonlinearity coefficients (as in [10]) are given by $\xi = \omega\kappa^2/2$ and

$$\zeta = \omega(k^5 - k^3\ell^2 - 3k\ell^4 - 2k^4\kappa + 2k^2\ell^2\kappa + 2\ell^4\kappa)/2\kappa^2(k - 2\kappa).$$

It is now amply confirmed that the two-dimensional CNLS equations (1) and (2) has temporal solutions

$$A_{eq} = A_0 \exp[-i(\xi|A_0|^2 + 2\zeta|B_0|^2)t] \quad \text{and} \quad B_{eq} = B_0 \exp[-i(\xi|B_0|^2 + 2\zeta|A_0|^2)t],$$

and we may use these soloutions to derive the nonlinear dispersion relation. Thus, assuming a small linear harmonic perturbation with the wavevector $\mathbf{K} = (K, L)$ and the frequency Ω, around the equilibrium solution given by

$$A = [A_0 + \epsilon A_1 + O(\epsilon^2)] \exp[-i(\xi|A_0|^2 + 2\zeta|B_0|^2)t]$$

and

$$B = [B_0 + \epsilon B_1 + O(\epsilon^2)] \exp[-i(\xi|B_0|^2 + 2\zeta|A_0|^2)t]$$

where $\epsilon \ll 1$ is a real parameter. Substituting theses into equations (1) and (2), linearizing in ϵ, then separating the real and imaginary parts, combining the resulting equations, and Fourier transforming, we obtain the nonlinear dispersion relation

$$[(\Omega - C_xK - C_yL)^2 - \Omega_1^2]\,[(\Omega - C_xK + C_yL)^2 - \Omega_2^2] = \Omega_c^4\,, \tag{3}$$

where

$$\Omega_1^2 = (\alpha K^2 + \beta L^2 - \gamma KL)(\alpha K^2 + \beta L^2 + \gamma KL + 2\xi|A_0|^2),$$
$$\Omega_2^2 = (\alpha K^2 + \beta L^2 + \gamma KL)(\alpha K^2 + \beta L^2 - \gamma KL + 2\xi|B_0|^2)$$

and

$$\Omega_c^4 = 16\,\zeta^2\,|A_0|^2|B_0|^2\,(\alpha K^2 + \beta L^2 - \gamma KL)(\alpha K^2 + \beta L^2 + \gamma KL).$$

For computational purposes, it is convenient to make variables dimensionless. Thus, defining the wave steepness by κA and κB, we make the wave amplitudes dimensionless according to $A_0 = A_0'/\kappa$ and $B_0 = B_0'/\kappa$. Similarly, the wavenumbers and frequencies are

 ISBN: 978-84-614-6167-7

made dimensionless in the following manner: $K' = K/\kappa$, $L' = L/\kappa$, $k' = k/\kappa$, $\ell' = \ell/\kappa$, and $\Omega' = \Omega/\omega$. The remaining coefficients are also made dimensionless, using the adoptation

$$C'_x = \frac{C_x\kappa}{\omega} = \frac{k'}{2}, \qquad C'_y = \frac{C_y\kappa}{\omega} = \frac{\ell'}{2},$$

$$\alpha' = \frac{\alpha\kappa^2}{\omega} = \frac{2\ell'^2 - k'^2}{8}, \qquad \beta' = \frac{\beta\kappa^2}{\omega} = \frac{2k'^2 - \ell'^2}{8},$$

$$\gamma' = \frac{\gamma\kappa^2}{\omega} = -\frac{3\ell' k'}{4}, \qquad \xi' = \frac{\xi}{\omega k^2} = \frac{1}{2},$$

and

$$\zeta' = \frac{\zeta}{\omega k^2} = \frac{(k')^5 - (k')^3(\ell')^2 - 3k'(\ell')^4 - 2(k')^4 + 2(k')^2(\ell')^2 + 2(\ell')^4}{2(k'-2)}.$$

Hence, Eqs. (1) and (2) will remain the same except all the variables are now replaced with their primed counterparts. Note that, $k' = \cos\theta$ and $\ell' = \sin\theta$, where θ is the angle between the wave directions.

We remark that, in what follows, we will drop the 'dash' notation for the sake of clarity.

3 Numerical Approach

The nonlinear strongly coupled system of equations (1) and (2) will be computed using a fast numerical algorithm based on the spectral method [3, 14] which is explained below.

3.1 Fourier Spectral Method

Let $S = A + B$ and $D = A - B$, and consider the following system of equations obtained from (1) and (2)

$$i\left(\frac{\partial S}{\partial t} + C_x\frac{\partial S}{\partial x} + C_y\frac{\partial D}{\partial y}\right) + \alpha\frac{\partial^2 S}{\partial x^2} + \beta\frac{\partial^2 S}{\partial y^2} + \gamma\frac{\partial^2 D}{\partial x\partial y} = g(S,D) \tag{4}$$

$$i\left(\frac{\partial D}{\partial t} + C_x\frac{\partial D}{\partial x} + C_y\frac{\partial S}{\partial y}\right) + \alpha\frac{\partial^2 D}{\partial x^2} + \beta\frac{\partial^2 D}{\partial y^2} + \gamma\frac{\partial^2 S}{\partial x\partial y} = g(D,S) \tag{5}$$

where

$$g(u,v) = \frac{1}{8}\left[(\xi+2\eta)\left(|u+v|^2 + |u-v|^2\right)u + (\xi-2\eta)\left(|u+v|^2 - |u-v|^2\right)v\right] \tag{6}$$

First, we reduce the above system of PDEs (4)–(5) into a system of ODEs using Fourier transform. The Fourier transform of $u(x,y)$ is defined by

$$\mathcal{F}(u)(k_x,k_y) = \widehat{u}(k_x,k_y) = \frac{1}{2\pi}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-i(k_x x + k_y y)}u(x,y)\,dx\,dy, \tag{7}$$

 ISBN: 978-84-614-6167-7

with the corresponding inverse

$$\mathcal{F}^{-1}(\widehat{u})(x,y) = u(x,y) = \frac{1}{2\pi}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{i(k_x x + k_y y)}\widehat{u}(k_x,k_y)\,dk_x\,dk_y. \tag{8}$$

The function $\widehat{u}(k_x,k_y)$ can be interpreted as the amplitude density of u for wavenumbers k_x, k_y. Now, we take the Fourier transform of both (4) and (5) as

$$i\frac{d\widehat{S}_t}{dt} - \left(k_x C_x + \alpha k_x^2 + \beta k_y^2\right)\widehat{S} - k_y\left(C_y + \gamma k_x\right)\widehat{D} = \widehat{g(S,D)}, \tag{9}$$

$$i\frac{d\widehat{D}_t}{dt} - \left(k_x C_x + \alpha k_x^2 + \beta k_y^2\right)\widehat{D} - k_y\left(C_y + \gamma k_x\right)\widehat{S} = \widehat{g(D,S)}, \tag{10}$$

Let $k_x C_x + \alpha k_x^2 + \beta k_y^2 = p$ and $k_y\left(C_y + \gamma k_x\right) = r$. Then, the equations (9) and (10) can be written in the matrix form as

$$i\frac{d}{dt}\begin{pmatrix}\widehat{S}\\ \widehat{D}\end{pmatrix} = \begin{pmatrix} p & r\\ r & p\end{pmatrix}\begin{pmatrix}\widehat{S}\\ \widehat{D}\end{pmatrix} + \begin{pmatrix}\widehat{g(S,D)}\\ \widehat{g(D,S}\end{pmatrix} \tag{11}$$

Computing the eigenvalues and eigenvectors the solution to (11) can be written as

$$\begin{aligned}\begin{pmatrix}\widehat{S}\\ \widehat{D}\end{pmatrix} &= \frac{1}{2}\begin{pmatrix} e^{-i\lambda_1 t} + e^{-i\lambda_2 t} & -e^{-i\lambda_1 t} + e^{-i\lambda_2 t}\\ -e^{-i\lambda_1 t} + e^{-i\lambda_2 t} & e^{-i\lambda_1 t} + e^{-i\lambda_2 t}\end{pmatrix}\begin{pmatrix}\widehat{S}(0)\\ \widehat{D}(0)\end{pmatrix}\\ &\quad + \frac{1}{2}\int_0^t \begin{pmatrix} e^{i\lambda_1\tau}\left(\widehat{g(S,D)} - \widehat{g(D,S)}\right)\\ e^{i\lambda_2\tau}\left(\widehat{g(S,D)} + \widehat{g(D,S)}\right)\end{pmatrix} d\tau\end{aligned} \tag{12}$$

with $\lambda_1 = k_x C_x + \alpha k_x^2 + \beta k_y^2 - k_y\left(C_y + \gamma k_x\right)$ and $\lambda_2 = k_x C_x + \alpha k_x^2 + \beta k_y^2 + k_y\left(C_y + \gamma k_x\right)$.

3.2 Spatial discretization (Discrete Fourier Transform)

We discretize the spatial domain $\Omega = [-L/2, L/2]\times[-L/2, L/2]$ into $n\times n$ uniformly spaced grid points $X_{ij} = (x_i, y_j)$ with $\Delta x = \Delta y = L/n$, n even, and L the length of the rectangular mesh Ω. Given $u(X_{ij}) = U_{ij}$, $i,j = 1,2,\cdots,n$, we define the 2D Discrete Fourier transform (2DFT) of u as

$$\widehat{u}_{k_x k_y} = \Delta x\Delta y\sum_{i=1}^{n}\sum_{j=1}^{n} e^{-i(k_x x_i + k_y y_j)}U_{ij}, \quad k_x, k_y = -\frac{n}{2}+1,\cdots,\frac{n}{2} \tag{13}$$

and its inverse 2DFT as

$$U_{ij} = \frac{1}{(2\pi)^2}\sum_{k_x=-n/2+1}^{n/2}\sum_{k_y=-n/2+1}^{n/2} e^{i(k_x x_i + k_y y_j)}\widehat{u}_{k_x k_y}, \quad i,j = 1,2,\cdots,n. \tag{14}$$

In equation (13) and (14) the wavenumbers k_x and k_y, and the spatial indexes i and j, take only integer values.

 ISBN: 978-84-614-6167-7

3.3 Temporal discretization

We solve the initial value problem of the ODE system (11) using the classical fourth order Runge-Kutta (RK4) method combined with the Super–Time–Stepping (STS) [1] and exact treatment for the linear part [3].

Given $t_{\max}$, we discretize the time domain $[0, t_{\max}]$ with equal time steps of size Δt with $t_n = n\Delta t$, $n = 0, 1, 2, \cdots$, and define $S^n = S(x, y; t_n)$ and $D^n = D(x, y; t_n)$. Initializing $\widehat{S}^n = \widehat{S}(t_n)$ and $\widehat{D}^n = \widehat{D}(t_n)$, we compute the Fourier transforms of the nonlinear terms $\mathcal{F}\left(g\left(\mathcal{F}^{-1}(\widehat{S}^n), \mathcal{F}^{-1}(\widehat{D}^n)\right)\right)$ and $\mathcal{F}\left(g\left(\mathcal{F}^{-1}(\widehat{D}^n), \mathcal{F}^{-1}(\widehat{S}^n)\right)\right)$, and advanced the ODE (11) in time with time step Δt using the explicit RK4 for the nonlinear part, together with an exact solution for the linear part as shown in (12). We exploit symmetry of the nonlinear function g in developing a numerical code to solve the system of ODEs (11). To overcome stability restriction on the stepsize Δt, we employ Super–Time–Stepping (STS) strategy [1].

The main idea behind the STS is to demand stability restriction only at the end of every N steps, consisting one super–step, instead of at every single step. The intermediate steps are chosen non–uniformly from a simple formula in terms of some modified Chebychev polynomials as

$$\tau_i = \frac{\Delta t}{(-1+\nu)\cos\left(\frac{2i-1}{N}\frac{\pi}{2}\right) + 1 - \nu}, \quad i = 1, 2, \cdots N, \quad 0 < \nu < 1. \tag{15}$$

As $\nu \to 0$, the duration of the superstep $\Delta t_{\text{sup}} = \sum_{i=1}^{N} \tau_i \to N^2\,\Delta t$. Thus, N substeps of a super step cover a time interval N times bigger than N explicit steps when $\nu \to 0$. For each choice of N, the scheme is stable for large enough ν. The larger the damping factor ν, the shorter the Δt_{sup} becomes, improving the accuracy at the expense of more computations. The length of the superstep Δt_{sup}, which is determined by the choice of N, ν and Δt, is only restricted by accuracy, just like in any unconditionally stable implicit methods.

The numerical code for the above procedure is implemented in Fortran 90 and executated on a linux cluster (of 128 nodes with dual Xeon 3.2GHz processors, 1024 KB cache 4GB with Myrinet) at Embry-Riddle Aeronautical University.

3.4 Simulation setup

The initial profiles for A and B were taken as the bell-shaped functions,

$$A(x, y; 0) = (A_0 + random(O(10^{-3}/\kappa))\, e^{-\sigma(x^2+y^2)} \tag{16}$$

$$B(x, y; 0) = (B_0 + random(O(10^{-3}/\kappa))\, e^{-\sigma(x^2+y^2)} \tag{17}$$

In the simulations reported here, we used the parameter values $\theta_0 = \pi/6$, $g = 9.81$, $w = 0.56$, $k = 0.33$, $A_0 = 0.1/\kappa$, $B_0 = A_0, 0$, $\sigma = 1, 0$, $L = 2$ and a grid of 256×256 nodes in the computational domain $[-1, 1] \times [-1, 1]$ with the time stepsize $\Delta t = 0.01$.

 ISBN: 978-84-614-6167-7

For each simulation we monitor the energies $Q_A(t)$ and $Q_B(t)$, calculated as

$$Q_A(t) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} |A(x,y;t)|^2\, dx\, dy = \sum_{i=1}^{n}\sum_{j=1}^{n} |A_{ij}|^2 \Delta x \Delta y \tag{18}$$

$$Q_B(t) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} |B(x,y;t)|^2\, dx\, dy = \sum_{i=1}^{n}\sum_{j=1}^{n} |B_{ij}|^2 \Delta x \Delta y \tag{19}$$

Observing a finite energy will reveal stability of a solution. As soon as the solution becomes unstable, the energy diverges. When the solution dissipates the energy approaches to zero.

4 Results

We commence this section by emphasizing that the results presented in this paper represents a very preliminary findings on dynamics of interacting nonlinear water waves. In due course, the full account of our findings will be reported elsewhere. The main emphasis here is the numerical methodologies for solution of equations (1) and (2).

The problem considered here comprises the dynamics of nonlinear interacting water wave packets through solving the coupled system of equations (1) and (2). The results of the simulation are displayed in Fig. 2 and Fig. 3. In this simulation, we have adopted the normalization $A' = A/\kappa$, $B' = B/\kappa$, $t' = \omega t$, $x' = \kappa x$, and $y' = \kappa y$ (the other scaled parameters are as those given above), for a single value of $\theta = \pi/6$. The results that are shown in Fig. 2 are all in dimensional units, where the two interacting waves initially have the amplitude $A = B = 0.1/\kappa + \text{ran}$, with ran representing a random low-amplitude noise, equal to $10^{-3}/\kappa$, in order to enhance instability. The results shown in Fig. 2 represent different time steps (starting on the left-hand panel and going downwards) for $t = 300/\omega$, $t = 600/\omega$, $t = 900/\omega$ then (right-hand panel) $t = 1200/\omega$, $t = 1500/\omega$; the last figure on the right-hand panel is at the same time as that above it but plotted from a different prospective reflecting the maximum growth rate in the y direction. For our simulations we have taken typical data from ocean waves [5]. Thus, choosing the frequency to be 0.09 Hz, we have $\omega = 0.56\,\text{s}^{-1}$, and $\kappa = \omega^2/g \approx 0.033\,\text{m}^{-1}$. The waves A and B in Fig. 2, then have the initial amplitudes $|A| = |B| = 0.1/\kappa \approx 3$ meters. From these figures, we see at $t = 1500/\omega$ (≈ 2680 seconds) that wave A focouses as a localized wave packets with a maximum amplitude of $\approx 0.35/\kappa \approx 10$ meters. We remark for considerable period after the initial step, waves A and B are qualitatively the same (with $|A| > |B|$) before the nonlinear wave-wave interactions set in which results to wave break-up.

We next consider the case of a single wave we have set B to zero, being the same as the standard Benjamin-Feir instability. In this case (see Fig. 3.), we do not see the formation of well-defined wave-packets, but the instability gave rise to a wide spectrum of waves in

 ISBN: 978-84-614-6167-7

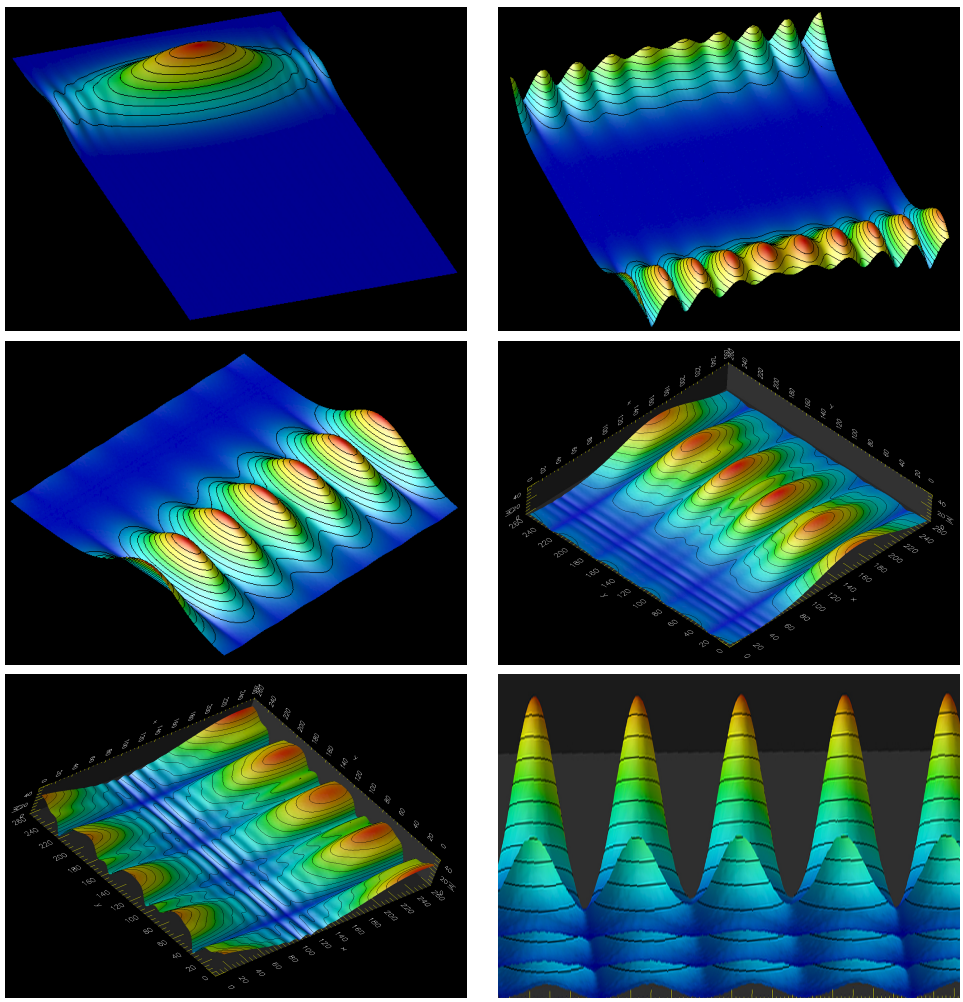


Figure 2: The interaction between two waves, with equal initial amplitudes $|A| = |B| = 0.1\,\kappa^{-1}$ which are propagating at an angle of $\theta = \pi/6$. A low-amplitude noise equal to $10^{-3}/\kappa$ is added to the initial amplitude in order to enhance the modulation instability.

 ISBN: 978-84-614-6167-7

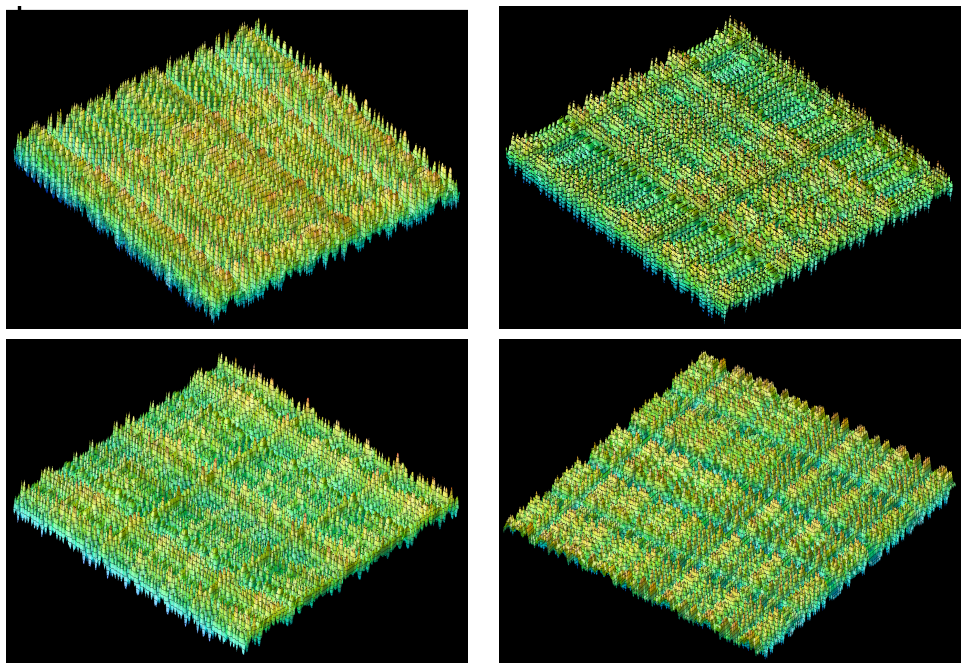


Figure 3: The amplitude $|A|$ (with $B = 0$ initially) for time $t = 300/\omega$, $t = 900/\omega$ (left-hand panel) and $t = 1200/\omega$, $t = 1500/\omega$ (right-hand panel).

different directions, in agreement with the standard linear analysis. This is in contrast with the new instability due to the coupling of the two waves, shown in Fig. 2., which has a well-defined maximum in the y direction, with the concentration of wave energy into localized wavepackets.

Hence, in summary, we have presented a theoretical and computational study of the modulational instabilities of a pair of nonlinearly interacting two-dimensional waves in deep water. we have demonstrated that the full dynamics of these interacting waves gives rise to localized large-amplitude wavepackets. Starting from the CNLS equations of [10] and following [12], we have derived a nonlinear dispersion equation which give rise to new class of modulational instabilities demonstrating the dependence of obliqueness of the interacting waves. Furthermore, the numerical analysis of the full dynamical system reveals that even waves that are separately modulationally stable can, when nonlinear interactions are taken into account, give rise to novel behavior such as the formation of large-amplitude coherent wave packets with amplitudes several times the initial waves. This behavior is quite different from that of a single wave (the case for the original Benjamin-Feir instability) which disintegrates into a wide spectrum of waves. These results are relevant to the nonlinear

 ISBN: 978-84-614-6167-7

instability arising from colliding water waves thereby producing large-amplitude oceanic freak waves.

Acknowledgements

This work has been partially supported by the Department of Mathematics, Embry-Riddle Aeronautical University.

References

[1] V. Alexiades, G. Amiez, and P.A. Gremaud, *Super-Time-Stepping acceleration of explicit schemes for parabolic problems*, Communications in Numerical Methods in Engineering, **12**, (1996), 31–42.

[2] T. B. Benjamin and J. E. Feir, *The desintegration of wave trains on deep water. Part 1. Theory*, J. Fluid Mech. **27** (1967) 417–430.

[3] G. Beylkin, J.M. Keiser, L. Vozovoi, *A New Class of Time Discretization Schemes for the Solution of Nonlinear PDEs*, J. Comp. Physics. **147** (1998) 362–387.

[4] K. B. Dysthe, *Note on a modification to the nonlinear Schrodinger equation for application to deep water waves*, Proc. Roy. Ser. A **369** (1979) 105–114.

[5] K. Hasselmann, W. Sell, D. B. Ross, P. Müller, *A parametric wave prediction model*, J. Phys. Oceanogr. **6**, 200 (1976).

[6] V. I. Karpman, *Nonlinear Waves in Dispersive Media*, Pergamon Press, New York, 1975.

[7] M. J. Lighthill, *Contribution to the theory of waves in nonlinear dispersive systems*, J. Inst. Math. Appl. **1** (1965) 269–306.

[8] D J Nicholas and S G Sajjadi, *Numerical simulation of filamentation in laser-plasma interactions*, J. Phys. D: Appl. Phys. **19** (1986) 737–749.

[9] D. J. Nicholas and S. G. Sajjadi, *The effect of light filamentation on uniformity of energy deposition in laser plasmas*, J. Plasma Physics **41** (1989) 209–218.

[10] M. Onorato, A. R. Osborne, and M. Serio, *Modulational Instability in Crossing Sea States: A Possible Mechanism for the Formation of Freak Waves*, Phys. Rev. Lett. **96** (2006) 014503-6.

 ISBN: 978-84-614-6167-7

[11] H. Peregrine, *Interaction of water waves and currents*, Adv. Appl. Mech. **16** (1976), 9–117.

[12] P. K. Shukla, I. Kourakis, B. Eliasson, M. Marklund, and L. Stenflo, *Instability and Evolution of Nonlinearly Interacting Water Waves*, Phys. Rev. Lett. **97** (2006) 094501-4.

[13] R. Smith, *Giant Waves*, Fluid Mech. **77** (1976), 417– 431.

[14] L. N. Trefethen, *Spectral Methods in Matlab*, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2000.

[15] V. E. Zakharov, *Stability of periodic waves of finite amplitude on a surface of deep fluid*, J. Appl. Mech. Tech. Phys. **9** (1968) 190–194.

[16] V. E. Zakharov, A. B. Shabat, *Exact theory of two-dimensional self-focusing and one-dimensional self-modulation of waves in nonlinear media*, Soviet Physics JETP **34** (1972) 62–69.

[17] V. E. Zakharov and E. A. Kuznetsov, *Optical solitons and quasisolitons*, JETP **86** (1998) 1035– 1046.

[18] V. E. Zakharov, F. Dias and A. N. Pushkarev, *One-Dimensional Wave Turbulence*, Phys. Reports **398** (2004) 1–65.

[19] V.E. Zakharov and L.A. Ostrovsky, *Modulation instability: The beginning*, Physica D **238** (2009) 540–548.

 ISBN: 978-84-614-6167-7

Proceedings of the 11th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2011
26–30 June 2011.

The ETD-CN Scheme for Reaction-Diffusion Problems

Britta Kleefeld[1], Abdul Q.M. Khaliq[2] and Bruce A. Wade[3]

[1] *Institut für Mathematik, Brandenburgische Technische Universität Cottbus, Postfach 101344, 03013 Cottbus, Germany.*

[2] *Department of Mathematical Sciences, Middle Tennessee State University, Murfreesboro, TN 37132-0001, USA.*

[3] *Department of Mathematical Sciences, University of Wisconsin–Milwaukee, Milwaukee, WI 53201-0413, USA.*

emails: britta.kleefeld@TU-Cottbus.de, akhaliq@mtsu.edu, wade@uwm.edu

Abstract

A novel Exponential Time Differencing (ETD) Crank-Nicolson method is developed which is stable, second order convergent, and highly efficient. In the nonsmooth data case we employ a positivity-preserving initial damping scheme to recover the full rate of convergence.

Key words: Exponential Time Differencing, exotic options, nonlinear Black-Scholes equation
MSC 2000: 65M12, 65M15, 65Y05, 65Y20

1 Introduction

Various types of Exponential Time Differencing Runge-Kutta schemes (ETDRK) for nonlinear parabolic equations have been proposed and investigated, though there is no complete theoretical analysis and the focus has not been on any specific efficient version like the one introduced in this article. This Exponential Time Differencing Crank-Nicolson (ETD-CN) scheme utilizes an exponential time differencing Runge-Kutta approach followed by a (1,1)–diagonal Padé approximation of matrix exponential functions. This is an extension of several previous papers on Exponential Time Differencing schemes, in particular [2, 3, 5, 6, 8, 10, 13, 16, 17, 18].

 ISBN: 978-84-614-6167-7

Consider the following nonlinear parabolic initial-boundary value problem:

$$\begin{aligned} u_t + Au &= F(t,u) \qquad \text{in } \Omega, \qquad t \in (0,T\,), \\ u(\cdot,0) &= u_0 \qquad\quad \text{in } \Omega, \end{aligned} \tag{1}$$

where Ω is a bounded domain in $\mathbb{R}^d$ with Lipschitz continuous boundary, A represents a uniformly elliptic operator, and F is a sufficiently smooth, nonlinear reaction term. One should have in mind the following type of differential operator:

$$A := -\sum_{j,k=1}^{d} \frac{\partial}{\partial x_j}\left(a_{j,k}(x)\frac{\partial}{\partial x_k}\right) + \sum_{j=1}^{d} b_j(x)\frac{\partial}{\partial x_j} + b_0(x),$$

where the coefficients $a_{j,k}$ and b_j are C^∞ (or sufficiently smooth) functions on $\overline{\Omega}$, $a_{j,k} = a_{k,j}$, $b_0 \geq 0$, and for some $c_0 > 0$

$$\sum_{j,k=1}^{d} a_{j,k}(\cdot)\xi_j\xi_k \geq c_0|\xi|^2, \quad \text{on } \overline{\Omega}, \quad \text{for all } \xi \in \mathbb{R}^{\mathrm{d}}.$$

However, we shall use A and F based on an abstract formulation for convenience of the development of the numerical scheme and its analysis. The initial value problem (1) is reset to be posed in a Banach space $\mathcal{X}$, as follows. Consider now A to be a linear, self-adjoint, positive definite, closed operator with a compact inverse, defined on a dense domain $D(A) \subset \mathcal{X}$. The operator A could represent any of $\{A_h\}_{0<h\leq h_0}$, obtained through spatial discretization, and $\mathcal{X}$ could be $\mathcal{X} = S_h$, an appropriate finite dimensional subspace of $L^2(\Omega)$.

We assume *cf.* [12, Remark 1.1, p. 324] that D is a locally closed subset of $\mathcal{X}$ and $F : [0,\infty) \times D \to \mathcal{X}$ is continuous. Then, it follows that for $z \in D$ there are positive numbers R, M, and T such that

$$\begin{aligned} S_R &\equiv \{x \in D : \|x - z\| \leq R\} \text{ is closed}, \\ \|F(t,x)\| &\leq M - 1, \quad \text{if } (t,x) \in [0,T] \times S_R, \quad \text{and} \\ \|E(t)z + y - z\| &\leq R/2, \qquad \text{if } t \in [0,T] \text{ and } y \in \mathcal{X} \text{ with } \|y\| \leq T(M-1)e^{\omega T}, \end{aligned} \tag{2}$$

where ω is the least number such that $\|E(t)\| \leq e^{\omega t}\ \forall\, t \geq 0$.

We also assume that $x \in S_R$ implies $E(t)x \in D$, $\forall\, t \in [0,T]$ and that

$$\liminf_{h\to 0^+} \frac{d(x + hF(t,x); D)}{h} = 0, \qquad \forall\ (t,x) \in [0,T] \times D.$$

By the Duhamel principle we know that each solution of (1) must be of the form

$$u(t) = E(t)u_0 + \int_0^t E(t-s)\, F(s,u(s))\, ds. \tag{3}$$

We are assuming

 ISBN: 978-84-614-6167-7

Assumption 1 *Let $F : [0,T] \times \mathcal{X} \to \mathcal{X}$ and U be an open subset of $[0,T] \times \mathcal{X}$. For every $(t,x) \in U$ there exists a neighborhood $V \subset U$ and a real number L_T such that*

$$\|F(t_1,x_1) - F(t_2,x_2)\|_{\mathcal{X}} \leq L_T\Big(|t_1 - t_2| + \|x_1 - x_2\|_{\mathcal{X}}\Big) \tag{4}$$

for all $(t_i,x_i) \in V$.

Let $0 < k \leq k_0$, for some k_0, and $t_n = nk$, $0 \leq n \leq N$. Replacing t by $t+k$, using basic properties of E and by the change of variable $s - t = k\tau$, we arrive at the following recurrence formula for the exact solution:

$$u(t_{n+1}) = e^{-kA}u(t_n) + k\int_0^1 e^{-kA(1-\tau)}F(t_n + \tau k, u(t_n + \tau k))\,d\tau. \tag{5}$$

This is the basis for deriving ETD schemes.

There are several ways of approximating the integral representation of the exact solution (5). Cox and Matthews [2] developed time stepping schemes by using polynomial formulae which give a Runge-Kutta type higher order approximation. Du and Zhu [4] and Kassam and Trefethen [10], and Nie, Zhang, and Zhao [14] pointed out some implementation difficulties. These ETD schemes are studied in many articles, *cf.,* [4, 5, 6, 8, 13]. Prior treatments of ETD methods do not discretize the matrix exponentials, or else employ contour integration techniques. In this work we focus on the development of a highly efficient scheme by employing Padé approximations for some crucial terms. The proofs given in previous articles omit the fully discrete case, thus leaving a gap in the convergence theory. In this section we derive a highly efficient, fully discrete second order version of the ETD schemes and in the next section we address the theory.

We will approximate e^{-kA} using the $(0,1)-$ and $(1,1)-$ Padé schemes, $R_{0,1}(kA)$ and $R_{1,1}(kA)$, respectively, as follows. Specifically, we define $R_{0,1}(kA) := (I + kA)^{-1}$ and $R_{1,1}(kA) := (2I - kA)(2I + kA)^{-1}$, which is commonly called the 'Crank-Nicolson,' or CN, scheme.

Denoting the semi-discrete approximation to $u(t_n)$ by u_n (note that only the time-variable is discretized) and $F(t_n,u_n)$ by F_n, the simplest approximation to the integral is to impose that F is constant for $t \in [t_n, t_{n+1}]$, i.e. $F \approx F_n$. This yields (from (5))

$$\begin{aligned} u_{n+1} &\approx e^{-kA}u_n + e^{-kA}k\int_0^1 e^{kA\tau}\,d\tau\,F_n \\ &= e^{-kA}u_n - A^{-1}\Big(e^{-kA} - I\Big)\,F_n. \end{aligned} \tag{6}$$

This semi-discrete scheme *cf.* [2, 13] is not useful until the matrix exponential is discretized

 ISBN: 978-84-614-6167-7

. Noting that

$$\begin{aligned} -A^{-1}\big(e^{-kA}-I\big) &\approx -A^{-1}\big((I+kA)^{-1}-I\big) \\ &= -A^{-1}\big(I-(I+kA)\big)(I+kA)^{-1} \\ &= k\,(I+kA)^{-1} \\ &= kR_{0,1}\big(kA\big), \end{aligned} \tag{7}$$

we arrive at the following fully discrete first order scheme, where v now denotes the fully discrete solution. This is the same as a standard first order linearly implicit scheme, in particular, [2, 4, 5, 6], the interesting aspect now being that this ETD derivation leads to an extended family of similar type of ETD schemes. on of ETD-CN. To obtain a second order accurate RK-scheme, we employ (6) as intermediate prediction of $u(t_{n+1})$, letting

$$a_n := e^{-kA}u_n - A^{-1}\left(e^{-kA}-I\right)F(t_n,u_n).$$

We then approximate the integral in (5) by

$$F \approx F_n + (t-t_n)\frac{F(t_n+k,a_n)-F_n}{k} \qquad t\in[t_n,t_{n+1}].$$

Using (5), a short calculation yields the following:

$$\begin{aligned} u_{n+1} &= e^{-kA}u_n + ke^{-kA}\int_0^1 e^{kA\tau}\left(F_n + k\tau\frac{F(t_n+k,a_n)-F_n}{k}\right)d\tau \\ &= a_n + \frac{1}{k}A^{-2}\big(e^{-kA}-I+kA\big)\big(F(t_n+k,a_n)-F(t_n,u_n)\big). \end{aligned}$$

The Second Order ETD Semi-discrete Scheme. Thus a second order exponential time differencing Runge-Kutta semi-discrete type scheme is given by

$$u_{n+1} = a_n + \frac{1}{k}A^{-2}\big(e^{-kA}-I+kA\big)\big(F(t_n+k,a_n)-F(t_n,u_n)\big) \tag{8}$$

where

$$a_n = e^{-kA}u_n - A^{-1}(e^{-kA}-I)F(t_n,u_n). \tag{9}$$

The computational challenge now is to efficiently compute terms like $-A^{-1}\left(e^{-kA}-I\right)$ and $\frac{1}{k}A^{-2}\left(e^{-kA}-I+kA\right)$. Kassam and Trefethen [10] and Du and Zhu [4] have developed a contour integration technique for this problem [4, 10], while Hochbruck and Osterman [5, 6] do not deal with the problem, leaving the computation of polynomial functions of matrix exponentials to standard software at the time of implementation. The approach we introduce in this work deals directly with the full discretization of the underlying matrix exponentials with an eye on efficiency. Tests show that the fully-discrete ETD-CN version

 ISBN: 978-84-614-6167-7

performs with significantly less CPU time than using a standard routine for the exponential of A.

Similar to (7), except now with $R_{1,1}(kA)$ instead of $R_{0,1}(kA)$ for e^{-kA} to achieve higher spatial accuracy, we compute that

$$\begin{aligned} -A^{-1}\big(e^{-kA}-I\big) &\approx -A^{-1}\big((2I-kA)\,(2I+kA)^{-1}-I\big) \\ &= 2k\,(2I+kA)^{-1} \\ &= kR_{0,1}\Big(\frac{1}{2}kA\Big), \end{aligned} \tag{10}$$

and

$$\begin{aligned} \frac{1}{k}A^{-2}\big(e^{-kA}-I+kA\big) &\approx \frac{1}{k}A^{-2}\big((2I-kA)\,(2I+kA)^{-1}-I+kA\big) \\ &= k\,(2I+kA)^{-1}. \end{aligned} \tag{11}$$

In the expression (9) for a_n we now use b_n in order to distinguish between the semi-discrete case (with e^{-kA}) and the fully discrete predictor stage where e^{-kA} is replaced by an appropriate Padé approximation. Next, we substitute (10, 11) into (8, 9), which gives the **ETD-CN scheme**:

$$v_{n+1} = b_n + \frac{1}{2}kR_{0,1}\left(\frac{1}{2}kA\right)\big[F(t_n+k,b_n)-F(t_n,v_n)\big], \tag{12}$$

$$b_n = R_{1,1}(kA)v_n + kR_{0,1}\left(\frac{1}{2}kA\right)F(t_n,v_n). \tag{13}$$

Theorem 1 *If the listed assumptions are satisfied and $F(t,u(t)) \in \mathcal{D}(A)$ as well as $F \in C^2([0,T];L^1)$, then for the numerical solution the following error bound holds*

$$\begin{aligned} \|u(t_n)-v_n\|_{\mathcal{X}} \leq\ & Ck^2 \max\Big(\sup_{0\le\tau\le T}\Big\|F'(\xi,u(\xi))\Big\|_{\mathcal{X}}, \sup_{0\le\tau\le T}\Big\|F^{(2)}(\tau,u(\tau))\Big\|_{\mathcal{X}}, \\ & \|u_0\|_{\mathcal{X}}, \|Au_0\|_{\mathcal{X}}\Big) + Ck^3\sum_{j=0}^{n-1}\|AF(t_j,u_j)\|_{\mathcal{X}} + CDk^2 \end{aligned}$$

uniformly on $0 \le t_n \le T$. The constant C depends on T, but is independent of n and k.

2 A Standard Example for Baseline Comparison

We consider the Brusselator in one spatial variable describing a chemical reaction with two components as given by the following system of PDE's:

$$\begin{aligned} \frac{\partial u}{\partial t} &= A + u^2v - (B+1)u + \alpha\frac{\partial^2 u}{\partial x^2} \\ \frac{\partial v}{\partial t} &= Bu - u^2v + \alpha\frac{\partial^2 v}{\partial x^2} \end{aligned}$$

 ISBN: 978-84-614-6167-7

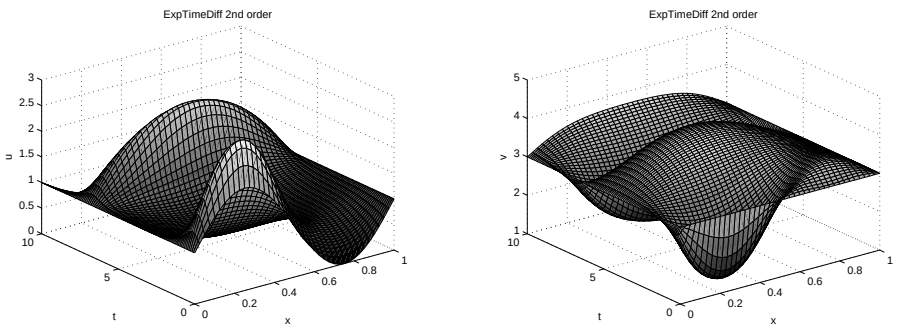
with $0 \leq x \leq 1, A = 1, B = 3, \alpha = 1/50$, and boundary conditions

$$\begin{aligned} u(0,t) &= u(1,t) = 1 \\ v(0,t) &= v(1,t) = 3 \end{aligned}$$

and initial conditions

$$\begin{aligned} u(x,0) &= 1 + \sin(2\pi x) \\ v(x,0) &= 3. \end{aligned}$$

We integrate the problem for $0 \leq t \leq 10$ in order to demonstrate the performance of the scheme. Figure 1 contains a solution profile.



(a) Solution profile u *(b) Solution profile v*

Figure 1: Example 2 with $h = k = \frac{1}{40}$.

Table 1 shows the observed ℓ_2–errors and rates for this problem. As expected the rate of convergence is 2.

Table 2 shows an observed timing comparison between ETD-CN and the standard Crank-Nicolson and BDF-2 schemes. Here, for the Crank-Nicolson and BDF-2 schemes we employ a modified Newton's method as in the previous paragraph. The data show how significant the improvement can be when using the ETD version of the Crank-Nicolson scheme as compared to the two standard schemes on this well-known test case. However, if the Crank-Nicolson and BDF(2) schemes are applied in a linearly implicit manner, they will use less CPU time but their convergence rates will seriously deteriorate. Our aim is to keep the convergence order of all the test examples the same (second order).

3 Conclusion

We derive a new fully discrete Exponential Time Differencing Runge-Kutta method followed

 ISBN: 978-84-614-6167-7

Table 1: Example 2 Numerical errors (ℓ_2) and rates using ETD-CN for the Brusselator equation.

h	$k = h$	Error ETD-CN	Rates
0.1	0.1	3.0×10^{-3}	
0.05	0.05	7.7718×10^{-4}	1.9499
0.025	0.025	1.9751×10^{-4}	1.9763
0.0125	0.0125	4.9833×10^{-5}	1.9867
0.00625	0.00625	1.2481×10^{-5}	1.9974
0.003125	0.003125	3.0816×10^{-6}	2.0179

h	$k = h$	Error ETD-CN	CPU	Error CN	CPU	Error BDF-2	CPU
0.1	0.1	3.0×10^{-3}	0.016	3.1×10^{-3}	0.078	3.3×10^{-3}	0.062
0.05	0.05	7.77×10^{-4}	0.031	7.96×10^{-4}	0.297	1.0×10^{-3}	0.313
0.025	0.025	1.98×10^{-4}	0.078	1.96×10^{-4}	2.156	2.53×10^{-4}	2.204
0.0125	0.0125	4.98×10^{-5}	0.531	4.91×10^{-5}	22.078	6.34×10^{-5}	22.437

Table 2: Numerical errors (ℓ_2) and CPU-time (sec) using ETD-CN, Crank-Nicolson, and the second order Backward Differentiation Formula (BDF-2) for Brusselator equation.

by a (1,1)–diagonal Padé scheme to solve nonlinear parabolic partial differential equations. Convergence of the new scheme is second order in the semilinear case. A comparison the new scheme ETD-CN to other standard second order codes like the Crank-Nicolson and BDF-2 schemes shows the effectiveness of the new algorithm. ETD-CN is faster due to the fact that it does not have to solve nonlinear systems in each time step, yet it remains second order accurate. In applications one often encounters nonsmooth initial data, which in most well-known second order codes inflicts oscillations if not treated carefully. Between two and four steps of initial damping suffices to restore the good numerical properties of ETD-CN. The new scheme is comparable to other well known second order schemes in accuracy, yet is more effective with regard to CPU time.

Acknowledgements

The second author was supported in part by the U.S. National Security Agency under Grant Agreement Number H98230-09-1-0002. The United States Government is authorized to reproduce and distribute reprints notwithstanding any copyright notation herein.

References

[1] G. Barles and H. M. Soner, *Option Pricing with Transaction Costs and a Nonlinear Black-Scholes Equation,* Finance Stochast. **2** (1998), 369–397.

[2] S.M. Cox and P.C. Matthews, *Exponential Time Differencing for Stiff Systems,* J. Computational Physics **176** (2002) 430–455.

[3] B. Düring, M. Fournié, and A. Jüngel, *High order compact finite difference schemes for a nonlinear Black-Scholes equation,* Intern J. Theor. Appl. Finance **6** (2003), 767–789.

[4] Q. Du and W. Zhu, *Analysis and Applications of the Exponential Time Differencing Schemes and Their Contour Integration Modifications,* BIT Numer. Math. **45** (2005), 307–328.

[5] M. Hochbruck, A. Ostermann, *Explicit exponential Runge-Kutta methods for semilinear parabolic problems,* SIAM J. Numer. Anal. **43** (2005) 1069-1090.

[6] M. Hochbruck and A. Ostermann, *Exponential Runge-Kutta Methods for Parbolic Problems,* Appl. Numer. Math. **53** (2005) 323-339.

[7] T. Hoggard, A. E. Whalley, and P. Wilmott, *Hedging option portfolios in the presence of transaction costs,* Adv. Futures Opt. Res., **7** (1994) 21–35.

[8] W. Hundsdorfer and J. Verwer, *Numerical Solution of Time-Dependent Advection-Diffusion-Reaction Equations,* Springer-Verlag, Berlin. 2003.

[9] H. Imai, N. Ishimura, I. Mottate and M. Nakamura, *On the Hoggard-Whalley-Wilmott equation for the pricing of options with transaction costs,* Asia Pacific Financial Markets, **13** (2006) 315–326.

[10] A.K. Kassam and Ll. N. Trefethen, *Fourth-Order Time Stepping for Stiff PDEs,* SIAM J. Sci. Comput. **26** (2005) 1214–1233.

[11] B. Kleefeld, A.Q.M. Khaliq and B.A. Wade, *An ETD Crank-Nicolson Method for Reaction-Diffusion Systems* Num. Meth. PDE, to appear, 2011.

[12] R.H. Martin, *Nonlinear operators and differential equations in Banach spaces,* Wiley, New York (1976).

[13] B. Minchev, *Exponential Integrators for Semilinear Problems,* Ph.D. Thesis, Dept. of Informatics, Univ. of Bergen, 2004.

 ISBN: 978-84-614-6167-7

[14] Q. Nie, Y.-T. Zhang and R. Zhao, *Efficient semi-implicit schemes for stiff systems,* J. Comput. Phys. **214** (2006) 521-537.

[15] J. Vigo-Aguiar and J. M. Ferrándiz, *A General Procedure for the Adaption of Multistep Algorithms to the Integration of Oscillatory Problems,* SIAM J. Numer. Anal., **35(4)** (1998) pp. 1684-1708.

[16] B.A. Wade and A.Q.M. Khaliq, *On Smoothing of the Crank-Nicolson Scheme for Nonhomogeneous Parabolic Problems*, J. of Comput. Meth. in Sci & Eng. **1** (2001) 107–124.

[17] B.A. Wade, A.Q.M. Khaliq, M. Siddique, and M. Yousuf, *Smoothing with Positiviy-Preserving Padé Schemes for Parabolic Problems with Nonsmooth Data,* Numer. Meth. PDE **21(3)**(2005) 553–573.

[18] B.A. Wade, A.Q.M. Khaliq, M. Yousuf, J. Vigo-Aguiar, and R. Deininger, *On Smoothing of the Crank-Nicolson Scheme and Higher Order Schemes for Pricing Barrier Options*, Journal of Computational and Applied Mathematics (JCAM) **204(1)** (2007) 144–158.

 ISBN: 978-84-614-6167-7

Proceedings of the 11th International Conference on Computational and Mathematical Methods in Science and Engineering, CMMSE 2011 26–30 June 2011.

Two Dimensional Node Optimization in Piecewise High Dimensional Model Representation

Evrim Korkmaz Özay[1] **and Metin Demiralp**[1]

[1] *Computational Science and Engineering Program, Informatics Institute, İstanbul Technical University*

emails: evrim.korkmaz@be.itu.edu.tr, metin.demiralp@be.itu.edu.tr

Abstract

Multivariate functions play important roles in almost all branches of science and engineering. In this sense the evaluations on multivariate functions with less effort become important. This is the reason why High Dimensional Model Representation (HDMR), which is a decomposition of a multivariate function to components with ascending multivariance, has been widely used in last two decades with increasing tendency from year to year. HDMR applications are based on the truncations, which are kept basically at most bivariance level to avoid computational complexity as long as the approximation quality becomes satisfactory, and various HDMR varieties have been developed to this end. One of these works is Piecewise High Dimensional Model Representation (PHDMR), which is based on splitting HDMR geometry into appropiate segments in accordance with the structure of the target multivariate function. In this manner PHDMR increases approximation quality of plain HDMR and decreases computational complexity since the growth in constant term dominancy results in no need to higher variate terms. PHDMR uses nodes on coordinates such that their locations are optimised to get maximum constancy in the components of the representation. In this work we present theoretical backround and certain properties of PHDMR with some illustrative examples.

Key words: Multivariate Analysis, Node Optimization, High Dimensional Model Representation, Approximation

1 Introduction

The difficulties in dealing with multivariate functions, in both mathematical and computational sense, urged the scientists to propose High Dimensional Model Representation (HDMR) and its varieties for applying to problems in statistics, chemistry, and, various similar areas. HDMR is a divide-and-conquer algorithm, which allows us to additively represent multivariate functions in terms of less-variate functions ordered in

ascending multivariance. The plain HDMR expansion for a multivariate function is given as below

$$f(x_1,...,x_N) = f_0 + \sum_{i_1=1}^{N} f_{i_1}(x_{i_1}) + \sum_{\substack{i_1,i_2=1 \\ i_1<i_2}}^{N} f_{i_1,i_2}(x_{i_1},x_{i_2}) + \cdots + f_{12\cdots N}(x_1,...,x_N) \quad (1)$$

HDMR expansion of a function with N variables have 2^N components at the right hand side, a constant term f_0, N univariate terms $f_i(x_i), (1 \le i \le N)$, $N(N-1)/2$ bivariate terms $f_{i_1,i_2}(x_{i_1},x_{i_2})$, $(1 \le i_1 \le i_2)$ and so on. The determinations of the components in this expansion is based on the multiplication with a weight function and then multivariate integration of both sides in (1). To get the constant term, integration is performed on all independent variables. This gives a single equation to determine 2^N components and urges us to impose some conditions on the integrals of the HDMR components. If we assume that anyone of the HDMR components except the constant one vanishes under the integration with respect to its any argument. By doing so the constant component can be determined as

$$f_0 = \int_{a_1}^{b_1} dx_1 \cdots \int_{a_N}^{b_N} dx_N W(x_1,...,x_N) f(x_1,...,x_N) \quad (2)$$

where the weight function denoted by W should be of a product type each factor of which is a univariate function depending on a different independent variable. For the univariate component determination, the multiplication with the weight function discards the univariate weight factor whose argument is same as the argument of the component to be determined and the integration of both sides is performed over all independent variables except the one which appears as the argument of the component to be determined. The same vanishing integral condition enables us to write

$$\begin{aligned} f_i(x_i) &= \int_{a_1}^{b_1} dx_1 \cdots \int_{a_{i-1}}^{b_{i-1}} dx_{i-1} \int_{a_{i+1}}^{b_{i+1}} dx_{i+1} \cdots \int_{a_N}^{b_N} dx_N W^{(1)}(x_1,...,x_{i-1},x_{i+1},...,x_N) \\ &\quad \times f(x_1,...,x_N) - f_0, \qquad 1 \le i \le N \end{aligned} \quad (3)$$

where superscripted W contains all univariate factors of W except the one depending on the argument of the univariate component to be determined. The higher variate terms can also be determined in a similar way such that the both sides of (1) is integrated over the all independent variables except the ones appearing as the arguments of the kth variate component after multiplying both sides with the weight which contains all the univariate factors of W except the ones depending on the arguments on the component to be determined, and, the vanishing integration conditions are used. This procedure allows us to determine all HDMR components uniquely. It is possible to define some projection operators to get more concise formulation even though we do not get into the further details of derivations.

 ISBN: 978-84-614-6167-7

As can be immediately noticed the most important issue is to impose the "vanishing integral conditions" to get uniqueness. They were first proposed by Sobol[1]. In its original form the integration was on the unit hypercube whose one corner is located at the origin and axes are all in positive directions and also unit weight function was used. Those conditions were extended to all types of rectangular hyperprismatic geometries and to the utilization of product type non-unit weight functions whose factors are univariate functions eah of which depends on a different independent variable, by Rabitz group[2, 3, 4]. Denoting the univariate weight functions by $W_j(x_j)$s we can write the "Vanishing Integral Conditions" as follows

$$\int_{a_j}^{b_j} dx_j W_j(x_j) f i_1 ... i_k(x_{i_1}, ..., x_{i_k}) \quad = 0, \quad x_j \in \{x_{i_1}, x_{i_2}, ..., x_{i_k}\}, \quad 1 \leq j, k \leq N \quad (4)$$

Then the overall weight function of HDMR can be written as follows

$$W(x_1, ..., x_N) \equiv \prod_{i=1}^{N} W_i(x_i), \quad 1 \leq i \leq N \quad (5)$$

where $W_i(x_i)$ is assumed to satisfy the following normalization condition

$$\int_{a_i}^{b_i} dx_i W_i(x_i) = 1, \quad 1 \leq i \leq N \quad (6)$$

which facilitates the formulation by getting rid of some denominators in the resulted expressions.

Another important property is the orthogonality of HDMR terms. Orthogonality was first proven by Demiralp [5] and used by both Demiralp and Rabitz frequently. Orthogonality of HDMR terms can be shown by using the following inner product on related geometry (where $\mathcal{H}$ stands for the Hilbert space of the multivariate functions square integrable over the hyperprism on which the following integrals are performed)

$$(f, g) \quad \equiv \quad \int_{a_1}^{b_1} dx_1 \cdots \int_{a_N}^{b_N} dx_N W(x_1, \ldots, x_N) f(x_1, ..., x_N) g(x_1, ..., x_N), \quad f, g \in \mathcal{H} \quad (7)$$

which permits us to get the benefits of Hilbert space tools utilization. This inner product enables us to write

$$(f_{i_1 i_2 ... i_k}, f_{j_1 j_2 ... j_l}) = 0, \quad (k \neq l) \vee (i_1 \neq j_1) \vee \cdots \vee (i_k \neq j_l) \quad (8)$$

If HDMR expansion is truncated at some level of multivariance then the HDMR approximants are obtained. Only the zeroth and first order HDMR approximants are explicitly given below

$$\begin{aligned} s_0(x_1, \ldots, x_N) &= f_0 \\ s_1(x_1, \ldots, x_N) &= s_0(x_1, \ldots, x_N) + \sum_{i_1=1}^{N} f_{i_1}(x_{i_1}) \end{aligned} \quad (9)$$

 ISBN: 978-84-614-6167-7

Higher level approximants show similar structure. The following parameters which are called basically measurers were defined by Demiralp to measure how the approximants represent a given multivariate function.

$$\begin{aligned} \sigma_0 &\equiv \frac{1}{\|f\|^2}\|f_0\|^2 \\ \sigma_1 &\equiv \frac{1}{\|f\|^2}\sum_{i=1}^{N}\|f_i\|^2 + \sigma_0 \\ &\vdots \\ \sigma_N &\equiv \frac{1}{\|f\|^2}\|f_{12\ldots N}\|^2 + \sigma_{N-1} \end{aligned} \tag{10}$$

where σ_0 is called Constancy Measurer while σ_1 is named Univariance Measurer. The general term σ_k is called "kth Order Additivity Measurer". Additivity Measurers form a well-ordered sequence and the sufficiently closedness to 1 for any additivity measurer means that the relevant approximant is sufficient to represent the original function satisfactorily. In other words, there is no need to add higher variable terms of HDMR to construct an approximation. In practice, the first few terms (in fact at most bivariate terms) are taken into consideration from the HDMR expansion, to avoid computational complexity. Obviously the first HDMR component can not represent the function efficienlty unless the target function is almost constant. Constant component evaluation is the easiest part of HDMR and urges us to find an answer to the question " Is it possible that a given multivariate function is represented by both using the constant term and obtaining a desired approximation quality?". In this study we show the fact that the answer to this question is yes and present the way to get the fruits of this possibility. To this end we partition HDMR geometry into optimized subgeometries and we evaluate the constant term in each subgeometry. Next section includes the presentation of the mathematical backround of the optimization procedure. The third section involves certain illustrative numerical implementations while the fourth (and last) section presents the concluding remarks.

2 Two Dimensional Node Optimization in Piecewise High Dimensional Model Representation

Dividing geometry is not a new scheme in HDMR history [6, 7, 8]. Previous works were based on HDMR geometries which are divided into huge number of equal subgeometries whose sizes are sufficiently small [9, 10].

Piecewise HDMR (PHDMR), being a new variant of HDMR, is based on the utilization of constant HDMR approximant in each piece of geometry with a possibly different constant value. The determination of these constants turns into an optimization problem, solution of which gives the optimal nodes where the HDMR geometry is partitioned to subgeometries. Using optimal nodes we can evaluate constant terms in

all subgeometries. In this way we avoid from high complexity of evaluating univariate and higher variate terms. In this section we will focus on the HDMR on two independent variables for simplicity in the presentation. However, it is almost straightforward to extend what we obtain in this limited, lowest level multivariate, case to the higher multivariate cases. To construct the problem for finding the optimal nodes, we assume that there are unknown nodes on each interval defining the two dimensional space, that is, plane. In other words we split HDMR geometry of $f(x_1, x_2)$ into pieces and then interpolate that function whose values are known only for the selected nodes. The nodes on the coordinate x_1 are symbolized by $c_1, c_2, ..., c_m$ while the ones on the coordinate x_2 are denoted by $d_1, d_2, ..., d_n$. The following norm is needed to be minimized to find the optimal nodes where $a_1 \leq x_1 \leq b_1$ and $a_2 \leq x_2 \leq b_2$.

$$\Upsilon = \int_{a_1}^{b_1} dx_1 w_1(x_1) \int_{a_2}^{b_2} dx_2 w_2(x_2) \left[f(x_1, x_2) - f_0 \right]^2 \tag{11}$$

Here we take m and n number of points on the first and second coordinates respectively. Then we differentiate Υ with respect to each node and set equal to zero to get related equations to 5 determine c and d parameters.

$$\frac{\partial \Upsilon}{\partial c_i} = 0 \qquad and \qquad \frac{\partial \Upsilon}{\partial d_j} = 0, \quad i = 1...m, \quad j = 1...n \tag{12}$$

Nodes define planar subregions in the HDMR geometry. The constant HDMR component on the kth planar subregion $[\, c_{i-1}, c_{i+1} \,] \times [\, d_{j-1}, d_{j+1} \,], \quad 1 \leq i \leq m, \quad 1 \leq j \leq n$ can be evaluated as below

$$f_0^k = \int_{c_{i-1}}^{c_{i+1}} dx_1 \int_{d_{j-1}}^{d_{j+1}} dx_2 \frac{w_1(x_1)}{\int_{c_{i-1}}^{c_{i+1}} dx_1 w_1(x_1)} \frac{w_2(x_2)}{\int_{d_{j-1}}^{d_{j+1}} dx_2 w_2(x_2)} f(x_1, x_2) \tag{13}$$

where k stands to give the location in an ordering of all possible subregions and therefore it depends on i and j values. To obtain optimized nodes, (12) is used in (11) via Leibniz's differentiation rule of integrals. After some intermediate manipulations and algebraic operations we reach two general expressions which contain the local constant terms in the vicinity of the coordinates of the optimized nodes.

$$\begin{aligned} &\sum_{k=1}^{n+1} \int_{d_{k-1}}^{d_k} dx_2 w_2(x_2) f(c_i, x_2) \left[-2 f_0^{i+(k-1)(n+1)} + 2 f_0^{i+(k-1)(n+1)+1} \right] \\ &= \sum_{k=1}^{n+1} \left[-\left(f_0^{(i+(k-1)(n+1)} \right)^2 + \left(f_0^{i+(k-1)(n+1)+1)} \right)^2 \right] \end{aligned} \tag{14}$$

 ISBN: 978-84-614-6167-7

$$\sum_{k=1}^{m+1} \int_{c_{k-1}}^{c_k} dx_1 w_1(x_1) f(x_1, d_i) \left[-2f_0^{(i-1)(m+1)+k} + 2f_0^{(i-1)(m+1)+(k+m+1)} \right]$$

$$= \sum_{k=1}^{m+1} \left[-\left(f_0^{(i-1)(m+1)+k} \right)^2 + \left(f_0^{(i-1)(m+1)+(k+m+1)} \right)^2 \right] \quad (15)$$

It is possible to obtain all c_i and d_i values by solving the related equations with different ways like iterative solution constructions or Gröbner basis set utilization. In this work we use fixed point iteration, firstly we make an initial guess for the c_is and d_js in the intervals $[\, a_1, b_1 \,]$ and $[\, a_2, b_2 \,]$ respectively. To be able to obtain the optimized nodes within the acceptable tolerance, we need a criterion to stop the iteration process. Thus, the optimized separation of HDMR geometry has been realized.

3 NUMERICAL IMPLEMENTATION

We can implement the optimization procedure for different kinds of functions. Here we chose exponential and sine functions for testing and showing the efficiency of PHDMR. We use (13) and (14) to get appropriate nodes and then evaluate the constant terms on the planar subregions determined by the optimal nodes. Procedure code written in MUPAD within 20-digits precision [11].

The first test function is $f\,(x_1, x_2) = e^{x_1+x_2}$ and in the following first figure the test function and the constant term truncation results of PHDMR on each planar subregion are shown. Node numbers were chosen as $m = n = 3$, this means that there are 16 planar subregiond in Figure 1. According to the results obtained from this example, it is possible to say that nodes tend to locate towards the high curvilinearity surface of the given function. This is an expected behaviour for the optimal nodes, because when curvilinearity increases at a point it is better to split the area into two parts for calculation of the constant term, since the constant term of HDMR can be considered as the mean of the function at the given geometry. Figure 1 also shows that 16 planar subregions are enough for a good representation of the exponential function because all the planar subregions overlap with the function surface.

Second test function is a sine function, $f\,(x_1, x_2) = \sin(x_1 + x_2)$ which forms a difficult example because of the periodic structure and the sharp curvilinearity. At first function tested for $m = n = 3$, it can be seen that the approximation quality is not good enough on figure 2, because overlapping of the function surface and the planar subregion is quite weak. In these situations there may be different reasons of the weakness and different solutions can be sought to improve the approximation quality. For example, we give the starting values for nodes at the first iteration on two coordinate lines and these values affect the locations of optimized nodes at the last iteration because of stopping the application of the criteria. Another reason may be the insufficient node number. If so, increasing the number of nodes profits for a good approximation with constant term truncations. This idea can be supported by the observations on figure 3 where, this time, we use the same test function and more

 ISBN: 978-84-614-6167-7

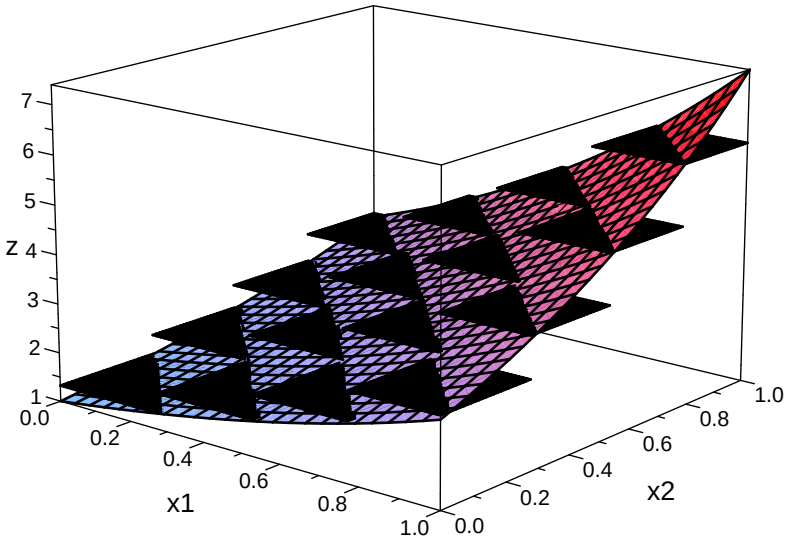


Figure 1: $f(x_1, x_2) = e^{x_1+x_2}$ vs. Constant Terms of PHDMR on 16 subplanes

nodes ($m = n = 5$). Compatibility of planar subregions for surface of sine function can be easily seen on figure 3.

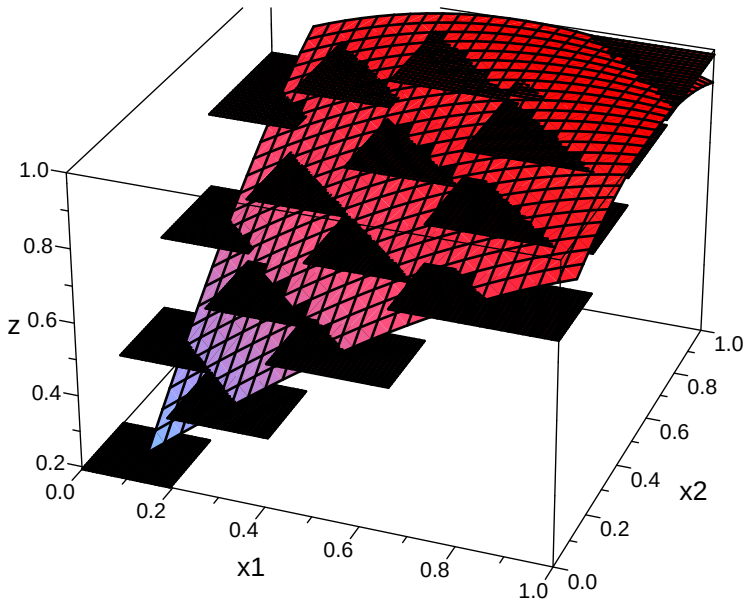


Figure 2: $f(x_1, x_2) = sin(x_1 + x_2)$ vs. Constant Terms of PHDMR on 16 sub-planes

 ISBN: 978-84-614-6167-7

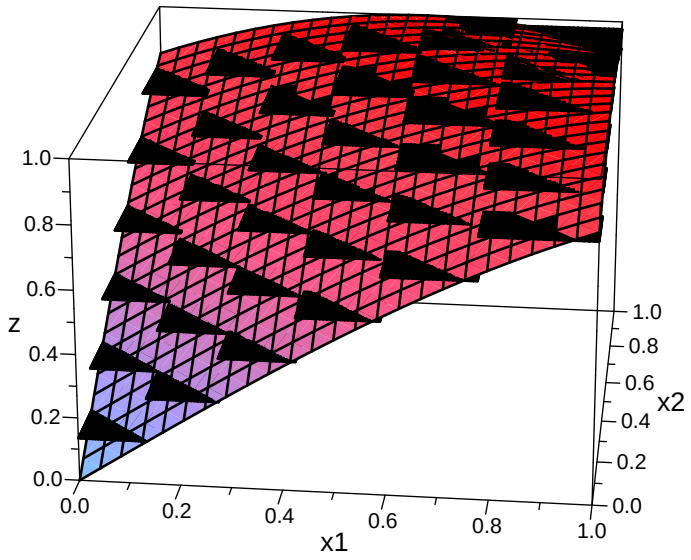


Figure 3: $f(x_1, x_2) = sin(x_1 + x_2)$ vs. Constant Terms of PHDMR on 36 sub-planes

4 Conclusions

In this work we have tried to increase the quality of approximation to get the sufficiency of HDMR's constant term, which enables us not to need the evaluation of higher variate components of HDMR. For this purpose we parttion HDMR geometry amongst optimal nodes which are found by fixed point iteration. For this work we confine ourselves to two dimensional space for presentation simplicity in theory and for simply illustrating implementation results on figures and we work with planar segments. We have dealt with only with the constancy whereas univariance could also be tackled with by accepting a little bit higher computational complexity. The confinement ourselves to the two dimensional HDMR can be relaxed and the number of the variables can be increased. The basic philosophy remains same also for those cases. However, the calculations becomes more comprehensive as the number of independent variables increases.

According to results increasing the number of nodes provide better approximation quality and represent more sensitive results in the areas including high curvature of the given function. The main reason for this is due to the fact we have observed in our illustrative implementations. The optimal nodes more intensely locate in higher curvilinearity parts of the target function. This behaviour of the method can be mathematically proven although we have not attempted to do so. Nevertheless we can say that this is very important because of its implications that the target function domain must be more intensely divided on the areas where curvilinearity is high.

Acknowledgements

First author is grateful to İstanbul Technical University (İTÜ) of Turkey for its support. The second author sends warmest thanks to Turkish Academy of Sciences for its support and motivation.

References

[1] I.M. Sobol, *Sensitivity estimates for nonlinear mathematical models*, Mathematical Modelling and Computational Experiments (MMCE). **1** (1993) 407–414.

[2] Ö.F. Alış and H. Rabitz, *Additive and Multiplicate High Dimensional Representation General Foundations of High Dimensional Model Representations*, Journal of Mathematical Chemistry. **25** (1999) 197–233.

[3] G. Li, C. Rosenthal and H. Rabitz, *High Dimensional Model Representations*, J. Phys. Chem. **105** (2001) 7765–7777.

[4] Ö.F. Alış and H. Rabitz, *Efficient Implementation of High Dimensional Model Representations*, J. Math. Chem. **29** (2001) 127–142.

[5] M. Demiralp, *High Dimensional Model Representation and its application varieties*, Tools for Mathematical Methods, Mathematical Research. **9** (2003) 146–159.

[6] N. A. Baykara and M. Demiralp, *Hyperspherical or Hyperellipsoidal Coordinates in the Evalua- tion of High Dimensional Model Representation Ap- proximants* ,Mathematical Research. **9** (2003) 49–62.

[7] M. A. Tunga and M. Demiralp, *A Factorized High Dimensional Model Representation on the Nodes of a Finite Hyperprismatic Regular Grid*, Applied Mathematics and Computation. **164** (2005) 865–883.

[8] M. Demiralp, *Illustrative Implementations to Show How Logarithm Based High Dimensional Model Representation Works for Various Function Structures.* WSEAS Transaction on Computers. **5** (2006) 1333–1338.

[9] E. Korkmaz, M. Demiralp, *Small Scale High Dimensional Model Representation.* WSEAS Transaction on Multivariate Analysis and its Application in Science and Engineering. **1** (2008) 232–236.

[10] E. Korkmaz Özay, M. Demiralp, *Small Scale Enhanced Multivariance Product Representation.* Proceedings of the International Conference on Applied Computer Science. **1** (2010) 350–356.

[11] W. Oevel, F. Postel, S. Wehmeier and J. Gerhard , *The Mupad Tutorial.* Springer. (2000)

 ISBN: 978-84-614-6167-7

Proceedings of the 11th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2011
26–30 June 2011.

An $O(N^3)$ implementation of Hedin's GW approximation

Peter Koval[1], Dietrich Foerster[2] and Daniel Sánchez-Portal[1]

[1] *Centro de Física de Materiales CFM-MPC, Centro Mixto CSIC-UPV/EHU and DIPC, E-20018 San Sebastián, Spain*

[2] *CPMOH/LOMA, University of Bordeaux, France*

emails: koval.peter@gmail.com, d.foerster@cpmoh.u-bordeaux1.fr, sqbsapod@sq.ehu.es

Abstract

Organic electronics is a rapidly developing technology. Typically, the molecules involved in organic electronics are made up of hundreds of atoms, prohibiting a theoretical description by wavefunction-based ab-initio methods. Density-functional and Green's function type of methods scale less steeply with the number of atoms. Therefore, they provide a suitable framework for the theory of such large systems.

In this contribution, we describe an implementation, for molecules, of Hedin's GW approximation. The latter is the lowest order solution of a set of coupled integral equations for electronic Green's and vertex functions that was found by Lars Hedin half a century ago.

Our implementation of Hedin's GW approximation has two distinctive features: i) it uses sets of localized functions to describe the spatial dependence of correlation functions, and ii) it uses spectral functions to treat their frequency dependence. Using these features, we were able to achieve a favorable computational complexity of this approximation. In our implementation, the number of operations grows as N^3 with the number of atoms N.

Key words: Hedin's GW approximation, basis of dominant products, large molecules.

1 Introduction

The promising field of organic electronics deals with large molecules of several tens or even hundreds of atoms [1]. For instance, fullerene C_{60} is a frequently used subunit in organic electronics and it alone consist of 60 atoms (see figure 1).

 ISBN: 978-84-614-6167-7

Each individual molecule may be used in a device in many different ways and there is an astronomically large number of different promising molecules. As in many cases there is a limited knowledge of the relevant physical parameters, and it might be also interesting to explore the potential of candidate molecules theoretically, before these molecules has been actually synthesised.

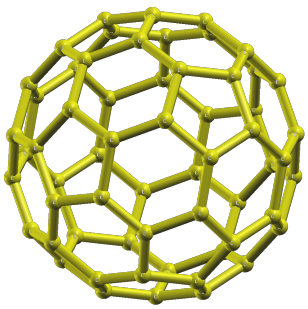


Figure 1: Ball and stick model of fullerene C_{60} produced with XCrysDen package [2].

The geometry of large organic molecules can be reliably predicted by density-functional theory (DFT)[3]. However, the properties of their excited states such as the energy of the highest occupied (HOMO) and lowest unoccupied molecular orbitals (LUMO), corresponding to adding and subtracting one electron from the system respectively, require a description of electronic correlations better than that provided by current functionals of DFT and its time-dependent counterpart, TDDFT.

Such effects can be efficiently incorporated with the help of Hedin's method that is based on Green's function. Hedin's GW approximation for one-electron Green's function is computationally cheaper than wavefunction-based methods, although it remains computationally more expensive than DFT and TDDFT within linear response.

The goal of our work is to develop a practical algorithm for Hedin's GW approximation which is suitable for large organic molecules, allowing to access the excited states of such molecules.

2 Theoretical framework for Hedin's GW approximation

Electronic Green's function (propagators) are useful in condensed matter physics because many simple observables can be computed in terms of them. At the same time, such Green's functions remain simpler than many-body wavefunction.

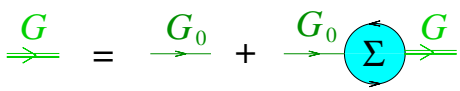


Figure 2: Feynman diagram of Dyson equation (1).

Hedin's GW is a useful approximation for the so-called self-energy $\Sigma(\boldsymbol{r}, \boldsymbol{r}, \omega)$ that enters Dyson's equation for an interacting electronic propagator $G(\boldsymbol{r}, \boldsymbol{r}', \omega)$

$$G^{-1}(\boldsymbol{r}, \boldsymbol{r}', \omega) = G_0^{-1}(\boldsymbol{r}, \boldsymbol{r}', \omega) - \Sigma(\boldsymbol{r}, \boldsymbol{r}, \omega). \tag{1}$$

Here, the inversions must be understood in operator sense $\int G^{-1}(\boldsymbol{r}, \boldsymbol{r}'', \omega) G(\boldsymbol{r}'', \boldsymbol{r}', \omega) d\boldsymbol{r}'' = \delta(\boldsymbol{r} - \boldsymbol{r}')$ and $G_0(\boldsymbol{r}, \boldsymbol{r}', \omega)$ stands for Green's function where electron-electron interactions

 ISBN: 978-84-614-6167-7

have been switched off. It is obtained from an effective one-particle Hamiltonian

$$(\omega - H(\boldsymbol{r}))G_0(\boldsymbol{r},\boldsymbol{r}',\omega) = \delta(\boldsymbol{r}-\boldsymbol{r}'). \tag{2}$$

In this work we use a Kohn-Sham Hamiltonian [3], although Hartree-Fock Hamiltonian also proves to be useful at this point [4]. Hedin's GW approximation for the self-energy $\Sigma(\boldsymbol{r},\boldsymbol{r},\omega)$ reads

$$\Sigma(\boldsymbol{r},\boldsymbol{r}',t) = \mathrm{i}G_0(\boldsymbol{r},\boldsymbol{r}',t)W_0(\boldsymbol{r},\boldsymbol{r}',t). \tag{3}$$

It involves the non interacting electronic Green's function $G_0(\boldsymbol{r},\boldsymbol{r}',t)$ and a screened Coulomb interaction $W_0(\boldsymbol{r},\boldsymbol{r}',t)$. This approximation is a solution of a truncated version Hedin's equations [5, 6]. The name of this approximation is taken from the simple form of the electronic self-energy $\Sigma = \mathrm{i}GW$.

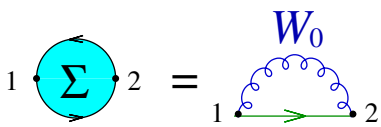


Figure 3: Feynman diagram of self-energy (3).

The screened Coulomb interaction W_0 can be easily calculated in frequency domain using the so-called RPA approximation [7]

$$W_0(\boldsymbol{r},\boldsymbol{r}',\omega) = \left[\delta(\boldsymbol{r}-\boldsymbol{r}''') - v(\boldsymbol{r},\boldsymbol{r}'')\chi_0(\boldsymbol{r}'',\boldsymbol{r}''',\omega)\right]^{-1} v(\boldsymbol{r}''',\boldsymbol{r}'), \tag{4}$$

where $v(\boldsymbol{r},\boldsymbol{r}') \equiv |\boldsymbol{r}-\boldsymbol{r}'|^{-1}$ is the bare Coulomb interaction. Here and in the following we assume integration over repeated spatial coordinates ($\boldsymbol{r}''$ and $\boldsymbol{r}'''$ in equation (4)) on the right hand side of an equation if they do not appear on its left hand side. The screened interaction (4) is the sum of the bare Coulomb interaction created by a point charge at $\boldsymbol{r}'$, plus a correction due to the redistribution of charge induced in response to the total field [7, 6]. The non-interacting response function $\chi_0(\boldsymbol{r},\boldsymbol{r}',t)$ is related to the non-interacting Green's function

$$\mathrm{i}\chi_0(\boldsymbol{r},\boldsymbol{r}',t) = 2G_0(\boldsymbol{r},\boldsymbol{r}',t)G_0(\boldsymbol{r}',\boldsymbol{r},-t), \tag{5}$$

where a factor 2 arises because of the summation over spin variable.

As we mentioned already, we construct the non-interacting Green's function using an effective Kohn-Sham Hamiltonian [3]

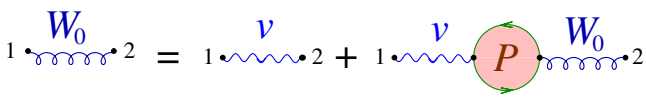


Figure 4: Feynman diagram of screened Coulomb interaction (4).

$$H_{\mathrm{KS}} = -\frac{1}{2}\nabla^2 + V_{\mathrm{KS}}, \tag{6}$$

$$V_{\mathrm{KS}} = V_{\mathrm{ext}} + V_{\mathrm{Hartree}} + V_{\mathrm{xc}}, \text{ where } V_{\mathrm{xc}}(\boldsymbol{r}) = \frac{\delta E_{\mathrm{xc}}}{\delta n(\boldsymbol{r})}.$$

E_{xc} is a functional of the electronic density that includes the effects of exchange and correlation in an effective way. Its functional derivative $V_{\mathrm{xc}}(\boldsymbol{r})$ is the so-called exchange-correlation potential and it must be subtracted from $\Sigma(\boldsymbol{r},\boldsymbol{r}',t)$ to avoid including the

 ISBN: 978-84-614-6167-7

exchange-correlation interaction twice in equation (3). This is accomplished with the substitution

$$\Sigma(\boldsymbol{r},\boldsymbol{r}',t) \to \Sigma(\boldsymbol{r},\boldsymbol{r}',t) - \delta(\boldsymbol{r}-\boldsymbol{r}')\delta(t)V_{\rm xc}(\boldsymbol{r})$$

in Dyson's equation (1).

3 A basis set of localized functions

Having the equations (1,3,4,5) at hand we introduce a basis set of localized functions and rewrite the system of equations in the basis. We start with linear combinations of atom orbitals (LCAO) to represent the non-interacting Green's function $G_0(\boldsymbol{r},\boldsymbol{r}',t)$

$$G_0(\boldsymbol{r},\boldsymbol{r}',t) = \sum_{ab} G^0_{ab}(t) f^a(\boldsymbol{r}) f^b(\boldsymbol{r}'), \tag{7}$$

where $f^a(\boldsymbol{r})$ are atom centered orbitals. The frequency (and time) dependence has been factorized in the last equation. The treatment of the frequency (and time) dependence by spectral functions will be explained in section 4. Inserting equation (7) into the equation (5), we obtain

$$\mathrm{i}\chi_0(\boldsymbol{r},\boldsymbol{r}',t) = 2\sum_{abcd} G^0_{ab}(t)G^0_{cd}(-t)\, f^a(\boldsymbol{r})f^d(\boldsymbol{r})\, f^b(\boldsymbol{r}')f^c(\boldsymbol{r}'). \tag{8}$$

Products of localized orbitals such as $f^a(\boldsymbol{r})f^d(\boldsymbol{r})$ appear in the last equation. Although a product of localized orbitals is also a localized function, such products do not form a suitable basis because they contain many collinear functions. Several methods have been proposed to construct more efficient basis to span the products of localized orbitals [4, 9, 10]. Here we use a basis of dominant products [11] that is constructed individually for each atom pair. The dominant products are identified as certain linear combinations of the original orbital products and they are free of any collinearity within a given atom pair (with respect to a given metric, here we have used the Coulomb metric). Moreover, the original orbital products can be expressed as linear combinations of dominant products

$$f^a(\boldsymbol{r})f^b(\boldsymbol{r}) = V^{ab}_\mu F^\mu(\boldsymbol{r}). \tag{9}$$

The three-index coefficient V^{ab}_μ will be referred to as the *product vertex*. The product vertex is local or sparse by construction and indeed the *locality* of our construction is its main characteristic.

Considering Dyson's equation (1), we arrive at its tensor counterpart

$$G_{ab}(\omega) = G^0_{ab}(\omega) + G_{aa'}(\omega)\Sigma^{a'b'}(\omega)G^0_{b'b}(\omega), \tag{10}$$

 ISBN: 978-84-614-6167-7

where matrix elements of the self-energy $\Sigma^{ab}(\omega)$ must be used

$$\Sigma^{ab}(\omega) = \iint f^a(\boldsymbol{r})\Sigma(\boldsymbol{r},\boldsymbol{r}',\omega)f^b(\boldsymbol{r}')\,d^3r d^3r'. \tag{11}$$

Calculating the matrix elements of the self-energy by equation (3) and using (7) for the non interacting Green's function, we arrive at

$$\Sigma^{ab}(\omega) = \mathrm{i}\sum_{a'b'} G^0_{a'b'}(t)\int f^a(\boldsymbol{r})f^{a'}(\boldsymbol{r})W_0(\boldsymbol{r},\boldsymbol{r}',t)f^{b'}(\boldsymbol{r}')f^b(\boldsymbol{r}')\,d^3r d^3r'. \tag{12}$$

Using the identity (9), we rewrite the latter equation as

$$\Sigma^{ab}(\omega) = \mathrm{i}G^0_{a'b'}(t)V_\mu^{aa'}W_0^{\mu\nu}(t)V_\nu^{b'b}, \tag{13}$$

where the matrix elements of the screened Coulomb interaction appear

$$W_0^{\mu\nu}(t) = \iint F^\mu(\boldsymbol{r})W_0(\boldsymbol{r},\boldsymbol{r}',t)F^\nu(\boldsymbol{r}')\,d^3r d^3r'. \tag{14}$$

Finally, the equation (4) gives rise to the corresponding tensor expression

$$W_0^{\mu\nu}(\omega) = (\delta^\mu_{\nu'} - v^{\mu\mu'}\chi^0_{\mu'\nu'}(\omega))^{-1}v^{\nu'\nu}. \tag{15}$$

The last expression can be elucidated by developing the operator $[1-v\chi_0]^{-1}$ in a geometric series $[1-v\chi_0]^{-1} = 1 + v\chi_0 + v\chi_0 v\chi_0 + v\chi_0 v\chi_0 v\chi_0 \ldots$ The expressions (8), (10), (13) and (15) are tensor counterparts of Hedin's equations in coordinate space (5), (1), (3) and (4), correspondingly. In the next section, we will present our method for treating the frequency (and time) dependence of these tensor equations.

4 Spectral function technique

Because of the discontinuities of the electronic Green's functions, a direct, straightforward and accurate computation of the response function (8) is practically impossible both in the time domain and in the frequency domain. However, one can use an imaginary time technique [12] or spectral function representations to recover a computationally feasible approach. In this work, we use spectral function techniques and rewrite the time ordered

 ISBN: 978-84-614-6167-7

operators as follows

$$\begin{aligned}
G^0_{ab}(t) &= -\mathrm{i}\theta(t)\int_0^\infty ds\,\rho^+_{ab}(s)e^{-\mathrm{i}st} + \mathrm{i}\theta(-t)\int_{-\infty}^0 ds\,\rho^-_{ab}(s)e^{-\mathrm{i}st};\\
\chi^0_{\mu\nu}(t) &= -\mathrm{i}\theta(t)\int_0^\infty ds\,a^+_{\mu\nu}(s)e^{-\mathrm{i}st} + \mathrm{i}\theta(-t)\int_{-\infty}^0 ds\,a^-_{\mu\nu}(s)e^{-\mathrm{i}st};\\
W^{\mu\nu}_0(t) &= -\mathrm{i}\theta(t)\int_0^\infty ds\,\gamma^{\mu\nu}_+(s)e^{-\mathrm{i}st} + \mathrm{i}\theta(-t)\int_{-\infty}^0 ds\,\gamma^{\mu\nu}_-(s)e^{-\mathrm{i}st};\\
\Sigma^{ab}(t) &= -\mathrm{i}\theta(t)\int_0^\infty ds\,\sigma^{ab}_+(s)e^{-\mathrm{i}st} + \mathrm{i}\theta(-t)\int_{-\infty}^0 ds\,\sigma^{ab}_-(s)e^{-\mathrm{i}st},
\end{aligned} \tag{16}$$

where "positive" and "negative" spectral functions define the whole spectral function by means of Heaviside functions $\theta(t)$. For instance, the spectral function of the electronic Green's function reads

$$\rho_{ab}(s) = \theta(s)\rho^+_{ab}(s) + \theta(-s)\rho^-_{ab}(s).$$

Transforming the first of equations (16) to the frequency domain, we obtain the familiar expression for the spectral representation of a Green's function

$$G^0_{ab}(\omega) = \int_{-\infty}^\infty \frac{\rho_{ab}(s)\,ds}{\omega - s + \mathrm{i}\,\mathrm{sgn}(s)\varepsilon}. \tag{17}$$

Here ε is a small line-broadening constant. In practice, the choice of ε is related to the spectral resolution $\Delta\omega$ of the numerical treatment.

As first application of representations (16), we derive the spectral function of the non interacting response $a_{\mu\nu}(s)$ using equation (5) as a starting point

$$a^+_{\mu\nu}(s) = \iint V^{ab}_\mu \rho^+_{bc}(s_1)V^{cd}_\nu \rho^-_{da}(-s_2)\delta(s_1+s_2-s)ds_1ds_2. \tag{18}$$

Here, the convolution can be computed with fast Fourier methods and the (time-ordered) response function $\chi^0_{\mu\nu}(\omega)$ can be obtained with a Cauchy transformation

$$\chi^0_{\mu\nu}(\omega) = \chi^+_{\mu\nu}(-\omega) + \chi^+_{\mu\nu}(\omega), \text{ where } \chi^+_{\mu\nu}(\omega) = \int_0^\infty ds\,\frac{a^+_{\mu\nu}(s)}{\omega + \mathrm{i}\varepsilon - s}. \tag{19}$$

The calculation of the screened interaction $W^{\mu\nu}_0(\omega)$ must be done with functions, rather than with spectral functions, because of the inversion in equation (15). The spectral function of the screened interaction $\gamma^{\mu\nu}(\omega)$ can be easily recovered from the screened interaction itself [6]. Since $\mathrm{Im}\frac{1}{\omega+\mathrm{i}\varepsilon-s}$ is a representation of Dirac δ-function when ε goes to zero, then $\gamma^{\mu\nu}(\omega) = -\frac{1}{\pi}\mathrm{Im}\,W^{\mu\nu}_0(\omega)$.

 ISBN: 978-84-614-6167-7

Deriving the spectral function $\sigma(\omega)$ of the self-energy, we arrive at

$$\sigma_+^{ab}(s) = \int_0^\infty \int_0^\infty \delta(s_1+s_2-s)\, V_\mu^{aa'} \rho_{a'b'}^+(s_1) V_\nu^{b'b} \gamma_+^{\mu\nu}(s_2) ds_1 ds_2, \tag{20}$$
$$\sigma_-^{ab}(s) = -\int_{-\infty}^0 \int_{-\infty}^0 \delta(s_1+s_2-s) V_\mu^{aa'} \rho_{a'b'}^-(s_1) V_\nu^{b'b} \gamma_-^{\mu\nu}(s_2) ds_1 ds_2.$$

These expressions show that the spectral function of a convolution is given by a convolution of the corresponding spectral functions.

4.1 Discretization of frequency-dependent quantities

The spectral functions in equation (18) are merely a set of poles at (eigen)energies E

$$\rho_{ab}^+(\omega) = \sum_{E>0} \delta(\omega - E) X_a^E X_b^E,\ \rho_{ab}^-(\omega) = \sum_{E<0} \delta(\omega - E) X_a^E X_b^E. \tag{21}$$

Here the eigenvectors X_a^E diagonalize the corresponding Kohn-Sham Hamiltonian

$$H^{ab} X_b^E = E S^{ab} X_b^E,$$

where the Hamiltonian and the overlap matrices of atomic orbitals $f^a(\boldsymbol{r})$ enter

$$H^{ab} = \int f^a(\boldsymbol{r}) H(\boldsymbol{r}) f^b(\boldsymbol{r}) d^3r, \text{ and } S^{ab} = \int f^a(\boldsymbol{r}) f^b(\boldsymbol{r}) d^3r. \tag{22}$$

In practice, we use the SIESTA package [13] that gives the orbitals $f^a(\boldsymbol{r})$, eigenvectors X_a^E and eigenvalues E for a given molecule as the output of a DFT calculation.

The use of fast Fourier techniques for convolution, for instance in equation (18), requires that the spectral functions $\rho_{bc}^+(\omega)$, $\rho_{da}^-(\omega)$ be known at equidistant grid points $\omega_j = j\Delta\omega, j = -N_\omega \dots N_\omega$, rather than at a set of energies resulting from a diagonalization procedure. The solution for this problem (discretization of spike-like functions) is known and well tested [14]. We define a grid of points that covers the whole range of eigen energies E. Going through the poles E, we assign their spectral weight $X_a^E X_b^E$ to the neighboring grid points n and $n+1$ such that $\omega_n \le E < \omega_{n+1}$ according to the distance between the pole and the grid points $p_{n,ab} = \dfrac{\omega_{n+1} - E}{\Delta\omega} X_a^E X_b^E,\ p_{n+1,ab} = 1 - p_{n,ab}$. Such a discretization keeps both the spectral weight and the center of mass of a pole. It also reduces the number of operations that are needed to calculate the non interacting response function $\chi_{\mu\nu}^0(\omega)$. This is so because the number of frequencies N_ω can be kept small (typically a few hundred points) even for large molecules while the number of states N_{orb} grows linearly with the size of the system.

 ISBN: 978-84-614-6167-7

4.2 The second window technique

The discretization of spectral weight helps to control the computational complexity for large molecules. However, we are actually interested in the properties of low lying levels (HOMO and LUMO and several levels below and above). At first sight one might think that one could neglect the contributions of high energy spectral weights in the Cauchy transformation. However, neglecting the high energy spectral weight actually results in a wrong real part of the functions. Fortunately, the high energy spectral weight tolerates a coarser grid [8]. Therefore, we calculate each spectral function twice: once with a higher resolution in a low frequency range, and a second time with a lower resolution but in the whole range. The Cauchy transformation for such a two-window representation must be modified as follows

$$\begin{aligned}\chi^0_{\mu\nu}(\omega+\mathrm{i}\varepsilon_{\text{small}}) &= \int_{-\lambda}^{\lambda} ds \frac{a_{\mu\nu}(s)}{\omega+\mathrm{i}\varepsilon_{\text{small}}-s} + \left(\int_{-\Lambda}^{-\lambda}+\int_{\lambda}^{\Lambda}\right) ds \frac{a_{\mu\nu}(s)}{\omega+\mathrm{i}\varepsilon_{\text{large}}-s} \\ &= \chi^{\text{small window}}_{\mu\nu}(\omega+\mathrm{i}\varepsilon_{\text{small}}) + \left[\chi^{\text{large window}}_{\mu\nu}(\omega+\mathrm{i}\varepsilon_{\text{large}})\right]_{\text{truncated spectral function}}. \quad (23)\end{aligned}$$

After the calculation of spectral functions in both windows, we truncate the spectral function in the second window in the range $0\ldots\lambda$, do Cauchy transform of both spectral functions and update (by a linear interpolating procedure) the function in the first window with the truncated function from the second window.

We use the second window technique both for the non interacting response function $\chi^0_{\mu\nu}(\omega)$ and for the self-energy $\Sigma^{ab}(\omega)$.

5 Non-local compression of the product basis

The basis of dominant products is optimal within a given atom pair, but unfortunately, there is still a lot of collinearity between dominant products belonging to different pairs. This collinearity is an indication that the size of the product basis can be strongly reduced. Even for the molecules of modest size considered in Section 7.1 the basis set of dominant product becomes so large that hampers the storage of the (non-interacting) response function (19) and slows down the inversion in the calculation of the screened interaction (15). In order to improve the situation we perform a non-local (global) contraction of the basis of dominant product. We start by considering a sum-over-states expression for the non-interacting response function in the basis of dominant products

$$\chi^0_{\mu\nu}(\omega) = 2\sum_{E,F} V^{EF}_{\mu} \frac{n_F - n_E}{\omega + i\varepsilon - (E-F)} V^{EF}_{\nu}, \text{ where } V^{EF}_{\mu} = X^E_a V^{ab}_{\mu} X^F_b. \quad (24)$$

 ISBN: 978-84-614-6167-7

The response $\chi^0_{\mu\nu}(\omega)$ is built up from vectors V^{EF}_μ that represent electron-hole pair excitations. One can use these vectors to identify important directions in the space of dominant products. The number of electron–hole pairs EF grows as N^2 with the molecular size while the dimension of dominant product basis is $O(N)$ by construction (due to the localization of the basis orbitals). Therefore, one has to limit the set of electron-hole pairs EF from the beginning to keep the efficiency of the algorithm, particularly if one uses a diagonalization-based procedure for generating the (globally) optimal basis. Because of the inherent limitations of LCAO to represent high energy features, and the fact that we are mainly interested in the lowest lying excitations, we choose $O(N)$ low-energy electron-hole pairs

$$\{X^n_\mu\} \equiv \text{subset of } \{V^{EF}_\mu\} \text{ limited by } |E-F| < E_{\text{threshold}}, n = 1 \dots N_{\text{rank}}. \tag{25}$$

After the initial selection according to the energy criterion $|E-F| < E_{\text{threshold}}$, we define a metric g^{mn}

$$g^{mn} = X^m_\mu v^{\mu\nu} X^n_\nu, \text{ where } v^{\mu\nu} = \iint F^\mu(\boldsymbol{r}) |\boldsymbol{r}-\boldsymbol{r}'|^{-1} F^\mu(\boldsymbol{r}') d^3r d^3r'. \tag{26}$$

After diagonalizing the metric $g^{mn}\xi^\lambda_n = \lambda \xi^\lambda_m$, we can identify important directions (like in the construction of the basis of dominant products [11, 15]) by building linear combinations of the original vectors X^m_μ and by choosing only eigenvectors with eigenvalues above a suitable threshold value

$$Z^\lambda_\mu \equiv X^m_\mu \xi^\lambda_m / \sqrt{\lambda}. \tag{27}$$

These linear combinations can be used to expand the original response function $\chi^0_{\mu\nu}(\omega)$ in terms of fewer functions

$$\chi^0_{\mu\nu}(\omega) = Z^m_\mu \chi^0_{mn}(\omega) Z^n_\nu. \tag{28}$$

In order to express $\chi^0_{mn}(\omega)$ in terms of $\chi^0_{\mu\nu}(\omega)$ we multiply equation (28) with $Z^m_\mu v^{\mu\nu}$ from both sides and notice that $Z^m_\mu v^{\mu\nu} Z^n_\nu \equiv Z^m_\mu Z^\mu_n = \delta^m_n$. Therefore, the response function can be "compressed" by using basis vectors $Z^\nu_n \equiv v^{\mu\nu} Z^n_\nu$

$$\chi^0_{mn}(\omega) = Z^\mu_m \chi^0_{\mu\nu}(\omega) Z^\nu_n. \tag{29}$$

The particular choice of the Coulomb metric $v^{\mu\nu}$ in equation (26) simplifies the computation of the Coulomb screened interaction (15). We can rewrite the equation (15) in terms of a Taylor series

$$W^{\mu\nu}_0 = v^{\mu\nu} + v^{\mu\mu'} \chi^0_{\mu'\nu'} v^{\nu'\nu} + v^{\mu\mu'} \chi^0_{\mu'\nu'} v^{\nu'\mu''} \chi^0_{\mu''\nu''} v^{\nu''\nu} + \cdots \tag{30}$$

Inserting here the response function $\chi^0_{\mu\nu}$ according to equation (28) and recalling the identity $Z^m_\mu Z^\mu_n = \delta^m_n$, one arrives at

$$\begin{aligned} W^{\mu\nu}_0 &= v^{\mu\nu} + Z^\mu_m \chi^0_{mr} \left[\delta_{rn} + \chi^0_{rn} + \chi^0_{rs}\chi^0_{sn} + \cdots \right] Z^\nu_n = \\ &= v^{\mu\nu} + Z^\mu_m \chi^{\text{RPA}}_{mn} Z^\nu_n, \text{ where } \chi^{\text{RPA}}_{mn} \equiv \left(\delta_{mk} - \chi^0_{mk} \right)^{-1} \chi^0_{kn}. \end{aligned} \tag{31}$$

 ISBN: 978-84-614-6167-7

At this point it should be also be noted that the self-energy $\Sigma_x^{ab}(\omega)$ that corresponds to the instantaneous part of the screened interaction $v^{\mu\nu}$ is computed separately [6, 8] without any non local compression.

6 Computational complexity of the algorithm

The number of mathematical operations spent in different parts of the approach presented above can be estimated if the dimensions of the corresponding matrices are known. The numbers that determine the complexity of the algorithm are the number of atomic orbitals N_{orb}, the number of dominant functions N_{prod} and the number of frequencies N_ω. The number of orbitals and the number of dominant products are proportional to the number of atoms N by construction. The number of frequencies affects the run time linearly, but it is independent of the number of atoms. The non-local basis of section 5, can be constructed in $O(N^3)$ operations because N_{rank} in equation (25) can be kept proportional to number of orbitals. In practical calculations we have found that converged results are achieved with $N_{\text{rank}} \sim 5N_{\text{orb}}$. For large molecules, the number of important eigenvectors N_{subrank} after dropping small eigenvalues λ in equation (27) is approximately N_{orb}. No part of the algorithm scales worse than $O(N^3)$ [8]. There are several portions of the code where $O(N^3)$ operations are needed. However, only two of them have an appreciable impact on the run time: the computation of the response function and the computation of the self-energy. Both of them scale as $O(N_{\text{prod}}^2 N_{\text{subrank}} N_\omega)$ and give rise to an overall $O(N^3)$ scaling of the run time.

7 Applications to organic molecules

The methods described in the previous sections were carefully tested on several molecules. In this paper, we present two examples: calculations of HOMO and LUMO levels of three aromatic hydrocarbons (benzene, naphthalene and anthracene) and a calculation of the HOMO and LUMO levels of fullerene C_{60}.

7.1 Aromatic hydrocarbons

From the interacting Green's function $G_{ab}(\omega)$ we calculate the density of states (DOS) $\rho(\omega) = -S^{ab}\text{Im}G_{ab}(\omega)/\pi$ and we then determine the positions of the HOMO and LUMO levels from the DOS.

 ISBN: 978-84-614-6167-7

The results of this procedure for aromatic hydrocarbons are collected in table 1. One can see that our (LDA+G_0W_0) approach delivers *qualitatively correct predictions* for the ionization potentials (IP) and electron affinities (EA) of benzene and naphthalene (donors) and anthracene (acceptor). On the other hand, the LUMO of the underlying DFT calculation is always below the vacuum level. The calculations have been done on top of DFT-SIESTA calculations. We used pseudo potentials of Troullier-Martins type [17] and the Perdew-Zunger exchange-correlation functional [18]. We found that rather extended atomic orbitals must be used to achieve converged results in our *GW* approach. The energy shift parameter [19], that controls the spatial extension of atomic orbitals has been set to 3 meV for benzene, and to 20 meV for naphthalene and anthracene. The spectral functions have been discretized in two energy windows, with each window containing $N_\omega = 64$ frequency points.

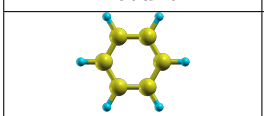
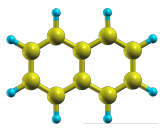
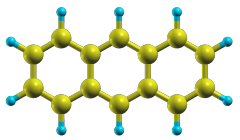
Picture	IP, eV	EA, eV
	8.82 (9.25)	-1.43 (-1.12)
	7.58 (8.14)	-0.15 (-0.19)
	6.87 (7.44)	0.73 (0.530)

Table 1: The ionization potentials (IP) and electron affinities (EA) of benzene, naphthalene and anthracene. Experimental data [16] are given in brackets.

7.2 Fullerene C_{60}

The fullerene C_{60} and its derivatives are very popular ingredients in organic semiconductors and extensive experimental data and theoretical computations are available for the basic fullerene. We found that our LDA+G_0W_0 results are in very good agreement with experimental data (see table 2). The computational parameters of this calculation are the same as in subsection 7.1, while the energy shift parameter is chosen to be 15 meV. The number of frequency points was chosen rather large $N_\omega = 128$ and the calculation has been done with 8 cores of a Nehalem machine (Intel®E5520 2.27GHz, Cache 8M/DDR3 RAM 24GB). The current version of the code consumed 26 hours of wall clock time.

Source	IP, eV	EA, eV
Our LDA+G_0W_0	7.33	2.97
Experimental [16]	7.58	2.65

Table 2: The IP and EA of fullerene C_{60} calculated with our method and corresponding experimental data.

A comparison of DOS calculated with DFT LDA Hamiltonian and with our LDA+G_0W_0 approach is shown in figure 5. Such a result is a typical when Hedin's *GW* approach is applied on top of a LDA calculation. *GW* HOMO has lower energy than DFT HOMO. Therefore, the change density $n(\boldsymbol{r})$ will be more localized in the *GW* calculation. *GW* LUMO has higher energy than DFT LUMO. Therefore, the change density $n(\boldsymbol{r})$ will be more delocalized in the *GW* calculation.

 ISBN: 978-84-614-6167-7

8 Conclusions

We have described our approach to Hedin's GW approximation for finite systems. This approach allows to compute the interacting Green's function on a frequency grid. The density of states is our output and it provides HOMO and LUMO levels in reasonable agreement with experiment. The complexity of the approach scales with the third power of the number of atoms, while the needed memory scales with the second power of the number of atoms. These features make our approach suitable for treating the large molecules that are used in organic semiconductors.

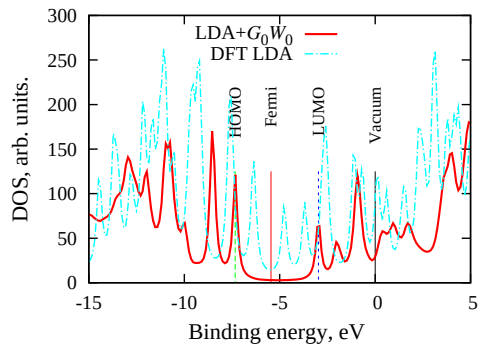


Figure 5: DOS of fullerene C_{60} computed with our LDA-G_0W_0 approach.

Acknowledgments

We thank James Talman for inspiring discussions, correspondence and essential algorithms and computer codes [20], that are used in our implementation. We are indebted to the organizers of the ETSF2010 meeting at Berlin for feedback and perspective on the ideas of this paper. Arno Schindlmayr, Xavier Blase and Michael Rohlfing helped with extensive correspondence on various aspects of the GW method. DSP and PK acknowledge financial support from the Consejo Superior de Investigaciones Científicas (CSIC), the Basque Departamento de Educación, UPV/EHU (Grant No. IT-366-07), the Spanish Ministerio de Ciencia e Innovación (Grants No. FIS2010-19609-C02-02) and, the ETORTEK program funded by the Basque Departamento de Industria and the Diputación Foral de Guipuzcoa.

References

[1] See for example the review paper, D. Braga and G. Horowitz, *High-Performance Organic Field-Effect Transistors*, Adv. Mater. **21** 1473 (2009).

[2] A. Kokalj, Comp. Mater. Sci. **28**, 155 (2003).

[3] P. Hohenberg and W. Kohn, Phys. Rev. **136**, B864 (1964); W. Kohn and L. J. Sham, Phys. Rev. **140**, A1133 (1965).

[4] X. Blase, C. Attaccalite, V. Olevano, Phys. Rev. B **83**, 115103 (2011); C. Rostgaard, K. W. Jacobsen, and K. S. Thygesen, Phys. Rev. B **81**, 085103 (2010).

 ISBN: 978-84-614-6167-7

[5] L. Hedin, Phys. Rev. **139**, A796 (1965). For a review see F. Aryasetiawan and O. Gunnarsson, Rep. Prog. Phys. **61**, 237 (1998).

[6] C. Friedrich and A. Schindlmayr, *Many-Body Perturbation Theory: The GW Approximation*, NIC Series, **31**, 335 (2006).

[7] D. Pines, *Elementary Excitations in Solids* (Wiley, New York, 1964).

[8] D. Foerster, P. Koval, and D. Sánchez-Portal, arXiv `http://arxiv.org/abs/1101.2065`, submitted (2011).

[9] M. E. Casida, in *Recent Advances in Density Functional Theory*, edited by D. P. Chong (World Scientific, Singapore, 1995, p. 155).

[10] P. Umari, G. Stenuit and S. Baroni, Phys. Rev. B **79**, 201104R (2009); Phys. Rev. B **81**, 115104 (2009).

[11] D. Foerster, J. Chem. Phys. **128**, 34108 (2008).

[12] M. M. Rieger, L. Steinbeck, I. D. White, H. N. Rojas, R. W. Godby, Comp. Phys. Comm. **117**, 211 (1999).

[13] J. M. Soler, E. Artacho, J. D. Gale, A. García, J. Junquera, P. Ordejón and D. Sánchez-Portal, J. Phys.: Condens. Matter **14**, 2745 (2002); E. Artacho, E. Anglada, O. Dieguez, J. D. Gale, A. García, J. Junquera, R. M. Martin, P. Ordejón, J. M. Pruneda, D. Sánchez-Portal and J. M. Soler, J. Phys.: Condens. Matter **20**, 064208 (2008).

[14] M. Shishkin and G. Kresse, Phys. Rev. B **74**, 035101 (2006); D. Foerster and P. Koval, J. Chem. Phys. **131**, 044103 (2009).

[15] F. Aryasetiawan and O. Gunnarsson, Phys. Rev. B **49**, 16214 (1994).

[16] For hydrocarbons: `http://cccbdb.nist.gov/` and J. C. Rienstra-Kiracofe, Ch. J. Barden, Sh. T. Brown, and H. F. Schaefer, J. Phys. Chem. A **105**, 524 (2001); for fullerene C_{60} at `http://sesres.com/PhysicalProperties.asp`.

[17] N. Troullier and J. L. Martins, Phys. Rev. B **43**, 1993 (1991).

[18] J. P. Perdew and A. Zunger, Phys. Rev. B **23**, 5048 (1981).

[19] J. Junquera, Ó. Paz, D. Sánchez-Portal, and E. Artacho, Phys. Rev. B **64**, 235111 (2001).

[20] J. D. Talman, J. Chem. Phys. **80**, 1984 (2000); J. Comput. Phys. **29**, 35 (1978); Comput. Phys. Commun. **30**, 93 (1983); Comput. Phys. Commun. **180**, 332 (2009).

 ISBN: 978-84-614-6167-7

Proceedings of the 11th International Conference on Computational and Mathematical Methods in Science and Engineering, CMMSE 2011 26–30 June 2011.

Algorithm for computing matrices that involve some of their powers and an involutory matrix

Leila Lebtahi[1], Óscar Romero[2] and Néstor Thome[1]

[1] *Instituto Universitario de Matemática Multidisciplinar, Universitat Politècnica de València*

[2] *Departamento de Comunicaciones, Universitat Politècnica de València*

emails: leilebep@mat.upv.es, oromero@dcom.upv.es, njthome@mat.upv.es

Abstract

In this paper, we deal with $\{K, s+1\}$-potent matrices. These matrices generalize all the following classes of matrices: k-potent matrices, periodic matrices, idempotent matrices, involutory matrices, centrosymmetric matrices, mirrorsymetric matrices, circulant matrices, etc. Several applications of these classes of matrices can be found in the literature. We develop an algorithm in order to compute $\{K, s+1\}$-potent matrices by using spectral theory. In addition, some examples are presented in order to show the numerical performance of the method.

Key words: potent matrix, involutory matrix, algorithm
MSC 2000: AMS codes 15A24

1 Introduction

In the last years, real applications by using different classes of matrices have been developed. Specifically, the problem of multiconductor transmission lines has been studied by means of mirror symmetric matrices in [4] and [5]. Also, circulant matrices have been applied to solve problems in several areas such as numerical computation, solid state-physics, image process, signal processing, coding theory, mathematical statistics, and molecular vibration [2]. Some applications of centrosymmetric matrices have been given in [1], for example, for solving problems in pattern recognition, antenna theory, mechanical and electrical systems, and quantum physics. In this case, symmetric and skew-symmetric eigenvectors have been used.

Related to the aforementioned classes of matrices, another type of matrices was introduced in [3], namely $\{K, s+1\}$-potent matrices. For a given involutory matrix $K \in \mathbb{C}^{n\times n}$ ($K^2 = I$) and $s \in \{1, 2, 3, \dots\}$, we recall that a matrix $A \in \mathbb{C}^{n\times n}$ is called $\{K, s+1\}$-potent if it satisfies

$$KA^{s+1}K = A$$

 ISBN: 978-84-614-6167-7

It can be seen that $\{K, s+1\}$-potent matrices generalize all the following classes of matrices: k-potent matrices, periodic matrices, idempotent matrices, involutory matrices, centrosymmetric matrices, mirrorsymmetric matrices, circulant matrices, etc. [3]. Hence, it is interesting to know how to construct a sufficient number of this new class of matrices. The main aim of this paper is to develop a method to construct them.

Throughout this work, K will stand for an involutory matrix and s a positive integer. We will denote by Ω_k the set of all k^{th} roots of unity with k a positive integer, that is, if we define $\omega_k = e^{2\pi i/k}$ then $\Omega_k = \{\omega_k, \omega_k^2, \ldots, \omega_k^k\}$.

The following function will be required. Let $\mathbb{N}_s = \{0, 1, 2, \ldots, (s+1)^2 - 2\}$ and

$$\varphi : \mathbb{N}_s \to \mathbb{N}_s$$

be the bijective function given by $\varphi(j) = b_j$ where b_j is the smallest nonnegative integer such that $b_j \equiv j(s+1) \pmod{((s+1)^2 - 1)}$ [3].

2 Algorithm for computing $\{K, s+1\}$-potent matrices

In this section an algorithm for computing $\{K, s+1\}$-potent matrices will be developed.

Given $K \in \mathbb{C}^{n\times n}$ and s as stated before, we want to find a $\{K, s+1\}$-potent matrix $A \in \mathbb{C}^{n\times n}$. It is clear that the cases $K = \pm I$ only provide the well-known results corresponding to $A^{s+1} = A$ and we are not interested in these situations.

Since K is involutory, there is a nonsingular matrix $T = \begin{bmatrix} t_1 & \ldots & t_n \end{bmatrix}$ such that

$$K = T \begin{bmatrix} -I_r & O \\ O & I_{n-r} \end{bmatrix} T^{-1} \tag{1}$$

where the first r eigenvectors of K are associated with the eigenvalue -1. Without loss of generality, we will assume that $r \leq n - r$. Otherwise, we pick $-K$ instead of K obtaining the same solution. It is well-known that the eigenvalues of A are included in the following set $\Lambda = \left\{0, w^1_{(s+1)^2-1}, \ldots, w^{(s+1)^2-2}_{(s+1)^2-1}, 1\right\}$ and A is diagonalizable [3], i.e.

$$A = S\mathrm{diag}(\lambda_1, \ldots, \lambda_n)S^{-1} \qquad \text{with} \quad S = \begin{bmatrix} s_1 & \ldots & s_n \end{bmatrix} \quad \text{and} \quad S^{-1} = \begin{bmatrix} y_1^T \\ \vdots \\ y_n^T \end{bmatrix}$$

Then, denoting by $P_i = s_i y_i^T$ we have $A = \sum_{i=1}^n \lambda_i P_i$.

The idea of the method is based on the construction of the s_i's and y_i's in terms of the t_i's. The most representative cases are presented in the following table for different involutory matrices $K_i = TD_iT^{-1}$ for $i = 1, 2, 3, 4, 5$ where $D_1 = \mathrm{diag}(-1, 1)$, $D_2 = \mathrm{diag}(-1, 1, 1)$, $D_3 = \mathrm{diag}(-1, 1, 1, 1)$, $D_4 = \mathrm{diag}(-1, -1, 1, 1)$, $D_5 = \mathrm{diag}(-1, -1, 1, 1, 1)$.

Although we are going to construct only one $\{K, s+1\}$-potent matrix A, it is clear that this method allows us to construct more of them (e.g., by changing adequately the ω's in $\Omega_{(s+1)^2-1}$).

 ISBN: 978-84-614-6167-7

Table 1. The most representative cases.

	Construction of s_i''s	$s=1$	$s\geq 2$
D_1	$s_1=s_1'=t_1+t_2$ $s_2=s_{\varphi(1)}'=-t_1+t_2$	$A=\omega P_1+\omega^{\varphi(1)}P_2$	$A=\omega^j P_1+\omega^{\varphi(j)}P_2$
D_2	$s_1=s_1'=t_1+t_2$ $s_2=s_{\varphi(1)}'=-t_1+t_2$ $s_3=t_3$	$A=\omega P_1+\omega^{\varphi(1)}P_2+P_3$	$A=\omega^j P_1+\omega^{\varphi(j)}P_2+P_3$
D_3	$s_1=s_1'=t_1+t_2$ $s_2=s_{\varphi(1)}'=-t_1+t_2$ $s_3=t_3$ $s_4=t_4$	$A=\omega P_1+\omega^{\varphi(1)}P_2+P_3+P_4$	$A=\omega^j P_1+\omega^{\varphi(j)}P_2+P_3+P_4$
D_4	$s_1=s_1'=t_1+t_3$ $s_2=s_{\varphi(1)}'=-t_1+t_3$ $s_3=s_a'=t_2+t_4$ $s_4=s_{\varphi(a)}'=-t_2+t_4$	$A=\omega(P_1+P_3)+$ $+\omega^{\varphi(1)}(P_2+P_4)$	$A=\omega^j(P_1+P_3)+$ $+\omega^{\varphi(j)}(P_2+P_4)$
D_5	$s_1=s_1'=t_1+t_3$ $s_2=s_{\varphi(1)}'=-t_1+t_3$ $s_3=s_a'=t_2+t_4$ $s_4=s_{\varphi(a)}'=-t_2+t_4$ $s_5=s_b'=t_5$	$A=\omega(P_1+P_3)+$ $+\omega^{\varphi(1)}(P_2+P_4)+P_5$	$A=\omega^j(P_1+P_3)+$ $+\omega^{\varphi(j)}(P_2+P_4)+P_5$

where $j\in\mathbb{N}_s-\{0\}$, $w=w_{(s+1)^2-1}$, and moreover $a\in\mathbb{N}_s-\{0,1,\varphi(1)\}$ such that $\varphi(a)\neq a$, and $b\in\mathbb{N}_s-\{0\}$ such that $\varphi(b)=b$.

Algorithm

Step 1 Diagonalize K as in (1).

Step 2 If $r>n-r$, change K to $-K$ and rearrange adequately as in Step 1.

Step 3 For $i=1,\ldots,r$, compute $s_{2i-1}=t_i+t_{i+r}$ and $s_{2i}=-t_i+t_{i+r}$.

Step 4 For $i=2r+1,\ldots,n$, set $s_i=t_i$.

Step 5 Solve the linear systems $Sy_i=e_i$, where e_i are the canonical basis vectors of $\mathbb{R}^n$ for $i=1,\ldots,n$.

Step 6 Compute $P_i=s_i y_i^T$ for $i=1,\ldots,n$.

Step 7 For $i=1,\ldots,r$, compute $Q_i=\omega P_{2i-1}+\omega^{\varphi(1)}P_{2i}$.

Step 8 Compute $A=\sum_{i=1}^{r}Q_i+\sum_{i=2r+1}^{n}P_i$.

End

3 Numerical examples

In this section we present some numerical examples in order to show the performance of our algorithm. The algorithm has been implemented by using MATLAB R2010b.

 ISBN: 978-84-614-6167-7

Example 1 $s = 2$, $n = 4$, D_4

$$K = \begin{bmatrix} 1.2989 & -1.2069 & 3.5632 & 3.0460 \\ 1.3793 & -2.7241 & 4.1379 & 4.8276 \\ -0.2759 & 1.3448 & -3.8276 & -3.9655 \\ 0.6437 & -2.1379 & 4.5977 & 5.2529 \end{bmatrix}$$

$$A = \begin{bmatrix} -0.3820 - 0.0731i & 1.2435 - 0.0853i & -0.9103 + 0.4877i & -0.9834 + 1.1582i \\ 0.3657 + 0.5202i & 0.6035 - 0.4145i & -1.0241 - 0.3251i & -1.0180 + 0.4064i \\ -0.2845 - 0.1300i & 0.4145 - 0.7803i & 0.0894 + 0.7884i & -0.6421 + 0.9591i \\ 0.3657 + 0.5202i & -0.1036 + 0.2926i & -1.0241 - 0.3251i & -0.3109 - 0.3007i \end{bmatrix}$$

Example 2 $s = 2$, $n = 5$, D_5

$$K = \begin{bmatrix} 1.4023 & 0.6228 & 0.3946 & 0.2166 & -0.1702 \\ -0.0290 & 0.2147 & -0.9323 & 0.2921 & 0.1277 \\ -0.6180 & -1.3269 & -0.0580 & -0.1789 & 0.3191 \\ -2.3269 & -1.3810 & 0.4294 & -1.1760 & 0.6383 \\ 1.1779 & 1.0832 & 0.2515 & 0.9420 & 0.6170 \end{bmatrix}$$

$$A = \begin{bmatrix} -1.1912 - 0.1115i & -1.3429 + 0.6127i & 0.1868 - 0.5649i & -1.2304 + 0.3207i & 0.0331 - 0.0752i \\ 0.5194 + 0.0841i & 1.0709 + 0.2968i & -0.0923 + 0.1573i & 0.3341 - 0.2811i & -0.0226 + 0.1956i \\ 0.1227 - 1.0996i & 0.5879 - 0.9136i & -0.8167 + 0.4445i & 0.4375 - 0.8097i & 0.4258 + 0.0301i \\ 0.1996 - 0.5314i & 0.3129 - 1.5127i & -0.2300 + 0.1792i & 0.9591 - 0.4493i & 0.2046 + 0.0752i \\ -0.3164 + 0.0410i & -0.0175 + 0.4035i & -0.1457 - 0.0957i & -0.0577 + 0.2941i & 0.9779 - 0.1805i \end{bmatrix}$$

Acknowledgements

This work has been partially supported by grant DGI MTM2010-18228 and grant Universitat Politècnica de València, PAID-06-09, Ref.: 2659.

References

[1] L. Datta, S.D. Morgera, *On the Reducibility of Centrosymmetric Matrices - Applications in Engeneering Problems*, Circuits Systems Signal Process, **8** 1 (1989) 71–96.

[2] B. Gellai, *Determination of molecular symmetry coordinates using circulant matrices*, Journal of Molecular Structure, **114** (1984) 21–26.

[3] L. Lebtahi, O. Romero, N. Thome, *Characterizations of $\{K, s+1\}$-Potent Matrices and Applications*, To appear in Linear Algebra and its Applications DOI: 10.1016/j.laa.2010.11.034.

[4] G.L. Li, Z.H. Feng, *Mirrorsymmetric Matrices, Their Basic Properties, and An Application on Odd/Even-Mode Decomposition of Symmetric Multiconductor Transmission Lines*, SIAM J. Matrix Anal. Appl., **24** 1 (2002) 78–90.

[5] G.L. Li, Z.H. Feng, *Mirror-Transformations of Matrices and Their Application on Odd/Even Modal Decomposition of Mirror-Symmetric Multiconductor Transmission Line Equations*, IEEE Transactions on Advanced Packaging **26** 2 (2003) 172–181.

 ISBN: 978-84-614-6167-7

Proceedings of the 11th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2011
26–30 June 2011.

Performance evaluation of GPU memory hierarchy using the FFT

Jacobo Lobeiras[1], Margarita Amor[2] and Ramón Doallo[3]

[1] *Computer Architecture Group (GAC), University of A Coruña (UDC), Spain*

emails: jacobo.lobeiras@udc.es, margarita.amor@udc.es, ramon.doallo@udc.es

Abstract

Modern GPUs (*Graphics Processing Units*) are becoming more relevant in the world of *HPC* (*High Performance Computing*) thanks to their large computing power and relative low cost, however their special architecture results in more complex programming. To take advantage of their computing resources and develop efficient implementations is essential to have certain knowledge about the architecture and memory hierarchy. In this paper we use the FFT (*Fast Fourier Transform*) as a benchmark tool to analyze different aspects of *GPU* architectures, like the influence of the memory access pattern or the impact of the register pressure. The *FFT* is a good tool for performance analysis because it is used in many real applications that require digital signal processing and has a good balance between computational cost and memory bandwidth requirements. The work presents a comparison of two *CUDA* architectures to analyze the evolution of the memory hierarchy, studying which are the most efficient solutions for each case.

Key words: Signal processing, performance analysis, FFT, GPGPU, CUDA.

1 Introduction

The specialized hardware design of modern *GPUs* (*Graphics Processing Units*), optimized for running graphics tasks, can perform much faster than normal *CPUs* (*Central Processing Units*) in many general purpose parallel applications. However, from a programmability standpoint, *GPU* programming is still complex as it requires special languages (like *NVIDIA's CUDA* [12] or *OpenCL* [8]) that often expose some limitations or hardware specific features. This restricts the flexibility of *GPUs* and forces the programmer to have some knowledge about the hardware to efficiently exploit the *GPU* resources. *GPUs* expose different memory types (like global, local, shared, constant and texture memory) and

 ISBN: 978-84-614-6167-7

let the user decide some advanced options like cache configuration. Due to the fast *GPU* evolution, the execution parameters or the most suited algorithm may even vary from one hardware generation to another. Even standard languages for heterogeneous computing, like *OpenCL*, may require different implementations depending on the underlying hardware to obtain a good efficiency, specially if we consider or require specific vendor extensions.

Detailed hardware specification and theoretical performance study provide information that can be used by the programmer in the optimization of applications. Many works were proposed about the elaboration of models for *GPU* performance analysis [15, 14] and automatic performance tuning [6]. Other works study concrete *GPU* architectures through micro-benchmarks [16, 2, 13]. However, empiric performance analysis of real applications also has great interest, as it provides valuable information to the programmer. Thanks to the study of real algorithms and applications it is possible to have a more general view of the architecture as a whole. For example, for *CPUs* there are well-known benchmarks, like the *SPEC* (*Standard Performance Evaluation Corporation*) suite. Currently, there are no standard benchmarks for *GPU* computing, and usually only very specific applications that take fully advantage of the hardware are studied.

The *FFT* is a very important operation for many applications, such as image and digital signal processing, filtering and compression, partial differential equation resolution or large number manipulation. There are several efficient proposals for *CPUs* like [4, 3, 9, 10] and a few *GPU FFT* implementations have started to appear, like [1], [5] or *NVIDIA's CUFFT*. The *FFT* algorithm has a fair computational cost, as well as notable bandwidth requirements with good flexibility in the memory access pattern and data distribution, which makes it an adequate tool for performance analysis. In this work a *CUDA* based *FFT* implementation is used to analyze the different memory types. This implementation is not centered on obtaining the highest performance, but it was designed focusing on flexibility to allow this analysis. More specifically, the *FFT* is used as a tool to study several performance aspects of the memory hierarchy in two *NVIDIA GPUs*, the *GeForce 280*, based on the *Tesla* architecture, and the *GeForce 480*, based on the *Fermi* architecture. Both *GPUs* were compared to find out the performance limiting factors and the most appropriate implementation strategies for each architecture.

This work is structured as follows. Section 2 introduces the architecture of the *GPUs* used in this work. Section 3 describes the different implementations of the *FFT* using *CUDA*. In Section 4 the experimental results are shown. Finally, in Section 5, the main conclusions are summarized.

2 NVIDIA CUDA GPU Architecture

NVIDIA GPUs are based on the *CUDA* architecture (*Computer Unified Device Architecture*) [12], which enables the execution of general purpose code on the *GPU* offering great

 ISBN: 978-84-614-6167-7

performance in tasks where a high degree of parallelism can be exploited. Each chip has several independent processing clusters, which are composed by a set of *SMs* (*Streaming Multiprocessors*). Each *SM* consists of many *SPs* (*Streaming Processors*), which process instructions in a *SIMD* fashion (*Single Instruction Multiple Data*). The number of *SMs* and *SPs* can vary for each specific model. *CUDA GPUs* have a specialized architecture with several memory types and distinct properties, that will be described.

The main memory of *GPUs* is the global memory. It can be accessed from all the *SPs* in every *SM*. Depending on the model it can accommodate one or more *GB*. It has two orders of magnitude more latency than on-chip memory, however it is optimized to simultaneously handle a lot of memory requests providing a considerable large aggregate bandwidth.

The fastest memory access is provided by hardware registers. *GPUs* have much more registers than *CPUs*, for instance, in the case of the *Fermi* architecture there are about 2 *MB* of registers, while current *CPUs* only have a few *KB*.

GPUs have a small memory within each *SM* that is common to a group of *SPs* and can be concurrently accessed by several threads. This shared memory is much faster than global memory but it offers less effective bandwidth than registers. It can be used as a user-managed cache to reduce the number of slow global memory accesses, to communicate several threads within a group so they can collaborate in a given task, or simply to temporally store data and reduce register pressure. Physically this memory is distributed in banks, and if several threads within a warp try to access different locations within the same bank, a bank conflict occurs and the access is serialized. While the *Tesla* architecture has 16 *KB* of shared memory per *SM*, *Fermi* allows the user to choose between having either 48 *KB* of shared memory but only 16 *KB* of *L1* cache, or just 16 *KB* of shared memory and 48 *KB* of *L1*. Furthermore, the *Fermi* architecture has improved memory hierarchy with the addition of general *L2* cache.

GPUs commonly store the texture data in the device global memory and access it using an spatially coherent access pattern. To render smooth graphics they sample textures using data spatial interpolation and decompression on the flight. For this purpose, *GPUs* have dedicated hardware with a texture cache. The texture memory can be read by all the *SPs*.

Finally, constant cache is a fast on-die memory and can be read by all the *SPs*, however it cannot be modified by the *GPU* kernels, only by the *CPU*.

3 FFT benchmarks using CUDA

To analyze the performance of the different storage types in the *Tesla* and *Fermi GPU* architectures a set of different *FFT* proposals were executed in *CUDA*. The *FFT* requires a significant amount of computation, and using transforms of different sizes and distributions [7, 11] it is possible to study several influence factors separately. Our implementation is based on the *Cooley-Tukey* algorithm [7], characterized by its regular structure and easy

 ISBN: 978-84-614-6167-7

implementation. This algorithm performs a bit-reversal operation at the beginning of the process, and then it operates on the data in pairs while increasing the stride in each stage.

In our FFT benchmarks each thread calculates an independent FFT and each block is composed by T_b threads which operate on different input data in batch mode. Therefore, we have developed a set of kernels for each signal, $N = \{4, 8, 16, 32\}$. All FFT kernels are recursively subdivided into smaller problems until reaching the base case of $N = 2$. Twiddle factors w_N^{ik} are stored in constant memory. This memory can be used to store commonly used values or the result of precalculated formulas, thus avoiding redundant computations or expensive global memory requests. Constant memory can store up to 64 KB of data, being optimized for broadcast memory access, where many threads read the same location, otherwise the requests may be serialized.

Figure 1 shows an example of the kernel used to compute the FFT for a signal of $N = 8$ data. It presents a complex input signal which can be processed in either registers or shared memory, recursively performing an in-place FFT (subkernels in lines 2, 9 and 18). Global and texture memory can also be used directly as inputs for the FFT_time kernel (line 32), however in our tests data will reside locally in the *SPs* before calling the FFT. FFT_time is part of a bigger function $FFTy$ (see Figure 2), which is the main kernel that manages the storage type and will be called by the host. In line 5 $FFTy$ reads the stride that will be used to access the different input signals within the batch, and in line 6 this stride is used to obtain a pointer to the problem that will be processed by the current thread. Following, the memory to store the signal is reserved, either using shared memory (lines 9 and 10) or registers (line 12), and then the data is copied to the current thread from texture memory (line 16) or global memory (line 18). Next, the FFT is performed, calling a forward FFT function (line 23) or a reverse and scale function (lines 25 and 26) depending on the direction DIR, which is a compile-time parameter. Finally, in line 29 the data is copied back to global memory. In line 32 the template $FFTy$ (line 1) is instantiated with different parameters in a table of function pointers.

Table 1 depicts some information for each kernel compiled for *CUDA* 1.3 and 2.0 capabilities using the verbose flag *–ptxas-options=-v* and allowing the compiler to take as many registers as necessary, up to a maximum of 127 registers per thread for the *Tesla* architecture and 63 registers for *Fermi*. If the maximum number of registers is reached, the compiler will resort to local memory to supply enough space for the private thread data, for example *Fermi* will require 16 bytes of local memory for $N = 16$ and 336 bytes for $N = 32$. Table 1 also displays information about the total constant memory reserved by the kernel (expressed in bytes), for example used by twiddle factors.

Our tests will be executed with $T_b = 32$ threads per block. The utilization with just 32 threads may seem rather low but, according to our tests, in most cases there is no significant advantage using 64 threads per block, and with 128 a small performance drop is experienced. Also note that the amount of shared memory in some of the tests may be too tight, thus

 ISBN: 978-84-614-6167-7

```
inline __host__ __device__ void
FFT2(Complex &a0, Complex &a1) {
    COMPLEX c0 = a0;
    a0 = c0 + a1;
    a1 = c0 - a1;
}

inline __host__ __device__ void
FFT4(Complex &a0, Complex &a1, Complex &a2, Complex &a3) {
    FFT2(a0, a2);
    FFT2(a1, a3);
      a3 = make_COMPLEX(a3.y, -a3.x);
    FFT2(a0, a1);
    FFT2(a2, a3);
}

inline __host__ __device__ void
FFT8(Complex &a0, Complex &a1, Complex &a2, Complex &a3,
       Complex &a4, Complex &a5, Complex &a6, Complex &a7) {
    FFT4(a0, a2, a4, a6);
    FFT4(a1, a3, a5, a7);
      a3 =       make_COMPLEX(       a3.y,      -a3.x);
      a5 = a5 * make_COMPLEX( ANG_4_8,-ANG_4_8);
      a7 = a7 * make_COMPLEX(-ANG_4_8,-ANG_4_8);
    FFT2(a0, a1);
    FFT2(a2, a3);
    FFT2(a4, a5);
    FFT2(a6, a7);
}

template<> inline __host__ __device__ void
FFT_time<8>(Complex *a) {
      FFT8(a[0], a[4], a[2], a[6],
           a[1], a[5], a[3], a[7]);
}
```

Figure 1: FFT kernel for N=8

```
template<int N, int DIR> __global__ void
FFTy(Complex* src) {
    int posX = get_global_id(0);
    int posY = get_global_id(1);
    int stride = get_global_size(0);    // Stride among elements
    src += posX + posY * N * stride;    // First pos in current batch

    #if SHM_MODE == 1                   // Registers or Shared Mem.
        __shared__ Complex _tmp[(N + 1) * 32];
        Complex* tmp = _tmp + (N + 1) * get_local_id(0);
    #else
        Complex tmp[N];
    #endif

    #if TEX_MODE == 1                   // Texture or Global Mem.
        copyY<N>(tmp, posX, posY * N);
    #else
        copy<N>(tmp, src, stride);
    #endif

    rev<N>(tmp);                        // Bit reversal
    if(DIR > 0) {                       // Compile-time condition
        FFT_time<N>(tmp);               // Forward FFT
    } else {
        IFFT_time<N>(tmp);              // Inverse FFT
        scale<N>(tmp);                  // Scaling
    }

    copy<N>(src, stride, tmp);          // Write back to Global Mem.
}

void(*fft_funptrs[2][4])(Complex*) = {
    { FFTy<4, 1>, FFTy<8, 1>, FFTy<16, 1>, FFTy<32, 1> },
    { FFTy<4,-1>, FFTy<8,-1>, FFTy<16,-1>, FFTy<32,-1> }
};
```

Figure 2: General kernel template

Table 1: Compiler information for the FFT kernel

	Tesla (CUDA cap. 1.3)		Fermi (CUDA cap. 2.0)	
N	Registers	Const (bytes)	Registers	Const (bytes)
4	14	4	18	0
8	28	12	36	4
16	58	28	63	12
32	123	56	63	28

restricting the maximum block size, so a common size of 32 was used for all the executions. Using just 32 threads it is possible to take advantage of the maximum number of registers allowed by the architecture without resorting to local memory, and if enough resources are available the *GPU* will be able to transparently execute several blocks per *SM*.

Following, three decisive parameters will be analyzed for an optimal *CUDA* implementation: Local Data *SPs*, storage type for input data, and access pattern of global data. The different configurations considered in our tests are shown in Figure 3. Each test will be assigned a three letter code according to its configuration parameters. For example, *STC* will mean that the test was performed using **S**hared memory, reading data from **T**extures with a **C**oalescent memory access pattern.

With respect to the register-based solution (R), a key feature of this implementation is how the register pressure of an algorithm may affect performance. As more registers than the maximum allowed by the compiler configuration or available in the architecture are

 ISBN: 978-84-614-6167-7

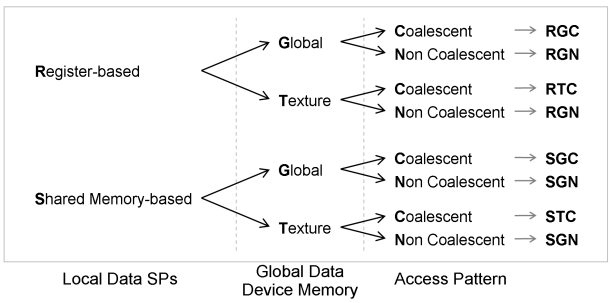


Figure 3: Test configuration

required, local memory is allocated. Local memory is private in the scope of each thread and offers quite poor performance on *Tesla* because at the hardware level is implemented using normal uncached global memory, but on the *Fermi* architecture the access is cached.

In the shared memory implementation (S), to avoid bank conflicts the data is stored in the shared memory using one element of padding between consecutive signals. This slightly increases the amount of shared memory required by the test from $N \times T_b$ to $(N+1) \times T_b$ elements, but it is far more efficient.

Storage type for input data is also analyzed. There are two different memory spaces accessible by all the *SPs*: Global (G) and Texture (T) memory. Texture memory can only be used for reading data, but the texture cache is specially efficient if there are redundant read operations or there is spatial coherency in the access pattern.

On the other hand, the impact of the memory access pattern and coalescence is studied using two different data distributions. Coalescent memory access is used to group several global memory requests in a single one, thus reducing effective bandwidth usage and also pressure in the memory controller, that will receive less requests. In a coalescent access pattern the tasks within a half-warp access data in the same memory segment, and in a non coalescent access two or more tasks access different segments, so they are not performed simultaneously. Figure 4 shows the two signal distributions used in this work. The first data distribution (see Figure 4(a)) is a coalescent memory access pattern (C), where the data of the input signals is stored sequentially, so each thread t_i is assigned to read $\{x_0^i, x_1^i, \ldots, x_N^i\}$, therefore the data read by the corresponding warp in each iteration is located in the same segment. For example, the first read request of the first warp will be $\{x_0^i, x_0^{i+1}, \ldots, x_0^{i+32}\}$ (elements shaded in Figure 4(a)). The total amount of bytes required by the block will be $NB = T_b \times N \times 8$ bytes per complex value, and the total number of read operations is $NB/128$ bytes per transaction for aligned data. The second data distribution is a non-coalescent pattern (N). Figure 4(b) displays this distribution in which each signal is stored sequentially, so the accesses in the same segment are only $segment\_size/N$. For example, in the case of Figure 4(b) (assuming $N \times batch_x > segment\_size$), the first read performed by

 ISBN: 978-84-614-6167-7

(a) Coalescent (b) Non-coalescent

Figure 4: Memory access patterns

the first warp will be composed by the shaded elements $\{x_0^0, x_0^1, \ldots, x_0^{T_b}\}$, which in principle will originate T_b read requests in different segments. Depending on the data alignment and the hardware *CUDA* capabilities, a non-coalescent access may generate up to $N \times T_b$ different memory requests per block. Both access patterns are important, as sometimes applications may require complex changes in order to prevent non-coalescent data access.

4 Experimental results

All the tests were run in single precision using input signals in the range $N = \{4, 8, 16, 32\}$ and batch execution to perform several *FFTs* each time. The size of the batch depends on the input size and is given by the expression $b = 2^{24}/N$, so as the input signal increases the number of batch executions decreases. All the data resides on the *GPU* device memory at the beginning of each test, so there are no data transfers to *CPU* during the benchmarks to prevent interactions with the study of the memory hierarchy. The performance of the experiments is measured in *GFLOPS* through the commonly used expression: $5Nlog_2N \cdot b \cdot 10^{-9}/t$, were N is the size of the input, b is the number of signals processed in batch mode and t is the time in seconds.

Our test platform is composed by a *Core 2 Duo E8400* processor running at 3.0 *GHz* and 2 *GB DDR3 1333 CL9* memory. Two *GPUs* were used for the tests, the *GeForce 280* (based on the *Tesla* architecture) and the *GeForce 480* (based on the *Fermi* architecture). The software setup is *Windows XP x64* operating system, using *Microsoft Visual C++ 2008* compiler (x64, release profile) and *CUDA 3.0 SDK* with the *260.99 GPU* driver.

4.1 Registers vs Shared memory

The performance of the register implementation (RGC) is compared to the shared memory version (SGC), using coalesced access to global memory. Figure 5 presents the results for both the *GeForce 480* and the *GeForce 280 GPUs*. First, we analyze the impact of the two available cache configurations on the Fermi architecture (48 *KB* L1 + 16 *KB* shared

 ISBN: 978-84-614-6167-7

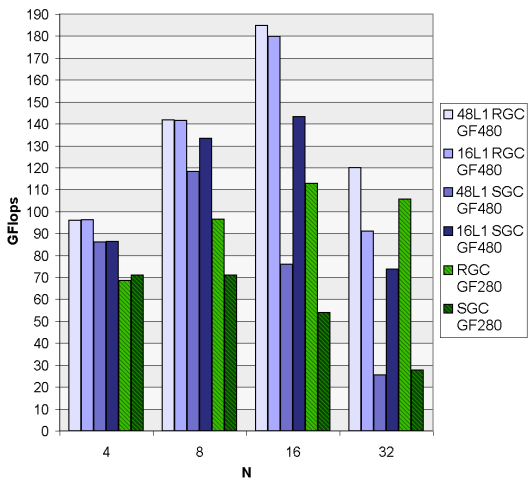


Figure 5: RGC vs SGC

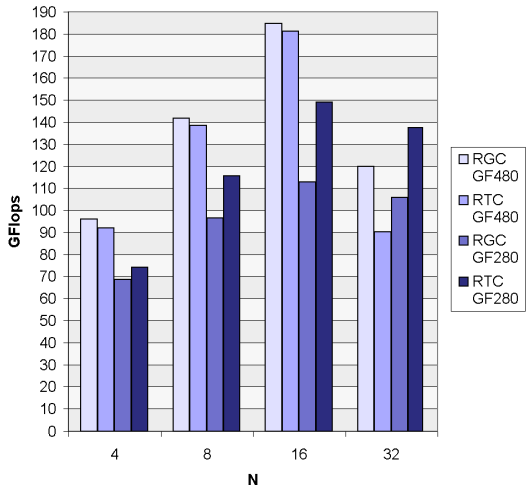


Figure 6: RGC vs RTC

memory vs 16 *KB* L1 + 48 *KB* shared memory).

Regarding the influence of the cache configuration on the *Fermi* architecture, for small problems the difference using *RGC* is not significant, but for $N = 16$ the bigger *L1* cache configuration ($48L1$) offers a bit more performance, while *SGC* loses about a 12% for $N = 8$ and nearly a 50% for $N = 16$. For $N = 32$ this difference increases, using the $48L1$ cache configuration it is possible to improve the result in about a 33% in *RGC*, while losing 65% of the performance for *SGC*. The big difference between cache configurations in this case points to a limitation in the number of simultaneous blocks per *SM*: Each thread requires enough shared memory to fit a whole *FFT*, thus more than 8 *KB* of shared memory are reserved for $N = 32$ and only one block will be executed with 16 *KB* of shared memory, so latency hiding techniques will not work as expected. The most efficient configuration depends on the particular application, programs that were not designed to take advantage of shared memory will probably experience performance gains with the additional *L1*.

Observe that nearly in all the cases (except for $N = 4$ in the *GeForce 280*), the performance of *RGC* is always higher than *SGC* for both architectures (see Figure 5). For example, for $N = 16$ *RGC* achieves 185 *GFlops* on the *GeForce 480*, while *SGC* obtains just 149 *GFlops*. The bandwidth of the shared memory is lower than the register bandwidth, therefore it results in reduced performance for the *SGC* test. The results are quite similar for small N, but if N increases the difference between *SGC* and *RGC* also increases. On the *GeForce 280* for $N = 32$ *SGC* barely reaches a 26% of the *RGC* performance, since allocating such big portions of shared memory for a block will reduce too much the number of simultaneous blocks per *SM*. Also observe how the performance improves for both architectures until $N = 16$ but then decreases for $N = 32$, about a 6% drop for the *GeForce 280*

 ISBN: 978-84-614-6167-7

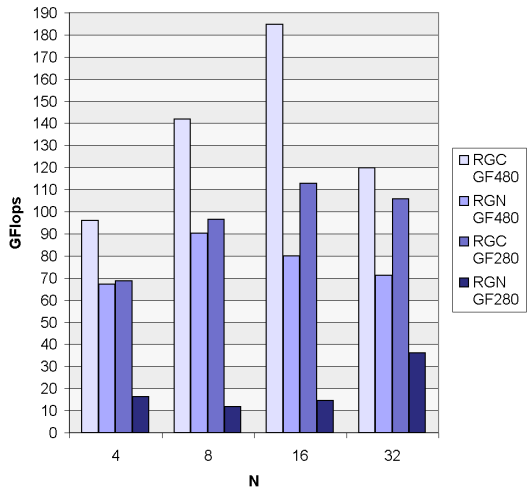


Figure 7: RGC vs RGN

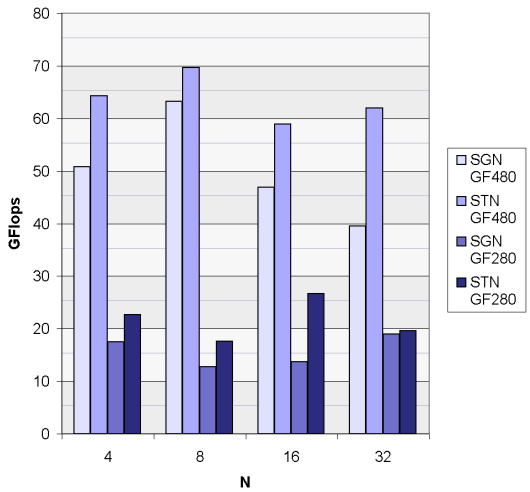


Figure 8: SGN vs STN

and more than a 40% for the *GeForce 480*. The number of operations per thread increases with the size of the problem, therefore the *GPU* can make better usage of the execution resources. However, if a thread has a big working set which does not completely fit in the registers, its content is spilled to the next level in the *GPU* memory hierarchy, the slower *local memory*. Additionally, when kernels require too many registers, less blocks may be simultaneously scheduled in the *GPU*. To avoid the performance degradation experienced for $N \geq 32$ due to the high register pressure, a different implementation where several threads cooperate in the same problem would be more suited.

4.2 Global memory vs Texture memory

The second configuration parameter that will be studied is the impact of the choice between texture memory (*RTC*) and global memory (*RGC*). Only test results using register configuration (R) are shown, as it was the best performing option according to Section 4.1. As seen in Figure 6, the *GeForce 480* experiences a small performance loss using texture memory (under 5% for $N \leq 16$ and a 25% for $N = 32$). According to the documentation, the *L1* data cache of *Fermi* has higher bandwidth than the texture cache. In contrast, for the *GeForce 280* the texture memory provides up to a 32% improvement over the global memory version. This is the normal behavior, as on the *Tesla* architecture the global memory is uncached, so the texture cache can be used to reduce the number of memory fetches. Furthermore, for the particular case of $N = 32$, *RTC* executed in the *GeForce 280* is able to outperform both *RGC* and *RTC* configurations on the *GeForce 480*, thanks to the greater number of registers per thread available in *Tesla*.

 ISBN: 978-84-614-6167-7

4.3 Coalescent memory access vs Non-coalescent

The last factor to study is the impact of the access pattern. In Figure 7 *RGC* and *RGN* test configurations are compared. Changing the global memory access pattern to force a non-coalescent access causes a performance drop in both architectures, albeit more noticeable in the case of the *GeForce 280*, which become about seven times slower for some problem sizes. Even with the advances in the *Fermi* architecture with the cached memory access, there is a big performance impact for non-coalescent memory access and the *GeForce 480* loses approximately half the performance. For example, for $N = 16$ *RGN* is more than 56% slower than *RGC*, which is even lower than the *GeForce 280* in the *RGC* test.

An interesting comparison is the impact of the coalescence when making heavy use of shared memory and trying to minimize the performance cost of the uncoalesced access through texture memory. In this sense, the texture cache can be exploited playing a similar role to the *L1* cache in *Fermi*. Figure 8 compares the performance of *SGN* and *STN*. Notice how in this case (in contrast to Figure 6), texture cache memory improves performance for the *GeForce 480* when configured with just 16 *KB* of *L1* at least in a 10%, because instead of using *L1* to minimize the coalescence problem it is possible to take advantage of the texture cache for this work. For the *GeForce 280* a similar behavior is observed except for $N = 32$, where the dispersion of data and pressure on the texture cache is too big.

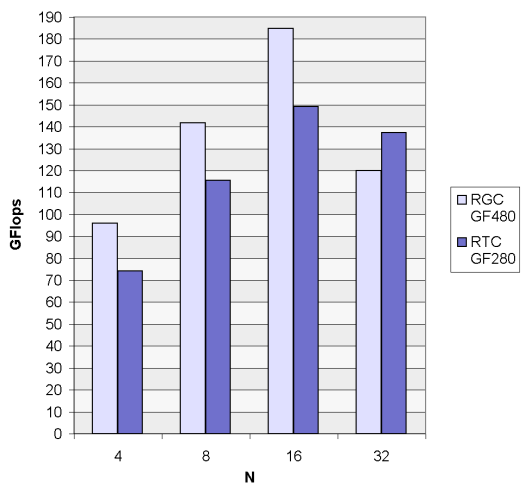


Figure 9: Best performing configurations

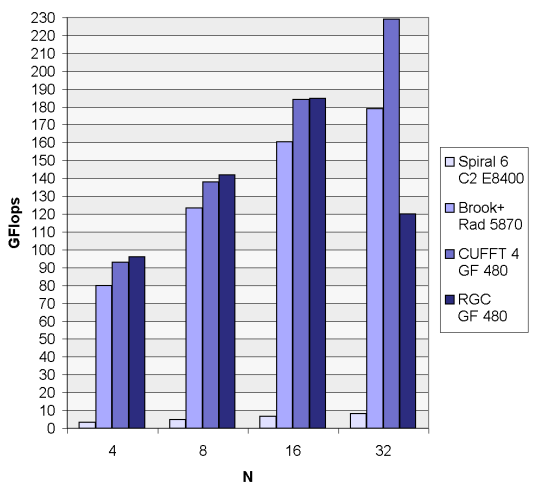


Figure 10: Comparison of different solutions

4.4 Comparison with other state-of-the-art implementations

Figure 9 shows the performance of the best configurations for each tested *GPU*. Observe that up to $N = 16$ the good scaling reveals that the limiting factor is not the computing

 ISBN: 978-84-614-6167-7

power but the effective memory bandwidth, and the typical advantage of the newer *GeForce 480* is about a 24%. For the particular case of $N = 32$ the limiting factor is the on-chip memory and the opposite happens, the *RTC* configuration on the *GeForce 280* offers a 14% advantege over *RGC* in the *GeForce 480*, which needs to spill more data to local memory.

Finally, although the objective of this work was to define a *FFT* focusing on flexibility and programmability instead of performance, observe that the resulting implementation is very competitive for the addressed problem sizes. Figure 10 compares our *RGC* implementation with the *Spiral 6.0* library, the *Brook+ GPU* version presented in [5], and the *CUFFT 4.0*. As can be observed, *GPU* based solutions offer a clear advantage over *CPU* implementations (in this case *Spiral*, but other solutions like the *Intel IPP* library [3] offer very similar performance), resulting at least twelve times faster. Furthermore, the *FFT* proposed in this work is slightly faster than the *CUFFT* for small signals up to 16 elements.

5 Conclusions

In order to keep scaling further, *GPU* memory hierarchy is becoming more complex. Currently, there are no standard benchmarks for *GPU* architectures, and the utilization of real applications provides valuable information. The *FFT* is a very important operation part of many applications, and in this work it was used as a tool to study the different memory types of the *Tesla* and *Fermi* architectures. Our analysis was focused in three configuration parameters: local data *SP*s, global data device memory, and access pattern. It was observed that, when applicable, the performance of the register based solutions is better than shared memory due to the higher register bandwidth. The use of texture cache was also studied, resulting in better performance on *Tesla* but lower performance on *Fermi*, except when only 16 *KB* of *L*1 were selected. It also was proved how the access pattern can have a huge performance impact, even considering the global memory cache in the *Fermi* architecture.

Acknowledgments

This work was supported by the Xunta de Galicia under projects 08TIC001206PR and INCITE08PXIB105161PR, the Ministry of Science and Innovation, cofunded by the FEDER funds of the European Union under the grant TIN2010-16735, and the Consolidation of Competitive Research Groups ref. 2010/06.

References

[1] A. Nukada and S. Matsuoka. Auto-tuning 3-D FFT Library for CUDA GPUs. In *SC '09: Proc. of the Conference on High Performance Computing Networking, Storage and Analysis* (2009), pp. 1–10.

 ISBN: 978-84-614-6167-7

[2] H. Wong, M.-M. Papadopoulou, M. Sadooghi-Alvandi, A. Moshovos. Demystifying GPU Microarchitecture through Microbenchmarking. In *2010 IEEE Int. Symp. on Performance Analysis of Systems Software (ISPASS)* (2010), pp. 235–246.

[3] Intel. *Intel Integrated Performance Primitives for Intel Architecture, Reference Manual*, 2009. Volume 1: Signal Processing.

[4] Intel. *Intel Math Kernel Library, Reference Manual*, 2009. v10.2.

[5] J. Lobeiras, M. Amor and R. Doallo. FFT implementation on a streaming architecture. In *PDP '11: Proc. of the 19th Euromicro Conference On Parallel, Distributed and Network-based Processing* (2011), IEEE Computer Society, pp. 381–388.

[6] J.W. Choi, A. Singh and R.W. Vuduc. Model-Driven Autotuning of Sparse Matrix-Vector Multiply on GPUs. In *Proc. of the 15th ACM SIGPLAN symposium on Principles and practice of parallel programming (PPoPP 2010)* (2010), vol. 45, pp. 115–126.

[7] J.W. Cooley and J.W. Tukey. An Algorithm for the Machine Calculation of Complex Fourier Series. *Mathematics of Computation 19*, 90 (1965), 297–301.

[8] Khronos OpenCL Working Group. *The OpenCL Specification*, 2009.

[9] M. Frigo and S. G. Johnson. The Design and Implementation of FFTW3. *Proceedings of the IEEE 93*, 2 (2005), 216–231.

[10] M. Püschel et al. SPIRAL: Code Generation for DSP Transforms. *Proc. of the IEEE, special issue on "Program Generation, Optimization, and Platform Adaptation" 93*, 2 (2005), 232–275.

[11] M.C. Pease. An Adaptation of the Fast Fourier Transform for Parallel Processing. *Journal of the Association for Computing Machinery (ACM) 15*, 2 (1968), 252–264.

[12] NVIDIA. *CUDA Compute Unified Device Architecture*, 2010. v3.0.

[13] R. Taylor, X. L. A Micro-benchmark Suite for AMD GPUs. In *2010 39th International Conference on Parallel Processing Workshops (ICPPW)* (2010), pp. 387–396.

[14] S. Hong and H. Kim. An Analytical Model for a GPU Architecture with Memory-Level and Thread-Level Parallelism Awareness. In *Proc. of the 36th Int. Symposium on Computer Architecture (ISCA '09)* (2009).

[15] S.S. Baghsorkhi et al. An Adaptive Performance Modeling Tool for GPU Architectures. In *Proceedings of the 15 th ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming (PPoPP 2010).* (2010), pp. 105–114.

[16] Y. Zhang and J.D. Owens. A Quantitative Performance Analysis Model for GPU Architectures. In *Proc. of the 17th IEEE Int. Symposium on High-Performance Computer Architecture (HPCA 17)* (2011).

 ISBN: 978-84-614-6167-7

Proceedings of the 11th International Conference on Computational and Mathematical Methods in Science and Engineering, CMMSE 2011 26–30 June 2011.

A consistent second order theory on the self-gravitatory potential in the equilibrium figures of deformable celestial bodies

José A. López Ortí[1], Manuel Forner Gumbau[1] and Miguel Barreda Rochera[1]

[1] *Departamento de Matemáticas, Universidad Jaume I de Castellón*

emails: lopez@mat.uji.es, fornerm@mat.uji.es, barreda@mat.uji.es

Abstract

The main aim of this paper is the construction of a self-consistent second order theory about the potential of the equilibrium configuration of celestial deformable extended bodies. The classical Clairaut method to solve this problem involves convergence problems in a point of a specific layer arround the equipontial surface that contains the referred point. To solve this problem a method based on the analytical expansions has been developed. This new theory is applied to a particular case of the equilibrium configurations in close binary systems

Key words: Celestial Mechanics, Perturbation Theory, Potential Theory, Computational Algebra.
MSC 2000: 70F15, 76E07.

1 Introduction

The main aim of this paper is the study of the equilibrium configurations of the components of a close binary system. This equilibrium state implies that the system is rotating like a rigid body [1],[5]. This condition is given when each component of the system reach the hydrostatic equilibrium.

$$d\,P = \rho\, d\,\Psi \tag{1}$$

where P is the pressure, ρ the density and Psi the total potential.

To study a component of a close binary system it is convenient to use a coordinate system $OXYX$ defined as: O is the centre of mass of the component, the axe OX is running from the centre of mass of the system, the axe OZ is parallel to angular rate

 ISBN: 978-84-614-6167-7

$\overrightarrow{\omega}$ and the axe OY is chosen in order to define a direct thrihedron $OXYZ$. In this system the total potential Ψ is given by

$$\Psi = \Omega + V_c + V_t \tag{2}$$

where Ω is the self-gravitatory potential due to the primary component, V_c is the centrifugal potential due to the rotation, and V_t is the tidal potential due to the other component of the system [7].

To integrate these equations may be completed with an equation of state $P = P(\rho)$ which connect pressure and density. The hydrostatic equilibrium condition implies the coincidence between the equipotential, the isobaric and the isopycnic surfaces. To solve this problem, the Clairaut method [4] will be applied.

Let be (r, θ, λ) the spherical coordinates defined through $x = r\cos\lambda\cos\theta$, $y = r\sin\lambda\cos\theta$, $z = r\sin\theta$, and let be (r, θ, λ) the spherical coordinates of a point in the primary component. The potential in the referred point due to the primary component is given by

$$\Omega = G\int_D \frac{dm'}{\Delta} \tag{3}$$

where (r', θ', λ') are the coordinates of an arbitrary point in the primary component, $dm' = \rho r'^2\cos\theta' dr' d\theta' d\lambda'$ the element of mass of that point, and C the space region occupied by the primary component. The inverse of the distance can be written as $\Delta^2 = r^2 + r'^2 - 2r\,r'\,cos\gamma$, where γ is the angle between the vector radius $\overrightarrow{r}$, and $\overrightarrow{r'}$. The inverse of the distance can be given by

$$\frac{1}{\Delta} = \begin{cases} \frac{1}{r'}\sum\limits_{n=0}^{\infty}\left(\frac{r}{r'}\right)^n P_n(\cos\gamma) & r < r' \\ \frac{1}{r}\sum\limits_{n=0}^{\infty}\left(\frac{r'}{r}\right)^n P_n(\cos\gamma) & r' < r \end{cases} \tag{4}$$

and from this equation

$$\Omega = \sum_{n=0}^{\infty} U_n r^n + \sum_{n=0}^{\infty} V_n r^{-n} \tag{5}$$

where

$$\begin{aligned} U_n &= G\int_r^{r_1}\int_{-\frac{\pi}{2}}^{\frac{\pi}{2}}\int_0^{2\pi} \rho r'^{1-n} P_n(\cos\gamma)\cos\theta' dr\, d\theta'\, d\lambda' \\ V_n &= G\int_0^{r}\int_{-\frac{\pi}{2}}^{\frac{\pi}{2}}\int_0^{2\pi} \rho r'^{n+2} P_n(\cos\gamma)\cos\theta' dr\, d\theta'\, d\lambda' \end{aligned} \tag{6}$$

The equipotential surfaces can be parametrized as $r = r(a, \theta, \lambda)$ where each of these surfaces is determinated by a value of a and the total potential is given by $\Psi = \Psi(a)$. In general, the radius r of the equipotential surface of the parameter a can be developed as

$$r = a\left(1 + \sum_{n=0}^{\infty}\sum_{m=-n}^{n} f_{n,m}(a) Y_{n,m}(\theta, \lambda)\right) \tag{7}$$

 ISBN: 978-84-614-6167-7

where $f_{n,m}(a)$ are called amplitude functions, and $Y_{n,m}(\theta,\lambda)$ the spherical functions [2]. For symmetry reasons, $F_{n,m}(a)=0$ if $m<0$ or $n+m$ is even.

Classical methods [4], [5], [6] assume that the equations (6) can be rewritten according to the Clairaut coordinates (a,θ,λ) as

$$\begin{aligned} U_n &= \frac{G}{2-n}\int_a^{a_1}\rho\,\frac{\partial}{\partial a'}\int_0^{2\pi}\int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} r'^{2-n}\,P_n(\cos\gamma)\,\cos\theta'\,d\theta'\,d\lambda'\,da' \;\text{ if }\; n\neq 2\\ U_2 &= G\int_a^{a_1}\rho\,\frac{\partial}{\partial a'}\int_0^{2\pi}\int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \ln r'\,P_2(\cos\gamma)\,\cos\theta'\,d\theta'\,d\lambda'\,da'\\ V_n &= \frac{G}{n+3}\int_0^{a}\rho\,\frac{\partial}{\partial a'}\int_0^{2\pi}\int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} r'^{n+3}\,P_n(\cos\gamma)\,\cos\theta'\,d\theta'\,d\lambda'\,da' \end{aligned} \tag{8}$$

These methods assume the desideratum of Laplace [3],[9] about the convergence series.

2 Analytical method

In this section a new method to solve the proposed problem is presented; this new method does not require the use of the Laplace conjecture. Let S the equipotential surface that contains the point of Clairaut coordinates (a,θ,λ). The self-gravitational potential Ω can be obtained as

$$\Omega = G\int_0^{a}\int_{-\frac{\pi}{2}}^{\frac{\pi}{2}}\int_0^{2\pi}\frac{1}{\Delta}dm' + G\int_a^{a_1}\int_{-\frac{\pi}{2}}^{\frac{\pi}{2}}\int_0^{2\pi}\frac{1}{\Delta}dm' \tag{9}$$

Let us define $\Sigma=\sum_{n=0}^{\infty}\sum_{m=0}^{\infty} f_{n,m}(a)Y_{n,m}(\theta,\lambda)$ and so, $r=a+a\Sigma$. Classical method involves convergence problems in the development of $\frac{1}{\Delta}$ in the region $r'\in[m,M]$, where m, M are the minimum and the maximum values of the set $\{r'(a',\theta,\lambda),\theta\in[-\frac{\pi}{2},\frac{\pi}{2}],\lambda\in[0,2\pi]\}$.

Let us define $T(a,a',\cos\gamma)$ [8] as

$$T(a,a',\cos\gamma)=\frac{1}{\sqrt{a^2+a'^2-2\,a\,a'\cos\gamma}} \tag{10}$$

the inverse of the distance $\frac{1}{\Delta}$ can be approach up second orden in amplitudes by

$$\frac{1}{\Delta}=T(a,a')+T_a(a,a')a\Sigma+T_{a'}a'\Sigma'+T_{a,a}a^2\Sigma^2+2\,T_{a,a'}\,a\,a'\Sigma\Sigma'+T_{a',a'}a'^2\Sigma'^2 \tag{11}$$

where subindex a, a' denote partial derivative. To evaluate the partial derivatives it is necessary to use the developments

$$T(a,a',\cos\gamma)=\begin{cases}\frac{1}{a'}\sum\limits_{n=0}^{\infty}\left(\frac{a}{a'}\right)^n P_n(\cos\gamma) & a<a'\\ \frac{1}{a}\sum\limits_{n=0}^{\infty}\left(\frac{a'}{a}\right)^n P_n(\cos\gamma) & a'<a\end{cases} \tag{12}$$

 ISBN: 978-84-614-6167-7

Replacing the developments of Σ and Σ' up to second order in amplitudes, and replacing the products of spherical functions by their developements, according to these ones, by integration we obtain the development

$$\Omega = \sum_{n=0}^{\infty} \sum_{m=0}^{n} \Omega_{n,m}(a) Y_{n,m}(\Theta, \lambda) \tag{13}$$

3 Concluding remarks

The method presented here allows us to obtain the value of the autogravitatory potential up to second order in amplitudes without using the Laplace conjecture. This method is suitable to be extended up to third of higher order.

Acknowledgements

This research has been partially supported by Grant GV/2009/027 from the Generalitat Valenciana and Grant P1-06I455.01/1 from Fundación Caja Castellón (Bancaja).

References

[1] E. Finlay-Frendulich, *Celestial Mechanics*, Pergamon Press Inc., New York 1958.

[2] E. W. Hobson *Theory of spherical and elliptical harmonics*. New York: Chelsea, 1955.

[3] Jardetzky, W.,*Theories of figures of celestial bodies*, Interscience Publishers, Inc., New York, 1958.

[4] Kopal, Z., *Figures of Equilibrium of Celestial Bodies*, Univ Wisconsin Press, Madison, 1960.

[5] Kopal, Z., *Dynamics of close binary systems*, Kluwer, Dordrecht, Holland (D. Reidel Publishing Company), 1978.

[6] López Ortí, J.A.; López García, A. y López Machí, R., *Figures of equilibrium in close binary systems*, Celestial Mechanics and Dynamical Astronomy **53**, pp 311-322, 1992.

[7] López Ortí, J.A.; Forner Gumbau, M. y Barreda Rochera, M., *A method to improve the computation of tidal potential in the equilibrium configuration of a close binary system*, International Journal of Computer Mathematics **86**, pp 1831-1840, 2009.

 ISBN: 978-84-614-6167-7

[8] 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. First published on 11 March 2011. DOI: 10.1080/00207160.2010.521817

[9] Tisserand, F.F., *Traité de Mécanique Celeste. Tome II pp 317*, Gautier-Vilar, París, 1889.

 ISBN: 978-84-614-6167-7

Proceedings of the 11th International Conference on Computational and Mathematical Methods in Science and Engineering, CMMSE 2011 26–30 June 2011.

A parallel solver using the Fast Multipole Method for noise problems

M. López-Portugués[1], Jesús A. López-Fernández[1], José Ranilla[2], Rafael G. Ayestarán[1] and Fernando Las-Heras[1]

[1] *Departamento de Ingeniería Eléctrica, Electrónica, de Computadores y de Sistemas, Universidad de Oviedo. Spain.*

[2] *Departamento de Informática, Universidad de Oviedo. Spain.*

emails: mlopez@tsc.uniovi.es, jelofer@tsc.uniovi.es, ranilla@uniovi.es, rayestaran@tsc.uniovi.es, flasheras@tsc.uniovi.es

Abstract

In this work, we present a heterogeneous solver of the *Fast Multipole Method* (FMM) applied to acoustic scattering. The developed tool supports multiGPU configurations. The GPU deals with compute-bound parts of the FMM meanwhile the CPU tackles the memory-bound part. We have evaluated the accuracy of the developed approach using a direct solver as a reference. Finally, the performance of the implemented tool is measured on a workstation with 2 NVIDIA GTX 480 GPUs. Time to solution is reduced an order of magnitude compared to our optimized parallel CPU solver.

Key words: Heterogeneous, GPGPU, BEM, GMRES, Fast Multipole Method, acoustic scattering.

1 Introduction

The design of silent aircraft configurations to reduce the environmental noise [1] stimulates the implementation of accurate and efficient mathematical models that simulate the acoustical behavior of the system.

The *Boundary Elements Method* (BEM) [2] is low-frequency method that provides a precise numerical formulation of acoustic scattering problems. Nevertheless, the solution of the linear system produced by the BEM may be very expensive from a computational viewpoint. The direct solution of the above-mentioned linear system, with N equations and N unknowns, has a time cost $O\left(N^3\right)$ and a memory cost $O\left(N^2\right)$. The use of iterative solvers reduces the time cost to $O\left(N^2\right)$ per iteration (a *Matrix-Vector Product* –MVP– is computed in each iteration). The computation of the whole

 ISBN: 978-84-614-6167-7

system matrix may be avoided by the use of the *Fast Multipole Method* (FMM) [3, 4] which yields a dramatic reduction in the iteration time without significantly affecting the solution accuracy.

Nowadays, one of the most outstanding trends in *High Performance Computing* is the spreading of heterogeneous computing. For the last years, researchers involved in scientific and technical computing have been trying to improve the performance of computational simulations by means of using specialized processors like graphics processing units (GPUs). In this manner, initiatives like general-purpose computation on graphics hardware (GPGPU [5]) are getting closer to the mainstream. Since modern GPUs have been demonstrated to be suitable for problems with large computational requirements that are prone to parallelism [6], acoustic scattering problems seem a good field to put into practice GPGPU.

In this work, we present a heterogeneous acoustic scattering solver with multiGPU support that enhances the solver proposed by the authors in [7]. BEM is used to model numerically the physical problem. In order to iteratively solve the linear system of equations posed by the BEM, we have chosen the *Generalized Minimum Residual* (GMRES) method [8] due to its robustness for acoustic scattering [9]. In addition, we speed up the GMRES solution by means of the FMM.

2 Physical problem and the Fast Multipole Method

In this work, we try to predict the acoustic pressure on the space surrounding a 3-D obstacle impinged by an acoustic wave. We use the integral form of the Helmholtz equation [2], also known as *Conventional Boundary Integral Equation* (CBIE), combined with its normal derivative [10] to model the acoustic problem. This formulation is commonly known as the Burton and Miller equation [10] and overcomes the non-uniqueness difficulty [10] that appears in the CBIE at resonant frequencies. In addition we use the BEM to discretise, over the obstacle surface S, the Burton and Miller equation into N facets. In this manner, the problem is formulated in terms of a linear system of N equations. The size of the problem, N, is proportional to the acoustic size of S and, therefore, it increases with the frequency squared $\left(f^2\right)$.

In order to efficiently solve the linear system of equations posed by the BEM, we use the GMRES method [8] in addition to the FMM that speeds up the computation of the GMRES iterations. In the setup step, the FMM makes N_{g} groups from the N facets (elements) of the problem. On the one hand, the interactions between elements that pertain to non-neighboring groups, *far interactions*, are efficiently calculated by means of the *Addition Theorems* and *plane wave decomposition* [11] performing three operations: *aggregation*, *translation*, and *disaggregation*. On the other hand, the interactions between elements that pertain to nearby groups, those that share at least a border point, must be computed directly [3].

 ISBN: 978-84-614-6167-7

3 Heterogeneous FMM solver with multiGPU support

Since three of the FMM steps (near interactions, aggregation, and dissagregation) are compute bound, they are the best candidates to be performed using GPUs. On the contrary, the remaining step (translation) is memory bound due to the translation operator size, so it is most suitable to be performed in the CPU. We have proved that this approach delivers a better performance than the strategy followed in [7]. To fully exploit NVIDIA[1] latest generation GPUs (Fermi), the *Compute Unified Device Architecture* (CUDA) API is used to tackle the computation of the accelerated steps.

One of the most time-consuming parts of the FMM is the computation of the near interactions. To perform this step, we have developed a kernel with a fine grain approach. In this manner, each CUDA thread tackles the near interactions of a single element, in a cyclic manner, until all the contributions are calculated. When compiling the kernel with the parameter "-arch=sm_20" to match the Fermi architecture, the nvcc compiler reports a register file usage of 63 registers per thread. Moreover, since the memory accesses are mostly data dependent, we configure the kernel with L1 cache preference, that is, 48 KB of L1 cache. Finally, by means of several experimental tests, we have verified that this kernel achieves the best performance when using 1024 blocks with 256 threads per block.

Just like the near interactions, the translation is also very expensive in terms of runtime. Although we have decided to keep the translation in CPU, our approach delivers a significant performance gain due to computing overlapping. The translation is performed in CPU, whereas the calculation of the near interactions is performed in GPU. Moreover, this step is related to the far field contributions, whereas the near interactions are related to the near field contributions. Thus, the calculation of both translation and near interactions is totally independent. As a consequence, both steps may be done in parallel, at the same time, allowing an efficient use of the available resources.

In order to tackle the aggregation, we have developed a kernel that also operates at a fine grain level. In this case, each thread tackles the operations related to one single direction, in a cyclic manner, until all the aggregated contributions (each direction for every group) are computed. When compiling the kernel for devices with computing capability 2.0, the compiler shows a register file usage of 51 registers per thread. In this FMM step, the memory accesses are also data dependent, so L1 cache preference configuration yields best performance. Thanks to experimentation, we have checked that using 1024 blocks and 192 threads per block leads to best runtime in this kernel.

The last FMM step is the disaggregation which is performed in GPU. The kernel developed to deal with this step also uses a fine grain approach, so each CUDA thread tackles the disaggregation of a single element at a time, in a cyclic manner, until all the far interactions are calculated. With regard to resource use, the register file usage is 42 registers per thread when compiling for Fermi type devices. Furthermore, this kernel also benefits from the L1 cache preference. Finally, taking into account the conducted

[1] http://www.nvidia.com/

 ISBN: 978-84-614-6167-7

tests, we have verified that this kernel yields the best performance when using 1024 blocks with 128 threads per block.

With the aim of reducing even further the runtime, we have decided to add support for multiple GPUs, since it is useful for the solution of real problems (usually with millions of unknowns). In order to deal with multiple GPUs, the outermost loop in each kernel is divided into disjoint subsets of adjoining items (elements or directions), using so many subsets as used GPUs. In this manner, each GPU tackles its part of the workload without dependence on data stored in the rest of GPUs, avoiding costly communications between host and devices.

4 Validation and computational results

The results shown in this work have been obtained by using a workstation with GPGPU capabilities. This workstation consists of an Intel Core i7 930 CPU (4 cores at 2.8 GHz), 12 GB of DDR3 RAM, and 2 NVIDIA GTX 480 GPUs. Each GTX 480 has 480 cores running at 1.4 GHz and 1.5 GB of GDDR5 RAM. On the software side, Intel icc 11.1 and NVIDIA nvcc 3.1 have been used to compile the source code for CPU and GPU, respectively. Since the workstation CPU has 4 cores with *Hyper-Threading*, we have used 8 threads to perform the computation that takes place in the CPU. It is also worth noting that single-precision arithmetic is used throughout our codes.

In order to demonstrate the correctness of our heterogeneous implementation, we have compared some results against a direct solution implementation. Since the direct solution is prohibitive in terms of time, $O(N^3)$, and memory, $O(N^2)$, the chosen problem has a moderate size. The problem entails the analysis of a 2 m diameter sphere with an impinging plane wave at 950 Hz ($N = 10\,106$ elements). The accuracy of the implementation is evaluated taking into account different values for the residual error (ϵ), which is defined as follows:

$$\epsilon = \frac{\left\| \bar{\bar{K}}\bar{p} - \bar{g} \right\|_2}{\|\bar{g}\|_2}, \tag{1}$$

where $\|\cdot\|_2$ is de euclidean norm (norm-2), $\bar{\bar{K}}$ is the system matrix (stiffness matrix), $\bar{g}$ is the excitation vector (related to the incident pressure), and $\bar{p}$ is the final iterate (solution pressure). Both the relative error (η) and the *Root Mean Square Error* (RMSE) (σ) are used to measure the difference between our heterogeneous FMM implementation and the direct solution:

$$\eta = \frac{\left\| \bar{p}^{(d)} - \bar{p}^{(f)} \right\|_2}{\left\| \bar{p}^{(d)} \right\|_2}, \qquad \sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N} \left| \bar{p}_i^{(d)} - \bar{p}_i^{(f)} \right|^2}, \tag{2}$$

where N is the number of elements or unknowns, $\bar{p}^{(d)}$ is the solution pressure obtained with the direct solution, and $\bar{p}^{(f)}$ is the solution pressure obtained by using our FMM implementation.

 ISBN: 978-84-614-6167-7

Problem size	Stop condition	# of iterations	ϵ	η	σ
$N = 10\,106$	$\epsilon \leq 10^{-2}$	5	$7.7 \cdot 10^{-3}$	$9.6 \cdot 10^{-3}$	$1.3 \cdot 10^{-2}$
	$\epsilon \leq 10^{-3}$	9	$7.1 \cdot 10^{-4}$	$8.6 \cdot 10^{-4}$	$1.2 \cdot 10^{-3}$
	$\epsilon \leq 10^{-4}$	13	$6.3 \cdot 10^{-5}$	$7.7 \cdot 10^{-5}$	$1.1 \cdot 10^{-4}$
	$\epsilon \leq 10^{-5}$	16	$9.8 \cdot 10^{-6}$	$5.5 \cdot 10^{-5}$	$7.7 \cdot 10^{-5}$

Table 1: Accuracy of the heterogeneous implementation of the FMM vs direct solution.

In table 1, some results related to the accuracy of the proposed solution are shown. The achieved accuracy is enough when the target residual error is up to 10^{-4}, which is usually the case. If a residual error less or equal than 10^{-5} is required then single-precision arithmetic seems insufficient. In such cases, double-precision arithmetic should be more adequate.

In order to show the computational achievements, two different 3D objects (scatterers) have been analysed. The first object is a real size model of an aircraft (Airbus A300 series). This scatterer has been used to obtain the distribution of the acoustic pressure at two different frequencies, 500 Hz and 1000 Hz. Thus, two different triangular meshes have been generated to model the aircraft (each triangle corresponds to one unknown of the problem). One consists of 533 176 triangles (500 Hz), and the other one consists of 2 132 704 triangles (1000 Hz). The second object is a 2 m diameter sphere, which has been also analysed at two different frequencies, 4 kHz and 8 kHz. Therefore, two different meshes must be used to model the sphere, one mesh of 750 004 triangles (4 kHz) and another one of 2 999 428 triangles (8 kHz).

	A300 ($N = 533\,176$)		A300 ($N = 2\,132\,704$)	
System	Iteration	Total	Iteration	Total
1 CPU	54.5	5 347.0	401.3	64 201.6
1 CPU + 1 GPU	4.4	440.8	32.3	5 218.7
1 CPU + 2 GPU	3.0	304.7	23.0	3 724.1

Table 2: A300 model analysis runtime, per iteration and total (in seconds).

The tables 2 and 3 are used for comparing the runtime of a parallel CPU implementation to the runtime of the heterogeneous solver presented in here. It is remarkable that, in all the test cases, the use of a GPU delivers an order-of-magnitud performance gain. The best case (A300 with 2 132 704 elements) achieves a 12.3x speedup over the CPU, whereas the worst case (sphere with 750 004 elements) results in a 11.2x speedup over the quad-core processor. It is also worth noting the performance gains over the CPU when using two GPUs, that vary between 14.5x and 17.5x.

 ISBN: 978-84-614-6167-7

	Sphere ($N = 750\,004$)		Sphere ($N = 2\,999\,428$)	
System	Iteration	Total	Iteration	Total
1 CPU	87.0	1 125.8	641.5	9 634.5
1 CPU + 1 GPU	6.8	100.2	50.0	815.7
1 CPU + 2 GPU	5.1	77.8	37.1	615.7

Table 3: Sphere analysis runtime, per iteration and total (in seconds).

5 Conclusions

In this work, we present a heterogeneous solver based on FMM for acoustic scattering problems. We use the CUDA API to take advantage of latest NVIDIA GPUs (Fermi). The most outstanding result is the runtime reduction of an order of magnitude when comparing a parallel CPU solver to the heterogeneous one presented in here. Furthermore, when using two GPUs, we achieve a speedup around 15x compared to a quad-core CPU with Hyper-Threading (8 simultaneous threads). Thus, the solution developed may be considered a useful engineering tool for noise control applications.

Acknowledgements

This work has been supported by the "Ministerio de Ciencia e Innovación" from Spain/FEDER —under the research projects TEC2008-01638/TEC (INVEMTA) and TIN2010-14971—, and by "Cátedra Telefónica - Universidad de Oviedo". Financial support (grant: UNOV-10-BECDOC) given by the University of Oviedo is also acknowledged. The Airbus A300 series geometry has been provided by the research project GRD1-2001-40147 financed by the European Union.

References

[1] Advisory Council for Aeronautics Research in Europe, *2008 Addendum to the Strategic Research Agenda*, available online at: `http://www.acare4europe.org/` (2008).

[2] T. W. Wu, *Boundary Element Acoustics*, Advances in Boundary Elements, WIT Press, 2000.

[3] V. Rokhlin, *Diagonal Forms of Translation Operators for the Helmholtz Equation in Three Dimensions*, Applied and Computational Harmonic Analysis, **vol. 1(1)** (1993) 82–93.

 ISBN: 978-84-614-6167-7

[4] J. Song and W. Chew, *Multilevel Fast-Multipole Algorithm for Solving Combined Field Integral Equations of Electromagnetic Scattering*, Microwave and Optical Technology Letters **10(1)** (1995) 14–19.

[5] GPGPU.org, *General-Purpose computation on Graphics Processing Units*, available online at: `http://gpgpu.org/` (2011).

[6] J.D. Owens, M. Houston, D. Luebke, S. Green, J.E. Stone and J.C. Phillips, *GPU Computing*, Proceedings of the IEEE **96(5)** (2008) 879–899.

[7] M. López-Portugués, J.A. López-Fernández, A. Rodríguez-Campa and J. Ranilla, *A GPGPU Solution of the FMM Near Interactions for Acoustic Scattering Problems*, Journal of Supercomputing (DOI: 10.1007/s11227-011-0584-6).

[8] Y. Saad and M. H. Schultz, *GMRES: A Generalized Minimal Residual Algorithm for Solving Nonsymmetric Linear Systems*, SIAM J. of Sci. and Statist. Comput. **7** (1986) 856–869.

[9] S. Marburg and S. Schneider, *Performance of iterative solvers for acoustic problems. Part I. Solvers and effect of diagonal preconditioning*, Engineering Analysis with Boundary Elements **27(7)** (2003) 727–750.

[10] A. J. Burton and G. F. Miller, *The Application of Integral Equation Methods to the Numerical Solution of Some Exterior Boundary-Value Problems*, Proc. of the Royal Society of London **323(1553)** (1971) 201–210.

[11] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions*, Dover Publications, 1972.

 ISBN: 978-84-614-6167-7

Proceedings of the 11th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2011
26–30 June 2011.

Non-linear harmonic modelling of geocenter variations caused by continental water flux

Pedro A. Martínez-Ortiz[1] **and J. M. Ferrándiz**[1]

[1] *Department of Applied Mathematics, University of Alicante*

emails: pedroa.martinez@ua.es, jm.ferrandiz@ua.es

Abstract

It is known that continental water storage and flux are not steady, regular processes. Their variations are not uniform either in time or in space and they affect climate, weather, land and the position of the geocenter, among other phenomena. By using the *Land Data Assimilation System* (LDAS) model, we obtain a multidimensional time series that represent the monthly variations of the geocenter position caused by the continental water storage changes. After that, a non-linear harmonic analysis is performed over the three one-dimensional time series in order to understand the periodic behavior of the phenomenon. Finally, some conclusions are derived and we suggest candidate causes of the changes in the geocenter variations detected.

Key words: Non-linear harmonic analysis, continental water flux, time series analysis, geocenter variations.

1 Introduction

The *geocenter* is defined as the center of mass of the Earth system, including the solid earth, oceans, and atmosphere [6]. In a reference frame linked to the solid Earth and defined by a set of mean geodetic station coordinates, the center of mass of the Earth system moves because of mass redistribution inside the system [5].

On timescales ranging from intraseasonal to interannual, this mass redistribution mainly results from fluid redistribution within and among atmosphere, oceans, continental water reservoirs and ice sheets [5]. The question that needs to be answered now is to what extent this redistribution of mass can affect the position of the geocenter. Several studies were carried out about this issue [6, 7]. They concluded that the contribution of atmospheric

 ISBN: 978-84-614-6167-7

pressure, ocean mass and continental waters produce an annual geocenter variation whose amplitude is less than 5 millimeters (mm).

In this paper we show the results of the study of the behavior of the geocenter due to the redistribution of the continental water. To carry out this task we have considered the *Land Data Assimilation System* (LDAS) model produced by the *Climate Prediction Center* (CPC) of the *National Oceanic and Atmospheric Administration* (NOAA). On the other hand, the technique used for the analysis is based on the non-linear harmonic (NLH) method developed by W. Harada and T. Fukushima [3], which allows the recursive detection of frequencies and their associated amplitudes and phases as well as the secular mixed Fourier terms when found in the signal.

2 Non-linear harmonic method

Let us consider a time series with $N \in \mathbb{N}$ observations, represented by $\{t_n, d_n\}_{n=1,\ldots,N}$ where d_n is the measurement of the phenomenon we are interested in at epoch t_n. Henceforth, in order to facilitate subsequent calculations, given a time series, we will consider its temporal translation $\{\tau_n, d_n\}_{n=1,\ldots,N}$ where:

$$\tau_n = t_n - \frac{t_1 + t_N}{2} = t_n - t_1 - \frac{T}{2} \tag{1}$$

where $T = t_N - t_1$ is the range of the time domain. The technique, which we will describe briefly, is about fitting a time series using the least-squares method to a function of the form:

$$h_n = \sum_{l=1}^{L} a_l \cdot \varphi_l(t_n) \tag{2}$$

where $L \in \mathbb{N}$ represents the number of basis functions $\{\varphi_l\}_{l=1,2,\ldots,L}$, and $\{a_l\}_{l=1,2,\ldots,L}$ are the linear coefficients to solve for. This set of base functions consists of:

1. *Three polynomial functions* which set up the trend of the series:

$$\varphi_1(\tau) = 1 \tag{3}$$
$$\varphi_2(\tau) = \frac{4\tau}{T} \tag{4}$$
$$\varphi_3(\tau) = \left[\frac{3 \cdot (N-1)}{4 \cdot (N+1)}\right] \cdot \varphi_2^2(\tau) - 1 \tag{5}$$

2. A couple of *Fourier terms* for the angular frequency ω_k:

$$\varphi_{2k+2}(\tau) = \sin(\omega_k \tau) \tag{6}$$
$$\varphi_{2k+3}(\tau) = \cos(\omega_k \tau) \tag{7}$$

 ISBN: 978-84-614-6167-7

3. A couple of the so-called *mixed secular terms* for the frequency ω_{k_i}:

$$\varphi_{2K+2i+2}(\tau) = \tau \cdot \sin(\omega_{k_i}\tau) \quad (8)$$

$$\varphi_{2K+2i+3}(\tau) = \tau \cdot \cos(\omega_{k_i}\tau) \quad (9)$$

for $i = 1, 2, \ldots, S$. This kind of basis functions does not have to appear necessary in the model for each frequency.

First, we assume that a trend (if necessary) and K angular frequencies ω_k are already included in the functional model, the set of them being denoted as a vector $\vec{\omega}^{(K)} = (\omega_1, \omega_2, \ldots, \omega_K)^T$. Let us also consider $W' = \left\{\omega_{k_v} \in \vec{\omega}^{(K)}\right\}_{v=1,\ldots,S}$, the subset of frequencies already associated to mixed secular terms. We search additional frequencies that could be added to the functional model by studying the Lomb periodogram of the residuals obtained from the least squares fit of the temporary functional model.

A criterion is needed to guess which frequencies are linked to mixed secular terms and which not, so as to construct our objective function. The procedure that allows us to elucidate such an association is based on the Lomb periodogram and an extension of it. The algorithm increases the number of frequencies one by one and adds them to the model. First, we have to compute the spectrum of the Lomb periodogram, which is given by the formula:

$$P(\omega) = \frac{\left[\sum_{n=1}^{N} d_n \cdot \sin(\omega\tau_n)\right]^2}{\sum_{n=1}^{N} \sin^2(\omega\tau_n)} + \frac{\left[\sum_{n=1}^{N} d_n \cdot \cos(\omega\tau_n)\right]^2}{\sum_{n=1}^{N} \cos^2(\omega\tau_n)} \quad (10)$$

The peak of this spectrum will point out a significant angular frequency to be included in the model. In the next stage, when the Fourier term of the frequency is already selected as a basis function, we compute the extended periodogram given by the equation:

$$Q(\omega) = \frac{\left[\sum_{n=1}^{N} d_n\tau_n \cdot \sin(\omega\tau_n)\right]^2}{\sum_{n=1}^{N} \tau_n^2 \sin^2(\omega\tau_n)} + \frac{\left[\sum_{n=1}^{N} d_n\tau_n \cdot \cos(\omega\tau_n)\right]^2}{\sum_{n=1}^{N} \tau_n^2 \cos^2(\omega\tau_n)} \quad (11)$$

If this maximum of the extended periodogram is larger than the maximum of the Lomb periodogram, mixed secular terms will be linked to this frequency. In other case, we will only select classic Fourier terms.

As we draw each frequency, an adjustment of the data is made by using the least-squares method just to solve for the value of the linear coefficients $\{a_l\}_{l=1,\ldots,L}$. After each estimation of these coefficients that we denote by $\tilde{a} = \{\tilde{a}_l\}_{l=1,\ldots,L}$ and before adding a new angular frequency, we can regard the objective function of the least square problem for our model as a function in the space of frequencies, that is:

$$\phi(\tilde{a}, \vec{\omega}) = \sum_{n=1}^{N} \left[d_n - \sum_{l=1}^{L} \tilde{a}_l \cdot \varphi_l(\tau_n, \vec{\omega})\right]^2 = \hat{\phi}(\vec{\omega})$$

 ISBN: 978-84-614-6167-7

with $\vec{\omega} \in \mathbb{R}^K$. Therefore, our problem is reduced to a minimization problem $(P1)$ that depends non-linearly of the parameters, namely:

$$\begin{array}{rll} (P1) & Min & \hat{\phi}\left(\vec{\omega}\right) \\ & s.t. & \vec{\omega} \in \mathbb{R}^K \end{array} \tag{12}$$

To carry out this task we use the *Broyden-Fletcher-Goldfarb-Shanno* (BFGS) algorithm [1, 2] where we consider $\vec{\omega}^{(K)}$ as a seed point. Let us denote by $\vec{\omega}_s$ the solution of $(P1)$. Next stage deals with the adjustment of the model (2) to the data by considering $\vec{\omega}_s$ as a vector of frequencies and by using the least squares method again. This cycle recurs until the difference between the seed point and the solution of $(P1)$ becomes small. After that, we can continue extracting another frequency or we can stop the process if the weighted root mean square of the residuals becomes smaller than a fixed level. We can also stop the algorithm by reaching a preset number of frequencies given by the user.

3 Data

As we have already mentioned, we have used the data of the LDAS model to perform the analysis of the variations of the geocenter due to the redistribution of continental water mass. LDAS is forced by observed precipitation, derived from CPC daily and hourly precipitation analysis, downward solar and long-wave radiation, surface pressure, humidity, 2-m temperature and horizontal wind speed from the *National Centers for Environmental Prediction* (NCEP) reanalysis. The output consists of soil temperature and soil moisture in four layers below the ground. At the surface, it includes all components affecting energy and water mass balance, including snow cover, depth, and albedo. The data[1] represent the monthly averaged soil water storage changes (in *centimeters (cm) of equivalent water thickness*). They are provided on a 1x1 degree grid that covers the whole Earth's surface (although no estimate is provided over Antarctica) during the period of time ranging from January, 1948 to December 2007.

We will consider each grid of data as a system of particles, where each one corresponds to a 1x1 degree square and whose mass depends on its area and the amount of water in it. The cartesian coordinates of the geocenter for this system of particles will be given by the

[1] Available at ftp://ftp.crs.utexas.edu/pub/ggfc/water/CPC/

 ISBN: 978-84-614-6167-7

equations (see [6]):

$$X = \frac{R_E}{M_E} \sum_{\phi=\frac{-\pi}{2}}^{\frac{\pi}{2}} \sum_{\lambda=0}^{2\pi} \cos\phi \cos\lambda \cdot L(\phi,\lambda) \cdot \Delta s = R_E \cdot C_{1,1} \tag{13}$$

$$Y = \frac{R_E}{M_E} \sum_{\phi=\frac{-\pi}{2}}^{\frac{\pi}{2}} \sum_{\lambda=0}^{2\pi} \cos\phi \sin\lambda \cdot L(\phi,\lambda) \cdot \Delta s = R_E \cdot S_{1,1} \tag{14}$$

$$Z = \frac{R_E}{M_E} \sum_{\phi=\frac{-\pi}{2}}^{\frac{\pi}{2}} \sum_{\lambda=0}^{2\pi} \sin\phi \cdot L(\phi,\lambda) \cdot \Delta s = R_E \cdot C_{1,0} \tag{15}$$

In these expressions ϕ is the latitude (expressed in radians), λ is the East longitude (expressed in radians), $M_E = 5.9742 \times 10^{27}$ grams is the mass of the Earth, $L(\phi,\lambda)$ represents the water storage (in cm of equivalent water thickness) of the region of the grid that contains the point with coordinates (ϕ,λ), $R_E = 6.371 \times 10^8$ cm is the Earth's mean radius and Δs is the area of the surface linked to $L(\phi,\lambda)$ given by the equation:

$$\Delta s = R_E^2 \cos\phi \cdot \Delta\phi \cdot \Delta\lambda \tag{16}$$

By using the equations (13)–(15) we build the scalar time series that represent the variations of the geocenter in each cartesian coordinate due to the redistribution of continental water mass.

4 Analysis and results

We have performed a non-linear harmonic analysis of each scalar time series, X, Y and Z representing the cartesian coordinates of the center of mass of the continental water system. The time domain considered goes from January 1970 to December 2007 (throughout 456 monthly observations). We extract up to 15 spectral lines for each component and we consider a polynomial linear trend component to be included in the functional model.

After the analysis with the available routines created with MATLAB, we get the following results. The trend component for each coordinate is represented in Figures 1a, 1c and 1e. Their analytical expressions are given by the equations:

$$T_X(\tau_n) = (0.7550 \pm 0.0006) + (-0.0190 \pm 0.0006) \cdot \varphi_2(\tau_n) \tag{17}$$

$$T_Y(\tau_n) = (0.2299 \pm 0.0006) + (0.0056 \pm 0.0005) \cdot \varphi_2(\tau_n) \tag{18}$$

$$T_Z(\tau_n) = (1.3795 \pm 0.0004) + (-0.0181 \pm 0.0004) \cdot \varphi_2(\tau_n) \tag{19}$$

where $\tau_n = t_n - t_c$ is a translation of the time domain centered on t_c that corresponds to mid-January, 1989 and $\varphi_2(\tau_n)$ is the basis function given at equation (4). The estimated coefficients and uncertainties of the trend component are expressed in cm.

 ISBN: 978-84-614-6167-7

Table 1: Fourier terms. Harmonic content for the variations on the X coordinate of the geocenter. Columns refer to the extracting order, frequency (cycles per month), uncertainty of the frequency (cycles per month), period (months) and coefficients (cm) linked to sinus and cosinus terms, respectively.

No.	f	σ_f	Π	S	C
1	0.083429	0.16×10^{-4}	11.99±0.00	0.0644±0.0008	0.0086±0.0008
2	0.009690	0.43×10^{-4}	103.20±0.46	-0.0186±0.0009	0.0162±0.0009
3	0.005511	0.43×10^{-4}	181.44±1.40	0.0087±0.0009	-0.0047±0.0009
4	0.031854	0.48×10^{-4}	31.39±0.05	-0.0013±0.0009	0.0014±0.0009
5	0.023275	0.84×10^{-4}	42.97±0.15	0.0093±0.0009	0.0110±0.0009
6	0.020115	1.03×10^{-4}	49.71±0.25	0.0110±0.0009	0.0010±0.0009
7	0.049815	0.78×10^{-4}	20.07±0.03	0.0056±0.0009	-0.0120±0.0009
8	0.055741	0.99×10^{-4}	17.94±0.03	-0.0109±0.0009	0.0063±0.0009
9	0.249955	1.02×10^{-4}	4.00±0.00	0.0021±0.0008	0.0097±0.0008
10	0.027124	0.96×10^{-4}	36.87±0.13	-0.0078±0.0009	0.0080±0.0009
11	0.039632	0.94×10^{-4}	25.23±0.06	0.0058±0.0009	0.0080±0.0008
12	0.093058	1.04×10^{-4}	10.75±0.01	0.0022±0.0008	0.0037±0.0008
13	0.062207	0.91×10^{-4}	16.08±0.02	0.0013±0.0008	-0.0029±0.0008
14	0.016199	1.26×10^{-4}	61.73±0.48	-0.0040±0.0009	-0.0073±0.0009
15	0.123874	1.14×10^{-4}	8.07±0.01	-0.0008±0.0008	0.0001±0.0008

As far as the harmonic content is concerned, this is shown in Tables 1, 3 and 5. The frequencies that are linked to mixed secular terms in each coordinate can be found in Tables 2, 4 and 6. Moreover, Figures 1b, 1d and 1f show the graphical of the residuals generated by the model. In order to understand how the RMS is reduced as new parameters are added to the model, we show the Figures 2a, 2b and 2c where this reduction is represented.

5 Conclusions

The estimated harmonic models are able to reduce the value of the RMS to 0.0168 *cm*, 0.0153 *cm* and 0.0117 *cm* for the X, Y and Z coordinates of the center of mass of the continental water system. However, all models are not formed by the same number of parameters and they not have the same harmonic content. Thus, the model for the X coordinate uses 57 parameters, whereas the model for the Y coordinate uses 51, and the Z-model includes 59 parameters. Those models are able to explain 71.92%, 82.48% and 83.05% of the variability

 ISBN: 978-84-614-6167-7

Table 2: Same as Table 1 but for the mixed secular terms. It appears the extracting order, period (months) and coefficients (mm) linked to the sinus and cosinus mixed secular term, respectively.

No.	Π	*SS*	*CC*
3	181.44±1.40	-1.485×10^{-4}±7.0×10^{-6}	4.93×10^{-5}±8.0×10^{-6}
4	31.39±0.05	1.074×10^{-4}±6.6×10^{-6}	7.58×10^{-5}±6.9×10^{-6}
12	10.75±0.01	4.75×10^{-5}±6.4×10^{-6}	-2.83×10^{-5}±6.4×10^{-6}
13	16.08±0.02	7.4×10^{-6}±6.6×10^{-6}	-6.40×10^{-5}±6.6×10^{-6}
15	8.07±0.01	-3.97×10^{-5}±6.4×10^{-6}	-3.96×10^{-5}±6.4×10^{-6}

Table 3: Fourier terms. Harmonic content for the variations on the Y coordinate of the geocenter. Columns refer to the extracting order, frequency (cycles per month), uncertainty of the frequency (cycles per month), period (months) and coefficients (cm) linked to sinus and cosinus terms, respectively.

No.	f	σ_f	Π	S	C
1	0.083343	0.09×10^{-4}	12.00±0.00	-0.1061±0.0008	-0.0174±0.0008
2	0.166688	0.30×10^{-4}	6.00±0.00	0.0272±0.0008	-0.0146±0.0008
3	0.008098	0.48×10^{-4}	123.49±0.73	-0.0079±0.0008	0.0186±0.0008
4	0.011226	0.54×10^{-4}	89.08±0.43	0.0136±0.0008	0.0147±0.0008
5	0.033622	0.61×10^{-4}	29.74±0.05	-0.0098±0.0008	0.0138±0.0008
6	0.003636	0.74×10^{-4}	275.02±5.56	0.0089±0.0008	-0.0101±0.0008
7	0.023732	0.51×10^{-4}	42.14±0.09	-0.0048±0.0008	0.0054±0.0008
8	0.040596	0.75×10^{-4}	24.63±0.05	0.0013±0.0008	-0.0009±0.0008
9	0.018081	0.77×10^{-4}	55.31±0.23	0.0083±0.0008	0.0100±0.0008
10	0.013789	0.98×10^{-4}	72.52±0.51	0.0006±0.0008	0.0115±0.0008
11	0.077288	1.11×10^{-4}	12.94±0.02	0.0073±0.0008	0.0006±0.0008
12	0.045162	1.11×10^{-4}	22.14±0.05	0.0089±0.0008	-0.0015±0.0008
13	0.036035	1.14×10^{-4}	27.75±0.09	-0.0079±0.0008	-0.0024±0.0008
14	0.102659	1.43×10^{-4}	9.74±0.01	-0.0015±0.0008	-0.0066±0.0008
15	0.050884	1.67×10^{-4}	19.65±0.06	-0.0060±0.0008	0.0031±0.0008

 ISBN: 978-84-614-6167-7

Table 4: Same as Table 3 but for the mixed secular terms. It appears the extracting order, period (months) and coefficients (mm) linked to the sinus and cosinus mixed secular term, respectively.

$No.$	Π	SS	CC
7	42.14±0.09	$8.11\times10^{-5}\pm5.8\times10^{-6}$	$-5.88\times10^{-5}\pm6.0\times10^{-6}$
8	24.63±0.05	$7.46\times10^{-5}\pm6.3\times10^{-6}$	$-6.1\times10^{-6}\pm6.4\times10^{-6}$

Table 5: Fourier terms. Harmonic content for the variations on the Z coordinate of the geocenter. Columns refer to the extracting order, frequency (cycles per month), uncertainty of the frequency (cycles per month), period (months) and coefficients (cm) linked to sinus and cosinus terms, respectively.

$No.$	f	σ_f	Π	S	C
1	0.083337	0.09×10^{-4}	12.00±0.00	-0.0534±0.0006	0.0626±0.0006
2	0.005061	0.17×10^{-4}	197.59±0.66	-0.0019±0.0006	-0.0052±0.0006
3	0.014783	0.44×10^{-4}	67.65±0.20	-0.0041±0.0006	0.0025±0.0006
4	0.021570	0.58×10^{-4}	46.36±0.13	0.0081±0.0006	-0.0114±0.0006
5	0.166726	0.62×10^{-4}	6.00±0.00	-0.0103±0.0006	-0.0040±0.0006
6	0.032726	0.51×10^{-4}	30.56±0.05	-0.0144±0.0007	-0.0058±0.0007
7	0.030539	0.48×10^{-4}	32.74±0.05	-0.0026±0.0006	0.0044±0.0006
8	0.055593	0.82×10^{-4}	17.99±0.03	-0.0015±0.0006	0.0035±0.0006
9	0.042804	1.14×10^{-4}	23.36±0.06	0.0069±0.0006	0.0022±0.0006
10	0.011200	0.98×10^{-4}	89.29±0.78	-0.0059±0.0006	0.0059±0.0006
11	0.107123	0.95×10^{-4}	9.34±0.01	0.0002±0.0006	-0.0008±0.0006
12	0.061892	1.23×10^{-4}	16.16±0.03	-0.0055±0.0006	0.0008±0.0006
13	0.090079	1.02×10^{-4}	11.10±0.01	-0.0026±0.0006	-0.0019±0.0006
14	0.026472	1.25×10^{-4}	37.78±0.18	-0.0029±0.0006	0.0055±0.0006
15	0.068766	1.83×10^{-4}	14.54±0.04	-0.0032±0.0006	0.0040±0.0006

 ISBN: 978-84-614-6167-7

Table 6: Same as Table 5 but for the mixed secular terms. It appears the extracting order, period (months) and coefficients (mm) linked to the sinus and cosinus mixed secular term, respectively.

No.	Π	*SS*	*CC*
2	197.59±0.66	$2.249\times10^{-4}\pm5.3\times10^{-6}$	$1.117\times10^{-4}\pm4.9\times10^{-6}$
3	67.65±0.20	$-1.030\times10^{-4}\pm4.7\times10^{-6}$	$-3.69\times10^{-5}\pm4.9\times10^{-6}$
7	32.74±0.05	$-7.80\times10^{-5}\pm5.8\times10^{-6}$	$4.75\times10^{-5}\pm5.9\times10^{-6}$
8	17.99±0.03	$4.01\times10^{-5}\pm4.6\times10^{-6}$	$3.30\times10^{-5}\pm4.6\times10^{-6}$
11	9.34±0.01	$1.13\times10^{-5}\pm4.5\times10^{-6}$	$4.51\times10^{-5}\pm4.5\times10^{-6}$
13	11.10±0.01	$4.05\times10^{-5}\pm4.6\times10^{-6}$	$1.20\times10^{-5}\pm4.5\times10^{-6}$

of the data for the X, Y and Z components, respectively.

As we might expect from the graphical display of the data, the trend component does not contain a significant amount of information of the geocenter variations. This can be understood by looking at the small estimated value of the trend coefficients and the percentage of variability that this component explains (1.63%, 0.32% and 2.84% for the X, Y and Z coordinates, respectively).

In this study, a frequency that deserves special attention is that one linked to an annual period. This signal is detected in X, Y and Z coordinates of the center of mass, and furthermore, it always appears at first position in the harmonic content. On the other hand, this frequency explains the largest percentage of variability (36.76%, 51.46% and 51.11% for the X, Y and Z coordinates, respectively). From Tables 1, 3 and 5 we can observe that at annual frequency the largest contribution arises in the Y component with an annual cycle of 1.07 mm, showing a maximum between late April and early May. The annual cycle for the remaining coordinates is smaller, with a minimum of 0.65 mm in the X component. Note that these results are consistent with other studies such as that one carried out by [5]. The only substantial difference lies in the epoch of maximum influence of the annual signal.

The semiannual period, is clearly shown in the Z and Y coordinates with amplitudes of 0.11 mm and 0.31 mm, respectively. However, this spectral line is not detected for the X component. The most similar fluctuations correspond to periods of 4 months (with 0.0999 mm of amplitude) and 8 months (0.0083 mm of amplitude). The reason why the semiannual period does not appear for this coordinate is because it has an amplitude less than 0.0083 mm (which corresponds to the amplitude of the last frequency considered in the model for the X component). In fact, if we perform a harmonic analysis by extracting a

 ISBN: 978-84-614-6167-7

larger number of frequencies or considering a vectorial analysis of (X, Y, Z), the semiannual frequency appears as part of the harmonic content and it does with an amplitude of 0.0061 mm, approximately.

Many other frequencies associated with interannual periods also appear in the harmonic content of the estimated models. Among them, we find a frequency close to a 2-year period (around 23–25 months). This signal has an amplitude of 0.099 mm, 0.016 mm and 0.072 mm for the X, Y and Z coordinates, respectively. If we look for its origin, this frequency might be associated with the so-called *Quasi Biennial Oscillation* (QBO) [4]. This phenomenon is an atmospheric oscillation that takes place every 20–36 months or so, and among its effects we can highlight the change of monsoon rains, which directly affects the distribution of continental water.

Another interannual frequency that appears in the estimated models is linked to a period of 3.5–3.8 years (42–46 months). This signal was also detected by [8], although the estimated value of its amplitude differs from that one presented in this chapter. Its nature, as well as the origin of the 55–68 months signal (4.5–5.5 years), is unknown but, perhaps, it might be attributed to the climatological phenomenon ENSO which is repeated every 3 to 7 years (with an average of 5 years).

There are also fluctuations associated with long periods of time that range between 123 and 198 months. Thus, the X coordinate seems to be affected by a 15.12-year signal whose amplitude is 0.099 mm; the Y component has a 10.3-year fluctuation of 0.2 mm of amplitude and finally, the Z coordinate shows a period of 16.47 years whose amplitude is 0.055 mm, approximately. These spectral lines are some of the most important signals in the models because they explain a significant proportion of the variability of the data and they have a remarkable amplitude if we compared them with other harmonics. The origin of these signals is difficult to guess. It is unknown to what extent monthly data can affect the detection of the spectral lines because the time between two consecutive months is not always the same. If we assume that this fact affects the extraction frequency process, we can say that this period of 10-16 years might be related to the lunar nodes period and semi-period (18.6 and 9.3 years). Obviously, other studies are required to check this hypothesis.

Finally, we note that other shorter periods (80–100 months) are included together with these aforementioned spectral lines. At this moment, its origin is unknown so, in the future, it would be interesting to study the harmonic content of some climate phenomena in order to find similarities and answers.

 ISBN: 978-84-614-6167-7

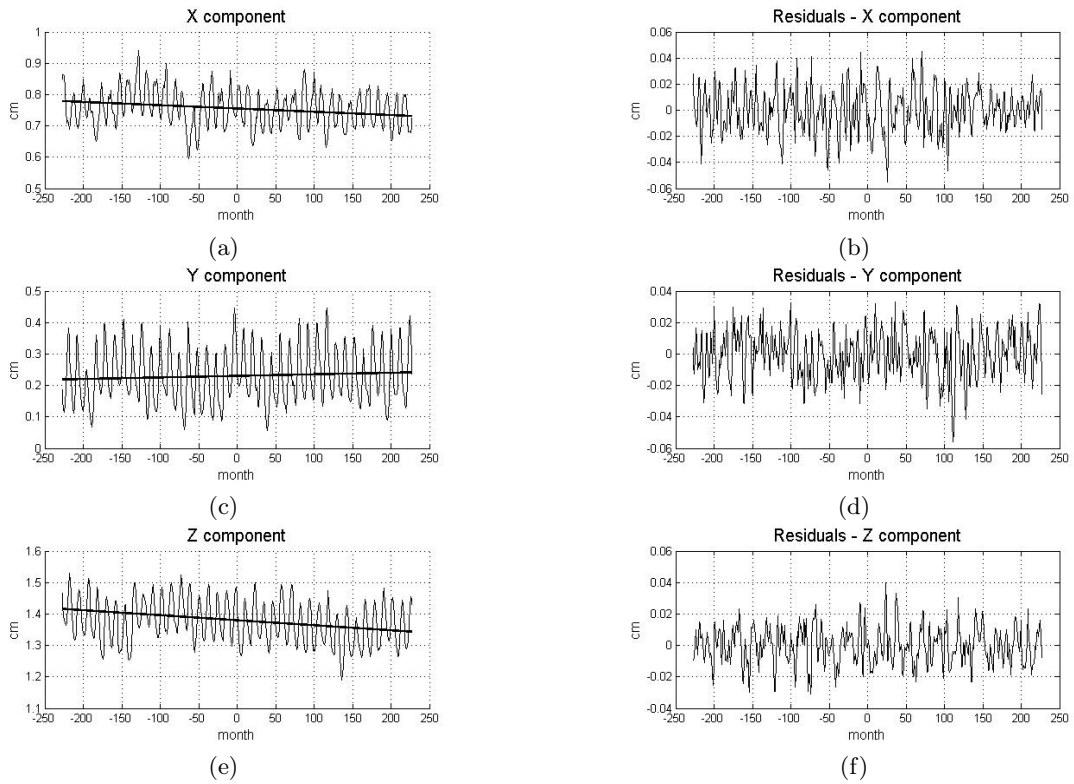


Figure 1: Graphical results of the non-linear harmonic analysis for the geocenter variations caused by continental water changes. (a) Raw data and trend component (straight line) for X coordinate, (b) residuals for the X coordinate, (c) raw data and trend component (straight line) for Y coordinate, (d) residuals for the Y coordinate, (e) raw data and trend component (straight line) for Z coordinate, (b) residuals for the Z coordinate.

6 Figures

Acknowledgements

This research was sponsored by a grant from Generalitat Valenciana, BEFPI 2010-021. Partial support of the Spanish Projects AYA-2009-07981 and AYA 2010-22039-C02-01 of MICIN and ACOMP 2011/196 from Generalitat Valenciana, is also acknowledged.

References

[1] W. H. Press, B. P. Flannery, S. A. Teukolsky and W. T. Vetterling, *Nu-*

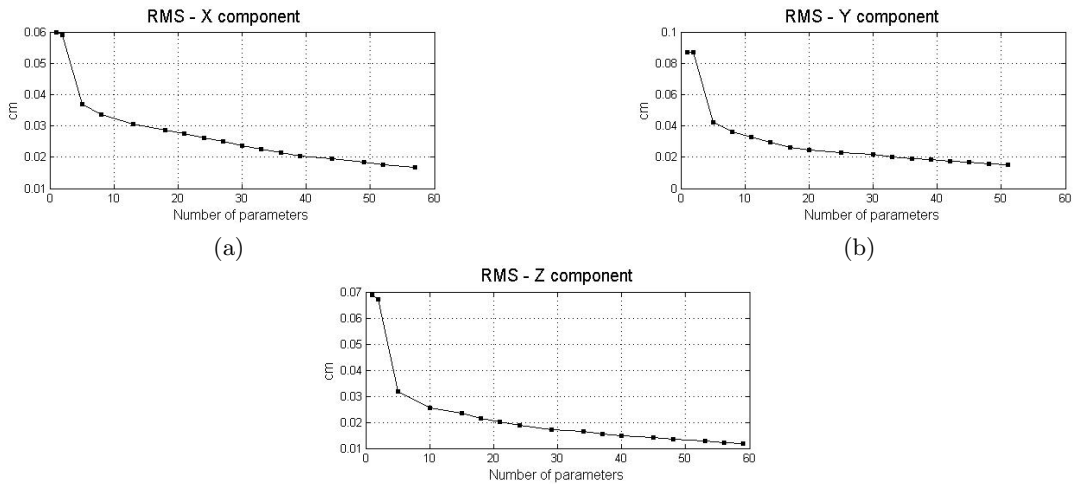


Figure 2: Reduction of the RMS as new parameters are added to the (a) X coordinate model, (b) Y coordinate model and (c) Z coordinate model.

merical recipes in C, The Art of Scientific Computing, 2nd ed., Cambridge University Press, 1967.

[2] C.G. Broyden, *Mathematics of Computation*, 21, 1967.

[3] W. Harada, *A method of non-linear harmonic analysis and its application to dynamical astronomy.*, Master's thesis. Department of Astronomy, Graduate School of Science, University of Tokyo (Japan), 2003.

[4] M. Alexander and K. Weickman, *Biennial variability in an atmospheric general circulation model*, Journal of Climatology **8** (1995) 431–440.

[5] F. Bouillé, A. Cazenave, J. M. Lemoine and J. F. Crétaux, *Geocentre motion from the DORIS space system and laser data to the Lageos satellites: comparison with surface loading data*, Geophys. J. Int. (2000) 71–82.

[6] J. L. Chen, C. R. Wilson, R. J. Eanes and R. S. Nerem, *Geophysical interpretation of observed geocenter variations*, Journal of Geophysical Research **104** (1999) 2683–2690.

[7] D. Dong, J. O. Dickey, Y. Chao and M. K. Cheng, *Geocenter variations caused by atmosphere, ocean and surface ground water*, Geophysical Research Letter **24** (1997) 1867–1870.

 ISBN: 978-84-614-6167-7

[8] J. Huang, H. M. Van Den Dool and K. P. Georgakalos, *Analysis of model-calculated soil-moisture over the United States (1931 - 1993) and applications to long-range temperatures forecasts*, Journal of Climatology **9** (1996) 1350–1362.

 ISBN: 978-84-614-6167-7

Proceedings of the 11th International Conference on Computational and Mathematical Methods in Science and Engineering, CMMSE 2011 26–30 June 2011.

Parallel Discrete Dynamical Systems on Maxterms and Minterms Boolean Functions

S. Martinez[1], F.L. Pelayo[2] and J.C. Valverde[1]

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

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

emails: Silvia.Martinez5@alu.uclm.es, Fernandol.Pelayo@uclm.es, jose.valverde@uclm.es

Abstract

This work is devoted to giving a complete characterization of the orbit structure of parallel discrete dynamical systems with maxterms and minterms Boolean functions as global functions. In this sense, first of all, we analyze what happens with the simplest maxterm and minterm and then we extend our analysis to the rest of maxterms and minterms. As a result, it is shown that the orbit structure does not remain when the system is perturbed.

Key words: Discrete dynamical systems, parallel dynamical systems, orbit structure, Boolean functions, maxterms, minterms

MSC 2000: 37B99; 37E15; 37N99; 68R10; 94C10

1 Introduction

The work of many scientists and technicians consists of finding the future states of processes whose present states they are observing. Certainly, the future states of many biological, ecological, physical and even computer processes can be predicted if their present states and the laws governing their evolution are known, provided that these laws do not change in time. Dynamical systems are the mathematical formalization of deterministic processes, created to deal with these kinds of challenges.

Computer processes are used very often in our modern technological society and this claims for a mathematical modeling of them in order to obtain information about their evolution in time. In computer processes, there are many entities an each entity has a state at a given time (see [3, 4, 5]). Usually, in order to get a graphical idea of the situation, any entity is represented by a vertex of a graph and two vertices are adjacent if their states influence each other in the update of the system. The undirected graph so built is called the *dependency graph* of the system.

 ISBN: 978-84-614-6167-7

If we denominate this graph $G = (V, E)$, where $V = \{1, 2, \ldots, n\}$ is the vertex set and E is the edge set, then, for each vertex/entity $1 \leq i \leq n$, it is natural to consider that its state $x_i \in \{0, 1\}$. That is, the entity could be activated or deactivated.

On the other hand, for every vertex/entity $1 \leq i \leq n$ and every subset $W \subset V$, we need to consider all the vertices that interfere with them. Thus, we denote

$$A_G(i) = \{j \in V | \{i, j\} \in E\} \tag{1}$$

$$A_G(W) = \bigcup_{i \in W} A_G(i) \tag{2}$$

the sets of vertices that are adjacent to the vertex i and to the subset W, respectively.

The evolution or update of the system is implemented by local functions which are the restrictions of a global function. In this context, for updating the state of any entity, the corresponding local function acts only on the state of that entity itself and the states of the entities related to it.

If the states of the entities are updated in a parallel manner, the system is named a *parallel dynamical system* (PDS), while if they are updated in a sequential order, the system is named *sequential dynamical system* (SDS) (see [2, 9]). In this work, we are concerned with parallel dynamical systems. Actually, it can be stated the following definition.

Definition 1. Let $G = (V, E)$ be a graph on $V = \{1, 2, \ldots, n\}$. Then the following map

$$F : \{0, 1\}^n \to \{0, 1\}^n, \qquad F(x_1, x_2, \ldots, x_i, \ldots, x_n) = (y_1, y_2, \ldots, y_i, \ldots, y_n),$$

where y_i is the updated state of the entity/vertex i by applying a local function f_i over $\{i\} \cup A_G(i)$, constitutes a discrete dynamical system called *parallel (discrete) dynamical system* over $\{0, 1\}^n$.

Another important concept for our purposes in this work is the following.

Definition 2. Let B be the set $\{0, 1\}$ and $n \in \mathbb{N}$. A *Boolean function* of n variables is a function of the form

$$F : B^n \to B,$$

where $F(x_1, x_2, \ldots, x_n) \in B$ is obtained from $x_1, x_2, \ldots, x_n \in B$ using the logical AND, the logical OR, the logical NOT and the elements $0, 1 \in B$.

A Boolean function describes how to determine a Boolean output from some Boolean inputs. Thus, such functions play a fundamental role in questions as design of circuits or computer processes. In our context, they represent the evolution operator of the corresponding system.

Maxterms and minterms are both special cases of Boolean functions.

 ISBN: 978-84-614-6167-7

Definition 3. A Boolean function of n variables, $x_1, x_2, \ldots, x_n$ that only uses the disjunction operator, where each of the n variables appears once in either its direct or its complemented (logical NOT) form is called a *maxterm*.

In this sense, the simplest maxterm corresponds to the one that only uses the disjunction operator, where each of the n variables appears once in its direct form, i.e.,

$$OR(x_1, x_2, \ldots, x_n) = x_1 \vee x_2 \vee \ldots \vee x_n. \tag{3}$$

Minterm is the dual of the maxterm concept. That is, instead of using the disjunction operator and complements, the conjunction operator and complements are used in a similar way.

Definition 4. A Boolean function of n variables that only uses the conjunction operator, where each of the n variables appears once in either its direct or its complemented form is called a *minterm*.

The simplest minterm corresponds to the one that only uses the conjunction operator, where each of the n variables appears once in its direct form, i.e.,

$$AND(x_1, x_2, \ldots, x_n) = x_1 \wedge x_2 \wedge \ldots \wedge x_n. \tag{4}$$

As one can check, there exist exactly 2^n maxterms of n variables, since a variable in a maxterm expression can be in either its direct or its complemented form, and dually 2^n minterms.

As is well known, see [6], any Boolean function, except $F \equiv (1, 1, \ldots, 1)$ (resp. $F \equiv (0, 0, \ldots, 0)$), can be expressed in a canonical form as a conjunction (resp. disjunction) of *maxterms* (resp. *minterms*). Therefore, it is natural to begin the study of the dynamics with these basic Boolean functions.

Following [1, 10], the main goals in the study of a dynamical system are giving a complete characterization of its orbit structure and analyzing whether or not this structure remains when the system is perturbed slightly. Since in our particular case of a discrete dynamical system, we have a finite state space, it is obvious that every orbit is either periodic or eventually periodic. Therefore, every orbit is an invariant set of the system. However, it is not so easy to determine a priori the different coexistent periods of its orbits.

In this sense, we prove that the PDS induced by the simplest maxterm or minterm function have only fixed points, which are described, or eventually fixed points whose transients to arrive in the corresponding fixed point is at most as a certain number. Moreover, we show that if other general maxterms and minterms are considered as global functions in order to define the evolution of the systems, the dynamics can change, giving as a result an absence of structural stability of the systems so perturbed.

 ISBN: 978-84-614-6167-7

Acknowledgements

This work has been partially supported by the Grants MTM2008-03679/MTM, TIN2009-14312-C02-02 and PEII09-0184-7802.

References

[1] D.K. Arrowsmith, C.M. Place, *An Introduction to Dynamical Systems*, Cambridge University Press, 1990.

[2] C.L. Barret, W.Y.C. Chen and M.J. Zheng, *Discrete dynamical systems on graphs and Boolean functions*, Math. Comput. Simul. 66 (2004), pp. 487–497.

[3] C.L. Barret and C.M. Reidys, *Elements of a theory of computer simulation I*, Appl. Math. Comput. 98 (1999), pp. 241–259.

[4] C.L. Barret, H.S. Mortveit and C.M. Reidys, *Elements of a theory of computer simulation II*, Appl. Math. Comput. 107 (2002), pp. 121–136.

[5] C.L. Barret, H.S. Mortveit and C.M. Reidys, *Elements of a theory of computer simulation III*, Appl. Math. Comput. 122 (2002), pp. 325-340.

[6] E. A. Bender and S. G. Williamson, *A Short Course in Discrete Mathematics*, Dover Publications, Inc., Mineola, New York, 2005.

[7] O. Colón-Reyes, R. Laubenbacher, B. Pareigis, *Boolean monomial dynamical systems*, Ann. Combin. 8 (2004), pp. 425–439.

[8] R. Laubenbacher, B. Pareigis, *Decomposition and simulation of sequential dynamical systems*, Adv. Appl. Math. 30 (2003), pp. 655678.

[9] H.S. Mortveit and C.M. Reidys, *Discrete, sequential dynamical systems*, Discrete Math. 226 (2002), pp. 281-295.

[10] S. Wiggins, *Introduction to Applied Nonlinear Systems and Chaos*, Springer, New York, 1990.

 ISBN: 978-84-614-6167-7

Proceedings of the 11th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2011
26–30 June 2011.

Comparing DES and DESL from an MRHS point of view

Kenneth R. Matheis and Rainer Steinwandt

Center for Cryptology and Information Security, Florida Atlantic University

emails: kmatheis@fau.edu, rsteinwa@fau.edu

Abstract

Multiple Right Hand Sides (MRHS) is an algebraic attack which seems particularly well-suited for block ciphers having a Feistel structure. We compare the effectivity of this attack against DES and its lightweight variant DESL. Unlike the Data Encryption Standard, DESL uses only a single S-box, and our experiments indicate that for an MRHS-based attack this modification does not significantly affect the workload faced by an adversary. As a side, our experimental results falsify a conjecture of Schoonen about MRHS in connection with DES.

Key words: cryptography, block cipher, algebraic attack

1 Introduction

The block cipher DESL is a *lightweight extension* of the Data Encryption Standard [1] and has been proposed by Leander et al. at FSE 2007 [2]. The structure of this lightweight version is basically identical with DES, but introduces one major change: only a single S-box is used. While this modification may seem attractive from an implementation point of view, one faces the question of the cryptanalytic consequences of such a design choice. Here we describe experimental results with an algebraic attack known as MRHS when applied to DES and to DESL.

MRHS stands for *Multiple Right Hand Sides* and is discussed in detail by Raddum and Semaev [4]. The technique appears to be well-suited for Feistel ciphers, including DES and DESL. From the point of view of MRHS, our experiments do not exhibit a design weakness of DESL. As a side, the results falsify a conjecture about MRHS when being applied to DES, however, which might be of independent interest.

 ISBN: 978-84-614-6167-7

2 Preliminaries

2.1 A lightweight variant of DES

With the structure of DESL and the Data Encryption Standard being identical up to two modificatons, werefer to [2] for a detailed discussion of DESL. The only deviations in DESL from the original DES are the following:

- The (key-independent) initial and final permutations in DES are omitted.
- All S-boxes are replaced with a single (new) S-box.

While the former change does not obviously affect the security against algebraic attacks, the implications of the latter modification are less clear.

2.2 Multiple right-hand sides (MRHS)

Here we restrict to a short review of the main components of MRHS and refer to Raddum and Semaev's work [4] for a more elaborate discussion.

2.2.1 Basic terminology

For a column vector $x = (x_1\ x_2\ \ldots\ x_y)^{\mathrm{T}} \in \mathbb{F}_2^y$, a $k \times y$ binary matrix A of rank k, and column vectors $b_1, b_2, \ldots, b_s \in \mathbb{F}_2^k$ consider the following type of equation:

$$Ax = b_1, b_2, \ldots, b_s \tag{1}$$

We refer to such an equation as *MRHS system of linear equations* with *right hand sides* $b_1, b_2, \ldots, b_s$. By a *solution* to (1) we mean a vector in $\mathbb{F}_2^y$ satisfying at least one particular linear system of equations $Ax = b_i$. The set of *all solutions to* (1) is the union of the solutions to the individual systems $Ax = b_i$ $(1 \leq i \leq s)$. To work with MRHS systems of linear equations, we juxtapose the above column vectors b_i to form a matrix L and rewrite Equation (1) as $Ax = [L]$. The pair (A, L) is called a *symbol*, and when writing equations, the brackets around L emphasize that we are not working with an ordinary equation of matrices. Given a system of symbols

$$\begin{array}{rlcl} S_1: & A_1x & = & [L_1] \\ & & \vdots & \\ S_n: & A_nx & = & [L_n] \end{array}, \tag{2}$$

a *solution to such a system* is a vector $x \in \mathbb{F}_2^y$ satisfying all of the underlying n MRHS systems of linear equations.

 ISBN: 978-84-614-6167-7

2.2.2 Solving a system of symbols

There are three main components to MRHS: *agreeing*, *gluing*, and *extracting equations*. Since memory is finite in any actual implementation of the algorithm, it may also happen that we have to *guess variables*, and sometimes an *equation symbol* is used.

Agreeing The basic idea of an agreeing phase is to remove columns b in a right hand side L_i, if no one solution of $A_i x = b$ can be a solution to the system (2). To achieve this, pairwise *agreeing* of symbols is employed. Namely, let $S_i : A_i x = [L_i]$ and $S_j : A_j x = [L_j]$ be two symbols. Then we say that S_i and S_j *agree* if for every $b \in L_i$, there exists a $b' \in L_j$ such that the linear system

$$\begin{pmatrix} A_i \\ A_j \end{pmatrix} x = \begin{pmatrix} b \\ b' \end{pmatrix} \tag{3}$$

is consistent, and, vice versa, for each $b' \in L_j$ there exists a $b \in L_i$ such that (3) is consistent. In a situation where S_i and S_j do not agree, we remove those columns b from L_i for which the linear system $A_i x = b$ is inconsistent with $A_j x = [L_j]$. Dually, those columns b' from L_j are removed, for which $A_j x = b'$ is inconsistent with $A_i x = [L_i]$.

If two symbols S_h and S_i agree, but S_i and S_j disagree, columns may be deleted in one or both of L_i and L_j. After this happens, it may well happen that S_h does not agree with either of the modified symbols, and it becomes necessary to *re-agree* S_h with them. During the latter agreement, columns from L_h may have to be deleted, and so on, possibly resulting in a chain reaction of column deletions. To ensure that a system of symbols reaches a pairwise-agreed state, we perform the *Agreeing1 Algorithm* in Figure 1 (see [4, Section 3.1]).

While the symbols in a System (2) do not pairwise agree,

1. find S_i and S_j which do not agree
2. agree S_i and S_j.

Figure 1: Agreeing1 Algorithm.

Gluing When a system of symbols is in a pairwise-agreed state, we may choose to apply a different operation: The *gluing* of two symbols $S_i = (A_i, L_i)$ and $S_j = (A_j, L_j)$ results in a new symbol $Bx = [L]$ whose set of solutions is the set of common solutions to $A_i x = [L_i]$ and $A_j x = [L_j]$. After having formed this new symbol, it is inserted into the system at hand and the two symbols S_i and S_j which formed (B, L) are no longer necessary and removed from the system.

 ISBN: 978-84-614-6167-7

Gluing an L_i of width s_i with a matrix L_j of width s_j may yield a matrix L with as many as $s_i \cdot s_j$ columns. In an implementation, computing certain glues might therefore turn out to be infeasible, and one restricts to gluing only pairs of symbols where the number of columns in the resulting symbol does not exceed a certain *threshold*. Once several glues have been performed, the symbols in the resulting system will usually no longer be pairwise-agreed, so the Algorithm in Figure 1 can be run again, initiating another round of agreeing and gluing. The eventual goal of iterated agreeing and gluing steps is to obtain a system of symbols which consists of a single symbol.

Extracting equations From a given symbol $S : Ax = [L]$ we can try to extract *Unique Right Hand Side (URHS)* equations, and if this is done, the resulting linear equations are placed in a dedicated symbol S_0 to which we refer as *equation symbol*. The equation symbol is checked for consistency and size. The A-part of S_0 has the same number of columns as the A-parts of the other symbols, but its L-part has only one column. The symbol S_0 is not considered a proper part of the system (2) and does not take part in the Agreement1 Algorithm, nor is it removed after being glued to a symbol in the system. However, various implementations will involve S_0 in an agreement or gluing step. Further, information from guessing variables may also be reflected by S_0.

Guessing variables It may happen that all symbols in a system are pairwise-agreed, no new URHS equations can be extracted, and no pair of symbols can be glued without exceeding the threshold. In such a situation one can guess the (one bit) value of a variable. Before performing a guess, the system of symbols—to which we will refer as the *state*—is stored. After the guess has been made, pairwise agreeing, gluing, and equation extraction are performed as normal. If after some steps the state, again, does not allow for any new URHS equation to be computed or pair of symbols to be glued, the state is again saved, and we guess the value of another variable.

A guess for a variable can be incorrect, and this manifests as follows: during the agreement of two symbols, all right hand sides of at least one of the symbols get removed, indicating that the system has no solution. When this happens, the state can be rolled back to a previously saved state, to make a different guess. In the sequel, we denote by δ the number of key bits we must guess before we discover the whole key through an MRHS attack.

3 Applying MRHS to DES and DESL

Since the structure of DES and DESL is the same, the process for creating the A-parts of MRHS symbols for DESL is the same as that for DES, which is described in [5, pp. 50–53]. The only difference is that the L-part of each symbol will not correspond to a DES S-box,

 ISBN: 978-84-614-6167-7

but instead to the DESL S-box. For our experiments, we used a PC with two quad-core Xeon E5520 2.26 GHz processors (though only one core was used), 24 GB of RAM, using Windows 7 Server (Standard Edition). The ciphertext was `0123456789ABCDEF`, and the key was the first 56 bits of the SHA-1 hash of "Katalina" (without quotes). Under these conditions, DES and DESL were attacked varying both the number of rounds of the cipher and the threshold of MRHS. The results are summarized in Table 1, with the note that the threshold listed is actually the base 2 logarithm of the actual threshold, so we always choose a power of 2 for the number of columns each L-part is allowed to grow to. The table lists the δ-value for DES followed by a value in parantheses specifying the difference to DESL. For instance, with threshold 20, to attack twelve rounds of DES 41 key bits had to be guessed, whereas in the same scenario for DESL only 40 guesses were needed.

	Rounds of DES						
Threshold	4	6	8	10	12	14	16
20	1 (+1)	35 (+1)	36 (+0)	36 (+0)	41 (+1)	41 (+3)	40 (+0)
21	0 (+0)	35 (+1)	39 (+0)	37 (+0)	39 (+0)	40 (+1)	39 (-3)
22	0 (+0)	32 (-1)	39 (+0)	37 (+0)	38 (+0)	40 (-3)	38 (+0)
23	0 (+0)	33 (+0)	39 (+1)	43 (-2)	46 (+0)	48 (+0)	46 (+0)

Table 1: Value of δ for DES (and difference to DESL), varying number of rounds and threshold

It was conjectured by Schoonen in [5, Hypothesis 5.1] that, for 7 through 16 rounds of DES, δ would always be 56 minus the (base 2 logarithm of the) threshold, but Table 1 demonstrates that this is not the case.

4 Conclusion

Our experimental results indicate that from an MRHS perspective DESL offers comparable security as DES. On two occasions, DESL actually required three more bits to be guessed before recovering the entire key.

References

[1] William M. Daley and Raymond G. Kammer. Data Encryption Standard (DES). Federal Information Processing Standards Publication, U.S. Department of Commerce/National Institute of Standards and Technology, October 1999.

[2] Gregor Leander, Christof Paar, Axel Poschmann, and Kai Schramm. New Lightweight DES Variants. In Alex Biryukov, editor, *Fast Software Encryption, 14th International*

 ISBN: 978-84-614-6167-7

Workshop, FSE 2007, volume 4593 of *Lecture Notes in Computer Science*, pages 196–210. International Association for Cryptologic Research, Springer, 2007.

[3] Håvard Raddum and Igor Semaev. Solving MRHS linear equations. Cryptology ePrint Archive, Report 2007/285, 2007. Available at `http://eprint.iacr.org/2007/285`.

[4] Håvard Raddum and Igor Semaev. Solving Multiple Right Hand Sides linear equations. *Designs, Codes and Cryptography*, 49:147–160, 2008. Preprint available in [3].

[5] A. C. C. Schoonen. Multiple right-hand side equations. Master's thesis, Eindhoven University of Technology, Department of Mathematics and Computer Science, May 2008. Available at `http://alexandria.tue.nl/extra1/afstversl/wsk-i/schoonen2008.pdf`.

 ISBN: 978-84-614-6167-7

Proceedings of the 11th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2011
26–30 June 2011.

Towards dual multi-adjoint concept lattices

Jess Medina[1]

[1] *Depart. Matemáticas, Universidad de Cádiz. Spain*

emails: jesus.medina@uca.es

Abstract

Several papers relate the classical property-oriented and object-oriented concept lattices and the dual concept lattices, although a negation is needed. This paper presents a fuzzy generalization of the dual concept lattice, the dual multi-adjoint concept lattice in which the philosophy of the multi-ajoint paradigm is applied and where no negation on the lattices is needed.

Key words: Concept lattices, Galois connection, implication triples

1 Introduction

Wille introduced formal concept analysis (FCA) in [28] and it has become an important and appealing research topic both from the theoretical perspective [27] and from the applicative one. Regarding applications, we can find papers ranging from ontology merging [23], to applications to the semantic web by using the notion of concept similarity [9], and from processing of medical records in the clinical domain [14] to the development of recommender systems [6].

This important tool, in knowledge representation and knowledge discovery in relational information systems, has been related with another important tool, the rough set theory, and different results presented in a formal concept analysis framework have been applied to a rough set theory [5, 26].

Pawlak introduced rough set (RS) theory in [22] as a formal tool for modelling and processing incomplete information in information systems. This theory was extended by Düntsch and Gediga in [8, 11] and Yao in [29] in order to consider two different sets, the set of objects and the set of attributes. These extensions are called *property-oriented concept lattice* and *object-oriented concept lattices* [5].

Another interesting concept lattice framework introduced in [5] was the dual formal concept lattice, which is building from the dual sufficient modal operator.

There exists several fuzzy extensions of the FCA and RS [3, 1, 24, 7, 25, 16, 12]. In the framework of fuzzy FCA, multi-adjoint concept lattices, were introduced [20]

 ISBN: 978-84-614-6167-7

as a new general approach to formal concept analysis, in which the philosophy of the multi-adjoint paradigm is applied (see [21] for more information). With the idea of providing a general framework in which different fuzzy approaches could be conveniently accommodated, the authors worked in a general non-commutative environment; and this naturally leads to the consideration of adjoint triples as the main building blocks of a multi-adjoint concept lattice.

Recently, the philosophy of the multi-adjont paradigm is used to present a fuzzy generalization of the property-oriented concept lattice [18]. A similar idea can be used to obtain a fuzzy framework of the object-oriented concept lattice.

This paper presents a technical generalization of the other interesting crisp concept lattice, the dual formal concept lattice. The multi-adjoint paradigm [21] is adapted to this new fuzzy environment where no negations are needed, the carriers may be complete lattices, different adjoint triples can be assumed, etc. Applications and practical examples on this framework will be studied further.

The generalization proposed is very interesting since, at the moment, the introduction of this kind of concept lattice has been made using a negation on the carrier. Moreover, this environment provides a new point of view to obtain information from databases with both incomplete information and imprecise information, which will give more flexibility than the existing procedures.

This paper is structured as follows: a summary of formal concept analysis and derivation operators is introduced in Section 2. Section 3 recalls the main computation operators, the adjoint triples, and a general and flexible fuzzy concept lattice structure, the multi-adjoint concept lattices; the "dual" of this structure that embeds the crisp definition given in [5] is presented in Section 4. Lastly, the paper ends with diverse conclusions and prospects for future work.

2 Formal concept analysis and derivation operators

Formal concept analysis considers a set of **a**ttributes A, a set of o**b**jects B and a crisp relation between them $R\colon A \times B \to \{0,1\}$, where, for each $a \in A$ and $b \in B$, we have that $R(a,b) = 1$, if a and b are related, or $R(a,b) = 0$, otherwise. We will also write aRb when $R(a,b) = 1$. The triple (A,B,R) is called a *formal context* and the mappings $^{\triangle}\colon 2^B \to 2^A$, $^{\triangle}\colon 2^A \to 2^B$, are defined, for each $X \subseteq B$ and $Y \subseteq A$, as:

$$\begin{aligned} X^{\triangle} &= \{a \in A \mid \text{for all } b \in X, aRb\} \\ &= \{a \in A \mid \text{if } x \in X, \text{ then } aRb\} \qquad (1) \\ Y^{\triangle} &= \{b \in B \mid \text{for all } a \in Y, aRb\} \\ &= \{b \in B \mid \text{if } a \in Y, \text{ then } aRb\} \qquad (2) \end{aligned}$$

These operators are so-called *sufficient operators*, although in order to distinguish about which carriers are defined, they are also called the *extent* and *intent* mappings, respectively.

Given a context (A,B,R), a *concept* in (A,B,R) is defined to be a pair (X,Y),

 ISBN: 978-84-614-6167-7

where $X \subseteq B$, $Y \subseteq A$, and which satisfy that $X^{\triangle} = Y$ and $Y^{\triangle} = X$. The element X of the concept (X, Y) is the *extent* and Y the *intent*.

The set of concepts in a context (A, B, R) is denoted as $\mathcal{B}(A, B, R)$ and it is a complete lattice [10], with the inclusion order on the left argument or the opposite of the inclusion order on the right argument, that is, given $(X_1, Y_1), (X_2, Y_2) \in \mathcal{B}(A, B, R)$, we have that $(X_1, Y_1) \leq (X_2, Y_2)$ if $X_1 \subseteq X_2$ (or, equivalently, $Y_2 \subseteq Y_1$).

The more important characteristic of the mappings ${}^{\triangle}\colon 2^B \to 2^A$ and ${}^{\triangle}\colon 2^A \to 2^B$, is that they form a Galois connection.

Proposition 1 *Given a formal context (A, B, R) and the mappings ${}^{\triangle}\colon 2^B \to 2^A$ and ${}^{\triangle}\colon 2^A \to 2^B$, defined above, the pair $({}^{\triangle}, {}^{\triangle})$ is a Galois connection, that is,* Galois connection *between P_1 and P_2 if and only if:*

1. *${}^{\triangle}\colon 2^B \to 2^A$ and ${}^{\triangle}\colon 2^A \to 2^B$ are order-reversing.*
2. *$X \subseteq_B X^{\triangle\triangle}$, for all $X \subseteq B$.*
3. *$Y \subseteq_A Y^{\triangle\triangle}$, for all $Y \subseteq A$.*

These definitions of extent and intent operators are the classical in FCA but not the unique derivation operators. There exist three extra definitions considered in several frameworks: qualitative data analysis [11, 8], crisp rough set theory [30], fuzzy rough set theory [4, 17]. Some extra motivations about these operators are also introduced in [16, 12].

Considering the sets A, B, and a crisp relation $R\colon A \times B \to \{0, 1\}$, that is, which can be assumed as a formal concept, the derivation operators ${}^{\pi}\colon 2^B \to 2^A$, ${}^{N}\colon 2^B \to 2^A$, ${}^{\nabla}\colon 2^B \to 2^A$ are defined, for each $X \subseteq B$, as:

$$\begin{aligned} X^{\pi} &= \{a \in A \mid \text{ there exists } b \in X, \text{ such that } aRb\} \\ X^{N} &= \{a \in A \mid \text{ for all } b \in B, \text{ if } aRb, \text{ then } b \in X\} \\ X^{\nabla} &= \{a \in A \mid \text{ there exists } b \in X^c, \text{ such that } aR^c b\} \end{aligned}$$

where X^c is the complement of X and R^c is the complement relation of R. Analogously, abusing of notation, we can define the mappings: ${}^{\pi}\colon 2^A \to 2^B$, ${}^{N}\colon 2^A \to 2^B$ and ${}^{\nabla}\colon 2^A \to 2^B$.

These operators are so-called *possibility*, *necessity* and *dual sufficiency operators*, respectively. They are composed in order to form Galois connections or closure operators [10, 28, 5, 11, 8] and diferent concept lattices are obtained: the *property-oriented concept lattice*, *object-oriented concept lattice* and *dual formal concept lattice* [5].

Clearly, the dual sufficiency operator satisfies that: $X^{\nabla} = ((X^c)^{\triangle})^c$, for each $X \subseteq B$, therefore these operators are not independent and the concept lattices given from them are related, specifically we can obtain one from the other. Observe that we are relating both operators using a negation, the complement operator.

Moreover, the necessity and possibility operators are related with the sufficient operators (for details see [16]).

 ISBN: 978-84-614-6167-7

3 Adjoint triples and multi-adjoint concept lattices

This section recalls the multi-adjoint concept lattice as well as its main building blocks, the adjoint triples. These triples are a generalization of the well-known t-norm and its residuated implication satisfying the adjoint property [13]. A triple is obtained since we do not assume that the conjunctors verify the commutative property. This directly provides two different ways of applying the adjoint property, depending on which argument is fixed.

Definition 1 *Let* $(P_1, \leq_1)$, $(P_2, \leq_2)$, $(P_3, \leq_3)$ *be 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 an* adjoint triple *with respect to* P_1, P_2, P_3 *if:*

1. $\&$ *is order-preserving in both arguments.*
2. $\swarrow$ *and* $\nwarrow$ *are order-preserving on the first argument and order-reversing on the second argument.*
3. $x \leq_1 z \swarrow y \quad$ *iff* $\quad x \,\&\, y \leq_3 z \quad$ *iff* $\quad y \leq_2 z \nwarrow x, \quad$ *where* $x \in P_1$, $y \in P_2$ *and* $z \in P_3$.

The following definition presents the basic structure which allows the existence of several adjoint triples for a given triplet of lattices.

Definition 2 *A* multi-adjoint frame $\mathcal{L}$ *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 complete lattices,* $(P, \leq)$ *is a poset and, for all* $i = 1, \ldots, n$, $(\&_i, \swarrow^i, \nwarrow_i)$ *is an adjoint triple with respect to* L_1, L_2, P.
Multi-adjoint frames are denoted $(L_1, L_2, L, \&_1, \ldots, \&_n)$.

Considering a multi-adjoint frame, a *multi-adjoint context* is a tuple consisting of a set of objects, a set of attributes and a fuzzy relation among them; in addition, the multi-adjoint approach also includes a function which assigns an adjoint triple to each object (or attribute).

Definition 3 *Let* $(L_1, L_2, P, \&_1, \ldots, \&_n)$ *be a multi-adjoint frame, a* context *is a tuple* (A, B, R, σ) *such that* A *and* B *are non-empty sets (usually interpreted as attributes and objects, respectively),* R *is a* P*-fuzzy relation* $R\colon A \times B \to P$ *and* $\sigma\colon B \to \{1, \ldots, n\}$ *is a mapping which associates any element in* B *with some particular adjoint triple in the frame.*[1]

[1] A similar theory could be developed by considering a mapping $\tau\colon A \to \{1, \ldots, n\}$ which associates any element in A with some particular adjoint triple in the frame.

 ISBN: 978-84-614-6167-7

Given a multi-adjoint frame and context, the mappings $^{\uparrow_\sigma}: L_2^B \longrightarrow L_1^A$ and $^{\downarrow^\sigma}: L_1^A \longrightarrow L_2^B$ are defined, for all $g \in L_2^B$ and $f \in L_1^A$, as:

$$g^{\uparrow_\sigma}(a) = \inf\{R(a,b) \swarrow^{\sigma(b)} g(b) \mid b \in B\} \quad (3)$$

$$f^{\downarrow^\sigma}(b) = \inf\{R(a,b) \nwarrow_{\sigma(b)} f(a) \mid a \in A\} \quad (4)$$

which generalize the classical definitions given in (1), (2), and that can be seen as extensions of those fuzzy given in [2, 15]. Moreover, as the classical ones, these two arrows generate a Galois connection [20].

A *multi-adjoint concept* is a pair $\langle g, f\rangle$ satisfying that $g \in L_2^B$, $f \in L_1^A$ and that $g^{\uparrow_\sigma} = f$ and $f^{\downarrow^\sigma} = g$; with $(^{\uparrow_\sigma}, ^{\downarrow^\sigma})$ being the Galois connection defined above. The set of all multi-adjoint concepts is called multi-adjoint concept lattice [20].

Definition 4 *The* multi-adjoint concept lattice *associated to a multi-adjoint frame* $(L_1, L_2, P, \&_1, \ldots, \&_n)$ *and a context* (A, B, R, σ) *is the set*

$$\mathcal{M} = \{\langle g, f\rangle \mid g \in L_2^B, f \in L_1^A \text{ and } g^{\uparrow_\sigma} = f, f^{\downarrow^\sigma} = g\}$$

in which the ordering is defined by $\langle g_1, f_1\rangle \preceq \langle g_2, f_2\rangle$ *if and only if* $g_1 \preceq_2 g_2$ *(equivalently* $f_2 \preceq_1 f_1$*).*

The pair $(\mathcal{M}, \preceq)$ is indeed a complete lattice [20]. Moreover, a representation theorem to multi-adjoint concept lattices was proved, which generalizes the classical one and some other fuzzy generalizations.

4 Dual multi-adjoint concept lattice

We commented above, in Section 2, that the operator ∇ can be obtained from the $\triangle$ operator, specifically $X^\nabla = ((X^c)^\triangle)^c$, for all $X \subseteq B$. As a consequence, all properties given to $\triangle$ can be transfer to obtain properties to ∇.

Following this fact, in this section, we aim to introduce a fuzzy extension of the dual sufficient operator and relate it with the fuzzy definition of the sufficient operator, in order to translate the properties from one to another. This relation could not be so direct as in the classical case, since we are in a fuzzy environment.

First of all, we need to recall the definition and notation of dual order. Given a set P and an order relation, $\leq$, on P, the *dual* order of $\leq$ is the relation $\leq^\partial$, defined as $x_1 \leq^\partial x_2$ if and only if $x_2 \leq x_1$, for all $x_1, x_2 \in P$. Usually, we will write P instead of the partial ordered set $(P, \leq)$, P^∂ instead of $(P, \leq^\partial)$, and we will say that P^∂ is the *dual* of P.

Now, the frame in this environment must be defined. Given two complete lattices $(L_1, \preceq_1)$ and $(L_2, \preceq_2)$, a poset $(P, \leq)$ and adjoint triples with respect to $L_1^\partial, L_2^\partial, P$, $(\&_i, \swarrow^i, \nwarrow_i)$, for all $i = 1, \ldots, n$, a *dual multi-adjoint frame* $\mathcal{L}$ is the tuple

$$(L_1, L_2, P, \preceq_1, \preceq_2, \leq, \&_1, \swarrow^1, \nwarrow_1, \ldots, \&_n, \swarrow^n, \nwarrow_n)$$

Dual multi-adjoint frames are denoted $(L_1, L_2, P, \&_1, \ldots, \&_n)$.

 ISBN: 978-84-614-6167-7

The definition of context is analogous to the one given in the previous section. Assumed a dual multi-adjoint frame, $(L_1, L_2, P, \&_1, \ldots, \&_n)$, a *context* is a tuple (A, B, R, σ) such that A and B are non-empty sets (usually interpreted as attributes and objects, respectively), R is a P-fuzzy relation $R\colon A \times B \to P$ and $\sigma\colon B \to \{1, \ldots, n\}$ is a mapping which associates any element in B with some particular adjoint triple in the frame.

From now on, we will fix a dual multi-adjoint frame, $(L_1, L_2, P, \&_1, \ldots, \&_n)$ and context, (A, B, R, σ).

Now, we will introduce the mappings which will build the dual multi-adjoint formal concept lattice, ${}^{\uparrow\nabla}\colon L_2^B \to L_1^A$ and ${}^{\downarrow^\nabla}\colon L_1^A \to L_2^B$. Given mappings $g\colon B \to L_2$, $f\colon A \to L_1$, we consider the "dual" mappings $g^\partial\colon B \to L_2^\partial$, $f^\partial\colon A \to L_1^\partial$, defined as g and f but interpreted in the opposite lattice, respectively, and we apply the mappings ${}^{\uparrow}\colon (L_2^\partial)^B \to (L_1^\partial)^A$, ${}^{\downarrow}\colon (L_1^\partial)^A \to (L_2^\partial)^B$ defined in Equations (3) and (4), that is:

$$\begin{aligned} (g^\partial)^{\uparrow}(a) &= \inf_{1,\partial}\{R(a,b) \swarrow^b g^\partial(b) \mid b \in B\} \\ (f^\partial)^{\downarrow}(b) &= \inf_{2,\partial}\{R(a,b) \nwarrow_b f^\partial(a) \mid a \in A\} \end{aligned}$$

where $(\&_i, \swarrow^i, \nwarrow_i)$ are adjoint triples on L_1^∂, L_2^∂ and P, and $\inf_{1,\partial}$, $\inf_{2,\partial}$ are the infimum on L_1^∂ and L_2^∂, respectively.

Finally, the mappings ${}^{\uparrow\nabla}\colon L_2^B \to L_1^A$ and ${}^{\downarrow^\nabla}\colon L_1^A \to L_2^B$ are defined as: $g^{\uparrow\nabla} = ((g^\partial)^{\uparrow})^\partial$, $f^{\downarrow^\nabla} = ((f^\partial)^{\downarrow})^\partial$, for all $g \in L_2^B$, $f \in L_1^A$. Hence, these definitions are equivalent to:

$$\begin{aligned} g^{\uparrow\nabla}(a) &= \sup_1\{R(a,b) \swarrow^b g^\partial(b) \mid b \in B\} \\ f^{\downarrow^\nabla}(b) &= \sup_2\{R(a,b) \nwarrow_b f^\partial(a) \mid a \in A\} \end{aligned}$$

where $\sup_1$, $\sup_2$ are the supremum on L_1^∂ and L_2^∂, respectively.

The pair $({}^{\uparrow\nabla}, {}^{\downarrow^\nabla})$ is not a Galois connection but satisfies the closure property, which is sufficient condition to form a concept lattice. A *dual concept* is a pair $\langle g, f\rangle$, where $g \in L_2^B$, $f \in L_1^A$ and that satisfies $g^{\uparrow\nabla} = f$, $f^{\downarrow^\nabla} = g$.

Definition 5 *Given a dual multi-adjoint frame $(L_1, L_2, P, \&_1, \ldots, \&_n)$ and a context (A, B, R, σ), a* dual multi-adjoint concept lattice *is the pair $(M^\nabla, \leq^\nabla)$, where*

$$\mathcal{M}^\nabla = \{\langle g, f\rangle \mid g \in L_2^B, f \in L_1^A \text{ and } g^{\uparrow\nabla} = f, f^{\downarrow^\nabla} = g\}$$

is the set of dual concepts, and $\leq^\nabla$ is the order defined $\langle g_1, f_1\rangle \leq^\nabla \langle g_2, f_2\rangle$ if and only if $g_1 \preceq_2 g_2$ (or, equivalently, $f_2 \preceq_1 f_1$).

As $({}^{\uparrow\nabla}, {}^{\downarrow^\nabla})$ is not a Galois connection, the proof that $(M^\nabla, \leq^\nabla)$ is indeed a complete lattice is not direct. In order to prove this fact, we will consider a particular multi-adjoint concept lattice.

Specifically, the complete lattices L_1^∂ and L_2^∂ will be considered together with the adjoint triples $(\&_i, \swarrow^i, \nwarrow_i)$ on L_1^∂, L_2^∂ and P. Therefore, given $g^\partial \in (L_2^\partial)^B$ and $f^\partial \in$

 ISBN: 978-84-614-6167-7

$(L_1^\partial)^A$, the pair $\langle g^\partial, f^\partial\rangle$ is a *multi-adjoint concept* with respect to the Galois connection $(^\uparrow, ^\downarrow)$, if $(g^\partial)^\uparrow = f^\partial$ and $(f^\partial)^\downarrow = g^\partial$, and we obtain the lattice:

$$\mathcal{M}' = \{\langle g^\partial, f^\partial\rangle \mid \text{ if } \langle g^\partial, f^\partial\rangle \text{ is a multi-adjoint concept}\}$$

with order, $(g_1^\partial, f_1^\partial) \leq (g_2^\partial, f_2^\partial)$ if and only if $g_1^\partial \preceq_2^\partial g_2^\partial$ (or, equivalently, $f_2^\partial \preceq_1^\partial f_1^\partial$), where:

$$\begin{aligned} \inf\{\langle g_i^\partial, f_i^\partial\rangle \mid i \in I\} &= \langle \inf_{2,\partial}\{g_i^\partial \mid i \in I\}, (\sup_{1,\partial}\{f_i^\partial \mid i \in I\})^{\downarrow\uparrow}\rangle \quad (5)\\ \sup\{\langle g_i^\partial, f_i^\partial\rangle \mid i \in I\} &= \langle (\sup_{2,\partial}\{g_i^\partial \mid i \in I\})^{\uparrow\downarrow}, \inf_{1,\partial}\{f_i^\partial \mid i \in I\}\rangle \quad (6)\end{aligned}$$

such that $\sup_{j,\partial}$ and $\inf_{j,\partial}$ are the supremum and infimum on L_j^∂, respectively, with $j \in \{1, 2\}$.

The next proposition relates the concept lattices above to dual multi-adjoint concept lattices, justifying why the name of "dual multi-adjoint concept lattice" has been considered for this new environment of concept lattices.

Proposition 2 *Let $(L_1, L_2, P, \&_1, \ldots, \&_n)$ be a dual multi-adjoint frame, (A, B, R, σ) a context and $(\mathcal{M}', \leq)$, $(\mathcal{M}^\nabla, \leq^\nabla)$ the concept lattices defined above. The pair $\langle g, f\rangle$ is a dual concept, that is, $\langle g, f\rangle \in (\mathcal{M}^\nabla, \leq^\nabla)$, if and only if $\langle g^\partial, f^\partial\rangle \in (\mathcal{M}', \leq)$. Moreover, given $\langle g_1, f_1\rangle, \langle g_2, f_2\rangle \in (\mathcal{M}^\nabla, \leq^\nabla)$ we obtain that $\langle g_1, f_1\rangle \leq^\nabla \langle g_2, f_2\rangle$ if and only if $\langle g_2^\partial, f_2^\partial\rangle \leq \langle g_1^\partial, f_1^\partial\rangle$.*

As a consequence of this result, the expressions (5) and (6) are equivalent to:

$$\begin{aligned} \sup\{\langle g_i, f_i\rangle \mid i \in I\} &= \langle \sup_2\{g_i \mid i \in I\}, (\inf_1\{f_i \mid i \in I\})^{\downarrow^\nabla\uparrow^\nabla}\rangle\\ \inf\{\langle g_i, f_i\rangle \mid i \in I\} &= \langle (\inf_2\{g_i \mid i \in I\})^{\uparrow^\nabla\downarrow^\nabla}, \sup_1\{f_i \mid i \in I\}\rangle\end{aligned}$$

for each family of dual concepts $\langle g_i, f_i\rangle \in (\mathcal{M}^\nabla, \leq^\nabla)$, with i in an index set I and, therefore, the pair $(\mathcal{M}^\nabla, \leq^\nabla)$ is a complete lattice.

Theorem 1 *Given a dual multi-adjoint frame $(L_1, L_2, P, \&_1, \ldots, \&_n)$ and a context (A, B, R, σ), the dual multi-adjoint formal concept lattice $(M^\nabla, \leq^\nabla)$ is a complete lattice, where*

$$\begin{aligned} \inf\{\langle g_i, f_i\rangle \mid i \in I\} &= \langle (\inf_2\{g_i \mid i \in I\})^{\uparrow^\nabla\downarrow^\nabla}, \sup_1\{f_i \mid i \in I\}\rangle\\ \sup\{\langle g_i, f_i\rangle \mid i \in I\} &= \langle \sup_2\{g_i \mid i \in I\}, (\inf_1\{f_i \mid i \in I\})^{\downarrow^\nabla\uparrow^\nabla}\rangle\end{aligned}$$

Note that any negation operator has been used, although adjoint triples on the dual lattices of $(L_1, \preceq_1)$ and $(L_2, \preceq_2)$ have been needed.

Moreover, considering the classical case, that is, $L_1 = L_2 = P = \{0, 1\}$, we obtain: $g^{\uparrow^\nabla} = g^\nabla$ and $f^{\downarrow^\nabla} = f^\nabla$, for all g and f crisp subsets of X and A, respectively.

 ISBN: 978-84-614-6167-7

5 Conclusions and future work

It is very important to extract information from databases. Obtaining this knowledge in information systems is a necessity. Two particular tools for that are formal concept analysis and rough sets theory.

In this paper, we have generalized the classical dual concept lattices to a fuzzy environment, where we can use different adjoint triples defined on non-linear sets and no negation is needed.

This is very interesting because, at the moment, the introduction of this kind of concept lattice has been made using a negation on the considered carrier. Moreover, this environment provides a new point of view to obtain information from databases with both incomplete information and imprecise information, which will give more flexibility than the existence procedures. Indeed, some times could be more efficient to compute the dual concept than the standard one.

In the future, applications and practical examples will be studied where the theory developed in this paper can be used.

Acknowledgements

This work has been partially supported by Junta de Andalucía grant P09-FQM-5233, and by the EU (FEDER), and the Spanish Science and Education Ministry (MEC) under grant TIN2009-14562-C05-03.

References

[1] R. Bělohlávek. Fuzzy concepts and conceptual structures: induced similarities. In *Joint Conference on Information Sciences*, pages 179–182, 1998.

[2] R. Bělohlávek. Concept lattices and order in fuzzy logic. *Annals of Pure and Applied Logic*, 128:277–298, 2004.

[3] A. Burusco and R. Fuentes-González. The study of L-fuzzy concept lattice. *Mathware & Soft Computing*, 3:209–218, 1994.

[4] X. Chen and Q. Li. Construction of rough approximations in fuzzy setting. *Fuzzy Sets and Systems*, 158(23):2641–2653, 2007.

[5] Y. Chen and Y. Yao. A multiview approach for intelligent data analysis based on data operators. *Information Sciences*, 178(1):1–20, 2008.

[6] P. du Boucher-Ryana and D. Bridge. Collaborative recommending using formal concept analysis. *Knowledge-Based Systems*, 19(5):309–315, 2006.

[7] D. Dubois and H. Prade. Putting fuzzy sets and rough sets together. In R. Slowiński, editor, *Intelligent Decision Support*, pages 203–232. Kluwer Academic, Dordrecht, 2004.

[8] I. Düntsch and G. Gediga. Approximation operators in qualitative data analysis. In *Theory and Applications of Relational Structures as Knowledge Instruments*, pages 214–230, 2003.

[9] A. Formica. Concept similarity in formal concept analysis: An information content approach. *Knowledge-Based Systems*, 21(1):80–87, 2008.

 ISBN: 978-84-614-6167-7

[10] B. Ganter and R. Wille. *Formal Concept Analysis: Mathematical Foundation*. Springer Verlag, 1999.

[11] G. Gediga and I. Düntsch. Modal-style operators in qualitative data analysis. In *Proc. IEEE Int. Conf. on Data Mining*, pages 155–162, 2002.

[12] G. Georgescu and A. Popescu. Non-dual fuzzy connections. *Arch. Math. Log.*, 43(8):1009–1039, 2004.

[13] P. Hájek. *Metamathematics of Fuzzy Logic*. Trends in Logic. Kluwer Academic, 1998.

[14] G. Jiang, K. Ogasawara, A. Endoh, and T. Sakurai. Context-based ontology building support in clinical domains using formal concept analysis. *International Journal of Medical Informatics*, 71(1):71–81, 2003.

[15] S. Krajči. A generalized concept lattice. *Logic Journal of IGPL*, 13(5):543–550, 2005.

[16] H. Lai and D. Zhang. Concept lattices of fuzzy contexts: Formal concept analysis vs. rough set theory. *International Journal of Approximate Reasoning*, 50(5):695–707, 2009.

[17] G. L. Liu. Construction of rough approximations in fuzzy setting. *Information Sciences*, 178(6):1651–1662, 2008.

[18] J. Medina. Towards multi-adjoint property-oriented concept lattices. *Lect. Notes in Artificial Intelligence*, 6401:159–166, 2010.

[19] J. Medina and M. O. Aciego. Towards attribute reduction in multi-adjoint concept lattices. In *The 7th International Conference on Concept Lattices and Their Applications*, pages 92–103, 2010.

[20] J. Medina, M. Ojeda-Aciego, and J. Ruiz-Calviño. Formal concept analysis via multi-adjoint concept lattices. *Fuzzy Sets and Systems*, 160(2):130–144, 2009.

[21] J. Medina, M. Ojeda-Aciego, and P. Vojtáš. Similarity-based unification: a multi-adjoint approach. *Fuzzy Sets and Systems*, 146:43–62, 2004.

[22] Z. Pawlak. Rough sets. *International Journal of Computer and Information Science*, 11:341–356, 1982.

[23] V. Phan-Luong. A framework for integrating information sources under lattice structure. *Information Fusion*, 9:278–292, 2008.

[24] S. Pollandt. *Fuzzy Begriffe*. Springer, Berlin, 1997.

[25] A. M. Radzikowska and E. E. Kerre. A comparative study of fuzzy rough sets. *Fuzzy Sets and Systems*, 126(2):137–155, 2002.

[26] L. Wang and X. Liu. Concept analysis via rough set and afs algebra. *Information Sciences*, 178(21):4125–4137, 2008.

[27] X. Wang and W. Zhang. Relations of attribute reduction between object and property oriented concept lattices. *Knowledge-Based Systems*, 21(5):398–403, 2008.

[28] R. Wille. Restructuring lattice theory: an approach based on hierarchies of concepts. In I. Rival, editor, *Ordered Sets*, pages 445–470. Reidel, 1982.

[29] Y. Y. Yao. A comparative study of formal concept analysis and rough set theory in data analysis. *Lect. Notes in Artificial Intelligence*, 3066:59–68, 2004.

[30] Y. Y. Yao and Y. Chen. Rough set approximations in formal concept analysis. In *Transactions on Rough Sets V*, volume 4100 of *Lect. Notes in Computer Science*, pages 285–305, 2006.

 ISBN: 978-84-614-6167-7

Proceedings of the 11th International Conference on Computational and Mathematical Methods in Science and Engineering, CMMSE 2011 26–30 June 2011.

Python Interface-Library using OpenMP and CUDA for solving Nonlinear Systems

Héctor Migallón[1], Violeta Migallón[2] and José Penadés[2]

[1] *Departamento de Física y Arquitectura de Computadores, Universidad Miguel Hernández, 03202 Elche, Alicante, Spain*

[2] *Departamento de Ciencia de la Computación e Inteligencia Artificial, Universidad de Alicante, 03071 Alicante, Spain*

emails: hmigallon@umh.es, violeta@dccia.ua.es, jpenades@dccia.ua.es

Abstract

In this paper we present new features of PyPANCG. PyPANCG is a parallel library treated as a high-level interface for solving nonlinear systems. Using the new features, PyPANCG is able to exploit the parallelism offered by shared memory platforms and graphics processing units (GPUs). The new library still has two modules, PySParNLCG and PySParNLPCG, which include new features, both modules are backward-compatible with the earlier versions of PyPANCG. The PySparNLCG module parallelizes the conjugate gradient method for solving mildly nonlinear systems, and the PySParNLPCG module implements the preconditioning technique based on block two-stage methods. In order to create the high-level interfaces, we have chosen the Python language. Experimental results report the behavior and the parallel performance of our approach on both the shared memory platforms and the GPUs.

Key words: CUDA, OpenMP, parallel libraries, nonlinear algorithms, Python high-level interfaces

1 Introduction

In this paper we present new features of PyPANCG (http://atc.umh.es/PyPANCG), a Python based high-level parallel interface-library for solving mildly nonlinear systems of the form

$$Ax = \Phi(x), \tag{1}$$

where $A \in \Re^{n \times n}$ and $\Phi : \Re^n \to \Re^n$ is a nonlinear diagonal mapping, i.e., the ith component ϕ_i of ϕ is a function only of the ith component x_i of x.

 ISBN: 978-84-614-6167-7

This library, distributed as a standard Python package, provides parallel implementations of both the nonlinear conjugate gradient method (NLCG) and the nonlinear preconditioned conjugate gradient method (NLPCG). PyPANCG earlier versions could work with different tools to manage a distributed memory platform through *MPI* (www-unix.mcs.anl.gov/mpi). The PyPANCG current version allows to work with shared memory platforms through OpenMP and, using CUDA, this library is able to work with GPUs.

This paper is structured as follows. Section 2 introduces both the nonlinear conjugate gradient method (NLCG) parallelized in the PySParNLCG module of PyPANCG, and the nonlinear preconditioned conjugate gradient parallelized in the PySParNLPCG module. In Sections 3, 4 and 5 we explain the main tools used in order to build PyPANCG, the involved parameters and the way to implement the nonlinearity, respectively. In Section 6 some examples of using the features of PyPANCG are reported while in Section 7 the behavior of this library is illustrated by means of numerical experiments. Finally, concluding remarks are presented in Section 8.

2 Nonlinear methods

Consider the problem of solving the nonlinear system (1), where $A \in \Re^{n\times n}$ is a symmetric positive definite matrix and $\Phi : \Re^n \to \Re^n$ is a nonlinear function with certain local smoothness properties. Let $\Psi : \Re^n \to \Re$ be a nonlinear mapping and consider $\langle x, y\rangle = x^T y$ the inner product in $\Re^n$. The minimization problem of finding $x \in \Re^n$ such that

$$J(x) = \min_{y\in\Re^n} J(y), \tag{2}$$

where $J(x) = \frac{1}{2}\langle Ax, x\rangle - \Psi(x)$, is equivalent to find $x \in \Re^n$ such that $F(x) = Ax - \Phi(x) = 0$, where $\Phi(x) = \Psi'(x)$.
An effective approach for solving nonlinear system (1), by considering the connection with the minimization problem (2), is the Fletcher-Reeves version [5] of the nonlinear conjugate gradient method (NLCG), which takes the following form:

Algorithm 1 *(Fletcher-Reeves Nonlinear Conjugate Gradient)*
Given an initial vector $x^{(0)}$
$r^{(0)} = \Phi(x^{(0)}) - Ax^{(0)}$
$p^{(0)} = r^{(0)}$
For $i = 0, 1, \ldots,$ *until convergence*
$\alpha_i =\to$ *see below*
$x^{(i+1)} = x^{(i)} + \alpha_i p^{(i)}$
$r^{(i+1)} = r^{(i)} - \Phi(x^{(i)}) + \Phi(x^{(i+1)}) - \alpha_i A p^{(i)}$
Convergence test
$\beta_{i+1} = -\frac{\langle r^{(i+1)}, r^{(i+1)}\rangle}{\langle r^{(i)}, r^{(i)}\rangle}$
$p^{(i+1)} = r^{(i+1)} - \beta_{i+1} p^{(i)}$

 ISBN: 978-84-614-6167-7

Note that, in Algorithm 1, α_i is chosen to minimize the associated functional J in the direction $p^{(i)}$. This is equivalent to solve the one dimensional zero-point problem $\frac{dJ(x^{(i)}+\alpha_i p^{(i)})}{d\alpha_i} = 0$. From the definition of J it follows that

$$J(x^{(i)} + \alpha p^{(i)}) = \frac{1}{2}\left\langle A(x^{(i)} + \alpha_i p^{(i)}), x^{(i)} + \alpha_i p^{(i)} \right\rangle - \Psi(x^{(i)} + \alpha_i p^{(i)}).$$

Then a simple differentiation with respect to α_i yields

$$\frac{dJ(x^{(i)} + \alpha_i p^{(i)})}{d\alpha_i} = \alpha_i \left\langle Ap^{(i)}, p^{(i)} \right\rangle - \left\langle r^{(i)}, p^{(i)} \right\rangle + \left\langle \Phi(x^{(i)}) - \Phi(x^{(i)} + \alpha_i p^{(i)}), p^{(i)} \right\rangle,$$

where $r^{(i)} = \Phi(x^{(i)}) - Ax^{(i)}$ is the nonlinear residual.

On the other hand, it is easy to see that the second derivative with respect to α_i takes the form

$$\frac{d^2 J(x^{(i)} + \alpha_i p^{(i)})}{d\alpha_i^2} = \left\langle Ap^{(i)}, p^{(i)} \right\rangle - \left\langle \Phi'(x^{(i)} + \alpha_i p^{(i)})p^{(i)}, p^{(i)} \right\rangle.$$

Then, using the Newton method for solving the zero-point problem for α_i, we obtain $\alpha_i^{(k+1)} = \alpha_i^{(k)} - \delta^{(k)}$, where

$$\delta^{(k)} = \frac{dJ(x^{(i)} + \alpha_i^{(k)} p^{(i)})/d\alpha_i}{d^2 J(x^{(i)} + \alpha_i^{(k)} p^{(i)})/d\alpha_i^2} =$$

$$\frac{\alpha_i^{(k)} \left\langle Ap^{(i)}, p^{(i)} \right\rangle - \left\langle r^{(i)}, p^{(i)} \right\rangle + \left\langle \Phi(x^{(i)}) - \Phi(x^{(i)} + \alpha_i^{(k)} p^{(i)}), p^{(i)} \right\rangle}{\left\langle Ap^{(i)}, p^{(i)} \right\rangle - \left\langle \Phi'(x^{(i)} + \alpha_i^{(k)} p^{(i)})p^{(i)}, p^{(i)} \right\rangle}.$$

Note that in order to obtain $\delta^{(k)}$, the inner products $\langle Ap^{(i)}, p^{(i)} \rangle$ and $\langle r^{(i)}, p^{(i)} \rangle$ can be computed once at the first Newton iteration. Moreover $Ap^{(i)}$ is available from the conjugate gradient iteration.

In order to generate efficient algorithms to solve the nonlinear system (1), we have designed a parallel version of Algorithm 1 and a parallel nonlinear preconditioned conjugate gradient algorithm, based on both Algorithm 1 and a polynomial preconditioner type based on block two-stage methods [3]; see [4] and [7] for detailed description.

Preconditioning is a technique for improving the condition number (cond) of a matrix. Suppose that M is a symmetric, positive definite matrix that approximates A, but is easier to invert. We can solve $Ax = \Phi(x)$ indirectly by solving $M^{-1}Ax = M^{-1}\Phi(x)$. If $\text{cond}(M^{-1}A) << \text{cond}(A)$ we can iteratively solve $M^{-1}Ax = M^{-1}\Phi(x)$ more quickly than the original problem. In this case we obtain the following nonlinear preconditioned conjugate gradient algorithm (NLPCG).

Algorithm 2 *(Nonlinear Preconditioned Conjugate Gradient)*
Given an initial vector $x^{(0)}$
$r^{(0)} = \Phi(x^{(0)}) - Ax^{(0)}$
Solve $Ms^{(0)} = r^{(0)}$

 ISBN: 978-84-614-6167-7

$p^{(0)} = s^{(0)}$
For $i = 0, 1, \ldots,$ *until convergence*
$\alpha_i =\rightarrow$ *see Algorithm 1*
$x^{(i+1)} = x^{(i)} + \alpha_i p^{(i)}$
$r^{(i+1)} = r^{(i)} - \Phi(x^{(i)}) + \Phi(x^{(i+1)}) - \alpha_i A p^{(i)}$
Solve $Ms^{(i+1)} = r^{(i+1)}$
Convergence test
$\beta_{i+1} = -\frac{\langle s^{(i+1)}, r^{(i+1)} \rangle}{\langle s^{(i)}, r^{(i)} \rangle}$
$p^{(i+1)} = r^{(i+1)} - \beta_{i+1} p^{(i)}$

Since the auxiliary system $Ms = r$ must be solved at each conjugate gradient iteration, this system needs to be easily solved. Moreover, in order to obtain an effective preconditioner, it wants M to be a good approximation of A. One of the general preconditioning techniques for solving linear systems is the use of the truncated series preconditioning [1]. These preconditioners consist of considering a splitting of the matrix A as

$$A = P - Q \tag{3}$$

and performing m steps of the iterative procedure defined by this splitting toward the solution of $As = r$, choosing $s^{(0)} = 0$. It is well known that the solution of the auxiliary system $Ms = r$ is effected by $s = (I + R + R^2 + \ldots + R^{m-1})P^{-1}r$, where $R = P^{-1}Q$ and the preconditioning matrix is $M_m = P(I+R+R^2+\ldots+R^{m-1})^{-1}$, cf. [1]. In order to obtain the preconditioners, choosing $s^{(0)} = 0$, we use m steps of the block-Jacobi type two-stage methods toward the solution of $As = r$. In order to obtain the inner splittings of these block methods, incomplete LU factorizations are considered; see e.g., [4].

3 Development resources

This section analyzes the basic resources used in the building process of the designed library. The main language used for the development of the basic routines and on which the final library will be based is Fortran. However, C language is also used in order to develop CUDA-based routines. The desired objective is to unite the development features offered by Python in a single platform and to approach the execution features offered by, in this case, Fortran and CUDA.

In order to access the routines developed in Fortran from Python, the F2PY tool (cens.ioc.ee/projects/f2py2e) has been used. These routines were developed using OpenMP extensions to run on shared memory platforms. An enhanced feature is the least influence on the behavior of the method of both the use and handling of arrays or vectors and the communication between Python and Fortran. However, two equivalent options can still be used: the Python modules for vector management *Numeric* and *numarray* (*numarray* is part of *NumPy*). The new features are not able to use

 ISBN: 978-84-614-6167-7

main routines developed in Python. Hence, vector communications between languages remain an important aspect to consider in order to achieve the best performance.

On the other hand, to access the CUDA-based routines developed in C language from Python, the PyCUDA package was used. PyCUDA is a package that offers access to Nvidia's CUDA parallel computation API from Python in such a way that it is not necessary to access to a set of CUDA-based routines included in a library to link to from Python. The PyCUDA package uses CodePy, a C/C++ metaprogramming toolkit for Python. CodePy compiles C source code and dynamically loads it into the Python interpreter, a key aspect in the nonlinearity implementation.

4 Parameters of the methods and platform

This section deals with the parameters which have to be passed to the Python functions which solve a sparse nonlinear system using the NLCG or NLPCG method. The only indispensable parameters are the parameters of the system to be solved ($Ax = \phi(x)$), which are the size of the system, the matrix A stored in CSR (Compressed Sparse Row) format, and the nonlinear mapping $\phi(x)$. In addition the derivative of $\phi(x)$ ($\phi'(x)$) is required for computing δ as seen in Section 2. There is also a set of optional parameters to modify the NLCG and NLPCG methods. If values for the optional parameters are not specified, default values are used. The optional parameters are (see [7] for more information):

- The initial vector to start procedure (*initial_vector*).
- The stopping criterion to stop procedure (*global_stopping_error*).
- The stopping criterion to stop the iterative procedure to compute α (*alfa_stopping_error*).
- The maximum number of iterations performed in the iterative procedure to compute α (*iter_alfa*).
- Way to communicate integers from Python to Fortran or C (*trash_int*).
- Way to communicate doubles from Python to Fortran or C (*trash_double*).
- Level of the incomplete LU factorization performed in the NLPCG method (*level*).
- Number of outer iterations in the NLPCG method (*niter_2e*).
- Number of inner iterations in the NLPCG method (*val_q*).

Another important parameter, that the library can calculate, is the size of the problem assigned to each process; this is given by the parameter *block_dimensions*. This parameter is an integer vector whose dimension corresponds to the number of processes and which stores the block size assigned to each process. In the examples provided by PyPANCG, the parameter is internally calculated, such that a load balancing is achieved. On the other hand, this parameter has no relevance if CUDA is used, since

 ISBN: 978-84-614-6167-7

in this case the shared memory multiprocessor is not used even if it is available. In this case a single process manages the GPU computing.

The parameter *For_or_Py* selects the set of routines and the platform to be used. In [7] we can see the options to use in a distributed memory platform: *Python_full, Python, Fortran or Fortran_full*. The following options can be chosen with regard to this parameter in order to use a shared memory platform or a GPU:

1. *Fortran_mp*: The routines are codified in Fortran using OpenMP. Moreover ϕ and ϕ' are codified independently.
2. *Fortran_mp_full*: All of the routines are codified in Fortran using OpenMP but ϕ and ϕ' are not codified independently.
3. *GPU*: All of the routines are codified in C as CUDA kernels. Moreover ϕ and ϕ' are codified as CUDA kernels independently.

Using OpenMP or CUDA, the nonlinear functions must be codified in Fortran or C respectively. Using *Fortran_mp* or *Fortran_mp_full* implies codifying the nonlinear functions in Fortran and the compilation of the Fortran library linked to from Python. However, the use of *GPU* avoids the explicit recompilation of the nonlinear functions developed as CUDA kernels, by exploiting the CodePy features.

Finally, there are new parameters to work with the new options of the *For_or_Py* parameter, i.e. to work using OpenMP and CUDA. The first parameter, *nprocs_mp*, is the number of processes used in the shared memory platform when OpenMP is selected. The rest of the new parameters are used by CUDA. In a CUDA kernel calling, in addition to classical function parameters, there are two parameters that define the structure of threads that will be generated to run the CUDA kernel. Both parameters are the number of blocks to be generated (*grid*) and the number of threads in each block (*block*), see for example [8] to obtain detailed description. Moreover there are two global variables in order to tune the inner products, *VECTOR_N* and *ELEMENT_N*, see [4] for detailed description. Note that, both considered algorithms intensively use inner product, which is also a special operation involving a reduction process.

5 Encoding nonlinear functions

In a library for solving nonlinear systems it is important how to implement the nonlinearity of the problem to be solved. In [7] we show that PyPANCG can work either at component level and at vector level. However, when OpenMP or CUDA is used, for usability reasons and for the nature of the GPU computing, the library works at component level. The example below shows the Fortran code for the function $\phi(x)$ used in the examples of PyPANCG.

```
double precision function phi(input,trash_int,trash_double)
    implicit none
```

 ISBN: 978-84-614-6167-7

```
    real*8 input,trash_double(*),sc
    integer trash_int(*)
    sc = trash_double(1)
    phi = -sc*exp(input)
    return
```

The C CUDA kernel code to compute the same function is:

```
__device__ double Fi_x(double x,double sc){
    return (-sc*__expf(x))
}
```

We would like to note that both functions require a parameter transfer (sc) for the computation of ϕ. Using Fortran and OpenMP, in order to realize this transfer -both real values and integer values if needed- we use two vectors, one integer vector *trash_int* and one double precision real vector *trash_double*. These vectors are dynamic and thus all parameters required for the computation can be passed to functions ϕ and ϕ'. Naturally, these functions must always be implemented in order to adapt them to the problem to be solved. On the other hand, using CUDA the memory allocation and the GPU-CPU communication processes can be expensive, therefore the memory used is the strictly necessary memory. In the previous example we only communicate the necessary double parameter.

6 Python examples using OpenMP and CUDA

The use of the modules PySparNLCG and PySParNLPCG using OpenMP or CUDA is closely similar to earlier version presented in [7]. In order to use the library the size of the system (nrow), the matrix A in CSR format (tcol, trow, tval), and the nonlinear functions (ϕ and ϕ') must be passed at the very least, and optionally, the block size assigned to each process (block_dimensions). Moreover, as we have mentioned, additional parameters, if needed, can be passed by using the variables *trash_int* and *trash_double*. The following code shows the most simple NLCG function call using OpenMP.

```
1  from math import exp
2  import numpy
3  import PyPANCG
4  import PyPANCG.PySParNLCG as PySparNLCG

5  nprocs = 4
6  trash_double = numpy.zeros(((1),),float)
7  trash_double[0] = 6/(float(49)**3)
8  nrow = 125000

9  nrow,block_dimensions,bls = _
       PyPANCG.MakeBlockStructure(nrow=nrow)
10 nnz,tcol,trow,tval = PyPANCG.PartialMatrixA _
       (Mx=Mx,s=nrow,d=nrow)
```

 ISBN: 978-84-614-6167-7

```
11 x,error,time,iter = PySParNLCG.nlcg(nrow=nrow,tcol=tcol,trow=trow,tval=tval, _
              block_dimensions = block_dimensions,Fi_x=Fi_x,Fi_prime_x=Fi_prime_x, _
              For_or_Py='Fortran_mp',trash_double = trash_double,nprocs_mp=nprocs)
```

The number of processes to use is established in line 5. Note that the number of processes must be fixed before the block structure be defined in line 9. The matrix A is obtained in line 10 following the block structure above defined. This matrix is included in PyPANCG as an example or test. It is important to point out that in line 10 root process computes the full matrix A, we maintain the *PartialMatrixA* as routine name in order to accomplish the compatibility. In line 11, the actual call to the NLCG method takes place, whereby we assume that *Fi_x* (ϕ) and *Fi_prime_x* (ϕ') were developed in Fortran and the vector *trash_double* is passed, in this case of a single component. Note that OpenMP is not used in the development of *Fi_x* (ϕ) and *Fi_prime_x* (ϕ') because they are implemented at component level.

The most simple NLPCG using an OpenMP function call is similar to the NLCG example above showed. In this case, the PySParNLPCG module must be imported instead of the PySParNLCG module in line 4, and line 11 must be modified by the main function of the PySParNLPCG module.

```
4  import PyPANCG.PySParNLPCG as PySparNLPCG

11 x,error,time,iter = PySParNLPCG.nlpcg(nrow=nrow,tcol=tcol,trow=trow,tval=tval, _
              block_dimensions = block_dimensions,Fi_x=Fi_x,Fi_prime_x=Fi_prime_x, _
              For_or_Py='Fortran_mp',trash_double = trash_double,procs_mp=nprocs)
```

The following example shows the simplest example to call NLCG method using CUDA.

```
1  from math import exp
2  import numpy
3  import PyPANCG
4  import PyPANCG.PySParNLCG as PySparNLCG
5  import PyPANCG.PySparNLCG_ModGPU as PySParNLCG_ModGPU

6  nprocs = 1
7  trash_double = numpy.zeros(((1),),float)
8  trash_double[0] = 6/(float(49)**3)
9  nrow = 125000

10 nnz,tcol,trow,tval = PyPANCG.PartialMatrixA _
      (Mx=Mx,s=nrow,d=nrow)

11 x,error,time,iter = PySParNLCG.nlcg(nrow=nrow,tcol=tcol,trow=trow,tval=tval, _
              Fi_x=0,Fi_prime_x=0,For_or_Py='GPU', _
              trash_double = trash_double,nprocs_mp=nprocs)
```

In line 5 the *PyPANCG.PySparNLCG_ModGPU* module is imported. This module contains all CUDA kernels needed by the NLCG method, including *Fi_x* (ϕ) and

 ISBN: 978-84-614-6167-7

Fi_prime_x (ϕ'). Note that the encoding of the nonlinear functions must be done in that module without any compiling process. In the NLCG method, the CPU only performs the management of the GPU, therefore only one process is used (see line 6). The full matrix A is computed in line 10 by the CPU. In line 11, the call to the NLCG method to be computed in the GPU takes place. Note that the nonlinear functions *Fi_x* (ϕ) and *Fi_prime_x* (ϕ') are not defined in this call because, as we have mentioned, they are included inside CUDA kernels from *PyPANCG.PySparNLCG_ModGPU* module.

Finally, the most simple NLPCG function call, using CUDA, is similar to the previous example. In this case the PySParNLPCG module must be imported instead of the PySParNLCG module in line 4. The *PyPANCG.PySparNLPCG_ModGPU* module containing the CUDA kernels used by the NLPCG method, is also imported in line 5.

```
4  import PyPANCG.PySParNLPCG as PySparNLPCG
5  import PyPANCG.PySparNLPCG_ModGPU as PySParNLPCG_ModGPU
```

The main function for the NLPCG method takes, in this case, the following form:

```
11 x,error,time,iter = PySParNLPCG.nlpcg(nrow=nrow,tcol=tcol,trow=trow,tval=tval, _
              Fi_x=0,Fi_prime_x=0,For_or_Py='GPU', _
              trash_double = trash_double,procs_mp=nprocs)
```

In [4] some aspects of GPU computing are pointed out in order to tune the performance of the NLCG and NLPCG methods. Essentially this improvements are related to the inner products computation and the number of threads by block in the CUDA kernels. The following lines show an improvement of the NLPCG method with the use of some parameters (grid and block) as described in [4].

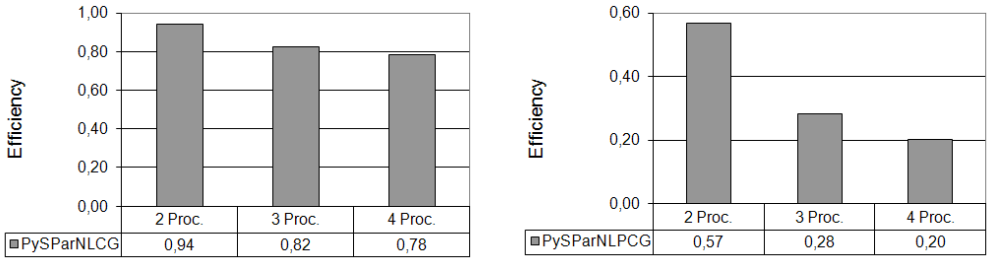
```
11 if (nrow == 125000)
              VECTOR_N = 128
              ELEMENT_N = 2916
              grid = (1458,1,1)
              block = (256,1)
12 x,error,time,iter = PySParNLPCG.nlpcg(nrow=nrow,tcol=tcol,trow=trow,tval=tval, _
              Fi_x=0,Fi_prime_x=0,For_or_Py='GPU', _
              trash_double = trash_double,procs_mp=nprocs, _
              block=block,grid=grid)
```

7 Numerical experiments

In order to illustrate the behavior of PyPANCG, we have tested the algorithms provided by this library on an Intel Core 2 Quad Q6600, 2.4 GHz, with 4 GB of RAM, called SULLI. The GPU available in SULLI is a GeForce GTX 280. The performed analysis is based on the run-times measured on the GeForce GTX 280, and on the parallel run-times measured on SULLI using OpenMP, when pure Fortran code (using OpenMP) or pure C code (using CUDA) are used, compared with the times obtained by PyPANCG.

As our illustrative example we have considered a nonlinear elliptic partial differential equation, known as the Bratu problem. In this problem, heat generation from

 ISBN: 978-84-614-6167-7



(a) PySParNLCG (b) PySParNLPCG

Figure 1: Efficiency using OpenMP, $n = 373248$.

a combustion process is balanced by heat transfer due to conduction. The three-dimensional model problem is given as

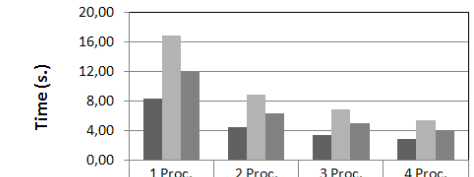
$$\nabla^2 u - \lambda e^u = 0, \tag{4}$$

where u is the temperature and λ is a constant known as the Frank-Kamenetskii parameter; see e.g., [2]. There are two possible steady-state solutions to this problem for a given value of λ. One solution is close to $u = 0$ and it is easy to obtain. A starting point near to the other solution is needed to converge to it. For our model case, we consider a 3D cube domain Ω of unit length and $\lambda = 6$. To solve equation (4) using the finite difference method, we consider a grid in Ω of d^3 nodes. This discretization yields a nonlinear system of the form $Ax = \Phi(x)$, where $\Phi : \Re^n \to \Re^n$ is a nonlinear diagonal mapping, i.e., the ith component Φ_i of Φ is a function only of the ith component of x. The matrix A is a sparse matrix of order $n = d^3$ and the typical number of nonzero elements per row of this matrix is seven, with fewer in rows corresponding to boundary points of the physical domain.

First, we analyze the efficiency for both methods using OpenMP. The NLCG method is performed by PySParNLCG module and the NLPCG method is performed by PySParNLPCG module. Optimal values of the parameters are used when the NLPCG method is computed, these parameters are the level of fill-in of the incomplete LU factorization (*level*), the number of outer iterations of the block two-stage method (*niter_2e*), and the number of inner iterations of the block two-stage method (*val_q*). On the other hand, in all experiments reported here the values of *global_stopping_error* and *alfa_stopping_error* are 10^-7, and we set *iter_alfa* parameter equal to 2. Figure 1 shows the efficiency of both methods using OpenMP and up to 4 cores available in SULLI. The efficiency behavior of the methods is not influenced by the use of the Python library; a pure Fortran code obtains similar efficiencies. For the NLCG method we obtain a good efficiency with a slight decrease when the number of processes is increased. However, as we showed in [6], the NLPCG method is a very good algorithm but with poor scalability, even for very large systems.

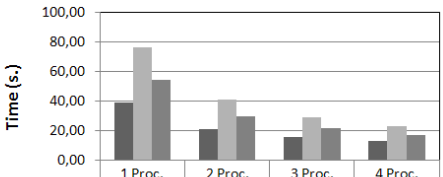
In order to select OpenMP, we have two options for parameter *For_or_Py* (see Section 4). In Figure 2 we can observe the behavior for both options. Setting *For-*

 ISBN: 978-84-614-6167-7



	1 Proc.	2 Proc.	3 Proc.	4 Proc.
Pure Fortran	8,33	4,43	3,45	2,87
Fortran_mp	16,83	8,92	6,81	5,36
Fortran_mp_full	11,96	6,4	4,94	4,03

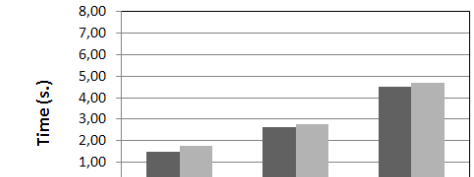
(a) 125000



	1 Proc.	2 Proc.	3 Proc.	4 Proc.
Pure Fortran	39,20	21,27	15,59	12,91
Fortran_mp	76,60	40,76	28,97	22,99
Fortran_mp_full	54,18	29,41	21,38	16,93

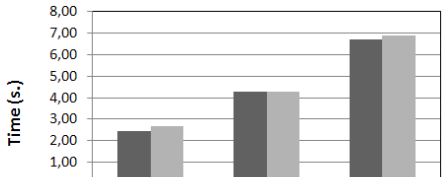
(b) 373248

Figure 2: PySparNLCG using OpenMP.



	n=373248	n=592704	n=884736
Pure CUDA	1,48	2,64	4,51
Python + CUDA	1,76	2,78	4,71

(a) PySParNLCG



	n=373248	n=592704	n=884736
Pure CUDA	2,46	4,26	6,70
Python + CUDA	2,66	4,30	6,88

(b) PySParNLPCG

Figure 3: PyPANCG using CUDA.

tran_mp_full option we obtain results closer to those obtained with pure Fortran than using *Fortran_mp* option. Note that the development effort using *Fortran_mp_full* is higher than using *Fortran_mp*.

Finally, we compare the results obtained by both modules when CUDA is used. Concretely, Figure 3 shows the results of these PyPANCG modules compared with a pure CUDA code for several problem sizes. As it can be appreciated both implementations report similar execution times.

8 Conclusion

In this paper we have presented new features of PyPANCG, a Python library-interface that implements both the conjugate gradient method and the preconditioned conjugate gradient method for solving nonlinear systems. The aim of this library is to develop high performance scientific codes for high-end computers hiding many of the underlying low-level programming complexities from users with the use of a high-level Python interface. The new features are designed to allow PyPANCG to be able to work on both shared memory platforms and GPUs. We have described the use of the library and its advantages in order to get fast development. The library has been designed for

 ISBN: 978-84-614-6167-7

adapting to different stages of the design process, depending on whether the purpose is computational performance or fast development. We have achieved both objectives at once using the GPU as a computation platform, which is also the platform on which the proposed algorithms have better performance.

Acknowledgements

This research was supported by the Spanish Ministry of Science and Innovation under grant number TIN2008-06570-C04-04.

References

[1] L. Adams, *M-step preconditioned conjugate gradient methods*, SIAM Journal on Scientific and Statistical Computing **6** (1985) 452–462.

[2] B. M. Averick, R. G. Carter, J. J. More and G. Xue, *The MINPACK-2 Test Problem Collection*, Technical Report MCS-P153-0692, Mathematics and Computer Science Division, Argonne, 1992.

[3] R. Bru, V. Migallón, J. Penadés and D.B. Szyld, *Parallel, Synchronous and Asynchronous Two-Stage Multisplitting Methods*, Electronic Transactions on Numerical Analysis **3** (1995) 24–38.

[4] V. Galiano, H. Migallón, V. Migallón and J. Penadés, *GPU-Based Parallel Nonlinear Conjugate Gradient Algorithms*, Proceedings of the Second International Conference on Parallel, Distributed, Grid and Cloud Computing for Engineering (2011) Paper 24.

[5] R. Fletcher and C. Reeves, *Function Minimization by Conjugate Gradients*, The Computer Journal **7** (1964) 149–154.

[6] H. Migallón, V. Migallón, J. Penadés, *Parallel Nonlinear Conjugate Gradient Algorithms on Multicore Architectures*, Proceedings of the 9th International Conference on Computational and Mathematical Methods in Science and Engineering (2009) 689–700.

[7] H. Migallón, V. Migallón and J. Penadés, *PyPANCG: A Parallel Python Interface-Library for solving Mildly Nonlinear Systems*, Proceedings of the 10th International Conference on Computational and Mathematical Methods in Science and Engineering (2010) 646–657.

[8] NVIDIA Corporation, *NVIDIA CUDA C Programming Guide, Version 3.2, 2010*, `http://developer.download.nvidia.com/compute/cuda/3_2/toolkit/docs/CUDA_C_Programming_Guide.pdf`

 ISBN: 978-84-614-6167-7

Proceedings of the 11th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2011
26–30 June 2011.

Local versus Global Implementation of Hyperspectral Anomaly Detection Algorithms: A Parallel Processing Perspective

J. M. Molero[1], E. M. Garzón[1], I. García[2] and A. Plaza[3]

[1] *Department of Computer Architecture and Electronics, University of Almería, Spain*

[2] *Department of Computer Architecture, University of Málaga, Spain*

[3] *Hyperspectral Computing Laboratory, University of Extremadura, Spain*

emails: jmp384@ual.es, gmartin@ual.es, igarciaf@uma.es, aplaza@unex.es

Abstract

Anomaly detection is an important task for remotely sensed hyperspectral data exploitation. Its goal is to identify anomalous pixels, i.e. pixels with spectral characteristics different to those of its neighboring pixels. In this paper, we explore global versus local approaches for implementation of anomaly detection algorithms. Our introspection is that global approaches exhibit more complexity in parallel implementation due to intensive communications between the nodes of the parallel system, while local approaches can offer improved detection capabilities and, at the same time, be more suited for parallel implementation. Experimental results with synthetic and real images are given to validate these remarks.

Key words: Hyperspectral imaging, anomaly detection, parallel computing.

1 Introduction

Hyperspectral imaging instruments are now able to record the visible and near-infrared spectrum (wavelength region from 0.4 to 2.5 micrometers) of the reflected light of an area 2 to 12 kilometers wide and several kilometers long using hundreds of spectral bands [1]. The resulting "image cube" is a stack of images in which each pixel (vector) has an associated spectral signature or *fingerprint* that uniquely characterizes the underlying objects [2]. The data volume typically comprises several GBs per flight [3].

 ISBN: 978-84-614-6167-7

The array of techniques that can be applied for hyperspectral image processing is enormous [4]. Anomaly detection is highly relevant in many application domains, including detection of forest fires, pollutants in water, rare minerals in geology, or military targets in defense and security applications. [5]. A popular algoritm for hyperspectral anomaly detection was developed by Reed and Xiaoly (RX algorithm [6]). This algorithm follows a global approach, i.e., it uses the full hyperspectral image in order to derive statistics. To adapt this algorithm to real-time application scenarios in which the full hyperspectral image may not be available, variations of this algorithm have been focused on using small subsets of the image (e.g. those already collected by the imaging instrument [2]. In this case, a sliding-window approach can be used to include spatial information in the analysis [7]. Depending on the complexity and dimensionality of the input scene, the aforementioned algorithms may be computationally expensive, a fact that limits the possibility of utilizing those algorithms in time-critical applications [3].

In this paper, we explore global versus local approaches for the implementation of hyperspectral anomaly detection algorithms. Our introspection is that global approaches exhibit more complex parallelization that requires intensive communications between the nodes of the parallel system, while local approaches may offer improved detection capabilities and still be more suitable for parallel implementation.

2 RX Algorithm: Local versus Global Implementations

The RX algorithm has been widely used in signal and image processing [6]. The filter implemented by this algorithm is referred to as RX filter (RXF) and defined by the following expression:

$$\delta^{\mathrm{RXF}}(\mathbf{x}) = (\mathbf{x}-\mu)^T \mathbf{K}_{n\times n}^{-1}(\mathbf{x}-\mu), \tag{1}$$

where $\mathbf{x} = \left[x^{(0)}, x^{(1)}, \cdots, x^{(n)}\right]$ is a sample, n-dimensional hyperspectral pixel (vector), μ is the sample mean and $\mathbf{K}_{n\times n}$ is the sample data covariance matrix. As we can see, the form of δ^{RXF} is actually the well-known Mahalanobis distance [2]. It is important to note that the images generated by the RX algorithm are generally gray scale images. In this case, the anomalies can be categorized in terms of the probability value returned by RXF.

In this work, we consider also two variations of the original RX implementation. The first one replaces the sample covariance matrix $\mathbf{K}_{n\times n}$ by the sample correlation matrix $\mathbf{R}_{n\times n}$. Since the covariance relies on the full image statistics, using the correlation matrix instead allows adapting the algorithm to a real-time scenario in which not the full data set might be available at a certain point. In this case, the correlation matrix is calculated using only the available samples. The second variation considers a local approach to determine whether or not the image pixels are anomalous. The idea is to place a window about each pixel in the image and use local image statistics (i.e., μ, $\mathbf{K}_{n\times n}$ and $\mathbf{R}_{n\times n}$ would be calculated in the local window instead of in the full image). Our introspection is that this approach

 ISBN: 978-84-614-6167-7

is more suitable for parallel implementation but it can also suffer from several problems, including the fact that window pixel vectors are almost never statistically independent (thus introducing conditioning problems in the calculated matrices), or the fact that outliers (i.e., the anomalies we are looking for) can compromise the integrity of the local statistics, particularly the covariance matrix.

3 Experiments and Discussion

In this section we inter-compare local versus global implementations of different variations of the RX algorithm for anomaly detection, including implementations with the covariance matrix and with the correlation matrix, and also implementations using the full hyperspectral image and small windows centered around each image pixel when calculating the RX statistics. Our experiments have been conducted with a synthetic and a real hyperspectral image. The real hyperspectral image was colected by the HYperspectral Digital Image Collection Experiment (HYDICE) described in [2]. It is an image scene with a size of 64×64 pixels with 15 panels in the scene and a ground-truth map indicating the position of the panels. The real hyperspectral image was acquired by 210 spectral bands with a spectral coverage from 0.4 to 2.5 microns. Low signal/high noise bands: 1-3 and 202-210; and water vapor absorption bands: 101-112 and 137-153 were removed prior to experiments, so a total of 169 bands were used. The spatial resolution is 1.56 meters and the spectral resolution is 10 nanometers. The synthetic hyperspectral image was designed to mimic the spatial and the spectral characteristics of the original hyperspectral image using different spectral signatures, in this case, minerals obtained from the U.S. Geological Survey (USGS)[1]. It has a size of 64×64 pixels with 224 spectral bands. Random noise with signal-to-noise ratio of 30:1 [2] was then added to the synthetic scenes to simulate the contribution of instrumental noise.

Fig. 1 shows the receiver operating characteristics (ROC) curve obtained after comparing the output provided by different implementations of the RX algorithm (using the covariance versus the correlation matrix, and also using a global versus a local approach) with the ground-truth available for the scene [6]. The ROC curve plots the probability of detection versus the probability of false alarm, so that the area under the ROC curve serves as a good indicator to evaluate the probability of detection achieved by the different implementations independently of the selection of specific threshold values. On the other hand, Fig. 2 shows the ROC curves obtained for the real HYDICE hyperspectral image for which ground-truth information is also available. As shown by Figs. 1 and 2, the local versions generally offer the best detection results (area under the ROC curve). This is an important conclusion since the local versions are *embarrasingly parallel* (i.e., they can scale to any number of processors in a parallel system due to the lack of inter-processor communications

[1]The USGS mineral library can be downloaded from http://speclab.cr.usgs.gov.

 ISBN: 978-84-614-6167-7

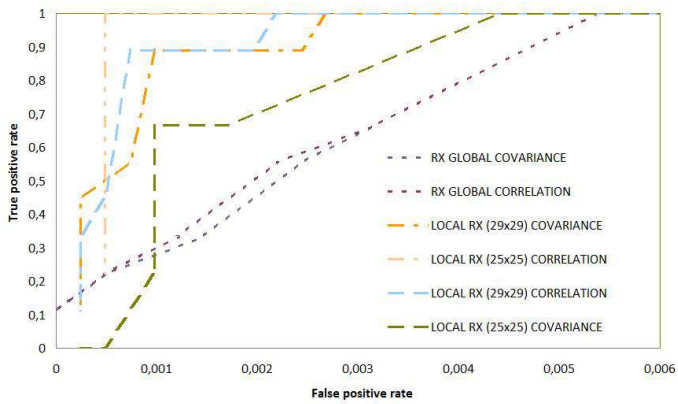


Figure 1: ROC curve illustrating the performance of different anomaly detection algorithms with a synthetic hyperspectral image.

[8]). Also, the good results provided by the correlation matrix in the local implementation allow us to circumvent the problems related with the conditioning of the matrix used in the calculations. Further experiments will be focused on evaluating the performance of the different approaches on parallel computing architectures.

References

[1] A. F. H. Goetz, G. Vane, J. E. Solomon, and B. N. Rock, "Imaging spectrometry for Earth remote sensing," *Science*, vol. 228, pp. 1147–1153, 1985.

[2] C.-I. Chang, *Hyperspectral Imaging: Techniques for Spectral Detection and Classification.* Norwell, MA: Kluwer, 2003.

[3] A. Plaza and C.-I. Chang, *High performance computing in remote sensing.* Boca Raton: CRC Press, 2007.

[4] A. Plaza, J. A. Benediktsson, J. Boardman, J. Brazile, L. Bruzzone, G. Camps-Valls, J. Chanussot, M. Fauvel, P. Gamba, J. Gualtieri, M. Marconcini, J. C. Tilton, and G. Trianni, "Recent advances in techniques for hyperspectral image processing," *Remote Sensing of Environment*, vol. 113, pp. 110–122, 2009.

[5] A. Paz and A. Plaza, "Clusters versus GPUs for parallel automatic target detection in remotely sensed hyperspectral images," *EURASIP Journal on Advances in Signal Processing*, vol. 2010, pp. 1–18, 2010.

[6] I. Reed and X.Yu, "Adaptive multiple-band cfar detection of an optical pattern with unknown spectral distrihution." *IEEE Trans. Acoustics, Speech and Signal Processing*, vol. 38, pp. 1760–1770, 1990.

[7] Y. P. Taitano, B. A. Geier, and K. W. B. Jr., "A locally adaptable iterative rx detector," *EURASIP Journal on Advances in Signal Processing*, vol. 2010, pp. 1–10, 2010.

[8] J. M. Molero, A. Paz, E. M. Garzon, J. A. Martinez, A. Plaza, and I. Garcia, "Fast anomaly detection in hyperspectral images with rx method on heterogeneous clusters," *Journal of Supercomputing, accepted for publication*, 2010.

 ISBN: 978-84-614-6167-7

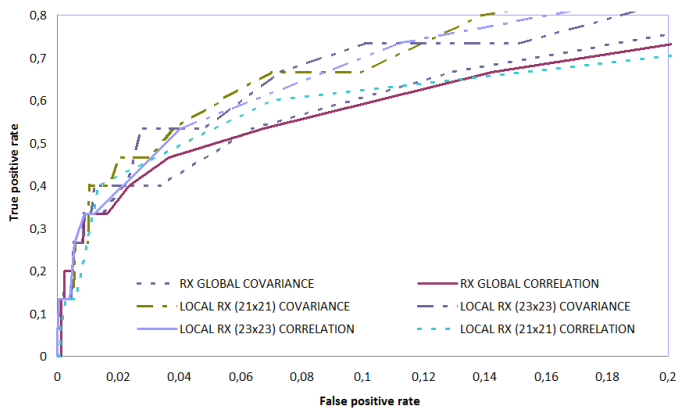


Figure 2: ROC curve illustrating the performance of different anomaly detection algorithms with a real hyperspectral image.

 ISBN: 978-84-614-6167-7

Proceedings of the 11th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2011
26–30 June 2011.

Towards an efficient execution of Multiple Sequence Alignment in multi-core systems

Alberto Montañola[1], Concepció Roig[1], Porfidio Hernández[2], Antonio Espinosa[2], Yandi Naranjo[2] and Cedric Notredame[3]

[1] *Distributed Computing Group*
Computer Science and Industrial Engineering Department, Universitat de Lleida, Spain

[2] *Computer Architecture and Operating Systems Department, Universitat Autònoma de Barcelona, Spain*

[3] *Bioinformatics Programme, Centre de Regulació Genòmica (CRG-UPF), Barcelona, Spain*

emails: alberto@diei.udl.cat, roig@diei.udl.cat, Porfidio.Hernandez@uab.es, antonio.espinosa@caos.uab.es, yandi.naranjo@caos.uab.es, cedric.notredame@crg.es

Abstract

Nowadays multi-core systems are being proposed as a new environment, in order to increase performance of modern computer systems. The new challenge is to use them efficiently during the execution of applications. In this work we focus on the exploitation of multi-core systems in the execution of Multiple Sequence Alignment (MSA) applications. The MSA is a high computing demanding process, were the aim is to find similar regions in biological sequences. The goal is to align as much sequences as possible in an acceptable amount of time and with a level of quality that makes the alignment biologically meaningful. This paper presents an efficiency study of different implementations based on T-Coffee, an application that carries out MSA. The study is focused on finding new optimizations that may improve the average execution time on multi-core parallel systems. In order to evaluate the efficiency and scalability of T-Coffee, memory and total computing time have been measured on different system configurations by gradually increasing the number of input sequences. We have found that the current message passing implementations have several issues that affect negatively the performance and scalability of T-Coffee. Finally we present a proposal modification over the T-Coffee parallelization, where threading functions were introduced with the aim of optimizing the execution of T-Coffee in multi-core systems.

 ISBN: 978-84-614-6167-7

1 Introduction

Traditionally, the increase of performance in computer systems has been based on the increment of the processor clock frequency. This has a technological limit, being that increasing the frequency causes an increment of the energy cost and the dissipated heat by the system [hill08]. Currently, multi-cores (up to 12 cores per processor) are part of most general purpose systems [geer05].

The current challenge is to use multi-core systems in an efficient way during the execution of applications. There is a wide range of scientific problems that can benefit from an efficient exploitation of parallel systems, and specially of multi-core systems. One of the scientific fields with large scale problems to be solved, is computational biology. In this work, we focus on the exploitation of multi-core systems for solving the biological problem of Multiple Sequence Alignment (MSA). MSA is one of the most used techniques in the study of biological sequences of genomes. The main goal is to find coincidences between different sequences. This allows biologists to study the differences between genomes and study the evolution of the studied species. The current MSA implementations are few, and they are not completely written to fully use efficiently the current multi-core architectures. MSA implementations found in literature are MUSCLE [edgar04], MAFFT [katoh08] and ClustalW [larkin07].

One MSA implementation currently used by the bioinformatics community is T-Coffee [notredame00]. T-Coffee, a sequential sequence aligner, has been designed to work on standalone systems with one processor. Although the latest implementation creates as many sequential processes as processors cores are found in the system, the implementation used is not the most efficient for a cluster environment. Parallel-T-Coffee [zola07] and Clus-T-Coffee [naranjo09] are two parallel implementations based on T-Coffee, that can be deployed over a cluster. The first one is a parallel re-implementation of T-Coffee, when the second one, performs a different heuristic where the set of input sequences is clusterized and processed by different parallel instances of T-Coffee. The goal of this work is to analyse the behaviour of these implementations on different systems, in order to detect possible ways to improve the overall process.

In order to achieve our goal, we have performed an study of both parallel implementations of T-Coffee in a set of different system configurations with different input data. Then, the results are studied in order to determine the benefits and drawbacks of each implementation. After this study, we are proposing the implementation of a shared memory version of T-Coffee combining threads with MPI (messaging passing interface).

This paper is organized as follows: Section 2 presents the application under study. Section 3 presents the experimentation and results of the study. Section 4 presents a proposed implementation with threads. Finally, Section 5 presents the main conclusions.

 ISBN: 978-84-614-6167-7

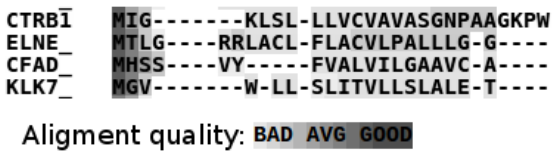


Figure 1: Example of a multiple alignment of a set of sequences

2 Multiple Sequence Alignment

The Multiple Sequence Alignment problem (MSA) consists in finding which patterns have in common at least three or more biological sequences. These sequences are formed by a structure of proteins, or nucleic acids DNA/RNA, and they are represented as plain ASCII characters. For example, on DNA sequences, there are four possible bases, Adenine, Thymine, Cytosine and Guanine. These will be encoded with the characters A,T,C and G respectively. The generated alignment will show shared subsequences from a common ancestor and insertions, deletions or different bases representing alterations like mutations or single nucleotide polymorphisms.

In Figure 1, we can see the output of a MSA of some sequences, the displayed sequences are a portion of a 100 sequences output file. Matching characters are aligned on the same column, gaps, insertions and/or deletions of characters may be performed to enforce the alignment of the sequence. A gap is a space between two characters, a insertion is the addition of one or more characters, and a deletion is the suppression of one or more characters. On this example, the alignment quality is represented in a different tonality color. The color tone goes from light to dark, were the light colors are the worst values of quality and dark ones are the bests one.

2.1 T-Coffee

T-Coffee (Tree-based Consistency Objective Function For alignment Evaluation), is the MSA application most extensively used among biologists [notredame00]. The original standalone version of T-Coffee is a sequential implementation, although there are some parallelization optimizations performed to use a multi-core system only. The other implementations presented on the next sections, use the message-passing library MPI, to parallelize the different stages of the algorithm.

T-Coffee can be used to align Protein, DNA and RNA sequences, and it can also combine the output generated by other MSA like Clustal, Muscle, Mafft, etc.. into one unique alignment using another mode of operation called M-Coffee. It processes input plain text files containing unaligned sequences, and generates another file with the resulting

 ISBN: 978-84-614-6167-7

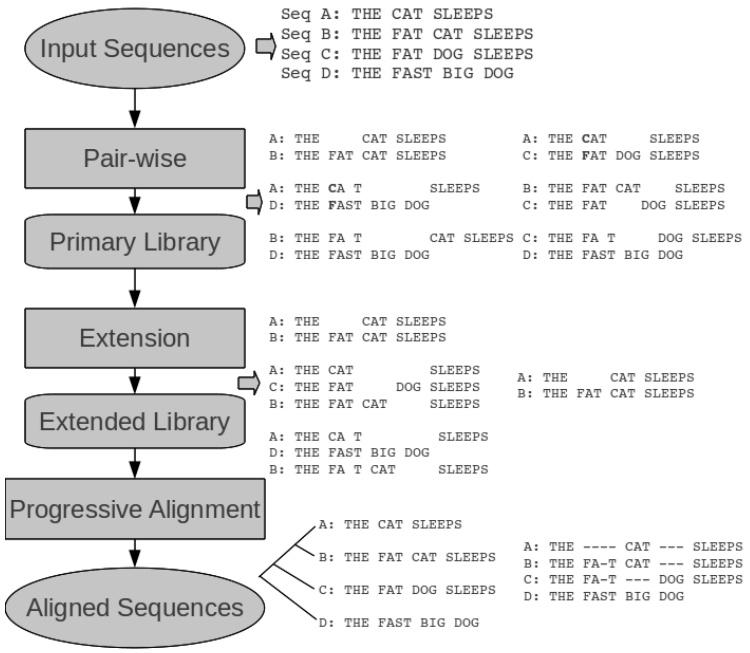


Figure 2: Steps and example of T-Coffee (TC) process. The circle represents a set of sequences, the square a computational process or step and the rounded squares are the name of the data structures (library) generated by the process

alignment of all the involved sequences. It tries to improve the accuracy of the results taking into account biological information obtained from pair-wise alignments.

In T-Coffee, sequences are processed through the three following stages: The pairwise alignment, the extension and the progressive alignment. Figure 2 illustrates the process of these stages of T-Coffee for a example of four input sequences (A, B, C, D).

- Pairwise alignment: Construction of library, referred as Primary Library (PL), containing all the pair-wise alignments from all input sequences. In our example, all the pairs of A, B, C and D sequences are aligned. Information about $N(N-1)/2$ pairs is stored in this library, where N is the number of input sequences. Alignments are represented as a list of matching pairs of residues. Each of these correlations is considered by T-Coffee as a restriction. A weight value from a scoring scheme is assigned to each restriction. The computing cost of this phase is very high because it is necessary to obtain the pair-wise alignments for all input sequences. The alignment of two sequences in particular has a computing complexity in space of $O(L)$ and in time $O(L^2)$, here L is the average length of sequences to align. Aligning N sequences

 ISBN: 978-84-614-6167-7

requires $N(N-1)/2$ combinations. This is equivalent to a complexity in general of $O(N^2)$, in time of $O(N^2L^2)$ and in space $O(N^2+L)$.

- Extension: The extension is the heuristic used by T-Coffee to adjust the restrictions group of the PL into multiple sequence alignments. The analysis in the extension is based on taking from the PL each pair of aligned residues in two sequences and check whether these are aligned with residue of the third sequence. The weight of the aligned residue in the third sequence in the PL is added to the weight assigned to the pair of residues in PL. The process in the extension is made for each pair of residues with all the input sequences. The cost of this process is very high. The extension is designed to generate a list of restrictions used to evaluate the construction of the final multiple sequence alignment. The complexity is in time $O(N^3L^2)$ and in space $O(N^3+L)$. The generated library, is called extended library (EL).

- Progressive Alignment: To obtain the progressive alignment it is necessary to generate a distance matrix (DM) from all input sequences. The DM is built with the values of similarity between pairs of sequences and is used to construct a guide tree. The guide tree is build as a Directed Acyclic Graph [kwok99] and its precedences are used to establish the order in the progressive alignment. In the example of the illustration the order established by the guide tree indicates a first alignment of sequences A and B, next with C and finally with D.

2.2 Parallel-T-Coffee

PTC, see [zola07], is a parallel implementation of T-Coffee using MPI. It is based in T-Coffee version 3.79. The constraints library is distributed across the different tasks, and performs alignment operations in parallel using dynamic scheduling techniques. As seeing in Figure 3, PTC performs the same process as T-Coffee but parallellized across a set of different tasks whose computation can be distributed among different nodes.

- Library Generation: In PTC the library generation is computed in a distributed form. First the pairwise computation is distributed, then the constraints are grouped and reweighted. Library extension is not performed as in original T-Coffee, it is postponed and done on the fly in the progressive alignment phase. Finally, a 3d lookup table is built from the library. Caching techniques are applied on this table.

- Progressive Alignment: Is the most difficult step to parallelize. Computations in PA follow a tree order, thus its parallelization is reduced to a DAG (Directed acyclic graph) scheduling problem [kwok99].

 ISBN: 978-84-614-6167-7

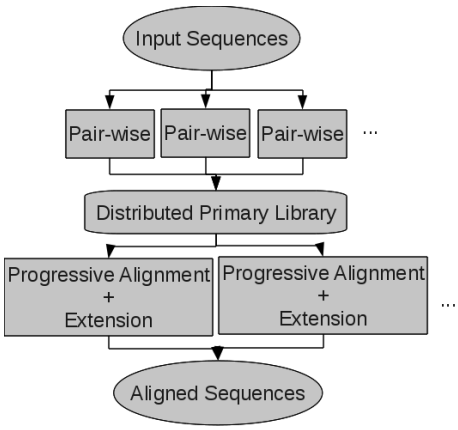


Figure 3: Steps of Parallel-T-Coffee (PTC) process

2.3 Clus-T-Coffee

CTC is a new parallel implementation of T-Coffee developed in our group [naranjo09]. It is based on version 7.81 of T-Coffee. It uses a divide and conquer technique. See Figure 4 for the schematics of the process. The main problem is divided in a set of small problems. In CTC, a new phase is introduced. The clustering phase builds a set of different clusters (groups) of sequences. The number of generated clusters and the amount of sequences depends on a cut-off value. This value is used to determine the similarity of the sequences to group in each cluster. As the initial task to compute the clusters, CTC needs to compute the pair-wise alignments between all input sequences. In the following phase, each node will receive a precomputed cluster and will run the whole T-Coffee algorithm on it, at the end it will generate a partial alignment. Finally all partial alignments are merged on a last invocation of T-Coffee.

3 Testing the performance of MSA

One of the basic goals of biologists researching on MSA, is to be able to align as much sequences as possible with an acceptable quality level. Thus, we carried out a test of performance of these implementations of T-Coffee in order to study the consumption of memory usage and computing time of CTC and PTC, in order to find optimizations to improve its efficiency.

In order to perform the experimentation, a set of system monitoring tools were used to extract information about CPU and memory usage. A set of different tests were performed on one of the clusters of the research group. This cluster had 20 nodes and one frontal, but

 ISBN: 978-84-614-6167-7

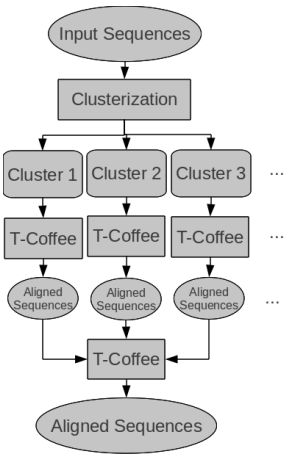


Figure 4: Different steps of Clus-T-Coffee (CTC) process

in our experiments we are going to use only up to four nodes per test, all nodes were Intel Core 2 Quad Q6600 @ 2.4GHz, they had 4 GBytes of RAM, 150 GBytes of disk and the network speed was at 1 Gbit. The version of T-Coffee used in all tests is 8.99.

The following configurations were used:

- 4 processes in total:
  - CTC and PTC executed in one node and four MPI processes per node.
  - CTC and PTC executed in four nodes and one MPI process per node.
- 8 processes in total:
  - CTC and PTC executed in two nodes and four MPI processes per node.

On these tests we can see the impact of the network communications, notice that we are intentionally leaving 3 cores of the system idle on some tests.

We have generated a set of 50 input files containing from 100 to 1000 sequences. The origin data used to generate these datasets comes from different BAliBASE sequence files. BAliBASE [thompson99], is a benchmarking library which contains several input sequences divided in a different set of benchmarks. There is not a fixed sequence length on these tests, as different lengths were concatenated. BAliBASE is used to test the quality level of a set of precomputed MSAs, its benchmarking tool gives an score of how good is an alignment. In our study, we are only using BAliBASE sequences for generating a set of input sequences.

Table 1 shows a global summary of all the performed tests in our study. For each test it is shown the number of MPI processes that were launched, the system configuration

 ISBN: 978-84-614-6167-7

Table 1: Relation of different tests, xn/yp where x is number of nodes and y number of MPI processes launched

Number Processes.	Test	Max. Seqs.
1	TC	300
4	PTC 1n/4p	500
4	CTC 1n/4p	1000
4	PTC 4n/1p	400
4	CTC 4n/1p	700
8	PTC 2n/4p	400
8	CTC 2n/4p	200

expressed in terms of number of nodes and number of processors per node, and the maximum number of sequences that we were able to execute for each test. Tests with a higher number of sequences failed mainly due to complete exhaustion of all the system memory, or due to exhaustion of the imposed time limit of 60 hours.

As can be observed in the table, the sequential implementation of T-Coffee (TC) is able to process a maximum of 300 sequences. The parallel versions PTC and CTC are able to increase this limit up to different values depending on the implementation. These differences between the parallel implementations are caused by two reasons. First, on PTC's implementation, as seen before, memory handling of the constraints library allows a bit more of space for processing more sequences. Second, when it runs with more than one node, the penalty of the message passing communications is so high that it increases considerably its processing time, causing the expiration of the time limit that we set on all tests. CTC handling of memory and communications is different than in PTC, due to the way the problem is divided into a subset of different parallel problems, but its limitation to process sequences is due to some stability problems in the current implementation. Thus stability of CTC decreases when the number of processes is increased.

A more detailed analysis of performance, while varying the number of processed sequences, is shown in Figure 5 and Figure 6. Relevant results found are: when CTC processes more than 300 sequences, it starts to consume less memory than PTC. Furthermore CTC is consuming less time than PTC to complete. This difference, as commented before, is caused by the fact that PTC is continuously communicating to access the distributed constraints list. On the other hand CTC only communicates at the beginning and at the end of its execution because its parallel execution does not have any kind of data requirement from the other siblings. In addition, we expected that with the introduction of more execution nodes the application would have to run faster, but the results show higher computing times

 ISBN: 978-84-614-6167-7

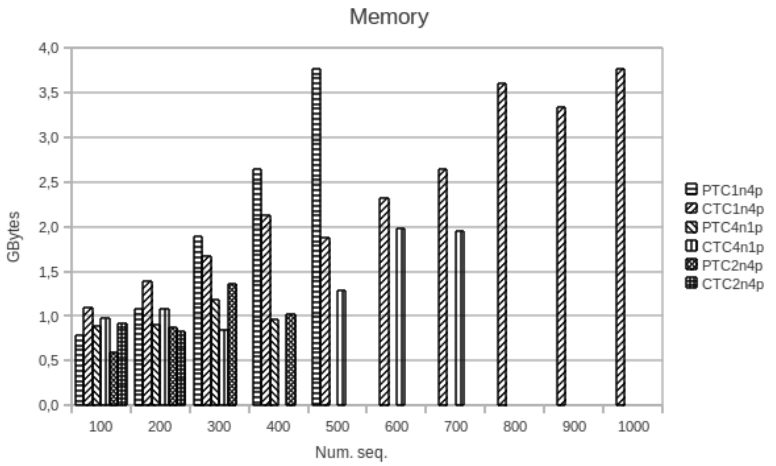


Figure 5: Memory required to process n sequences

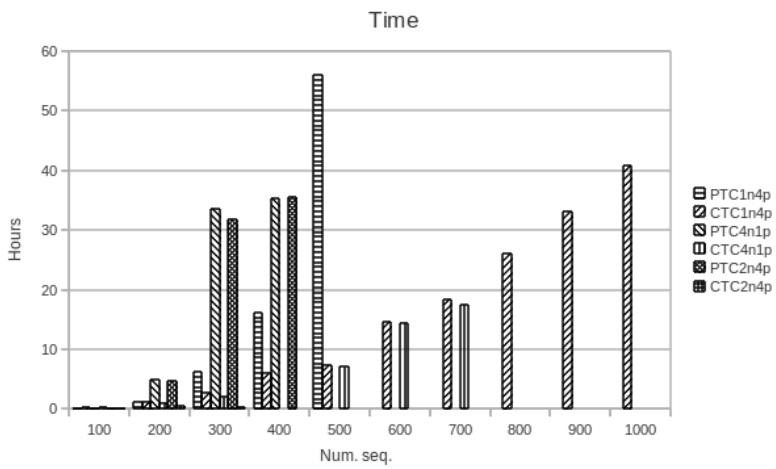


Figure 6: Computing time required to process n sequences

 ISBN: 978-84-614-6167-7

due to communication overheads. Summarizing the results obtained by these tests, we can state the following things:

- By comparing executions in a single node with ones done in 2 or 4 nodes, both CTC and PTC have some communication issues, mainly because single node executions are able to process more sequences.
- CTC consumes less memory than PTC and its performance is more predictable than PTC.
- PTC takes more time to process than CTC, mainly due to the communications.

In order to improve the memory usage, and reduce the overhead of the current available parallelizations we propose to optimize them using the multithreading pthreads POSIX standard as is exposed in next section.

4 Towards a pthreads implementation

One of the main problems of T-Coffee is the memory consumption, and one approach left to resolve is the better usage of parallelization of the whole process in a multi-core machine. It is already possible to run T-Coffee in parallel in a single machine. The original version is capable to launch a set of processes using the system fork call. PTC and CTC are capable of running multiple MPI processes on the same machine. On all these parallel implementations we are not making use of a shared memory model such as threads. At the same time the overhead generated by the sub-process creation and by the interprocess communications should be reduced by applying a different model, more suitable for a multi-core system. Thus, we proposed a reorientation of the implementation of parallel versions of T-Coffee in such a way that they rely on the best of the studied implementations by combining a message passing technology such as MPI with a threading API such as pthreads. This proposed implementation substitutes the usage of the *fork/wait* system calls currently used in the original T-Coffee implementation with *pthread_create/pthread_join* with the goal of reducing the overhead of such parallelization on a per node based execution. The current prototype has been focused in the parallelization of the construction of the primary library, the pair-wise initial step of T-Coffee.

In order to perform the transformation of the current code, a thread safety study has been performed. In this study, we found a set of different problems that are adding complexity to the implementation of a fully multi-threaded implementation. The main existing problem is that code is not thread safe due to the usage of non-reentrant system calls, which should be substituted with the corresponding reentrant ones. Moreover, the most extended issue on the code, is the usage of static variables used as global variables.

 ISBN: 978-84-614-6167-7

These variables are being used within the most important functions of T-Coffee. The program state is being reused on subsequent system calls. In order to be able to run this code in parallel, we have created a memory container that stores all this global data, in a per thread context. Each separate thread has a separate memory container which its own memory. On all affected functions, the static variables are substituted by normal ones, and its state is recovered at the beginning of its execution, and its stored back at the end of the function. Additionally, we have found a lot of synchronization problems due to incorrectly referred memory locations accessed from the wrong threads. T-Coffee has its own memory management system, mainly for debugging purposes. The memory management code has been modified in a way that each memory allocation keeps track of its own thread caller identification. Then the code performs checks over the memory usage in a way that it avoids and reports illegal access to memory regions of other threads. This verification code, causes a considerable penalty on the efficiency on all the memory allocation code.

Another problem, is how T-Coffee deals with memory. It relies on fork assuming that all memory is freed on process termination. When this code is migrated to pthreads, this assumption causes an uncontrolled memory leakage. To avoid this problem, we have to keep track of all memory allocated within a thread, deallocating it after thread termination.

In order to study our implementation we have performed several tests using sequences from the BAliBASE benchmarking data set. These tests were run on a Intel Core 2 Quad machine, with has four cores and eight Gigabytes of physical memory. Table 2 displays the average time in seconds consumed by these sequences, each row is a different BAliBASE input data directory, each one, has different biological meanings. The first column shows the average computing time of the execution of T-Coffee running in serial, the second one shows the time consumed by the pthreads implementation, running four worker threads. As we can see on the table, the time required by the threaded implementation is obviously lower than the one required by the serial one, on average the threaded implementation is 2.6 times faster. These performance results encourages us to go towards a global multi-threading implementation of T-Coffee. On these experiments we are not looking at the number of sequences, since we are more focused on the timings and BAliBASE benchmark has very few sequences on its input data files to perform this kind of study. In addition as we are using BAliBASE sequences for comparing two implementations of the same algorithm, the BAliBASE score remains identical and its not considered.

5 Conclusions

MSA (Multiple Sequence Alignment) is a high computing demanding process, in the biology field. In this work we focussed on the analysis of performance of the application T-Coffee. This carries out MSA and its use is very extended among the scientific community. Two different parallel implementations of T-Coffee, using a message passing library, have been

 ISBN: 978-84-614-6167-7

Table 2: Consumption of computing time, in seconds, by T-Coffe serial, and T-Coffee with pthreads

Test	Serial	Threads
RV11	2.51	1.68
RV12	5.36	3.58
RV20	82.95	31.49
RV30	198.43	76.72
RV40	73.32	30.42
RV50	63.46	23.36
Total	71.01	27.88

analysed, to determine their scalability. PTC (Parallel-T-Coffee), that parallelizes the internal steps and, CTC (Clust-T-Coffee) that adds an initial step to divide the global input sequences on different sets and, then, T-Coffee is executed in parallel for each set. With the different tests we could verify that the parallel implementations were able to process more sequences than the serial one. However, there are still some issues to confront, with the volume of communications, on the message passing implementations. Finally we are proposing to introduce a multi-threading solution in one of the current T-Coffee implementations in order to increase its efficiency, reduce the consumed memory and reduce the overhead of using a message passing library for computation on a multi-core machine.

Acknowledgements

This work was supported by the MEyC-Spain under contract TIN 2008-05913 and Consolider CSD2007-0050. The CUR of DIUE of GENCAT and the European Social Fund.

References

[geer05] Geer D, *Chip makers turn to multicore processors*, Computer, vol.**38**, no.5, pp. 11- 13, May 2005; doi: 10.1109/MC.2005.160.

[hill08] Hill M.D, Marty M.R, *Amdahl's Law in the Multicore Era* Computer, vol.**41**, no.7, pp.33-38, July 2008; doi: 10.1109/MC.2008.209.

[notredame00] Notredame C, Higgins DG, Heringa J, *T-Coffee: A novel method for fast and accurate multiple sequence alignment.*, J Mol Biol. 2000 Sep 8;302(1):205-17. PMID: 10964570 [PubMed - indexed for MEDLINE].

[zola07] Zola J, Yang X, Rospondek S, Aluru S;, *Parallel T-Coffee: A Parallel Multiple Sequence Aligner*, In Proc. of ISCA PDCS-2007, pp. 248-253, 2007.

[naranjo09] Naranjo Y., *Alineamiento mltiple de secuencias con T-Coffee: una aproximacin paralela* master thesis, UAB, 2009.

[kwok99] Kwok Y. K. and Ahmad. I. *Benchmarking and comparison of the task graph scheduling algorithms.*, Journal of Parallel and Distributed Computing, **59**:381-422, 1999.

[edgar04] Edgar, R.C., *MUSCLE: multiple sequence alignment with high accuracy and high throughput.*, Nucleic Acids Research. **32(5)**: 1792-1797. (2004).

[katoh08] Katoh, K. and Toh, H., *Recent developments in the MAFFT multiple sequence alignment program.*, Briefings in Bioinformatics., **9(4)**:286-298. (2008)

[larkin07] Larkin, M.A., Blackshields, G., Brown, N.P., Chenna, R., McGettigan, P.A., McWillian, H., Valentin, F., Wallace, I.M., Wilm, A., Lopez, R., Thompson, J.D., Gibson, T.J. and Higgins, D.G., (2007). *ClustalW and ClustalX version 2.*, Bioinformatics. **23(21)**: 2947-2948. (2007)

[thompson99] Thompson J. D. Plewniak F. Poch, O., *BAliBASE: a benchmark alignment database for the evaluation of multiple alignment programs*, BIOINFORMATICS -OXFORD- Bibliographic details 1999, VOL **15**; NUMBER 1, pages 87-88.

 ISBN: 978-84-614-6167-7

Proceedings of the 11th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2011
26–30 June 2011.

Comparing different theorem provers for modal logic K

Angel Mora[1], Emilio Muñoz-Velasco[1], Joanna Golińska-Pilarek[2] and Sergio Martín[1]

[1] *Dept. Applied Mathematics., University of Málaga. Spain*

[2] *National Institute of Telecommunications, Warsaw, Poland.*
Institute of Philosophy. University of Warsaw . Poland

emails: amora@ctima.uma.es, emilio@ctima.uma.es, j.golinska@uw.edu.pl, animasergio@gmail.com

Abstract

Several provers are proposed in the literature for modal logic. We give a first step on an exhaustive comparative of the provers for modal logic K designed by the authors with other provers well known in the literature. A comparative is complicate because each prover has been implemented for different platforms using different programming languages and with several input formats for formulas. We designed a common framework to facilitate a comparison among provers with some tools developed for this purpose.

Key words: therorem proving, modal logic, implementation

1 Introduction

Modal logics are very useful in many areas of computer science. These areas include knowledge and belief representation, database theory and distributed systems, program verification, cryptography and agent based systems, computational linguistics, and nonmonotonic formalisms [7]. Moreover, the recent development Description Logics [2] has increased the interest for modal logics. The basis of all these logics is modal logic K. For this reason, many provers have been constructed for this logic (see for example, [1,4,8,9]). Most of these provers are based either on tableaux or in sequent calculus [3] and use external techniques such as backtracking, backjumping, etc. For this reason, we developed in [11] a theorem prover for modal logic K (based on dual tableaux [12]) which is a deterministic decision

 ISBN: 978-84-614-6167-7

procedure but does not use any external technique. The main intention was, at the beginning, purely theoretical and we assumed a practical cost in terms of complexity. As a consequence, we did some improvements in [5] by using modal clauses and reducing the number of rules applied.
In this paper, we give a first step on an exhaustive comparative of our provers for modal logic K with other provers in the literature. The first thing we notice is that a comparative is complicate because each prover has been implemented for different platforms using different programming languages and with several input formats for formulas. We designed a common framework to facilitate a comparison among provers with some tools developed for this purpose.

The provers are designed in order to check the *validity* of a formula in modal logic K. We introduce now the basic notions needed. The language of logic K consists of the symbols from the following pairwise disjoint sets:

- $\mathbb{V} = \{p_1, p_2, p_3, \ldots\}$ - an ordered countable infinite set of propositional variables indexed with natural numbers
- $\{\neg, \vee\}$ - the set of classical propositional operations of negation ($\neg$) and disjunction ($\vee$)
- $\{\Box\}$ - the set consisting of modal propositional operation called the *necessity* operation.

The set of K-formulas is the smallest set including the set of propositional variables and closed with respect to all the propositional operations.

A K-*model* is a structure $\mathcal{M} = (U, R, m)$ such that:

- U is a non-empty set (of states)
- R is a binary relation on U
- m is a meaning function such that $m(p) \subseteq U$, for every propositional variable $p \in \mathbb{V}$.

The relation R is referred to as the *accessibility relation*. The *satisfaction relation* is defined as usual in modal logics. Recall that for $\Box$-formulas it is defined as:

$\mathcal{M}, s \models \Box\varphi$ if and only if for all $s' \in U$, $(s, s') \in R$ implies $\mathcal{M}, s' \models \varphi$.

A K-formula φ is said to be *true* in a K-model $\mathcal{M} = (U, R, m)$, $\mathcal{M} \models \varphi$, whenever for every $w \in U$, $\mathcal{M}, w \models \varphi$, and it is K-*valid* whenever it is true in all K-models. A formula φ is said to be K-*satisfiable* whenever there exist a K-model $\mathcal{M} = (U, R, m)$ and $w \in U$ such that $\mathcal{M}, w \models \varphi$.

 ISBN: 978-84-614-6167-7

The paper is organized as follows. In Section 2, we present some modal provers for modal logic K. In Section 3 we present the tools we have developed for doing the comparative, which are explained in Section 4. Finally, some conclusions and prospects of future works are presented in Section 5.

2 Provers for modal logics

There are several provers for modal logics in the literature. An exhaustive list of these provers can be found in `http://www.cs.man.ac.uk/~schmidt/tools/`. We focus our attention in TWB [1], developed in the Computer Science Laboratory at the Australian National University; LWB [2], developed in the University of Bern; and in Lotrec [3], developed in the IRIT. The goal in this paper is to obtain a comparative among these provers and the two provers [5, 11] we have developed in Swi-Prolog.

TWB has a demo version executable via web. The format of formulas is inorder and it uses the operators $\rightarrow$, &, $\vee$, *box*, and *dia* . A limitation is the length of formulas that it is possible to prove. A big formula is cut when you try to write it or copy-paste it from another editor. The system informs about the total number of applied rules but a trace with information about the rules applied is not given nor real information about the execution time.

LWB is a project of the University of Bern for several logics. An executable version of the prover via web is provided. The format of the formulas is similar to TWB but using $\rightarrow$, &, v, [], <> as operators. The prover returns information about the rules applied but the execution time is not showed.

Lotrec is a project of IRIT and the LoTREC Web Start is available, together with an executable Java file to be downloaded. As a difference with other provers, the formulas are written in preorder. Probably Lotrec is the more powerful prover regarding the possibilities of extension to other logics, operators, rules, etc. No execution time is showed and the user needs count the number of rules applied by hand.

Our first prover is named $RePML_K$ and it is a proof system in the style of dual tableaux for the relational logic associated to modal logic K. It was the first implementation of a specific relational prover for a standard modal logic [11]. This prover is itself a deterministic decision procedure verifying validity of K-formulas, that is, it does not use any external technique such as backtracking, backjumping, etc. $RePML_K$ has been developed in Prolog and the formulas are represented as Prolog predicates. Moreover, in [10] we proposed a front-end for our theorem prover providing an user-friendly environment which could be very useful both for research and educational applications. Our prover returns the execution time

[1] `http://twb.rsise.anu.edu.au/`
[2] `http://www.lwb.unibe.ch/`
[3] `http://www.irit.fr/Lotrec/`

 ISBN: 978-84-614-6167-7

and the full trace of the rules applied in order to use the prover as a learning environment of modal logic.

We have also implemented in Prolog the prover $\mathcal{RLK}$ which is also a dual tableau system, which is itself a deterministic decision procedure verifying validity of K-formulas, without any external technique [6]. $\mathcal{RLK}$ improves $RePML_K$ by using modal clauses to represent the formulas and by the reduction of the number of rules to be applied.

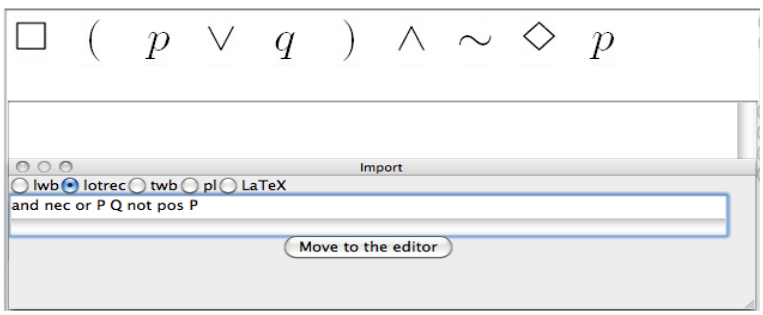


Figure 1: Translation tool

3 Tools for a comparative

First of all, we need a tool for translating formulas from each format to the other ones. We have developed this tool using the Python programming language. In the tool, we select the option `import` from the editor and the Figure 1 appears. We introduce the formula, then we select the adequate format of the formula to be imported and push the buttom `Move to the editor`. Then the formula is displayed at the top of the window. In this process, the formula is stored in an internal tree which facilitates the posterior process of conversion to others formats. In Figure 1, the modal formula `and nec or P Q not pos P` in the format used for the Lotrec prover is displayed as $\Box\,(p \vee q) \wedge \sim \Diamond p$ when you push the buttom `Move to the editor`. At this stage, we think that a random generator of modal formulas could allow us to make a first interesting comparative. Figure 2 shows the tool developed. We introduce the name of the file, the path, the number of formulas will be generated and three parameters in order to prepare bigger comparatives. We can select the maximum number of the atoms that appears in the formula and also the modal length of the formula using the parameters maximum and minimum level of the tree of the formula. Finally, in order to automatize the process of conversion to the rest of formats, another tool in Python has been developed. Figure 3 shows this important tool. A file from the random generator of formulas, or a formula can be selected and the output format can be choosen. Then, you

 ISBN: 978-84-614-6167-7

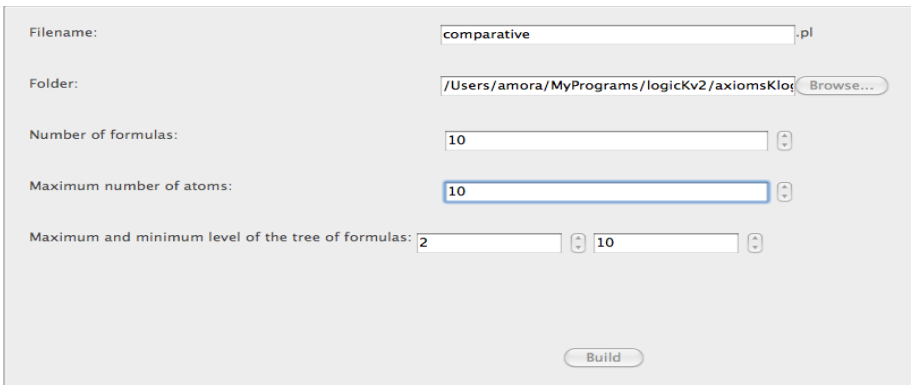


Figure 2: Random generator tool

push `Start translation` and the result is stored in different files with the same name and the appropriated extension: .lwb, .twb, .lot, .pl.

4 The comparative

First, we translate the following formulas from TWB [4] by using our Python tool for translation.

NT1: $\Box p1 \to p1$

```
LWB:     (box p1)->p1
TWB:     [] p1 ->  p1
Lotrec:    imp nec P1 P1
Prolog: formulaKLogic(implication(square(p1),p1)).
```

NT2: $\Box p1 \to \Box\Box p1$

```
LWB:     (box p1)->(box(box p1))
TWB:     [] p1 -> [] [] p1
Lotrec:   imp nec P1 nec nec P1
Prolog:  formulaKLogic(implication(square(p1),square(square(p1)))).
```

NT3: $\neg\Box p1 \to \Box\neg\Box p1$

```
LWB:    (~(box p1))->(box(~(box p1)))
TWB:    (~([] p1))->([](~([] p1)))
Lotrec:  imp not nec P1 nec not nec P1
Prolog:  formulaKLogic(implication(not(square(p1)),square(not(square(p1))))).
```

T1: $\Diamond p \to \Diamond p$

[4] http://twb.rsise.anu.edu.au/modal_logic_k_0

 ISBN: 978-84-614-6167-7

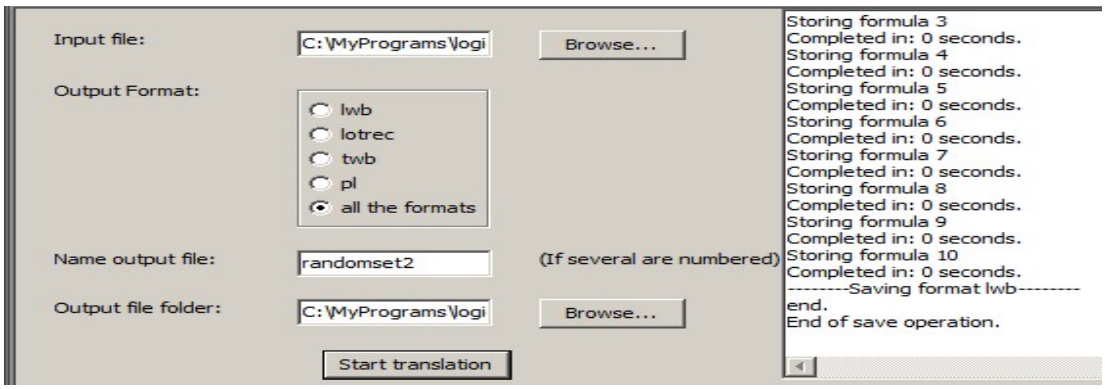


Figure 3: Translator tool

```
LWB:     (dia p)->(dia p)
TWB:     <> p -> <> p
Lotrec:  imp pos P pos P
Prolog:  formulaKLogic(implication(diamond(p),diamond(p))).
```

T2: $\Diamond p \rightarrow \Box(\Box\neg q \vee \Diamond q)$

```
LWB:     (dia p)->(box((box(~q))v(dia q)))
TWB:     <> p -> [] ( [] ~ q v <> q)
Lotrec:  imp pos P nec or nec not Q pos Q
Prolog:  formulaKLogic(implication(diamond(p),square(or(square(not(q)),diamond(q))))).
```

This way, we have translated all the following formulas provided in TWB.

T3: $\neg(\Diamond p \wedge \Box\neg p)$

T4: $\neg(\Diamond p \wedge \Diamond(\Diamond q \wedge \neg\Diamond q))$

T5: $\Box(a \rightarrow b) \rightarrow (\Box a \rightarrow \Box b)$

T6: $(\Box(p0 \wedge p1) \leftrightarrow (\Box p0 \wedge \Box p1))$

T7: $(\neg\Box\neg p0 \leftrightarrow \Diamond p0)$

T8: $(\Box(p0 \rightarrow p1)) \rightarrow (\Box p0 \rightarrow \Box p1)$

T9: $(\Box p0 \rightarrow \Diamond p0) \wedge (\Box p0 \rightarrow \Box\Box p0) \wedge (\neg p0 \rightarrow \Box\Diamond\neg p0) \rightarrow (\Box p0 \rightarrow p0)$

Similarly, form the webpage of Lotrec [5] we obtain the following formulas:

Lot1: $(\Box p) \wedge ((\Diamond q) \wedge (\Diamond(r \vee \neg q)))$

Lot2: $\sim\sim \Diamond p \wedge \Diamond q \wedge \Box(s \vee \sim p)$

Lot3: $p \wedge \sim \Box p$

Lot4: $\Diamond p \wedge \Box(p \rightarrow \Diamond q) \wedge \Diamond \sim p$

[5] http://www.irit.fr/Lotrec/

 ISBN: 978-84-614-6167-7

Lot5: $\Diamond p \wedge \Box (p \rightarrow \Diamond q) \wedge \Box \sim q$

Lot6: $\Box\Diamond \sim p \wedge \Diamond\Box \sim q \wedge \Box\Diamond\Box (p \vee q)$

The following table shows the result of the comparative of the number of rules applied by our provers $RePML_K$, $\mathcal{RLK}$ and the rest of provers TWB, Lotrec, and LWB.

Modal formula	TWB	Lotrec	LWB	$RePML_K$	$\mathcal{RLK}$	Output
NT1	1	5	1	2	1	Not valid
NT2	3	4	3	6	1	Not valid
NT3	3	3	5	6	1	Not valid
T1	3	4	1	1	1	Valid
T2	5	3	6	6	2	Valid
T3	3	6	4	3	1	Valid
T4	5	4	6	3	2	Valid
T5	7	9	4	7	2	Valid
T6	18	7	11	17	2	Valid
T7	8	11	9	8	1	Valid
T8	7	8	4	7	2	Valid
T9	31	37	11	14	2	Valid
Lot1	2	2	6	2	3	Not valid
Lot2	3	1	9	2	1	Not valid
Lot3	1	1	3	1	1	Not valid
Lot4	1	10	4	1	2	Not valid
Lot5	1	1	8	1	2	Not valid
Lot6	2	2	6	3	2	Not valid

Observe that we have improved the number of applied rules in practically all the cases: in 11 of 18 formulas, we use a lower number of rules, in 4 cases we use the same number of rules of any prover and only in 3 cases we use a rule more than any prover. We have obtained also a bank of formulas by using the random generator with different values for the number of atoms and for the modal length of a formula. The formulas generated, its translation to the rest of formats, together with the results of the execution of all the provers can be found in `http://www.matap.uma.es/~amora/randomcomparative.zip`. Figure 4 shows a comparative of the results obtained and a picture with the average of number of used rules of all provers. Notice that TWB can not be used in formulas F14 and F18 because of its length. Our provers have a good behavior and specially we have excellent results with $\mathcal{RLK}$. In 11 cases of 19, $\mathcal{RLK}$ applies less number of rules than the rest, in 5 cases it applies the same number of rules than any other prover, and only in 3 cases it applies more rules than any other prover.

 ISBN: 978-84-614-6167-7

5 Conclusions and future works

We presented a first step on an exhaustive comparative of provers for modal logic K which could be extended to other logics. We developed some tools in order to do this comparative, such as translation tool using the Python programming language to import from a format and convert to the rest of the formats; and a random generator of modal formulas. The results of this first comparison show that prover $\mathcal{RLK}$ improves the number or rules applied in almost all the cases.

As a future work, we consider first the extension of this comparative to greater formulas to check how the different provers scale as the problem size increases. We are preparing a framework for future bigger comparatives. In `http://www.lwb.unibe.ch/run.html` a benchmark for modal logics is available. However, the size of the formulas is blocking the use of some provers: TWB in the web version does not allow us to introduce some formulas and Lotrec is deadlocked also with some big formulas. Of course, an automatic tool to translate this benchmark to the rest of formats should be implemented.

Acknowledgements

This work is partially supported by the Spanish research projects TIN09-14562-C05-01 and P09-FQM-5233. The first author of the paper is supported by the Polish Ministry of Science and Higher Education grant Iuventus Plus IP2010 010170.

References

[1] P. Abate and R. Goré, *The Tableau Workbench*, Electronic Notes in Theoretical Computer Science Vol. 231, 55–67 (2009).

[2] F. Baader, D. Calvanese, D. McGuinness, D. Nardi, P. Patel-Schneider (eds), The Description Logic Handbook, Cambridge University Press, Cambridge, London (2002).

[3] M. Fitting, *Modal Proof Theory*, in: P. Blackburn, J. van Benthem, and F. Wolter (eds), Handbook of Modal Logic, Studies in Logic and Practical Reasoning, Volume 3, Elsevier, Amsterdam, 85–138 (2007).

[4] O. Gasquet, A. Herzig, D. Longin, and M. Sahade, *Lotrec: Logical tableaux research engineering companion*, Lecture Notes in Artificial Intelligence, Vol. 3702, 318–322 (2005).

[5] J. Golińska-Pilarek, E. Muñoz-Velasco, and A. Mora, *A new deduction system for deciding validity in modal logic* K, Logic Jnl of IGPL, Vol. 19 No. 2, 425–434 (2011).

 ISBN: 978-84-614-6167-7

	TWB	LOTREC	LWB	REPMLK	RLK	RESULT
F1	0	4	0	0	0	NOT VALID
F2	6	0	7	7	3	NOT VALID
F3	5	1	5	6	1	NOT VALID
F4	3	8	4	4	3	NOT VALID
F5	0	3	0	0	0	NOT VALID
F6	4	10	4	5	1	NOT VALID
F7	5	0	5	7	2	NOT VALID
F8	4	4	3	3	1	VALID
F9	2	3	2	2	1	NOT VALID
F10	3	3	3	4	1	NOT VALID
F11	2	6	2	3	0	NOT VALID
F12	1	0	1	2	2	NOT VALID
F13	4	9	4	6	2	NOT VALID
F14	NOT	42	11	6	1	NOT VALID
F15	1	19	1	1	1	NOT VALID
F16	3	3	4	1	3	NOT VALID
F17	6	9	9	7	3	NOT VALID
F18	NOT	48	2	2	1	NOT VALID
F19	4	39	6	6	1	NOT VALID
AVERAGE	3,1	11,105	3,8	3,7895	1,4	

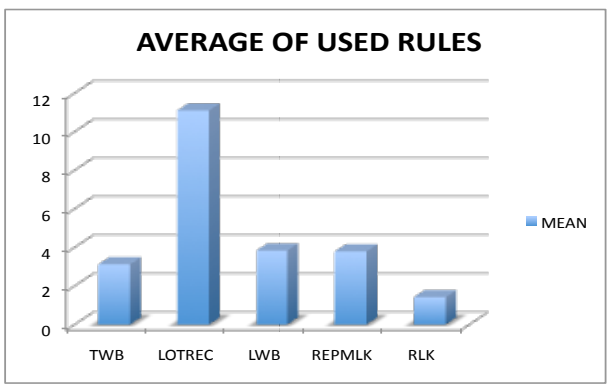


Figure 4: Table of the comparative and picture with the mean of the results

[6] J. Golińska-Pilarek, E. Muñoz-Velasco, and A. Mora, *Relational dual tableau decision procedure for modal logic* K, Logic Jnl of IGPL, doi: 10.1093/jigpal/jzr019 (2011).

[7] I. Horrocks, U. Hustadt, U. Sattler, and R. Schmidt, *Computational Modal Logic*, in: P. Blackburn, J. van Benthem, and F. Wolter (eds), Handbook of Modal Logic, Studies in Logic and Practical Reasoning, Volume 3, Elsevier, Amsterdam, 181–245 (2007).

[8] A. Heuerding, LWBtheory: Information about some propositional logics via the WWW, Logic J. IGPL 4, 169–174,(1996).

[9] I. Horrocks, *The FaCT System*, in Proc. of the 2nd Int. Conf. on Analytic Tableaux and Related Methods (TABLEAUX 98), Lecture Notes in Artificial Intelligence, Vol. 1397, H. de Swart, ed., Springer, Berlin, Heidelberg, 307–312 (1998).

 ISBN: 978-84-614-6167-7

[10] A. Mora, J. Golińska-Pilarek, S. Martín, and E. Muñoz-Velasco, *A front-end for theorem proving with modal logics*, International Conference Computational and Mathematical Methods in Science and Engineering, Volume 2, 658–663 (2010).

[11] A. Mora, E. Muñoz-Velasco, and J. Golińska-Pilarek, *Implementing a Relational Theorem Prover for Modal Logic K*, Int. Jnl. of Computer Mathematics, doi: 10.1080/00207160.2010.49321. (2011).

[12] E. Orłowska and J. Golińska-Pilarek, *Dual Tableaux: Foundations, Methodology, Case Studies*, Trends in Logic 33, Springer Science (2011).

 ISBN: 978-84-614-6167-7

Proceedings of the 11th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2011
26-30 June 2011.

Dedekind-MacNeille Completion and Multi-adjoint Lattices

P.J. Morcillo, G. Moreno, J. Penabad and C. Vázquez[1]

[1] *University of Castilla-La Mancha, Faculty of Computer Science Engineering*
02071, Albacete (Spain),

emails: {pmorcillo,cvazquez}@dsi.uclm.es
{Gines.Moreno,Jaime.Penabad}@uclm.es

Abstract

Among other applications, *multi-adjoint lattices* have been successfully used for modeling flexible notions of truth-degrees in the fuzzy extension of logic programming called MALP (*Multi-Adjoint Logic Programming*). In this paper we focus in the completion of such mathematical construct by adapting the classical notion of *Dedekind-MacNeille* in order to relax this usual hypothesis on such kind of ordered sets. On the practical side, we show too the role played by multi-adjoint lattices into the *"Fuzzy LOgic Programming Environment for Research" FLOPER* that we have developed in our research group.

Key words: Lattice, Completion, Multi-adjoint Logic Programming.
MSC 2000: Lattice, Dedekind-MacNeille Completion.

1 Introduction

In essence, the notion of multi-adjoint lattice considers a carrier set L (whose elements verify a concrete ordering $\leq$) equipped with a set of connectives like implications, conjunctions, disjunctions and other *hybrid aggregators*, with the particularity that for each implication symbol there exists its *adjoint conjunction* used for modeling the *modus ponens* inference rule in a fuzzy setting. For instance, some adjoint pairs, i.e. conjunctors and implications, in the lattice $([0,1],\leq)$ are presented below, where labels `L`, `G` and `P` mean respectively *Łukasiewicz logic*, *Gödel intuitionistic logic* and *product logic* (with different capabilities for modeling *pessimist*, *optimist* and *realistic scenarios*, respectively):

$\&_{\mathtt{P}}(x,y)$	$\triangleq$	$x * y$		$\leftarrow_{\mathtt{P}}(x,y)$	$\triangleq$	$\min(1, x/y)$	*Product*
$\&_{\mathtt{G}}(x,y)$	$\triangleq$	$\min(x,y)$		$\leftarrow_{\mathtt{G}}(x,y)$	$\triangleq$	$\begin{cases} 1 & \text{if } y \leq x \\ x & \text{otherwise} \end{cases}$	*Gödel*
$\&_{\mathtt{L}}(x,y)$	$\triangleq$	$\max(0, x+y-1)$		$\leftarrow_{\mathtt{L}}(x,y)$	$\triangleq$	$\min\{x-y+1, 1\}$	*Łukasiewicz*

 ISBN: 978-84-614-6167-7

Moreover, in the MALP framework [25, 23, 24], each program has its own associated multi-adjoint lattice and each program rule (very similar to a Prolog clause[1]) is "weighted" with an element of L, whereas the components in its body are *linked* with connectives of the lattice. For instance, in the following propositional MALP program (where obviously $@_{\texttt{aver}}$ refers to the classical average aggregator):

$$\begin{array}{llll} p & \leftarrow_{\texttt{P}} & @_{\texttt{aver}}(q,r) & with \quad 0.9 \\ q & \leftarrow & & with \quad 0.8 \\ r & \leftarrow & & with \quad 0.6 \end{array}$$

the last two rules directly assign truth values 0.8 and 0.6 to propositional symbols q and r, respectively, and the execution of p using the first rule, simply consists in evaluating the expression "$\&_{\texttt{P}}(0.9, @_{\texttt{aver}}(0.8, 0.6))$", which returns the final truth degree 0.63.

Anyway, although the class of multi-adjoint lattices is wide enough to model real-world application written with the MALP language [2], in [28, 26] we have proposed some debugging/tracing techniques based on lattices (whose elements are strings of characters) which do not fully accomplish with the hypothesis of complete lattice required by multi-adjoint lattices.

Motivated by this fact, in Sections 2 and 3 of this paper, we give a first step in solving such problem, inspired by the Dedekind-MacNeille completion of an ordered set P (also known as the normal completion of P and the completion by cuts) which was originally proposed by M. MacNeille in 1937 (see [18]) as an extension of the famous definitions of real numbers conceived as cuts from rational ones due to Dedekind[2] in 1872 [8].

The Dedekind-MacNeille completion is directly related to the concept of the canonical extension that was firstly introduced, for Boolean algebras, in [13] and that arises from Stone's duality theorem. Although out of the scope of this paper, in the future we plan to analyze some canonical extensions for multi-adjoint lattices, formally introduced in [9] (see also [10, 29], which study the completion of an n-ordered set), that have associated monotone operators and analyze the results especially for the habitual domains in multi-adjoint logic programming of bilattices and trilattices [20, 21, 19, 4, 5, 3, 6].

On the other hand, the last part of this paper is concerned with implementation and practical developments achieved in our group. More exactly, in Section 4 we present the FLOPER tool [27, 28, 26], which currently is useful for compiling (to standard Prolog code), executing and debugging MALP programs in a safe way and it is ready for being extended in the near future with powerful transformation and optimization techniques designed in our research group in the recent past [14, 11]. In this paper, we will focus in the management of multi-adjoint lattices performed by FLOPER, where such constructs can be easily expressed by means of a set Prolog clauses. Moreover, for a given program and goal, we will see too that different solution could be achieved depending on the currently loaded lattice (which can be changed as much as wanted even in a single work session).

[1] We assume familiarity with pure Logic Programming and its most popular language Prolog [16].

[2] This German mathematician was pupil of Gauss in Gotinga and nowadays is considered one of the founders of modern algebra.

 ISBN: 978-84-614-6167-7

2 Dedekind-MacNeille Completion

We start this section by giving some basic definitions before addressing the concept of Dedekind-MacNeille completion.

Definition 2.1. *Let $(P, \leq)$ be an ordered set and $Q \subset P$.*

i) Q is a down-set *(also called decreasing set and order ideal) if whenever $x \in Q, y \in P$ with $y \leq x$, we have $y \in Q$.*

ii) Dually, Q is an up-set *(also called increasing set and order filter) if whenever $x \in Q, y \in P$ with $y \geq x$, we have $y \in Q$.*

In what follows we will use the set, read as "down" $Q, \downarrow Q = \{y \in P : \exists x \in Q \ / \ y \leq x\}$, in particular, $\downarrow x = \{y \in P : y \leq x\}$ that is called *principal down-set* and also *principal ideal generated by* x (obviously, $\downarrow \{x\} = \downarrow x$). The set of all down-sets of P is denoted by $\mathcal{O}(\mathcal{P})$ which is an ordered set under the usual inclusion ordering. Similary, "up" Q, $\uparrow Q = \{y \in P : \exists x \in Q \ / \ x \leq y\}$.

If P is an ordered set and $X = \{\downarrow x : x \in P\}$ (ordered by inclusion), $Y = \{\uparrow x : x \in P\}$ (ordered by reverse inclusion), then the maps $\triangleright : X \to Y$ given by $\downarrow x \mapsto \uparrow x$ and $\triangleleft : Y \to X$ given by $\uparrow x \mapsto \downarrow x$ forms a *Galois connection* between X and Y. This notion appears in [22, 19], where a fuzzy generalization of the formal concept analysis was presented. In particular, multi-adjoint concept lattices were introduced into the MALP framework for application to formal concept analysis.

Definition 2.2. *Let $(P, \leq), (Q, \leq)$ be ordered sets. A map $\varphi : P \to Q$ is said to be*

i) order-preserving (or monotone) if $x \leq y$ in P implies $\varphi(x) \leq \varphi(y)$ in Q.

ii) order-embedding if $x \leq y$ in P if and only if $\varphi(x) \leq \varphi(y)$ in Q.

ii) order-isomorphism if it is an order-embedding which maps P in Q.

Definition 2.3. *Given a partially ordered set $(P, \leq)$, we define for every subset A of P, two subsets of P as follows: $A^u = \{x \in P : a \leq x, \forall a \in A\}$ and $A^l = \{x \in P : x \leq a, \forall a \in A\}$.*

The sets A^u, A^l are called A *upper* and A *lower*, respectively. A^u is the set of all upper bounds[3] of A and A^l is the set of all lower bounds[4]. Moreover, A^u is an up-set and A^l is a down-set.

Definition 2.4. *Let $(P, \leq)$ be ordered set and $Q \subset P$. Q is join-dense (similary, meet-dense) in P if for all $a \in P$ exists $A \subset Q$ such that $a = \bigvee A$[5] (similary, $a = \bigwedge A$).*

[3]By definition, an element $x \in P$ is an upper bound of A if $a \leq x$ for all $a \in A$.

[4]By definition, an element $x \in P$ is an lower bound of A if $x \leq a$ for all $a \in A$.

[5]We also write $\bigvee A$ for the *joint* or *supremum* of A instead of $sup(A)$ and $\bigwedge A$ for the *met* or *infimum* of A instead of $inf(A)$ where these exist.

 ISBN: 978-84-614-6167-7

The following proposition is elementary, but in later Theorems 2.10, 2.11 y 3.2 we find interesting examples of isomorphisms guarantees.

Proposición 2.5. *All ordered set $(P, \leq)$ is isomorphic to a subset of set $(2^P, \subset)$.*

Proof. It suffices to note that the map $f : P \to 2^P$, given by $f(x) = \{y \in P : y \leq x\}$ is injective and is order-preserving. On the other hand, the image $f(x)$ is a down-set of ordered set $(2^P, \subset)$. □

We shall be interested in the ordered sets in which the infimum and the supremum exist for all subsets.

Definition 2.6. *Let $(P, \leq)$ be a non-empty ordered set. If $inf(S)$ and $sup(S)$ exist for all $S \subset P$, then P is called a* complete lattice.

It is straightforward to prove that a non-empty P is complete lattice if and only if $inf(S)$ exists in P for every subset S of P.

On the other hand, there are many options for the embedding of an ordered set into a complete lattice. We examine here one such embedding that generalizes Dedekind's construction of $\mathbb{R}$ by cuts of $\mathbb{Q}$, in order to apply it to the case of multi-adjoint lattice.

Definition 2.7. *Let P be an ordered set. If C is a complete lattice and $\varphi : P \to C$ is an order-embedding, then we say that C is a* completion of P *(via φ).*

Since the map $\varphi : x \longmapsto \ \downarrow x$ is trivially an order-embeding of P into the complete lattice $\mathcal{O}(\mathcal{P})$ of all down-sets of P (with the inclusion order), this one is a natural completion of P. However, it is unnecessarily large: it is sufficient to take into account that if P is a complete lattice then P is a completion of itself (via the identity map), while $\mathcal{O}(\mathcal{P})$ is much larger. Another completion of an ordered set is the ideal completion. In what follows, we consider the smallest complete lattice containing P, namely the Dedekind-MacNeille completion.

Definition 2.8. *[7] The* Dedekind-MacNeille completion *of an ordered set P is the set $DM(P) = \{A \subset P : A^{ul} = A\}$.*

Moreover, it is also known as the completion by cuts and the normal completion of P (see [10]). By means of the following theorem, we can give equivalent definitions of the above concept, in terms of principal ideals of the notion of cut.

Theorem 2.9. *Let $DM(P)$ be the the Dedekind-MacNeille completion of P. Then,*

i) $DM(P) = \{A \subset P : \triangle A \subset A\}$, *where* $\triangle A = \bigcap\{\downarrow x : x \in A^{\uparrow}\}$.

ii) $DM(P) = \{A \subset P : (A, B)$ *is a cut of* P, *for some* $B \subset P\}$.

$(DM(P), \subset)$ is a complete lattice and, moreover, the map $\varphi : x \longmapsto \downarrow x$ is an order-embedding of P into $DM(P)$. Then, it is easy to prove the following theorem.

Theorem 2.10. *Let P be an ordered set and let $\varphi : P \to DM(P)$ be such that $\varphi(x) = \downarrow x$ for all $x \in P$. Then, $DM(P)$ is a completion of P via the map φ.*

 ISBN: 978-84-614-6167-7

$DM(P)$ is known as the Dedekind-MacNeille completion of an ordered set P. In $DM(P)$, since $inf(A)$ and $sup(A)$ exist for any subset, $A \subset DM(P)$, is a complete lattice. Moreover, this process can be readily applied to any lattice, if we define a completion of a lattice. The fundamental theorem that follows can be used to characterize the Dedekind-MacNeille completion.

Theorem 2.11. *[7] Let P be an ordered set and let $\varphi : P \to DM(P)$ be the order-embedding of P into its Dedekind-MacNeille completion given by $\varphi(x) = {\downarrow} x$ for all $x \in P$. Then*

i) $\varphi(P)$ is both join-dense and meet-dense in $DM(P)$.

ii) If C is a complete lattice and P is a subset of C which is both join-dense and meet-dense in C, then $C \approx DM(P)$ via an order-isomorphism which agrees with φ on P.

3 Completion of a Quasimulti-adjoint Lattice

In this section, we analyze the specific properties of the Dedekind-MacNeille completion in the case of the lattices used by MALP programs, starting with their formal definition.

Definition 3.1. *Let $(L, \leq)$ be a lattice. A* multi-adjoint lattice *is $(L, \leq, \leftarrow_1, \&_1, \ldots, \leftarrow_n, \&_n)$ such that:*

i) $(L, \leq)$ is a complete lattice, namely, $\forall S \subset L, \exists\ inf(S), sup(S)$[6].

ii) $\&_i$ is increasing in both arguments, for all i, $i = 1, \ldots, n$.

iii) $\leftarrow_i$ is increasing in the first argument and decreasing in the second, for all i.

iv) If $\langle \&_i, \leftarrow_i \rangle$ is an adjoint pair *in $(L, \leq)$ then, for any $x, y, z \in L$, we have that: $x \leq (y \leftarrow_i z)$ if and only if $(x \&_i z) \leq y$.*

This last condition, called *adjoint property*, is the most important feature of the framework. Moreover, if $(L, \leq, \leftarrow_1, \&_1, \ldots, \leftarrow_n, \&_n)$ is bounded and satisfy only $ii), iii), iv)$, we call *quasimulti-adjoint* lattice.

The following theorem guarantees that the Dedekind-MacNeille completion of a quasimulti-joint lattice has a quasimulti-adjoint sublattice isomorphic to the initial one. Also, in this theorem it can be viewed in detail the particular properties of the embedding φ.

Theorem 3.2. *If $(L, \leq, \leftarrow_1, \&_1, \ldots, \leftarrow_n, \&_n)$ is a* quasimulti-adjoint lattice*, then $Im(\varphi)$ is a quasimulti-adjoint sublattice of the complete lattice $DM(P)$ (via φ) that is isomorphic to L. Moreover, the order-embedding φ is a lattice homomorphism of lattices; preserves all joins and meets which exist in P; for any adjoint pair $(\leftarrow, \&)$ in L, there exists an adjoint pair $(\leftarrow_{\bar{L}}, \&_{\bar{L}})$ in $Im(\varphi)$. Finally, for any connective in L there exists an associated connective in $\bar{L}$.*

[6] Then, it is a bounded lattice, i.e. it has bottom and top elements, denoted by $\bot$ and $\top$, respectively.

 ISBN: 978-84-614-6167-7

Proof. We will prove first that the map $\varphi : L \to DM(L)$, given by $\varphi(x) =\downarrow x$ is an order-embedding, homomorphism and preserves the indicated joins and meets. Indeed:

i) φ is a map since $\varphi(x) =\downarrow x \in DM(L)$ for all $x \in DM(L)$: one has only to consider $(\downarrow x)^{\uparrow\downarrow} =\downarrow x$.

ii) φ is injective: if we assume that $\varphi(x) =\downarrow x =\downarrow y = \varphi(y)$, then $x \in\downarrow x =\downarrow y$, so $x \leq y$. Similarly, shows that $y \leq x$, and we obtain $x = y$ by the antisymmetric property.

iii) φ is order-preserving: if $x \leq y$ in L implies $\varphi(x) \leq \varphi(y)$ in $DM(L)$, by definition of lower bound.

iv) Also, φ is an order-embedding: if $\varphi(x) \leq \varphi(y)$, then $\downarrow x \subset\downarrow y$. Since $x \in\downarrow x$, $x \in\downarrow y$ and therefore $x \leq y$.

iv) φ is a lattice homomorphism, i.e., $\varphi(x \wedge y) = \varphi(x) \cap \varphi(y)$, $\varphi(x \vee y) = \varphi(x) \cup \varphi(y)$. Certainly, we shall prove the equality of both sets. If $z \in \varphi(x \wedge y)$ it holds that $z \leq x \wedge y$ and, by definition of greatest lower bound, $z \leq x, z \leq y$. Then, we have $z \in \varphi(x), z \in \varphi(y)$, that is, $z \in \varphi(x) \cap \varphi(y)$. Thus, we obtain, $\varphi(x \wedge y) \subset \varphi(x) \cap \varphi(y)$. The reverse inclusion, $\varphi(x) \cap \varphi(y) \subset \varphi(x \wedge y)$, is analogous, like the dual result $\varphi(x \vee y) = \varphi(x) \cup \varphi(y)$.

v) φ preserves all joins and meets wich exist in P. Let A be a subset of L and assume that $\bigvee A$ exists in L. We shall prove that $\varphi(\bigvee A) = \bigvee \varphi(A)$, namely, $\downarrow (\bigvee A) = \bigvee\{\downarrow a : a \in A\}$. It is easy to prove that $\downarrow (\bigvee A)$ is an upper bound for $\{\downarrow a : a \in A\}$. Moreover, if $B \in DM(L)$ is an upper for the set $\{\downarrow a : a \in A\}$, we have $a \in\downarrow a \subset B$ for all $a \in A$, and therefore $A \subset B$. On the other hand, if $\downarrow (\bigvee A)$ exists in L, $\downarrow (\bigvee A) = {}^{7}A^{ul}$, and so $\downarrow (\bigvee A) = A^{ul} \subset B^{ul} = B$.

Likewise, if $\bigwedge A$ exists in L, we shall prove that $\varphi(\bigwedge A) = \bigwedge \varphi(A)$, that is, $\downarrow (\bigwedge A) = \bigwedge\{\downarrow a : a \in A\}$. Since, $\bigwedge\{\downarrow a : a \in A\} = \bigcap\{\downarrow a : a \in A\}$, we have the intended result.

Furthermore, we shall show that for any adjoint pair $(\leftarrow, \&)$ in L, there exists an adjoint pair $(\leftarrow_{\bar{L}}, \&_{\bar{L}})$ in $Im(\varphi)$, set denoted by $\bar{L}$. First, let $A, B, C \in \bar{L}$, $A = \varphi(x), B = \varphi(y), C = \varphi(z)$ be, for $x, y, z \in L$; then, we define the conjunction $\&_{\bar{L}}$ and the implication $\leftarrow_{\bar{L}}$ as

$$A\&_{\bar{L}}B = \varphi(x)\&_{\bar{L}}\varphi(y) := \varphi(x\&y) \qquad B \leftarrow_{\bar{L}} C = \varphi(y) \leftarrow_{\bar{L}} \varphi(z) := \varphi(y \leftarrow z)$$

resulting the following properties:

i) $\&_{\bar{L}}$ is increasing in both arguments: we shall show that if $A_1 \subset A_2$, then $A_1\&_{\bar{L}}B \subset A_2\&_{\bar{L}}B$. Since $A_1 = \varphi(x_1), A_2 = \varphi(x_2), B = \varphi(y)$, with $x_1, x_2, y \in L$, we have that $A_1\&_{\bar{L}}B = \varphi(x_1\&y) \subset \varphi(x_2\&y) = A_2\&_{\bar{L}}B$ being as φ is order-preserving and $\&$ is increasing in the first argument. Likewise, the increase in the second component is obtained.

[7] By definition of least upper bound and since A^{ul} is a down-set.

 ISBN: 978-84-614-6167-7

ii) $\&_{\bar{L}}$ has identity element, in particular the identity of $\bar{L}$ denoted by $\top_{\bar{L}}$ and is the set $\top_{\bar{L}} = \downarrow \top = \{z \in L : z \leq \top\} = L$. We need to check that $\top_{\bar{L}} \&_{\bar{L}} A = A$, for all $A \in \bar{L}$. Certainly, if $A = \varphi(x), x \in L$, we have $\top_{\bar{L}} \&_{\bar{L}} A = L \&_{\bar{L}} A = \varphi(\top \& x) = \varphi(x) = A$, because $\&$ is a conjunction in $\bar{L}$ and $\top$ is the identity element of $\bar{L}$.

iii) $\leftarrow_{\bar{L}}$ is increasing in the first argument and decreasing in the second argument or, more accurately, $\leftarrow_{\bar{L}}$ is order-preserving in the consequent and order-reversing in the antecedent. Regarding the antecedent, we need to prove that if $C_1 \subset C_2$, then $B \leftarrow_{\bar{L}} C_1 \supset B \leftarrow_{\bar{L}} C_2$. Since $C_1 = \varphi(z_1), C_2 = \varphi(z_2), B = \varphi(y)$, with $z_1, z_2, y \in L$, we have that $B \leftarrow_{\bar{L}} C_1 = \varphi(y \leftarrow z_1) \supset \varphi(y \leftarrow z_2) = B \leftarrow_{\bar{L}} C_2$ because $\leftarrow_L$ is an implication and φ is order-preserving. Similarly, the behavior in the consequent is obtained.

iv) $(\leftarrow_{\bar{L}}, \&_{\bar{L}})$ is an adjoint pair: we need to check that for any $A, B, C \in \bar{L}$, $A \subset (B \leftarrow_{\bar{L}} C) \Leftrightarrow A \&_{\bar{L}} C \subset B$ is fulfilled. Given $A, B, C \in \bar{L}$, $A = \varphi(x), B = \varphi(y), C = \varphi(z)$, with $x, y, z \in L$. For the first expression, we have $A \subset (B \leftarrow_{\bar{L}} C) \Leftrightarrow \varphi(x) \subset \varphi(y \leftarrow z) \Leftrightarrow x \subset (y \leftarrow z)$, where we use in the last step that φ is an order-embedding. On the other hand, using the definition of $\&_{\bar{L}}$ and again the character of order-embedding of φ, $A \&_{\bar{L}} C \subset B \Leftrightarrow \varphi(x \& z) \subset \varphi(y) \Leftrightarrow x \& z \leq y$, and we have the indeed equality in virtue of the adjoint property of pair $(\leftarrow, \&)$ in lattice L.

Finally, each connective in L defines a connective in $\bar{L}$, more detailed:

a) if $\wedge$ is a conjunction in L, there exists an associated conjunction $\bar{\wedge}$ in $\bar{L}$. The commutative and associative properties of $\bar{\wedge}$ are derived from the respective of $\wedge$. Moreover, $\bar{\wedge}$ verifies claims $i), ii)$ before that we have shown for $\&_{\bar{L}}$. All is routine and we omit it.

b) if $\vee$ is a disjunction in L, there exists a disjunction $\bar{\vee}$ in $\bar{L}$ associated. Similarly *a)*.

c) if @ is a aggregator in L, there exists an associated aggregator $\bar{@}$ in $\bar{L}$..

It is easy to prove that φ is suryective if φ preserves all joins and meets. In this case, L is isomorphic to (complete) multi-adjoint lattice $DM(P)$. □

4 Multi-adjoint Lattices in Practice using FLOPER

From now, we proceed with more practical aspects regarding multi-adjoint lattices and implementation issues. The parser of our FLOPER tool [27, 28] has been implemented by using the Prolog language. Once the application is loaded inside a Prolog interpreter, it shows a menu which includes options for loading/compiling, parsing, listing and saving fuzzy programs, as well as for executing/debugging fuzzy goals. Moreover, in [27] we explain that FLOPER has been recently equipped with new options, called

 ISBN: 978-84-614-6167-7

"`lat`" and "`show`", for allowing the possibility of respectively changing and displaying the multi-adjoint lattice associated to a given program, as we are going to explain.

When modeling a lattice to be loaded into FLOPER, all its relevant components must be encapsulated inside a Prolog file which must necessarily contain the definitions of a minimal set of predicates defining the set of valid elements (including special mentions to the "`top`" and "`bottom`" ones), the full or partial ordering established among them, as well as the repertoire of fuzzy connectives which can be used for their subsequent manipulation. In order to simplify our explanation, assume that file "bool.pl" refers to the simplest notion of (a binary) adjoint lattice, thus implementing the following set of predicates:

- `member/1` which is satisfied when being called with a parameter representing a valid truth degree. For instance, in the Boolean case, both predicates can be simply modeled by the Prolog facts: `member(0).`, `member(1).` and `members([0,1]).`
- `bot/1` and `top/1` obviously answer with the top and bottom element of the lattice, respectively. Both are implemented into "bool.pl" as `bot(0).` and `top(1).`
- `leq/2` models the ordering relation among all the possible pairs of truth degrees, and obviously it is only satisfied when it is invoked with two elements verifying that the first parameter is equal or smaller than the second one. So, in our example it suffices with including into "bool.pl" the facts: `leq(0,X).` and `leq(X,1).`
- Finally, if we have some fuzzy connectives of the form $\&_{label_1}$ (conjunction), $\vee_{label_2}$ (disjunction) or $@_{label_3}$ (aggregation) with arities n_1, n_2 and n_3 respectively, we must provide clauses defining the *connective predicates* "`and_`$label_1$`/(`n_1`+1)`", "`or_`$label_2$`/(`n_2`+1)`" and "`agr_`$label_3$`/(`n_3`+1)`", where the extra argument of each predicate is intended to contain the result achieved after the evaluation of the proper connective. For instance, in the Boolean case, the following two facts model in a very easy way the behaviour of the classical conjunction operation: `and_bool(0,_,0).` `and_bool(1,X,X).`

The reader can easily check that the use of lattice "bool.pl" when working with MALP programs whose rules have the form: "$A \leftarrow_{bool} \&_{bool}(B_1, \ldots, B_n)$ *with* 1", being A and B_i typical atoms, successfully mimics the behaviour of classical Prolog programs where clauses accomplish with the shape "A :- $B_1, \ldots, B_n$". As a novelty in the fuzzy setting, the outputs associated to the evaluation of goals will contain the corresponding Prolog's substitution (i.e., the *crisp* notion of computed answer obtained by means of classical SLD-resolution) together with the maximum truth degree 1.

On the other hand, and following the Prolog style regulated by the previous guidelines, in file "num.lat" we have included the clauses shown in Figure 1. Here, we have modeled the more flexible lattice which enables the possibility of working with truth degrees in the real interval $[0, 1]$, allowing too the possibility of using conjunction and disjunction operators recasted from the three typical fuzzy logics proposals described before (i.e., the *Łukasiewicz*, *Gödel* and *product* logics), as well as a useful description for the hybrid aggregator *average*.

 ISBN: 978-84-614-6167-7

```
member(X) :- number(X),0=<X,X=<1.

bot(0).              top(1).             leq(X,Y) :- X=<Y.

and_luka(X,Y,Z) :- pri_add(X,Y,U1),pri_sub(U1,1,U2),pri_max(0,U2,Z).
and_godel(X,Y,Z):- pri_min(X,Y,Z).
and_prod(X,Y,Z) :- pri_prod(X,Y,Z).

or_luka(X,Y,Z)  :- pri_add(X,Y,U1),pri_min(U1,1,Z).
or_godel(X,Y,Z) :- pri_max(X,Y,Z).
or_prod(X,Y,Z)  :- pri_prod(X,Y,U1),pri_add(X,Y,U2),pri_sub(U2,U1,Z).

agr_aver(X,Y,Z) :- pri_add(X,Y,U),pri_div(U,2,Z).

pri_add(X,Y,Z)  :- Z is X+Y.     pri_min(X,Y,Z) :- (X=<Y,Z=X;X>Y,Z=Y).
pri_sub(X,Y,Z)  :- Z is X-Y.     pri_max(X,Y,Z) :- (X=<Y,Z=Y;X>Y,Z=X).
pri_prod(X,Y,Z) :- Z is X * Y.   pri_div(X,Y,Z) :- Z is X/Y.
```

Figure 1: Multi-adjoint lattice modeling truth degrees in the real interval [0,1].

Note also that we have included definitions for auxiliary predicates, whose names always begin with the prefix "`pri_`". All of them are intended to describe primitive/arithmetic operators (in our case $+$, $-$, $*$, $/$, min and max) in a Prolog style, for being appropriately called from the bodies of clauses defining predicates with higher levels of expressivity (this is the case for instance, of the three kinds of fuzzy connectives we are considering: conjuntions, disjunctions and agreggations).

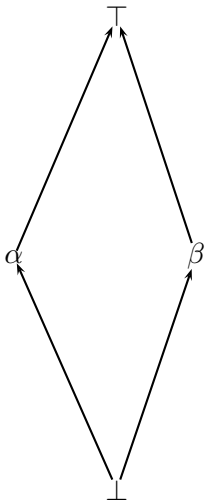
Assume that "new_num.pl" contains the same Prolog code than "num.pl" with the exception of the definition regarding the average aggregator. Now, instead of computing the average of two truth degrees, let us consider a new version which computes the average between the results achieved after applying to both elements the disjunctions operators described by Gödel and Łukasiewicz, that is: $@_{\mathtt{aver}}(x_1, x_2) = (\vee_{\mathtt{G}}(x_1, x_2) + \vee_{\mathtt{L}}(x_1, x_2))/2$ (where $\vee_{\mathtt{G}}(x_1, x_2) = max(x_1, x_2)$ and $\vee_{\mathtt{L}}(x_1, x_2) = min(1, x_1, x_2)$. The corresponding Prolog clause modeling such definition into the "new_num.pl" file is:

```
agr_aver(X,Y,Z)  :-  or_godel(X,Y,Z1),or_luka(X,Y,Z2),
                     pri_add(Z1,Z2,Z3), pri_div(Z3,2,Z).
```

And now, if with the new lattice we execute goal p w.r.t. the same program seen in Section 1 (introduction), instead of obtaining 0.63, the new solution will be 0.81 since now $\&_{\mathtt{P}}(0.9, @_{\mathtt{aver}}(0.8, 0.6)) = \&_{\mathtt{P}}(0.9, (\vee_{\mathtt{G}}(0.8, 0.6) + \vee_{\mathtt{L}}(0.8, 0.6))/2) = 0.9 * (max(0.8, 0.6) + min(1, 0.8 + 0.6))/2) = 0.9 * ((0.8 + 1)/2) = 0.81$.

To finish this section, in our last example we consider the partially ordered multi-adjoint lattice of Figure 2, for which the conjunction and implication connectives based on the *Gödel* intuistionistic logic conforms and adjoint pair (in the general case, the *Gödel*'s conjunctor is expressed in terms of "inf" instead of "min").

 ISBN: 978-84-614-6167-7




```
member(bottom).    member(alpha).
member(beta).      member(top).

leq(bottom,X). leq(alpha,alpha).  leq(alpha,top).
leq(beta,beta).  leq(beta,top).  leq(X,top).

and_godel(X,Y,Z) :- pri_inf(X,Y,Z).

pri_inf(bottom,X,bottom):-!.
pri_inf(alpha,X,alpha):-leq(alpha,X),!.
pri_inf(beta,X,beta):-leq(beta,X),!.
pri_inf(top,X,X):-!.
pri_inf(X,Y,bottom).
```

Figure 2: Partially-ordered multi-adjoint lattice.

Conclusions and Future Work

This paper has been mainly concerned with the Dedekind-MacNeille completion, a relevant and elegant mathematical concept which might help us to adapt some lattices for being safely used into the multi-adjoint logic programming framework. In particular, we have shown a technique which let us to "skip" in some cases the hypothesis of complete lattice usually required in multi-adjoint lattices, being this hypothesis mandatory, for instance, when describing the fix-point and model-theoretic declarative semantics of MALP [15]. In particular, the results achieved in this paper are useful to justify the safe use into FLOPER, according the methodology explained in the last part of this paper, of those lattices (composed by all finite strings whose elements are formed from an arbitrary alphabet of symbols) used in [28, 26] for documenting with declarative traces the execution of goals at a very low computational cost.

Since many standard completions of a lattice arise for suitable choices of the sets of up-subset and down-subset of P, in the future we plan to consider the concept of canonical extensions (i.e., dense and compact completions) of lattices with additional operations (introduced in [9]) and its application in the case of the quasimulti-adjoint lattices. Lattices with additional operations emerge from linear logics([1, 12, 17]). The results obtained can be applied for bilattices and trilattices (common MALP domains).

Acknowledgements

This work was supported by the EU (FEDER), and the Spanish Science and Innovation Ministry (MICINN) under grant TIN 2007-65749 as well as by the Castilla-La Mancha Administration under grant PII1I09-0117-4481.

References

[1] G. Allwein and J. M. Dunn. Kripke models for linear logic. *The Journal of Symbolic Logic*, 58(2):514–545, 1993.

[2] J.M. Almendros-Jiménez, A. Luna, and G. Moreno. A Flexible XPath-based Query Language Implemented with Fuzzy Logic Programming. In *Proc. of 5th Int. Symp. RuleML'11.* Springer LNCS (8 pages, in press), 2011.

[3] I. P. Cabrera, P. Cordero, and M. Ojeda-Aciego. Non-deterministic algebraic structures for soft computing. In *Proceedings of International Work-Conference on Artificial Neural Networks, IWANN'11*, 2011. Accepted.

[4] I.P. Cabrera, P. Cordero, G. Gutiérrez, J. Martínez, and M. Ojeda-Aciego. A coalgebraic approach to non-determinism: applications to multilattices. *Information Sciences*, 180(22):4323–4335, 2010.

[5] I.P. Cabrera, P. Cordero, G. Gutiérrez, J. Martínez, and M. Ojeda-Aciego. Coalgebras and non-determinism: an application to multilattices. In *Proceedings of Physics and Computation*, 2010.

[6] C.V. Damásio, N. Madrid, and M. Ojeda-Aciego. On the notions of residuated-based coherence and bilattice-based consistence. In *Proceedings of International Workshop on Fuzzy Logic and Applications WILF'11*, 2011. Accepted.

[7] B.A. Davey and H.A. Priestley. *Introduction to Lattices and Order.* Cambrige, University Press, UK, 2002. Second Edition.

[8] R. Dedekind. Stetigkeit und irrationale Zahlen. *Braunscheweig*, 1872.

[9] M. Gehrke and J. Harding. Bounded Lattice Expansions. *Journal of Algebra*, 238:345–371, 2001.

[10] M. Gehrke, J. Harding, and Y. Venema. MacNeille completions and canonical extensions. *Transactions of the American Mathematical Soc.*, 358(2):573–590, 2005.

[11] J.A. Guerrero and G. Moreno. Optimizing fuzzy logic programs by unfolding, aggregation and folding. *Elec. Notes in Theoretical Comp. Sci.*, 219:19–34, 2008.

[12] C. Hartonas. Duality for lattice-ordered algebras and for normal algebraizable logics. *Estudia Logica*, 58:403–450, 1997.

[13] B. Jonnson and A. Tarski. Boolean algebras with operators I. *American Journal of Mathematics*, 73:891–939, 1951.

[14] P. Julián, G. Moreno, and J. Penabad. On Fuzzy Unfolding. A Multi-adjoint Approach. *Fuzzy Sets and Systems, Elsevier*, 154:16–33, 2005.

 ISBN: 978-84-614-6167-7

[15] P. Julián, G. Moreno, and J. Penabad. On the declarative semantics of multi-adjoint logic programs. In *Proceedings of the 10th Int. Work-Conference on Artificial Neural Networks, IWANN'09*, pages 253–260, 2009. Springer, LNCS 5517.

[16] J.W. Lloyd. *Foundations of Logic Programming*. Springer-Verlag, Berlin, 1987.

[17] W. MacCaull. Kripke semantics for logics with bck implication. *Bulletin of the Section of Logic of the University of Lódź*, 25(1):41–51, 1996.

[18] H. M. MacNeille. Partially ordered sets. *Transactions of the American Mathematical Society*, pages 416–460, 1937.

[19] J. Medina and M. Ojeda-Aciego. Towards attribute reduction in multi-adjoint concept lattices. In *Proceedings of Concept Lattices and Applications, CLA'10*, pages 92–103, 2010.

[20] J. Medina, M. Ojeda-Aciego, and J. Ruiz-Calviño. Fuzzy logic programming via multilattices. *Fuzzy Sets and Systems*, 158(6):674–688, 2007.

[21] J. Medina, M. Ojeda-Aciego, and J. Ruiz-Calviño. On reachability of minimal models of multilattice-based logic programming. *LNAI*, 4827:271–282, 2007.

[22] J. Medina, M. Ojeda-Aciego, and J. Ruiz-Calviño. Formal concept analysis via multi-adjoint concept lattices. *Fuzzy Sets and Systems*, 160(2):130–144, 2009.

[23] J. Medina, M. Ojeda-Aciego, and P. Vojtáš. Multi-adjoint logic programming with continuous semantics. *Proceedings of Logic Programming and Non-Monotonic Reasoning, LPNMR'01, Springer, LNAI*, 2173:351–364, 2001.

[24] J. Medina, M. Ojeda-Aciego, and P. Vojtáš. A procedural semantics for multi-adjoint logic programing. *Progress in Artificial Intelligence, EPIA'01, Springer-Verlag, LNAI*, 2258(1):290–297, 2001.

[25] J. Medina, M. Ojeda-Aciego, and P. Vojtáš. Similarity-based Unification: a multi-adjoint approach. *Fuzzy Sets and Systems*, 146:43–62, 2004.

[26] P. J. Morcillo, G. Moreno, J. Penabad, and C. Vázquez. Declarative Traces into Fuzzy Computed Answers. In *Proceedings of 5th Int. Symp. RuleML'11*, Springer, LNCS (16 pages, in press), 2011.

[27] P.J. Morcillo, G. Moreno, J. Penabad, and C. Vázquez. A Practical Management of Fuzzy Truth Degrees using FLOPER. In *Proceedings of 4th Intl Symp. RuleML'10*, pages 119–126. Springer, LNCS 6403, 2010.

[28] P.J. Morcillo, G. Moreno, J. Penabad, and C. Vázquez. Fuzzy Computed Answers Collecting Proof Information. In *Proceedings of the 11th Int. Work-Conference on Artificial Neural Networks, IWANN'11* LNCS (in press), page 8. Springer, 2011.

[29] G. Voutsadakis. Dedekind-Macneille Completion of n-ordered Sets. *Order*, 24(1):15–29, 2007.

 ISBN: 978-84-614-6167-7

Proceedings of the 11th International Conference on Computational and Mathematical Methods in Science and Engineering, CMMSE 2011 26–30 June 2011.

Modeling the effect of bipolar trapping dopants on the current and efficiency of organic semiconductor devices

Luís F. Morgado[1], Luís Alcácer[2] and Jorge Morgado[2]

[1] *Dep. Física, Escola de Ciências e Tecnologia, Universidade de Trás-os-Montes e Alto Douro*

[2] *Intituto de Telecomunicações and Departamento de Engenharia Química e Biológica, Instituto Superior Técnico*

emails: lmorgado@utad.pt, alcacer@lx.it.pt, jorge.morgado@lx.it.pt

Abstract

A mathematical model is proposed to study the effect of bipolar doping of the emissive layer of organic light emitting diodes in the steady state situation. Considering the disordered nature of organic semiconductors, the carrier transport model takes into account the hopping nature of carrier transport, by considering that the density of states distribution function is Gaussian. Singlet excitons diffusion and energy transfer, formed from both free and trapped carriers are also included. Carriers injection into the emissive layer is determined self-consistently, considering the electrodes electronic characteristics.

Key words: organic light emitting diodes, charge carrier traps, device models, carrier density, carrier mobility, excitons

MSC 2000: 85.60.Jb,72.20.Jv, 71.35.-y

1 Introduction

Doping of the emissive layer of organic light-emitting diodes (OLEDs) has been mostly used to tune their emission colour and to increase the electroluminescence efficiency. In the first case, the process relies on the energy transfer from the host (which, in most cases, is the charge transporting material) to the guest molecules. In the second case, the increase of efficiency may result from an exciton confinement or due to the use of phosphorescent dopants in order to harvest the triplets. The presence of either guest or dopant traps has, usually, a detrimental effect on the charge transport, via charge localization or disorder, and leads also to an increase of the dopant emission

 ISBN: 978-84-614-6167-7

with respect to its contribution to the photoluminescence spectrum, as a result of on-site exciton formation promoted by that charge localization. In this study we model the use of dopants with a wider gap than the host polymer, thereby avoiding energy transfer from the host to the guest and promoting instead the reverse energy transfer process [4]. In this case, any on-guest formed exciton would be transferred to the host.

This model follows the work of Chih-Chien Lee *et al.* [3], for single layer devices, and includes some ideas from transport model of Refs. [2] and [7, 8], regarding the hopping nature of the carrier transport in organic semiconductors and self-consistency of the problem.

2 Transport model

Considering an OLED as a structure consisting of a very thin layer of luminescent organic material ($L \sim$100 nm thick) sandwiched between two metallic contacts. This configuration is usually modelled considering a one-dimensional in space formulation, along the perpendicular direction to the layers and considering hole (electron) injection at the left (right) where $x = 0$ ($x = L$). The carrier transport model in OLEDs consists of Poisson, current continuity and current density drift-diffusion equations.

2.1 Poisson-Drift-Diffusion equation

Taking into account the trapped carriers, the Poisson equation can be written as

$$\frac{\partial}{\partial x}\left(\varepsilon_r \frac{\partial \psi}{\partial x}\right) = -\frac{q}{\varepsilon_o}\left(p + p_t + p_d - n - n_t - n_d\right). \tag{1}$$

Here q stands for the elementary charge, $\varepsilon(x) = \varepsilon_r(x)\,\varepsilon_o$ is the dielectric constant of the organic material and T is the temperature. $\psi(x)$ is the electric potencial, $p(x)$, $p_t(x)$ and $p_d(x)$ ($n(x)$, $n_t(x)$ and $n_{td}(x)$) are, respectively positive (negative) free, trapped by host and trapped by guest carrier densities. The current continuity equations are

$$\frac{\partial p}{\partial t} = -\frac{1}{q}\frac{\partial J_p}{\partial x} - R_p, \tag{2}$$

$$\frac{\partial n}{\partial t} = \frac{1}{q}\frac{\partial J_n}{\partial x} - R_n, \tag{3}$$

and the drift-diffusion equations verified by the hole (electron) current density J_p (J_n) are

$$J_p = -q\mu_p p\frac{\partial \psi}{\partial x} - qD_p\frac{\partial p}{\partial x}, \tag{4}$$

$$J_n = -q\mu_n n\frac{\partial \psi}{\partial x} + qD_n\frac{\partial n}{\partial x}. \tag{5}$$

μ_p and μ_n are the hole and electron mobilities, respectively, considered to be, simultaneoulsly, electric field (Poole-Frenkel type) and carrier density-dependent [6],

$$\mu_{p,n}(x) = a_{p,n}\left\{\begin{array}{c} p(x)^{b_p} \\ n(x)^{b_n} \end{array}\right\}\exp\left(\gamma_{n,p}\sqrt{E(x)}\right), \tag{6}$$

 ISBN: 978-84-614-6167-7

where $E = -\frac{\partial \psi}{\partial x}$ and $\gamma_{n,p}$ are the electric field activation constants of the mobilities. The electric field dependence of the mobility, frequently observed in amorphous materials, including conjugated polymers, has been explained as arising from energetic disorder due to the interaction of charge carriers with randomly oriented permanent dipoles [1]. $D_p(D_n)$ is the hole (electron) diffusion coefficient, which are assumed to verify a generalized Einstein relation [5].

2.2 Recombination

The term $R_{p,n}(x)$ in the current continuity equations consists of Langevin-type recombination rates

$$R_p = \frac{q}{\varepsilon}\left\{(\mu_n + \mu_p)np + \mu_p p(n_d + n_t)\right\}, \quad (7)$$

$$R_n = \frac{q}{\varepsilon}\left\{(\mu_n + \mu_p)np + \mu_n n(p_d + p_t)\right\}, \quad (8)$$

where the recombination of the free with trapped carriers is included, considering the attractive Coulombic interaction between them.

References

[1] D. H. Dunlap, P. E. Parris, and V. M. Kenkre. Charge-dipole model for the universal field dependence of mobilities in molecularly doped polymers. *Phys. Rev. Lett.*, 77:542–545, 1996.

[2] E. Knapp, R. Hausermann, H. U. Schwarzenbach, and B. Ruhstaller. Numerical simulation of charge transport in disordered organic semiconductor devices. *Journal of Applied Physics*, 108(5):054504, 2010.

[3] Chih-Chien Lee, Mei-Ying Chang, Ping-Tsung Huang, Yen Chun Chen, Yih Chang, and Shun-Wei Liu. Electrical and optical simulation of organic light-emitting devices with fluorescent dopant in the emitting layer. *Journal of Applied Physics*, 101(11):114501, 2007.

[4] Luis Morgado, Luis Alcácer, , and Jorge Morgado. Polymer light-emitting diodes efficiency dependence on bipolar charge traps concentration. *Research Letters in Materials Science*, vol. 2009:Article ID 503042, 2009.

[5] Y. Preezant and N. Tessler. Self-consistent analysis of the contact phenomena in low-mobility semiconductors. *J. Appl. Phys.*, 93(4):2059–2064, 2003.

[6] M. C. J. M. Vissenberg and M. Matters. Theory of the field-effect mobility in amorphous organic transistors. *Phys. Rev. B*, 57(20):12964–12967, May 1998.

[7] S. V. Yampolskii, Yu. A. Genenko, C. Melzer, K. Stegmaier, and H. von Seggern. Bipolar charge-carrier injection in semiconductor/insulator/conductor heterostructures: Self-consistent consideration. *Journal of Applied Physics*, 104(7):073719, 2008.

 ISBN: 978-84-614-6167-7

[8] S. V. Yampolskii, Yu. A. Genenko, C. Melzer, and H. von Seggern. Self-consistent model of unipolar transport in organic semiconductor diodes: Accounting for a realistic density-of-states distribution. *Journal of Applied Physics*, 109(7):073722, 2011.

 ISBN: 978-84-614-6167-7

Volume III

 ISBN: 978-84-614-6167-7

 ISBN: 978-84-614-6167-7

Contents:

Volume I

 ISBN: 978-84-614-6167-7

Contents:

Volume II

 ISBN: 978-84-614-6167-7

 ISBN: 978-84-614-6167-7

 ISBN: 978-84-614-6167-7

Contents:

Volume III

 ISBN: 978-84-614-6167-7

 ISBN: 978-84-614-6167-7

 ISBN: 978-84-614-6167-7

Contents:
Volume IV

 ISBN: 978-84-614-6167-7

 ISBN: 978-84-614-6167-7

Contents:
Abstracts & Late Papers

 ISBN: 978-84-614-6167-7

 ISBN: 978-84-614-6167-7

Proceedings of the 11th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2011
26–30 June 2011.

Analysis of linear delay fractional differential initial value problems

M. L. Morgado [1,2], N. J. Ford[3] and P. M. Lima[2]

[1] *Department of Mathematics, University of Trás-os-Montes e Alto Douro, Portugal*

[2] *Centro de Matemática e Aplicações, Instituto Superior Técnico, Portugal*

[3] *Department of Mathematics, University of Chester, UK*

emails: luisam@utad.pt, n.ford@chester.ac.uk, pedro.t.lima@ist.utl.pt

Abstract

For a class of linear delay fractional differential initial value problems, we study the existence and the uniqueness of the solution by using the method of steps. An analytical representation of the solution is given based on the Mittag-Leffler functions. Also, we analyse how the solution is influenced by small perturbations on the initial function.

Key words: Fractional differential equation, delay differential equation, method of steps, Mittag-Leffler functions, initial value problem
MSC 2000: 65L05

1 Introduction

In the past last decades it has been observed an increasing interest in the study of fractional differential equations, mainly because recent investigations in science and engineering have demonstrated that the dynamic of many systems is described more accurately by using differential equations of non-integer order. On the other hand, in real world systems, delay is recognized everywhere. However, fractional delay differential equations (FDDEs) is a very recent topic. Although it seems natural to model certain processes and systems in engineering and other sciences (with memory and heritage properties) with this kind of equations, only in the last decade, some attention in the scientific community has been devoted to them.

Concerning the existence of solution issues we refer to [7], [8] and [9]. In [7], Lakshmikantham provides sufficient conditions for the existence of solutions to initial value problems to single term nonlinear delay fractional differential equations, with the fractional derivative in the Riemann-Liouville sense. In [9], Ye *et al* investigate the existence of positive solutions for a class of single term delay fractional differential

 ISBN: 978-84-614-6167-7

equation. Later, in [8], for the same class of equations, sufficient conditions for the uniqueness of the solution are reported. Similarly to classical differential equations, the stability issues are a very important task in the fractional setting. We believe that one of the first results dealing with the stability analysis for delay fractional differential equations was provided in [4], where the authors gave an analytic stability bound for a simple class of single term equation, by using the Lambert function. Later, [10], taking into account the usefulness of models with FDDEs in control of robotics, the authors of that paper investigated the finite-time stability analysis of FDDEs. Remarking also the application of such models in robotics, in [6], Krol studies the asymptotic properties of a n-dimensional linear FDDE and proposes necessary and sufficient conditions for asymptotic stability. In [5], the authors study the stability of a system of FDDEs with multiple delays. In all these papers the order of the derivative lies between 0 and 1. In these three last works, the fractional derivative is given in the Caputo sense, since it is more convenient in applications.

In this paper, we focus on the following initial value problem for a linear fractional differential equation with finite delay $\tau > 0$:

$$\begin{aligned} D^\alpha y(t) &= ay(t-\tau) + by(t) + f(t), \quad t > 0, &(1)\\ y(t) &= \phi(t), t \in [-\tau, 0], &(2)\end{aligned}$$

where a and b are constants, f is a continuous function on $[0, T]$, $T > 0$, the initial function ϕ is continuous on $[-\tau, 0]$, and D^α is the Caputo derivative of order α:

$$D^\alpha y(t) :=^{RL} D^\alpha (y - T[y])(t)$$

where $T[y]$ is the Taylor polynomial of degree $\lfloor \alpha \rfloor$ for y, centered at 0, and ${}^{RL}D^\alpha$ is the Riemann-Liouville derivative of order α [1]. The latter is defined by ${}^{RL}D^\alpha := D^{\lceil \alpha \rceil} J^{\lceil \alpha \rceil - \alpha}$, where J^β is the Riemann-Liouville integral operator,

$$J^\beta y(t) := \frac{1}{\Gamma(\beta)} \int_0^t (t-s)^{\beta-1} y(s) ds$$

and $D^{\lceil \alpha \rceil}$ is the classical integer order derivative. Here we follow the usual notation, so $\lfloor \alpha \rfloor$ denotes the greatest integer smaller than α and $\lceil \alpha \rceil$ is the smallest integer greater or equal than α. Here we consider the case $0 < \alpha < 1$.

Let us denote by $C = C[\tau, 0]$ the space of continuous real functions φ defined on $[-\tau, 0]$, equipped with the norm

$$\|\varphi\| = \max_{-\tau \le t \le 0} |\varphi(t)|.$$

The paper organizes as follows: In section 2, we analyse the existence and uniqueness of the solution by using the method of steps. Also an explicit representation of such solution is provided by means of the Mittag-Leffler functions. In section 3 we analyse how the solution is influenced under small perturbations on the initial function.

We end with some conclusions and a resume of some ongoing work.

 ISBN: 978-84-614-6167-7

2 Existence and uniqueness of solution

In this section we investigate the existence and uniqueness of solution, by extending to the fractional case, the method of steps [3].

Let us consider first that $0 \leq t \leq \tau$. In this case, since $t - \tau \in [-\tau, 0]$, then $y(t-\tau) = \phi(t-\tau)$, and therefore equation (1) can be rewritten as

$$D^\alpha y(t) = by(t) + g_\tau(t), \quad 0 < t \leq \tau,$$

where $g_\tau(t) = a\phi(t-\tau) + f(t)$. Since f and ϕ are continuous functions, g_τ is also continuous and then this equation is solvable. Its general solution may be written as (see, for example Theorem 5.15 of [2])

$$y_\tau(t) = \int_0^t (t-s)^{\alpha-1} E_{\alpha,\alpha}\left(b(t-s)^\alpha\right)\left(a\phi(s-\tau) + f(s)\right) ds + c_\tau E_\alpha\left(bt^\alpha\right),$$

where c_τ is a constant and the functions $E_{\alpha,\beta}$ are the so-called Mittag-Leffler functions defined by

$$\begin{aligned} E_{\alpha,\beta}(z) &= \sum_{k=0}^{\infty} \frac{z^k}{\Gamma(\alpha k + \beta)}, \quad z, \beta \in \mathbb{C}, Re(\alpha) > 0 \\ E_\alpha(z) &= E_{\alpha,1} \end{aligned}$$

Note that taking (2) into account, the solution is uniquely determined since we must have $c_\tau = \phi(0)$.

Hence in the interval $[0, \tau]$ the solution of (1)-(2) exists and is unique. We can now proceed analogously in the interval $[\tau, 2\tau]$. In this interval, (1) may be rewritten as

$$D^\alpha y(t) = by(t) + g_{2\tau}(t), \quad \tau < t \leq 2\tau,$$

where $g_{2\tau}(t) = ay_\tau(t-\tau) + f(t)$. If g is continuous in $[\tau, 2\tau]$, then the solution in this interval exists, is unique and given by

$$y_{2\tau}(t) = \int_0^t (t-s)^{\alpha-1} E_{\alpha,\alpha}\left(b(t-s)^\alpha\right)\left(ay_\tau(s-\tau) + f(s)\right) ds + c_{2\tau} E_\alpha\left(bt^\alpha\right).$$

We can then easily conclude the following theorem:

Theorem 1 *Let k be the greatest positive integer such that the function*

$$g_{k\tau}(t) = ay_{(k-1)\tau}(t-\tau) + f(t) \quad \textit{with} \quad y_{0\tau}(t) = \phi(t)$$

is continuous.

Then the initial value problem (1)-(2) has on the interval $[0, k\tau]$ a unique solution that can be represented by $y(t) = y_{i\tau}(t)$, if $(i-1)\tau \leq t \leq i\tau$, where

$$y_{i\tau}(t) = \int_0^t (t-s)^{\alpha-1} E_{\alpha,\alpha}\left(b(t-s)^\alpha\right) g_{i\tau}(s) ds + c_{i\tau} E_\alpha\left(bt^\alpha\right), \quad t \in [(i-1)\tau, i\tau]$$

and $c_{i\tau}$ is a constant, $i = 1, \ldots, k$.

 ISBN: 978-84-614-6167-7

3 Dependence of the solution on the initial function

Assume that y and z are the respective solutions of the initial value problems (1)-(2) and

$$\begin{aligned} D^{\alpha}z(t) &= az(t-\tau)+bz(t)+f(t), \quad t>0, &(3)\\ z(t) &= \tilde{\phi}(t), t\in[-\tau,0], &(4)\end{aligned}$$

where $\tilde{\phi}$ is also continuous on $[-\tau,0]$. Assume also that

$$\left\|\phi-\tilde{\phi}\right\| = \max_{-\tau\le t\le 0}|\phi(t)-\tilde{\phi}(t)|\le\epsilon, \tag{5}$$

for some $\epsilon>0$ small. Note that since for every t where both solutions y and z exist, we have

$$\begin{aligned} y(t)-z(t) &= \phi(0)-\tilde{\phi}(0)+\frac{a}{\Gamma(\alpha)}\int_0^t(t-s)^{\alpha-1}\left(y(s-\tau)-z(s-\tau)\right)ds+\\ &+ \frac{b}{\Gamma(\alpha)}\int_0^t(t-s)^{\alpha-1}\left(y(s)-z(s)\right)ds \end{aligned} \tag{6}$$

we can easily determine how $|y(t)-z(t)|$ is influenced by $\left\|\phi-\tilde{\phi}\right\|$. This dependence is stated in the following theorem.

Theorem 2 *Let $y(t)$ and $z(t)$ be the respective solutions of the initial value problems (1)-(2) and (3)-(4), existing both on the interval $[0,k\tau]$, for some $k\ge 1$. Assuming that (5) holds, we have*

$$|y(t)-z(t)|\le A_k e^{\frac{|b|}{\Gamma(\alpha+1)}(k\tau)^{\alpha}}\epsilon, \quad t\in[(k-1)\tau,k\tau], \tag{7}$$

where

$$\begin{aligned} A_0 &= 1,\\ A_k &= 1+\frac{|a|}{\Gamma(\alpha+1)}\sum_{j=0}^{k-1}((k-j)\tau)^{\alpha}A_j e^{\frac{|b|}{\Gamma(\alpha+1)}(j\tau)^{\alpha}}. \end{aligned}$$

Proof. The proof will be done by induction on k.

Let us prove first that (7) holds for $k=1$.

If $t\in[0,\tau]$, taking (6) into account, we have

$$\begin{aligned} y(t)-z(t) &= \phi(0)-\tilde{\phi}(0)+\frac{a}{\Gamma(\alpha)}\int_0^t(t-s)^{\alpha-1}\left(\phi(s-\tau)-\tilde{\phi}(s-\tau)\right)ds+\\ &+ \frac{b}{\Gamma(\alpha)}\int_0^t(t-s)^{\alpha-1}\left(y(s)-z(s)\right)ds. \end{aligned}$$

 ISBN: 978-84-614-6167-7

Therefore,

$$\begin{aligned}
|y(t)-z(t)| &\leq |\phi(0)-\tilde{\phi}(0)| + \frac{|a|}{\Gamma(\alpha)} \max_{s\in[0,t]} |\phi(s-\tau)-\tilde{\phi}(s-\tau)| \int_0^t (t-s)^{\alpha-1} ds + \\
&+ \frac{|b|}{\Gamma(\alpha)} \int_0^t (t-s)^{\alpha-1} |y(s)-z(s)|\, ds \\
&\leq |\phi(0)-\tilde{\phi}(0)| + \frac{|a|}{\Gamma(\alpha)} \max_{s\in[0,\tau]} |\phi(s-\tau)-\tilde{\phi}(s-\tau)| \int_0^t (t-s)^{\alpha-1} ds + \\
&+ \frac{|b|}{\Gamma(\alpha)} \int_0^t (t-s)^{\alpha-1} |y(s)-z(s)|\, ds \\
&\leq \left\|\phi-\tilde{\phi}\right\| + \frac{|a|}{\Gamma(\alpha)} \left\|\phi-\tilde{\phi}\right\| \frac{t^\alpha}{\alpha} + \frac{|b|}{\Gamma(\alpha)} \int_0^t (t-s)^{\alpha-1} |y(s)-z(s)|\, ds \\
&\leq \left(1 + \frac{|a|}{\Gamma(\alpha+1)}\tau^\alpha\right)\epsilon + \frac{|b|}{\Gamma(\alpha)} \int_0^t (t-s)^{\alpha-1} |y(s)-z(s)|\, ds \\
&= A_1\epsilon + \frac{|b|}{\Gamma(\alpha)} \int_0^t (t-s)^{\alpha-1} |y(s)-z(s)|\, ds.
\end{aligned}$$

It follows, by a Gronwall inequality, that

$$|y(t)-z(t)| \leq A_1 e^{\frac{|b|}{\Gamma(\alpha+1)}\tau^\alpha}\epsilon, \quad t\in[0,\tau],$$

proving that (7) holds for $k=1$.

Next suppose that (7) holds for $(k-1)$, that is, let us assume that for $t\in[(k-2)\tau,(k-1)\tau]$, the following inequalty is satisfied:

$$|y(t)-z(t)| \leq A_{k-1} e^{\frac{|b|}{\Gamma(\alpha+1)}((k-1)\tau)^\alpha}\epsilon, \tag{8}$$

with

$$A_{k-1} = 1 + \frac{|a|}{\Gamma(\alpha+1)} \sum_{j=0}^{k-2} ((k-1-j)\tau)^\alpha A_j e^{\frac{|b|}{\Gamma(\alpha+1)}(j\tau)^\alpha}.$$

Let us the prove that it will also be valid for k.

Whenever $t\in[(k-1)\tau,k\tau]$, $k\geq 2$, taking (6) into account, we have

$$\begin{aligned}
y(t)-z(t) &= \phi(0)-\tilde{\phi}(0) + \frac{a}{\Gamma(\alpha)} \int_0^\tau (t-s)^{\alpha-1} \left(\phi(s-\tau)-\tilde{\phi}(s-\tau)\right) ds + \\
&+ \sum_{j=2}^{k-1} \frac{a}{\Gamma(\alpha)} \int_{(j-1)\tau}^{j\tau} (t-s)^{\alpha-1} \left(y(s-\tau)-z(s-\tau)\right) ds + \\
&+ \frac{a}{\Gamma(\alpha)} \int_{(k-1)\tau}^t (t-s)^{\alpha-1} \left(y(s-\tau)-z(s-\tau)\right) ds \\
&+ \frac{b}{\Gamma(\alpha)} \int_0^t (t-s)^{\alpha-1} \left(y(s)-z(s)\right) ds.
\end{aligned}$$

 ISBN: 978-84-614-6167-7

Hence

$$\begin{aligned}
|y(t)-z(t)| &\leq |\phi(0)-\tilde{\phi}(0)| + \frac{|a|}{\Gamma(\alpha)} \max_{s\in[0,\tau]} |\phi(s-\tau)-\tilde{\phi}(s-\tau)| \int_0^{\tau} (t-s)^{\alpha-1} ds + \\
&+ \sum_{j=2}^{k-1} \frac{|a|}{\Gamma(\alpha)} \max_{s\in[(j-1)\tau,j\tau]} |y(s-\tau)-z(s-\tau)| \int_{(j-1)\tau}^{j\tau} (t-s)^{\alpha-1} ds + \\
&+ \frac{|a|}{\Gamma(\alpha)} \max_{s\in[(k-1)\tau,k\tau]} |y(s-\tau)-z(s-\tau)| \int_{(k-1)\tau}^{t} (t-s)^{\alpha-1} ds + \\
&+ \frac{|b|}{\Gamma(\alpha)} \int_0^t (t-s)^{\alpha-1} |y(s)-z(s)| \, ds.
\end{aligned}$$

By (8),

$$\begin{aligned}
|y(t)-z(t)| &\leq \epsilon + \frac{|a|}{\Gamma(\alpha+1)} \epsilon \left[t^{\alpha} - (t-\tau)^{\alpha}\right] + \\
&+ \sum_{j=2}^{k-1} \frac{|a|}{\Gamma(\alpha+1)} A_{j-1} \epsilon e^{\frac{|b|}{\Gamma(\alpha+1)}((j-1)\tau)^{\alpha}} \left((t-(j-1)\tau)^{\alpha} - (t-j\tau)^{\alpha}\right) + \\
&+ \frac{|a|}{\Gamma(\alpha+1)} A_{k-1} \epsilon e^{\frac{|b|}{\Gamma(\alpha+1)}((k-1)\tau)^{\alpha}} (t-(k-1)\tau)^{\alpha} + \\
&+ \frac{|b|}{\Gamma(\alpha)} \int_0^t (t-s)^{\alpha-1} |y(s)-z(s)| \, ds.
\end{aligned}$$

Since $(k-1)\tau \leq t \leq k\tau$,

$$\begin{aligned}
|y(t)-z(t)| &\leq \epsilon + \frac{|a|}{\Gamma(\alpha+1)} \epsilon (k\tau)^{\alpha} + \sum_{j=2}^{k-1} \frac{|a|}{\Gamma(\alpha+1)} A_{j-1} \epsilon e^{\frac{|b|}{\Gamma(\alpha+1)}((j-1)\tau)^{\alpha}} ((k-(j-1))\tau)^{\alpha} + \\
&+ \frac{|a|}{\Gamma(\alpha+1)} A_{k-1} \epsilon e^{\frac{|b|}{\Gamma(\alpha+1)}((k-1)\tau)^{\alpha}} \tau^{\alpha} + \frac{|b|}{\Gamma(\alpha)} \int_0^t (t-s)^{\alpha-1} |y(s)-z(s)| \, ds \\
&= \epsilon + \frac{|a|}{\Gamma(\alpha+1)} \epsilon (k\tau)^{\alpha} + \frac{|a|}{\Gamma(\alpha+1)} \sum_{i=1}^{k-2} A_i \epsilon e^{\frac{|b|}{\Gamma(\alpha+1)}(i\tau)^{\alpha}} ((k-i)\tau)^{\alpha} + \\
&+ \frac{|a|}{\Gamma(\alpha+1)} A_{k-1} \epsilon e^{\frac{|b|}{\Gamma(\alpha+1)}((k-1)\tau)^{\alpha}} \tau^{\alpha} + \frac{|b|}{\Gamma(\alpha)} \int_0^t (t-s)^{\alpha-1} |y(s)-z(s)| \, ds \\
&= \left(1 + \frac{|a|}{\Gamma(\alpha+1)} \sum_{j=0}^{k-1} ((k-j)\tau)^{\alpha} A_j e^{\frac{|b|}{\Gamma(\alpha+1)}(j\tau)^{\alpha}}\right) \epsilon + \frac{|b|}{\Gamma(\alpha)} \int_0^t (t-s)^{\alpha-1} |y(s)-z(s)| \, ds \\
&= A_k \epsilon + \frac{|b|}{\Gamma(\alpha)} \int_0^t (t-s)^{\alpha-1} |y(s)-z(s)| \, ds,
\end{aligned}$$

and the result follows by a Gronwall inequality.

 ISBN: 978-84-614-6167-7

4 Conclusions and ongoing work

For a class of linear delay fractional initial value problems an explicit representation of the solution is given, by means of the Mittag-Leffler functions. This is also used to provide the existence and uniqueness results. The influence on such solution of small perturbations of the initial function is also investigated.

Being one of our main purposes the construction of reliable numerical methods for FDDEs, we continue our investigation on the analysis of these kind of problems (the linear ones, for now), namely investigating some other issues on the structural stability, as for example the dependence of the solution under small perturbations in the order of the derivative, and on the right-hand function in (1). The smoothness of the solution is another ongoing investigation. Finally we will provide a comparison of several numerical schemes for this kind of problems.

Acknowledgements

M. L. Morgado acknowledges financial support from FCT, Fundação para a Ciência e Tecnologia, through grant SFRH/BPD/46530/2008.

References

[1] S. G. Samko, A. A. Kilbas and O. I. Marichev, *Fractional Integrals and Derivatives: Theory and Applications*, Gordon and Breach, Yverdon, 1993.

[2] A. A. Kilbas, H. M. Srivastava and J. J. Trujillo, *Theory and Applications of Fractional Differential Equations*, North-Holland Mathematics Studies, 2006.

[3] R. Bellman and K. L. Cooke, *Differential-Difference Equations*, New York-London: Academic Press, 1963.

[4] Y. Chen and K. L. Moore, *Analytical Stability Bound for a Class of Delayed Fractional-Order Dynamic Systems*, Nonlinear Din., **29** (2002) 191-200.

[5] W. Deng, C. Li and J. Lu, *Stability analysis of linear fractional differential system with multiple time delays*, Nonlinear Din., **48** (2007) 409-416.

[6] K. Krol, *Asymptotic Properties of Fractional Delay Differential Equations*, submitted, available at http://www.mathematik.hu-berlin.de/publ/pre/2008/P-08-18.pdf.

[7] V. Lakshmikantham, *Theory of fractional functional differential equations*, Nonlinear Science, **69** (2008) 3337-3343.

[8] C. Liao and H. Ye, *Existence of positive solutions of nonlinear fractional delay differential equations*, Positivity, **13** (2009) 601-609.

[9] H. Ye, Y. Ding and J. Gao, *The existence of a positive solution of* $D^{\alpha}[x(t) - x(0)] = x(t)f(t, xt)$, Positivity, **11** (2007) 341-350.

[10] Mihailo P. Lazarević and Aleksandar M. Spasić, *Finite-time stability analysis of fractional order time-delay systems: Gronwall's approach*, Mathematical and Computer Modelling, **49** (2009) 475-481.

 ISBN: 978-84-614-6167-7

Proceedings of the 11th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2011
26–30 June 2011.

Numerical solution of high order differential equations with Bernoulli boundary conditions

Anna Napoli[1]

[1] *Department of Mathematics, University of Calabria, 87036 Rende (Cs), Italy*

emails: a.napoli@unical.it

Abstract

For the numerical solution of high order boundary value problems with special boundary conditions a general collocation method is derived and studied.

Key words: Boundary value problem, Bernoulli polynomials, interpolation
MSC 2000: 65L10, 65D05

We consider the general n-th order boundary value problem for nonlinear ordinary differential equations

$$y^{(n)}(x) = f(x, \mathbf{y}(x)) \qquad x \in [a,b] \tag{1}$$

where $\mathbf{y}(x) = \left(y(x), y'(x), \ldots, y^{(q)}(x)\right), 0 \le q \le n-1$ and $f : [a,b] \times \mathbb{R}^{q+1} \to \mathbb{R}$. The particular boundary conditions are the so called *Bernuolli boundary conditions* ([6])

$$\begin{aligned} &y(a) = \beta_0 \\ &y^{(k)}(b) - y^{(k)}(a) = \beta_{k+1} \qquad k = 0, \ldots, n-2 \end{aligned} \tag{2}$$

with β_k, $k = 0, \ldots, n-1$ real constants.

In [5] a non constructive proof of the existence and uniqueness of solution is given, under the hypothesis that the function satisfies Lipschitz condition in a certain domain interval of $[a,b] \times \mathbb{R}^{q+1}$. In [6] Picard's method is applied in connection with Newton's method for the numerical solution of (1)-(2). Here a family of general collocation methods for the numerical solution of (1)-(2) is proposed.

Let $B_k(x)$ be the Bernoulli polynomial of degree k and let us set $h = b - a$, $S_k = B_k(t) - B_k(0)$, $\Delta f_a^{(k)} = f^{(k)}(b) - f^{(k)}(a)$. The boundary value problem (1)-(2) is equivalent to the nonlinear Fredholm integral equation ([6])

$$y(x) = P_{n-1}[y,x] + \int_a^b G_{n-1}(x,t) f(t, \mathbf{y}(t)) dt \tag{3}$$

where $$G_n(x,t) = \frac{1}{n!}\left[(x-t)_+^n - \sum_{k=1}^{n} S_k\left(\frac{x-a}{h}\right)\frac{h^{k-1}}{k}\binom{n}{k-1}(b-t)^{n-k+1}\right]$$

 ISBN: 978-84-614-6167-7

and $$P_n[f,x] = f(a) + \sum_{k=1}^{n} S_k\left(\frac{x-a}{h}\right)\frac{h^{k-1}}{k!}\Delta f_a^{(k-1)}.$$

$P_n[f,x]$ satisfies the Bernoulli interpolation problem ([2])

$$P_n[f,a] = f(a), \qquad P_n[f,b] = f(b),$$
$$\Delta P_n^{(k)} = P_n^{(k)}[f,b] - P_n^{(k)}[f,a] = f^{(k)}(b) - f^{(k)}(a) = \Delta f_a^{(k)}, \quad k = 1,\dots,n-1.$$

Let $y(x)$ be the solution of (1)-(2) and let x_i, $i = 1,\dots,m$, be m distinct points in $[a,b]$. We prove that the polynomial

$$y_{n,m}(x) = P_{n-1}[y,x] + \sum_{i=1}^{m} p_{n,i}(x) f\left(x_i, \mathbf{y}_{n,m}(x_i)\right) \tag{4}$$

where $\mathbf{y}_{n,m}(x) = \left(y_{n,m}(x), y'_{n,m}(x), \dots, y^{(q)}_{n,m}(x)\right)$, $0 \le q \le n-1$, and

$$p_{n,i}(x) = \int_a^b G_{n-1}(x,t) l_i(t)\, dt \qquad i = 1,\dots,n-1$$

is a collocation polynomial for (1)-(2) on the nodes x_i.

An a-priori estimation of the global error is possible: let $Q_m = \max\limits_{0\le s\le q}\left\{\max\limits_{a\le x\le b}\sum\limits_{i=1}^{m}\left|p_{ni}^{(s)}(x)\right|\right\}$, $\overline{R}_m(y,x)$ be the Lagrange interpolation remainder, and $L = \sum\limits_{k=0}^{q} L_k$ where L_k are the Lipschitz constants of the function f. If $LQ_m < 1$, then

$$\|y - y_{n,m}\| \le \frac{R\Delta}{1 - LQ_m}$$

where $$\|y\| = \max_{0\le s\le q}\left\{\max_{a\le t\le b}\left|y^{(s)}(t)\right|\right\}, \qquad R = \max_{a\le x\le b}\left|\overline{R}_m(y,x)\right|,$$

$$D_{n,s} = \max_{a\le x\le b}\sum_{k=s}^{n-1}\frac{\left|B_{k-s}\left(\frac{x-a}{h}\right)\right|}{(k-s)!(n-k)!}, \quad \Delta = \max_{0\le s\le q}\left\{\frac{h^{n-s+1}}{(n-s+1)!} + h^{n-s-1} D_{n,s}\right\}.$$

Then the the fifth-order BVPs are considered. These problems generally arise in the mathematical modeling of viscoelastic flows and other branches of mathematical, physical and engineering sciences ([7]). In this case

$$y_{5,m}(x) = P_4[y,x] + \sum_{i=1}^{m} p_{5,i}(x) f\left(x_i, y_{5,m}(x_i)\right)$$

where

$$\begin{aligned} 4!\, p_{5,i}(x) = {} & \left(x^4 - 2x^3 + x^2\right)\left[F_{i2}(x) - M_{i2}(x)\right] - 2x(2x^2 - 3x + 1)\left[F_{i3}(x) + M_{i3}(x)\right] \\ & + 12x(x-1)\left[F_{i4}(x) - M_{i4}(x)\right] + 24(1-x)F_{i5}(x) - 24xM_{i5}(x) \end{aligned}$$

 ISBN: 978-84-614-6167-7

$$F_{i1}(x) = \int_0^x l_i(t)\,dt, \qquad M_{i1}(x) = \int_x^1 l_i(t)\,dt$$

$$F_{ik}(x) = \int_0^x F_{i,k-1}(t)dt, \qquad M_{ik}(x) = \int_x^1 M_{i,k-1}(t)dt \qquad k = 2,\dots,5.$$

To calculate the approximate solution of problem (1)-(2) by (4) at $x \in [a,b]$, we need the values $y_i^{(k)} = y_{n,m}^{(k)}(x_i)$, $i = 1,\dots,m$, $k = 0,\dots,q$. These values can be calculated by solving the following system

$$y_i^{(k)} = P_{n-1}^{(k)}[y,x_i] + \sum_{j=1}^{m} p_{nj}^{(k)}(x_i) f\left(x_j, \mathbf{y}_j\right) \qquad i = 1,\dots,m,\ k = 0,\dots,q \tag{5}$$

with $\mathbf{y}_j = \left(y_{j,m}, y'_{j,m}, \dots, y_{j,m}^{(q)}\right)$, $0 \le q \le n-1$. We prove that it has a unique solution.

To calculate $F_{ik}(x_j), M_{ik}(x_j)$, $i,j = 1,\dots m$, $k = 1,\dots,n$, it suffices to compute

$$\int_a^x r_{m,i}(t)dt\ , \qquad \underbrace{\int_a^x \cdots \int_a^x}_{k} r_{m,i}(t)\,dt \cdots dt$$

where $r_{0,0}(t) = 1, \quad r_{m,i}(t) = (t-x_1)\cdots(t-x_{i-1})(t-x_{i+1})\cdots(t-x_m) \quad i = 1,2,\dots,m$. Following the idea of Omar and Suleiman ([8]), we propose a recursive algorithm: for each $i = 1,\dots,m-1$, let us define the new points $z_j^{(i)} = x_j$ if $j < i$ and $z_j^{(i)} = x_{j+1}$ if $j \ge i$, $j = 1,\dots,m-1$. Moreover, let us define $g_{0,1,a}^{(i)}(x) = x-a$, and, for $s = 1,\dots,m-1$,

$$g_{s,j,a}^{(i)}(x) = \int_a^x \underbrace{\int_a^x \cdots \int_a^x}_{j-1} \left(t - z_1^{(i)}\right)\left(t - z_2^{(i)}\right)\cdots\left(t - z_s^{(i)}\right) dt \cdots dt\,. \tag{6}$$

For the computation of (6) the following recurrence formula holds

$$g_{s,j,a}^{(i)}(x) = \left(x - z_s^{(i)}\right) g_{s-1,j,a}^{(i)}(x) - j g_{s-1,j+1,a}^{(i)}(x)\,.$$

Numerical experiments demonstrate the practical usefulness of the proposed method.

Example 1.

$$\begin{cases} y^{(v)}(t) = e^{-t}y^2(t) & t \in [-1,1] \\ y\,(0) = 1, \quad y^{(k)}(1) - y^{(k)}(0) = e-1 & k = 0,\dots,3 \end{cases} \tag{7}$$

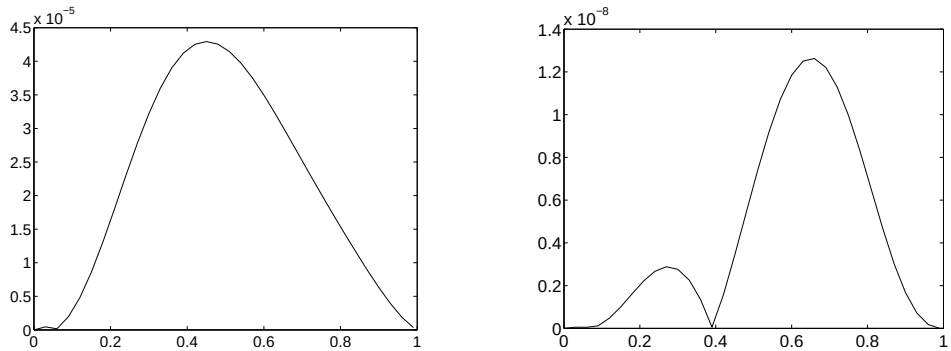
with solution $y\,(t) = e^t$. Figure (a) shows the graph of the error function $e(x)$.

Example 2.

$$\begin{cases} y^{(v)}(t) = -24e^{-5y} + \frac{48}{1+t^5} & t \in [0,1] \\ y\,(0) = 0 & y(1) = \lg 2 \\ y'(1) - y'(0) = -\frac{1}{2} & y''(1) - y''(0) = \frac{3}{4} \\ y'''(1) - y'''(0) = -\frac{7}{4} \end{cases} \tag{8}$$

with solution $y\,(t) = e^t$. The graph of $e(x)$ is plotted in Figure (b).

 ISBN: 978-84-614-6167-7



(a) (b)

References

[1] R. P. Agarwal, *Boundary value problems for higher order differential equations*, World Scientific, Singapore, 1986.

[2] F. Costabile, *Expansion of real functions in Bernoulli polynomials and applications*, Conferences Seminars Mathematics, Univesty of Bari, **273** (1999) 1–13.

[3] F. Costabile and A. Napoli, *Collocation for high order differential equations with Hermite boundary conditions*, submitted

[4] F. Costabile and A. Napoli, *A class of collocation methods for numerical integration of initial value problems*, submitted

[5] F. Costabile, A. Serpe and A. Bruzio, *No Classic Boundary Conditions*, in Proceedings of World Congress on Engineering 2007, Hong Kong (2007) 918-921.

[6] F. Costabile and A. Serpe, *On Bernoulli boundary value problems*, Le Matematiche, **LXII**(2) (2007) 163–173.

[7] A. Karageorghis, T. N. Phillips and A. R. Davies, *Spectral collocation methods for the primary two point boundary value problem in modelling viscoelastic flows*, Int. J. Numer. Methods Eng., **26** (1988) 805–813.

[8] Z. B. Omar and M. Suleiman, *Solving first order system of ordinary differential equations using parallel r-point block method of variable step size and order*, Chiang Mai J. Sci. **36**(1) (2009) 9–23.

 ISBN: 978-84-614-6167-7

Proceedings of the 11th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2011
26–30 June 2011.

Symbolic computation of the solution to a complete ODE

Juan F. Navarro[1] **and A. Pérez–Carrió**[1]

[1] *Department of Applied Mathematics, University of Alicante*

emails: jf.navarro@ua.es, a.perez@ua.es

Abstract

The aim of this contribution is to introduce a symbolic technique for the computation of the solution to a complete ordinary differential equation with constant coefficients. The symbolic solution is computed via the variation of parameters method, and thus, constructued over the exponential matrix of the linear system associated to the homogeneous equation. This matrix is also symbolically determined. The accuracy of the symbolic solution is tested by comparing it with the exact solution of a test problem.

Key words: template, instructions Perturbation methods, symbolic computation, differential equations

1 Introduction

Perturbation theories for differential equations containing a small parameter ϵ are quite old. The small perturbation theory originated by Sir Isaac Newton has been highly developed by many others, and an extension of this theory to the asymptotic expansion, consisting on a power series expansion in the small parameter, was devised by Poincaré (1892). The main point is that for the most of the differential equations, it is not possible to obtain an exact solution. In cases where equations contain a small parameter, we can consider it as a perturbation parameter to obtain an asymptotic expansion of the solution. In practice, the work involved in the application of this approach to compute the solution to a differential equation cannot be performed by hand, and algebraic processors result to be a very useful tool.

As explained in Henrard (1989), the first symbolic processors were developed to work with Poisson series, that is, multivariate Fourier series whose coefficients are multivariate Laurent series,

$$\sum_{i_1,\ldots,i_n} \sum_{j_1,\ldots,j_m} C_{i_1,\ldots,i_n}^{j_1,\ldots,j_m} x_1^{i_1} \cdots x_n^{i_n} \frac{\cos}{\sin}(j_1\phi_1 + \cdots + j_m\phi_m)\,,$$

 ISBN: 978-84-614-6167-7

where $C_{i_1,\ldots,i_n}^{j_1,\ldots,j_m} \in \mathbb{R}$, $i_1,\ldots,i_n,j_1,\ldots,j_m \in \mathbb{Z}$, and $x_1,\ldots,x_n$ and $\phi_1,\ldots,\phi_m$ are called polynomial and angular variables respectively. These processors were applied to problems in non–linear mechanics or non–linear differential equations problems, in the field of Celestial Mechanics.

In order to achieve better accuracies in the applications of analytical theories, high orders of the approximate solution must be computed, making necessary a continuous maintenance and revision of the existing symbolic manipulation systems, as well as the development of new packages adapted to the peculiarities of the problem to be treated. Recently, Navarro (2008a, 2008b) developed a symbolic processor to deal with the solution to perturbed second order differential equations. To face this problem, the algebraic processor handles objects called quasi–polynomials, and it has resulted to be a useful tool in the computation of the solution applying the asymptotic expansion method. A modificacion of this processor has been employed to compute periodic solutions in perturbed second order differential equations via the Poincaré–Lindstedt method in Navarro (2008a, 2008b). To that end, the idea is to expand both the solution and the modified frequency with respect to the small parameter, allowing to kill secular terms which appear in the recursive scheme. The elimination of secular terms is performed through a manipulation system which works with modified quasi–polynomials, that is, quasi–polynomials containing undetermined constants:

$$u(t) = \sum_{\nu \geq 0} \tau_1^{\sigma_1} \times \cdots \times \tau_Q^{\sigma_Q}\, t^{n_\nu} e^{\alpha_\nu t} \times \left(\lambda_\nu \cos(\omega_\nu t) + \mu_\nu \sin(\omega_\nu t)\right), \tag{1}$$

where $n_\nu \in \mathbb{N}$, α_ν, ω_ν, λ_ν, $\mu_\nu \in \mathbb{R}$, $\sigma_\nu \in \mathbb{Z}$, and $\tau_1,\ldots,\tau_Q$ are real constants with unknown value.

One year later, Navarro (2009) presented a symbolic computation package based on the object–oriented philosophy for manipulating matrices whose elements lie on the set of quasi–polynomials. The kernel of the symbolic processor was developed in C++, defining a class for this new object as well as a set of functions that operate on the data structure: addition, substraction, differentiation and integration with respect to t, substitution of an undetermined coefficient by a series, and many others. The goal of this processor is to provide a tool to solve a perturbed n–order differential equation of the class

$$x^{(n)} + a_{n-1}x^{(n-1)} + \cdots + a_0 x = u(t) + \epsilon f(x, \dot{x}, \ldots, x^{(n-1)}), \tag{2}$$

with initial conditions

$$x(0) = x_{10}, \quad \dot{x}(0) = x_{20}, \quad \ldots, \quad x^{(n-1)}(0) = x_{n0},$$

where ϵ is a small real parameter, $a_0, a_1, \ldots, a_{n-1} \in \mathbb{R}$, $u(t)$ is a quasi–polynomial, and f is such that

$$f(x, \dot{x}, \ldots, x^{(n-1)}) = \sum_{0 \leq \nu_1,\ldots,\nu_n \leq M} f_{\nu_1,\ldots,\nu_n}\, x^{\nu_1} \cdots (x^{(n-1)})^{\nu_n},$$

 ISBN: 978-84-614-6167-7

with $M \in \mathbb{N}$, $\nu_1, \ldots, \nu_n \in \mathbb{N}$ and $f_{\nu_1,\ldots,\nu_n} \in \mathbb{R}$.

As a first step, Navarro and Pérez (2010) have developed a symbolic technique for the computation of the principal matrix of the linear system associated to an homogeneous ordinary differential equation with constant coefficients of the form

$$x^{(n)} + a_{n-1}x^{(n-1)} + \cdots + a_1\dot{x} + a_0 x = 0\,, \tag{3}$$

where $a_0, a_1, \ldots, a_{n-1} \in \mathbb{R}$, and with initial conditions

$$x(0) = x_{10}\,, \quad \dot{x}(0) = x_{20}\,, \quad \ldots\,, \quad x^{(n-1)}(0) = x_{n0}\,,$$

being $x_{10}, \ldots, x_{n0} \in \mathbb{R}$. This method provides a final analytical solution which can be completely computed in a symbolic way.

In this contribution, we detail a symbolic procedure for computing the solution to the non–homogeneus equation

$$x^{(n)} + a_{n-1}x^{(n-1)} + \cdots + a_0 x = u(t)\,, \tag{4}$$

where, as above, $u(t)$ is a quasi–polynomial. In next section, we summarize the scheme proposed in (Navarro and Pérez, 2010) to calculate the solution to (3), which is needed to express the solution to the complete problem (4), and then, to apply a perturbation method to face equation (2).

2 Solution to the homogeneous problem

As mentioned abbove, Navarro and Pérez (2010) have proposed a symbolic method to compute the solution to the equation (3),

$$x^{(n)} + a_{n-1}x^{(n-1)} + \cdots + a_1\dot{x} + a_0 x = 0\,,$$

where $a_0, a_1, \ldots, a_{n-1} \in \mathbb{R}$, and with initial conditions

$$x(0) = x_{10}\,, \quad \dot{x}(0) = x_{20}\,, \quad \ldots\,, \quad x^{(n-1)}(0) = x_{n0}\,,$$

being $x_{10}, \ldots, x_{n0} \in \mathbb{R}$. This method provides a final analytical solution which can be completely computed in a symbolic way. This technique can be useful for academic purposes and it is also a necessary step to treat more involved situations as the perturbed differential equation (2).

 ISBN: 978-84-614-6167-7

2.1 Description of the method

With the aid of the substitutions $x_1 = x\,, x_2 = \dot{x}\,, \ldots\,, x_n = x^{(n-1)}$, equation (3) is transformed into the system of differential equations given by

$$\dot{X}(t) = A\,X(t)\,, \quad X(0) = X_0\,, \tag{5}$$

where A is the companion matrix,

$$A = \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ -a_0 & -a_1 & -a_2 & \cdots & -a_{n-1} \end{pmatrix}, \quad X(t) = \begin{pmatrix} x_1(t) \\ x_2(t) \\ \vdots \\ x_n(t) \end{pmatrix}, \quad X_0 = \begin{pmatrix} x_{10} \\ x_{20} \\ \vdots \\ x_{n0} \end{pmatrix}.$$

To compute the exponential of A, the matrix is splitted into $B + C$, where

$$B = \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 0 \end{pmatrix}, \quad C = \begin{pmatrix} 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ -a_0 & -a_1 & -a_2 & \cdots & -a_{n-1} \end{pmatrix}.$$

Then, we use the approximation

$$e^A \approx (e^{B/m} e^{C/m})^m\,,$$

with $m \in \mathbb{N}$. This approach to obtain e^A is of potential interest when the exponentials of matrices B and C can be efficiently computed, and requires a_{n-1} to be non–equal to zero. In our case,

$$e^{B/m} = \begin{pmatrix} g_0(m) & g_1(m) & g_2(m) & \cdots & g_{n-1}(m) \\ 0 & g_0(m) & g_1(m) & \cdots & g_{n-2}(m) \\ 0 & 0 & g_0(m) & \cdots & g_{n-3}(m) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & g_0(m) \end{pmatrix},$$

and

$$e^{C/m} = I - \frac{1}{a_{n-1}} \left(e^{-a_{n-1}/m} - 1 \right) C = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \omega_0 & \omega_1 & \omega_2 & \cdots & \omega_{n-1} \end{pmatrix},$$

being, for any $n \in \mathbb{N}$,

$$g_n(m) = \frac{1}{n! m^n}\,,$$

 ISBN: 978-84-614-6167-7

and, for any $i = 0, \ldots, n-2$,

$$\omega_i = \frac{a_i}{a_{n-1}}(e^{-a_{n-1}/m} - 1)\,, \quad \omega_{n-1} = e^{-a_{n-1}/m}\,.$$

As it is established in Moler (2003), for a general splitting $A = B + C$, m can be determine from the inequality

$$\|e^A - (e^{B/m}e^{C/m})^m\| \leq \frac{1}{2m}\|[B,C]\|e^{\|B\|+\|C\|}\,. \tag{6}$$

Thus, being A the companion matrix, we get that

$$\|e^A - (e^{B/m}e^{C/m})^m\| \leq \frac{1}{m}e^{1+\|a\|}\|a\|\,,$$

where $a = (a_0, a_1, \ldots, a_{n-1})$. Here,

$$\|x\| = \|x\|_1 = |x_1| + \cdots + |x_n|\,,$$

for any $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$, and $\|A\| = \max_{\|x\|_1=1} \|A\,x\|_1 = \max_\nu (|a_{1\nu}| + \cdots + |a_{n\nu}|)$, for any matrix

$$A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{pmatrix}.$$

2.2 Adaptation to a symbolic formalism

Navarro and Pérez (2008, 2009, 2010) have proposed an adaptation of the method described in section 2.1 in order to compute the matrix e^{At} instead of e^A. If we do so, we obtain the principal matrix of (5), whose elements lie on the set of quasi–polynomials, and the symbolic processor results to be suitable to work with those matrices. The approach is to compute

$$e^{At} \simeq (e^{Bt/m}e^{Ct/m})^m\,, \tag{7}$$

taking into account that

$$e^{Bt/m} = \begin{pmatrix} g_0(t,m) & g_1(t,m) & \cdots & g_{n-1}(t,m) \\ 0 & g_0(t,m) & \cdots & g_{n-2}(t,m) \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & g_0(t,m) \end{pmatrix}, \tag{8}$$

 ISBN: 978-84-614-6167-7

and

$$e^{Ct/m} = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \omega_0(t) & \omega_1(t) & \omega_2(t) & \cdots & \omega_{n-1}(t) \end{pmatrix}, \tag{9}$$

being, for any $n \in \mathbb{N}$,

$$g_n(t,m) = \frac{t^n}{n!m^n}\,, \tag{10}$$

and, for any $i = 0, \ldots, n-2$,

$$\omega_i(t) = \frac{a_i}{a_{n-1}}(e^{-a_{n-1}t/m} - 1)\,, \quad \omega_{n-1}(t) = e^{-a_{n-1}t/m}\,.$$

Thus, e^{At} is a matrix of quasi-polynomials that can be completely computed through the symbolic processor developed by Navarro (2009). The procedure for the computation of the exponential matrix is substantially simplified by using the following equation, which avoids the symbolic multiplication of $e^{Bt/m}$ and $e^{Ct/m}$,

$$e^{Bt/m}e^{Ct/m} = e^{Bt/m} + H(t,m)\,, \tag{11}$$

where

$$H(t,m) = \frac{1}{a_{n-1}} f(t,m) \begin{pmatrix} g_{n-1}(t,m) \\ \vdots \\ g_0(t,m) \end{pmatrix} \times \begin{pmatrix} a_0 & a_1 & \cdots & a_{n-1} \end{pmatrix}, \tag{12}$$

and

$$f(t,m) = e^{-a_{n-1}t/m} - 1\,. \tag{13}$$

This relation can be obtained directly from the multiplication of matrices $e^{Bt/m}$ and $e^{Ct/m}$ as given in equations (8) and (9).

2.3 Symbolic expansion of the exponential matrix

As it has been stated in section 2.2, the exponential matrix e^{At} can be symbolically calculated through equations (7) and (11),

$$e^{At} \simeq (e^{Bt/m}e^{Ct/m})^m = (e^{Bt/m} + H(t,m))^m\,. \tag{14}$$

This expression can be calculated symbolically. Let us express $H(t,m)$ as

$$H(t,m) = \lambda(t,m)J(t,m)\,, \tag{15}$$

 ISBN: 978-84-614-6167-7

with

$$\lambda(t,m) = \frac{1}{a_{n-1}} f(t,m)\,, \tag{16}$$

and

$$J(t,m) = \begin{pmatrix} g_{n-1}(t,m) \\ \vdots \\ g_0(t,m) \end{pmatrix} \begin{pmatrix} a_0 & a_1 & \cdots & a_{n-1} \end{pmatrix}. \tag{17}$$

Thus,

$$(e^{Bt/m} + H(t,m))^m = \sum_{k=0}^{m} \lambda^k(t,m) \left(e^{Bt/m}\,, J(t,m)\right)_k^m\,, \tag{18}$$

being $(A\,,B)_k^m$ the so-called *non-commutative parenthesis*, defined as

$$(A\,,B)_k^m = \sum_{i=1}^{\binom{m}{k}_{\sigma(m,k)}} \prod_{j=1} \left(A^{(1-(-1)^j)/2} B^{(1+(-1)^j)/2}\right)^{c_{ij}}\,,$$

where A and B are two non-commutative $n \times n$ matrices, $m, k, c_{ij} \in \mathbb{Z}^+ \cup \{0\}$,

$$\sigma(m,k) = \begin{cases} 2k+1 & \text{if} \quad 2k \leq m \\ 2(m-k+1) & \text{if} \quad 2k > m \end{cases}\,,$$

and c_{ij} is determined by the following properties.

1. $\sum_{j=1}^{\sigma(m,k)} c_{ij} = m \;\forall\; i = 1, 2, \ldots, \binom{m}{k}$
2. $\sum_{j=1}^{[\sigma(m,k)/2]} c_{i(2j)} = m - k \;\forall\; i = 1, 2, \ldots, \binom{m}{k}$, being $[\sigma(m,k)/2]$ the entire part of $\sigma(m,k)/2$
3. $c_{i2} \neq 0 \;\forall\; i = 1, 2, \ldots, \binom{m}{k}$
4. If $c_{ij} \neq 0$ and $c_{i(j+2)} \neq 0$, then $c_{i(j+1)} \neq 0 \,\forall\, i = 1, 2, \ldots, \binom{m}{k}$ and $j = 1, 2, \ldots, \sigma(m,k) - 2$

In the following, we summarize some expressions which simplify the way in which the matrix $(e^{Bt/m} + H(t,m))^m$ is simbolically computed. First, let us introduce the following matrices

$$S_{n-1}(t,m) = \begin{pmatrix} a_0 & a_1 & \cdots & a_{n-1} \end{pmatrix} F(t,m)\,, \tag{19}$$

 ISBN: 978-84-614-6167-7

and

$$F(t,m)=\begin{pmatrix} t^{n-1} \\ (n-1)t^{n-2}m \\ (n-1)(n-2)t^{n-3}m^2 \\ \vdots \\ (n-1)!t^0m^{n-1} \end{pmatrix}. \tag{20}$$

Lemma 1 *For any $k\in\mathbb{Z}$ such that $k>1$,*

$$(e^{Bt/m})^k=\begin{pmatrix} G_{0,k}(t,m) & G_{1,k}(t,m) & \cdots & G_{n-1,k}(t,m) \\ 0 & G_{0,k}(t,m) & \cdots & G_{n-2,k}(t,m) \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & G_{0,k}(t,m) \end{pmatrix}, \tag{21}$$

with

$$G_{\nu,k}(t,m)=k^\nu g_\nu(t,m)\,,$$

for each $\nu=0,\ldots,n-1$, and where $g_\nu(t,m)$ is given by equation (10).

Lemma 2 *For any $p\in\mathbb{Z}$ such that $p>1$,*

$$(H(t,m))^p=\left(\frac{f(t,m)}{a_{n-1}(n-1)!m^{n-1}}\right)^p(S_{n-1}(t,m))^{p-1}\times \\ \times F(t,m)\begin{pmatrix} a_0 & a_1 & \cdots & a_{n-1}\end{pmatrix}, \tag{22}$$

where $S_{n-1}(t,m)$ and $F(t,m)$ are given by equations (19) and (20).

Lemma 3 *For any $p,k\in\mathbb{Z}$ such that $p,k\geq 1$,*

$$\left(e^{Bt/m}\right)^k(H(t,m))^p=\left(\frac{f(t,m)}{a_{n-1}(n-1)!m^{n-1}}\right)^p(S_{n-1}(t))^{p-1}\times \\ \times\,\Omega\,F(t,m)\begin{pmatrix} a_0 & a_1 & \cdots & a_{n-1}\end{pmatrix},$$

where $S_{n-1}(t,m)$ and $F(t,m)$ are given by equations (19) and (20), and

$$\Omega=\begin{pmatrix} (k+1)^{n-1} & 0 & \cdots & 0 \\ 0 & (k+1)^{n-2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix}.$$

 ISBN: 978-84-614-6167-7

Lemma 4 *For any $p, k \in \mathbb{Z}$ such that $p, k \geq 1$,*

$$H^p(t,m)\left(e^{Bt/m}\right)^k = \left(\frac{f(t,m)}{a_{n-1}(n-1)!m^{n-1}}\right)^p \times (S_{n-1}(t))^{p-1} F(t,m)A(t,m)\,,$$

where $S_{n-1}(t,m)$ and $F(t,m)$ are given by equations (19) and (20),

$$A(t,m) = \begin{pmatrix} A_0(t,m) & A_1(t,m) & \cdots & A_{n-1}(t,m) \end{pmatrix},$$

$$A_\mu(t,m) = \sum_{\nu=0}^{\mu} a_\nu k^{\mu-\nu} g_{\mu-\nu}\,,$$

for any $\mu = 0, \ldots, n-1$, and g_ν is given by equation (10).

For the sake of simplicity, we have omitted the dependence on m and t of g_ν.

3 Solution to the non-homogeneous problem

The general solution to a non–homogeneous linear differential equation of order n can be expressed as the sum of the general solution to the corresponding homogeneous, linear differential equation and any solution to the complete equation. The symbolic manipulation system calculates the solution to a non–perturbed differential equation with initial conditions of the form (4),

$$x^{(n)} + a_{n-1}x^{(n-1)} + \cdots + a_1\dot{x} + a_0 x = u(t)\,,$$

with initial conditions

$$x(0) = x_{10}\,, \quad \ldots\,, \quad x^{(n-1)}(0) = x_{n0}\,,$$

where $a_0, a_1, \ldots, a_{n-1} \in \mathbb{R}$, $x_{10}, \ldots, x_{n0} \in \mathbb{R}$ and

$$u(t) = \sum_{\nu \geq 0} t^{n_\nu} e^{\alpha_\nu t}(\lambda_\nu \cos(\omega_\nu t) + \mu_\nu \sin(\omega_\nu t))\,,$$

being $n_\nu \in \mathbb{N}$, and α_ν, ω_ν, λ_ν, $\mu_\nu \in \mathbb{R}$. With the aid of the substitutions

$$x_1 = x, \quad x_2 = \dot{x}\,, \quad \ldots\,, \quad x_n = x^{(n-1)}\,,$$

equation (4) is reduced to the system of differential equations given by

$$\dot{X}(t) = A\,X(t) + B(t)\,, \quad X(0) = X_0\,, \tag{23}$$

 ISBN: 978-84-614-6167-7

where

$$A = \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ -a_0 & -a_1 & -a_2 & \cdots & -a_{n-1} \end{pmatrix},$$

and

$$B(t) = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ u(t) \end{pmatrix}, \quad X(t) = \begin{pmatrix} x_1(t) \\ x_2(t) \\ \vdots \\ x_n(t) \end{pmatrix}, \quad X_0 = \begin{pmatrix} x_{10} \\ x_{20} \\ \vdots \\ x_{n0} \end{pmatrix}.$$

The computation of the solution to the constant coefficients linear part requires the calculation of the exponential of the matrix A, $\Phi(t) = e^{At}$:

$$X(t) = \Phi(t)X_0 + \Phi(t)\int_0^t \exp(A\tau)\, B(\tau)\, d\tau\,. \tag{24}$$

3.1 Symbolic expansion of the solution

In the following, we give a formula for the symbolic expansion of the solution to the complete orinary differential equation. To that end, let us express the non–commutative parenthesis $\left(e^{Bt/m}, J(t,m)\right)_k^m$ as

$$\left(e^{Bt/m}, J(t,m)\right)_k^m = \begin{pmatrix} p_{r_{11}}^{(m,k)}(t) & \cdots & p_{r_{1n}}^{(m,k)}(t) \\ \vdots & \ddots & \vdots \\ p_{r_{n1}}^{(m,k)}(t) & \cdots & p_{r_{nn}}^{(m,k)}(t) \end{pmatrix} = \left(p_{r_{ij}}^{(m,k)}(t)\right)_{i,j=1}^n = P(m,k)(t)\,,$$

where each $p_{r_{1n}}^{(m,k)}(t)$ is a polynomial of degree $r_{ij} \leq m(n-1)$ in the indeterminate t with coefficients from $\mathbb{R}$, and $m \neq 0$.

Let us also express

$$\lambda^k(t,m) = \left(\frac{1}{a_{n-1}}\left(e^{a_{n-1}t/m} - 1\right)\right)^k = \sum_{\nu=0}^{k} \alpha_\nu e^{\beta_\nu t}\,,$$

being

$$\alpha_\nu = \left(\frac{1}{a_{n-1}}\right)^k \begin{pmatrix} k \\ \nu \end{pmatrix} (-1)^\nu\,,$$

and

$$\beta_\nu = -\frac{a_{n-1}}{m}(k-\nu)\,.$$

 ISBN: 978-84-614-6167-7

Thus, taking into account equation (14), the exponential matrix of A can be arranged as

$$e^{At} \simeq \begin{pmatrix} \sum_{k=0}^{m} \phi_k(t) p_{r_{11}}^{(m,k)}(t) & \cdots & \sum_{k=0}^{m} \phi_k(t) p_{r_{1n}}^{(m,k)}(t) \\ \vdots & \ddots & \vdots \\ \sum_{k=0}^{m} \phi_k(t) p_{r_{n1}}^{(m,k)}(t) & \cdots & \sum_{k=0}^{m} \phi_k(t) p_{r_{nn}}^{(m,k)}(t) \end{pmatrix},$$

where

$$\phi_k(t) = \sum_{\nu=0}^{k} \alpha_\nu e^{\beta_\nu t}\,.$$

In order to develop a symbolic expression of the solution to the complete differential equation, let us express $u(t)$ as

$$u(t) = e^{\delta t}(Q_a(t)\cos(\omega t) + T_b(t)\sin(\omega t))\,,$$

with $a, b \in \mathbb{Z} \cup \{0\}$, $\delta, \omega \in \mathbb{R}$, and being $Q_a(t)$ and $T_b(t)$ real polinomyals of degree a and b respectively, in the indeterminate t.

The product $e^{-A\tau}B(\tau)$ can be arranged as follows:

$$e^{-A\tau}B(\tau) = \begin{pmatrix} \displaystyle\int_0^t \left(\sum_{k=0}^{m} \left(\sum_{\nu=0}^{k} \alpha_\nu e^{\beta_\nu \tau} \right) p_{r_{1n}}^{(m,k)}(-\tau) \right) u(\tau)\, d\tau \\ \vdots \\ \displaystyle\int_0^t \left(\sum_{k=0}^{m} \left(\sum_{\nu=0}^{k} \alpha_\nu e^{\beta_\nu \tau} \right) p_{r_{nn}}^{(m,k)}(-\tau) \right) u(\tau)\, d\tau \end{pmatrix}.$$

Here, the integrand of each element of the matrix above can be written in the form of a quasi–polynomial. Hence, we arrive to the formulae

$$\int_0^t \left(\sum_{k=0}^{m} \left(\sum_{\nu=0}^{k} \alpha_\nu e^{\beta_\nu \tau} \right) p_{r_{jn}}^{(m,k)}(-\tau) \right) u(\tau)\, d\tau = \\ = \sum_{\alpha_{jn}} \left(\eta_{\alpha_{jn}} C_{\alpha_{jn},\beta_{jn},\gamma_{jn}}(t) + \nu_{\alpha_{jn}} S_{\alpha_{jn},\beta_{jn},\gamma_{jn}}(t) \right),$$

where $\alpha_{jn} \in \mathbb{Z}^+ \cup \{0\}$, $\beta_{jn}, \gamma_{jn} \in \mathbb{R}$, $j = 1, \ldots, n$, and

$$C_{m,\beta,\gamma}(t) = \int t^m e^{\beta t}\cos(\gamma t)\, dt\,, \quad S_{m,\beta,\gamma}(t) = \int t^m e^{\beta t}\sin(\gamma t)\, dt\,.$$

These functions are computed recursively through

$$\begin{aligned} C_{0,\beta,\gamma}(t) &= \frac{\gamma}{\beta^2+\gamma^2} e^{\beta}\sin(\gamma t) + \frac{\beta}{\beta^2+\gamma^2} e^{\beta}\cos(\gamma t)\,, \\ S_{0,\beta,\gamma}(t) &= \frac{\beta}{\beta^2+\gamma^2} e^{\beta}\sin(\gamma t) - \frac{\gamma}{\beta^2+\gamma^2} e^{\beta}\cos(\gamma t)\,, \end{aligned}$$

 ISBN: 978-84-614-6167-7

and, for any $m \geq 1$,

$$\begin{aligned} C_{m,\beta,\gamma}(t) &= t^m C_{0,\beta,\gamma}(t) - m\frac{\gamma}{\beta^2+\gamma^2} S_{m-1,\beta,\gamma} - m\frac{\beta}{\beta^2+\gamma^2} C_{m-1,\beta,\gamma}(t)\,, \\ S_{m,\beta,\gamma}(t) &= t^m S_{0,\beta,\gamma}(t) - m\frac{\beta}{\beta^2+\gamma^2} S_{m-1,\beta,\gamma} + m\frac{\gamma}{\beta^2+\gamma^2} C_{m-1,\beta,\gamma}(t)\,. \end{aligned}$$

Therefore, equation (24) can be expressed via a quasi–polynomial and, thus, obtained through the designed symbolic system.

3.2 Description of the program

In order to describe the algebraic processor, let us introduce the following test problem.

$$\ddot{x} + \dot{x} + x = \sin t\,,$$

with initial conditions $x(0) = 0$, $\dot{x}(0) = 1$. This equation describes the (small) angular position in radians $x(t)$ of a forced damped pendulum with a periodic driving force. In this equation, $\ddot{x}(t)$ represents the inertia, $\dot{x}(t)$ represents the friction and $\sin(t)$ represents a sinusoidal driving torque applied at the pivot of the pendulum. The exact solution to this problem is given by

$$x(t) = -\cos t + e^{-t/2}\cos\left(\frac{\sqrt{3}}{2}\,t\right) + \sqrt{3}e^{-t/2}\sin\left(\frac{\sqrt{3}}{2}\,t\right)\,. \tag{25}$$

The program proceeds as follows. The first window allows us introducing the definition of variables, the order of the ODE and its coefficients and the function of the non–homogeneous term $u(t)$. Then, a new window visualizes the expression of the companion matrix related to the ODE. Next, matrices B, $\exp(Bt/m)$, C and $\exp(Ct/m)$, which are used in the calculation of the exponential matrix, are computed by using a value of m satisfying equation (6).

In figure 1, we compare the solution computed via the symbolic processor with the exact solution (25) (left pannel) and a numeric solution computed through a Runge–Kutta 4^{th} order method with step $h = 0.1$ with the exact one (right pannel). From these two figures, it is obvious that the solution symbolically computed adjunts better to the exact solution than the described numerical solution. Nevertheless, it is not the goal of this paper to develop a numerical tool. The symbolic tecnique we have developed provides an analytical solution which can be used as a kernel to apply perturbation methods to computed the solution to a perturbed differential equation depending on a small parameter.

Acknowledgements

This work has been partially supported by grants AYA2010-22039-C02-01 (MICINN) and ACOMP/2011/196 (Generalitat Valenciana).

 ISBN: 978-84-614-6167-7

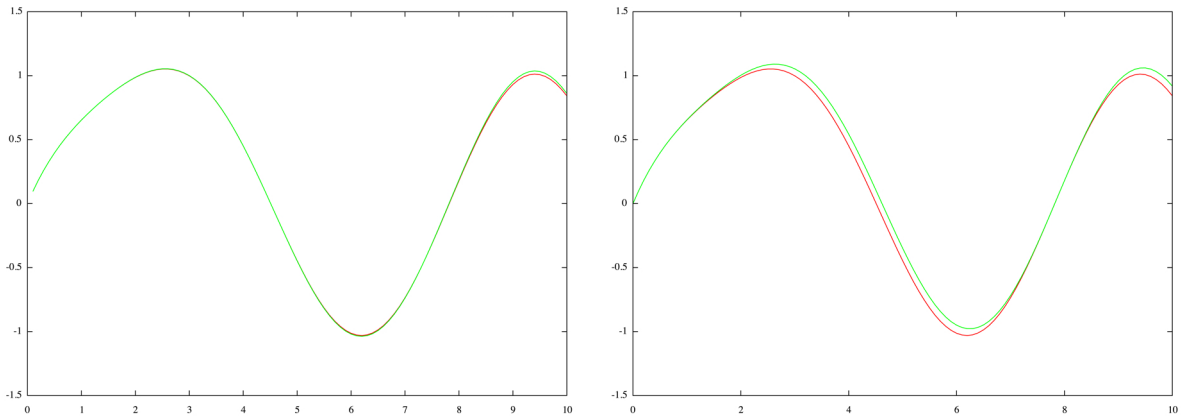


Figure 1: Comparison between numeric, symbolic and exact solution

References

[1] Jacques Henrard, *A survey of Poisson series processors*, Celestial Mech. Dynam. Astron. **45** (1989) 245–253

[2] C. Moler and C. V. Loan, Nineteen dubious ways to compute the exponential of a matrix, twenty-five years later, SIAM Review **45** (2003) 1

[3] Juan F. Navarro, *On the implementation of the Poincaré-Lindstedt technique*, Applied Mathematics and Computation **195** (2008a) 183–189

[4] Juan F. Navarro, *Computation of periodic solutions in perturbed second order ODEs*, Applied Mathematics and Computation **202** (2008b) 171–177

[5] Juan F. Navarro, *On the symbolic computation of the solution to a differential equation*, Advances and Applications in Mathematical Sciences **1** (2009) 1–21

[6] Juan F. Navarro and Antonio Pérez, *Principal matrix of a linear system symbolically computed*, Numerical Analysis and Applied Mathematics, International Conference (2008) 400–402

[7] Juan F. Navarro and Antonio Pérez, *Symbolic computation of the solution to an homogeneous ODE with constant coefficients*, Numerical Analysis and Applied Mathematics, International Conference (2009) 400–402

[8] Juan F. Navarro and Antonio Pérez, *Symbolic computation of the exponential matrix of a linear system*, submitted to *Applied Mathematics and Computation* (2010)

 ISBN: 978-84-614-6167-7

[9] H. Poincaré, *Les Methodes Nouvelles de la Mécanique Celeste I*, Gauthiers-Villars, 1892

 ISBN: 978-84-614-6167-7

Proceedings of the 11th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2011
26–30 June 2011.

A fractal method for numerical integration of experimental signals

M.A. Navascués[1] and M.V. Sebastián[2]

[1] *Departmento 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

Many of the methods of advanced mathematics and statistics are based on the computation of integrals. Most of the standard numerical procedures of integration and the corresponding theorems of error representation are argued on the basis of a high degree of regularity of the function involved. However, it is clear that not all the natural phenomena enjoy this property. One of the goals of this paper is the study of procedures for the quadrature and representation of functions defined through their samples, where the original "signal" is not explicitly known, but it shows experimentally some kind of fractal complexity (for instance through the numerical computation of fractal parameters). The methods proposed in this paper are supported on a fractal interpolation of the function.

Key words: Fractal interpolation functions, numerical integration
MSC 2000: 28A80, 65D05, 41A10

1 Affine fractal interpolation functions

Let $t_0 < t_1 < ... < t_N$ be real numbers, and $I = [t_0, t_N]$ the closed interval that contains them. Let a set of data points $\{(t_n, x_n) \in I \times \mathbb{R} : \ n = 0, 1, 2, ..., N\}$ be given.
Let us consider an Iterated Function System $w_n : I \times \mathbb{R} \to I \times \mathbb{R}$ defined as

$$w_n(t, x) = (L_n(t), F_n(t, x))$$

for $n = 1, 2, \ldots, N$. The maps L_n and F_n are defined by the expressions

$$\begin{cases} L_n(t) = a_n t + b_n \\ F_n(t, x) = \alpha_n x + q_n(t), \end{cases} \tag{1}$$

where

$$a_n = \frac{t_n - t_{n-1}}{t_N - t_0}, \qquad b_n = \frac{t_N t_{n-1} - t_0 t_n}{t_N - t_0} \tag{2}$$

 ISBN: 978-84-614-6167-7

and $q_n(t) = q_{n1}t + q_{n0}$, with

$$q_{n1} = \frac{x_n - x_{n-1}}{t_N - t_0} - \alpha_n \frac{x_N - x_0}{t_N - t_0} \tag{3}$$

$$q_{n0} = \frac{t_N x_{n-1} - t_0 x_n}{t_N - t_0} - \alpha_n \frac{t_N x_0 - t_0 x_N}{t_N - t_0}. \tag{4}$$

The attractor of this system is the graph of a continuous function $f : I \to \mathbb{R}$ such that $f(t_n) = x_n$ for $n = 0, 1, \ldots, N$. The function f is called an Affine Fractal Interpolation Function and it satisfies the fixed point equation ([1]):

$$f(t) = \alpha_n f \circ L_n^{-1}(t) + q_n \circ L_n^{-1}(t), \tag{5}$$

for $t \in I_n = [t_{n-1}, t_n]$.The scalars α_n are termed scaling factors of the system.
These approximants were discussed in the references [3], [4], [5] and [6].

2 Fractal quadrature

In this section we propose a procedure for the numerical quadrature of functions which display some kind of fractal complexity. In specific subsections we prove the convergence of the method for continuous and Hölder-continuous functions (which include the smooth case) and a numerical example.
The method is obtained by means of a fractal interpolation of the signal. Let us denote M_0 the integral of an affine interpolant of the data on the interval I. The moment M_0 can be computed using (5):

$$M_0 = \int_I f(t)\ dt = \sum_{n=1}^{N} \int_{I_n} (\alpha_n f \circ L_n^{-1}(t) + q_n \circ L_n^{-1}(t))\ dt$$

that is to say,

$$M_0 = \sum_{n=1}^{N} \alpha_n \int_{I_n} f \circ L_n^{-1}(t)\ dt + Q_0$$

where

$$Q_0 = \int_I Q(t)\ dt \tag{6}$$

and

$$Q(t) = q_n \circ L_n^{-1}(t) \quad \text{if} \quad t \in I_n. \tag{7}$$

With the change $L_n^{-1}(t) = \widetilde{t}$, bearing in mind (1),

$$M_0 = \sum_{n=1}^{N} \alpha_n a_n M_0 + Q_0$$

and

$$M_0 = \frac{Q_0}{1 - \sum_{n=1}^{N} \alpha_n a_n}.$$

 ISBN: 978-84-614-6167-7

There are several criteria for the election of the scaling factors. They can be consulted in the references [3], [6]. The procedures use a uniform partition $\Delta : t_0 < t_1 < \ldots < t_N$ and subpartitions $\{\bar{t}_j\}_{j=1}^m$ of each interval $I_n = [t_{n-1}, t_n]$.

2.1 Rate of convergence

With the help of previous results published in the references quoted, the next Theorems of interpolation error can be proved.

Theorem 2.1 *If x is a continuous function providing the non-aligned data $\{(t_n, x_n)\}_{n=0}^N$ with a constant step $h = t_n - t_{n-1}$, the integral computed by the method described is convergent to the exact value as the step h tends to zero. Let $g_0(t)$ be the polygonal whose vertices are the data. An upper estimate of the error is given by the following inequality:*

$$\left| \int_I x - \int_I f \right| \leq T\omega_x(h)\left(1 + \frac{2K'_h\|x\|_\infty}{1 - \omega_x(h)K'_h}\right) \tag{8}$$

where $\omega_x(h)$ is the modulus of continuity of $x(t)$, $T = length(I)$, and

$$K'_h = (m-1)^{1/2}K_h^{-1}$$

with

$$K_h = \left(\sum_{j=1}^{m-1} \left(g_0(\frac{(m-j)t_0 + jt_N}{m}) - \frac{(m-j)x_0 + jx_N}{m} \right)^2 \right)^{1/2}. \tag{9}$$

Note: The former result states the convergence of the procedure since, for a continuous function x on a compact interval I, $\omega_x(h) \to 0$ as $h \to 0$ ([2]). The only hypothesis required is the continuity of the original signal.

Theorem 2.2 *If $x \in Lip\ \beta$, the error in the computation of the integral is bounded by the expression:*

$$\left| \int_I x - \int_I f \right| \leq \frac{kT^{\beta+1}}{N^\beta}\left(1 + \frac{2N^\beta(m-1)^{1/2}}{K_hN^\beta - kT^\beta(m-1)^{1/2}}\|x\|_\infty\right)$$

where k is the Lispchitz constant of $x(t)$, T is the length of the interval I, $N+1$ is the number of points of the partition, $(m-1)$ is the number of intermediate points in every subinterval, and K_h is given by the expression (9).

Consequence: The rate of convergence of the procedure is $\mathcal{O}(N^{-\beta})$. It is clear that if β is low the convergence may be slow.

 ISBN: 978-84-614-6167-7

2.2 An application

Let us consider a function of Weierstrass given by the expression

$$x(t) = 3t^2 + 2t + 0.7 - 5\sum_{k=1}^{\infty} \frac{\sin(6^k\ \pi t)}{2^k}$$

in the interval $I = [-1, 1]$. The mapping $x(t)$ is a Hölder-continuous function with exponent $\beta = \ln 2/\ln 6$. The integral M_0 has been computed numerically using the method described, with several steps and $m = 2$. The results are displayed in Table 1. For each step, the value of the integral and the fractal dimension corresponding to the fractal function used are shown. The latter scalar was computed according to the exact formula given in the reference [1] for an affine fractal function, in terms of the scale factors.
The exact value of the integral is 3.4 and the fractal dimension (box-counting) is $D = 2 - \beta = 1.61315$.

N	M_0	$\lvert\bar{\alpha}\rvert_\infty$	D
64	2.80812	0.75710	1.73334
128	3.09601	0.25797	1.57493
256	3.24591	0.57978	1.76810
512	3.32242	0.31603	1.67046
1024	3.36107	0.26477	1.68908
2048	3.38050	0.21263	1.69378
4096	3.39024	0.08399	1.59499

Table 1. Results corresponding to the map $x(t)$.

References

[1] M. F. Barnsley, *Fractals Everywhere*, Academic Press Inc., 1988.

[2] E. W. Cheney, *Approximation Theory*, AMS Chelsea Publ., 1966.

[3] M. A. Navascués and M. V. Sebastián, *Fitting curves by fractal interpolation: An application to the quantification of cognitive brain processes*, in: Thinking in Patterns: Fractals and Related Phenomena in Nature World Sci. (2004) 143–154.

[4] M. A. Navascués and M. V. Sebastián, *Error bounds for affine fractal interpolation functions*, Mathematical Inequalities Applications **9(2)** (2006) 273–288.

[5] M. A. Navascués and M. V. Sebastián, *Spectral and affine fractal methods in signal processing*, Int. Math. Forum **1(29-32)** (2006) 1405–1422.

[6] M. A. Navascués and M. V. Sebastián, *Construction of affine fractal functions close to classical interpolants*, J. of Comp. Anal. Appl. **9(3)** (2007) 271–283.

 ISBN: 978-84-614-6167-7

Proceedings of the 11th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2011
26–30 June 2011.

Exploiting the regularity of differential operators to accelerate solutions of PDEs on GPUs

G. Ortega[1], E. M. Garzón[1], F. Vázquez[1] and I. García[2]

[1] *Dpt. Computer Architecture and Electronics, Cra Sacramento s/n Almeria 04120 Spain, Almeria University*

[2] *Dpt. Computer Architecture, Málaga 29071 Spain, Málaga University*

emails: glortega@ual.es, gmartin@ual.es, f.vazquez@ual.es, igarcia@ac.uma.es

Abstract

Partial Differential Equations (PDEs) are involved in the study of a wide range of physical problems. The discretization of the differential operators in the space generates the solution of the linear systems of equations. These linear systems of equations are characterized by the regularity of the sparse matrices involved. Furthermore, the sparse matrix vector product (SpMV) is the operation with higher computational cost in the resolution of a PDE. So, in this work we have focusing on: (1) the implementation of a SpMV GPU kernel using a regular format, and (2) the GPU implementation of the Bi-Conjugate Gradient method which is based on the aforementioned SpMV kernel. Thus, in both implementations the regular pattern of the matrices involved are considered. The central question to be examined in this paper is the exploiting of these regularities in order to achieve a high performance and to reduce the memory requirements. Therefore, by means of our approach is possible to extend the dimension of the linear equation system to solve and can deal with matrices of larger dimension.

Key words: GPU computing, parallel computing, linear system of equations, SpMV, BiConjugate gradient method

1 Introduction

The solution of PDEs is involved in several scientific and engineering applications. In general, its high computational cost requires the use of High Performance Computing techniques (HPC). On the other hand, the discretization of the differential operators generates the solution of the linear systems of equations which are characterized by the regularity of

 ISBN: 978-84-614-6167-7

matrices involved. In this paper, the main goal is the simultaneous exploitation of both, the high regularity of the matrices and the massive parallelism supplied by the GPU platforms.

Our interest is focused on the solution of the Helmholtz differential equation since this equation is considered in a wide variety of applications from different areas of Physics. To be more specific, this equation arises in the study of physical problems involving Partial Differential Equations (PDEs) in both space and time. Helmholtz equation is linear and is characterized by the regularity of its matrices. Moreover, this kind of linear system equations can be solved by using iterative methods such as the BiConjugate Gradient Method (BCG) since it is suitable for solving complex and nonsymmetric linear systems. Nevertheless, from a computational point of view, BCG is very expensive due to the sparse matrix vector products (SpMV) included in the algorithm. Thus, in order to accelerate the BCG method, the exploitation of HPC is necessary. The GPU computing has emerged as a new HPC technique that offers massive parallelism and can be useful for accelerating this kind of algorithms.

CUDA (Computing Unified Device Architecture) is the Application Programming Interface developed by NVIDIA to facilitate the programming of GPUs. Using the CUDA interface, the GPU is considered by the programmer as a set of SIMT (Single Instruction, Multiple Threads) multiprocessors [9]. Each kernel (parallel code) is executed as a batch of threads organized as a grid of thread blocks whose configuration is defined by the programmer setting up specific parameters. One of these parameters is the threads block size. The blocks in turn are divided into sets of threads called warps. Currently, the SMs are composed by thirty-two SPs on the most extended Fermi architectures [9, 12].

An approach to facilitate the GPU programming is based on the use of basic routines or libraries which: (1) compute the most used operations in the applications and (2) are optimally accelerated by GPU. In this line, NVIDIA supplies a wide set of routines related to matrix computations such as CuBLAS [10].

This work shows that BCG method can be efficiently accelerated if the SpMV operations are computed on the GPU taking advantage of the regularity of the matrices related to the PDE solution. So, the implementation onto the GPU of SpMV using a regular format and a BCG approach based on the same format are developed and evaluated for a set of test complex matrices in single precision. Experimental results have shown that the implementation on GPU of both, SpMV and BCG using regular format for the storage of matrices, outperforming the implementation on CPU.

So, our interest is centered on accelerating SpMV to improve the performance of BCG method and the remaining vector operations. Our approach to accelerate this algorithm is based on GPU computation, since GPU platform can be used as accelerator of the SpMV [1, 11, 15].

Section 2 briefly reviews the regularity of the matrix form to solve PDEs. Section 3 is devoted to introduce the specific Regular Format considered the efficient computation of

 ISBN: 978-84-614-6167-7

the SpMV operation, focusing our interest on the solution of Helmholtz equation on the GPU architecture. Next, Section 4 evaluates the SpMV and the BCG method, both based on the Regular Format on the NVIDIA GPU GTX 480 with a set of representative test matrices. The results clearly show that the better performance for both, SpMV and BCG method using Regular Format for all the test matrices. Finally, Section 5 summarizes the main conclusions.

2 Solving PDEs. Analysis of the matrix form

The numerical solution of PDEs is based on the discretization of differential operators. Consequently, linear system equations in $\mathbb{R}$ or $\mathbb{C}$ (depending on the kind of application) have to been solved. In general, these linear systems are defined by a sparse matrix which exhibits a strong regularity in both, the pattern and the values of its nonzero elements.

This kind of linear system equations can be solved by using iterative methods such as the Conjugate Gradient (for real and positive definite systems) or the BiConjugate Gradient Method (for complex or real nonsymmetric linear systems).

Our interest is centered on the BiConjugate Gradient Method (BCG) method because it has a wide range of applications such as Electromagnetism [5, 14].

The BCG (proposed by Lanczos [6, 7] and discussed by Fletcher [3] and Jacobs [4]) is a nonstationary iterative method to solve systems of linear equations $Ax = b$ where the matrix $A \in \mathbb{C}^{N \times N}$ can be non-symmetric. For the given system of equations, A denotes the coefficient (sparse) matrix, b indicates the independent term and x is the vector solution.

The pseudocode for the BCG Method is given in Algorithm 1. Additionally, the complexity order for the most expensive operations are shown in Algorithm 1. Let us remark that nz denotes the number of nonzero elements of A and N the number of rows. The computational cost associated to the SpMV operation is larger that the remaining operations which are just inner products. It is noteworthy that every iteration involves the computation of SpMV operations using A and A^T (lines 9 and 13 of Algorithm 1). Notice that, in general, the SpMV operation over A^T represents a penalty on the performance of the BCG method.

3 SpMV based on Regular Format

In general, Compressed Row Storage (CRS) is the most extended format to store sparse matrices. Let N and Nz be the number of rows of the matrix and the total number of non-zero entries of the matrix, respectively; the data structure consists of the following arrays: (1) $A[]$ array of floats of dimension Nz, which stores the entries; (2) $J[]$ array of integers of dimension Nz, which stores their column index; and (3) $start[]$ array of integers of dimension $N + 1$, which stores the pointers to the beginning of every row in $A[]$ and $J[]$,

 ISBN: 978-84-614-6167-7

Algorithm 1 BiConjugate Gradient Method

Require: Define $EPS = Accuracy\ Threshold$

Ensure: The value of $x^{(i)}$.

1:	Compute $r^{(0)} = b - Ax^{(0)}$ for some initial guess $x^{(0)}$	
2:	Choose $r'^{(0)} = r^{(0)}$; $p'^{(0)} = 0$; $p^{(0)} = p'^{(0)}$; $\rho'^{(0)} = 1$	
3:	Calculate $\Delta^{(0)} = norm2(r^{(0)})$	$O(4N)$
4:	**for** $i = 1, 2, ...$ **do**	
5:	$\rho^{(i)} = (r'^{(i-1)}, r^{(i-1)})$	$O(8N)$
6:	$\beta^{(i)} = \rho^{(i)} / \rho'^{(i-1)}$	
7:	$p^{(i)} = r^{(i-1)} + \beta^{(i)} p^{(i-1)}$	$O(8N)$
8:	$p'^{(i)} = r'^{(i-1)} + \beta^{(i)} p'^{(i-1)}$	$O(8N)$
9:	$v^{(i)} = Ap^{(i)}$	$O(8nz)$
10:	$\alpha^{(i)} = \rho^{(i)} / (p'^{(i)}, v^{(i)})$	$O(8N)$
11:	$x^{(i)} = x^{(i-1)} + p^{(i)} \alpha^{(i)}$	$O(8N)$
12:	$r^{(i)} = r^{(i-1)} - v^{(i)} \alpha^{(i)}$	$O(8N)$
13:	$r'^{(i)} = r'^{(i-1)} - \alpha^{(i)} (A^T p'^{(i)})$	$O(8nz + 8N)$
14:	$\Delta^{(i)} = norm2(r^{(i)})$	$O(4N)$
15:	**if** $\Delta^{(i)} < \Delta^{(0)} EPS$ **then**	
16:	**return** $x^{(i)}$	
17:	**else**	
18:	$\rho'^{(i)} = \rho^{(i)}$	
19:	**end if**	
20:	**end for**	

both sorted by row index [2]. Moreover, other proposals to store the sparse matrices have been developed on the literature [1, 15].

However, this work aims at presenting and evaluating an approach to increase the performance of the SpMV operation. Bearing in mind that the matrix form of the differential operator involved in PDEs has several regularities, our proposal is based on a specific storage format for these regular sparse matrices. Hereinafter this kind of format is referred to as Regular Format.

The regularity of the matrix allow us to compact it better than other formats to store matrices. Thus, the goal of our Regular Format is to store the minimal information to define the sparse matrix. So, the computation time of SpMV can be reduced since the number of memory accesses to read the elements of the sparse matrix has a strong impact in its performance.

In order to illustrate our approach we deal with the Helmholtz equation

$$(\nabla^2(\mathbf{r}) + k(\mathbf{r})^2) E(\mathbf{r}) = 0 \tag{1}$$

 ISBN: 978-84-614-6167-7

where ∇^2 is the Laplace operator; E is a complex scalar function (potential) defined at a spatial point $r = (x, y, z) \in \mathbb{R}^3$ and k is some real or complex constant. This equation naturally appears from general conservation laws of physics and can be interpreted as a wave equation for monochromatic waves (wave equation in the frequency domain). Equation 1 can be numerically solved by means of the appropriated transformation based on Green's functions and the spatial discretization [8, 13].

As consequence, a complex linear system, characterized by the regularity of its matrix, must be solved by means of an iterative method suitable for solving complex linear systems, such as the BCG method.

In this example, the regularities of the matrix are: (1) The matrix is symmetric; (2) There is a maximum of seven nonzeros elements at every row; (3) The nonzeros values are located by seven diagonals in the matrix, where one is the main diagonal, two of them are the first lower and upper diagonals and four of them are located by $\pm d1$-th and $\pm d2$-th diagonals; (4) The nonzeros of every lateral diagonal are the same values (a, b, c); (5) The main diagonal is defined by a set of complex numbers.

Bearing in mind these characteristics (see Figure 1), the Regular Format consists of:

1. One array, $A[]$ (complex) of dimension N (where N represents the dimension matrix).
2. Three integer values (a, b, c) are included with the purpose of storing the value of lateral diagonals.
3. Two additional integer values $d1$, $d2$ in order to point at the location of the lateral diagonals in the first row.

It is relevant to underline the strong reduction of memory requirements of the Regular Format with respect to the CRS format.

Algorithm 2 illustrates the code of SpMV based on the Regular Format. The operations of Algorithm 2 can be highly accelerated by means of the GPU computing. It is due to the regularity of this kind of SpMV which allows to exploit the massive parallelism of the GPU platforms. A specific kernel to compute the SpMV operation on the GPU has been developed based on the Regular Format. In this kernel, every thread computes one element of the output vector ($u = Av$). So, the loop to compute every row has been unrolled. In this way, the parallelism of this specific SpMV kernel is very high.

In the next section, the kernel to compute the SpMV operation by means of the Regular Format is evaluated. Also, the BCG method based on the same Regular Format is carried out.

4 Evaluation

Our analysis is based on the run-times measured on a GeForce GTX 480 using a set of sparse complex matrices which define partial differential equations (PDEs). These matrices

 ISBN: 978-84-614-6167-7

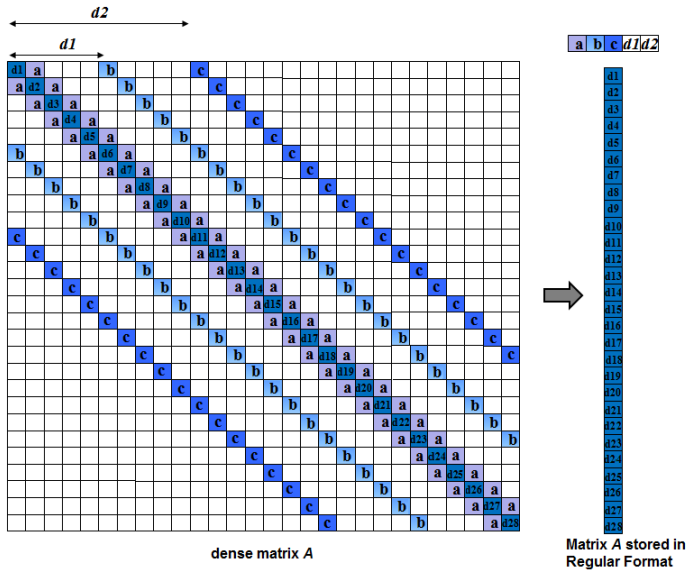


Figure 1: Sparse matrix related to the equation of Helmholtz and the Regular Format to storage this matrix taking advantage of its regularity.

Algorithm 2 Code for SpMV using Regular Format

Require: A, v, N, $d1$ and $d2$.
Ensure: The value of u.

1: **for** $i = 1, 2, ..., N$ **do**
2: $u^{(i)} = v^{(i-1)} + A^{(i)}v^{(i)} + v^{(i+1)}$
3: **if** $((i + d1) <= N)$ **then**
4: $u^{(i)} += v^{(i+d1)}$
5: **end if**
6: **if** $((i - d1) > 0)$ **then**
7: $u^{(i)} += v^{(i-d1)}$
8: **end if**
9: **if** $((i + d2) <= N)$ **then**
10: $u^{(i)} += v^{(i+d2)}$
11: **end if**
12: **if** $((i - d2) > 0)$ **then**
13: $u^{(i)} += v^{(i-d2)}$
14: **end if**
15: **end for**

 ISBN: 978-84-614-6167-7

have a very regular structure (see Figure 1) with small imbalances in the distribution of nonzeros per matrix row. Table 1 illustrates the set of test matrices used in this work and the characteristic parameters related to their specific pattern: number of rows (N), the total number of non-zero entries of the matrix (nz) and the pointers to the lateral diagonals ($d1$ and $d2$). Let us remark that the dimension for all matrices is N x N and the elements of all matrices are stored as two float numbers.

Table 1: Characteristics of test matrices.

Complex matrix	N	nz	$d1$	$d2$
NN70	729000	5086618	90	8100
NN90	1331000	9292578	110	12100
NN110	2197000	15344938	130	16900
NN130	3375000	23579698	150	22500
NN150	4913000	34332858	170	28900
NN170	6859000	47940418	190	36100

In our analysis, two implementations on GPUs are evaluated: (1) the SpMV kernel based on the Regular Format; and (2) the BCG kernel using Regular Format. The BCG implementation on GPU is based on the acceleration of SpMV and the inner products at every iteration of the method to improve its performance. So, the BCG kernel is based on the SpMV Regular Format kernel. Moreover, the inner products have been accelerated using CUBLAS Library [10].

In order to estimate the net gain provided by GPUs in the two implementations, we have taken the SpMV implementation based on Regular Format for modern processors and for GPUs. For the former, we have considered the sequential code for a computer based on a state-of-the-art superscalar core, Intel Xeon Westmere, and evaluated the computing times for the set of test matrices.

Table 2 shows the runtimes for computing the SpMV Regular Format on the GPU and CPU platforms considered in our evaluation. Moreover, Figure 2 lists the acceleration factor of the SpMV using Regular Format on the GPU GeForce GTX 480 against one core of Intel Xeon Westmere. Each test consists of 1000 executions of SpMV routine to obtain accurate timing. The languages used to design the codes are C and CUDA.

Table 2: Runtimes (seconds) for 1000 iterations of the SpMV operation using Regular Format.

	NN70	NN90	NN110	NN130	NN150	NN170
GPU	0,20	0,35	0,61	0,99	1,61	2,27
CPU	32,88	58,62	96,74	148,66	216,05	302,20

 ISBN: 978-84-614-6167-7

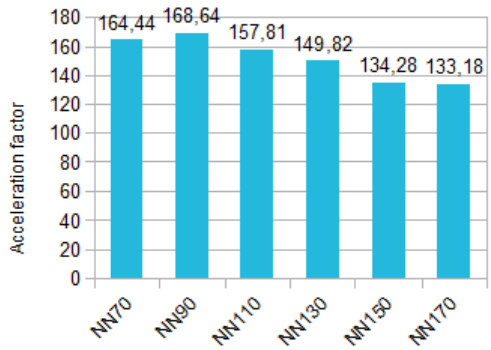


Figure 2: Acceleration factor of the SpMV based on Regular Format on GPU GTX 480 against one core of Intel Xeon Westmere.

Results in Table 2 and Figure 2 show the improvement of the runtimes and the acceleration factor for all test matrices reached by Regular Format format are very high since this kernel takes advantage of the regularities of the matrices and exhibits a high parallelism which can be exploit by the GPU platforms.

Table 3: Runtimes (seconds) for 1000 iterations of one resolution of the BCG method using Regular Format.

	NN70	NN90	NN110	NN130	NN150	NN170
GPU	2,47	3,84	5,90	8,71	12,59	17,49
CPU	152,80	275,99	454,78	698,02	1017,93	1422,52

Results in Table 3 and Figure 3 show the runtimes and the acceleration factor reached by the BCG method for solving complex linear systems using Regular Format. These results illustrate the high acceleration factor achieved by our implementation of the BCG method on the GPU against the version on one core, for a set of test matrices. Notice that the performance for the test matrices increases with the dimension of the problem to solve.

For the considered test matrices, the SpMV runtimes range from $0,20s$ to $2,27s$ on the GPU and from $32,88s$ to $302,20s$ on the CPU. For the BCG kernel based on the same format, the runtimes are between $2,47s$ and $17,49s$ on the GPU, and between $152,80s$ and $1422,52s$ on the CPU.

For the same set of matrices, the acceleration factor achieves values between $133\times$ and $168\times$ for the SpMV kernel based on Regular Format and between $61\times$ and $81\times$ for the BCG kernel based on the same format. So, our approach clearly improves the performance of solvers of Helmholtz equation.

 ISBN: 978-84-614-6167-7

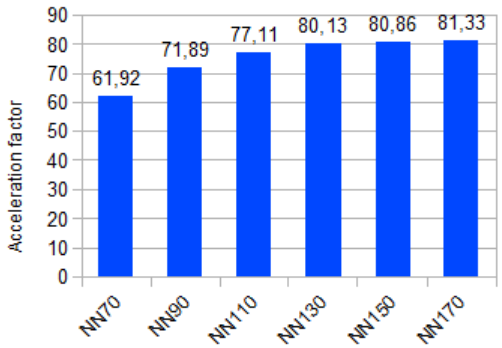


Figure 3: Acceleration factor of the BCG method using Regular Format on GPU GeForce GTX 480 against one core of Intel Xeon Westmere.

5 Conclussions

We have developed a specific implementation on GPUs of the BCG method for solving complex systems of linear equations related to the solution of the Helmholtz equation. The key of our approach is the exploitation of the regularity of the matrices related to the solver. Due to the fact that the SpMV involved in the BCG are the operations with higher cost computational a specific kernel based on the Regular Format to accelerate it has been developed. This kernel allow us to reduce the memory requirements and the runtimes relevantly. The results related to the test matrices with higher dimension, show that the GPU implementation can reduce the runtime from $1422,52s$ in the CPU (one core Intel Xeon Westmere) to $17,49s$ in the GPU (GTX 480). Therefore, our GPU implementation of the BCG method based on Regular Format allow us to extend the dimension of the linear equation system related to the solution of a wide range of applications.

Acknowledgements

This work has been funded by grants from the Spanish Ministry of Science and Innovation TIN2008-01117 and Junta de Andalucia (P08-TIC-3518, P10-TIC-6002), in part financed by the European Regional Development Fund (ERDF). Moreover, it has been developed in the framework of the network High Performance Computing on Heterogeneous Parallel Architectures (CAPAP-H3), supported by the Spanish Ministry of Science and Innovation (TIN2010-12011-E).

References

[1] Nathan Bell and Michael Garland. Implementing sparse matrix-vector multiplication on throughput-oriented processors. In *Proceedings of the Conference on High Performance Computing Networking, Storage and Analysis*, SC '09, pages 18:1–18:11, New York, NY, USA, 2009. ACM.

[2] R. H. Bisseling. *Parallel Scientific Computation.* Oxford University Press, 2004.

[3] R. Fletcher. Conjugate gradient methods for indefinite systems. volume 506, pages 73–89. Springer Berlin / Heidelberg, 1976.

[4] D. A. H. Jacobs. The exploitation of sparsity by iterative methods in sparse matrices and their uses. *I. S. Duff.*, pages 191–222, 1981.

[5] Jeffrey P. Thomas Kenneth C. Hall, Robert E. Kielb, editor. *Unsteady Aerodynamics, Aeroacoustics and Aeroelasticity of Turbomachines.* Springer, 2006.

[6] C. Lanczos. An iteration method for the solution of the eigenvalue problem of linear differential and integral operators. *J . Res. Nat. Bur. Stand.*, 45:255–282, 1950.

[7] Cornelius Lanczos. Solution of systems of linear equations by minimized iterations. *J. Res. Natl. Bur. Stand*, 49:33–53, 1952.

[8] Ramani Duraiswami Nail A. Gumerov. *Fast Multipole Methods for the Helmholtz Equation in Three Dimensions.* Elsevier Science, 2004.

[9] NVIDIA. Cuda programming guide. version 2.3. 2009.

[10] NVIDIA. Cublas library. Technical report, August 2010.

[11] NVIDIA. Cuda cusparse library. Technical report, September 2010.

[12] NVIDIA. Next generation CUDA architecture. Fermi Architecture. 2010.

[13] Matthew N.O. Sadiku. *Numerical Techniques in Electromagnetics. Second Edition.* CRC Press, Boca Raton, 2001.

[14] C. F. Smith, A. F. Peterson, and R. Mittra. The biconjugate gradient method for electromagnetic scattering. *IEEE Transactions on Antennas and Propagation*, 38:938–940, 1990.

[15] F Vázquez, J J Fernández, and E M Garzón. A new approach for sparse matrix vector product on NVIDIA GPUs. *Concurrency and Computation: Practice And Experience.*, 2010.

 ISBN: 978-84-614-6167-7

Proceedings of the 11th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2011
26–30 June 2011.

Introducing priorities in Rfuzzy: Syntax And Semantics

Víctor Pablos-Ceruelo[1] **and Susana Munoz-Hernandez**[1]

[1] *Facultad de Informática, Universidad Politécnica de Madrid (Spain)*

emails: vpablos@fi.upm.es, susana@fi.upm.es

Abstract

We present new syntax and semantics for Rfuzzy, a framework for fuzzy logic where priorities between the rules encoded by the developer are set and taken into account to choose always the one more close to the human way of thinking.

In previous works we presented three symbols, an order between them (▲ $<_a$ ♦ $<_a$ ▼) and an operator (∘) to combine them for this purpose. For working with small programs this is perfectly adequate, but when dealing with larger programs the intentions get lost due to the assignation of identical priority weights to clauses that depend on a small amount of default information and clauses that depend on a large amount of default information. Our goal with this contribution is to differentiate between them by using priorities so the inference process gets even closer to the human way of reasoning and solving problems.

Key words: fuzzy logic framework syntax semantics

1 Introduction

It was Lotfi Zadeh in 1965 who introduced fuzzy set theory [24], and its existence was justified in his paper "Is there a need for fuzzy logic?" [26]. We just outline here some ideas to make the contribution as self contained as possible.

It is usual when modeling real-world problems the necessity to represent not only if an individual belongs or not to a set, but the grade in which it belongs. This grade is what Zadeh tried to model by using a linguistic variable, a variable which can be assigned real-world adjectives[1] as values. For example, age takes the values young and old and temperature takes the values cold, warm and hot (Fig. 1.1).

[1] Please take into account that values for a linguistic variable are not always adjectives.

 ISBN: 978-84-614-6167-7

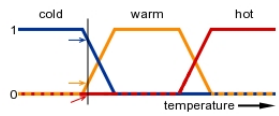


Figure 1.1: Temperature is a linguistic variable and (here) takes the values cold, warm and hot.

But this linguistic variables are no more than part of the fuzzy systems which, aimed at encoding in a computation the human way of solving problems, abstract the real-world facts into fuzzy facts by means of the fuzzification process, infer fuzzy solutions to the real-world problems by using fuzzy rules and defuzzify them to work in real-life scenarios.

To illustrate this description, suppose that temperature is 25 degrees and this measure is fuzzified into temperature(warm) by using the description for the linguistic variable temperature in Fig. 1.1. This fuzzy fact is then taken into account by the fuzzy rule if temperature(warm) then fan_speed(normal) and it is concluded fan_speed(normal). At last the conclusion is defuzzified by a linguistic variable description similar to the one in Fig. 1.1, obtaining the real-world decision to feed the fan with 5 volts[2].

In [25] Zadeh called (type-1) fuzzy sets and systems those using linguistic variables and applying "crisp rules" to infer the belonging value of another linguistic variable. He did that to distinguish them from type-2 fuzzy sets and systems, where rules can be not only satisfied or not but satisfied up to some degree.

A good example to illustrate this kind of rules is trying to determine if a train will stop because of its speed decrease and we should take our baggage, but without knowing if the speed reduction is high, normal or low[3]. We could say that speed_reduction is normal, but we are not completely sure about it and this should be taken into account. For that purpose we measure how much we trust our perception by means of a real number in the interval $[0,\ 1]$[4], 0.6 here. This number is called the credibility of the rule. The fuzzification process result ("speed_reduction(normal) with cred 0.6") is then taken into account by the fuzzy rule "if train_speed_reduction(normal) then train_stops(soon)" to infer the conclusion "train_stops(soon) with cred 0.6", which is still not the solution expected by the deffuzification process. We need to apply another rule to this result[5], "if train_stops(soon) then take_your_baggage(now)" and the result here is "take_your_baggage(now) with cred 0.6". Finally the defuzzification process tries to deffuzify the last conclusion, but it is not strong enough to fire a real-world action. So, it does not suggest us to take our baggage.

The inclusion of the credibility value increased the number of real-world problems that could

[2] suppose that the linguistic variable fan_speed has the following description: when its value is fast we feed the fan with 10 volts, when normal with 5 volts, when low with 2.5 volts and when stop with 0 volts.

[3] Suppose we are just passengers that feel something but can not measure it.

[4] As usually, 0 means we do not trust the rule and 1 we trust it completely. We could use here a linguistic variable again, but we think a number between 0 and 1 makes it easy to understand.

[5] This second inference step is included to highlight that we can model problems much more complex than the previous one, solved in only three inference steps.

 ISBN: 978-84-614-6167-7

be modeled and multiple fuzzy systems have been developed since then. The ones allowing the developer to code programs under the logic programming paradigm (called fuzzy logic systems) we know about are Flopper [19], Fuzzy Prolog [7], Rfuzzy [21] and FuzzyDL [2].

The basic difference between other paradigms and the logic programming paradigm is that it regards a computation as automated reasoning over a corpus of knowledge instead of actions that change the machine state in some way. Facts about the problem domain are expressed as logic formulas, and programs are executed by applying inference rules over them until an answer to the problem is found, or the collection of formulas is proved inconsistent. There are many reasons for using it, but maybe the most important ones are (1) the existing similarities with the human way of thinking and (2) the fact that since the invention of Prolog [11, 3] its use for modeling problems in Artificial Intelligence has not stopped growing up.

The number of problems we can represent by using fuzzy logic is huge but there still some that can not be simulated by just using it, as the one we present and try to overcome here: the existence of rule priorities that overwrite the normal ordering of results obtained from fuzzy logic inferences. Just suppose we want to have default values for some rules. This can not be coded in the current multi-adjoint framework because all rules have the same priority there. We try here to overcome this limitation.

There are many proposals on how to introduce priorities in logic programming (LP) [23, 1, 12, 13, 8, 6] but, as far as we know, there is no existing work on fuzzy logic programming, although it seems to be rather necessary its inclusion.

2 Syntax

We will use a signature Σ of function symbols and a set of variables V to "build" the *term universe* $\mathrm{TU}_{\Sigma,V}$ (whose elements are the *terms*). It is the minimal set such that each variable is a term and terms are closed under Σ-operations. In particular, constant symbols are terms. Similarly, we use a signature Π of predicate symbols to define the *term base* $\mathrm{TB}_{\Pi,\Sigma,V}$ (whose elements are called *atoms*). Atoms are predicates whose arguments are elements of $\mathrm{TU}_{\Sigma,V}$. Atoms and terms are called *ground* if they do not contain variables. As usual, the *Herbrand universe* $\mathbb{HU}$ is the set of all ground terms, and the *Herbrand base* $\mathbb{HB}$ is the set of all atoms with arguments from the Herbrand universe. A substitution σ or ξ is (as usual) a mapping from variables from V to terms from $\mathrm{TU}_{\Sigma,V}$[6].

To capture different interdependencies between predicates, we will make use of a signature Ω of *many-valued connectives*[7] formed by *conjunctions* $\&_1, \&_2, ..., \&_k$, *disjunctions* $\vee_1, \vee_2, ..., \vee_l$, *implications* $\leftarrow_1, \leftarrow_2, ..., \leftarrow_m$, *aggregations* $@_1, @_2, ..., @_n$ and tuples of real numbers in the interval $[0,\ 1]$ represented by $(\mathsf{p},\ \mathsf{v})$.

[6] Although we prefer using suffix notation ($(Term)\sigma$), note that it is equivalent to prefix notation ($\sigma(Term)$).

[7] In some works the term "aggregation operator" subsumes conjunctions, disjunctions and aggregations. In this work we distinguish between them and include a new one (implications).

 ISBN: 978-84-614-6167-7

While Ω denotes the set of connective symbols, $\hat{\Omega}$ denotes the set of their respective associated truth functions. Instances of connective symbols and truth functions are denoted by $\&_i$ and $\hat{\&}_i$ for conjunctors, $\vee_i$ and $\hat{\vee}_i$ for disjunctors, $\leftarrow_i$ and $\hat{\leftarrow}_i$ for implicators, $@_i$ and $\hat{@}_i$ for aggregators and $(\mathsf{p},\ \mathsf{v})$ and $(\hat{\mathsf{p}},\ \mathsf{v})$ for the tuples.

Truth functions for the connectives are then defined as $\hat{\&} : [0,1]^2 \to [0,1]$ monotone[8]and non-decreasing in both coordinates, $\hat{\vee} : [0,1]^2 \to [0,1]$ monotone in both coordinates, $\hat{\leftarrow} : [0,1]^2 \to [0,1]$ non-increasing in the first and non-decreasing in the second coordinate, $\hat{@} : [0,1]^n \to [0,1]$ as a function that verifies $\hat{@}(0,\ \ldots,\ 0) = 0$ and $\hat{@}(1,\ \ldots,\ 1) = 1$ and $(\mathsf{p},\ \mathsf{v}) \in \Omega^{(0)}$ are functions of arity 0 (constants) that coincide with the connectives.

Immediate examples for connectives that come to mind for conjunctors are: in Łukasiewicz logic ($\hat{F}(x,y) = max(0, x+y-1)$), in Gödel logic ($\hat{F}(x,y) = min(x,y)$), in product logic ($\hat{F}(x,y) = x \cdot y$), for disjunctors: in Łukasiewicz logic ($\hat{F}(x,y) = min(1, x+y)$), in Gödel logic ($\hat{F}(x,y) = max(x,y)$), in product logic ($\hat{F}(x,y) = x \cdot y$), for implicators: in Łukasiewicz logic ($\hat{F}(x,y) = min(1, 1-x+y)$), in Gödel logic ($\hat{F}(x,y) = y$ if $x > y$ else 1), in product logic ($\hat{F}(x,y) = x \cdot y$) and for aggregation operators[9]: arithmetic mean, weighted sum or a monotone function learned from data.

3 Semantics

The main idea behind our semantics is that if a rule has more priority than the other ones then the intended truth value for an inference where this rule is involved is the one it obtains.

For this purpose we attach to the usual truth value $\mathsf{v} \in [0,1]$ a real number $\mathsf{p} \in [0,1]$ denoting the (accumulated) priority, resulting in the tuple of real numbers between 0 and 1 symbolized by $(\mathsf{p},\ \mathsf{v}) \in \Omega^{(0)}$. As it can be noted from the symbols used, the first element indicates the priority and second one the "old" truth value. We represent the tuple by (p, v), although in some cases we use (pv) to highlight that the variable is only one and it can take the value $\perp$. The union between the set containing all possible combinations of two real numbers between 0 and 1 and $\{\perp\}$ is symbolized by $\mathbb{KT}$ and we define the ordering between elements from $\mathbb{KT}$ as follows:

Definition 3.1 ($\preccurlyeq_{\mathbb{KT}}$)**.**

$$\begin{aligned} &\perp \preccurlyeq_{\mathbb{KT}} \perp \\ &\perp \preccurlyeq_{\mathbb{KT}} (\mathsf{p},\ \mathsf{v}) \\ &(\mathsf{p}_1,\ \mathsf{v}_1) \preccurlyeq_{\mathbb{KT}} (\mathsf{p}_2,\ \mathsf{v}_2) \quad \leftrightarrow \quad (\ \mathsf{p}_1 < \mathsf{p}_2\ ) \quad or \quad (\ \mathsf{p}_1 = \mathsf{p}_2 \quad and \quad \mathsf{v}_1 \leq \mathsf{v}_2\ ) \end{aligned} \tag{3.1}$$

where $<$ is defined as usually (note that v_i and p_j are just real numbers between 0 and 1). It is obvious that the pair $(\ \mathbb{KT}, \preccurlyeq_{\mathbb{KT}})$ forms a complete lattice.

[8] As usually, a n-ary function $\hat{F}$ is called *monotonic in the i-th argument* ($i \leq n$), if $x \leq x'$ implies $\hat{F}(x_1,\ \ldots,\ x_{i-1},\ x,\ x_{i+1},\ \ldots,\ x_n) \leq \hat{F}(x_1,\ \ldots,\ x_{i-1},\ x',\ x_{i+1},\ \ldots,\ x_n)$ and a function is called *monotonic* if it is monotonic in all arguments.

[9] Note that the above definition of aggregation operators subsumes all kinds of minimum, maximum or mean operators.

 ISBN: 978-84-614-6167-7

The structure used to give semantics to our programs is the multi-adjoint algebra, presented in [14, 15, 16, 17, 18, 20] and somewhere else. The basic idea is that a multi-adjoint $\Omega-$algebra can be seen as an extension of a multi-adjoint lattice containing a number of extra operators provided by the signature Ω, and a multi-adjoint lattice is just a lattice with more than one pair of operations obeying the adjoint property. We start from the definition of adjoint property.

Definition 3.2 (Adjoint property)**.** Departing from a Poset (a partially ordered set) $< P, \leq >$ and introducing a pair of operations $(\&, \leftarrow)$, we say that the operations form an adjoint pair if (1) $\&$ is increasing in both arguments, (2) $\leftarrow$ is increasing in its first argument and decreasing in the second one and (3) (*the adjoint property*)[10] for any $x, y, z \in P$ we have that $z \leq (x \leftarrow y)$ holds if and only if $z \,\&\, y \leq x$.

A lattice with only one adjoint pair is called somewhere a residuated lattice (see [5, 4]), and when more that one pair is introduced we get to a multi-adjoint lattice.

Definition 3.3 (Multi-Adjoint Lattice)**.** A multi-adjoint lattice $\mathcal{L}$ is a tuple $(L, \leq, \leftarrow_1, \&_1, ..., \leftarrow_n, \&_n)$ satisfying (1) $< L, \leq >$ is a bounded lattice, (2) $(\leftarrow_i, \&_i)$ is an adjoint pair in $< L, \leq >$, for $i = 1, \ldots, n$ and (3) $\top \,\&_i\, \mathsf{v} = \mathsf{v} \,\&_i\, \top = \mathsf{v}$ for all $\mathsf{v} \in L$ and $i = 1, \ldots, n$.

Definition 3.4 (Multi-Adjoint Algebra)**.** Let $(L, \leq, \leftarrow_1, \&_1, ..., \leftarrow_n, \&_n)$ be a multi-adjoint lattice. The implication algebra Ω defining the operators $(\leftarrow_i, \&_i)$ for $i = 1, \ldots, n$ with respect to $\mathcal{L}$ is a multi-adjoint algebra.

It is usual to define a multi-adjoint logic program as a set of weighted rules $A \xleftarrow{F_c,\, c} F(B_1, \ldots, B_n)$ where $c \in [0, 1]$ and F_c is a conjunctor $\&$, but the semantics associated with this syntax is not capable to manage the priority issues we want to encode. To overcome this restriction we enrich this syntax by changing c by $(\mathsf{p}, \mathsf{v}) \in \mathbb{KT}$ and adding a condition $COND(A)$ that can be used to encode a truth value to a subset of individuals fulfilling the condition.

Definition 3.5 (Multi-Adjoint Logic Program)**.** A multi-adjoint logic program is a set of clauses of the form

$$A \xleftarrow{(\mathsf{p},\, \mathsf{v}),\, \&_i} @_i(B_1, \ldots, B_n) \text{ if } COND(A) \tag{3.2}$$

where $(\mathsf{p}, \mathsf{v}) \in \mathbb{KT}$, $\&_i$ is a conjunctor, $@_i$ an aggregator, A and B_i, $i \in [1..n]$, are atoms and $COND(A)$ is a first-order formula (a condition that needs to be satisfied for p to get the truth value v) formed by the predicates in $\mathrm{TB}_{\Pi,\Sigma,V}$, the predicates $=, \geq, \leq, >$ and $<$ restricted to terms from $\mathrm{TU}_{\Sigma,V}$, the symbol true and the conjunction $\wedge$ and disjunction $\vee$ in their usual meaning.

Definitions needed to understand the semantics are given in advance, as usually.

[10] Note that the adjoint property offers us a way to evaluate inference rules because $z \leq (x \leftarrow y)$ iff $z \,\&\, y \leq x$ defines the inference rule $\frac{(B,y) \quad (A \leftarrow B, z)}{(A,x)}$

 ISBN: 978-84-614-6167-7

Definition 3.6 (Valuation, Interpretation)**.** A *valuation* or *instantiation* $\sigma : V \rightarrow \mathbb{HU}$ is an assignment of ground terms to variables and uniquely constitutes a mapping $\hat{\sigma} : \mathrm{TB}_{\Pi,\Sigma,V} \rightarrow \mathbb{HB}$ that is defined in the obvious way.

A *fuzzy Herbrand interpretation* (or short, *interpretation*) of a fuzzy logic program is a mapping $I : \mathbb{HB} \rightarrow \mathbb{KT}$ that assigns an element in our lattice to ground atoms. The *domain* of an interpretation is the set of all atoms in the Herbrand Base, although for readability reasons we omit those atoms to which the truth value $\bot$ is assigned (interpretations are total functions). This mapping can be seen as a set of pairs $(A, (\mathsf{p}, \mathsf{v}))$ such that $A \in \mathbb{HB}$ and $(\mathsf{p}, \mathsf{v}) \in \mathbb{KT} \setminus \{\bot\}$.

It is possible to extend uniquely the mapping I defined on $\mathbb{HB}$ to the set of all ground formulas of the language by using the unique homomorphic extension. This extension is denoted $\hat{I}$ and the set of all interpretations of the formulas in a program $\mathbb{P}$ is denoted $\mathcal{I}_{\mathbb{P}}$.

Definition 3.7 (Interpretation Ordering, Minimum, Maximum, Infimum and Supremum)**.** For two interpretations I and J, we say I *is less than or equal to* J, written $I \sqsubseteq J$, iff $I(A) \preccurlyeq_{\mathbb{KT}} J(A)$ for all $A \in \mathbb{HB}$. Two interpretations I and J are *equal*, written $I = J$, iff $I \sqsubseteq J$ and $J \sqsubseteq I$. For all $A \in \mathbb{HB}$ minimum is defined as $min(I, J)(A) = I(A)$ if $I(A) \preccurlyeq_{\mathbb{KT}} J(A)$ and $min\,(I,\ J)\,(A) = J(A)$ if $J(A) \preccurlyeq_{\mathbb{KT}} I(A)$, maximum as $max\,(I,\ J)\,(A) = I(A)$ if $I(A) \succcurlyeq_{\mathbb{KT}} J(A)$ and $max\,(I,\ J)\,(A) = J(A)$ if $J(A) \succcurlyeq_{\mathbb{KT}} I(A)$, infimum (or intersection) as $(I \sqcap J)(A) := \min(I(A),\ J(A))$ and supremum (or union) as $(I \sqcup J)(A) := \max(I(A),\ J(A))$.

Lemma 3.8. *The pair $(\mathcal{I}_P, \sqsubseteq)$ of the set of all interpretations of a given program with the interpretation ordering forms a complete lattice.*

Proof. This follows readily from the fact that the underlying lattice set $\mathbb{KT}$ forms a complete lattice with the lattice values ordering $\preccurlyeq_{\mathbb{KT}}$. □

Up to now we have defined the underlying lattice we use for choosing the best interpretation between the available ones, but we still have not defined which is the expected one. For that purpose we define the model of a program, but before it we need to define a operator to combine the knowledge grades, $\circ$.

Definition 3.9 (The operator $\circ$)**.** The aim of taking into account the knowledge quality of every single root involved in the inference process removes the possibility to use mathematical operators in which the result remains unchanged when some input does not (i.e. min, max, etc).

Besides, the operator must be formed by a pair of functions ($\circ_{\&}$, $\circ_{\leftarrow}$) where the former is used when combining the knowledge under the application of a conjunction function and the latter when combining it under the implication function.

This is why we decided to use as operator $\circ_{\&}$ the mean, defining

$$x \circ_{\&} y = \frac{x + y}{2} \quad \text{and} \quad z \circ_{\leftarrow} y = 2 * z - y\,.$$

 ISBN: 978-84-614-6167-7

Remark. From here afterwards, the application of some conjunctor $\bar{\&}$ (resp. implicator $\leftarrow$, aggregator $\bar{@}$) to elements $(\mathsf{p},\ \mathsf{v}) \in \mathbb{KT} \setminus \{\bot\}$ refers to the application of the truth function $\hat{\&}$ (resp. $\hat{\leftarrow}$, $\hat{@}$) to the second elements of the tuples while $\circ_{\&}$ (resp. $\circ_{\leftarrow}$, $\circ_{\&}$) is the one applied to the first ones.[11]. When applied to the element $\bot$ all of them ($\bar{\&}$, $\leftarrow$ and $\bar{@}$) return $\bot$.

Definition 3.10 (Multi-Adjoint Satisfaction)**.** (Modified from the definition in [20]) Let $\mathbb{P}$ be a multi-adjoint logic program, $I \in \mathcal{I}_{\mathbb{P}}$ an interpretation and $A \in \mathbb{HB}$ a ground atom. A clause $Cl_i \in \mathbb{P}$ of the form $\{\ A \xleftarrow{(\mathsf{p},\ \mathsf{v}),\ \&_i} @_i(B_1, \ldots, B_n)\ \text{ if }\ COND(A)\ \}$ is satisfied by I iff

$$\begin{aligned}(\mathsf{p},\ \mathsf{v}) \preccurlyeq_{\mathbb{KT}} inf\ \{\ \hat{I}\ ((\ A \leftarrow_{\&_i} @_i(B_1,\ \ldots,\ B_n)\ )\ \xi)\ |\\ \xi\ \text{ any ground instantiation and }\ COND(\ A\ )\ \text{ is satisfied }\}\end{aligned} \tag{3.3}$$

which, by means of the adjoint property, is equivalent to

$$\begin{aligned}\hat{I}\ (\ A\ ) \succcurlyeq_{\mathbb{KT}}\ sup\ \{(\mathsf{p},\ \mathsf{v})\ \bar{\&}_{\&_i}\ \hat{I}((\ @_i(B_1,\ \ldots,\ B_n)\ )\ \xi)\ |\\ \xi\ \text{ any ground instantiation and }\ COND(\ A\ )\ \text{ is satisfied }\}\end{aligned} \tag{3.4}$$

Definition 3.11 (Satisfaction, Model)**.** Let $\mathbb{P}$ be a multi-adjoint logic program, $I \in \mathcal{I}_{\mathbb{P}}$ an interpretation and $A \in \mathbb{HB}$ a ground atom. We say that a clause $Cl_i \in \mathbb{P}$ is satisfied by I or *I is a model of the clause* Cl_i ($I \Vdash Cl_i$) iff for all ground atoms $A \in \mathbb{HB}$ and for all instantiations σ for which $B\sigma \in \mathbb{HB}$ (note that σ can be the empty substitution) it is true that

$$\hat{I}(A) \succcurlyeq_{\mathbb{KT}} (\mathsf{p},\ \mathsf{v})\ \bar{\&}_i\ \bar{@}_i(\ \hat{I}(B_1\sigma),\ \ldots,\ \hat{I}(B_n\sigma)\ )\ \text{ if }\ COND(\ A\ ) \tag{3.5}$$

Note that eq. 3.5 is equivalent to eq. 3.4. Finally, we say that *I is a model of the program* $\mathbb{P}$ and write $I \Vdash \mathbb{P}$ iff $I \Vdash Cl_i$ for all clauses in our multi-adjoint logic program $\mathbb{P}$.

Every program has a least model, which is usually regarded as the intended interpretation of the program, since it is the most conservative model. The following proposition will be an important prerequisite to define the least model semantics. It states that the infimum (or intersection) of a non-empty set of models of a program will again be a model. The existence of a least model is then obvious and easily defined as the intersection of all models.

Proposition 3.12 (Model intersection property)**.** *Let $\mathbb{P}$ be a multi-adjoint logic program and $\mathcal{I}_{\mathbb{P}}$ be a non-empty set of interpretations. Then*

$$I \Vdash \mathbb{P}\ \textit{ for all }\ I \in \mathcal{I}_{\mathbb{P}}\ \textit{ implies }\ \prod_{I \in \mathcal{I}_{\mathbb{P}}} I \Vdash \mathbb{P}$$

[11] Note that this new operators $\bar{\&}$, $\leftarrow$ and $\bar{@}$ still keep the properties exposed in Sec. 2, i.e. the first one is non-decreasing in both coordinates, the second one is non-increasing in the first and non-decreasing in the second coordinate and the last one verifies $\bar{F}(0,\ \ldots,\ 0) = 0$ and $\bar{F}(1,\ \ldots,\ 1) = 1$.

 ISBN: 978-84-614-6167-7

Proof. Suppose that for all $I \in \mathcal{I}_{\mathbb{P}}$ it is true that $I \Vdash \mathbb{P}$. We define $J = \bigsqcap_{I \in \mathcal{I}_{\mathbb{P}}} I$.

(step. 1) from the definition of model of a program (Def. 3.11) we have that $I \Vdash \mathbb{P}$ iff $I \Vdash Cl_i$ for all clauses in our program $\mathbb{P}$, and this results in

$$\hat{I}(A) \succcurlyeq_{\mathbb{KT}} (\mathsf{p},\ \mathsf{v})\ \bar{\&}_i\ \bar{@}_i(\ \hat{I}(B_1\sigma),\ \ldots,\ \hat{I}(B_n\sigma)\ )\ \text{ if }\ COND(\ A\ )$$

for all atoms $A \in \mathbb{HB}$ and for all instantiations σ for which $B\sigma$ is ground.

(step. 2) from $J = \bigsqcap_{I \in \mathcal{I}_{\mathbb{P}}} I$ and the definition of $\bigsqcap$ as the minimum between interpretations (Def. 3.7) we have that for all $I \in \mathcal{I}_{\mathbb{P}}$ it is true that for all $L \in \mathbb{HB}$ $(I \sqcap J)(L) := \min(I(L),\ J(L)) = J(L)$.

(step. 3) define for some ground atom $A \in \mathbb{HB}$ and some ground instantiation σ such that $B\sigma \in \mathbb{HB}$ the variables $(kv)_{I,\sigma}$ and $(kv)_{J,\sigma}$ as follows:

$$\begin{aligned}\hat{I}(A) \succcurlyeq_{\mathbb{KT}} (kv)_{I,\sigma} &= (\mathsf{p},\ \mathsf{v})\ \bar{\&}_i\ \bar{@}_i(\ \hat{I}(B_1\sigma),\ \ldots,\ \hat{I}(B_n\sigma)\ )\ \text{ if }\ COND(\ A\ )\\ (kv)_{J,\sigma} &= (\mathsf{p},\ \mathsf{v})\ \bar{\&}_i\ \bar{@}_i(\ \hat{J}(B_1\sigma),\ \ldots,\ \hat{J}(B_n\sigma)\ )\ \text{ if }\ COND(\ A\ )\end{aligned}$$

from the definition of $\bar{\&}_i$ and $\bar{@}_i$ as non decreasing functions and $\hat{I}(B_i\sigma) \succcurlyeq_{\mathbb{KT}} \hat{J}(B_i\sigma)$ (step. 2) it is clear that $(kv)_{I,\sigma} \succcurlyeq_{\mathbb{KT}} (kv)_{J,\sigma}$.

(step. 4) from $J = \bigsqcap_{I \in \mathcal{I}_{\mathbb{P}}} I$ and the definition of $\bigsqcap$ as the minimum between interpretations (Def. 3.7) we have that for some $I \in \mathcal{I}_{\mathbb{P}}$ and some $L \in \mathbb{HB}$ it is true that $(I \sqcap J)(L) := \min(I(L),\ J(L)) = J(L) = I(L)$.

(step. 5) from $(kv)_{I,\sigma} \succcurlyeq_{\mathbb{KT}} (kv)_{J,\sigma}$ (step. 3) and the fact that $\hat{J}(A)$ gets its value from some $\hat{I}(A)$ (step. 4) we can fix for some atom A and any substitution σ the order $(1,\ 1) \succcurlyeq_{\mathbb{KT}} \hat{I}(A) \succcurlyeq_{\mathbb{KT}} \hat{J}(A) \succcurlyeq_{\mathbb{KT}} (kv)_{I,\sigma} \succcurlyeq_{\mathbb{KT}} (kv)_{J,\sigma} \succcurlyeq_{\mathbb{KT}} \bot$,

(step. 6) The order in step. 5 defines $\hat{J}(A) \succcurlyeq_{\mathbb{KT}} (kv)_{J,\sigma}$, so

$$\hat{J}(A) \succcurlyeq_{\mathbb{KT}} (kv)_{J,\sigma} = (\mathsf{p},\ \mathsf{v})\ \bar{\&}_i\ \bar{@}_i(\ \hat{J}(B_1\sigma),\ \ldots,\ \hat{J}(B_n\sigma)\ )\ \text{ if }\ COND(\ A\ )$$

which proves Prop. 3.12. □

Definition 3.13. Let $\mathbb{P}$ be a well-defined fuzzy logic program. The *least model of* $\mathbb{P}$ is defined as $\mathsf{lm}(\mathbb{P}) := \bigsqcap_{I \Vdash \mathbb{P}} I$.

Definition 3.14 (T_P Operator)**.** Let $\mathbb{P}$ be a multi-adjoint logic program, $L \in \mathbb{HB}$ an atom and $I \in \mathcal{I}_{\mathbb{P}}$ an interpretation. The immediate consequences operator $T_P : \mathcal{I}_{\mathbb{P}} \to \mathcal{I}_{\mathbb{P}}$ is defined as follows:

$$\begin{aligned}T_P(I)(A) \doteq\ \ sup\{\ (\mathsf{p},\ \mathsf{v})\ \bar{\&}_i\ \bar{@}_i(\ \hat{I}(B_1\sigma),\ \ldots,\ \hat{I}(B_n\sigma)\ )\ \text{ if }\ COND(\ A\ )\ \mid\\ \{\ A \xleftarrow{(\mathsf{p},\ \mathsf{v}),\ \&_i}\ @_i(B_1,\ldots,B_n)\ \text{ if }\ COND(A)\ \}\ \in\ \mathbb{P}\ \}\end{aligned} \tag{3.6}$$

As it is usual in the logic programming framework, the semantics of a program $\mathbb{P}$ is characterized by the post-fixpoints of T_P.

 ISBN: 978-84-614-6167-7

Proposition 3.15. *Let $\mathbb{P}$ be a multi-adjoint logic program and $I \in \mathcal{I}_{\mathbb{P}}$ an interpretation.*

$$I \Vdash \mathbb{P} \;\Leftrightarrow\; T_P(I) \sqsubseteq I. \tag{3.7}$$

Proof. "if": Let $T_P(I) \sqsubseteq I$ and L be some arbitrary ground atom. Define for any ground instantiation σ the variable $(kv)_{I,\sigma}$ as follows:

$$\begin{aligned}(kv)_{I,\sigma} \doteq \quad & (\mathsf{p},\, \mathsf{v})\; \bar{\&}_i\; \bar{@}_i(\; \hat{I}(B_1\sigma),\; \ldots,\; \hat{I}(B_n\sigma)\;) \;\text{ if }\; COND(\;A\;) \;\mid \\ & \{\; A \xleftarrow{(\mathsf{p},\, \mathsf{v}),\, \&_i} @_i(B_1, \ldots, B_n) \;\text{ if }\; COND(A) \;\} \in \mathbb{P}\end{aligned} \tag{3.8}$$

so we can say that for any ground instantiation σ

$$\hat{I}(A) \succcurlyeq_{\mathbb{KT}} (kv)_{I,\sigma} \tag{3.9}$$

$$T_P(I)(A) \doteq= sup\,\{\; (kv)_{I,\sigma} \;\} \tag{3.10}$$

and from this and the definition of the symbols sup and $\succcurlyeq_{\mathbb{KT}}$ we can fix the order $\hat{I}(A) \succcurlyeq_{\mathbb{KT}} T_P(I)(A)$, proving that $I \Vdash \mathbb{P}$.

"only if": Let $I \Vdash \mathbb{P}$ and L be some arbitrary ground atom. We define for any ground instantiation σ the variable $(kv)_{I,\sigma}$ as in Eq. 3.8. Since $I \Vdash \mathbb{P}$, for any ground instantiation σ Eq. 3.9 has to be true. If we define our T_P operator from the variable $(kv)_{I,\sigma}$, as in Eq. 3.10, we know from the definition of the symbols sup and $\succcurlyeq_{\mathbb{KT}}$ that $\hat{I}(A) \succcurlyeq_{\mathbb{KT}} T_P(I)(A)$, so $T_P(I) \sqsubseteq I$. □

Proposition 3.16 (T_P is monotonic)**.** *Let $\mathbb{P}$ be a multi-adjoint logic program and $I_i \in \mathcal{I}_{\mathbb{P}}$ and $I_{i+1} \in \mathcal{I}_{\mathbb{P}}$ two interpretations. if $I_i \sqsubseteq I_{i+1} \;\Rightarrow\; T_P(I_i) \sqsubseteq T_P(I_{i+1})$.*

Proof. Suppose that $I_i \sqsubseteq I_{i+1}$. By definition of $\sqsubseteq$ this implies that for all atoms L $\hat{I}_i(L) \preccurlyeq_{\mathbb{KT}} \hat{I}_{i+1}(L)$. In the definition of T_P operator (Def. 3.14) $T_P(I)(L)$ is related to $I(L)$ by means of the operations sup, $\&_i$ and $@_i$. Since all of them are non-decreasing and monotone, we can assure that $T_P(I_i)(L) \preccurlyeq_{\mathbb{KT}} T_P(I_{i+1})(L)$ and conclude $T_P(I_i) \sqsubseteq T_P(I_{i+1})$ □

Proposition 3.17 (T_P is continuous)**.** *Let $\mathbb{P}$ be a multi-adjoint logic program and $I_0 \sqsubseteq I_1 \sqsubseteq \ldots$ a countable infinite increasing sequence of interpretations. Then $T_P\left(\bigsqcup_{n=0}^{\infty} I_n\right) = \bigsqcup_{n=0}^{\infty} T_P(I_n)$.*

Proof. We use the following facts:

(fact. 1) Since $I_0 \sqsubseteq I_1 \sqsubseteq \ldots$ and from definition of $\sqsubseteq$ we have that $I_i(A) \preccurlyeq_{\mathbb{KT}} I_{i+1}(A)$ for every ground term $A \in \mathbb{HB}$. As $\bigsqcup$ takes by definition the maximum interpretation, $\bigsqcup_{i=0}^{n} I_i = I_n$.

(fact. 2) We have that $T_P(I_0) \sqsubseteq T_P(I_1) \sqsubseteq \ldots$ since $I_0 \sqsubseteq I_1 \sqsubseteq \ldots$ and T_P is monotonic (Prop. 3.16). Again by using definitions of $\sqsubseteq$ and $\bigsqcup$ we obtain $\bigsqcup_{i=0}^{n} T_P(I_i) = T_P(I_n)$.

$$T_P\left(\bigsqcup_{n=0}^{\infty} I_n\right) \overset{fact.\ 1}{=} T_P\left(I_\infty\right) \overset{fact.\ 2}{=} \bigsqcup_{n=0}^{\infty} T_P(I_n)$$

□

 ISBN: 978-84-614-6167-7

Theorem 3.18. *Let $\mathbb{P}$ be a multi-adjoint logic program. Then the least fixpoint of T_P exists and is equal to $T_P{\uparrow}\omega$.*

Proof. The existence of the least fixpoint of T_P follows from the facts that $(\mathcal{I}_{\mathbb{P}}, \sqsubseteq)$ forms a complete lattice, T_P is monotone (Proposition 3.16), and the Knaster-Tarski fixpoint theorem [22, 10]. Its equality to $T_P{\uparrow}\omega$ follows from the facts that $(\mathcal{I}_{\mathbb{P}}, \sqsubseteq)$ forms a complete lattice, T_P is continuous (Proposition 3.17), and the Kleene fixpoint theorem [9]. □

Since the least fixpoint always exists, we can define a semantics based on it.

Definition 3.19. Let $\mathbb{P}$ be a multi-adjoint logic program. Then the *least fixpoint semantics of* $\mathbb{P}$ is defined as $\mathsf{lfp}(\mathbb{P}) = T_P{\uparrow}\omega(\bot)$. Here, $\bot$ denotes the interpretation mapping everything to $\bot$ (thus being the least element of the lattice $(\mathcal{I}_P, \sqsubseteq)$).

Theorem 3.20. *For a multi-adjoint logic program $\mathbb{P}$, we have* $\mathsf{lm}(\mathbb{P}) = \mathsf{lfp}(\mathbb{P})$.

Proof.

$$\mathsf{lm}(\mathbb{P}) \overset{1}{=} \bigsqcap_{I \Vdash \mathbb{P}} I \overset{2}{=} \bigsqcap_{T_P(I) \sqsubseteq I} I \overset{3}{=} T_P{\uparrow}\omega(\bot) \overset{4}{=} \mathsf{lfp}(\mathbb{P})$$

where (1) is by definition of least model of a program, (2) is by Prop. 3.15, (3) is by the Kleene fixpoint theorem [9] and (4) is by definition of least fixpoint semantics. □

4 Conclusions

We have presented syntax and semantics for a fuzzy logic framework in which priorities between the different rules in a fuzzy logic program can be set and are taken into account when computing the answer for a query. We hope this contribution fills the existing gap in fuzzy logic frameworks and enable them to develop programs more close to the human way of thinking, where the use of priorities is a reality.

Acknowledgements

This work is partially supported by research projects DESAFIOS10 (TIN2009-14599-C03-00) funded by Ministerio Ciencia e Innovación of Spain, PROMETIDOS (P2009/TIC-1465) funded by Comunidad Autónoma de Madrid and Research Staff Training Program (BES-2008-008320) funded by the Spanish Ministry of Science and Innovation. It is partially supported too by the Universidad Politécnica de Madrid entities Departamento de Lenguajes Sistemas Informáticos e Ingeniería de Software and Facultad de Informática.

 ISBN: 978-84-614-6167-7

References

[1] Anastasia Analyti and Sakti Pramanik. Reliable semantics for extended logic programs with rule prioritization. *J. Log. Comput.*, pages 303–324, 1995.

[2] Fernando Bobillo and Umberto Straccia. fuzzydl: An expressive fuzzy description logic reasoner. In *2008 International Conference on Fuzzy Systems (FUZZ-08)*, pages 923–930. IEEE Computer Society, 2008.

[3] Alain Colmerauer and Philippe Roussel. The birth of prolog. *SIGPLAN Not.*, 28:37–52, March 1993.

[4] Carlos Viegas Damásio and Luís Moniz Pereira. Hybrid probabilistic logic programs as residuated logic programs. In *Proceedings of the European Workshop on Logics in Artificial Intelligence*, JELIA '00, pages 57–72, London, UK, 2000. Springer-Verlag.

[5] Carlos Viegas Damásio and Luís Moniz Pereira. Monotonic and residuated logic programs. In *Proceedings of the 6th European Conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty*, ECSQARU '01, pages 748–759, London, UK, 2001. Springer-Verlag.

[6] James P. Delgrande, Torsten Schaub, and Hans Tompits. A framework for compiling preferences in logic programs. *Theory Pract. Log. Program.*, 3:129–187, March 2003.

[7] S. Guadarrama, S. Muñoz-Hernández, and C. Vaucheret. Fuzzy prolog: a new approach using soft constraints propagation. *Fuzzy Sets and Systems (FSS)*, 144(1):127 – 150, 2004. Possibilistic Logic and Related Issues.

[8] Bharat Jayaraman, Kannan Govindarajan, and Surya Mantha. Logic programming with preferences and constraints, 2007.

[9] Stephen Cole Kleene. *Introduction to Metamathematics*. D. Van Nostrand, New York, 1952.

[10] B. Knaster. Un théorème sur les fonctions d'ensembles. *Annales Soc. Polonaise*, 6:133–134, 1928.

[11] Robert A. Kowalski. The early years of logic programming. *Commun. ACM*, 31:38–43, January 1988.

[12] Els Laenens and Dirk Vermeir. A fixpoint semantics for ordered logic. *Journal of Logic and Computation*, 1:159–185, 1990.

[13] V. Wiktor Marek, Anil Nerode, and Jeffrey B. Remmel. Basic forward chaining construction for logic programs. In *Proceedings of the 4th International Symposium on Logical Foundations of Computer Science*, pages 214–225, London, UK, 1997. Springer-Verlag.

 ISBN: 978-84-614-6167-7

[14] Jesús Medina, Manuel Ojeda-Aciego, and Peter Vojtáš. A multi-adjoint approach to similarity-based unification. *Electronic Notes in Theoretical Computer Science*, 66(5):70 – 85, 2002. UNCL'2002, Unification in Non-Classical Logics (ICALP 2002 Satellite Workshop).

[15] Jesús Medina, Manuel Ojeda-Aciego, and Peter Vojtás. A completeness theorem for multi-adjoint logic programming. In *FUZZ-IEEE*, pages 1031–1034, 2001.

[16] Jesús Medina, Manuel Ojeda-Aciego, and Peter Vojtás. Multi-adjoint logic programming with continuous semantics. In Thomas Eiter, Wolfgang Faber, and Miroslaw Truszczynski, editors, *LPNMR*, volume 2173 of *Lecture Notes in Computer Science*, pages 351–364. Springer, 2001.

[17] Jesús Medina, Manuel Ojeda-Aciego, and Peter Vojtás. A procedural semantics for multi-adjoint logic programming. In Pavel Brazdil and Alípio Jorge, editors, *EPIA*, volume 2258 of *Lecture Notes in Computer Science*, pages 290–297. Springer, 2001.

[18] Jesús Medina, Manuel Ojeda-Aciego, and Peter Vojtás. Similarity-based unification: a multi-adjoint approach. *Fuzzy Sets and Systems*, 146(1):43–62, 2004.

[19] Pedro J. Morcillo and Gines Moreno. Programming with fuzzy logic rules by using the floper tool. In *RuleML '08: Proceedings of the International Symposium on Rule Representation, Interchange and Reasoning on the Web*, pages 119–126, Berlin, Heidelberg, 2008. Springer-Verlag.

[20] Jesus Medina Moreno and Manuel Ojeda Aciego. On first-order multi-adjoint logic programming. In *11th Spanish Congress on Fuzzy Logic and Technology*, 2002.

[21] Susana Munoz-Hernandez, Vctor Pablos-Ceruelo, and Hannes Strass. Rfuzzy: Syntax, semantics and implementation details of a simple and expressive fuzzy tool over prolog. *Information Sciences*, 181(10):1951 – 1970, 2011. Special Issue on Information Engineering Applications Based on Lattices.

[22] Alfred Tarski. A lattice-theoretical fixpoint theorem and its applications. *Pacific Journal of Mathematics*, 5:285–309, 1955.

[23] Kewen Wang, Lizhu Zhou, and Fangzhen Lin. Alternating fixpoint theory for logic programs with priority. In *Proceedings of the First International Conference on Computational Logic*, CL '00, pages 164–178, London, UK, 2000. Springer-Verlag.

[24] Lotfi A. Zadeh. Fuzzy sets. *Information and Control*, 8(3):338–353, 1965.

[25] Lotfi A. Zadeh. The concept of a linguistic variable and its application to approximate reasoning - i. *Inf. Sci.*, 8(3):199–249, 1975.

[26] Lotfi A. Zadeh. Is there a need for fuzzy logic? *Information Sciences*, 178(13):2751 – 2779, 2008.

 ISBN: 978-84-614-6167-7

Proceedings of the 11th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2011
26–30 June 2011.

Compartmental Mathematical Modelling of Immune System-Melanoma Competition

Marzio Pennisi[1], Carlo Bianca[2], Francesco Pappalardo[1] and Santo Motta[1]

[1] *Dipartimento di Matematica, Politecnico di Torino*

[2] *Dipartimento di Matematica & Informatica, Università di Catania*

emails: mpennisi@dmi.unict.it, carlo.bianca@polito.it, francesco@dmi.unict.it, motta@dmi.unict.it

Abstract

This paper deals with a preliminary model developed to sketch the immune response stimulated by the administration of OT1 activated cytotoxic T cells with Anti-CD 137 immunostimulatory monoclonal antibodies against melanoma cells. We model two compartments: the injection point compartment where the treatment is administered and the skin compartment where melanoma tumor cells proliferate. To model the migration of OT1 T cells and Antibodies from the injection to the skin compartment we use delay differential equations (DDE). We then present preliminary results showing the immune response entitled with the use of the treatment.

1 Introduction

Melanoma represents one of the most aggressive malignant tumors and it is due to the mutation of cells that produce the melanin (melanocytes). Many approaches in clinical and preclinical studies are actually based on the use and stimulation of cytotoxic T lymphocytes against melanoma cells. Immunological rejection of progressive tumors requires not only activation and expansion of tumor specific cytotoxic T lymphocytes (CTL), but also an efficient effector phase including migration of CTL in the tumor followed by conjugation and killing of target cells.

Accumulating evidence suggests that tumor-infiltrating lymphocytes are rendered anergic through the actions of co-inhibitory molecules expressed on the surface of tumor and stroma cells. Successful immunotherapy requires combined strategies that are able to

 ISBN: 978-84-614-6167-7

turn-off deleterious signals while enhancing CTL migration and overall killing capacity [1]. CD137, also known as 4-1BB, is a co-stimulatory protein expressed on activated T, NK, B-lymphocytes, dendritic cells and tumor endothelium [2]. CD137 natural ligand, CD137L is present on the surface of activated antigen presenting cells [3]. Recently, in vivo experiments executed in B16-OVA mice models [2] revealed that the combination of in vitro Activated-OT1 cytotoxic T cells with Anti-CD137 immunostimulatory monoclonal antibodies that improve cytotoxicity, duplication rate and chemotaxis sensitivity of activated cytotoxic T cells are able to prevent the melanoma formation.

To catch-up the dynamics of this biological process we sketched a two-compartments model. We used Delay Difference Equations to model two different compartments: the injection point compartment where both antibodies and OT1 cells are injected and the Skin compartment where melanoma develops.

2 The Model

The experiment runs for 30 days. At day 0 B16-OVA mice receive one injection of melanoma malignant cells. The therapeutic treatment is administered at day 3. We built a mathematical model describing the biological process.

Equation 1 and 2 refer to first compartment and simulate the time evolution of both injected activated OT1 cytotoxic T cells (E) and antibodies (Ab) where kin(t; r) is a known function that represents the number of inoculated entities r at the scheduled injection time t. Both the entities migrate to the skin compartment with given rates (terms $-a_{11}E$ and $-a_{11}Ab$) and are subject to death or natural degradation (terms $-a_8E$ and $-a_{12}Ab$). Equation 3 describes the melanoma cells (C) behavior in the skin compartment. The first term $((a_1 - a_2 \ln(C)) \cdot C)$ represents a gompertizan growth whereas the second term denotes killing of C by Activated OT1 T cells that are already in the skin compartment (E_s). With equation 4 we describe the tumor associated antigen (A) dynamics. Antigens are released in the skin compartment by killed melanoma cells ($a_4 \cdot (a_3 C E_s)$) and are subject to natural degradation ($-a_5 A$).

Activated OT1 T cells that have migrated to the skin compartment (E_s) are described by equation 5. The term $a_7 A_s E_s$ is used to model duplication of OT1 T cells. Anti CD-137 antibodies that reached the skin compartment (A_s) are able to boost OT1 T cells duplication rates. The term $a_{11}E(t-\tau)$ models migration of OT1 T cells from the injection point to the skin compartment. OT1 T cells in the skin compartment are supposed to be proportional to the number of OT1 T cells in the injection point compartment with a proportionality constant a_{11} and a time delay of τ. Some antigens released by killed melanoma cells may be captured by presenting cells such as macrophages and dendritic cells and presented to Naive cytotoxic T cells (N). After the right chain of steps (i.e. stimulation by T helper cells) these cells may become active and then able to kill melanoma tumor cells. This process

 ISBN: 978-84-614-6167-7

is not modeled since it involves to model other entities that are not essential at this first stage. We then estimate the number of newly activated OT1 cytotoxic T cells on the basis of released antigens with the term a_6NA. The last term $(-a_8E_s)$ reproduces natural death of OT1 T cells. Equation 6 models the behavior of Naive OT1 T cells that are already in the skin compartment. The term $h(M-N)$ models homeostasis. M is the number of circulating naive T cells under safe conditions given by the leukocyte formula. The second term (a_6NA) models the cytotoxic T cells state changing from naive to activated (E_s).

Antibodies that have reached the skin compartment (A_s) are modeled and described by equation 7. Antibodies in the skin compartment are supposed to be proportional to the number of antibodies in the injection point compartment (Ab) with a proportionality constant a_{11} and a time delay of τ. They also disappear by stimulating OT1 cells activities and are subject to a natural degradation (terms $-a_{12}A_s$ and $-a_9A_sE_s$).

Injection point compartment

$$\frac{dE}{dt} = kin(t,p) - a_{11}E - a_8E \quad (1)$$

$$\frac{dAb}{dt} = kin(t,k) - a_{11}Ab - a_{12}Ab \quad (2)$$

Skin compartment

$$\frac{dC}{dt} = (a_1 - a_2\ln(C))\cdot C - a_3E_sC \quad (3)$$

$$\frac{dA}{dt} = a_4\cdot(a_3CE_s) - a_5A \quad (4)$$

$$\frac{dEs}{dt} = a_7A_sE_s + a_{11}E(t-\tau) + a_6NA - a_8E_s \quad (5)$$

$$\frac{dN}{dt} = h(M-N) - a_6NA \quad (6)$$

$$\frac{dA_s}{dt} = a_{11}Ab(t-\tau) - a_9A_sE_s - a_{12}A_s \quad (7)$$

3 Preliminary Results and Conclusions

According to in vivo data from literature, our past experience in this field [4, 5] and experimental data coming from the experiment, we were able to find a preliminary tuning of the model able to qualitatively reproduce the time evolution of the system. In absence of therapy there is no induced immune response and thus melanoma cells grow up to their saturation threshold. In figure 1 we show the system behavior when the treatment is administered. Antibodies (Ab) and activated OT1 cells (E) are injected at day three and then migrate to the skin compartment (see fig.1 (g) and (d)). At the same time Antibodies (As) and activated OT1 cells (Es) in the skin compartment (fig.1 (f) and (c)) growth up to their maximum and cooperate to kill the melanoma cells (fig.1 (a)) entitling almost complete

 ISBN: 978-84-614-6167-7

eradication before day 20. Note here that the number of activated OT1 cells in the skin (E_s) is also boosted by (previously) naive OT1 T cells (N) (see fig.1 (e)) that are recruited thanks to killed melanoma cells antigens (A) (fig.1 (b)). As an initial conclusion we can therefore say that the treatment acts in two ways: directly by activated OT1 cytotoxic T cells that are able to kill melanoma and Antibodies that boost T cells activities, and indirectly by promoting recruitment of naive OT1 cytotoxic T cells thanks to the releasing of melanoma cells antigens captured by presenting cells and then presented to these.

Further improvement to the tuning as well as deeper analysis of the model are on the way and will be presented in due course.

4 Figures

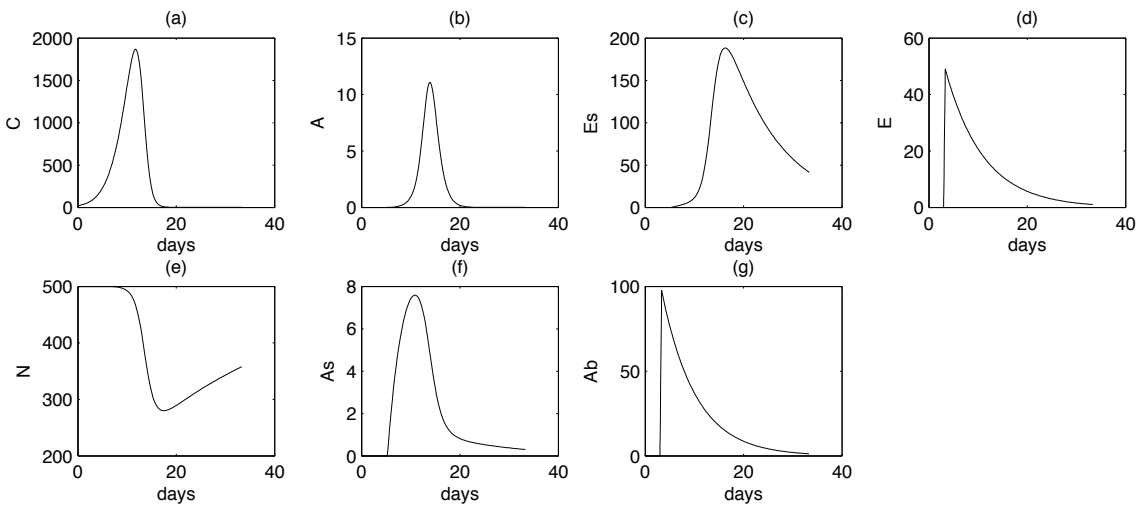


Figure 1: System behavior entitled with the use of activated OT1 cytotoxic T cells + AntiCD-137 monoclonal Antibodies.

Acknowledgements

CB was partially supported by the FIRB project-RBID08PP3J-Metodi matematici e relativi strumenti per la modellizzazione e la simulazione della formazione di tumori, competizione con il sistema immunitario, e conseguenti suggerimenti terapeutici.

 ISBN: 978-84-614-6167-7

References

[1] Perez-Gracia JL, Berraondo P, Martinez-Forero I, Alfaro C, et al. *Clinical development of combination strategies in immunotherapy: are we ready for more than one investigational product in an early clinical trial?* Immunotherapy, 1:845-853 (2009).

[2] Palazon A, Teijeira A, Martinez-Forero I, Hervas-Stubbs S, et al. *Agonist anti-CD137 mAb act on tumor endothelial cells to enhance recruitment of activated T lymphocytes,* Cancer Res, 71:801-811 (2011).

[3] Melero I, Murillo O, Dubrot J, Hervas-Stubbs S, Perez-Gracia JL *Multi-layered action mechanisms of CD137 (4-1BB)-targeted immunotherapies,* Trends Pharmacol Sci, 29:383-390 (2008).

[4] Pennisi M, Bianca C, Pappalardo F, Motta S. *Modeling artificial immunity against mammary carcinoma,* CMMSE 2010, ISBN 978- 84-613-5510-5, 753-756 (2010).

[5] Pennisi M, Pappalardo F, Palladini A, Nicoletti G, Nanni P, Lollini P-L, Motta S *Modeling the competition between lung metastases and the immune system using agents,* BMC Bioinformatics, 11(Suppl 7):S13 (2010).

Proceedings of the 10th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2010
27–30 June 2010.

Proper and weak efficiency for unconstraint vector optimization problems

Emilia-Loredana Pop[1] **and Dorel I. Duca**[1]

[1] *Faculty of Mathematics and Computer Science, Babeş-Bolyai University, 1 M. Kogălniceanu Street, 400084 Cluj-Napoca, România.*

emails: lory_pel@yahoo.com, dorelduca@yahoo.com

Abstract

Considering the vector optimization problem

$$\underset{x\in X}{\text{Min}}\{(f\circ A)(x)+g(x)\}$$

where $f:Y\to\overline{V}$, $g:X\to\overline{V}$ and $A\in\mathcal{L}(X,Y)$ invertible, we attach to it, by means of perturbation theory, new vector duals and establish duality results with respect to properly and weakly efficient solutions via linear scalarization. By extending the classical vector Wolfe and Mond-Weir duals we obtain different vector dual problems.

Key words: vector optimization problem, perturbation function, conjugate function, regularity conditions, strong duality, properly efficient solutions, weakly efficient solutions, Wolfe vector duals, Mond-Weir vector duals
MSC 2000: 49N15, 90C25, 90C29, 90C46

1 Introduction

We consider the optimization problem having the composition with a linear continuous mapping in the objective function and give some duality results referring to properly efficient solutions and weakly efficient solutions via linear scalarization. Then, for the unconstraint vector optimization problem we introduce the Wolfe and Mond-Weir type vector duals and obtain weak and strong duality results (see [4], [10], [11], [12]).

Consider two separated locally convex spaces X and Y and their topological dual spaces X^* and Y^*, respectively, endowed with the corresponding weak* topologies and denote by $\langle x^*,x\rangle=x^*(x)$ the value at $x\in X$ of the linear continuous functional $x^*\in X^*$. A cone $K\subseteq X$ is a nonempty set which fulfills $\lambda K\subseteq K$, for all $\lambda\geq 0$. A cone $K\subseteq X$ is called *pointed* if $K\cap(-K)=\{0\}$. For $K\subseteq X$ a nonempty convex cone, $K^*=\{x^*\in X^*:\langle x^*,x\rangle\geqq 0,\forall x\in K\}$ is the *positive dual cone.* In optimization

 ISBN: 978-84-614-6167-7

the *quasi interior of the dual cone* of K, $K^{*0} = \{x^* \in K^* : \langle x^*, x\rangle > 0, \text{ for all } x \in K \setminus \{0\}\}$ is used. The *algebraic interior* of a set $U \subseteq X$ is $core(U) = \{x \in X :$for each $y \in X\, \exists \delta > 0$ such that $x + \lambda y \in U, \forall \lambda \in [0, \delta]\}$. For a subset U of X by $\mathrm{cl}(U)$, $\mathrm{lin}(U)$, $\mathrm{cone}(U)$, $\mathrm{ri}(U)$ and $\dim(U)$ we denote its *closure, linear hull, conical hull, relative interior* and *dimension*, respectively. If U is convex, then $\mathrm{sqri}(U) = \{x \in U : \mathrm{cone}(U - x)$ is a closed linear subspace$\}$ denotes its *strong quasi relative interior*. On Y it is considered the partial ordering $\leqq_K$ induced by the convex cone $K \subseteq Y$ defined by $z \leqq_K y \Leftrightarrow y - z \in K$ when $z, y \in Y$. We denote also $z \leq_K y$ if $y - z \in K \setminus \{0\}$ and $z <_K y$ if $y - z \in \mathrm{int}(K)$, when $z, y \in Y$. The greatest element with respect to $\leqq_K$, which does not belong to Y, is denoted by ∞_K and let $\overline{Y} = Y \cup \{\pm\infty_K\}$.

For a function $f : X \to \overline{Y}$ the *domain* is defined by $\mathrm{dom}\, f = \{x \in X : f(x) \in Y\}$. For $f : X \to \overline{\mathbb{R}}$ the *epigraph* is given by $\mathrm{epi}\, f = \{(x, r) \in X \times \mathbb{R} : f(x) \leq r\}$ and the *conjugate function* $f^* : X^* \to \overline{\mathbb{R}}$ is defined by $f^*(x^*) = \sup\{\langle x^*, x\rangle - f(x) : x \in X\}$. The function $f : X \to \overline{\mathbb{R}}$ is *lower semicontinuous* at $\overline{x} \in X$ if $\liminf_{x \to \overline{x}} f(x) \geq f(\overline{x})$. The function $f : X \to \overline{Y}$ is called *proper* if its domain is nonempty. Between a function and its conjugate there is the *Young-Fenchel inequality* $f^*(x^*) + f(x) \geq \langle x^*, x\rangle$, for all $x \in X$ and $x^* \in X^*$. For $f : X \to \overline{\mathbb{R}}$ a given function and an arbitrary $x \in X$ such that $f(x) \in \mathbb{R}$, the set $\partial f(x) = \{x^* \in X^* : f(y) - f(x) \geq \langle x^*, y - x\rangle \text{ for all } y \in X\}$ is the (convex) *subdifferential of f at x* and its elements are called *subgradients of f at x* and the function f is *subdifferential at x* if $\partial f(x) \neq \emptyset$ (see [5], [6], [7], [9], [13]). The *adjoint operator* of a linear continuous mapping $A : X \to Y$ is $A^* : Y^* \to X^*$ given by $\langle A^* y^*, x\rangle = \langle y^*, Ax\rangle$, for any $(x, y^*) \in X \times Y^*$. For a vector function $f : X \to \overline{Y}$, f is *$K-$convex* if $f(tx + (1-t)y) \leqq_K tf(x) + (1-t)f(y)\ \forall x, y \in X, \forall t \in [0, 1]$.

Let X and Y be two Hausdorff locally convex spaces and $G : X \to \overline{\mathbb{R}}$ a given function and the general optimization problem

$$\inf_{x \in X} G(x). \tag{$\overline{P}$}$$

To this problem it is assigned a conjugate dual problem introduced by making use of the perturbation approach (see [1], [2]). The perturbation function of Problem $(\overline{P})$, $\Psi : X \times Y \to \overline{\mathbb{R}}$ with $\Psi(x, 0) = G(x)$, for all $x \in X$, is considered. The equivalent problem of $(\overline{P})$ becomes

$$\inf_{x \in X} \Psi(x, 0) \tag{$\overline{P}_0$}$$

and the conjugate dual problem can be formulated as

$$\sup_{x \in X} \{-\Psi^*(0, y^*)\}, \tag{$\overline{D}$}$$

where $\Psi^* : X^* \times Y^* \to \overline{\mathbb{R}}$ is the conjugate function of Ψ and X^* and Y^* are the topological dual spaces of the feasible variables X and of the perturbation variables Y, respectively. By $v(\overline{P}), v(\overline{D})$ are denoted the optimal objective values of the problems $(\overline{P})$ and $(\overline{D})$. The infimal value function of Ψ is $h : Y \to \overline{\mathbb{R}}$, $h(y) = \inf\limits_{x \in X}\{\Psi(x, y)\}$ and we have that $v(\overline{P}) = h(0)$. Problem $(\overline{P})$ is called *stable* if $h(0) \in \mathbb{R}$ and h is subdifferentiable at 0.

 ISBN: 978-84-614-6167-7

In [8] to the optimization problem

$$\underset{x\in X}{\text{Min}}\{\varphi(Ax)+\psi(x)\}, \tag{P}$$

where X and Y are Hausdorff locally convex spaces, $\varphi : Y \to \overline{\mathbb{R}}$ and $\psi : X \to \overline{\mathbb{R}}$ are proper functions fulfilling $\operatorname{dom}\psi \cap A^{-1}(\operatorname{dom}\varphi) \neq \emptyset$ and $A \in \mathcal{L}(X,Y)$ is invertible, we attached its Fenchel dual problem

$$\underset{y^*\in X^*}{\text{Max}}\{-\varphi^*((A^{-1})^*(-y^*)) - \psi^*(y^*)\} \tag{D}$$

and we established some duality theorems.

Theorem 1 *Let $\varphi : Y \to \overline{\mathbb{R}}$, $\psi : X \to \overline{\mathbb{R}}$ be proper and convex functions and $A \in \mathcal{L}(X,Y)$ invertible such that $\operatorname{dom}\psi \cap A^{-1}(\operatorname{dom}\varphi) \neq \emptyset$. If the regularity condition*

$$\mid\ \exists x' \in \operatorname{dom}\psi \cap A^{-1}(\operatorname{dom}\varphi) \text{ such that } \psi \text{ is continuous at } x' \tag{RC}$$

is fulfilled, then $v(P) = v(D)$ and the dual (D) has an optimal solution.

In this paper we consider the vector optimization problem

$$\underset{x\in X}{\text{Min}}\, F(x) = (f\circ A)(x) + g(x) \tag{VP}$$

where X, Y and V are Hausdorff locally convex spaces, K is a nontrivial pointed convex cone in V, V is partially ordered by K, $f : Y \to \overline{V}$ and $g : X \to \overline{V}$ are proper functions and $A \in \mathcal{L}(X,Y)$ is invertible.

Definition 1 (*See [2]*) *An element $\overline{x} \in X$ is called:*

(*i*) *efficient solution to (VP) if $\overline{x} \in \operatorname{dom}F$ and there is no $x \in \operatorname{dom}F$ such that $F(x) - F(\overline{x}) \in K \setminus \{0\}$.*

(*ii*) *weakly efficient solution to (VP) if $\overline{x} \in \operatorname{dom}F$ and there is no $x \in \operatorname{dom}F$ such that $F(x) - F(\overline{x}) \in \operatorname{core}K$.*

(*iii*) *properly efficient solution to (VP) if $\overline{x} \in \operatorname{dom}F$ and there exists $v^* \in K^{*0}$ such that $\langle v^*, F(\overline{x})\rangle \leq \langle v^*, F(x)\rangle$, for all $x \in X$.*

It follows that:

(*a*) the element $\overline{x} \in \operatorname{dom}F$ is an efficient solution to (VP) iff $(F(\overline{x}) - K) \cap F(\operatorname{dom}F) = \{F(\overline{x})\}$;

(*b*) the element $\overline{x} \in \operatorname{dom}F$ is a weakly efficient solution to (VP) iff $(F(\overline{x}) - \operatorname{core}(K)) \cap F(\operatorname{dom}F) = \emptyset$.

In this paper we establish connections between the vector optimization problem (VP) and the vector dual problems attached. In section 2 we give weak duality, strong duality and converse duality theorems for Problem (VP) and the vector dual problems to (VP) with respect to properly efficient solutions and weakly efficient solutions. In section 3 we study the connections between properly efficient solutions for vector optimization problem (VP) and efficient solutions for Wolfe and Mond-Weir vector dual problems attached.

 ISBN: 978-84-614-6167-7

2 Duality via scalarization

In this section we study the duality with respect to properly efficient solutions and weakly efficient solutions.

2.1 Duality with respect to properly efficient solutions

In this subsection, to the vector optimization problem (VP), we attach the vector dual problem

$$\underset{(v^*,y^*,v)\in\mathcal{B}}{\text{Max}}\, h(v^*,y^*,v) \tag{VD}$$

where

$$\mathcal{B}=\{(v^*,y^*,v)\in K^{*0}\times Y^*\times V:\langle v^*,v\rangle \leqq -(v^*f)^*((A^{-1})^*(-y^*))-(v^*g)^*(y^*)\}$$

and

$$h(v^*,y^*,v)=v.$$

In what follows some connections between the properly efficient solutions to problems (VP) and (VD) are given.

From Definition 1 we have that $\overline{x}\in X$ is a *properly efficient solution* to Problem (VP) if and only if $\overline{x}\in \operatorname{dom} g\cap A^{-1}(\operatorname{dom} f)$ and there exists $v^*\in K^{*0}$ such that $\langle v^*,(f\circ A+g)(\overline{x})\rangle \leqq \langle v^*,(f\circ A+g)(x)\rangle,$ for all $x\in X$.

Theorem 2 *There is no $x\in X$ and no $(v^*,y^*,v)\in\mathcal{B}$ such that $(f\circ A+g)(x)\leq_K h(v^*,y^*,v)$.*

Proof. We assume by contradiction, that there exist $\overline{x}\in X$ and $(\overline{v}^*,\overline{y}^*,\overline{v})\in\mathcal{B}$ such that

$$\overline{v}=h(\overline{v}^*,\overline{y}^*,\overline{v})\geq_K (f\circ A+g)(\overline{x}). \tag{1}$$

Then $\overline{x}\in\operatorname{dom} g\cap A^{-1}(\operatorname{dom} f)$, $\overline{v}^*\in K^{*0}$ and

$$\langle\overline{v}^*,\overline{v}\rangle \leqq -(\overline{v}^*f)^*((A^{-1})^*(-\overline{y}^*))-(\overline{v}^*g)^*(\overline{y}^*). \tag{2}$$

On the other hand, from (1) and $\overline{v}^*\in K^{*0}$ we deduce that

$$\langle\overline{v}^*,\overline{v}\rangle>\langle\overline{v}^*,(f\circ A)(\overline{x})\rangle+\langle\overline{v}^*,g(\overline{x})\rangle. \tag{3}$$

Since the functions $(\overline{v}^*f):Y\to\overline{\mathbb{R}}$ and $(\overline{v}^*g):X\to\overline{\mathbb{R}}$ are proper and convex with $\operatorname{dom}(\overline{v}^*f)=\operatorname{dom} f$ and $\operatorname{dom}(\overline{v}^*g)=\operatorname{dom} g$, the dual of the optimization problem

$$\inf_{x\in X}\{(\overline{v}^*f)(Ax)+(\overline{v}^*g)(x)\} \tag{SP}$$

is the problem

$$\sup_{y^*\in Y^*}\{-(\overline{v}^*f)^*((A^{-1})^*(-y^*))-(\overline{v}^*g)^*(y^*)\}. \tag{DSP}$$

 ISBN: 978-84-614-6167-7

In view of the weak duality theorem, applied to the pair of scalar problems $(SP)-(DSP)$, we have $(\overline{v}^*f)(Ax)+(\overline{v}^*g)(x) \geqq -(\overline{v}^*f)^*\ ((A^{-1})^*(-y^*))-(\overline{v}^*g)^*(y^*)$, for all $x\in X$ and $y^*\in Y^*$. Then

$$(\overline{v}^*f)(A\overline{x})+(\overline{v}^*g)(\overline{x}) \geqq -(\overline{v}^*f)^*((A^{-1})^*(-\overline{y}^*))-(\overline{v}^*g)^*(\overline{y}^*), \tag{4}$$

because $\overline{x}\in X$, $\overline{y}^*\in Y^*$. Consequently,

$$\langle \overline{v^*},\overline{v}\rangle \overset{(3)}{>} \langle \overline{v}^*,(f\circ A)(\overline{x})\rangle+\langle \overline{v}^*,g(\overline{x})\rangle \overset{(4)}{\geqq} -(\overline{v}^*f)^*((A^{-1})^*(-\overline{y}^*))-(\overline{v}^*g)^*(\overline{y}^*),$$

which contradicts (2). ■

Theorem 3 *If $\overline{x}\in X$ is a properly efficient solution to (VP) and the regularity condition*

$$\mid\ \exists x'\in \operatorname{dom} g\cap A^{-1}(\operatorname{dom} f) \text{ such that } g \text{ is continuous at } x' \tag{RCV}$$

is fulfilled, then there exists an efficient solution $(\overline{v}^*,\overline{y}^*,\overline{v})$ to (VD) such that $(f\circ A+g)(\overline{x})=h(\overline{v}^*,\overline{y}^*,\overline{v})=\overline{v}$.

Proof. Let $\overline{x}\in X$ be a properly efficient solution to (VP); then $\overline{x}\in \operatorname{dom} g\cap A^{-1}(\operatorname{dom} f)$ and there exists $\overline{v}^*\in K^{*0}$ such that $\langle \overline{v}^*,(f\circ A+g)(\overline{x})\rangle=\inf\limits_{x\in X}\{(\overline{v}^*f)(Ax)+(\overline{v}^*g)(x)\}$, hence $\overline{x}$ is a solution to Problem (SP).

Obviously, the functions $(\overline{v}^*f):Y\to\overline{\mathbb{R}}$ and $(\overline{v}^*g):X\to\overline{\mathbb{R}}$ are proper and convex functions with $\operatorname{dom}(\overline{v}^*f)=\operatorname{dom} f$ and $\operatorname{dom}(\overline{v}^*g)=\operatorname{dom} g$, and the regularity condition (RCV) yields that $(\overline{v}^*g)$ is continuous at some x'. Then, by strong duality theorem (Theorem 1) applied to the pair of scalar problems $(SP)-(DSP)$ we have $\inf\limits_{x\in X}\{(\overline{v}^*f)(Ax)+(\overline{v}^*g)(x)\}=\sup\limits_{y^*\in Y^*}\{-(\overline{v}^*f)^*((A^{-1})^*(-y^*))-(\overline{v}^*g)^*(y^*)\}$ and (DSP) has a solution $\overline{y}^*\in Y^*$, i.e. we have $\sup\limits_{y^*\in Y^*}\{-(\overline{v}^*f)^*((A^{-1})^*(-y^*))-(\overline{v}^*g)^*(y^*)\}=-(\overline{v}^*f)^*((A^{-1})^*(-\overline{y}^*))-(\overline{v}^*g)^*(\overline{y}^*)$. Then

$$\begin{aligned}&\langle \overline{v}^*,(f\circ A+g)(\overline{x})\rangle=\inf_{x\in X}\{(\overline{v}^*f)(Ax)+(\overline{v}^*g)(x)\}\\&=\sup_{y^*\in Y^*}\{-(\overline{v}^*f)^*((A^{-1})^*(-y^*))-(\overline{v}^*g)^*(y^*)\}=-(\overline{v}^*f)^*((A^{-1})^*(-\overline{y}^*))-(\overline{v}^*g)^*(\overline{y}^*).\end{aligned}$$

hence $\overline{v}:=(f\circ A+g)(\overline{x})\in V$ has the property $(\overline{v}^*,\overline{y}^*,\overline{v})\in\mathcal{B}$. Moreover, the point $(\overline{v}^*,\overline{y}^*,\overline{v})$ is an efficient solution to (VD). Indeed, if $(\overline{v}^*,\overline{y}^*,\overline{v})$ is not an efficient solution to (VD), then there exists an element $(v^*,y^*,v)\in\mathcal{B}$ such that $(f\circ A+g)(\overline{x})=\overline{v}\leq_K v=h(v^*,y^*,v)$. But this contradicts Theorem 2 and the conclusion follows. ■

Theorem 4 *If $\mathcal{B}$ is nonempty and the regularity condition (RCV) is fulfilled, then $V\setminus \operatorname{cl}((f\circ A+g)(\operatorname{dom} g\cap A^{-1}(\operatorname{dom} f))+K)\subseteq \operatorname{core}(h(\mathcal{B}))$.*

 ISBN: 978-84-614-6167-7

Proof. Let $\overline{v} \in V \setminus \mathrm{cl}((f \circ A + g)(\mathrm{dom}\, g \cap A^{-1}(\mathrm{dom}\, f)) + K)$. Since $\mathrm{cl}((f \circ A + g)(\mathrm{dom}\, g \cap A^{-1}(\mathrm{dom}\, f)) + K) \subseteq V$ is a convex and closed set, by Tuckey separation Theorem (Theorem 2.1.5 from [2]), there exist $\overline{v}^* \in V^* \setminus \{0\}$ and $\alpha \in \mathbb{R}$ such that

$$\langle \overline{v}^*, \overline{v} \rangle < \alpha < \langle \overline{v}^*, v \rangle, \text{ for all } v \in \mathrm{cl}((f \circ A + g)(\mathrm{dom}\, g \cap A^{-1}(\mathrm{dom}\, f)) + K). \quad (5)$$

Obviously, $\overline{v}^* \in K^* \setminus \{0\}$.

On the other hand, from $\mathcal{B} \neq \emptyset$, there exists $(\widehat{v}^*, \widehat{y}^*, \widehat{v}) \in K^{*0} \times Y^* \times V$ such that

$$\langle \widehat{v}^*, \widehat{v} \rangle \leq -(\widehat{v}^* f)^*((A^{-1})^*(-\widehat{y}^*)) - (\widehat{v}^* g)^*(\widehat{y}^*) \leq \inf_{x \in X} \langle \widehat{v}^*, (f \circ A + g)(x) \rangle.$$

If $\gamma := \alpha - \langle \overline{v}^*, \overline{v} \rangle$, then $\gamma > 0$ and

$$\langle s\widehat{v}^* + (1-s)\overline{v}^*, \overline{v} \rangle = \langle \overline{v}^*, \overline{v} \rangle + s(\langle \widehat{v}^*, \overline{v} \rangle - \langle \overline{v}^*, \overline{v} \rangle) = \alpha - \gamma + s(\langle \widehat{v}^*, \overline{v} \rangle - \alpha + \gamma), \quad (6)$$

for all $s \in (0,1)$. Now from (5), we have

$$\langle s\widehat{v}^* + (1-s)\overline{v}^*, v \rangle > s\langle \widehat{v}^*, v \rangle + (1-s)\alpha = \alpha + s(\langle \widehat{v}^*, v \rangle - \alpha), \quad (7)$$

for all $v \in (f \circ A + g)(\mathrm{dom}\, g \cap A^{-1}(\mathrm{dom}\, f))$. Let $\overline{s} \in (0,1)$ such that

$$\overline{s}(\langle \widehat{v}^*, \overline{v} \rangle - \alpha + \gamma) < \gamma/2 \quad (8)$$

and

$$\overline{s}(\langle \widehat{v}^*, \widehat{v} \rangle - \alpha) > -\gamma/2. \quad (9)$$

Since $\underline{v}^* = \overline{s}\widehat{v}^* + (1-\overline{s})\overline{v}^* \in K^{*0}$, from (6) and (8) we deduce that $\langle \underline{v}^*, \overline{v} \rangle < \alpha - \frac{\gamma}{2}$ and from (7) and (9) we obtain that $\alpha - \frac{\gamma}{2} < \langle \underline{v}^*, v \rangle$, for all $v \in (f \circ A + g)(\mathrm{dom}\, g \cap A^{-1}(\mathrm{dom}\, f))$. The last two relations imply

$$\langle \underline{v}^*, \overline{v} \rangle < \inf_{x \in X} \langle \underline{v}^*, (f \circ A + g)(x) \rangle. \quad (10)$$

Now, in view of Theorem 1, there exists $\underline{y}^* \in Y^*$ such that

$$\begin{aligned} \inf_{x \in X} \langle \underline{v}^*, (f \circ A + g)(x) \rangle &= \sup_{y^* \in Y^*} \{-(\underline{v}^* f)^*((A^{-1})^*(-y^*)) - (\underline{v}^* g)^*(y^*)\} \\ &= -(\underline{v}^* f)^*((A^{-1})^*(-\underline{y}^*)) - (\underline{v}^* g)^*(\underline{y}^*). \end{aligned} \quad (11)$$

From (10) and (11) we have that $\varepsilon := \frac{1}{2}\left(-(\underline{v}^* f)^*((A^{-1})^*(-\underline{y}^*)) - (\underline{v}^* g)^*(\underline{y}^*) - \langle \underline{v}^*, \overline{v} \rangle\right) > 0$ has the property that $\langle \underline{v}^*, \overline{v} \rangle + \varepsilon < -(\underline{v}^* f)^*((A^{-1})^*(-\underline{y}^*)) - (\underline{v}^* g)^*(\underline{y}^*)$ and, for each $v \in V$ there exists $\delta > 0$ such that $\langle \underline{v}^*, \overline{v} + \lambda v \rangle \leqq \langle \underline{v}^*, \overline{v} \rangle + \varepsilon < -(\underline{v}^* f)^*((A^{-1})^*(-\underline{y}^*)) - (\underline{v}^g)^*(\underline{y}^*)$, for all $\lambda \in [0, \delta]$. Consequently, for all $\lambda \in [0, \delta]$ we have $(\underline{v}^*, \underline{y}^*, \overline{v} + \lambda v) \in \mathcal{B}$ and $\overline{v} + \lambda v \in h(\mathcal{B})$. Then, $\overline{v} \in \mathrm{core}(h(\mathcal{B}))$. ■

Next follows the proof of the converse duality theorem.

Theorem 5 *If the regularity condition (RCV) is fulfilled and the set* $(f \circ A + g)(\mathrm{dom}\, g \cap A^{-1}(\mathrm{dom}\, g)) + K$ *is closed, then for every efficient solution* $(\overline{v}^*, \overline{y}^*, \overline{v})$ *to* (VD) *there exists* $\overline{x} \in X$ *a properly efficient solution to* (VP), *such that* $(f \circ A + g)(\overline{x}) = h(\overline{v}^*, \overline{y}^*, \overline{v}) = \overline{v}$.

 ISBN: 978-84-614-6167-7

Proof. First we show that $\overline{v} \in (f \circ A + g)(\operatorname{dom} g \cap A^{-1}(\operatorname{dom} f)) + K$. Indeed, if $\overline{v} \notin (f \circ A + g)(\operatorname{dom} g \cap A^{-1}(\operatorname{dom} f)) + K$, then, by Theorem 4, $\overline{v} \in \operatorname{core}(h(\mathcal{B}))$. It follows that for each $k \in K \setminus \{0\}$ there exists $\lambda > 0$ such that $v_\lambda = \overline{v} + \lambda k \in h(\mathcal{B})$ and $v_\lambda \geq_K \overline{v}$, which contradicts the fact that $(\overline{v}^*, \overline{y}^*, \overline{v})$ is an efficient solution to (VD).

Hence $\overline{v} \in (f \circ A + g)(\operatorname{dom} g \cap A^{-1} \operatorname{dom}(f)) + K$. Then there exist $\overline{x} \in \operatorname{dom} g \cap A^{-1}(\operatorname{dom} f)$ and $\overline{k} \in K$ such that $\overline{v} = (f \circ A + g)(\overline{x}) + \overline{k}$. In view of Theorem 3 we have that $\overline{k} = 0$. Now we have

$$\begin{aligned} &\langle \overline{v}^*, (f \circ A + g)(\overline{x}) \rangle = \langle \overline{v}^*, \overline{v} \rangle \\ &\leqq -(\overline{v}^* f)^*((A^{-1})(-\overline{y}^*)) - (\overline{v}^* g)^*(\overline{y}^*) \leq \inf_{x \in X} \langle \overline{v}^*, (f \circ A + g)(x) \rangle \end{aligned}$$

and hence $\overline{x}$ is a properly efficient solution to (VP). ■

Remark 1 *The regularity condition (RCV) can be replaced with the weaker sufficient condition: for all $v^* \in K^{*0}$ one has that $\inf_{x \in X} \langle v^*, (f \circ A + g)(x) \rangle = \sup_{y^* \in Y^*} \{-(\overline{v}^* f)^* ((A^{-1})(-\overline{y}^*)) + (v^* g)^*(\overline{y}^*)\}$ and the supremum is attained. This mean, that for all $v^* \in K^{*0}$ the scalar optimization problem $\operatorname{Min}_{x \in X} \langle v^*, (f \circ A + g)(x) \rangle$ is stable.*

2.2 Duality with respect to weakly efficient solutions

In this subsection we attach to the Problem (VP) the following vector dual problem:

$$\operatorname*{Max}_{(v^*, y^*, v) \in \mathcal{B}^w} h^w(v^*, y^*, v) \qquad (VD^w)$$

where

$$\mathcal{B}^w = \{(v^*, y^*, v) \in (K^* \setminus \{0\}) \times Y^* \times V : \langle v^*, v \rangle \leqq -(v^* f)^*((A^{-1})^*(-y^*)) - (v^* g)^*(y^*)\}$$

and

$$h^w(v^*, y^*, v) = v.$$

In what follows some connections between the weakly efficient solutions to problems (VP) and (VD^w) are given.

An element $\overline{x} \in X$ is a *weakly efficient solution* to (VP) if and only if $\overline{x} \in \operatorname{dom} g \cap A^{-1}(\operatorname{dom} f)$ and $((f \circ A + g)(\overline{x}) - \operatorname{core}(K)) \cap (f \circ A + g)(\operatorname{dom} g \cap A^{-1}(\operatorname{dom} f)) = \emptyset$.

Theorem 6 *There is no $x \in X$ and no $(v^*, y^*, v) \in \mathcal{B}^w$ such that $(f \circ A + g)(x) <_K h^w(v^*, y^*, v)$.*

Proof. We assume the contrary, that there exist $\overline{x} \in X$ and $(\overline{v}^*, \overline{y}^*, \overline{v}) \in \mathcal{B}^w$ such that

$$\overline{v} - (f \circ A + g)(\overline{x}) = h^w(\overline{v}^*, \overline{y}^*, \overline{v}) - (f \circ A + g)(\overline{x}) >_K 0. \qquad (12)$$

Then $\overline{x} \in \operatorname{dom} g \cap A^{-1}(\operatorname{dom} f)$, $\overline{v}^* \in K^* \setminus \{0\}$ and $\langle \overline{v}^*, \overline{v} \rangle \leqq -(\overline{v}^* f)^*((A^{-1})^*(-\overline{y}^*)) - (\overline{v}^* g)^*(\overline{y}^*)\}$. On the other hand, from (12) and $\overline{v}^* \in K^* \setminus \{0\}$ we deduce that $\langle \overline{v}^*, \overline{v} \rangle > \langle \overline{v}^*, (f \circ A)(\overline{x}) \rangle + \langle \overline{v}^*, g(\overline{x}) \rangle$. Using the same idea as in the proof of Theorem 2 we obtain a contradiction. ■

 ISBN: 978-84-614-6167-7

Theorem 7 *If the regularity condition (RCV) is fulfilled and $\overline{x} \in X$ is a weakly efficient solution to (VP), then there exists a weakly efficient solution $(\overline{v}^*, \overline{y}^*, \overline{v})$ to (VD^w) such that $(f \circ A + g)(\overline{x}) = h(\overline{v}^*, \overline{y}^*, \overline{v}) = \overline{v}$.*

Proof. If $\overline{x} \in X$ is a weakly efficient solution to (VP), then $\overline{x} \in \operatorname{dom} g \cap A^{-1}(\operatorname{dom} f)$ and $(f \circ A + g)(\overline{x})$ is a weakly minimal element of the set $(f \circ A + g)(\operatorname{dom} g \cap A^{-1}(\operatorname{dom} f)) \subseteq V$. As $(f \circ A + g)(\operatorname{dom} g \cap A^{-1}(\operatorname{dom} f)) + K$ is a nonempty convex set, then using Theorem 4.1.7 from [2] there exists $\overline{v}^* \in K^* \setminus \{0\}$, such that

$$\langle \overline{v}^*, (f \circ A + g)(\overline{x}) \rangle = \inf_{x \in X} \{(\overline{v}^* f)(Ax) + (\overline{v}^* g)(x)\}. \tag{13}$$

It follows that $\overline{x}$ is a solution to Problem (SP). Then by strong duality theorem (Theorem 1) applied to the pair of scalar problems $(SP) - (DSP)$ we have

$$\inf_{x \in X} \{(\overline{v}^* f)(Ax) + (\overline{v}^* g)(x)\} = \sup_{y^* \in Y^*} \{-(\overline{v}^* f)^*((A^{-1})^*(-y^*)) - (\overline{v}^* g)^*(y^*)\}, \tag{14}$$

and (DSP) has a solution $\overline{y}^* \in Y^*$, i.e. we have

$$\sup_{y^* \in Y^*} \{-(\overline{v}^* f)^*((A^{-1})^*(-y^*)) - (\overline{v}^* g)^*(y^*)\} = -(\overline{v}^* f)^*((A^{-1})^*(-\overline{y}^*)) - (\overline{v}^* g)^*(\overline{y}^*). \tag{15}$$

Then

$$\begin{aligned} \langle \overline{v}^*, (f \circ A + g)(\overline{x}) \rangle &\overset{(13)}{=} \inf_{x \in X} \{(\overline{v}^* f)(Ax) + (\overline{v}^* g)(x)\} \\ &\overset{(14)}{=} \sup_{y^* \in Y^*} \{-(\overline{v}^* f)^*((A^{-1})^*(-y^*)) - (\overline{v}^* g)^*(y^*)\} \\ &\overset{(15)}{=} -(\overline{v}^* f)^*((A^{-1})^*(-\overline{y}^*)) - (\overline{v}^* g)^*(\overline{y}^*). \end{aligned}$$

hence $\overline{v} := (f \circ A + g)(\overline{x}) \in V$ has the property $(\overline{v}^*, \overline{y}^*, \overline{v}) \in \mathcal{B}^w$. Moreover, the point $(\overline{v}^*, \overline{y}^*, \overline{v})$ is an weakly efficient solution to (VD^w). Indeed, if $(\overline{v}^*, \overline{y}^*, \overline{v})$ is not an weakly efficient solution to (VD^w), then there exists an element $(v^*, y^*, v) \in \mathcal{B}^w$ such that $(f \circ A + g)(\overline{x}) = \overline{v} <_K v = h(v^*, y^*, v)$. But this contradicts Theorem 6 and the conclusion follows. ■

For proving the converse duality theorem, the next theorem is given.

Theorem 8 *If the regularity condition (RCV) is fulfilled, then $V \setminus \operatorname{cl}((f \circ A + g)(\operatorname{dom} g \cap A^{-1}(\operatorname{dom} f)) + K) \subseteq \operatorname{core}(h^w(\mathcal{B}^w))$.*

Proof. Let $\overline{v} \in V \setminus \operatorname{cl}((f \circ A + g)(\operatorname{dom} g \cap A^{-1}(\operatorname{dom} f)) + K)$. The set $\operatorname{cl}((f \circ A + g)(\operatorname{dom} g \cap A^{-1}(\operatorname{dom} f)) + K) \subseteq V$ is convex and closed, then there exist $\overline{v}^* \in K^* \setminus \{0\}$ and $\alpha \in \mathbb{R}$ such that $\langle \overline{v}^*, \overline{v} \rangle < \alpha < \inf_{x \in X} \{(\overline{v}^* f)(Ax) + (\overline{v}^* g)(x)\}$. By strong duality theorem applied to the pair of scalar problems $(SP) - (DSP)$, it follows that there exists $\overline{y}^* \in Y^*$ such that $\inf_{x \in X} \{(\overline{v}^* f)(Ax) + (\overline{v}^* g)(x)\} = -(\overline{v}^* f)((A^{-1})^*(-\overline{y}^*)) - (\overline{v}^* g)(\overline{y}^*)$.
Like in the proof of Theorem 4 one can obtain that $\overline{v} \in \operatorname{core}(h^w(\mathcal{B}^w))$. ■

Theorem 9 *If the regularity condition (RCV) is fulfilled and the set $(f \circ A + g)(\operatorname{dom} g \cap A^{-1}(\operatorname{dom} g)) + K$ is closed, then for every weakly efficient solution $(\overline{v}^*, \overline{y}^*, \overline{v})$ to (VD^w) one has that $\overline{v}$ is a weakly efficient solution to Problem (VP).*

 ISBN: 978-84-614-6167-7

Proof. We have $\overline{v} \in (f \circ A + g)(\operatorname{dom} g \cap A^{-1}(\operatorname{dom} g)) + K$. Indeed, if $\overline{v} \notin (f \circ A + g)(\operatorname{dom} g \cap A^{-1}(\operatorname{dom} g)) + K$, by Theorem 8 we have that $\overline{v} \in \operatorname{core}(h^w(\mathcal{B}^w))$. It follows that, for each element $k \in \operatorname{int}(K)$ there exists $\lambda > 0$ such that $v_\lambda = v + \lambda k \in h^w(\mathcal{B}^w)$ and $v_\lambda >_K \overline{v}$. But this contradicts the fact that $(\overline{v}^*, \overline{y}^*, \overline{v})$ is a weakly efficient solution to (VD^w). Consequently, we must have $\overline{v} \in (f \circ A + g)(\operatorname{dom} g \cap A^{-1}(\operatorname{dom} g)) + K$. If $\overline{v}$ is not a weakly efficient solution to (VP) there exist $x \in X$ and $k \in K$ such that $(f \circ A + g)(x) \leqq_K (f \circ A + g)(x) + k <_K \overline{v}$. A contradiction follows if we apply Theorem 6. Obviously, $\overline{v}$ is a weakly efficient solution to Problem (VP). ■

3 Wolfe and Mond-Weir type vector duals

In this section, to Problem (VP) we attach the vector perturbation function $\Phi : X \times Y \to \overline{V}$ defined by $\Phi(x, y) = (f \circ A)(x) + g(x + y)$ for all $(x, y) \in X \times Y$. The conjugate function of Φ is the function $\Phi^* : X^* \times Y^* \to \overline{V}$ defined by $\Phi^*(x^*, y^*) = f^*\left(\left(A^{-1}\right)^*(x^* - y^*)\right) + g^*(y^*)$ for all $(x^*, y^*) \in X^* \times Y^*$. In this manner we consider the following three vector duals (see [3]) :

Problem (DV_W) :

$$\operatorname*{Max}_{(v^*, y^*, u, y, r) \in \mathcal{B}_W} h_W(v^*, y^*, u, y, r) \tag{DV_W}$$

where

$$\mathcal{B}_W = \{(v^*, y^*, u, y, r) \in K^{*0} \times Y^* \times X \times Y \times (K \setminus \{0\}) : (0, y^*) \in \partial(v^*\Phi)(u, y)\}$$

and

$$h_W(v^*, y^*, u, y, r) = \Phi(u, y) - \frac{\langle y^*, y\rangle}{\langle v^*, r\rangle} r.$$

Problem (DV_M) :

$$\operatorname*{Max}_{(v^*, y^*, u) \in \mathcal{B}_M} h_M(v^*, y^*, u) \tag{DV_M}$$

where

$$\mathcal{B}_M = \{(v^*, y^*, u) \in K^{*0} \times Y^* \times X : (0, y^*) \in \partial(v^*\Phi)(u, 0)\}$$

and

$$h_M(v^*, y^*, u) = \Phi(u, 0),$$

Problem (DV_{W^r}):

$$\operatorname*{Max}_{(v^*, y^*, u, y) \in \mathcal{B}^G_{W^r}} h_{W^r}(v^*, y^*, u, y) \tag{DV^r_W}$$

where

$$\mathcal{B}_{W^r} = \{(v^*, y^*, u, y) \in K^{*0} \times Y^* \times X \times Y : (0, y^*) \in \partial(v^*\Phi)(u, y),\ \langle v^*, r\rangle = 1\}$$

and

$$h_{W^r}(v^*, y^*, u, y) = \Phi(u, y) - \langle y^*, y\rangle r,$$

with $r \in K \setminus \{0\}$.

 ISBN: 978-84-614-6167-7

For $v^* \in K^{*0}, y^* \in Y^*, u \in X\ y \in Y$ one has $(0, y^*) \in \partial(v^*\Phi)(u, y)$ if and only if $(v^*\Phi)^*(0, y^*) + (v^*\Phi)(u, y) = \langle y^*, y\rangle$. This is equivalent to $(v^*f)^*((A^{-1})^*(-y^*)) + (v^*g)^*(y^*) + (f \circ A)(u) + g(u + y) = \langle y^*, y\rangle$. Using the Young-Fenchel inequality, the condition $(0, y^*) \in \partial(v^*\Phi)(u, y)$ is true if and only if $y^* \in \partial(v^*g)(u + y)$ and $(A^{-1})^*(-y^*) \in \partial(v^*f)(Au)$. Then, the vector duals to (VP) follows:

Problem (DV_W) :

$$\underset{(v^*,y^*,u,y,r)\in\overline{\mathcal{B}}_W}{\text{Max}}\ \overline{h}_W(v^*, y^*, u, y, r) \qquad (\overline{DV}_W)$$

where

$$\begin{aligned}\overline{\mathcal{B}}_W = \ & \{(v^*, y^*, u, y, r) \in K^{*0} \times Y^* \times X \times Y \times (K \setminus \{0\}) : \\ & y^* \in A^*(-\partial(v^*f)(Au)) \cap \partial(v^*g)(u + y)\}\end{aligned}$$

and

$$\overline{h}_W(v^*, y^*, u, y, r) = (f \circ A)(u) + g(u + y) + \frac{\langle y^*, y\rangle}{\langle v^*, r\rangle}r,$$

Problem (DV_M) :

$$\underset{(v^*,u)\in\overline{\mathcal{B}}_M}{\text{Max}}\ \overline{h}_M(v^*, u) \qquad (\overline{DV}_M)$$

where

$$\overline{\mathcal{B}}_M = \ \{(v^*, u) \in K^{*0} \times X : 0 \in A^*(\partial(v^*f)(Au)) + \partial(v^*g)(u))\}$$

and

$$\overline{h}_M(v^*, u) = (f \circ A)(u) + g(u),$$

Problem (DV_{W^r}):

$$\underset{(v^*,y^*,u,y)\in\overline{\mathcal{B}}_{W^r}}{\text{Max}}\ \overline{h}_{W^r}(v^*, y^*, u, y) \qquad (\overline{DV}^r_W)$$

where

$$\begin{aligned}\overline{\mathcal{B}}_{W^r} = \ & \{(v^*, y^*, u, y) \in K^{*0} \times Y^* \times X \times Y : \langle v^*, r\rangle = 1, \\ & y^* \in A^*(-\partial(v^*f)(Au)) \cap \partial(v^*g)(u + y)\}\end{aligned}$$

and

$$\overline{h}_{W^r}(v^*, y^*, u, y) = (f \circ A)(u) + g(u + y) - \langle y^*, y\rangle r$$

with $r \in K \setminus \{0\}$.

Next we obtain the weak and strong duality statements.

Theorem 10 *There are no $x \in X$ and $(v^*, y^*, u, y, r) \in \overline{\mathcal{B}}_W$ such that $(f \circ A)(x) + g(x) \leq_K \overline{h}_W(v^*, y^*, u, y, r)$.*

Theorem 11 *There are no $x \in X$ and $(v^*, u) \in \overline{\mathcal{B}}_M$ such that $(f \circ A)(x) + g(x) \leq_K \overline{h}_M(v^*, u)$.*

Theorem 12 *Let $r \in K \setminus \{0\}$. There are no $x \in X$ and $(v^*, y^*, u, y) \in \overline{\mathcal{B}}_{W^r}$ such that $(f \circ A)(x) + g(x) \leq_K \overline{h}_{W^r}(v^*, y^*, u, y)$.*

 ISBN: 978-84-614-6167-7

In the formulation of strong duality convexity there are needed assumptions which guarantee the $K-$ convexity of the vector perturbation function and the regularity conditions given in [8] obtained by particularizing the classical ones from [2].

$$\left| \begin{array}{l} X \text{ and } Y \text{ are Fréchet spaces, } f \text{ and } g \text{ are } K-\text{lower semicontinuous} \\ \text{and } 0 \in \operatorname{sqri}(\operatorname{dom} g - A^{-1}(\operatorname{dom} f)), \end{array} \right. \qquad (RCV^1)$$

$(RCV^{1'})$ where sqri is replaced by core, $(RCV^{1''})$ where sqri is replaced by int,

$$\left| \ \dim(\operatorname{lin}(\operatorname{dom} g - A^{-1}(f))) < +\infty, 0 \in \operatorname{ri}(\operatorname{dom} g - A^{-1}(\operatorname{dom} f)) \right. \qquad (RCV^2)$$

and

$$\left| \begin{array}{l} f \text{ and } g \text{ are } K-\text{lower semicontinuous and } \operatorname{epi} f^* \circ (A^*)^{-1} + \operatorname{epi} g^* \\ \text{is closed in the topology } w(X^*, X) \times \mathbb{R}, \text{ for all } v^* \in K^{*0}. \end{array} \right. \qquad (RCV^3)$$

Theorem 13 *Let $\overline{r} \in K \setminus \{0\}$. Assume that f and g are $K-$convex vector functions and one of the regularity conditions (RCV), (RCV^i), $i \in \{1,2,3\}$ is fulfilled. If $\overline{x}$ is a properly efficient solution to (VP), then there exist $\overline{v}^* \in K^{*0}$ and $\overline{y}^* \in Y^*$ such that $(\overline{v}^*, \overline{y}^*, \overline{x}, 0, \overline{r})$ is an efficient solution to $(\overline{DV}_W)$, $(\overline{v}^*, \overline{x})$ is an efficient solution to $(\overline{DV}_M)$, $(\overline{v}^*, \overline{y}^*, \overline{x}, 0)$ is an efficient solution to $(\overline{DV}_{W^{\overline{r}}})$ and $(f \circ A)(\overline{x}) + g(\overline{x}) = \overline{h}_W(\overline{v}^*, \overline{y}^*, \overline{x}, 0, \overline{r}) = \overline{h}_M(\overline{v}^*, \overline{x}) = \overline{h}_{W^{\overline{r}}}(\overline{v}^*, \overline{y}^*, \overline{x}, 0)$.*

Remark 2 *In case $V = \mathbb{R}$ and $K = \mathbb{R}_+$, by taking the functions $f : Y \to \overline{\mathbb{R}}$ and $g : X \to \overline{\mathbb{R}}$ proper, we discover the Wolfe and Mond-Weir duality schemes for the unconstrained scalar optimization problem (P) mentioned above.*

4 Conclusions

This study starts with a scalar problem for which we formulate the corresponding vector problem. To this problem we attach vector duals using the method of linear scalarization and we give duality results with respect to properly efficient solutions and weakly efficient solutions. Moreover, we investigate the Wolfe and Mond-Weir type vector duals for our vector optimization problem and formulate some duality results referring to efficient and properly efficient solutions.

Acknowledgements

The author Emilia-Loredana Pop, Phd student, wishes to thank for the financial support provided from programs co-financed by The Sectoral Operational Program for Human Resources Development, Contract POSDRU /88/1.5/S/60185− "Innovative doctoral studies in a knowledge based society".

References

[1] R.I. Boţ, *Conjugate duality in convex optimization*, Lecture Notes in Economics and Mathematical Systems 637, Springer-Verlag, Berlin Heidelberg, 2010.

[2] R.I. Boţ, S.-M. Grad, G. Wanka, *Duality in vector optimization*, Springer-Verlag, Berlin Heidelberg, 2009.

[3] R. I. Boţ, S.-M. Grad, *Extending the classical vector Wolfe and Mond-Weir duality concepts via perturbations*, 2010, to appear.

[4] R. I. Boţ, S.-M. Grad, *Wolfe duality and Monde-Weir duality via perturbation*, Nonlinear Analysis: Theory, Methods and Applications **73(2)** (2010) 374–384.

[5] F. H. Clarke, *Optimization and nonsmooth analysis*, John Wiley and Sons Inc, New York, A Wiley-Interscience Publication, 1983.

[6] B.S. Mordukhovich, *Variational Analysis and Generalized Differentiation, Vol. I: Basic Theory. Grundlehren der mathematischen Wissenschaften (A series of Comprehensive Studies in Mathematics 330)*, Berlin, 2006.

[7] B.S. Mordukhovich, *Variational Analysis and Generalized Differentiation, Vol. II: Applications. Grundlehren der mathematischen Wissenschaften (A series of Comprehensive Studies in Mathematics 330)*, Berlin, 2006.

[8] E.-L. Pop, D.I. Duca, *On optimization problems with composed functions*, 2011, to appear.

[9] R.T. Rockafellar, *Convex analysis*, Princeton University Press, Princeton, 1970.

[10] T. Weir, *Proper efficiency and duality for vector valued optimization problems*, Journal of Australian Mathematical Society Series A **43(1)** (1987) 21–34.

[11] T. Weir, *On efficiency, proper efficiency and duality in multiobjective programming*, Asia-Pacific Journal of Operational Research **7(1)** (1990) 46–54.

[12] T. Weir, B. Mond, *Multiple objective programming duality without a constraint qualification*, Utilitas Mathematica **39** (1991) 41–55.

[13] C. Zălinescu, *Convex analysis in general vector spaces*, World Scientific, 2002.

 ISBN: 978-84-614-6167-7

Proceedings of the 11th International Conference on Computational and Mathematical Methods in Science and Engineering, CMMSE 2011 26–30 June 2011.

Comparison via stability regions of the Störmer-Cowell and Falkner methods in predictor-corrector mode

Higinio Ramos[1] **and Cesáreo Lorenzo**[1]

[1] *Departamento de Matemática Aplicada, Universidad de Salamanca*

emails: higra@usal.es, cesareo@usal.es

Abstract

The Störmer-Cowell and the Falkner methods are usually used in the numerical solution of initial-value problems of the special second order differential equation $y'' = f(x, y)$. In this paper we compare the stability intervals and the stability regions of the Störmer-Cowell and Falkner methods in predictor-corrector mode. Through this analysis it is observed that Falkner methods can be considered as an alternative to Störmer methods. The primary stability intervals for both methods are non-empty for alternating pairs of values of the number of steps, specifically for $k = 4, 5, 8, 9, 12, 13$, being the stability regions of the Falkner methods larger than that of the Störmer methods.

Key words: absolute stability regions, Störmer-Cowell methods, Falkner methods, special second order initial-value problem

MSC 2000: 65L05, 65L20

1 Introduction

Second-order differential equations appear frequently in applied sciences. Examples of that are the mass movement under the action of a force, problems of orbital dynamics, or in general, any problem involving Newton's law.

Among the general procedures for direct integration of the so-called *special second-order* initial value problem (I.V.P.)

$$y''(x) = f(x, y(x)), \quad y(x_0) = y_0, \quad y'(x_0) = y'_0\,, \tag{1}$$

the Störmer-Cowell methods [10, 11] or the Falkner methods [5] are well-known classes of schemes that may be used for this purpose.

Although it is possible to integrate a second-order I.V.P. by reducing it to a first-order system and applying one of the methods available for such systems, it seems more natural to provide numerical methods in order to integrate the problem directly.

 ISBN: 978-84-614-6167-7

The advantage of this procedure lies in the fact that they are able to exploit special information about ODES, resulting in an increase in efficiency. For instance, it is well-known that Runge-Kutta-Nyström methods for (1) involve a real improvement as compared to standard Runge-Kutta methods for a given number of stages ([9], p. 285), although the computational cost remains high because of the number of function evaluations. On the other hand, a linear k-step method for first-order ODEs becomes a $2k$-step method for (1), ([9], p. 461), increasing the computational work.

The k-step Störmer method with constant stepsize h may be written in the form (see [9], p. 413 or [10], p. 291)

$$y_{n+1} - 2y_n + y_{n-1} = h^2 \sum_{j=0}^{k-1} \sigma_j \bigtriangledown^j f_n \,, \tag{2}$$

where the y_i are the numerical approximations for the true values $y(x_i)$ with $x_i = x_0 + ih$, and $f_i = f(x_i, y_i)$ for $i = n-(k-1), \ldots, n$ and $\bigtriangledown^j f_n$ is the standard notation for the backward differences. The coefficients σ_j may be easily obtained using the method of generating functions (see [10], p. 291), being

$$G_\sigma(t) = \sum_{j=0}^{\infty} \sigma_j \, t^j = \frac{t^2}{(1-t)(Log(1-t))^2} \,.$$

The k-step Störmer method has order k and error constant σ_k (except for $k = 1$, which has order 2 being the error constant σ_2, (see [11], p. 71)).

The corresponding implicit formulas are named Cowell methods and read

$$y_{n+1} - 2y_n + y_{n-1} = h^2 \sum_{j=0}^{k} \sigma_j^* \bigtriangledown^j f_{n+1} \,, \tag{3}$$

where the coefficients σ_j^* may be obtained by means of the generating function

$$G_{\sigma^*}(t) = \sum_{j=0}^{\infty} \sigma_j^* \, t^j = \frac{t^2}{(Log(1-t))^2} \,.$$

The implicit k-step method has order $k+1$ and error constant σ_{k+1}^* (except when $k = 2$ for which we obtain the Numerov method, with order 4 and error constant σ_4^*, (see [11], p. 71)).

The explicit Falkner method of k steps consists in two formulas that can be written in the form [4]

$$y_{n+1} = y_n + h\, y'_n + h^2 \sum_{j=0}^{k-1} \beta_j \bigtriangledown^j f_n \,, \tag{4}$$

$$y'_{n+1} = y'_n + h \sum_{j=0}^{k-1} \gamma_j \bigtriangledown^j f_n \,, \tag{5}$$

 ISBN: 978-84-614-6167-7

where the y'_i are approximations to the true values of the derivative at x_i. The coefficients β_j and γ_j can be obtained using respectively the generating functions

$$G_\beta(t) = \sum_{j=0}^{\infty} \beta_j\, t^j = \frac{t+(1-t)\,Log(1-t)}{(1-t)Log^2(1-t)}\,,$$

$$G_\gamma(t) = \sum_{j=0}^{\infty} \gamma_j\, t^j = \frac{-t}{(1-t)Log(1-t)}\,,$$

which may be obtained similarly as that for the Störmer or Cowell methods (see [10], p. 291).

The implicit Falkner method of k steps consists of two formulas that may be written as [4]

$$y_{n+1} = y_n + h\, y'_n + h^2 \sum_{j=0}^{k} \beta^*_j \bigtriangledown^j f_{n+1}\,, \tag{6}$$

$$y'_{n+1} = y'_n + h \sum_{j=0}^{k} \gamma^*_j \bigtriangledown^j f_{n+1}\,, \tag{7}$$

with generating functions for the coefficients given respectively by

$$G_{\beta^*}(t) = \sum_{j=0}^{\infty} \beta^*_j\, t^j = \frac{t+(1-t)\,Log(1-t)}{Log^2(1-t)}\,,$$

$$G_{\gamma^*}(t) = \sum_{j=0}^{\infty} \gamma^*_j\, t^j = \frac{-t}{Log(1-t)}\,.$$

Note that the formulas in (5) and (7) are respectively the Adams-Bashforth and Adams-Moulton schemes for the problem $(y')' = f(x,y)$, which are used to follow the values of the derivative.

All the above formulas are of multistep type, specifically k-step formulas, and so k initial values must be provided in order to proceed with the methods (the Runge-Kutta methods are commonly used to obtain the starting values). In this paper, the above methods will be treated in predictor-corrector mode (P-C) where the predictor is used only once.

The formulas in (2-3) may be used in predictor-corrector mode for solving the problem in (1) [12, 6] similarly as the Adams methods for first order initial-value problems.

The implicit formulas in (6-7) may also be used to produce methods for solving the I.V.P. in (1). A well-known procedure of this type is the Wilson method [17], which is one of the Newmark family, and is commonly used in molecular dynamics calculations. This method uses the two-step formula in (6), given by

$$y_{n+1} = y_n + h\, y'_n + \frac{h^2}{6}\,\left(2y''_n + y''_{n+1}\right) \tag{8}$$

 ISBN: 978-84-614-6167-7

to obtain the positions, while the formula to update the velocities is the two-step method in (7) given by

$$y'_{n+1} = y'_n + \frac{h}{2}\left(y''_n + y''_{n+1}\right). \tag{9}$$

The application of this procedure results in an implicit system that must be solved at each step, involving a great computational cost, but in practice, an explicit formulation in predictor-corrector mode is frequently used. In this way the implicit methods in (6-7) are adequately combined with the explicit ones in (4-5) so as to avoid having to solve an algebraic system at each step. This P-C formulation may also be used as an starting method, as for example in [3], where the one step implicit Falkner method in predictor-corrector mode is used to provide the starting values in the application of the De Vogelaere's method. Others examples of such procedures may be found in [1], [8] or [16].

The paper is organized as follows. In the following section we present different implementations of the Störmer-Cowell and Falkner methods. The analysis of the explicit Falkner methods has been done in [15] and so here the analysis of the implicit Falkner formula in P-C mode will be done. Section 3 deals with the stability analysis, which results of vital importance in the application of the methods, and a table with the stability intervals for both methods is presented. In Section 4 we extend the stability analysis and present the regions of absolute stability. In the final section some conclusions put an end to the article.

2 Methods in P-C mode

In the application of P-C modes, P indicates the application of the predictor, in our case the corresponding method given by the explicit formula, and C indicates the application of the corrector, that is, the corresponding implicit formula. In case of Falkner methods we need a formula to follow the derivative, and so we will use C' to indicate the application of the implicit formula in (7). Finally, E refers to the evaluation of the function f. We have considered two methods, the Störmer-Cowell in PECE mode and the Falkner methods in PECEC' mode. They are summarized in what follows.

2.1 Störmer-Cowell methods in PECE mode

The usual implementation of the Störmer-Cowell method in P-C mode for solving the problem in (1) on each step is as follows

1. Evaluate y_{n+1} using the formula in (2)
2. Evaluate $f_{n+1} = f(x_{n+1}, y_{n+1})$
3. Evaluate y_{n+1} using the formula in (3)
4. Evaluate $f_{n+1} = f(x_{n+1}, y_{n+1})$

 ISBN: 978-84-614-6167-7

2.2 Falkner methods in PECEC' mode

The implementation of the Falkner methods in P-C mode for solving the problem in (1) consists in the following steps

1. Evaluate y_{n+1} using the formula in (4)
2. Evaluate $f_{n+1} = f(x_{n+1}, y_{n+1})$
3. Evaluate y_{n+1} using the formula in (6)
4. Evaluate $f_{n+1} = f(x_{n+1}, y_{n+1})$
5. Evaluate y'_{n+1} using the formula in (7)

Note that in the above formulations, once we have obtained the latest value for y_{n+1} we evaluate $f_{n+1} = f(x_{n+1}, y_{n+1})$, which will be used in the next step. This evaluation is indicated by the final E for the different modes.

Obviously, many more choices could have been considered, taking methods with different steps and therefore, with different orders. Or you might have considered using further corrections, obtaining respectively methods of the form $P(EC)^nE$ or $P(EC)^nEC'$. We have done different experiments, and in the latter case the interval of stability changes, resulting in an interval slightly higher, but with so little difference that the computational effort is not worth. We have considered only P-C methods with the same number of steps in all formulas, where it is performed only a correction, as is usually done in Adams methods in P-C mode (see [13]).

3 Stability

In the context of ordinary differential equations, the concept of stability refers to what extent a numerical scheme is appropriate for solving an initial-value problem. Roughly speaking, a given method can be said to be stable if small changes in the data result in small changes in the solution obtained.

A procedure commonly used to study stability (zero-stability) consists in writing the difference equations of the method as a one-step recurrence in a space with high dimension and in an adequate norm to bound the finite powers of the resulting matrices. For the different combinations of the above explicit and implicit methods to get the P-C modes, a similar procedure to that in [15] may be considered to obtain zero-stability. But, the zero stability is only a minimal condition for a numerical method; for second order equations the so-called P-stability is the type of concern together with A-stability.

In order to determine whether a numerical method will produce reasonable results with a given value of $h > 0$, we need a notion of stability that is different from zero-stability. The stability properties are analyzed by using the linear test equation introduced by Lambert and Watson [14]

$$y''(x) = -\mu^2 y(x)\,, \quad \text{with } \mu > 0\,. \tag{10}$$

 ISBN: 978-84-614-6167-7

When the method described in subsection (2.1) is applied to this problem we obtain the following recursion

$$Y_n = A\,Y_{n-1}\,, \tag{11}$$

where Y_n is the k-vector given by $Y_n = \left(y_{n+1}, y_n, \ldots, y_{n-(k-2)}\right)^T$, and A is a $k \times k$ matrix whose elements depends on H^2 and the coefficients of the Störmer and Cowell methods, being $H = \mu\, h$. The matrix A is called the *stability matrix*.

The entries of this stability matrix are given by

$$\begin{aligned}
A_{1\,1} &= 2 + H^2 \sum_{j=0}^{k} \left(j - 2 + H^2 \sum_{s=0}^{k-1} \sigma_s \right) \sigma_j^* \\
A_{1\,2} &= -1 + H^2 \sum_{j=0}^{k} \left(1 - H^2 \sum_{s=0}^{k-1} s\sigma_s - \frac{j(j-1)}{2} \right) \sigma_j^*\,, \\
A_{1\,l} &= (-1)^{l-1} H^2 \sum_{j=0}^{k} \left(\binom{j}{l} + H^2 \sum_{s=l-1}^{k-1} \binom{s}{l-1} \sigma_s \right) \sigma_j^*\,, \\
&\qquad l = 3, \ldots, k\,. \\
A_{j\,l} &= \delta_{j\,l+1}\,, \qquad j = 2, \ldots, k \quad l = 1, \ldots, k\,,
\end{aligned}$$

where $\delta_{j\,l+1}$ is the Kronecker delta and the notations of type $\binom{s}{l-1}$ here and in the sequel refer to the binomial coefficients.

As an example, in case $k = 4$ the equation in (11) results in

$$Y_n = \begin{pmatrix} \frac{133H^4 - 1452H^2}{1440} + 2 & \frac{12H^2 - 19H^4}{576} - 1 & \frac{19H^4 - 12H^2}{720} & \frac{12H^2 - 19H^4}{2880} \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix} Y_{n-1}\,.$$

The application of the predictor-corrector formulation of the Falkner methods described in subsection (2.2) to the problem in (10) yields the following recursion

$$\overline{Y}_n = B\,\overline{Y}_{n-1}\,, \tag{12}$$

where $\overline{Y}_n$ is the $k+1$-vector given by $\overline{Y}_n = \left(y_{n+1}, y_n, \ldots, y_{n-(k-2)}, h\,y'_{n+1}\right)^T$, and B is a $(k+1) \times (k+1)$ matrix whose elements depends on H^2 and the coefficients of the methods, with $H = \mu\, h$. The matrix B is now the corresponding *stability matrix*.

The entries of the stability matrix B are given by

$$B_{1\,1} \qquad = \quad 1 + H^2 \sum_{j=0}^{k} \left(j - 1 + H^2 \sum_{s=0}^{k-1} \beta_s \right) \beta_j^*$$

 ISBN: 978-84-614-6167-7

$$
\begin{aligned}
B_{1\,l+1} &= (-1)^l H^2 \sum_{j=0}^{k} \left(\binom{j}{l+1} + H^2 \sum_{s=l}^{k-1} \binom{s}{l} \beta_s \right) \beta_j^* \,, \\
&\quad l = 1, \ldots, k-1 \,. \\
B_{j\,l} &= \delta_{j\,l+1} \,, \quad j = 2, \ldots, k \quad l = 1, \ldots, k+1 \,. \\
B_{k+1\,1} &= H^2 \sum_{j=0}^{k} \left[j - 1 + H^2 \sum_{s=0}^{k} \left(1 - s - H^2 \sum_{t=0}^{k-1} \beta_t \right) \beta_s^* \right] \gamma_j^* \\
B_{k+1\,l+1} &= (-1)^l H^2 \sum_{j=0}^{k} \left[\binom{j}{l+1} - H^2 \sum_{t=0}^{k} \left(\binom{t}{l+1} + H^2 \sum_{s=l}^{k-1} \binom{s}{l} \beta_s \right) \beta_t^* \right] \gamma_j^* \,, \\
&\quad l = 1, \ldots, k-1 \,. \\
B_{k+1\,k+1} &= 1 - H^2 \sum_{j=0}^{k} \left(1 - H^2 \sum_{s=0}^{k} \beta_s^* \right) \gamma_j^* \,,
\end{aligned}
$$

with similar comments as before for the Kronecker delta and the binomial coefficients.

For example, when $k = 3$ the equation in (12) reads $\overline{Y}_n = B\,\overline{Y}_{n-1}$ with the matrix B given by

$$
\begin{pmatrix}
\frac{19\left(19H^2-132\right)H^2}{4320} + 1 & \frac{H^2}{10} - \frac{19H^4}{432} & \frac{H^2\left(19H^2-28\right)}{1440} & 1 - \frac{19H^2}{180} \\
1 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 \\
\frac{2508H^4-361H^6-13440H^2}{11520} & \frac{95H^6-216H^4+1200H^2}{5760} & \frac{28H^4-19H^6-160H^2}{3840} & \frac{19H^4-180H^2}{480} + 1
\end{pmatrix} .
$$

The behaviour of the numerical solution will depend on the eigenvalues of the stability matrix M (either A or B in each case), and the stability properties of the method will be characterized by the spectral radius $\rho(M)$. According to the terminology introduced by Coleman and Ixaru [2] we have:

- $(0, H_s)$ is an interval of stability for a given method if for all $H \in (0, H_s)$ it is $|r_i| < 1$ where the $r_i, i = 1, \ldots, k+1$ are the eigenvalues of the stability matrix.
- $(0, H_p)$ is an interval of periodicity for a given method if for all $H \in (0, H_p)$ the eigenvalues of the stability matrix, $r_i, i = 1, \ldots, k+1$, satisfy

  $$r_1 = e^{i\theta(H)} \,, \quad r_2 = e^{-i\theta(H)} \,, \quad |r_i| \leq 1 \,,\ i > 2 \,,$$

  where $\theta(H)$ is a real function.

In particular, if the interval of stability is $(0, \infty)$ the method is *A-stable*, and if the interval of periodicity is $(0, \infty)$ the method is *P-stable*.

The practical significance of the interval of stability is that, for a given μ in (10), there is no explosion of the error in the numerical solution when $0 < h < H_s/\mu$ [2].

 ISBN: 978-84-614-6167-7

The interval of periodicity defines the stepsize which can be used in order for the approximation of the solution of problems with high oscillatory or periodic solution to be of the same order as the algebraic order of the method. When $0 < h < H_p/\mu$ the numerical solutions defined by (11) or (12) are also periodic, as is the exact solution of the test model (10) for all non-trivial initial-conditions on y and y'.

A crucial difference between A-stability and P-stability is that for A-stable methods the stability matrix satisfies $(M)^n \to 0$ as $n \to \infty$ because $\rho(M) < 1$, but for P-stable methods this fact is not possible because $\rho(M) = 1$ [7]. Therefore, A-stable methods alleviate the initial errors whereas for P-stable methods the initial errors do not diminish when the integration progresses in time.

For the methods presented in the previous sections up to $k = 14$ we have obtained that they are not A-stable nor P-stable. There exists only one interval of periodicity for each method, when $k = 2$ in case of the Störmer-Cowell method this interval is $[0, \sqrt{12}]$, and for $k = 1$ in case of the Falkner methods the interval is $[0, \sqrt{6}]$. Figure 1 shows the corresponding absolute values of the roots of the characteristics polynomials and the periodicity intervals.

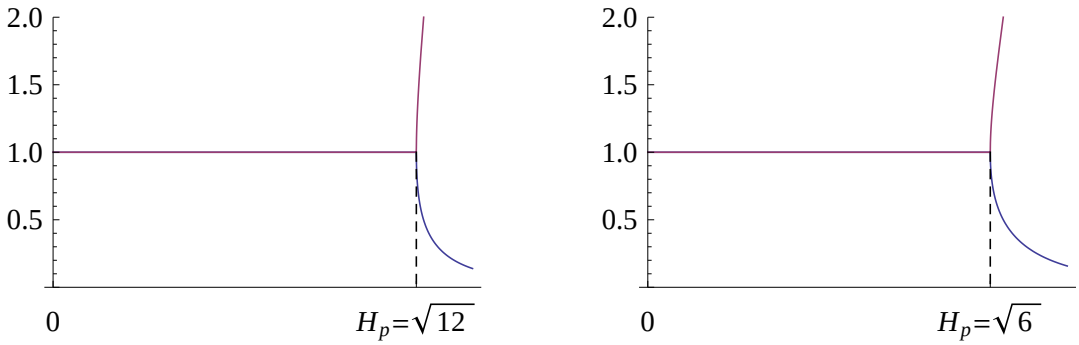


Figure 1: Absolute values of the eigenvalues $|r_i|$, $i = 1, 2$ of the stability matrices for the P-C modes: matrix A in case $k = 2$ (left) and matrix B in case $k = 1$ (right).

In Table 1 the intervals of stability are presented from $k = 2$ up to $k = 14$, where k refers to the number of steps, for the different implementations, named after SCk and FIk respectively, corresponding to the Störmer-Cowell and Falkner methods in the above section. All these values have been obtained with the help of the Mathematica program. We have considered only the primary stability intervals although for different methods and values of k it may exist secondary stability intervals.

4 Stability regions

If in equation (10) the parameter μ is let to take complex values, then the region in the H-complex plane within which all the eigenvalues of the stability matrix are less

 ISBN: 978-84-614-6167-7

k	SCk	FIk
2	$\varnothing$	$\varnothing$
3	$\varnothing$	$\varnothing$
4	$(0, 0.794719)$	$(0, 1.108998)$
5	$(0, 1.330257)$	$(0, 1.409664)$
6	$\varnothing$	$\varnothing$
7	$\varnothing$	$\varnothing$
8	$(0, 0.360863)$	$(0, 0.480033)$
9	$(0, 0.660841)$	$(0, 0.821956)$
10	$\varnothing$	$\varnothing$
11	$\varnothing$	$\varnothing$
12	$(0, 0.232478)$	$(0, 0.300133)$
13	$(0, 0.411350)$	$(0, 0.390144)$
14	$\varnothing$	$\varnothing$

Table 1: *Intervals of stability for the Störmer-Cowell and Falkner methods in P-C modes.*

than unity in absolute value is called the *stability region*. In case of a single differential equation or system of differential equations with real eigenvalues the stability region of interest reduces to the stability interval. For systems of differential equations with complex eigenvalues, however, the stability region plays an important role.

In Fig. 2 the stability regions for the Störmer-Cowell and Falkner methods in P-C mode when $k = 4$ are shown. Similarly, in cases of $k = 5, 8, 9$ the stability regions are represented in the figures 3, 4, 5. For $k = 12, 13$ there are also stability regions, but so small that they are of little interest.

Although the drawings are not very detailed near the origin all the stability regions reach the origin (not including) while around it they are very narrow. Furthermore, these regions are symmetric about the vertical axis. We note that the stability regions for the Falkner methods are larger than those for the Störmer-Cowell methods for the same number of steps (and hence for the same order). We also note that each pair of alternating groups for which there are non-empty intervals of stability, for the odd values the stability regions are larger. This makes us prefer the methods that correspond to the values of $k = 5, 9, 13$ versus $k = 4, 8, 12$. In view of the stability regions we suggest to choice the methods of Falkner versus the Störmer-Cowell methods. Note also that the number of evaluations of f in the two implementations is the same, two evaluations per step. Even more, in the Falkner methods the values of the derivative are obtained, while in the Störmer methods a procedure must be added if we want to obtain these values (see [12]).

 ISBN: 978-84-614-6167-7

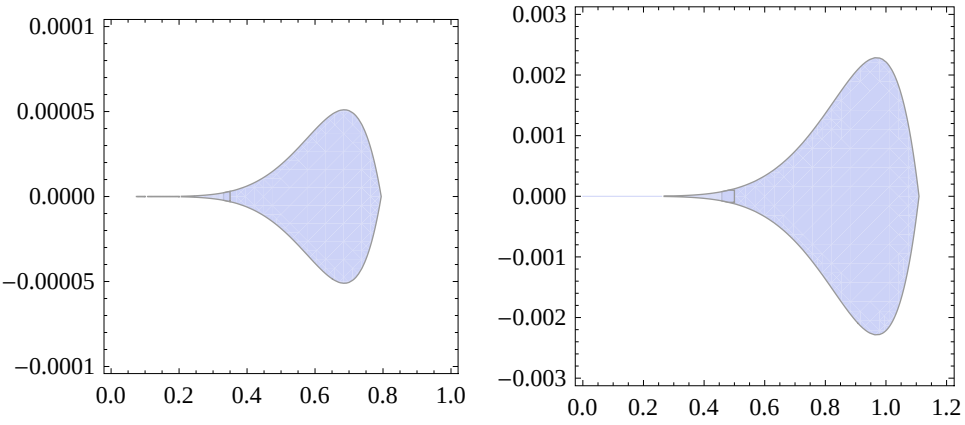


Figure 2: Stability regions for $k = 4$ in P-C mode: Störmer-Cowell (left) and Falkner (right) methods.

5 Concluding remarks

In the numerical treatment of initial-value problems through a numerical method the stepsize h has to be chosen in such a way that μh falls into the stability region of the method. Therefore the knowledge of the stability region is helpful during the integration. Comparisons between stability regions of the Störmer-Cowell and Falkner methods in PC mode show that the latter are wider, suggesting that these methods can be used as an alternative to the Störmer-Cowell methods.

Acknowledgements

This work has been partially supported by project MTM2008-05489 (Ministerio de Ciencia y Tecnología, Spain).

References

[1] D. Beeman, *Some multistep methods for use in molecular dynamics calculations*, J. Comput. Phys. **20** (1976) 130–139.

[2] Coleman J.P. and Ixaru L.Gr. *P-stability and exponential-fitting methods for* $y'' = f(x, y)$, IMA J. Numer. Anal. **16** (1996) ,179–199.

[3] Coleman J.P. and Mohamed J., *De Vogelaere´s methods with automatic error control*, Comput. Phys. Comm. **17** (1979) , 283–300

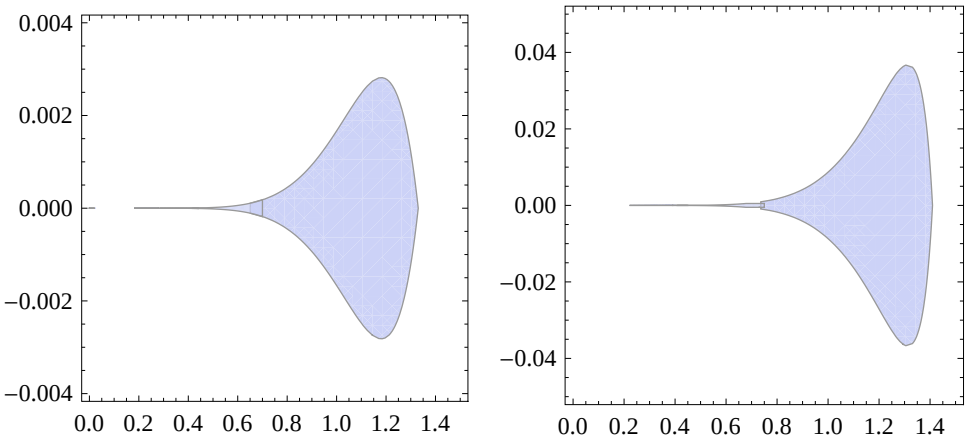


Figure 3: Stability regions for $k = 5$ in P-C mode: Störmer-Cowell (left) and Falkner (right) methods.

[4] L. Collatz, *The Numerical Treatment of Differential Equations*, Springer, Berlin, 1966.

[5] T. Feagin and P. R. Beaudet, *Multistep methods of numerical integration using back corrections*, Celestial Mechanics **13** (1976) 111–120.

[6] V. M. Falkner, *A method of numerical solution of differential equations*, Phil. Mag. S. **7** (1936) 621–640.

[7] J.M. Franco and I. Gómez *Accuracy and linear stability of RKN methods for solving second-order stiff problems*, Appl. Numer. Math. **59** (2009) , 959–975.

[8] I.Gladwell and R. Thomas, *Stability properties of the Newmark, Houbolt and Wilson θ methods*, I. J. Numer. and Anal. Meth. in Geom. **4** (1980) , 143–158.

[9] E. Hairer, S. P. Norsett and G. Wanner, *Solving Ordinary Differential Equations I* , Springer, Berlin, 1987.

[10] P. Henrici, *Discrete Variable Methods in Ordinary Differential Equations*, John Wiley, New York, 1962.

[11] L. Gr. Ixaru, *Numerical Methods for Differential Equations and Applications*, Editura Academiei, Romania, 1984.

[12] G.J. Kroes, *The royal road to an energy-conserving predictor-corrector method*, Comput. Phys. Comm. **70** (1992) 41–52.

 ISBN: 978-84-614-6167-7

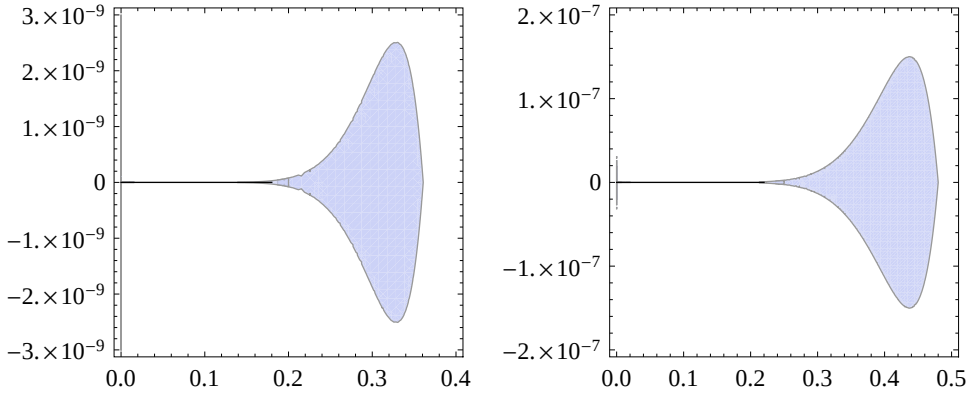


Figure 4: Stability regions for $k = 8$ in P-C mode: Störmer-Cowell (left) and Falkner (right) methods.

[13] J. D. Lambert *Numerical Methods for Ordinary Differential Systems*, John Wiley, England (1991)

[14] J. D. Lambert and I. A. Watson , *Symmetric multistep methods for periodic initial value problems*, IMA J. Numer. Anal. **18** (1976) , 189–202

[15] H. Ramos and C. Lorenzo, *Review of explicit Falkner methods and its modifications for solving special second-order I.V.P.s* , Comput. Phys. Comm. **181** (2010) 1833–1841.

[16] S. Toxvaerd, *A New Algorithm for Molecular Dynamics Calculations*, J. Comput. Phys. **47** (1982) 444–451.

[17] E. L. Wilson, *A computer program for the dynamic stress analysis of underground structures*, SESM Report No. 68–1, Division Structural Engineering and Structural Mechanics, University of California, Berkeley (1968)

 ISBN: 978-84-614-6167-7

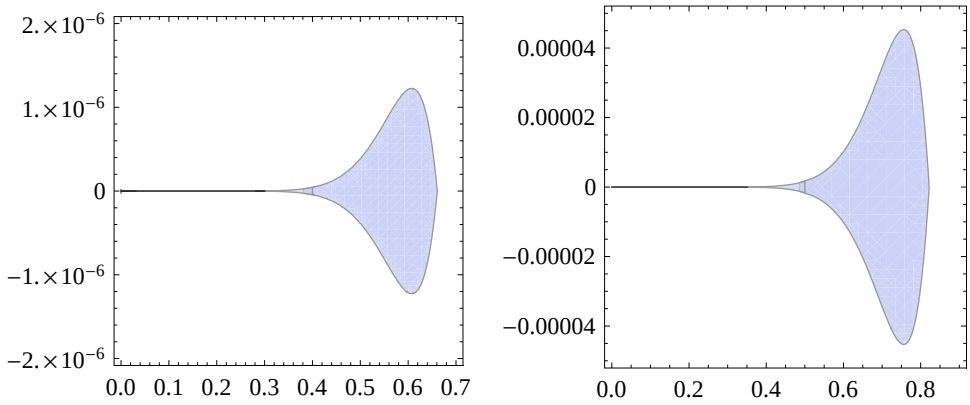


Figure 5: Stability regions for $k = 9$ in P-C mode: Störmer-Cowell (left) and Falkner (right) methods.

 ISBN: 978-84-614-6167-7

Proceedings of the 11th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2011
26–30 June 2011.

Multiscale modeling by anisotropic gaussian functions with applications to the corneal topography

Darío Ramos-López[1] and Andrei Martínez-Finkelshtein[1]

[1] *Department of Statistics and Applied Mathematics, University of Almería*

emails: dariorl@gmail.com, andrei@ual.es

Abstract

We describe a multi-scale and adaptive algorithm for the parsimonious fit of a surface from the elevation data contaminated by noise, collected at a discrete set of nodes. The surface is represented as a linear combination of anisotropic gaussian functions; the complexity of each basis function in this representation is the same, but their parameters vary to fit the current scale. This scale is determined only by the residual errors and not by the number of the iteration.

The procedure is applied to the reconstruction of the shape of the cornea from the data rendered by typical corneal topographers. The algorithm exhibits a steady exponential error decay, independently of the level of aberration of the cornea. The position and clustering of their centers, as well as the size of the shape parameters, provides an additional spatial information about the regions of higher irregularity.

Key words: surface reconstruction; surface modeling; corneal irregularities; gaussian functions; radial basis functions; multi-scale methods

1 Introduction

There is an increasing need of a reliable and precise modeling of complex surfaces, such as biological tissues, motivated both by technological and clinical applications. This applies in particular to the human cornea. Given the significance of the shape of the front surface of the cornea to the refraction of the eye [1] and the ability to correct refractive errors by laser ablation of the front surface of the cornea, a detailed wavefront error analysis of individual corneal topography data is crucial for the clinicians as a basis for a customized treatment.

 ISBN: 978-84-614-6167-7

Corneal modeling plays an essential role in diagnosing and managing corneal diseases such as keratoconus, to assess suitability of a subject for the treatment and prevent improper refractive surgeries [2]. Also, the great role of the reliable visualization tools in clinical practice should not be underestimated.

The vast majority of modern corneal topographers collect the data (either elevation, curvature, mire displacement or others) in a finite and discrete set of points.

Typically, the data is contaminated by the error, which stems from several sources: the natural device noise, measurement and digitalization errors, algorithm errors (like those converting the displacement in elevation), rounding errors and others. Hence, we face the problem of the parsimonious fit of the noisy surface data with a minimum number of coefficients or parameters, for its clinical and technological applications.

A standard approach follows the so-called modal paradigm, where the approximation is found as a linear combination of functions from the given dictionary. A standard functional basis for the modal reconstruction, used commonly in ophthalmology to express ocular wavefront error are the Zernike polynomials [3]. The coefficients of their expansions have interpretation in terms of the basic aberrations such as defocus, astigmatism or coma, along with higher order aberrations such as trefoil and spherical aberrations.

However, potential limitations in this approach have been reported in the literature [3, 4]. There is a growing concern that the Zernike fitting method itself may be inaccurate in abnormal conditions. Furthermore, it is very difficult to assess a priori how many terms are necessary to achieve acceptable accuracy in the Zernike reconstruction of any given corneal shape [5]. It is known [4] that limiting Zernike analysis to only a few orders may cause incorrect assessment of the severity of the more advanced stages of keratoconus [1]. This information is particularly needed in the discriminant analysis of the decease markers, or when selecting the numerical inputs for neural network–based diagnostic software such as corneal classification and grading utilities for condition severity.

Several alternatives to the modal least-square fit with Zernike polynomials have been recently suggested. In this paper we describe an adaptive and multi-scale working algorithm for the parsimonious fit of the surface data, based on residual iteration with knot insertion, that allows to adapt the number of functions used in the reconstruction to the conditions of each cornea.

1.1 The general setting

The input data is a 3D cloud (x_k, y_k, z_k), $k = 1, \ldots, N$, corresponding to the elevation z_k measured by a corneal topographer at the node $\boldsymbol{P}_k = (x_k, y_k)$ of the

 ISBN: 978-84-614-6167-7

anterior corneal surface. Taking into account the global shape of the cornea, the first global scale should "flatten" the data by fitting it with the best-fit sphere [6] of the form:

$$S(x,y) = z_0 + \sqrt{R^2 - (x-x_0)^2 - (y-y_0)^2}$$

where R and (x_0, y_0, z_0) are its radius and the Cartesian coordinates of its center, respectively. Best results are obtained with a weighted least square fit, using $(1 + \|\boldsymbol{P}_k\|)^{-1}$ as the weight, in accordance with the typical error distribution [7].

As a result of the previous step, the residual errors $\varepsilon_k^{(1)} = z_k - S(x_k, y_k)$ contain both the relevant information at different scales and noise. Our aim is to fit these residuals by a function $E(x,y)$ in such a way that an analytic expression for the corneal height is given by

$$\text{Cornea}(x,y) = S(x,y) + E(x,y), \tag{1}$$

In this way, S accounts for the global shape of the cornea, while E captures the small irregularities in the surface. Function E is given by a linear combination of n functions from a given dictionary,

$$E(x,y) = E_n(x,y) = \sum_{j=1}^{n} c_j\, h_j(x,y). \tag{2}$$

We use as basis functions the gaussians of the form

$$h(x,y) = \exp\left(-\|\boldsymbol{P} - \boldsymbol{Q}\|_A^2\right), \quad \boldsymbol{P} = (x,y)^T,$$

where the superscript T denotes the matrix transpose, $\boldsymbol{Q} = (Q_x, Q_y)^T$ is a certain point on the plane ("center"), and A is a positive-definite matrix in $\mathbb{R}^{2\times 2}$. For such a matrix the A-norm of a point (column vector) $\boldsymbol{P}$ in $\mathbb{R}^2$ is defined as

$$\|\boldsymbol{P}\|_A = \sqrt{\boldsymbol{P}^T A \boldsymbol{P}} = \sqrt{\alpha_x\, x^2 + \alpha_y\, y^2 + 2\alpha_{xy}\, xy},$$
$$\text{for } A = \begin{pmatrix} \alpha_x & \alpha_{xy} \\ \alpha_{xy} & \alpha_y \end{pmatrix} \text{ with } \alpha_x > 0 \text{ and } \alpha_x \alpha_y > \alpha_{xy}^2.$$

These anisotropic radial basis functions boil down to standard radial basis gaussian functions (RBGF) when both eigenvalues of A coincide (in other words, when A is a positive multiple of the identity matrix I_2). The size of the eigenvalues expresses the scale in the direction of the corresponding eigenvector.

Hence, we seek the expression of the form

$$\text{Cornea}(x,y) = S(x,y) + \sum_{j=1}^{n} c_j\, e^{-\|\boldsymbol{P} - \boldsymbol{Q}^{(j)}\|_{A_j}^2}, \tag{3}$$

 ISBN: 978-84-614-6167-7

where $\boldsymbol{P} = (x, y)^T$. A fitting routine should allow for an adequate selection of all parameters, namely centers $\boldsymbol{Q}^{(j)}$; shape matrices (or shape parameters) A_j; scaling factors c_j; and number of terms n in the functional representation.

We propose an iterative algorithm of reconstruction, such that in each step we fit partially the residual error by one anisotropic RBGF (A-RBGF), and compute the new residuals, which will become the input for the next iteration (residual iteration with knot insertion). To preserve the maximum possible degrees of freedom, the centers, the shape parameters and the scaling factors will be chosen dynamically depending on the residual data in each step.

1.2 Description of the iterative algorithm

Let E_{j-1} be already computed (we take $E_0 \equiv 0$). The input data for the j's iteration ($j = 1, 2, \dots$) is the cloud $(x_k, y_k, \varepsilon_k^{(j)})$ of nodes $\boldsymbol{P}_k = (x_k, y_k)^T$ and the corresponding residuals $\varepsilon_k^{(j)}$, $k = 1, 2, \dots, N$; recall that $\varepsilon_k^{(1)} = z_k - S(x_k, y_k)$.We perform the following steps:

Step 1: selection of the center $\boldsymbol{Q}^{(j)}$.

The simple criterion of maximal residual (strategy for so-called greedy approximation) proved to be very satisfactory, and at a minimum cost; it correlates also with the geometry of the A-RBGF. Hence, we choose

$$\boldsymbol{Q}^{(j)} = (x_{k_0}, y_{k_0})^T \quad \text{where} \quad k_0 = \arg\max_k |\varepsilon_k^{(j)}|$$

and denote $m^{(j)} := \varepsilon_{k_0}^{(j)}$.

Step 2: dynamical filtering.

As it was mentioned before, the altimetric data obtained from measuring devices such as a keratographer are contaminated by noise whose statistical distribution is difficult to estimate a priori. In order to cope with this problem we need to filter out those data that clearly correspond to the measurement error and thus spoil the quality of the reconstruction. For that, once the center $\boldsymbol{Q}^{(j)}$ has been selected, we check the number, ℓ_k, of nodes P_k lying in the largest disk, centered at $\boldsymbol{Q}^{(j)}$ and containing only nodes with the residues of the same sign as $m^{(j)}$. If $\ell_k < 20$, we consider $\boldsymbol{Q}^{(j)}$ an outlier and exclude it from consideration at this iteration. This can be done by simply setting $\varepsilon_{k_0}^{(j)} = 0$, after which we return to Step 1. Otherwise, we proceed to the next step.

Step 3: selection of the shape parameters and scaling factor.

We determine first the influence nodes $\mathcal{P}_j(s)$, defined as the maximal set of at most s nodes $\boldsymbol{P}_k$ closest to $\boldsymbol{Q}^{(j)}$ with residues of the same sign than $m^{(j)}$. It is convenient to parallelize the subsequent computations for several values of s, $s = [s_{\min}, 100, 150, 200, 300]$, where $s_{\min} = \min(\ell_k, 50)$, with ℓ_k defined in Step 2.

 ISBN: 978-84-614-6167-7

The interpolating conditions $\varepsilon_{k_0}^{(j)} h_j(x_k, y_k) = \varepsilon_k^{(j)}$, $k \in \mathcal{P}_j(s)$, are equivalent to the overdetermined linear system

$$\begin{aligned} &\alpha_x \left(x_k - x_{k_0}\right)^2 + 2\alpha_{xy} \left(x_k - x_{k_0}\right)\left(y_k - y_{k_0}\right) \\ &+\alpha_y \left(y_k - y_{k_0}\right)^2 = \log\left(\frac{\varepsilon_{k_0}^{(j)}}{\varepsilon_k^{(j)}}\right), \quad k \in \mathcal{P}_j(s) \end{aligned} \tag{4}$$

in the 3 unknown entries of the shape matrix A_j. We solve this system in the sense of weighted linear least squares (WLS), where the k-th equation is multiplied by the weight $\eta_k := (1 + \|\boldsymbol{P}_k - \boldsymbol{Q}^{(j)}\|^2)^{-1}$ in order to account for the bigger influence of the neighboring nodes on A_j. This solution is obtained by standard methods, using either the QR factorization of the collocation matrix corresponding to (4) or its singular value decomposition, see e. g. [8].

Observe that due to the selection of the active center $\boldsymbol{Q}^{(j)}$,

$$t_k := \log\left(\frac{\varepsilon_{k_0}^{(j)}}{\varepsilon_k^{(j)}}\right) \geq 0, \quad k \in \mathcal{P}_j(s).$$

However, this condition does not guarantee that the computed matrix A_j will be positive definite. This can typically fail in the periphery of the convex hull of the nodes, where the lack of data in some direction might yield non-positive definite A_j. Although the corresponding function h_j might fit the data correctly locally, it is not valid globally due to its exponential increase in the direction of the eigenvector of a negative eigenvalue of A_j. In this case we force h_j to be an isotropic (a bona fide) radial basis function: $A_j = \alpha I_2$. In this way, (4) is reduced to $\alpha \left\|\boldsymbol{P}_k - \boldsymbol{Q}^{(j)}\right\|^2 = t_k$, $k \in \mathcal{P}_j(s)$, whose solution in the sense of the WLS is $\alpha = \sum_{k \in \mathcal{P}_j(s)} \theta_k \, t_k$, with

$$\theta_k = \frac{\eta_k^2 \left\|\boldsymbol{P}_k - \boldsymbol{Q}^{(j)}\right\|^2}{\sum_{t \in \mathcal{P}_j(s)} \eta_t^2 \left\|\boldsymbol{P}_t - \boldsymbol{Q}^{(j)}\right\|^4}$$

and $\eta_k = \left(1 + \left\|\boldsymbol{P}_k - \boldsymbol{Q}^{(j)}\right\|^2\right)^{-1}$. Observe that in this case α is positive by construction, and we define

$$h_j(x, y) = \exp\left(-\alpha \left\|\boldsymbol{P} - \boldsymbol{Q}^{(j)}\right\|^2\right), \quad \boldsymbol{P} = (x, y)^T.$$

Finally, in order to compute the coefficient c_j we use the simple but effective interpolation condition $c_j = m^{(j)}$.

 ISBN: 978-84-614-6167-7

Step 4: computation of the new residuals.
With the values of c_j and A_j just computed we update

$$\varepsilon_k^{(j+1)} = \varepsilon_k^{(j)} - c_j\, h_j(x_k, y_k).$$

As it was mentioned before, all the computations have been performed in parallel for different values of s, and hence, different nested sets of influence nodes $\mathcal{P}_j(s)$. We now keep the value of s (and the corresponding values of c_j and A_j) that yields the smallest norm of the residue vector $(\varepsilon_k^{(j+1)})$, and discard the other values. As a result, we find the new approximation, $E_j = E_{j-1} + c_j\, h_j$, increment the iteration counter j in 1 and return to Step 1. This procedure is iterated a fixed number of times; in practice, the value $n = 100$ turns out to be totally satisfactory.

1.3 Multi-scale analysis and noise removal

In the real life situation of the data contaminated by errors, a very important problem is that of the model order selection: we want to capture all the relevant information without over-parametrizing the model and without fitting the noise. Many solutions to this problem are described in the literature. For instance, the choice of the number of Zernike polynomials for the modal reconstruction of the altimetric data has been discussed in [9, 10].

The statistical methods of selection of the appropriate number n in (2) usually make assumptions about the noise. However, a priori information about the measurement error bounds or measurement error distribution is very limited, in particular due to industrial secrets. According to [7], the errors cannot be assumed i.i.d. random variables, although the assumption that they are normally distributed (with the variance proportional to the square of the distance of the node to the center) is apparently reasonable. They are also computationally intensive, [10, 11].

The functional representation (3) allows also for a multi-resolution analysis, where bands of the values $\mathrm{Vol}_j := c_j/\sqrt{\det A_j}$ (proportional to the volume under the j-th gaussian component) can specify different levels of resolution. In particular, we can assume that the gaussians with small Vol_j correspond to the noise. This motivates the hard thresholding approach [12], [13] to the noise removal, setting to zero all coefficients c_j for which the condition $\mathrm{Vol}_j < \varepsilon$ holds. The selection of the parameter $\varepsilon > 0$ plays obviously the crucial role, and should be based either on the statistical distribution of noise or previous experiments.

2 Experimental results

The altimetric and curvature data from in-vivo corneas used for experiments described below were collected by the CSO topography system (CSO, Firenze, Italy),

 ISBN: 978-84-614-6167-7

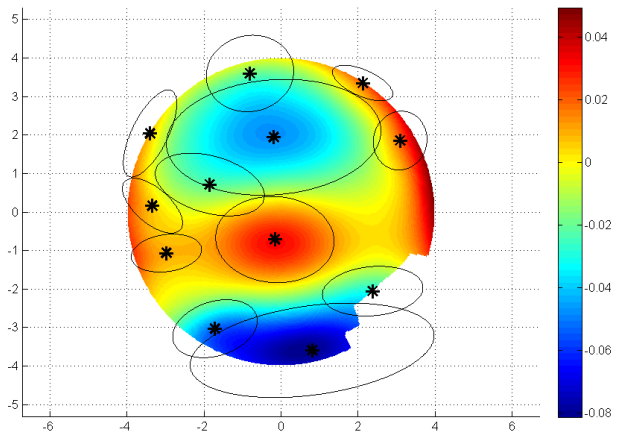


Figure 1: Corneal topography reconstructed with the A-RBGF algorithm. The asterisks show the locations of the centers $\boldsymbol{Q}^{(j)}$ in (3), while the ellipses are proportional to the values of the entries of the shape matrices A_j, showing the domains of influence of each gaussian (corresponding to the 75% of the volume under their surfaces).

which in ideal conditions digitizes up to 24 rings with 256 equally distributed points on each mire. All the procedures were implemented in Matlab 7 and run on standard platforms (Windows PC and Mac with average configuration). The execution time was always below 2 seconds, which makes it suitable for the real-time reconstruction of the corneal data.

Figure 1 shows the reconstructed corneal topography along with the centers and the "domains of influence" of the gaussian functions used in (3). Observe how the different scales of the functional components correspond to different scales of the features of the surface.

For comparison, we have reconstructed the same data with the A-RBGF algorithm and with the Zernike polynomials using the linear least squares, which is the standard procedure in the clinical practice, implemented in most topographers. The experiments illustrate that the Zernike polynomials easily capture the global shape of the reconstructed surface, which is expressed in a typical initial fast decay of the corresponding error. However, smaller scale details on the surface (such as areas of localized steepening) are much less suited for this tool and produce a typical saturation observed in the Zernike error behavior.

Another indication of a consistent behavior of the iterative algorithm proposed here is the evolution of the parameters computed dynamically in each iteration. While the spectral condition number of A_j remains essentially bounded, the scaling

 ISBN: 978-84-614-6167-7

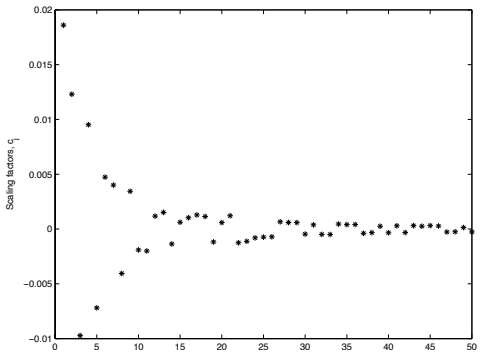


Figure 2: A typical evolution of the scaling factors c_j, plotted againts the number j of the function in (2).

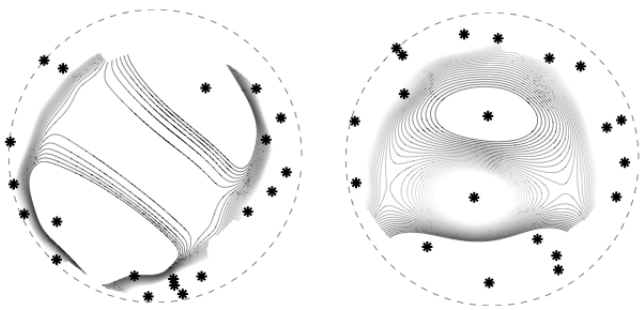


Figure 3: Center location for the reconstruction with 20 A-RBGF of a normal cornea (left) and a keratoconic one (right). The contours represent the level curves of the residues of the altimetric data with respect to the best fit sphere.

factors c_j steadily decrease, in concordance with a gradual reduction of the residual errors, see Figure 2.

It also should be pointed out that relevant correlation exists with the location and grouping of the centers $\boldsymbol{Q}^{(j)}$: for the normal corneas the centers typically cluster at the border of the area, where most of the oscillations occur, while for corneas affected by keratoconus we observe how some centers match the deformation already at the first iterations, see Figure 3.

3 Conclusion

In this work, we develop an adaptive fitting method for discrete altimetric data on scales independent of the iteration. It consists of a preliminary fit of the data with

 ISBN: 978-84-614-6167-7

some global function (in our case, a sphere) and an iterative procedure that adds terms to the analytic representation. Each term consists of a scaled anisotropic radial basis gaussian function whose coefficients are computed dynamically. The method comprises also both a noise reduction and a filtering procedure that discards the outliers (data clearly corresponding to measurement noise).

The numerical implementation of this algorithm in a standard personal computer is very fast (execution time below 2 seconds). Experimental results for a cornea reconstruction from keratometric measurements allow us to conclude that this iterative method exhibits a steady exponential error decay, independently of complexity of the cornea.

As a bonus, the position and clustering of the centers of A-RBGF, as well as the size of the shape parameters, provides an additional spatial information about the regions of higher irregularity.

In the case of the cornea, the iterative adaptive algorithm proposed here provides a method of obtaining a compact mathematical description of its shape at different scales. This description can be used for global visualization of the cornea or of its portions, as well as the input data for resampling and computation of some other relevant values via ray tracing, numerical integration and others.

Acknowledgements

A.M.-F. is supported in part by Junta de Andalucía grants FQM-229 and P09-FQM-4643, and by the Ministry of Science and Innovation of Spain (project code MTM2008-06689-C02-01). Both A.M.-F. and D.R.-L. are also supported in part by Junta de Andalucía grant P06-FQM-01735.

References

[1] J. T. Schwiegerling and J. E. Greivenkamp, "Keratoconus detection based on videokeratoscopic height data," *Optometry and Vision Science*, vol. 73, pp. 721–728, 1989.

[2] N. Maeda, S. D. Klyce, and M. K. Smolek, "Comparison of methods for detecting keratoconus using videokeratography," *Archives of Ophthalmology*, vol. 113, no. 7, pp. 870–874, July 1995.

[3] S. D. Klyce, M. D. Karon, and M. K. Smolek, "Advantages and disadvatages of the Zernike expansion for representing wave aberration of the normal and aberrated eye," *J. Refractive Surgery*, vol. 20, pp. S537–S541, 2004.

 ISBN: 978-84-614-6167-7

[4] M. K. Smolek and S. D. Klyce, "Goodness-of-prediction of Zernike polynomial fitting to corneal surfaces," *J. Cataract Refract. Surg.*, vol. 31, pp. 2350–2355, 2005.

[5] R. Iskander, M. J. Collins, and B. Davis, "Optimal modeling of corneal surfaces with Zernike polynomials," *IEEE Trans. Biomed. Eng.*, vol. 48, no. 1, pp. 87–95, January 2001.

[6] S. J. Ahn, W. Rauh, and H.-J. Warnecke, "Least-squares orthogonal distances fitting of circle, sphere, ellipse, hyperbola, and parabola," *Pattern Recognition*, vol. 34, no. 12, pp. 2283– 2303, 2001.

[7] W. Tang, M. J. Collins, L. G. Carney, and B. Davis, "The accuracy and precision performance of four videokeratoscopes in measuring test surfaces," *Optometry and Vision Science*, vol. 77, no. 9, pp. 483–491, September 2000.

[8] L. N. Trefethen and D. Bau III, *Numerical Linear Algebra.* SIAM, 1997.

[9] R. Iskander, M. R. Morelande, M. J. Collins, and B. Davis, "Modeling of corneal surfaces with radial polynomials," *IEEE Trans. Biomed. Eng.*, vol. 49, no. 4, pp. 320–328, 2002.

[10] D. Alonso-Caneiro, R. Iskander, and M. J. Collins, "Estimating corneal surface topography in videokeratoscopy in the presence of strong signal interference," *IEEE Trans. Biomed. Eng.*, vol. 55, no. 10, pp. 2381–2387, October 2008.

[11] D. R. Iskander, W. Alkhaldi, and A. M. Zoubir, "On the computer intensive methods in model selection," in *Proc. 33rd IEEE Int. Conf: Acoust., Speech Signal Process. (ICASSP)*, 2008, pp. 3461–3464.

[12] D. L. Donoho, I. M. Johnstone, G. Kerkyacharian, and D. Picard, "Wavelet shrinkage: asymptopia?" *J. Roy. Statist. Soc. Ser. B*, vol. 57, no. 2, pp. 301–369, 1995.

[13] D. L. Donoho, I. M. Johnstone, G. Kerkyacharian, and D. Picard, "Density estimation by wavelet thresholding," *Ann. Statist.*, vol. 24, no. 2, pp. 508–539, 1996. [Online]. Available: http://dx.doi.org/10.1214/aos/1032894451

 ISBN: 978-84-614-6167-7

Proceedings of the 11th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2011
26–30 June 2011.

On noncommutative semifields of odd characteristic

J. Ranilla[1], Elías F. Combarro[1] and I.F. Rúa[2]

[1] *Departamento de Informática, Universidad de Oviedo. Spain.*

[2] *Departamento de Matemáticas, Universidad de Oviedo. Spain.*

emails: ranilla@uniovi.es, efernandezca@uniovi.es, rua@uniovi.es

Abstract

Finite nonassociative division algebras (i.e., finite semifields) with 243 elements are completely classified.

1 Introduction

Finite division rings, also known as **finite semifields**, are nonassociative rings with identity such that the set of nonzero elements is closed under the product (i.e., a loop [11, 3]). In case it has no identity they are known as **presemifields**. These objects have been studied in different contexts: finite geometries (they coordinatize projective semifield planes [6]), coding theory [2, 8, 5], combinatorics and graph theory [13].

Computational methods have been considered in the study of these objects. Among others, the classification of finite semifields of orders 16 [10, 11], 32 [16, 11] and, more recently, orders 64 [14] and 81 [4] have been obtained with the help of computational tools. Presently, only the cases of orders 128, 243 and 256 remain to achieve the classification of semifield planes of order 256 or less suggested in [9].

In this paper we present a classification of semifields with **243 elements** up to isotopy. It is a computer-assited classification based on the algorithms introduced in [14].

2 Preliminaries

In this section we collect some definitions and facts on finite semifields, presemifields and planar functions (see, for instance [3, 11]). We restrict ourselves to the particular case of order $243 = 3^5$. The characteristic of a finite presemifield D with 3^5 elements is 3, and D is a $5-$dimensional algebra over $\mathbb{F}_3$. If D is a semifield, then $\mathbb{F}_3$ can be chosen to be contained in its associative-commutative center $Z(D)$. Other relevant subsets of

 ISBN: 978-84-614-6167-7

a finite semifield are the left, right, and middle nuclei (N_l, N_r, N_m), and the nucleus N which have to be field extensions $\mathbb{F}_{3^e}$ $(e \leq 5)$.

Classification of presemifields is usually considered up to *isotopy* (since this corresponds to classification of the corresponding projective planes up to isomorphism): If D_1, D_2 are two presemifields of order 3^5, an isotopy between D_1 and D_2 is a triple (F, G, H) of bijective $\mathbb{F}_3-$linear maps $D_1 \to D_2$ such that

$$H(ab) = F(a)G(b)\ ,\ \forall a, b \in D_1.$$

Any presemifield is isotopic to a finite semifield.

If $\mathcal{B} = [x_1, \ldots, x_5]$ is a $\mathbb{F}_3$-basis of a presemifield D, then there exists a unique set of constants $\mathbf{A}_{D,\mathcal{B}} = \{A_{i_1 i_2 i_3}\}_{i_1,i_2,i_3=1}^5 \subseteq \mathbb{F}_3$ such that

$$x_{i_1} x_{i_2} = \sum_{i_3=1}^{5} A_{i_1 i_2 i_3} x_{i_3}\ \forall i_1, i_2 \in \{1, \ldots, 5\}$$

This set of constants is known as **cubical array** or **3-cube** corresponding to D with respect to the basis $\mathcal{B}$, and it completely determines the multiplication in D. If D is a presemifield, and $\sigma \in S_3$ (the symmetric group on the set $\{1, 2, 3\}$), then the set

$$\mathbf{A}_{D,\mathcal{B}}^{\sigma} = \{A_{i_{\sigma(1)} i_{\sigma(2)} i_{\sigma(3)}}\}_{i_1,i_2,i_3=1}^5 \subseteq \mathbb{F}_3$$

is the 3-cube of a presemifield. Different choices of bases lead to isotopic presemifields. Up to six projective planes can be constructed from a given finite semifield D using the transformations of the group S_3. So, the classification of finite semifields can be reduced to the classification of the corresponding projective planes up to the action of the group S_3.

Finite semifields of order 243 can be constructed from sets of matrices with certain properties [4][7, Proposition 3].

Proposition 1. *There exists a finite semifield D of order 3^5 if, and only if, there exists a set of 5 matrices (a **standard basis** of D) $S_D = \{A_1, \ldots, A_5\} \subseteq GL(5, 3)$ (the set of invertible matrices of size 5 over $\mathbb{F}_3$) such that:*

1. *A_1 is the identity matrix I;*
2. *$\sum_{i=1}^5 \lambda_i A_i \in GL(5, 3)$, for all non-zero tuples $(\lambda_1, \ldots, \lambda_5) \in \mathbb{F}_3^5$, that is, $(\lambda_1, \ldots, \lambda_5) \neq \{\overrightarrow{0}\}$.*
3. *The first column of the matrix A_i is the column vector $e_i^{\downarrow}$ with a 1 in the i-th position, and 0 everywhere else.*

In such a case, the set $\{B_{ijk}\}_{i,j,k=1}^d$, where $B_{ijk} = (A_j)_{ik}$, is the 3-cube corresponding to D with respect to the canonical basis of $\mathbb{F}_3^5$.

If we identify the elements of $\mathbb{F}_3$ with the natural numbers $\{0, 1, 2\}$, then we can use the following convention to represent a semifield D of order 3^5. Let $S_D = \{A_1, \ldots, A_5\}$ be one of its standard bases. Since the first column of A_i has always a one in the i-th

 ISBN: 978-84-614-6167-7

position and zeroes elsewhere, we can encode A_i as the natural number $\sum_{j=0}^{19} a_j 3^j$, where

$$\left(\begin{array}{c|cccc} & a_{19} & a_{14} & a_9 & a_4 \\ & a_{18} & a_{13} & a_8 & a_3 \\ e_i^{\downarrow} & a_{17} & a_{12} & a_7 & a_2 \\ & a_{16} & a_{11} & a_6 & a_1 \\ & a_{15} & a_{10} & a_5 & a_0 \end{array}\right)$$

For a concrete representation of the semifield one can identify the semifield with $\mathbb{F}_3^5$, and the multiplication with $x*y = \sum_{i=1}^{5} x_i A_i y$, i.e., A_i is the matrix of left multiplication by the element e_i, where $\{e_1, \ldots, e_5\}$ is the canonical basis of $\mathbb{F}_3^5$. So, the elements of the standard basis are simply coordinate matrices of the linear maps $L_{e_i} : D \to D$, $L_{e_i}(y) = e_i * y$.

3 The Semifield Planes of order 243: a classification

We obtained a complete classification of finite semifields of order 243 with the help of the algorithm introduced in [14]. This algorithm searchs for standard bases of division algebras with 243 elements, and classify them according to equivalent S_3-equivalent semifields. This is done either for partial or for complete standard bases.

Our algorithm was processed in parallel in Magerit, a cluster of 1204 nodes eServer BladeCenter (1036 JS20 and 168 JS21, both PowerPC 64 bits). Each JS20 node has two processors IBM PowerPC single-core 970FX (two cores) with 2.2 GHz, 4 GB of RAM and 40 GB of local hard disk. On the other hand, each JS21 node has two processors IBM PowerPC dual-core 970FX (four cores) with 2.2 GHz, 8 GB of RAM and 80 GB of local hard disk. It was installed in 2006 and reached the 9th fastest in Europe and the 34th in the world (Top 500: List from November 2006). In May 2008 it was upgraded to reach 16 TFLOPS. This powerful cluster has allowed us to fill the gap between the commutative and the nonconmutative case.

Next we present the results obtained from our classification (Table 1). Let us compare the number of S_3-equivalence classes, semifield planes, and coordinatizing finite semifields which were found, with those previously known [15].

Number of classes	S_3**-action**	**Isotopy**	**Isomorphism**
Previously known	7	19	27313
Actual number	9	23	85877

Table 1: Number of division algebras with 243 elements

As we can see two new S_3-classes exist, that can not be constructed from commutative semifields. And four new semifield planes of such an order appear. Next we present standard bases of these classes (A_1 is always the identity matrix) (Table 2).

4 Further work

Currently we are working on appying this methodology to other cases. Namely, semifields of order 7^4 are being investigated.

 ISBN: 978-84-614-6167-7

#	A_2	A_3	A_4	A_5	# Semifield
I	129317742	43151760	25524498	2715668620	$\mathbb{F}_{3^5}$
II	129317638	44994959	28587138	1226007534	Albert's twisted field
III	129317781	52757047	20739470	3274303432	Albert's twisted field
IV	129317742	43393513	26923067	2713804376	Coulter-Matthews'
V	129317742	43215002	26537147	2719346408	Ding-Yuan's
VI	129317742	43185096	19259172	2718371119	[15]
VII	129317742	43215002	26558192	2719382129	[15]
VIII	129317636	14673002	1139489406	3073918154	-
IX	129317636	18089998	3416237282	1030364558	-

Table 2: Standard bases of division algebras with 243 elements (VIII and IX are new semifields)

Acknowledgments

This work has been partially supported by MEC - MTM - 2010 - 18370 - C04 - 01 and MICINN-TIN-2010-14971. The authors thankfully acknowledge the computer resources, technical expertise and assistance provided by the *Centro de Supercomputación y Visualización de Madrid (CeSViMa)*, the Spanish Supercomputing Network and *Clúster de Modelización Científica* at University of Oviedo.

References

[1] A. A. Albert, *Finite division algebras and finite planes*, Proceedings of Symposia in Applied Mathematics **10** (1960), 53-70.

[2] A. R. Calderbank, P. J. Cameron, W. M. Kantor, J. J. Seidel, $\mathbb{Z}_4$*-Kerdock codes, orthogonal spreads, and extremal Euclidean line-sets*, Proc. London Math. Soc **75** (1997), 436–480.

[3] M. Cordero, G. P. Wene, *A survey of finite semifields*, Discrete Mathematics **208/209** (1999), 125-137.

[4] U. Dempwolff, *Semifield Planes of Order 81*, J. of Geometry **89** (2008), 1-16.

[5] S. González, C. Martínez, I.F. Rúa, *Symplectic Spread based Generalized Kerdock Codes*, Designs, Codes and Cryptography **42 (2)** (2007), 213–226.

[6] M. Hall(Jr.), *The theory of groups*, Macmillan, (1959).

[7] I.R. Hentzel, I. F. Rúa, *Primitivity of Finite Semifields with 64 and 81 elements* , International Journal of Algebra and Computation **17 (7)** (2007), 1411-1429.

[8] W. M. Kantor, M. E. Williams, *Symplectic semifield planes and* $\mathbb{Z}_4$*-linear codes*, Transactions of the American Mathematical Society **356** (2004), 895–938.

 ISBN: 978-84-614-6167-7

[9] W. M. Kantor, *Finite semifields*, Finite Geometries, Groups, and Computation (Proc. of Conf. at Pingree Park, CO Sept. 2005), de Gruyter, Berlin-New York (2006).

[10] Kleinfeld, E. Techniques for enumerating Veblen-Wedderburn systems. J. Assoc. Comp. Mach. **1960**, 7, 330-337.

[11] D.E. Knuth, *Finite semifields and projective planes*, Journal of Algebra **2** (1965), 182-217.

[12] R. Lidl, H. Niederreiter, *Finite Fields*, Encyclopedia of mathematics and its applications 20, Addison-Wesley (1983).

[13] J. P. May, D. Saunders, Z. Wan, *Efficient Matrix Rank Computation with Applications to the Study of Strongly Regular Graphs*, Proceedings of ISSAC 2007, 277-284, ACM, New-York, 2007

[14] I. F. Rúa, Elías F. Combarro, J. Ranilla, *Classification of Semifields of Order 64*, J. of Algebra, **322 (11)** (2009), 941-961.

[15] I. F. Rúa, Elías F. Combarro, *Commutative semifields of order* 3^5, Communications in Algebra (to appear).

[16] R. J. Walker, *Determination of division algebras with 32 elements*, Proceedings in Symposia of Applied Mathematics **75** (1962), 83-85.

 ISBN: 978-84-614-6167-7

Proceedings of the 11th International Conference on Computational and Mathematical Methods in Science and Engineering, CMMSE 2011 26–30 June 2011.

Drug release from collagen matrices and transport phenomena in porous media including an evolving microstructure

Nadja Ray[1], Florin A. Radu[12] and Peter Knabner[1]

[1] *Department of Mathematics, Friedrich-Alexander University of Erlangen-Nuremberg, Martensstrasse 3, 91058 Erlangen, Germany*

[2] *UFZ-Helmholtz Center for Environmental Research, Permoser Straße 15, 04318 Leipzig, Germany*

emails: ray@am.uni-erlangen.de, raduf@am.uni-erlangen.de, knabner@am.uni-erlangen.de

Abstract

Biodegradable collagen matrices have been become a promising alternative to traditional drug delivery systems. The relevant mechanism in controlled drug release are the penetration of the collagen matrix by water, the swelling of the matrix where drug is released by diffusion and enzymatic degradation of the matrix with simultaneous drug release. These phenomena have been studied experimentally, via numerical simulations and also analytically extensively in the past. However, the adsorption processes that determine degradation later on have not been investigated in detail on the pore-scale by now. The sorption of solved particles on the collagen matrix due to particle interaction mechanism can change the underlying microstructure. This on the one hand can lead to a decrease of sorption sites for the enzyme followed by a decrease of the degradation rate. On the other hand, the physical entrapment of the active agent can be increased and therefore inhibits the drug release. We present an averaged model treating this phenomenon using formal two-scale asymptotic expansion in a level set framework. Thereby, we focus on the particle interaction with the collagen matrix.
Key words: drug release, porous media, homogenization, evolving microstructure

1 Introduction

Traditional drug delivery is characterized by its immediate release which leaves absorption to be controlled by the human body. Drug concentration thereby typically undergoes an abrupt increase followed by an abrupt decrease. Controlled release drug delivery systems such as collagen systems make it possible to change the drug release

 ISBN: 978-84-614-6167-7

profile and therefore improve the therapeutic efficacy.
The relevant mechanisms in controlled drug release are the penetration of the collagen matrix by water, the swelling of the matrix where drug is released by diffusion and the degradation of the matrix with simultaneous drug release. These mechanisms have been studied separately in [10], [13] and [14]. In [13] the water diffusion, the swelling of the matrices and the drug release by diffusion have been mathematically modeled, implemented and verified by experiments. Furthermore, a one-dimensional mathematical model based on two coupled partial differential equations with two moving fronts has been established in [13]. For the special case of constant diffusion coefficients even an analytical solution has been derived. As next step the enzymatic degradation of the polymer, which dramatically influences the drug release from the collagen matrices, because the drug molecules are mechanically inhibited to release by the collagen fibers, has been considered. A two-dimensional mathematical model which consists of a system of coupled partial and ordinary differential equations has been developed in [14]. This model has again been verified by experimental data, see [10]. An overall mathematical description of the two models described above has been investigated in [3].
When the water penetrates the collagen matrix there are several processes that are of interest. If an enzyme is solved within the water reactions with the matrix take place to form an enzyme-substrate complex which later on breaks up. This allows the drug that has been physically immobilized by the collagen matrix due to physical entrapment to be released. Furthermore, the dissolved species can sorb on the collagen matrix due to particle interaction mechanism. This may lead to a change of the underlying microstructure, to a decrease of available sorption sites for the enzyme and therefore to a decrease of the degradation rate and to an increased physical entrapment of the active agent and therefore to the inhibition of drug release. Although these effects strongly determine degradation later on, they have not been studied in detail on the pore-scale by now, i. e. in all the models mentioned above the change in the microstructure of the collagen matrix has not been considered explicitly. With our paper we want to contribute exactly to this point. It is thereby reasonable to separate the study of these phenomena not just for the sake of clearness but since the degradation occurs on a much larger time scale than the adsorption processes.
Besides the context of drug delivery, evolving microstructures are of great interest since they occur in variable applications such as swelling clays, concrete corrosion or dissolution and precipitation processes. In the literature various attempts to model this phenomena can be found. Swelling clays have been investigated in [11], [12] using formal two-scale asymptotic expansion. Applying hybrid mixture theory developments in the context of drug release have been studied by Weinstein in [22]. A direct treatment of a precipitation/dissolution front via level set function using formal two-scale asymptotic expansion can be found in [18].
As starting point of our investigations we mathematically set up a microscopic model at the pore scale for transport processes and fluid flow within the collagen matrix. Special attention is paid to the fact that particles that are solved in the penetrating water underly different interaction potentials such as electrostatic and van-der-Waals

 ISBN: 978-84-614-6167-7

forces. Thereby induced interaction with the porous matrix results in a change of the underlying microstructure of the collagen matrix. This is directly incorporated by using a level set formulation to describe the evolving microstructure. In our model, transport is caused by diffusion, convection and, in addition, drift due to the interaction potential. The number density of particles is therefore computed by a modified convection-diffusion equation, which is also known as Nernst-Planck equation. Since we pay special attention to the interaction effects, fluid flow is described by modified incompressible Stokes' equations with force density as right hand side. The aim of the paper is to derive an equivalent macroscopic model description that accounts for the evolving microstructure. We apply a formal upscaling procedure using two-scale asymptotic expansion handling the evolving microstructure directly by a level set function as was first supposed in [18] in the framework of precipitation/dissolution reactions in porous media. The evolution on the micro-scale is thereby dictated by the total interaction potential consisting of repulsive and attractive parts. An equivalent macroscopic model description for fluid flow and transport in porous media is presented for which consistency with well known standard models as well as symmetry, positive definiteness and ellipticity, resp., for the averaged coefficient functions, namely permeability and diffusivity can be proven.
The paper is organized as follows: In section 2.1, we state the description of the evolving microstructure. In section 2.2, a microscopic model at the pore scale for transport processes and fluid flow within the collagen matrix will be set up mathematically. In section 2.3, we apply a formal upscaling procedure using two-scale asymptotic expansion in a level set framework. In Theorem 2.1, we state the results of the averaging procedure, i.e. an equivalent macroscopic model description is presented. In section 2.4 and 3 we finally discuss the derived upscaled model.

2 Mathematical Model

2.1 Variable pore geometry

The mathematical model is designed here for a collagen matrix that consists of collagen fibers which have cylindrical shape. Due to homogeneity in one space dimension it is reasonable to restrict ourselves to a vertical cross section and therefore to a two-dimensional domain Ω that is supposed to be an idealized porous medium. It is the periodic composition of shifted and scaled unit cells $Y = \left[-\frac{1}{2}, +\frac{1}{2}\right]^2$ consisting of one centered spherical cross section of a collagen fiber Y_s and the penetrating water phase Y_l with $\bar{Y}_s \cup \bar{Y}_l = \bar{Y},\ Y_s \cap Y_l = \emptyset$. In order to describe the evolution of the underlying microstructure first within a unit cell, we consider the following simplified case: We regard the total interaction potential to be a-priori given and to be radial symmetrically and assume that the collagen fiber can only grow or shrink uniformly due to particle attachment/detachment. We assume furthermore that the influence of the inter-particle interaction on the total interaction potential only plays a tangential role compared to the particle fiber interaction potential since we want to concentrate our studies on the effects that occur due to the interaction of the particles with the collagen fibers. These

 ISBN: 978-84-614-6167-7

assumptions are for example reasonable if enough equally favorable sorption sites are available for attachment.
The goal is now to describe the evolution of the microstructure first within a unit cell by the change in size of the collagen fiber. The actual fiber radius R is thereby determined as equilibrium state of the system. Therefore the effective distance r^{eff} between the particles and the collagen fiber, which does depend on the fiber radius, is compared to an equilibrium distance that is determined by the total interaction potential: Depending on the distance from the collagen fiber, the total potential acting on the particle may be repulsive or attractive. As long as the total potential is attractive, attachment takes place whereas if the total potential is repulsive, the particles are stabilized within the solution. Altogether an equilibrium between attachment and detachment is reached for the equilibrium points a (primary or secondary minimum) of the total potential Φ. This means that neither repulsion nor attraction takes place. In the equilibrium case, the mean distance of particles to the collagen fiber is equal to this equilibrium distance given by the total potential. In the unit cell, the mean distance r^{eff} of the particles to the fiber can be calculated by

$$r^{\text{eff}} = \frac{\int_{Y_l} c(\boldsymbol{y}) r(\boldsymbol{y})\, d\boldsymbol{y}}{\int_{Y_l} c(\boldsymbol{y})\, d\boldsymbol{y}}$$

with $r(\boldsymbol{y}) = \|\boldsymbol{y}\| - R$ being the distance of any point y of the cell to the spherical fiber cross section with radius R and c denoting the number density of solved particles. The numerator in this formula then describes the weighted distance and the denominator describes the total number of particles. We now claim that $r^{\text{eff}} = a$ and may therefore calculate the equilibrium fiber radius $R = R(t, \boldsymbol{x})$ by

$$R = \frac{\int_{Y_l} c(\boldsymbol{y}) \|\boldsymbol{y}\|\, d\boldsymbol{y}}{\int_{Y_l} c(\boldsymbol{y})\, d\boldsymbol{y}} - a.$$

for given equilibrium point a of the total potential. This enables us to determine the distribution of pore water and (enlarged) collagen fibers within the unit cell. The coordinate $\boldsymbol{x}$ thereby denotes for example the position of the midpoint of the unit cell within the porous medium. The interface between water and fiber is given by $\Gamma(t, \boldsymbol{x}) = \{\boldsymbol{y} \in Y : \|\boldsymbol{y}\| = R(t, \boldsymbol{x})\} = \{\boldsymbol{y} \in Y : \|\boldsymbol{y}\| - R(t, \boldsymbol{x}) = 0\}$ and the corresponding level set function $S(t, \boldsymbol{x}, \boldsymbol{y})$ is defined by

$$S(t, \boldsymbol{x}, \boldsymbol{y}) := R(t, \boldsymbol{x}) - \|\boldsymbol{y}\| \begin{cases} < 0 & \text{liquid phase} \\ = 0 & \text{interface} \\ > 0 & \text{solid phase.} \end{cases}$$

2.2 Pore Scale Model

In this section we introduce the underlying mathematical model equations for the transport of the solved particles and the water flow within the collagen matrix. Moreover, we perform the upscaling the model equations involving the evolving microstructure

 ISBN: 978-84-614-6167-7

which is described by the level set function that has been introduced above.
The separation of the pore scale and macroscopic relevant scale is characteristic for a porous medium. Thus we define the ratio between the two length scales to be the scale parameter ε. We introduce the ε-scaled collagen matrix Ω_ε which is occupied by the idealized porous medium consisting of scaled unit cells with side length ε and fiber boundaries $\Gamma_\varepsilon := \varepsilon(\bar{Y}_s \cap \bar{Y}_l)$. Within this domain, the solved particles that penetrate the collagen matrix with the pore water are modeled in an Eulerian approach by some number density c_ε. Besides standard transport mechanisms, namely convection and diffusion, we incorporate various repulsive and attractive interaction potentials Φ_ε that may act on the particles in our model. This results in a modified convection-diffusion equation which is also known as Nernst-Planck equation, see [17], [6]. Furthermore, we incorporate these processes in the framework of the evolving microstructure that has been introduced in section 2.1 which leads to the following equation:

$$\partial_t c_\varepsilon + \boldsymbol{v}_\varepsilon^{\mathrm{hydr}} \cdot \nabla c_\varepsilon - \nabla \cdot (\nabla c_\varepsilon + c_\varepsilon \nabla \Phi_\varepsilon) = 0 \qquad \text{in } \Omega_\varepsilon(t),\ t \in [0,T], \tag{1a}$$

$$(-\boldsymbol{v}_\varepsilon^{\mathrm{hydr}} c_\varepsilon + \nabla c_\varepsilon + c_\varepsilon \nabla \Phi_\varepsilon) \cdot \boldsymbol{\nu}_\varepsilon + \varepsilon \partial_t R_\varepsilon (\rho_s - c_\varepsilon) = 0 \qquad \text{on } \Gamma_\varepsilon(t),\ t \in [0,T], \tag{1b}$$

$$c_\varepsilon = c^0 \qquad \text{in } \Omega_\varepsilon(0). \tag{1c}$$

Hereby $\boldsymbol{\nu}_\varepsilon$ denotes the outer normal to Γ_ε and c^0 is an appropriate initial condition. In order to determine the fluid velocity $\boldsymbol{v}_\varepsilon^{\mathrm{hydr}}$ and pressure p_ε we solve a modified Stokes' equations for incompressible fluid flow. On the pore level, we adopt the hydrology convention here for the penetrating water instead of using Fick's law of diffusion, opposed to e. g. [13]. This means that we distinguish between bulk flow of the penetrating water and the transport of the particles within the water phase. As force term on the right hand side we take into account the drift force density. These equations are supplemented by a no slip boundary condition at the collagen fiber:

$$-\varepsilon^2 \Delta \boldsymbol{v}_\varepsilon^{\mathrm{hydr}} + \nabla p_\varepsilon = -\varepsilon c_\varepsilon \nabla \Phi_\varepsilon \qquad \text{in } \Omega_\varepsilon(t),\ t \in [0,T], \tag{2a}$$

$$\nabla \cdot \boldsymbol{v}_\varepsilon^{\mathrm{hydr}} = 0 \qquad \text{in } \Omega_\varepsilon(t),\ t \in [0,T], \tag{2b}$$

$$\boldsymbol{v}_\varepsilon^{\mathrm{hydr}} = -\varepsilon \alpha \partial_t R_\varepsilon \boldsymbol{\nu}_\varepsilon \qquad \text{on } \Gamma_\varepsilon(t),\ t \in [0,T], \tag{2c}$$

The a priori given total interaction potential Φ_ε consists of a repulsive and an attractive part. The repulsive part may for example arise by reason of electric repulsion and the attractive part by van-der-Waals forces.

Remark 1. *Due to the no slip boundary condition, the velocity field $\boldsymbol{v}_\varepsilon^{hydr}$ is perpendicular to the boundary of the collagen fiber and is equal to $-\varepsilon\alpha\partial_t R_\varepsilon \boldsymbol{\nu}_\varepsilon$ on the moving boundary. The parameter α is defined to be $\frac{\rho_l - \rho_s}{\rho_l}$. This can be seen considering the conservation of mass and regarding the change in the density at the interface. Let thereby ρ_l denote the constant density of the liquid being composed of the concentration of particles and the density of the solvent and let ρ_s denote the density of the solid. The term $\varepsilon\partial_t R_\varepsilon(\rho_s - c_\varepsilon^{\pm})$ that occurs in the boundary condition of the Nernst-Planck equation and can also be developed regarding the conservation of mass. Furthermore the term $\partial_t R_\varepsilon$ denotes the time derivative of a shifted version of the radius R we have*

 ISBN: 978-84-614-6167-7

determined in section 2.1.
For the ease of presentation we do not regard the outer boundary conditions of the macroscopic domain explicitly since they play no crucial role in the formal homogenization process. Moreover, reasonable scalings with scale parameter ε could be motivated by performing a non-dimensionalization procedure.
The (unscaled) system (1), (2) can be found in a more general context. It occurs when determining ion distributions (for example around colloidal particles or in a ion channel). If the total potential is identified with the electrostatic potential and the convective term is neglected, the system is also well known in the framework of semiconductor devices. Analytical investigations of this system and extensions to the Navier-Stokes equations can be found in [9], [15] and [20]. Furthermore, several contributions to averaging such charged transport can be found in the literature. In rigid porous media formal upscaling for a linearized system has been considered for example by [8]. In [2], [8] and [11], [12], different homogenized models are furthermore developed and tested for the satisfaction of Onsager's reciprocity relation. Recent rigorous homogenization results in non-deformable media can be found in [21] and [1]. Swelling clays filled by an electrolyte are considered in [11], [12].

Considering the whole porous medium instead of one unit cell, a periodic extension of the above derived level set function for a unit cell to the whole scaled domain Ω_ε is necessary. This is done analogously to [19] with slight corrections: Assuming the total interaction potential being given, the fiber radius $R : \Omega \times [0,T] \to \left[0, \frac{1}{2}\right)$ may be calculated as described in section 2.1 for a unit cell. We define the integer part $\boldsymbol{x}^M := \begin{pmatrix} \lfloor x^1 + \frac{1}{2} \rfloor \\ \lfloor x^2 + \frac{1}{2} \rfloor \end{pmatrix}$ of any point $\boldsymbol{x} = \begin{pmatrix} x^1 \\ x^2 \end{pmatrix} \in \Omega$ which is the midpoint of the shifted unit cell containing $\boldsymbol{x}$. Note as a remark that these midpoints are situated at the integer grid points $(i,j) \in \mathbb{N}^2 \cap \Omega$. Moreover we denote the relative shift from the midpoint $\boldsymbol{x}^M$ of the corresponding cell by $\boldsymbol{x}^R := \boldsymbol{x} - \boldsymbol{x}^M$. Thus we may describe the evolution of the microstructure in Ω_ε by the level set function S_ε which is defined by

$$S_\varepsilon(t, \boldsymbol{x}) = R\left(t, \left(\frac{\boldsymbol{x}}{\varepsilon}\right)^M\right) - \left|\left|\left(\frac{\boldsymbol{x}}{\varepsilon}\right)^R\right|\right|. \tag{3}$$

2.3 Formal Upscaling

Our main goal is now to determine a macroscopic model description starting from system (1), (2). Therefore we intend to identify at least formally the limit $\varepsilon \to 0$. A widespread used method is the method of two-scale asymptotic expansion. A mathematical introduction can be found in [4], [16] or in [5], [7]. Concerning the scale separation, besides the global variable $\boldsymbol{x}$, a microscopic variable $\boldsymbol{y}$ is introduced. Both are connected via the relation $\boldsymbol{y} = \boldsymbol{x}/\varepsilon$. As a consequence, the expansion of the spatial gradient reads

$$\nabla_\varepsilon = \nabla_{\boldsymbol{x}} + \frac{1}{\varepsilon}\nabla_{\boldsymbol{y}}.$$

Higher order spatial derivatives may be calculated analogously applying the chain rule. Furthermore it is assumed that all variable functions can be expanded in series of the

 ISBN: 978-84-614-6167-7

scale parameter ε, i.e.

$$\phi_\varepsilon(t,\boldsymbol{x}) = \phi_0(t,\boldsymbol{x},\boldsymbol{y})+\varepsilon\phi_1(t,\boldsymbol{x},\boldsymbol{y})+\varepsilon^2\phi_2(t,\boldsymbol{x},\boldsymbol{y})+\dots,\ \boldsymbol{y}=\boldsymbol{x}/\varepsilon,\ \phi_\varepsilon\in\{c_\varepsilon,v_\varepsilon,\Phi_\varepsilon\}. \quad (4)$$

Additionally to the standard expansions introduced above in the framework of a level set description also the level set function S_ε itself and the normal vector $\boldsymbol{\nu}_\varepsilon$ have to be expanded due to the evolving microstructure. For a two-dimensional setting the expansion of the normal vector can be expressed in terms of the level set function and we obtain the following expressions, [18]:

$$S_\varepsilon(t,\boldsymbol{x}) = S_0(t,\boldsymbol{x},\boldsymbol{y}) + \varepsilon S_1(t,\boldsymbol{x},\boldsymbol{y}) + \varepsilon^2 S_2(t,\boldsymbol{x},\boldsymbol{y}) + \dots, \quad \boldsymbol{y}=\boldsymbol{x}/\varepsilon, \quad (5)$$

$$\boldsymbol{\nu}_\varepsilon = \boldsymbol{\nu}_0 + \varepsilon\boldsymbol{\nu}_1 + O(\varepsilon^2), \qquad \boldsymbol{\nu}_0 = \frac{\nabla_{\boldsymbol{y}} S_0}{|\nabla_{\boldsymbol{y}} S_0|}, \qquad \boldsymbol{\nu}_1 = \tau_0 \frac{\tau_0\cdot(\nabla_{\boldsymbol{x}} S_0 + \nabla_{\boldsymbol{y}} S_1)}{|\nabla_{\boldsymbol{y}} S_0|} \quad (6)$$

with $\boldsymbol{\tau}_0 := \boldsymbol{\nu}_0^\perp$ denoting the orthogonal complement of $\boldsymbol{\nu}_0$ in the two dimensional space. We now state and proof our main upscaling result:

Theorem 2.1. *The homogenized model of (1) - (2) consists of Darcy's law which describes the averaged water movement*

$$\begin{aligned} \overline{\boldsymbol{v}}_0 &= -K(t,\boldsymbol{x})\nabla_{\boldsymbol{x}} p_0 && \boldsymbol{x}\in\Omega \\ \nabla_{\boldsymbol{x}}\cdot\overline{\boldsymbol{v}}_0 &= -|\Gamma_0(t,\boldsymbol{x})|\partial_t R_0 && \boldsymbol{x}\in\Omega \end{aligned}$$

and is supplemented by a family of cell problem of Stokes type (7a). The transport of the transformed macroscopic concentration $u_0^\pm$ is described by an averaged convection-diffusion-equation

$$\partial_t(A(t,\boldsymbol{x})u_0^\pm) + \nabla_{\boldsymbol{x}}\cdot(\overline{\boldsymbol{V}}(t,\boldsymbol{x})u_0^\pm) - \nabla_{\boldsymbol{x}}\cdot(\overline{\mathbf{D}}(t,\boldsymbol{x})\nabla_{\boldsymbol{x}}u_0^\pm) + |\Gamma_0(t,\boldsymbol{x})|\partial_t R_0\rho_s = 0 \quad x\in\Omega$$

which is supplemented by two families of cell problems (9). Furthermore, the averaged coefficient functions are defined by (8) and (10).

Proof. We insert the expansions (4), (5) and (6) in (1), (2) and analyze the different orders in $\varepsilon^k,\ k\in\mathbb{Z}$:

Lowest Order Problems: The lowest order problem of the Stokes' equation is of order ε^{-1} and yields that p_0 is a macroscopic variable, i.e. $p_0(\boldsymbol{x},\boldsymbol{y}) = p_0(x)$ since $\nabla_{\boldsymbol{y}} p_0 = 0$. The Nernst-Planck equation of order ε^{-2} with corresponding boundary condition of order ε^{-1} reads

$$\begin{aligned} -\nabla_{\boldsymbol{y}}\cdot(\nabla_{\boldsymbol{y}} c_0 + c_0\nabla_{\boldsymbol{y}}\Phi_0) &= 0 \\ (\nabla_{\boldsymbol{y}} c_0 + c_0\nabla_{\boldsymbol{y}}\Phi_0)\cdot\boldsymbol{\nu}_0 &= 0. \end{aligned}$$

Applying the transformation $c_0(\boldsymbol{x},\boldsymbol{y}) = e^{-\Phi_0(\boldsymbol{x},\boldsymbol{y})}u_0(\boldsymbol{x},\boldsymbol{y})$ which is well known from the stationary theory of semiconductor devices [9], [15] leads to the following problem

$$\begin{aligned} -\nabla_{\boldsymbol{y}}\cdot(e^{-\Phi_0(\boldsymbol{x},\boldsymbol{y})}\nabla_{\boldsymbol{y}} u_0) &= 0 \\ e^{-\Phi_0(\boldsymbol{x},\boldsymbol{y})}\nabla_{\boldsymbol{y}} u_0\cdot\boldsymbol{\nu}_0 &= 0 \end{aligned}$$

 ISBN: 978-84-614-6167-7

which is uniquely solvable up to a constant (with respect to $\boldsymbol{y}$). Obviously every macroscopic solution $u_0(\boldsymbol{x})$ is a solution of this problem and therefore we derive the following form for the leading order concentration term: $c_0(\boldsymbol{x},\boldsymbol{y}) = e^{-\Phi_0(\boldsymbol{x},\boldsymbol{y})}u_0(\boldsymbol{x})$.

Next Order Problems: The next order problem for Stokes' equation is of order ε^0 supplemented by the incompressibility condition of order ε^{-1} and boundary condition of order ε^0:

$$\begin{aligned} -\Delta_y \boldsymbol{v}_0 + \nabla_{\boldsymbol{y}}(p_1 + c_0\Phi_0) &= -\nabla_{\boldsymbol{x}} p_0 \\ \nabla_{\boldsymbol{y}} \cdot \boldsymbol{v}_0 &= 0 \\ \boldsymbol{v}_0 &= \boldsymbol{0} \end{aligned}$$

With the definition of the modified pressure $\tilde{p}_1 := p_1 + c_0\Phi_0$ one proceeds as in [7] to set up the following $j = 1, 2$ cell problems of Stokes type:

$$-\Delta_{\boldsymbol{y}} \boldsymbol{w}_j + \nabla_{\boldsymbol{y}} \pi_j = e_j \qquad \text{in } Y_0(t,\boldsymbol{x}), \tag{7a}$$

$$\nabla_{\boldsymbol{y}} \cdot \boldsymbol{w}_j = 0 \qquad \text{in } Y_0(t,\boldsymbol{x}), \tag{7b}$$

$$\boldsymbol{w}_j = 0 \qquad \text{in } \Gamma_0(t,\boldsymbol{x}), \tag{7c}$$

$$\boldsymbol{w}_j \qquad \text{Y-periodic}, \tag{7d}$$

where $Y_0(t,\boldsymbol{x}) := \{\boldsymbol{y} : S_0(t,\boldsymbol{x},\boldsymbol{y}) < 0\}$, $\Gamma_0(t,\boldsymbol{x},\boldsymbol{y}) := \{\boldsymbol{y} : S_0(t,\boldsymbol{x},\boldsymbol{y}) = 0\}$, [18]. Following [7] the leading order velocity term can be expressed via $\boldsymbol{v}_0(\boldsymbol{x},\boldsymbol{y}) = -\sum_j \boldsymbol{w}_j \partial_{x_j} p_0$. For the macroscopic velocity $\overline{\boldsymbol{v}}_0(x)$ we derive Darcy's law, i. e. $\overline{\boldsymbol{v}}_0 := \int_{Y_0(t,\boldsymbol{x})} \boldsymbol{v}_0(\boldsymbol{x},\boldsymbol{y})dy = -\mathbf{K}\nabla_{\boldsymbol{x}} p_0$ with permeability tensor

$$\mathbf{K}_{ij} := \int_{Y_0(t,\boldsymbol{x})} w_j^i dy. \tag{8}$$

The supplementary compressibility condition has been derived in [18].
For the Nernst-Planck equation we consider the problem of order ε^{-1} with boundary condition of order ε^0. Since the boundary conditions have to be applied on $\Gamma_0(t,\boldsymbol{x})$ additional terms show up due to the evolution of the pore space. An abstract derivation of how to set up the right boundary condition can be found in [18].

$$\begin{aligned} \boldsymbol{v}_0 \cdot \nabla_{\boldsymbol{y}} c_0 - \nabla_{\boldsymbol{y}} \cdot (\nabla_{\boldsymbol{y}} c_1 + \nabla_{\boldsymbol{x}} c_0 + c_0 \nabla_{\boldsymbol{y}} \Phi_1 + c_1 \nabla_{\boldsymbol{y}} \Phi_0 + c_0 \nabla_{\boldsymbol{x}} \Phi_0) - \nabla_{\boldsymbol{x}} \cdot (\nabla_{\boldsymbol{y}} c_0 + c_0 \nabla_{\boldsymbol{y}} \Phi_0) = 0 \\ (-\boldsymbol{v}_0 c_0 + \nabla_{\boldsymbol{y}} c_1 + \nabla_{\boldsymbol{x}} c_0 + c_0 \nabla_{\boldsymbol{y}} \Phi_1 + c_1 \nabla_{\boldsymbol{y}} \Phi_0 + c_0 \nabla_{\boldsymbol{x}} \Phi_0) \cdot \boldsymbol{\nu}_0 + (\nabla_{\boldsymbol{y}} c_0 + c_0 \nabla_{\boldsymbol{y}} \Phi_0) \cdot \boldsymbol{\nu}_1 \\ + y \cdot \nabla_{\boldsymbol{x}} (\nabla_{\boldsymbol{y}} c_0 + c_0 \nabla_{\boldsymbol{y}} \Phi_0) \cdot \boldsymbol{\nu}_0 + \lambda \boldsymbol{\nu}_0 \cdot \nabla_{\boldsymbol{y}} (\nabla_{\boldsymbol{y}} c_0 + c_0 \nabla_{\boldsymbol{y}} \Phi_0) \cdot \boldsymbol{\nu}_0 = 0 \end{aligned}$$

Using the transformation for c_0 and an analogous transformation for c_1, namely $c_1(\boldsymbol{x},\boldsymbol{y}) = e^{-\Phi_0(\boldsymbol{x},\boldsymbol{y})}u_1(\boldsymbol{x},\boldsymbol{y})$ leads to the following problem:

$$\begin{aligned} -\nabla_{\boldsymbol{y}} \cdot (e^{-\Phi_0} \nabla_{\boldsymbol{y}} (u_1 + u_0\Phi_1)) &= -\boldsymbol{v}_0 \cdot \nabla_{\boldsymbol{y}} (e^{-\Phi_0} u_0) + \nabla_{\boldsymbol{y}} \cdot (e^{-\Phi_0} \nabla_{\boldsymbol{x}} u_0) \\ e^{-\Phi_0} \nabla_{\boldsymbol{y}} (u_1 + u_0\Phi_1) \cdot \boldsymbol{\nu} &= (\boldsymbol{v}_0 e^{-\Phi_0} u_0 - e^{-\Phi_0} \nabla_{\boldsymbol{x}} u_0) \cdot \boldsymbol{\nu} \end{aligned}$$

Defining the modified transformed first order concentration term $\tilde{u}_1 := u_1 + u_0\Phi_1$ and using the incompressibility and boundary condition of Stokes' equation leads to

$$\begin{aligned} -\nabla_{\boldsymbol{y}} \cdot (e^{-\Phi_0} \nabla_{\boldsymbol{y}} \tilde{u}_1) &= \nabla_{\boldsymbol{y}} \cdot (-e^{-\Phi_0} \boldsymbol{v}_0 u_0 + e^{-\Phi_0} \nabla_{\boldsymbol{x}} u_0) \\ e^{-\Phi_0} \nabla_{\boldsymbol{y}} \tilde{u}_1 \cdot \boldsymbol{\nu} &= -e^{-\Phi_0} \nabla_{\boldsymbol{x}} u_0 \cdot \boldsymbol{\nu}_0 \end{aligned}$$

 ISBN: 978-84-614-6167-7

We may now define two families of $j = 1, 2$ cell problems, one for every forcing term on the right hand side. Using furthermore the representation of v_0 in terms of the solutions of the cell problems w_j and their properties we derive

$$-\nabla_{\boldsymbol{y}}\cdot(e^{-\Phi_0}\nabla_{\boldsymbol{y}}\zeta_j^1)=\nabla_{\boldsymbol{y}}\cdot(e^{-\Phi_0}\boldsymbol{w}_j),\quad -\nabla_{\boldsymbol{y}}\cdot(e^{-\Phi_0}\nabla_{\boldsymbol{y}}\zeta_j^2)=\nabla_{\boldsymbol{y}}\cdot(e^{-\Phi_0}\boldsymbol{e}_j)\text{ in } Y_0(t,\boldsymbol{x}),\quad (9a)$$

$$(e^{-\Phi_0}\nabla_{\boldsymbol{y}}\zeta_j^1)\cdot\boldsymbol{\nu}_0=0,\qquad e^{-\Phi_0}\nabla_{\boldsymbol{y}}\zeta_j^2\cdot\boldsymbol{\nu}_0=-e^{-\Phi_0}\boldsymbol{e}_j\cdot\boldsymbol{\nu}_0\text{ in }\Gamma_0(t,\boldsymbol{x}),\quad (9b)$$

$$\zeta_j^1,\qquad \zeta_j^2\qquad \text{Y-periodic.}\quad (9c)$$

The modified first order term $\tilde{u}_1$ can then be expressed by linearity in the following way: $\tilde{u}_1 = \sum_j \zeta_j^1(\partial_{x_j}p_0)u_0 + \zeta_j^2\partial_{x_j}u_0$.

Zero Order Problems: The ε^0-order Nernst-Planck equation with boundary condition of order ε^1 reads with the additional term due to the evolving pore geometry analogously to [18]

$$\begin{aligned}
&\partial_t c_0 + \boldsymbol{v}_0\cdot\nabla_{\boldsymbol{x}}c_0 + \boldsymbol{v}_1\cdot\nabla_{\boldsymbol{y}}c_0 + \boldsymbol{v}_0\cdot\nabla_{\boldsymbol{y}}c_1 - \nabla_{\boldsymbol{x}}\cdot(\nabla_{\boldsymbol{x}}c_0+\nabla_{\boldsymbol{y}}c_1+c_0\nabla_{\boldsymbol{x}}\Phi_0+c_0\nabla_{\boldsymbol{y}}\Phi_1+c_1\nabla_{\boldsymbol{y}}\Phi_0)\\
&-\nabla_{\boldsymbol{y}}\cdot(\nabla_{\boldsymbol{x}}c_1+\nabla_{\boldsymbol{y}}c_2+c_0\nabla_{\boldsymbol{x}}\Phi_1+c_1\nabla_{\boldsymbol{x}}\Phi_0+c_1\nabla_{\boldsymbol{y}}\Phi_1+c_2\nabla_{\boldsymbol{y}}\Phi_0+c_0\nabla_{\boldsymbol{y}}\Phi_2)=0\\
&(-\boldsymbol{v}_1c_0-\boldsymbol{v}_0c_1+\nabla_{\boldsymbol{x}}c_1+\nabla_{\boldsymbol{y}}c_2+c_0\nabla_{\boldsymbol{x}}\Phi_1+c_1\nabla_{\boldsymbol{x}}\Phi_0+c_1\nabla_{\boldsymbol{y}}\Phi_1+c_2\nabla_{\boldsymbol{y}}\Phi_0+c_0\nabla_{\boldsymbol{y}}\Phi_2)\cdot\boldsymbol{\nu}_0\\
&+(-\boldsymbol{v}_0e^{-\Phi_0}u_0+e^{-\Phi_0}\nabla_{\boldsymbol{y}}(u_1+u_0\Phi_1)+e^{-\Phi_0}\nabla_{\boldsymbol{x}}u_0)\cdot\boldsymbol{\nu}_1\\
&+y\cdot\nabla_{\boldsymbol{x}}(-\boldsymbol{v}_0e^{-\Phi_0}u_0+e^{-\Phi_0}\nabla_{\boldsymbol{y}}(u_1+u_0\Phi_1)+e^{-\Phi_0}\nabla_{\boldsymbol{x}}u_0)\cdot\boldsymbol{\nu}_0\\
&+\lambda\boldsymbol{\nu}_0\cdot\nabla_{\boldsymbol{y}}(-\boldsymbol{v}_0e^{-\Phi_0}u_0+e^{-\Phi_0}\nabla_{\boldsymbol{y}}(u_1+u_0\Phi_1)+e^{-\Phi_0}\nabla_{\boldsymbol{x}}u_0)\cdot\boldsymbol{\nu}_0+\partial_tR_0(\rho_s-c_0)=0
\end{aligned}$$

Integrating with respect to y and applying the boundary condition and the transformations for c_0 and c_1, resp., gives

$$\begin{aligned}
&\int_{Y_0(t,\boldsymbol{x})}\partial_t(e^{-\Phi_0}u_0)dy+\int_Y\nabla_{\boldsymbol{x}}\cdot(\boldsymbol{v}_0e^{-\Phi_0}u_0-e^{-\Phi_0}\nabla_{\boldsymbol{x}}u_0-e^{-\Phi_0}\nabla_{\boldsymbol{y}}\tilde{u}_1)dy\\
&+\int_{\Gamma_0(t,\boldsymbol{x})}(-\boldsymbol{v}_0e^{-\Phi_0}u_0+e^{-\Phi_0}\nabla_{\boldsymbol{y}}(u_1+u_0\Phi_1)+e^{-\Phi_0}\nabla_{\boldsymbol{x}}u_0)\cdot\boldsymbol{\nu}_1\\
&+y\cdot\nabla_{\boldsymbol{x}}((-\boldsymbol{v}_0e^{-\Phi_0}u_0+e^{-\Phi_0}\nabla_{\boldsymbol{y}}(u_1+u_0\Phi_1)+e^{-\Phi_0}\nabla_{\boldsymbol{x}}u_0)\cdot\boldsymbol{\nu}_0)\\
&+\lambda\boldsymbol{\nu}_0\cdot\nabla_{\boldsymbol{y}}((-\boldsymbol{v}_0e^{-\Phi_0}u_0+e^{-\Phi_0}\nabla_{\boldsymbol{y}}(u_1+u_0\Phi_1)+e^{-\Phi_0}\nabla_{\boldsymbol{x}}u_0)\cdot\boldsymbol{\nu}_0)+\partial_tR_0(\rho_s-c_0)=0
\end{aligned}$$

Applying the transport theorem in order to interchange integration and spatial derivation we get

$$\begin{aligned}
&\int_{Y_0(t,\boldsymbol{x})}\partial_t(e^{-\Phi_0}u_0)dy+\nabla_{\boldsymbol{x}}\cdot\int_Y\boldsymbol{v}_0e^{-\Phi_0}u_0-e^{-\Phi_0}\nabla_{\boldsymbol{x}}u_0-e^{-\Phi_0}\nabla_{\boldsymbol{y}}\tilde{u}_1dy\\
&+\int_{\Gamma_0(t,\boldsymbol{x})}(-\boldsymbol{v}_0e^{-\Phi_0}u_0+e^{-\Phi_0}\nabla_{\boldsymbol{y}}(u_1+u_0\Phi_1)+e^{-\Phi_0}\nabla_{\boldsymbol{x}}u_0)\cdot\boldsymbol{\nu}_1do_y\\
&+\int_{\Gamma_0(t,\boldsymbol{x})}y\cdot\nabla_{\boldsymbol{x}}(-\boldsymbol{v}_0e^{-\Phi_0}u_0+e^{-\Phi_0}\nabla_{\boldsymbol{y}}(u_1+u_0\Phi_1)+e^{-\Phi_0}\nabla_{\boldsymbol{x}}u_0)\cdot\boldsymbol{\nu}_0do_y\\
&+\int_{\Gamma_0(t,\boldsymbol{x})}\lambda\boldsymbol{\nu}_0\cdot\nabla_{\boldsymbol{y}}(-\boldsymbol{v}_0e^{-\Phi_0}u_0+e^{-\Phi_0}\nabla_{\boldsymbol{y}}(u_1+u_0\Phi_1)+e^{-\Phi_0}\nabla_{\boldsymbol{x}}u_0)\cdot\boldsymbol{\nu}_0+\partial_tR_0(\rho_s-c_0)do_y\\
&+\int_{\Gamma_0(t,\boldsymbol{x})}\frac{\nabla_{\boldsymbol{x}}S_0}{|\nabla_{\boldsymbol{y}}S_0|}(\boldsymbol{v}_0e^{-\Phi_0}u_0-e^{-\Phi_0}\nabla_{\boldsymbol{x}}u_0-e^{-\Phi_0}\nabla_{\boldsymbol{y}}\tilde{u}_1)do_y=0
\end{aligned}$$

 ISBN: 978-84-614-6167-7

With the help of the representation of the normal vectors and the lemmas in [18] several terms cancel and we get

$$\int_{Y_0(t,\boldsymbol{x})} \partial_t(e^{-\Phi_0}u_0)dy + \nabla_{\boldsymbol{x}}\cdot\int_Y \boldsymbol{v}_0 e^{-\Phi_0}u_0 - e^{-\Phi_0}\nabla_{\boldsymbol{x}}u_0 - e^{-\Phi_0}\nabla_{\boldsymbol{y}}\tilde{u}_1 dy + \int_{\Gamma_0(t,\boldsymbol{x})} \partial_t R_0(\rho_s - c_0)do_y = 0$$

Using the transport theorem for the time derivative and the boundary condition for the normal velocity with appropriate scaling we get

$$\partial_t\left(\int_{Y_0(t,\boldsymbol{x})} e^{-\Phi_0}u_0 dy\right) + \nabla_{\boldsymbol{x}}\cdot\int_Y \boldsymbol{v}_0 e^{-\Phi_0}u_0 - e^{-\Phi_0}\nabla_{\boldsymbol{x}}u_0 - e^{-\Phi_0}\nabla_{\boldsymbol{y}}\tilde{u}_1 dy + \Gamma_0(t,\boldsymbol{x})\partial_t R_0\rho_s = 0$$

Here the cell problems derived above may be inserted as usual and we derive

$$\partial_t\left(Au_0\right) + \nabla_{\boldsymbol{x}}\cdot\left(\boldsymbol{V}u_0 - \overline{\mathbf{D}}\nabla_{\boldsymbol{x}}u_0\right) + \Gamma_0(t,\boldsymbol{x})\partial_t R_0\rho_s = 0$$

with averaged parameter defined by

$$A := \int_{Y_0(t,\boldsymbol{x})} e^{-\Phi_0}dy, \qquad \overline{\boldsymbol{V}} := -\int_{Y_0(t,\boldsymbol{x})} e^{-\Phi_0}\sum_j(\boldsymbol{w}_j + \nabla_{\boldsymbol{y}}\zeta_j^1)\partial_{x_j}p_0 dy, \tag{10a}$$

$$\overline{\mathbf{D}}_{ij} := \int_{Y_0(t,\boldsymbol{x})} e^{-\Phi_0}(\partial_{y_i}\zeta_j^2 + \delta_{ij})dy. \tag{10b}$$

□

2.4 Properties of Coefficient Functions

Theorem 2.2. *The porosity defined in (10a) is strictly positive, the permeability tensor defined in (8) is symmetric and positive definite. The diffusion tensor defined in (10b) is symmetric, positive definite and elliptic.*

Proof. The statement for the porosity is clear by definition. The statement for the permeability tensor $\mathbf{K}$ can be proven directly as in [7]. To proof the statement for the diffusion tensor the ideas in [7] and [5] can be applied with only slight modifications since the coefficient $e^{-\Phi_0}$ is strictly positive. □

Theorem 2.3. *If that there is no interaction and also no evolving microstructure, the averaged coefficient functions $A, \overline{\boldsymbol{V}}, \overline{\mathbf{D}}$ given in (10) reduce to*

$$A = |Y|, \quad \overline{\boldsymbol{V}} = \int_Y \boldsymbol{v}_0 dy = \overline{\boldsymbol{v}}_0, \quad \textit{and} \quad \overline{D}_{ij} = \int_Y (\partial_{y_i}\zeta_j^2 + \delta_{ij})dy$$

Proof. The occurrence of no interaction results in $e^{-\Phi_0} = e^0 = 1$. Furthermore the solutions ζ_j^1 of the additional family of cell problems are equal to zero. With the relation $\boldsymbol{v}_0 = -\sum_j w_j\partial_{x_j}p_0$ the statement of theorem 2.3 holds. □

 ISBN: 978-84-614-6167-7

Remark 2. *The homogenized model of theorem 2.1 is therefore consistent with the standard model for flow and transport in porous media without interaction and evolving microstructure. Theorem 2.3 shows that the coefficient functions reduce and obviously also the homogenized partial differential equation reduce to the well known two-scale description that can be found for example in [7].*

3 Conclusion

Applying the formal homogenization technique that is capable of an evolving microstructure by the level set formulation to our system of partial differential equations (1), (2), we obtain Darcy's law for a compressible fluid and a modified averaged convection-diffusion equation. These equations are supplemented by different families of microscopic cell problems determining averaged coefficient functions. The changes in the microscopic geometry finally result in a change of the porosity and permeability in the equivalent macroscopic description given in Theorem 2.1. Therefore a recoupling of the transport processes to the fluid phase takes place. The next step is to verify this theoretical model against experimental data.

Acknowledgements

N. Ray has been supported by Deutsche Telekom Stiftung.

References

[1] G. Allaire, A. Mikelić and A. Piatnitski, *Homogenization of the linearized ionic transport equations in rigid periodic porous media*, J. Math. Phys., **51** (123103), 2010.

[2] J.-L. Auriault and J. Lewandowska, *On the Cross-Effects of Coupled Macroscopic Transport Equations in Porous Media*, Transp. Porous Media, **16** (1994), 31–52.

[3] M. Bause, W. Friess, P. Knabner and F. Radu, *A comprehensive mathematical model describing drug release from collagen matrices*, Mathematical Modelling on Environmental and Life Sciences Problems. S. Ion et al. (editors), Editura Academiei Romane, Bucuresti, 2007, 26–34

[4] A. Bensoussan, J.-L. Lions and G. Papanicolau, "Asymptotic analysis of periodic structures", North-Holland,1978.

[5] D. Cioranescu and P. Donato, "An Introduction to Homogenization", Oxford University Press, 2000.

[6] M. Elimelech, J. Gregory, X. Jia, and R. Williams, "Particle Deposition and Aggregation, Measurement, Modelling and Simulation", Butterworth-Heinemann, 1995.

 ISBN: 978-84-614-6167-7

[7] U. Hornung, "Homogenization and Porous Media", Springer, 1997.

[8] J. Looker, "The Electrokinetics of Porous Colloidal Particles", Ph.D thesis, University of Melbourne, 2006.

[9] P. Markovich, "The Stationary Semiconductor Device Equations", Springer, 1986.

[10] I. Metzmacher, F. Radu, M. Bause, P. Knabner and W. Friess, *A Model Describing the Effect of Enzymatic Degradation on Drug Release from Collagen Minirods*, European Journal of Pharmaceutics and Biopharmaceutics, 67 (2007), 349-360

[11] C. Moyne and M. Murad, *Electro-Chemo-Mechanical Couplings in swelling Clays derived from a micro/macro-homogenization procedure*, International Journal of Solids and Structures, **39** (2006), 6159–6190.

[12] C. Moyne and M. Murad, *A Two-scale Model for coupled Electro-Chemo-Mechanical Phenomena and Onsager's Reciprocity Relation in Expansive Clays: I Homogenization Results*, Transp. Porous Media, **62** (2006), 333–380.

[13] F. Radu, M. Bause, P. Knabner, W. Friess and G. Lee, *Modelling of drug release from collagen matrices*, J. Pharm. Sci, **91** (2002), 964-972.

[14] F. Radu, M. Bause, P. Knabner, W. Friess and I. Metzmacher, *Numerical Simulation of Drug Release from Collagen Matrices by Enzymatic Degradation*, Comput. Vis. Sci., **12** (2007), 409420.

[15] T. Roubíček, "Nonlinear partial differential Equations with Applications", Birkhäuser, 2005.

[16] E. Sánchez-Palancia, "Nonhomogeneous media and vibration theory", Springer, 1980.

[17] T. van de Ven, "Colloidal Hydrodynamics", Academic Press, 1989.

[18] T. van Noorden, *Crystal precipitation and dissolution in a porous medium: effective equations and numerical experiments*, Multiscale Model. Simul., **7** (2009), 1220–1236.

[19] T. van Noorden and Adrian Muntean, *Homogenization of a locally-periodic medium with areas of low and high diffusivity*, CASA report 10-19, Eindhoven Universoty of Technology, 2010.

[20] M. Schmuck, "Modeling, Analysis and Numerics in Electrohydrodynamics", Ph.D thesis, University Tubingen, 2008.

[21] M. Schmuck, *Modeling and derinving porous media Stokes-Poisson-Nernst-Planckk equations by a multiple-scale appraoch*, Commun. Math. Sci., **3(9)** (2010), 685-710.

[22] T. Weinstein, "Three-phase hybrid mixture theory for swelling drug delivery systems", Ph.D thesis, University of Colorado at Denver, 2005.

 ISBN: 978-84-614-6167-7

Proceedings of the 11th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2011
26–30 June 2011.

How fast do stock prices adjust to market efficiency? Insights from detrended fluctuation analysis

Miguel A. Rivera-Castro[1], Juan Carlos Reboredo Nogueira[1] and Raquel García-Rubio[2]

[1] *Department of Economics, University of Santiago de Compostela - Spain*

[2] *Department of Finance, University of Salamanca - Spain*

emails: marc@ufba.br, juancarlos.reboredo@usc.es, rgr@usal.es

Abstract

In this paper we analyse price fluctuations with the aim of measuring how long the market takes to adjust prices to weak-form efficiency, i.e., how long it takes to become prices in a fractional Brownian motion with a Hurst exponent of 0.5. The Hurst exponent is estimated for different time horizons using detrended fluctuation analysis-an appropriate method for non-stationary series with trends-in order to identify at which time scale the Hurst exponent is consistent with the efficient market hypothesis. Using high frequency data for stock indices, exchange rates and security prices, we show that price dynamics exhibited important deviations from efficiency for time periods that lasted up to 10 minutes, but after this time the efficiency creating process ended and price dynamics was consistent with a geometric Brownian motion. Additionally, the intraday behaviour of the series corroborates that, in the opening and closing trading hours, price dynamics is hardly consistent with efficiency, thereby allowing investors to exploit price deviations from fundamental values. This result is consistent with the intraday pattern of volume, volatility and transaction time durations.

Key words: Market Efficiency, Hurst, Detrenden Fluctuation Analysis.

1 Introduction

The market efficiency hypothesis [10, 11]states that asset prices adjust to fully reflect all available information. Although the formulation of this hypothesis refers to a rapid and unbiased price adjustment process, in practice, prices tend not to adjust to new information instantly but after a certain amount of time. During this time investors take actions

 ISBN: 978-84-614-6167-7

to exploit temporary profit opportunities arising from new information, ultimately pushing prices towards efficiency. In fact, the speed of adjustment to market efficiency is an important dimension of the market efficiency hypothesis (see, e.g.,[7]). The adjustment of asset prices to information has been widely studied at the theoretical and empirical level. [12, 13, 8] developed models in which the incorporation of information in stock prices depends on the cost of information production. [5, 3] in a rational expectation framework, show how prices adjust in a sequence of trades to fully reveal all relevant information. In a model populated by Bayesian traders [6] found, in a simulation study, that the market usually converges more rapidly to an equilibrium price when arbitrageurs react to one another. Behavioural finance models have been developed by [2, 9, 14] to provide explanations for the empirically documented under- and over-reactions of stock prices to news. Empirically, several studies have examined market efficiency in terms of the speed with which prices react to new information arising from specific events (see, e.g., [4]) while other empirical studies have examined the speed of convergence of prices to efficiency in a more general setting, without identifying any specific new event (see, e.g.,[1]). In this paper we use a novel approach to address the problem of measuring the adjustment time of security prices towards weak-form market efficiency, based on using detrended fluctuation analysis (DFA) [16] for different time scales in order to identity the time horizon necessary for prices to adjust to a fBm with a Hurst exponent of 0.5.

2 Metodology

DFA, which was proposed by [16, 17] enables identification of a quantitative parameter that represents the correlation properties of a signal. DFA is based on analysing fluctuations in a time series at different scales. In practice, it removes the trend from a time series in different scales by analysing intrinsic fluctuations in the data. Fluctuations are, in effect a measure of variability in the signal and are associated with the variance for each segment in the series at different scales. The DFA algorithm implies six basic steps.

1. If we start with a series of equidistant time increments $\{x(t)\}, t = 1, ..., N$, we can obtain the path (or the profile)

$$y(t) = \sum_{j=1}^{t} x(t). \tag{1}$$

2. The entire interval $[1, N]$ can be divided into a series of M_ν boxes of length ν, not necessarily self-excluding. Each of such boxes receives a label $(m, \nu), m = 1, ..., M_\nu$. In our calculations, we considered a certain level of overlap between the boxes for the purpose of increasing the number of boxes where the method is applied and, hence,

 ISBN: 978-84-614-6167-7

to improve the statistics. To evaluate the magnitude of fluctuations in the box (m, ν) and, concomitantly, eliminate the trend of order q, we consider the difference

$$y_s(t) = y(t) - p_q(t; (m, \nu)), \tag{2}$$

3. where $p_q(t; (m, \nu))$ represents the polynomial of order q that minimizes the sum of $y_s(t)^2$ when t spans all points of the considered box. To be more precise, we consider the residue

$$f(m, \nu) = \frac{1}{\nu} \sum_{j=I_{min}(m,\nu)}^{I_{max}(m,\nu)} y_s^2(j), \tag{3}$$

4. where $I_{min}(m, \nu)$ and $I_{max}(m, \nu)$ are the lower and upper limit of the (m, ν) box. When $q = 0$, $f(m, \nu)$ corresponds to the roughness function $W(m, \nu)$ of the (m, ν) box. Subsequently, we consider the average

$$F(\nu) = \Big[\frac{1}{M_\nu} \sum_{m=1}^{M_\nu} f(m, \nu)\Big]^{1/2}, \tag{4}$$

5. which expresses the average detrended roughness at length scale ν of the entire profile. If the original series presents long-range correlations, it is expected that the values of $F(\nu)$ follows a power law

$$F(\nu) \sim \nu^H, \tag{5}$$

6. where the roughness exponent $H = 1 - \gamma/2$ is related to the exponent describing the decay of the correlation function $C(j) = E[y(t)y(t + j)] \sim j^{-\gamma}$.

In practice, this means the exponent can be calculated, with a linear adjustment in the logarithmic scale of ν in function of $F(\nu)$. The fluctuation exponent can be classified according to a dynamic band of values:

- $H < 0.5$: anti-correlated, anti-persistent signal.
- $H = 0.5$: non-correlated, white noise, no memory.
- $H > 0.5$: has long-range correlations.

 ISBN: 978-84-614-6167-7

3 Data

Intraday data for 35 days of trading in the DJIA and SP500 markets, the EUR-USD exchange rate and Telefónica España shares between April and June 2011 were used to analyse the 1- to 15-minute series, and intraday data for 46 days of EUR-USD exchange rate operations between October and December 2010 were used to calculate the mean intraday Hurst exponent.

4 Results

It can be observed (see annex) in [Fig.1] that the series showed long-term correlations (persistence) in the times below 5 and 4 minutes, indicating no random walk and convergence with a fractional Brownian motion (fBm), from 5 minutes for the SP500 series and from 4 minutes for the DJIA series. Behaviour of the EUR-USD exchange rate series [Fig. 2] was initially anti-persistent, indicating reversion to the mean, but later converged with an fBm from minute 2. Telefónica share behaviour, reflected in [Fig.3], was initially anti-persistent but converged with an fBm from minute 6. Behaviour of the Hurst mean of the intraday series for a time period of 46 days is shown in [Fig. 4]. It can be observed that the series showed anti-persistent behaviour in the initial and final periods of each day's trading, indicating reversion to the mean. From the mornings hours, midday and afternoon, behaviour fluctuated to yield a Hurst exponent of around 0.5, indicating random walk in these periods. In view of the volatility of these series in the initial and final periods, overall behaviour, concave in form, makes sense, according, as it does, with microstructure theories [15].

5 Final Considerations

For intraday data for market, exchange rate and share indices, we quantified quotation speed of adjustment to an fBm using DFA. The series converged rapidly to a Hurst exponent of 0.5 as the time of the series increased. In terms of mean daily exchange rate behaviour, indices tended to converge with an fBm in the morning hours, whereas the trading opening and closing periods tended not to behave like an fBm. The possibility remains, therefore, of abnormal returns that are greater than the risk assumed, Furthermore, it would seem that share pricing tools that assume price behaviour to be like an fBm are not accurate.

Acknowledgements

We thank financial support from MTM2008-03010 project and research grant INCITE09201042PR.

 ISBN: 978-84-614-6167-7

References

[1] Amihud, Y., Mendelson, H. 1987. *Trading mechanisms and stock returns: An empirical investigation.* Journal of Finance, 52(3), 533-553.

[2] Barberis, N., Shleifer, A., Vishny, R. 1998. *A model of investor sentiment.* Journal of Financial Economics 49, 307-343.

[3] Bruce D. Grundy and Maureen McNichols. 1989. *Trade and the revelation of information through prices and direct disclosure.* Review of Financial Studies, 2 (4): pp. 495-526.

[4] Busse, J. A. and Green, T.C. 2002. *Market efficiency in real time.* Journal of Financial Economics, 65, 415-437.

[5] Brown, D.P., Jennings, R.E. 1989. *On technical analysis.* Review of Financial Studies 2, 527-551.

[6] Chakrabarti, R. Roll, R. 1999. *Learning from others, reacting, and market quality.* Journal of Financial Markets 2, 153-178.

[7] Chordia T., Roll R., Subrahmanyam A., 2005. *Evidence on the speed of convergence to market effciency.* Journal of Financial Economics, 76, 271–292.

[8] Cornell, B., Roll, R., 1981. *Strategies for pairwise competitions in markets and organizations.* Bell Journal of Economics, 712, 201–213.

[9] Daniel, K., Hirshleifer, D., Subrahmanyam, A. 1998. *Investor psychology and security market under- and overreactions.* Journal of Finance 56, 921-965.

[10] Fama, E. 1970. *Efficient capital markets: a review of theory and empirical work.* Journal of Finance 25, 383-417.

[11] Fama, E. 1991. *Efficient capital markets: II.* Journal of Finance 46, 1575-1617.

[12] Grossman, S. 1976. *On the efficiency of competitive stock markets where trades have diverse information.* Journal of Finance 31 (2), 573-585.

[13] Grossman, S., Stiglitz, J. 1980. *On the impossibility of informationally efficient markets.* American Economic Review 70, 393-408.

[14] Hong, H., Stein, J. 1999. *A unified theory of underreaction, momentum trading, and overreaction in assets markets.* Journal of Finance 54, 2143-2184.

[15] O'Hara, M., 1995. *Market Microstructure Theory.* Blackwell, Cambridge.

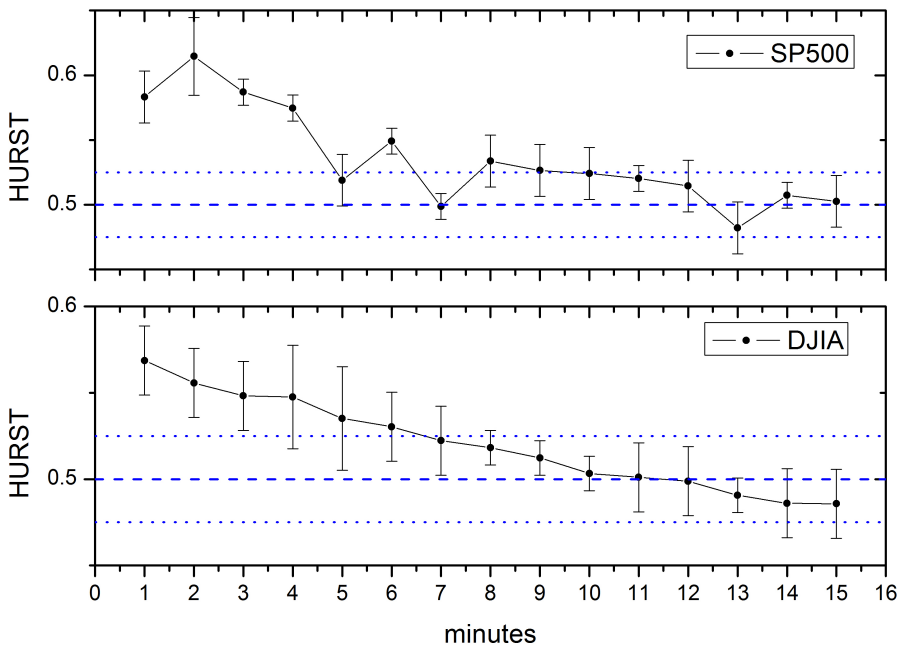
 ISBN: 978-84-614-6167-7

[16] Peng C-K, Buldyrev SV, Havlin S, Simons M, Stanley HE, Goldberger AL., 1994. *Mosaic organization of DNA nucleotides.* Phys Rev E, 49, 1685–1689.

[17] Peng C-K, Amaral LAN, Glass L, Hausdorff JM, Ivanov PCh, Mark RG, Mietus JE, Moody GB, Goldberger AL, Stanley HE., 2000. PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals. Circulation 101(23):e215-e220 [Circulation Electronic Pages; http://circ.ahajournals.org/cgi/content/full/101/23/e215], June 13.

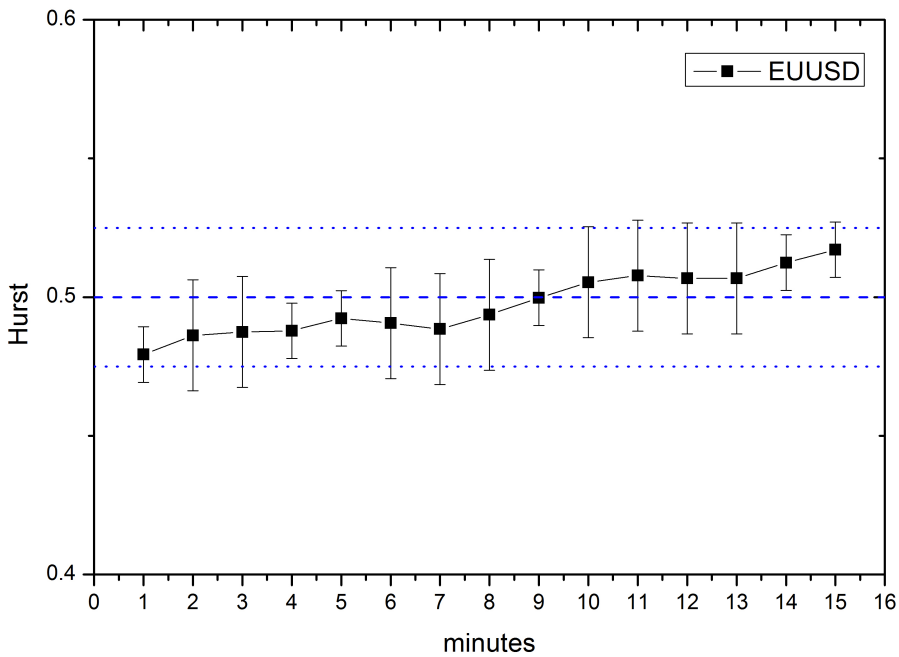
6 Annex

Fig. 1. Estimated Hurst minute by minute



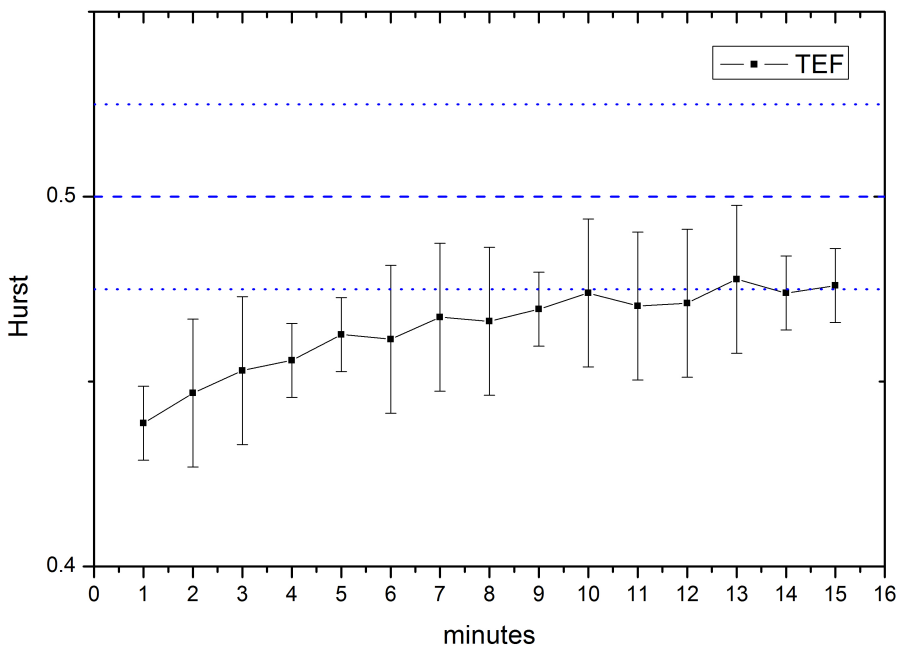
 ISBN: 978-84-614-6167-7

Fig. 2. Estimated Hurst minute by minute



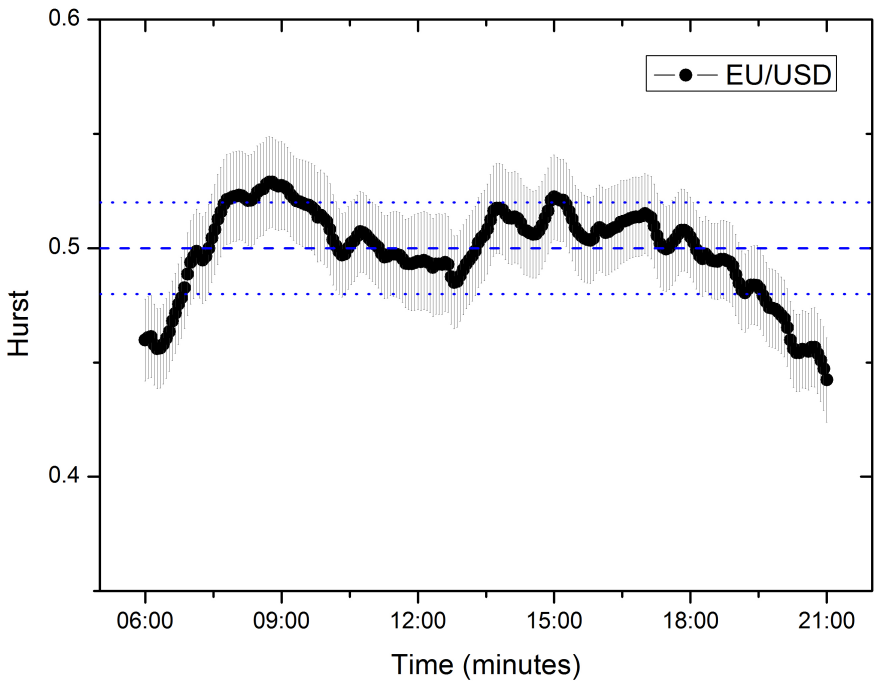
 ISBN: 978-84-614-6167-7

Fig. 3. Estimated Hurst minute by minute



 ISBN: 978-84-614-6167-7

Fig.4. Estimated mean Hurst for 46 intraday days



 ISBN: 978-84-614-6167-7

Proceedings of the 11th International Conference on Computational and Mathematical Methods in Science and Engineering, CMMSE 2011 26–30 June 2011.

New results on mathematical foundations of asymptotic complexity analysis of algorithms via complexity spaces

S. Romaguera[1], P. Tirado[1] and O. Valero[2]

[1] *Instituto Universitario de Matemática Pura y Aplicada, Universidad Politécnica de Valencia, Camino de Vera s/n, 46022 Valencia, Spain*

[2] *Departamento de Ciencias Matemáticas e Informática, Universidad de las Islas Baleares, Carretera de Valldemossa Km. 7.5, 07122 Palma de Mallorca, Spain*

emails: sromague@mat.upv.es, pedtipe@mat.upv.es, o.valero@uib.es

Abstract

In 1995, M.P. Schellekens introduced the theory of complexity (quasi-metric) spaces as a part of the development of a topological foundation for the asymptotic complexity analysis of programs and algorithms [Electron. Notes Theor. Comput. Sci. 1 (1995), 211-232]. The applicability of this theory to the asymptotic complexity analysis of Divide and Conquer algorithms was also illustrated by Schellekens in the same reference. In particular, he gave a new formal proof, based on the use of the Banach fixed point theorem, of the well-known fact that the asymptotic upper bound of the average running time of computing of Mergesort belongs to the asymptotic complexity class of $n \log_2 n$. Motivated by the utility of the quasi-metric approach for the asymptotic complexity analysis based on the use of fixed point techniques and complexity spaces, on one hand we extend Schellekens' method in order to yield asymptotic upper bounds for a class of algorithms whose running time of computing leads to recurrence equations different from the Divide and Conquer ones, and, on the other hand, we improve the original Schellekens method by introducing a new fixed point technique for providing lower asymptotic bounds for the running time of computing of the aforesaid algorithms. We illustrate and validate the developed method applying our results to provide the asymptotic complexity class (asymptotic upper and lower bounds), among others, of the celebrated recursive algorithm that solves the problem of Hanoi Towers.

Key words: quasi-metric, complexity space, fixed point, improver, worsener, complexity class.

MSC 2000: AMS codes: 47H10, 54E35, 54E50, 68Q25, 68Q30.

 ISBN: 978-84-614-6167-7

1 The fundamentals of asymptotic complexity analysis of algorithms via complexity spaces

Throughout this paper the letters $\mathbb{R}^+$ and $\mathbb{N}$ will denote the set of nonnegative real numbers and the set of positive integer numbers, respectively.

Our basic reference for complexity analysis of algorithms is [1].

In Computer Science the complexity analysis of an algorithm is based on determining mathematically the quantity of resources needed by the algorithm in order to solve the problem for which it has been designed. A typical resource, playing a central role in complexity analysis, is the running time of computing. Since there are often many algorithms to solve the same problem, one objective of the complexity analysis is to assess which of them is faster when large inputs are considered. To this end, it is required to compare their running time of computing. This is usually done by means of the asymptotic analysis in which the running time of an algorithm is denoted by a function $T : \mathbb{N} \to (0, \infty]$ in such a way that $T(n)$ represents the time taken by the algorithm to solve the problem under consideration when the input of the algorithm is of size n. Of course the running time of an algorithm does not only depend on the input size n, but it depends also on the particular input of the size n (and the distribution of the data). Thus the running time of an algorithm is different when the algorithm processes certain instances of input data of the same size n. As a consequence, in general it is necessary to distinguish three possible behaviors when the running time of an algorithm is discussed. These are the so-called best case, the worst case and the average case. The best case and the worst case for an input of size n are defined by the minimum and the maximum running time of computing over all inputs of the size n, respectively. The average case for an input of size n is defined by the expected value or average running time of computing over all inputs of the size n.

Given an algorithm, to determine exactly the function which describes its running time of computing is in general an arduous task. However, in most situations is more useful to know the running time of computing of an algorithm in an "approximate" way than in an exact one. For this reason the asymptotic complexity analysis of algorithms focus its interest in obtaining "approximate" running times of computing.

In order to recall pertinent notions from asymptotic complexity analysis, let us assume that $f : \mathbb{N} \to (0, \infty]$ denotes the running time of computing of a certain algorithm. In addition, consider that there exists a function $g : \mathbb{N} \to (0, \infty]$ such that there exist, simultaneously, $n_0 \in \mathbb{N}$ and $c > 0$ satisfying $f(n) \leq cg(n)$ for all $n \in \mathbb{N}$ with $n \geq n_0$ ($\leq$ and $\geq$ stand for the usual orders on $\mathbb{R}^+$). Then, the function g provides an asymptotic upper bound of the running time of the studied algorithm. Thus, if we do not know the exact expression of the function f, then the function g gives an "approximate" information of the running time of the algorithm in the sense that the algorithm takes a time to solve the problem bounded above by g. Following the standard notation, when g is an asymptotic upper bound of f we write $f \in \mathcal{O}(g)$.

Sometimes in the analysis of the complexity of an algorithm is useful to assess an asymptotic lower bound of the running time of computing. In this case the Ω-notation plays a central role. Thus the statement $f \in \Omega(g)$ means that there exist $n_0 \in \mathbb{N}$ and

 ISBN: 978-84-614-6167-7

$c > 0$ such that $cg(n) \leq f(n)$ for all $n \in \mathbb{N}$ with $n \geq n_0$. Of course, and similarly to the $\mathcal{O}$-notation case, when the time taken by the algorithm to solve the problem f is unknown, the function g yields an "approximate" information of the running time of the algorithm in the sense that the algorithm takes a time to solve the problem bounded below by g.

It is clear that the best situation, when the complexity of an algorithm is discussed, matches up with the case in which we can find a function $g : \mathbb{N} \to (0, \infty]$ in such a way that the running time f holds the condition $f \in \mathcal{O}(g) \cap \Omega(g)$, denoted by $f \in \Theta(g)$, because, in this case, we obtain a "tight "asymptotic bound of f and, thus, a total asymptotic information about the time taken by the algorithm to solve the problem under consideration. From now on, we will say that f belongs to the asymptotic complexity class of g whenever $f \in \Theta(g)$.

Hence, from an asymptotic complexity analysis viewpoint, to determine the running time of an algorithm consists of obtaining its asymptotic complexity class.

In 1995, M.P. Schellekens introduced a new mathematical framework, known as complexity spaces, as a part of the development of a topological foundation for the asymptotic complexity analysis of algorithms ([6]). This approach is based on the notion of quasi-metric space.

Following [4], a quasi-metric on a non-empty set X is a function $d : X \times X \to \mathbb{R}^+$ such that for all $x, y, z \in X$: (i) $d(x, y) = d(y, x) = 0 \Leftrightarrow x = y$; (ii) $d(x, y) \leq d(x, z) + d(z, y)$.

Of course a metric on a non-empty set X is a quasi-metric d on X satisfying, in addition, the following condition for all $x, y \in X$: (iii) $d(x, y) = d(y, x)$.

A quasi-metric space is a pair (X, d) such that X is a non-empty set and d is a quasi-metric on X.

Each quasi-metric d on X generates a T_0-topology $\mathcal{T}(d)$ on X which has as a base the family of open d-balls $\{B_d(x, \varepsilon) : x \in X, \varepsilon > 0\}$, where $B_d(x, \varepsilon) = \{y \in X : d(x, y) < \varepsilon\}$ for all $x \in X$ and $\varepsilon > 0$.

Given a quasi-metric d on X, the function d^s defined on $X \times X$ by $d^s(x, y) = \max(d(x, y), d(y, x))$ is a metric on X.

A quasi-metric space (X, d) is called bicomplete if the metric space (X, d^s) is complete.

The complexity (quasi-metric) space is the pair $(\mathcal{C}, d_{\mathcal{C}})$, where

$$\mathcal{C} = \{f : \mathbb{N} \to (0, \infty] : \sum_{n=1}^{\infty} 2^{-n} \frac{1}{f(n)} < \infty\}$$

and $d_{\mathcal{C}}$ is the bicomplete quasi-metric on $\mathcal{C}$ defined by

$$d_{\mathcal{C}}(f, g) = \sum_{n=1}^{\infty} 2^{-n} \max\left(\frac{1}{g(n)} - \frac{1}{f(n)}, 0\right).$$

(We adopt the convention that $\frac{1}{\infty} = 0$.)

 ISBN: 978-84-614-6167-7

According to [6], since every reasonable algorithm, from a computability viewpoint, must hold the "convergence condition" $\sum_{n=1}^{\infty} 2^{-n}\frac{1}{f(n)} < \infty$, it is possible to associate each algorithm with a function of $\mathcal{C}$ in such a way that such a function represents, as a function of the size of the input data, its running time of computing. Because of this, the elements of $\mathcal{C}$ are called complexity functions. Moreover, given two functions $f, g \in \mathcal{C}$, the numerical value $d_{\mathcal{C}}(f, g)$ (the complexity distance from f to g) can be interpreted as the relative progress made in lowering the complexity by replacing any program P with complexity function f by any program Q with complexity function g. Therefore, if $f \neq g$, the condition $d_{\mathcal{C}}(f, g) = 0$ can be read as the program P is at least as efficient as the program Q (indeed, note that $d_{\mathcal{C}}(f, g) = 0 \Leftrightarrow f(n) \leq g(n)$ for all $n \in \mathbb{N}$). In fact, the condition $d_{\mathcal{C}}(f, g) = 0$ implies that $f \in \mathcal{O}(g)$.

Notice that the asymmetry of the complexity distance $d_{\mathcal{C}}$ plays a central role in order to provide information about the increase of complexity whenever a program is replaced by another one. A metric will be able to yield information on the increase but it, however, will not reveal which program is more efficient.

The applicability of the theory of complexity spaces to the asymptotic complexity analysis of algorithms was illustrated by Schellekens in [6]. In particular, he gave, among other things, a new proof of the well-known fact that that the function $f \in \mathcal{C}$, given by $f(1) = c > 0$ and $f(n) = n \log_2 n$ for all $n \in \mathbb{N}$ with $n > 1$, is an asymptotic upper bound of the average running time of computing of Mergesort. To this end, he introduced a method, based on the below quasi-metric version of Banach's fixed point theorem, to analyze the running time of computing of the general class of Divide and Conquer algorithms (observe that Mergesort is a Divide and Conquer algorithm).

Theorem 1. *Let f be a mapping from a bicomplete quasi-metric space (X, d) into itself such that there exists $s \in [0, 1)$ satisfying*

$$d(f(x), f(y)) \leq sd(x, y), \tag{1}$$

for all $x, y \in X$. Then f has a unique fixed point.

Let us recall that a mapping f from a quasi-metric space (X, d) into itself holding inequality (1) is said to be contractive with contractive constant s.

Next we provide a general view of the aforenamed method with the aim of motivating our subsequent work.

A Divide and Conquer algorithm solves a problem of size n ($n \in \mathbb{N}$) splitting it into a subproblems of size $\frac{n}{b}$, for some constants a, b with $a, b \in \mathbb{N}$ and $a, b > 1$, and solving them separately by the same algorithm. After obtaining the solution of the subproblems, the algorithm combines all subproblem solutions to give a global solution to the original problem. The recursive structure of a Divide and Conquer algorithm leads to a recurrence equation for the running time of computing. In many cases the running time of a Divide and Conquer algorithm is the solution to a recurrence equation of the form

$$T(n) = \begin{cases} c & \text{if } n = 1 \\ aT(\frac{n}{b}) + h(n) & \text{if } n \in \mathbb{N}_b \end{cases}, \tag{2}$$

 ISBN: 978-84-614-6167-7

where $\mathbb{N}_b = \{b^k : k \in \mathbb{N}\}$, $c > 0$ denotes the complexity on the base case (i.e. the problem size is small enough and the solution takes constant time), $h(n)$ represents the time taken by the algorithm in order to divide the original problem into a subproblems and to combine all subproblems solutions into a unique one ($h \in \mathcal{C}$ with $h(n) < \infty$ for all $n \in \mathbb{N}$).

Notice that for Divide and Conquer algorithms, it is typically sufficient to obtain the complexity on inputs of size n with n ranges over the set $\mathbb{N}_b$ ([1]).

Mergesort is a typical and well-known example of a Divide and Conquer algorithm whose running time of computing satisfies the recurrence equation (2) (see [1] for a fuller description).

In order to compute the running time of computing of a Divide and Conquer algorithm satisfying the recurrence equation (2), it is necessary to show that such a recurrence equation has a unique solution and, later, to obtain the asymptotic complexity class of such a solution. The method introduced by Schellekens allows to show that the equation (2) has a unique solution, and provides an upper asymptotic complexity bound of the solution in the following way:

Denote by $\mathcal{C}_{b,c}$ the subset of $\mathcal{C}$ given by

$$\mathcal{C}_{b,c} = \{f \in \mathcal{C} : f(1) = c \text{ and } f(n) = \infty \text{ for all } n \in \mathbb{N}\backslash\mathbb{N}_b \text{ with } n > 1\}.$$

Since the quasi-metric space $(\mathcal{C}, d_{\mathcal{C}})$ is bicomplete ([5]) and the set $\mathcal{C}_{b,c}$ is closed in $(\mathcal{C}, d^s_{\mathcal{C}})$, we have that the quasi-metric space $(\mathcal{C}_{b,c}, d_{\mathcal{C}}|_{\mathcal{C}_{b,c}})$ is bicomplete.

Next we associate a functional $\Phi_T : \mathcal{C}_{b,c} \to \mathcal{C}_{b,c}$ with the recurrence equation (2) of a Divide and Conquer algorithm given as follows:

$$\Phi_T(f)(n) = \begin{cases} c & \text{if } n = 1 \\ \infty & \text{if } n \in \mathbb{N}\backslash\mathbb{N}_b \text{ and } n > 1 \\ af(\frac{n}{b}) + h(n) & \text{otherwise} \end{cases} . \tag{3}$$

Of course a complexity function in $\mathcal{C}_{b,c}$ is a solution to the recurrence equation (2) if and only if it is a fixed point of the functional Φ_T. Then, Schellekens proved ([6]) that

$$d_{\mathcal{C}}|_{\mathcal{C}_{b,c}}(\Phi_T(f), \Phi_T(g)) \leq \frac{1}{a} d_{\mathcal{C}}|_{\mathcal{C}_{b,c}}(f, g) \tag{4}$$

for all $f, g \in \mathcal{C}_{b,c}$. So, by Theorem 1, the functional $\Phi_T : \mathcal{C}_{b,c} \to \mathcal{C}_{b,c}$ has a unique fixed point and, thus, the recurrence equation (2) has a unique solution.

In order to obtain the upper asymptotic complexity bound of the solution to the recurrence equation (2), Schellekens introduced a special class of functionals known as improvers.

Let $C \subseteq \mathcal{C}$. A functional $\Phi : C \to C$ is called an improver with respect to a function $f \in C$ provided that $\Phi^n(f) \leq \Phi^{n-1}(f)$ for all $n \in \mathbb{N}$. Of course $\Phi^0(f) = f$. Observe that an improver is a functional which corresponds to a transformation on programs in such a way that the iterative applications of the transformation yield, from a complexity point of view, an improved program at each step of the iteration. Note that under the assumption that the functional Φ is monotone, to show that Φ is an improver with respect to $f \in C$ is equivalent to verify that $\Phi(f) \leq f$.

Taking into account the exposed facts, Schellekens stated the following result ([6]).

 ISBN: 978-84-614-6167-7

Theorem 2. *A Divide and Conquer recurrence of the form (2) has a unique solution f_T in $\mathcal{C}_{b,c}$. Moreover, if the functional Φ_T associated with (2) is an improver with respect to some function $g \in \mathcal{C}_{b,c}$, then the solution to the recurrence equation satisfies that $f_T \in \mathcal{O}(g)$.*

He also obtained an asymptotic upper bound of the running time of computing of Mergesort in order to illustrate the usefulness of Theorem 2. In the particular case of Mergesort (average case), the running time of computing satisfies the following particular case of recurrence equation (2):

$$T(n) = \begin{cases} c & \text{if } n = 1 \\ 2T(\frac{n}{2}) + \frac{n}{2} & \text{if } n \in \mathbb{N}_2 \end{cases} . \tag{5}$$

It is clear that Theorem 2 shows that the recurrence equation (5) has a unique solution f_T^M in $\mathcal{C}_{2,c}$. In addition, Schellekens proved that the functional Φ_T induced by the recurrence equation (5) is an improver with respect to a complexity function $g_k \in \mathcal{C}_{2,c}$ (with $k > 0$, $g_k(1) = c$ and $g_k(n) = kn\log_2(n)$ for all $n \in \mathbb{N}_2$) if and only if $k \geq \frac{1}{2}$. Therefore, by Theorem 2, we conclude that $f_T^M \in \mathcal{O}(g_{\frac{1}{2}})$, i.e. Theorem 2 provides a formal proof, based on fixed point techniques, of the well-known fact that the running time of computing (average case) f_T^M of Mergesort is in $\mathcal{O}(n\log_2 n)$, i.e. that the complexity function $g_{\frac{1}{2}}$, or equivalently $\mathcal{O}(n\log_2 n)$, gives an asymptotic upper bound of f_T^M. Furthermore, in [6] it is pointed out that an asymptotic lower bound of the running time of Mergesort (average case) belongs to $\Omega(n\log_2 n)$ (following standard arguments which are not based on the use of fixed point techniques). So Mergesort running time (average case) belongs to the complexity class $\Theta(n\log_2 n)$.

Of course, Schellekens' method, without meaning to compete with the standard and classical techniques to analyze the complexity of algorithms, has the advantage of allowing to apply similar ideas to those presented by D.S. Scott ([7], [8]) in modelling the meaning of recursive denotational specifications of algorithms via fixed point techniques in such a way that the notion of "iterative approximations", typical of the topological Scott framework, is captured trough the concept of improver functional.

2 Extending the applicability of complexity spaces: new algorithms and recurrence equations

In spite of it seems natural that the complexity analysis of Divide and Conquer algorithms always leads up to Divide and Conquer recurrence equations of type (2), this is not the case. Sometimes this kind of recursive algorithms yields recurrence equations that differ from (2). A well-known example of this sort of situations is provided by Quicksort (worst case) ([2]). Although the recurrence equations associated to the running time of computing of Mergesort and Quicksort do not belong to the same class, it is clear that the main relationship between both algorithms is given by the fact that them belong to the Divide and Conquer algorithms class and, thus, they are recursive algorithms.

 ISBN: 978-84-614-6167-7

Of course, the class of recursive algorithms is wider than the Divide and Conquer. An illustrative example of recursive algorithm, which does not belong to the Divide and Conquer family, is provided by Hanoi. Hanoi solves the Towers of Hanoi puzzle (see [2] and [3]).

The fact that the class of recursive algorithms is wider than the Divide and Conquer inspires to wonder two questions. On one hand, whether one can obtain a family of recurrence equations in such a way that the complexity analysis of those algorithms whose running time of computing is a solution either to recurrence equations associated with Quicksort and Hanoi (see Subsection 3.3) or to a Divide and Conquer one can be carried out from it. On the other hand, whether such a complexity analysis can be done via an extension of the fixed point technique of Schellekens.

Clearly, the recurrence equations that yield the running time of computing of the above aforesaid algorithms can be considered as particular cases of the following general one:

$$T(n) = \begin{cases} c & \text{if } n = 1 \\ aT(n-1) + h(n) & \text{if } n \geq 2 \end{cases}, \tag{6}$$

where $c > 0$, $a \geq 1$ and $h \in \mathcal{C}$ such that $h(n) < \infty$ for all $n \in \mathbb{N}$.

In particular, an easy modification of T in the recurrence equation (2) allows us to reduce this equation to one of type (6).

As well as the exposed advantage, the relevance of the family of recurrence equations of type (6) is intensified by the fact that the running time of certain non-recursive algorithms also matches up with the solution to a recurrence equation that can be retrieved as a particular case of the general recurrence equation (6). A good example is provided by Largetwo. This algorithm finds the two largest entries in one-dimensional array of size $n \in \mathbb{N}$ with $n > 1$ (for a deeper discussion we refer the reader to [2]).

In what follows our purpose is to demonstrate that the Schellekens fixed point technique can be used satisfactorily to discuss the complexity of those algorithms whose running time of computing yields with a recurrence equation of type (6). In particular, the aforesaid recurrence equation has a unique solution and, in addition, we obtain the complexity class (asymptotic upper and lower bounds) of such a solution. Similarly to Schellekens' approach, our technique to obtain the asymptotic upper bound is based on the use of the improver functional induced by the recurrence equation. Nevertheless, we introduce a new kind of functionals, that we have called "worsener" functionals, with the aim of obtaining the asymptotic lower bound of the solution to the recurrence equation. In order to provide the complexity class of an algorithm whose running time satisfies a recurrence equation of type (6) we prove that it is enough to search among all complexity functions for which the functional associated to the recurrence equation is simultaneously an improver and a worsener. Finally, in order, on one hand, to validate our new results and, on the other hand, to show the potential applicability of the developed theory to complexity analysis in Computer Science, we shall discuss the running time of Quicksort (worst case), Hanoi and Largetwo (average case), respectively.

 ISBN: 978-84-614-6167-7

3 The new fixed point technique in complexity analysis

Is this section we provide the new fixed point technique to show the existence and uniqueness of the solution to the recurrence equations of type (6) and the announced mathematical method to obtain the complexity class of those algorithms whose running time satisfies the recurrence equation under study.

3.1 The existence and uniqueness of solution

Consider the subset $\mathcal{C}_c$ of $\mathcal{C}$ given by

$$\mathcal{C}_c = \{f \in \mathcal{C} : f(1) = c\}.$$

Define the functional $\Psi_T : \mathcal{C}_c \to \mathcal{C}_c$ by

$$\Psi_T(f)(n) = \begin{cases} c & \text{if } n = 1 \\ af(n-1) + h(n) & \text{if } n \geq 2 \end{cases} \tag{7}$$

for all $f \in \mathcal{C}_c$. It is clear that a complexity function in $\mathcal{C}_c$ is a solution to the recurrence equation (6) if and only if it is a fixed point of the functional Ψ_T.

The next result supplies the bicompleteness of the pair $(\mathcal{C}_c, d_{\mathcal{C}}|_{\mathcal{C}_c})$.

Proposition 3. *The subset $\mathcal{C}_c$ is closed in $(\mathcal{C}, d^s_{\mathcal{C}})$.*

Since the metric space $(\mathcal{C}, d^s_{\mathcal{C}})$ is complete and, by the preceding proposition, the subset $\mathcal{C}_c$ is closed in $(\mathcal{C}, d^s_{\mathcal{C}})$ we immediately obtain the following consequence.

Corollary 4. *The quasi-metric space $(\mathcal{C}_c, d_{\mathcal{C}}|_{\mathcal{C}_c})$ is bicomplete.*

Theorem 5. *The functional Ψ_T is a contraction from $(\mathcal{C}_c, d_{\mathcal{C}}|_{\mathcal{C}_c})$ into itself with contractive constant $\frac{1}{2a}$.*

From the above theorem we can immediately gather that a recurrence equation of the form (6) has a unique solution f_T in $\mathcal{C}_c$ which matches up with the running time of computing of the algorithm under study considered in each case.

3.2 The complexity class of the solution

Next we provide a method (Theorem 7 below) to describe the complexity of those algorithms whose running time of computing satisfies a recurrence equation of type (6). To this end we need the below auxiliary result.

Lemma 6. *Let C be a subset of $\mathcal{C}$ such that the quasi-metric space $(C, d_{\mathcal{C}}|_C)$ is bicomplete and let $\Psi : C \to C$ be a contraction with fixed point $f \in C$ and contractive constant s. Then the following statements hold:*

1) If there exists $g \in C$ with $d_{\mathcal{C}}|_C(\Psi(g), g) = 0$, then $d_{\mathcal{C}}|_C(f, g) = 0$.

2) If there exists $g \in C$ with $d_{\mathcal{C}}|_C(g, \Psi(g)) = 0$, then $d_{\mathcal{C}}|_C(g, f) = 0$.

 ISBN: 978-84-614-6167-7

Note that if a complexity function f represents the running time of computing of an algorithm under study, the fact that there exists a complexity function g satisfying the condition $d_{\mathcal{C}}|_C(\Psi(g), g) = 0$ ($d_{\mathcal{C}}|_C(g, \Psi(g)) = 0$) in the preceding lemma provides an asymptotic upper (lower) bound of the aforesaid running time, since $d_{\mathcal{C}}|_C(f, g) = 0$ ($d_{\mathcal{C}}|_C(g, f) = 0$) implies that $f \in \mathcal{O}(g)$ ($f \in \Omega(g)$).

In the light of Lemma 6 we observe that in order to get an asymptotic upper bound of the running time of computing of an algorithm whose running time matches up with the fixed point of a contraction $\Psi : C \to C$ ($C \subseteq \mathcal{C}$), it is enough to check if such a mapping satisfies the condition $\Psi(g) \leq g$ for any complexity function even if Ψ is not monotone, i.e. it is unnecessary to check if Ψ is an improver with respect to a complexity function. Motivated by this reason, in the remainder of this paper, given $C \subseteq \mathcal{C}$ and a contraction $\Psi : C \to C$ we will say that Ψ is contractive improver (cont-improver for short) with respect to a complexity function $g \in C$ whenever $\Psi(g) \leq g$. Notice that an improver in our sense is an improver in the original sense of Schellekens. Moreover, the computational meaning of improver functionals remains valid for the cont-improver ones. Indeed, if Ψ is a cont-improver with respect to the complexity function g then $\Psi^n(g) \leq \Psi^{n-1}(g)$ for all $n \in \mathbb{N}$, since

$$d_{\mathcal{C}}|_C(\Psi^n(g), \Psi^{n-1}(g)) \leq \frac{1}{(2a)^{n-1}} d_{\mathcal{C}}|_C(\Psi(g), g) = 0$$

for all $n \in \mathbb{N}$.

Inspired by statement 2) in Lemma 6 we introduce a new kind of functionals that we call worseners. Let $C \subseteq \mathcal{C}$, a contraction $\Psi : C \to C$ is said to be a worsener with respect to a function $f \in C$ provided that $f \leq \Psi(f)$.

Observe that if Ψ is a worsener with respect to $f \in C$, then

$$d_{\mathcal{C}}|_C(\Psi^{n-1}(f), \Psi^n(f) \leq \frac{1}{(2a)^{n-1}} d_{\mathcal{C}}|_C(f, \Psi(f)) = 0$$

for all $n \in \mathbb{N}$. It follows that the computational meaning of a worsener functional is dual to the meaning of a cont-improver functional. In fact, a worsener is a functional which corresponds to a transformation on programs in such a way that the iterative applications of the transformation yield a worsened, from a complexity point of view, program at each step of the iteration.

In the next result we obtain the announced method to provide the complexity class of an algorithm whose running time of computing satisfies a recurrence equation of type (6).

Theorem 7. *Let $f_T \in \mathcal{C}_c$ be the (unique) solution to a recurrence equation of type (6). Then the following facts hold:*

1) *If the functional Ψ_T associated to (6), and given by (7), is a cont-improver with respect to some function $g \in \mathcal{C}_c$, then $f_T \in \mathcal{O}(g)$.*

2) *If the functional Ψ_T associated to (6), and given by (7), is a worsener with respect to some function $g \in \mathcal{C}_c$, then $f_T \in \Omega(g)$.*

 ISBN: 978-84-614-6167-7

Note that the solution to a recurrence equation of type (6) satisfies that $f_T \in \mathcal{O}(g) \cap \Omega(h)$ whenever Ψ_T is a cont-improver and a worsener with respect to $g \in \mathcal{C}_c$ and $h \in \mathcal{C}_c$, respectively. Consequently, Theorem 7 yields the complexity class of algorithms whose running time of computing satisfies a recurrence equation of type (6) when there exist $l \in \mathcal{C}_c$, $r, t > 0$ and $n_0 \in \mathbb{N}$ such that $g(n) = rl(n)$ and $h = tl(n)$ for all $n > n_0$ and, besides, Ψ_T is a cont-improver and a worsener with respect to g and h respectively, because, in such a case, $f_T \in \Theta(l)$.

3.3 Analyzing the running time computing of some algorithms

We end the paper showing that the developed method is useful to analyze the asymptotic complexity of Divide and Conquer algorithms, recursive algorithms and even non-recursive algorithms. To this aim we validate our results retrieving as an immediate consequence of Theorem 7 the well-known asymptotic complexity class of Quicksort (worst case), Hanoi and Largetwo (average case).

Quicksort: The running time of computing of Quicksort (worst case) is the solution to the recurrence equation

$$T(n) = \begin{cases} c & \text{if } n = 1 \\ T(n-1) + jn & \text{if } n \geq 2 \end{cases}, \tag{8}$$

where $c, j > 0$. It is clear that the preceding recurrence equation can be retrieved from (6) as a particular case when we fix $a = 1$ and $h(n) = jn$ for all $n \in \mathbb{N}$. Then, taking

$$\Psi_T(f)(n) = \begin{cases} c & \text{if } n = 1 \\ f(n-1) + jn & \text{if } n \geq 2 \end{cases} \tag{9}$$

for all $f \in \mathcal{C}_c$, Theorem 5 guarantees the existence and uniqueness of the solution (in $\mathcal{C}_c$), which matches up with the running time of computing of Quicksort (worst case), to the above recurrence equation. Denote such a solution by f_T^Q. It is not hard to see that Ψ_T is a cont-improver with respect to the complexity function $h_r \in \mathcal{C}_c$ (i.e. $\Psi_T(h_r) \leq h_r$) if and only if $r \geq \max\{\frac{3j}{5}, \frac{c}{4} + \frac{j}{2}\}$, where the complexity function h_r is given by $h_r(1) = c$ and $h_r(n) = rn^2$ for $n \geq 2$.

Hence we obtain, by statement 1) in Theorem 7, that the running time of Quicksort (worst case) holds $f_T^Q \in \mathcal{O}(h_{\max\{\frac{3j}{5}, \frac{c}{4}+\frac{j}{2}\}})$.

In addition, it is not hard to see that Ψ_T is a worsener with respect to the complexity function h_s (i.e. $h_s \leq \Psi_T(h_s)$) if and only if $s \leq \min\{\frac{j}{2}, \frac{c}{4} + \frac{j}{2}\}$, whence we deduce, by statement 2) in Theorem 7, that $f_T^Q \in \Omega(h_{\min\{d, \frac{2c+d}{3}\}})$.

Therefore we obtain that $f_T^H \in \mathcal{O}(h_{\max\{\frac{3j}{5}, \frac{c}{4}+\frac{j}{2}\}}) \cap \Omega(h_{\min\{\frac{j}{2}, \frac{c}{4}+\frac{j}{2}\}})$. Hence $f_T^Q \in \Theta(n^2)$, which is in accordance with the Quicksort (worst case) complexity class that can be found in the computational literature ([1], [2]).

Hanoi: The running time of computing of Hanoi is the solution, under the uniform cost criterion assumption, to the recurrence equation

 ISBN: 978-84-614-6167-7

$$T(n) = \begin{cases} c & \text{if } n = 1 \\ 2T(n-1) + d & \text{if } n \geq 2 \end{cases}, \tag{10}$$

where $c, d > 0$. It is clear that the preceding recurrence equation can be retrieved from (6) as a particular case when we fix $a = 2$ and $h(n) = d$ for all $n \in \mathbb{N}$. Then, taking

$$\Psi_T(f)(n) = \begin{cases} c & \text{if } n = 1 \\ 2f(n-1) + d & \text{if } n \geq 2 \end{cases} \tag{11}$$

for all $f \in \mathcal{C}_c$, Theorem 5 guarantees the existence and uniqueness of the solution (in $\mathcal{C}_c$), which matches up with the running time of computing of Hanoi, to the above recurrence equation. Next we denote such a solution by f_T^H. It is not hard to see that Ψ_T is a cont-improver with respect to the complexity function $h_r \in \mathcal{C}_c$ (i.e. $\Psi_T(h_r) \leq h_r$) if and only if $r \geq \max\{d, \frac{2c+d}{3}\}$, where the complexity function h_r is given by $h_r(1) = c$ and $h_r(n) = r(2^n - 1)$ for $n \geq 2$.

So, by statement 1) in Theorem 7, we deduce that the running time of Hanoi satisfies $f_T^H \in \mathcal{O}(h_{\max\{d, \frac{2c+d}{3}\}})$.

Furthermore, it is easily seen that Ψ_T is a worsener with respect to the complexity function h_s (i.e. $h_s \leq \Psi_T(h_s)$) if and only if $s \leq \min\{d, \frac{2c+d}{3}\}$, whence we deduce, by statement 2) in Theorem 7, that $f_T^H \in \Omega(h_{\min\{d, \frac{2c+d}{3}\}})$.

Therefore we deduce hat $f_T^H \in \mathcal{O}(h_{\max\{d, \frac{2c+d}{3}\}}) \cap \Omega(h_{\min\{d, \frac{2c+d}{3}\}})$. Thus $f_T^H \in \Theta(2^n)$, which is in accordance with the Hanoi complexity class that can be found in the computational literature ([2]).

Largetwo: The running time of computing of Largetwo (average case) is the solution to the recurrence equation

$$T(n) = \begin{cases} c & \text{if } n = 1 \\ T(n-1) + 2 - \frac{1}{n} & \text{if } n \geq 2 \end{cases}, \tag{12}$$

where $c > 0$. It is clear that the preceding recurrence equation can be retrieved from (6) as a particular case when we fix $a = 1$ and $h(n) = 2 - \frac{1}{n}$ for all $n \in \mathbb{N}$. Then, taking

$$\Psi_T(f)(n) = \begin{cases} c & \text{if } n = 1 \\ f(n-1) + 2 - \frac{1}{n} & \text{if } n \geq 2 \end{cases} \tag{13}$$

for all $f \in \mathcal{C}_c$, Theorem 5 guarantees the existence and uniqueness of the solution (in $\mathcal{C}_c$), which matches up with the running time of computing of Largetwo (average case), to the above recurrence equation. Let us denote such a solution by f_T^L. It is not hard to see that Ψ_T is a cont-improver with respect to the complexity function $h_r \in \mathcal{C}_c$ (i.e. $\Psi_T(h_r) \leq h_r$) if and only if $r \geq \max\{\frac{5}{6+3\log_2(\frac{2}{3})}, \frac{2c+3}{2+2d}\}$, where the complexity function h_r is given by $h_r(1) = c$ and $h_r(n) = r\left(2\left(n-1\right) - \log_2 n + d\right)$ for $n \geq 2$.

So we deduce, by statement 1) in Theorem 7, that the running time of Largetwo (average case) satisfies $f_T^L \in \mathcal{O}(h_{\max\{\frac{5}{6+3\log_2(\frac{2}{3})}, \frac{2c+3}{2+2d}\}})$.

Moreover, a straightforward computation shows that Ψ_T is a worsener with respect to the complexity function h_s (i.e. $h_s \leq \Psi_T(h_s)$) if and only if $s \leq \min\{1, \frac{2c+3}{2+2d}\}$, whence we deduce, by statement 2) in Theorem 7, that $f_T^L \in \Omega(h_{\min\{1, \frac{2c+3}{2+2d}\}})$.

 ISBN: 978-84-614-6167-7

Therefore we obtain hat $f_T^L \in \mathcal{O}(h_{\max\{\frac{5}{6+3\log_2(\frac{2}{3})}, \frac{2c+3}{2+2d}\}}) \cap \Omega(h_{\min\{1, \frac{2c+3}{2+2d}\}})$. Thus $f_T^L \in \Theta(n \log_2 n)$, which is in accordance with the Largetwo (average case) complexity class that can be found in the computational literature ([2]).

Acknowledgements

The authors thank the support from the Spanish Ministry of Science and Innovation, grant MTM2009-12872-C02-01.

References

[1] G. Brassard, P. Bratley, *Algorithms: Theory and Practice*, Prentice Hall, New Jersey, 1988.

[2] P. Cull, M. Flahive, R. Robson, *Difference equations: From rabbits to chaos*, Springer, New York, 2005.

[3] E.F. Ecklund Jr., P. Cull, *Towers of Hanoi and analysis of algorithms*, The American Mathematical Monthly **92** (1985), 407-420.

[4] H.P.A. Künzi, *Nonsymmetric distances and their associated topologies: About the origins of basic ideas in the area of asymmetric topology*, in: Handbook of the History of General Topology, ed. by C.E. Aull and R. Lowen, vol. 3, Kluwer, Dordrecht, 2001.

[5] S. Romaguera, M.P. Schellekens, *Quasi-metric properties of complexity spaces*, Topology Appl. **98** (1999), 311-322.

[6] M.P. Schellekens, *The Smyth completion: a common foundation for denonational semantics and complexity analysis*, Electron. Notes Theor. Comput. Sci. **1** (1995), 211-232.

[7] D.S. Scott, *Outline of a mathematical theory of computation*, in: Proc. 4th Annual Princeton Conference on Information Sciences and Systems, pp. 169-176 (1970).

[8] D.S. Scott, *Domains for denotational semantics*, LNCS **140** (1982), 577-613.

 ISBN: 978-84-614-6167-7

Proceedings of the 11th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2011
26–30 June 2011.

Van der Waals interactions in density functional theory: an efficient implementation for large systems

G. Román-Pérez[1], Félix Yndurain[1] and José M. Soler[1]

[1] *Departamento de Física de la Materia Condensada, Universidad Autónoma de Madrid, 28049 Spain*

emails: guillermo.roman@uam.es, felix.yndurain@uam.es, jose.soler@uam.es

Abstract

The recently developed functional, which includes van der Waals interactions in a full ab initio treatment of non-local correlation energy, has been shown to be very computationally demanding. This limits its use for small and medium systems. An implementation based on a suitable factorization and on the use of fast Fourier transforms, generalizing the use of the new functional to large systems. We present accurate results in small systems in order to validate the implementation, comparing the computational costs with the direct evaluation of the functional. We use this implementation to study properties in clathrate hydrates, where van der Waals interactions are crucial for stability.

Key words: DFT, van der Waals, clathrate hydrate

1 Introduction

In the last decades, one of the main aims in the density functional theory (DFT) has been to include the long range correlation contributions of the van der Waals interactions. Even though these interactions are much smaller than those responsible for the chemical bonds, they are important in many systems and processes, such as soft matter, molecular adsorption in surfaces and solids, biological reactions, etc [18]. Local and semi-local (DFT) approximations [21, 33] cannot describe the asymptotic behavior of the non-local dispersion correlation. These approximations have been found useful for finding binding distances in some cases, but its dependence on the specific parametrization of the different functionals compromises the method. The more extended solution to take into account these interactions in an ab initio framework was to add an interatomic potential, ensuring the correct

 ISBN: 978-84-614-6167-7

asymptotic behavior [9, 10]. These corrective terms depend on parameters which are fixed from experimental or high-quality quantum chemistry data. The results are good in many systems, but two objections can be given about its use: they are not ab initio and they are system dependent. The second one implies a lack of ful transferability, which is crucial in some cases.

Dion et al. [6] developed the first functional for general purposes with full ab initio treatment of the non local correlation energy. Other functionals were proposed for layered system in the 1990's [17, 1, 8, 12, 7, 26]. The first applications were post-GGA perturbations, using the density obtained in the general gradients approximation (GGA) as input to evaluate the new functional. It was not able to do molecular dynamics simulations, because Hellmann-Feymann forces were not calculated. Both limitations were resolved by the self-consistent version [28]. Other parametrizations [31, 30] and a new version of the vdW-DF functional has been recently reported [15]. The authors divide the correlation contribution in two terms and the general expression for exchange-correlation energy is,

$$E_{XC}[n(\mathbf{r})] = E_X[n(\mathbf{r})] + E_C[n(\mathbf{r})] = E_X^{GGA}[n(\mathbf{r})] + E_C^0[n(\mathbf{r})] + E_C^{nl}[n(\mathbf{r})] \tag{1}$$

where the brackets mean a functional dependence on the electron density, $n(\mathbf{r})$. The exchange energy, E_X, is described in the semi-local GGA. The local part of the correlation energy, E_C^0, is described in the local density approximation (LDA), and the non-local part (nl) E_C^{nl} is given by

$$E_C^{nl}[n(\mathbf{r})] = \frac{1}{2}\iint d^3\mathbf{r}_1\ d^3\mathbf{r}_2\ n(\mathbf{r}_1)n(\mathbf{r}_2)\phi(q_1, q_2, r_{12}) \tag{2}$$

where q_1 and q_2 are values of an universal function $q_0[n(\mathbf{r}), |\nabla n(\mathbf{r})|]$ evaluated in $\mathbf{r}_1$ and $\mathbf{r}_2$. $r_{12} = |\mathbf{r}_1 - \mathbf{r}_2|$.

Although the vdW-DF functional represents a big advance and it has been applied to many small systems [5, 23, 11, 4, 29], its use requires a very large computational resources. The direct evaluation of the double spatial integral of the Eq. 2 scales as $O(N^2)$, with the number N of integration points. This means a prohibitive demand for systems where van der Waals forces are important, that are typically large. The implementation explained here reduces the scaling to $O(NlogN)$ operations and it allows using the vdW-DF functional in large systems.

2 Efficient Implementation

The kernel ϕ has a universal and precise form [6, 13]. It obeys that E_C^{nl} is zero for any system with constant density, where the correlation energy is fully considered in E_C^0, and it has the correct dependence for long separations r^{-6}.

 ISBN: 978-84-614-6167-7

The integrand of Eq. 2 would be a convolution if q_1 and q_2 were fixed values and independent of $\mathbf{r}$. This would be very useful because Fourier methods can be used to evaluate it. We can write the kernel ϕ as an expansion,

$$\phi(q_1, q_2, r_{12}) \simeq \sum_{\alpha\beta} \phi(q_\alpha, q_\beta, r_{12}) p_\alpha(q_1) p_\beta(q_2) \tag{3}$$

where q_α and q_β are fixed values, chosen to ensure a good interpolation of ϕ. $p_\alpha(q_1)$ and $p_\beta(q_2)$ are the interpolation functions, which depend on the interpolation scheme and on the q_α and q_β values, respectively. In other words, Eq. 3 is the linear interpolation of the three-dimensional function ϕ, in its two first variables,q_α and q_β, being the interpolation coefficients still functions of the third variable, r_{12}. The shape of the function (shown in Fig. 1 of [6]) drives to a logarithmic mesh of points, q_α. This interpolation of ϕ is done up to the point where the original function, $q_0(n, |\nabla n|)$, reaches a value q_c. Above this cutoff, a saturated value is defined as

$$\begin{aligned} q_0^{sat}(n, \nabla n) &= h[q_0(n, |\nabla n|), q_c] \\ h(x, x_c) &= x_c \left[1 - exp\left(- \sum_{m=1}^{m_c} \frac{(x/x_c)^m}{m} \right) \right] \\ h(x, x_c) &= \begin{cases} \simeq x & \text{when} \quad x < x_c \\ \to x_c & \text{when} \quad x \to \infty \end{cases} \end{aligned} \tag{4}$$

where $h(x, x_c)$ is a soft function. m_c and q_c are determined to balance accuracy and computational time.

Although we are using ϕ as a three variable function, because is useful for interpolation, indeed it can be expressed as a function of two variables, $d_1 = q_1 r_{12}$ and $d_2 = q_2 r_{12}$ [6]. ϕ has a logarithmic divergence when $d_1, d_2 \to 0$. Instead of using this, we interpolate a modified ϕ [25]. This leads to a change in E_C^{nl}, which is corrected using a local density approximation, that can be calculated as in reference [2].

After substitution of the interpolated ϕ in Eq. 2, we get,

$$\begin{aligned} E_c^{nl} &= \frac{1}{2} \sum_{\alpha\beta} \iint d^3\mathbf{r}_1 \; d^3\mathbf{r}_2 \; \theta_\alpha(\mathbf{r}_1) \; \theta_\beta(\mathbf{r}_2) \; \phi_{\alpha\beta}(r_{12}) \\ &= \frac{1}{2} \sum_{\alpha\beta} \int d^3\mathbf{k} \; \theta_\alpha(\mathbf{k}) \; \theta_\beta(\mathbf{k}) \; \phi_{\alpha\beta}(k) \end{aligned} \tag{5}$$

where

$$\theta_\alpha(\mathbf{r}_1) = n(\mathbf{r}_1) p_\alpha \left[q_0^{sat}(n(\mathbf{r}_1), |\nabla n(\mathbf{r}_1)|) \right] \quad \text{and} \quad \theta_\beta(\mathbf{r}_2) = n(\mathbf{r}_2) p_\beta \left[q_0^{sat}(n(\mathbf{r}_2), |\nabla n(\mathbf{r}_2)|) \right]$$

 ISBN: 978-84-614-6167-7

$\theta_\alpha(\mathbf{k})$ is the Fourier transform of $\theta_\alpha(\mathbf{r})$, and $\phi_{\alpha\beta}(k)$ is the Fourier transform of $\phi_{\alpha\beta}(r) \equiv \phi(q_\alpha, q_\beta, r)$, which can be calculated in a fine radial mesh of k points for convenient interpolation.

In order to evaluate atomic forces, the full energy functional has to be minimized self-consistently. The method to handle the gradient dependence is based again in the technique reported in Ref. [2], and details can be found in Ref. [25]. This non-local potential, v_i^{nl}, is a functional of the density in every point $\mathbf{r}_i$ of the mesh, which is added to the semi-local terms in v_i^{XC} and to the rest of the effective potential.

Table 1 shows the algorithm. The main computational effort is done to calculate N_α transforms and N_α inverse transforms. The scaling of the fast Fourier transforms (FFT) and of the method is $O(N \log N)$. The method uses any electronic density defined in a uniform real-space grid $\mathbf{r}_i$ as input and returns exchange-correlation energy and potential in the same grid. The method does not depend on the basis set, and can be implemented in any electronic structure code which uses basis functions.

do, for each point i of the real-space mesh
    find n_i *and* ∇n_i
    find $q_i = q(n_i, \nabla n_i)$
    find $\theta_{\alpha i} = n_i p_\alpha(q_i)\ \forall\alpha$
end do
calculate Fourier Transform $\theta_{\alpha i} \rightarrow \theta_{\alpha k}\ \forall\alpha$
do, for each point k of the reciprocal-space mesh
    find $u_{\alpha k} = \sum_\beta \phi_{\alpha\beta}(k)\theta_{\beta k}\ \forall\alpha$
end do
calculate Inverse Fourier Transform $u_{\alpha k} \rightarrow u_{\alpha i}\ \forall\alpha$
do, for each point i of the real-space mesh
    find n_i, ∇n_i *and* q_i
    find $\theta_{\alpha i}$, $\partial\theta_{\alpha i}/\partial n_i$ *and* $\partial\theta_{\alpha i}/\partial\nabla n_i\ \forall\alpha$
    find v_i
end do

Table 1: Algorithmic for one self-consistent step of the implementation.

3 Results

This method was first tested in some of the systems studied originally with the non-local functional [6, 23, 29]. Fig. 1 a shows the interaction energy of a dimer of argon. We find a very good agreement between the direct evaluation of Eq. 2 [6] (dashed line) and

 ISBN: 978-84-614-6167-7

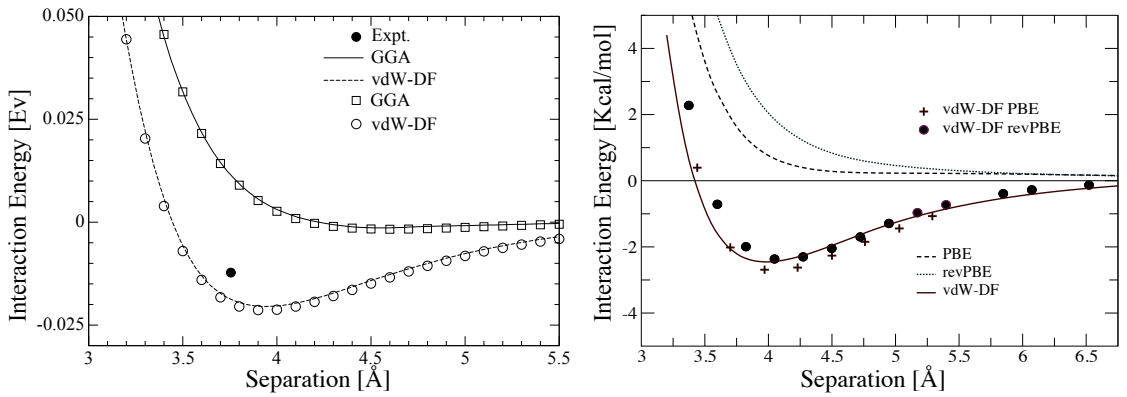


Figure 1: Interaction energy as function of separation of argon dimer (a) and benzene dimer in sandwich geometry (b). a) The solid and dashed lines represent the original calculation [6] using GGA and direct evaluation of de Eq. 2, respectively. Squares and circles show the results using our implementation. The black circle is the experimental equilibrium distance. b) Symbols represent direct evaluation of $E_{XC}^{nl}$from several authors [23, 14] and lines our calculations.

implementation explained above (circles). It is well known that the vdW-DF functional overestimates the experimental bond distances [29, 6, 23] (black circle [19]). Fig. 1 b shows the interaction energy of a dimer of benzene in a sandwich geometry, where the positions of atoms of different molecules differ only in the z coordinate. The symbols represent calculations by direct evaluation of E_{XC}^{nl} of different authors [14, 23] and the lines have been obtained with our implementation. Again, a good agreement is found. The small discrepancies may be due to the different basis-sets. The results from Ref. [23] (purple circles) were obtained with the non-selfconsistent version. We think that this is the reason why it shows smaller binding energies, especially when the dimer atoms are close. In both figures we include results obtained with the GGA, to show the qualitative and quantitative differences in the interaction energy.

This method was used in a study of double wall carbon nanotubes (DWNT) of different geometries [25] with a number of atoms from 60 to 168. Table 2 shows the computational times of calculating the exchange-correlation energy of two systems: a dimer of argon and a DWNT (8,2)(16,4) with 168 atoms. The relative differences between using GGA or vdW-DF are large for the dimer of argon but they are small for large systems.

Graphite is the paradigmatic example of a van der Waals solid. It is a layered material in which carbon atoms are arranged in a covalently bonded hexagonal lattice. The layers are stacked in different geometries and bonded by weak van der Waals forces. Its unit cell, in the AB stacking, has four carbon atoms. Table 3 shows the lattice parameter of a

 ISBN: 978-84-614-6167-7

System	Atoms	CPU time GGA-XC	CPU time vdW-XC	vdW/GGA
Ar_2	2	0.75 s (44%)	7.5 s (89%)	40%
DWNT	168	11.9 s (0.6%)	109 s (5.2%)	4%

Table 2: Comparison of the computational cost of two systems: a dimer of argon and a double wall carbon nanotube (8,2)(16,4). In parenthesis the percentage over the total time for one self-consistent step.

	vdW-DF*	vdW-DF	Experimento [32]
a (Å)	2.476	2.485	2.459
c (Å)	3.590	3.565	3.336
E_{exf} (meV/at)	-53	-48	-52 $\pm$ 5

Table 3: Lattice parameters and exfoliation energy of graphite. * Results obtained by direct evaluation of Eq. 2 [34].

monolayer, the interlayer equilibrium distance and the exfoliation energy. We labeled the data obtained by direct evaluation with "vdW-DF*" [34]. The small discrepancies with our calculations are again attributed to the different basis sets. To illustrate how important is a good description of the basis-sets, we include Table 4. It was found essential to correct the basis set superposition error (BSSE) [3], especially when we are interested in interaction or adsorption energies. This error has been also corrected in Fig. 1.

The clathrate hydrates are crystalline water solid phases, in which small molecular gases and hydrocarbons are trapped in cavities formed by hydrogen-bonded water molecules.

	TZP	TZP-optim	TZP bsse
a (Å)	2.485	2.485	2.485
c (Å)	3.425	3.445	3.565
E_{exf}	-67	-70	-48

Table 4: Lattice parameters and exfoliation energy of graphite using different basis-sets. In the last column BSSE has been corrected. TZP-optim basis was optimized in graphite system with simplex [22].

 ISBN: 978-84-614-6167-7

Fig. 2 shows the different cavities for two structures (I and H). In these systems, three

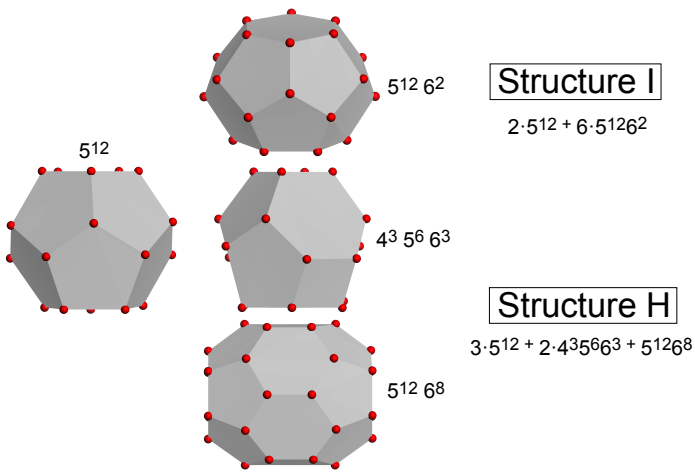


Figure 2: Structure I is formed by two 5^{12} cages and six $5^{12}6^2$ cages for a total of 46 H_2O molecules per unit cell. Structure H is formed by three 5^{12} cages , two $4^35^66^3$ cages and one $5^{12}6^8$ cage, and 34 H_2O molecules per unit cell. The dots indicate the oxygen atom positions. Hydrogen atoms are not shown. The cages are not scaled with respect to each other.

different interactions are combined: covalent interaction, responsible of the molecular bonds, hydrogen bonds, which keep the crystal structure, and van der Waals interactions, which dominate the adsorption energy of the guest molecules [24]. The adsorption energy is defined as the difference between the total energy of the clathrate, with the molecules inside, and that of the the empty clathrate plus the isolated molecules. Table 4 shows adsorption energy for methane and dioxide of carbon using GGA and vdW-DF. Although our definition of adsorption energy is useful to find the energy that the system gains, it cannot be directly interpreted in terms of stability. The structures without any guest molecules inside are not stable at any given temperature and pressure, and so we have to compare this energy gain with that of the stable phase of water (ice I_h). Taking this into account the structure is stabilized by van der Waals forces, when the number of guest molecules is higher than two [24]. This results were obtained with SIESTA (which uses numerical atomic orbitals as basis functions [27, 20]) and confirmed by plane waves calculations [16].

Fig. 3 shows rotational energies of methane in the $5^{12}6^2$ cavity (where the super-index indicates the faces with the shapes that are represent by the base number). The CH_4 rotates almost freely ($E_{barrier}$ <20 meV) inside all the cavities. Rotation of CO_2 shows higher energy barriers for all cavities (up to ∼600 meV in 5^{12}). This large difference can

 ISBN: 978-84-614-6167-7

Guest	Functional	Structure I		Structure H		
		5^{12}	$5^{12}6^2$	5^{12}	$4^35^66^3$	$5^{12}6^8$
CH_4	GGA	0.09	-0.07	0.07	0.05	-0.12
CH_4	vdW	-0.52	-0.59	-0.53	-0.54	-0.48
CO_2	GGA	0.35	0.10	0.29	0.22	0.01
CO_2	vdW	-0.41	-0.56	-0.41	-0.43	-0.38

Table 5: Adsorption energies (in eV per molecule) for single CH_4 and CO_2 molecules in one of the different cavities of clathrate structures I and H. Partially reproduced from [24].

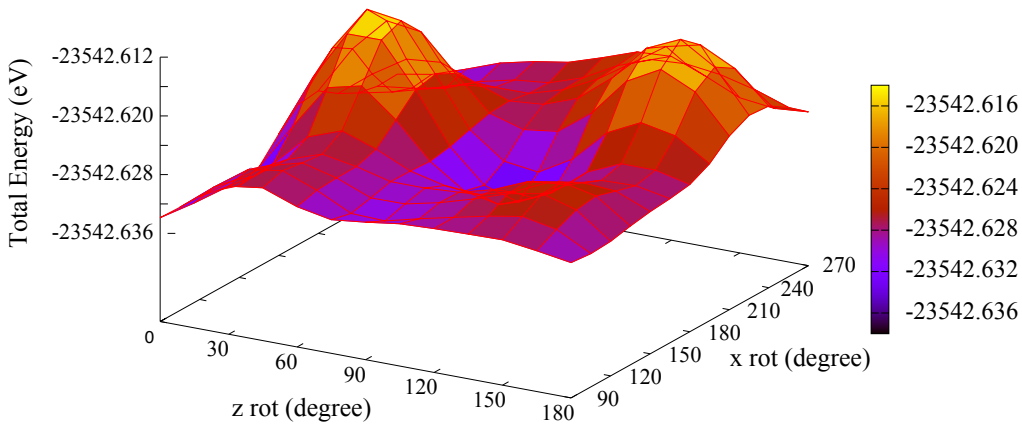


Figure 3: Rotation energy of CH_4 in $5^{12}6^2$ of structure I of clathrate hydrate.

be explained by the almost isotropic electronic density that methane shows, which interacts with an almost spherical cavity. However, this spherical potential is not dense in the sense that it comes from the position of the atoms of the water molecules. For this reason, CO_2 molecules with a cylindrical electronic density, show some easy direction orientations. Fig. 4 shows the rotation energy of CH_4 in the largest cavity of structure H. This cavity can hold up to five molecules. We see that the rotation is almost free when one molecule is in the cavity, but barriers increase with the number of molecules inside the cavity. When this number is higher than three, the interaction between molecules becomes very important and the adsorption energy decreases. This adsorption energy is exothermic up to the maximum occupation.

 ISBN: 978-84-614-6167-7

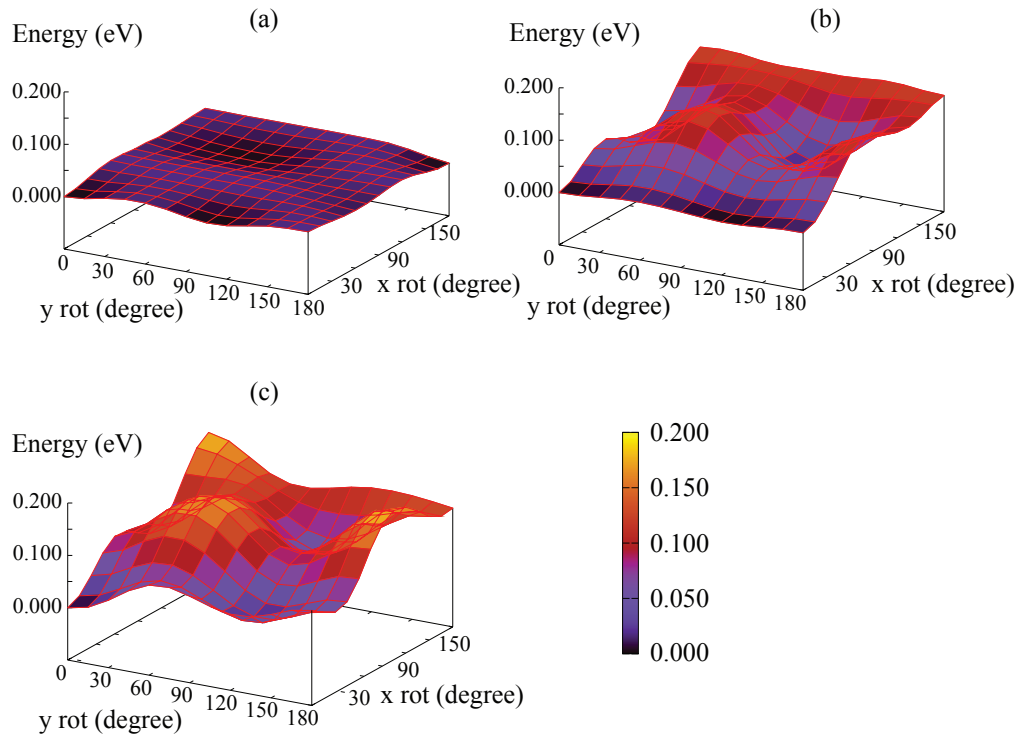


Figure 4: Rotation energy of CH_4 in the biggest cavity ($5^{12}6^8$) of structure H of clathrate hydrate as function of numbers of molecules(one, two, three molecules in a), b), c) respectively). Energy is given respect the relaxed energy.

4 Conclusions

We present a methodology for an accurate and efficient implementation of the vdW-DF functional. We show that it gives the same results as a direct evaluation of E_{XC}^{nl} with a small overhead in computational time for large systems, compared with popular GGA's. It can be concluded that any system suitable to be studied with local and semi-local approximations can be also studied using vdW-DF. Clathrate hydrates are stabilized by van der Waals forces. The movement and interactions of adsorbed molecules inside the cavities depend mainly on these interactions as well.

 ISBN: 978-84-614-6167-7

Acknowledgements

We would like to thank E. Artacho, D. C. Langreth and T. Thohansen for useful discussions. This work has been funded by grant FIS2009-12721 from Ministerio de Ciencia e Innovación de España.

References

[1] Y. Andersson, D. C. Langreth, and B. I. Lundqvist. van der waals interactions in density-functional theory. *Phys. Rev. Lett.*, 76(1):102–105, Jan 1996.

[2] L. C. Balbás, José Luís Martins, and José M. Soler. Evaluation of exchange-correlation energy, potential, and stress. *Phys. Rev. B*, 64(16):165110, Oct 2001.

[3] SF Boys and F Bernardi. Calculation of small molecular interactions by differences of separate total energies - some procedures with reduced errors. *Molec. Phys.*, 19(4):553–&, 1970.

[4] Svetla D. Chakarova-Käck, Elsebeth Schröder, Bengt I. Lundqvist, and David C. Langreth. Application of van der waals density functional to an extended system: Adsorption of benzene and naphthalene on graphite. *Phys. Rev. Lett.*, 96(14):146107, Apr 2006.

[5] Valentino R. Cooper, T. Thonhauser, and David C. Langreth. An application of the van der waals density functional: Hydrogen bonding and stacking interactions between nucleobases. *J. Chem. Phys.*, 128(20):204102, 2008.

[6] M. Dion, H. Rydberg, E. Schröder, D. C. Langreth, and B. I. Lundqvist. Van der waals density functional for general geometries. *Phys. Rev. Lett.*, 92(24):246401, Jun 2004.

[7] J. Dobson and J. Wang. Successful test of a seamless van der waals density functional. *Phys. Rev. Lett.*, 82(10):2123–2126, Mar 1999.

[8] J. F. Dobson and B. P. Dinte. Constraint satisfaction in local and gradient susceptibility approximations: Application to a van der waals density functional. *Phys. Rev. Lett.*, 76(11):1780–1783, Mar 1996.

[9] S. Grimme. Accurate description of van der waals complexes by density functional theory including empirical corrections. *J Comp. Chem.*, 25(12):1463–1473, 2004.

[10] S. Grimme. Semiempirical gga-type density functional constructed with a long-range dispersion correction. *J Comp. Chem.*, 27(15):1787–1799, 2006.

 ISBN: 978-84-614-6167-7

[11] Joe Hooper, Valentino R. Cooper, Timo Thonhauser, Nichols A. Romero, Frank Zerilli, and David C. Langreth. Predicting C-H/pi interactions with nonlocal density functional theory. *ChemPhysChem*, 9(6):891–895, 2008.

[12] W. Kohn, Y. Meir, and D. E. Makarov. van der waals energies in density functional theory. *Phys. Rev. Lett.*, 80(19):4153–4156, May 1998.

[13] D. C. Langreth, M. Dion, H. Rydberg, E. Schroder, P. Hyldgaard, and B. I. Lundqvist. Van der Waals density functional theory with applications. *J. Int. J. Quantum Chem.*, 101(5, Sp. Iss. SI):599–610, 2005.

[14] D. C. Langreth, B. I. Lundqvist, S. D. Chakarova-Käck, V. R Cooper, M. Dion, .P Hyldgaard, A. Kelkkanen, J. Kleis, L. Kong, S. Li, P. G. Moses, E. Murray, A. Puzder, H. Rydberg, E. Schröder, and T. Thonhauser. A density functional for sparse matter. *J. Phys.: Condens. Matter.*, 21(8):084203, 2009.

[15] K. Lee, É. D. Murray, L. Kong, D. I. Lundqvist, and D. C. Langreth. Higher-accuracy van der waals density functional. *Phys. Rev. B*, 82(8):081101, Aug 2010.

[16] Qi Li, Brian Kolb, Guillermo Román-Pérez, José M. Soler, Félix Yndurain, David C. Langreth, and Timo Thonhauser. *To be submitted*, 2011.

[17] B. I. Lundqvist, Y. Andersson, H. Shao, S. Chan, and D. C. Langreth. Density Functional Theory Includings van der Waals Forces. *J. Int. J. Quantum Chem.* , 56(4):247–255, 1995.

[18] Klaus Müller-Dethlefs and Pavel Hobza. Noncovalent interactions: A challenge for experiment and theory. *Chem. Rev.*, 100(1):143–168, 2000.

[19] J.F. Ogilvie and Frank Y.H. Wang. Potential-energy functions of diatomic molecules of the noble gases i. like nuclear species. *J. Mol. Struct.*, 273:277 – 290, 1992.

[20] P. Ordejón, E. Artacho, and J. M. Soler. Self-consistent order-n density-functional calculations for very large systems. *Phys. Rev. B*, 53:R10441, 1996.

[21] J. P. Perdew and Alex Zunger. Self-interaction correction to density-functional approximations for many-electron systems. *Phys. Rev. Lett.*, 23(10):5048–5079, May 1981.

[22] William H. Press, Saul A. Teukolsky, William T. Vetterling, and Brian P. Flannery. *Numerical Recipes*. Cambridge University Press, Cambridge, 1992.

[23] A. Puzder, M. Dion, and D. C. Langreth. Binding energies in benzene dimers: Nonlocal density functional calculations. *J. Chem. Phys.*, 124(16):164105, 2006.

 ISBN: 978-84-614-6167-7

[24] Guillermo Román-Pérez, Mohammed Moaied, Jose M. Soler, and Felix Yndurain. Stability, adsorption, and diffusion of ch_4, co_2, and h_2 in clathrate hydrates. *Phys. Rev. Lett.*, 105(14):145901, Sep 2010.

[25] Guillermo Román-Pérez and José M. Soler. Efficient implementation of a van der waals density functional: Application to double-wall carbon nanotubes. *Phys. Rev. Lett.*, 103(9):096102, Aug 2009.

[26] H. Rydberg, B. I. Lundqvist, D. C. Langreth, and M. Dion. Tractable nonlocal correlation density functionals for flat surfaces and slabs. *Phys. Rev. B*, 62(11):6997–7006, Sep 2000.

[27] J. M. Soler, E. Artacho, J. D. Gale, A. García, J. Junquera, P. Ordejón, and D. Sánchez-Portal. The Siesta method for *ab initio* order-N materials simulation. *J. Phys. Condens. Matter.*, 14:2745, 2002.

[28] T. Thonhauser, V. R. Cooper, S. Li, A. Puzder, P. Hyldgaard, and D. C. Langreth. Van der waals density functional: Self-consistent potential and the nature of the van der waals bond. *Phys. Rev. B*, 76(12):125112, Sep 2007.

[29] T. Thonhauser, Aaron Puzder, and David C. Langreth. Interaction energies of monosubstituted benzene dimers via nonlocal density functional theory. *J. Chem. Phys.*, 124(16):164106, 2006.

[30] O. A. Vydrov and T. Van Voorhis. Improving the accuracy of the nonlocal van der waals density functional with minimal empiricism. *J. Chem. Phys.*, 130(10):104105, 2009.

[31] O. A. Vydrov and T. Van Voorhis. Nonlocal van der waals density functional made simple. *Phys. Rev. Lett.*, 103(6):063004, Aug 2009.

[32] Renju Zacharia, Hendrik Ulbricht, and Tobias Hertel. Interlayer cohesive energy of graphite from thermal desorption of polyaromatic hydrocarbons. *Phys. Rev. B*, 69(15):155406, Apr 2004.

[33] Yingkai Zhang and Weitao Yang. Comment on "generalized gradient approximation made simple". *Phys. Rev. Lett.*, 80(4):890, Jan 1998.

[34] Eleni Ziambaras, Jesper Kleis, Elsebeth Schröder, and Per Hyldgaard. Potassium intercalation in graphite: A van der waals density-functional study. *Phys. Rev. B*, 76(15):155425, Oct 2007.

 ISBN: 978-84-614-6167-7

Proceedings of the 11th International Conference on Computational and Mathematical Methods in Science and Engineering, CMMSE 2011 26–30 June 2011.

Certificateless Secure Beaconing in Vehicular Ad-hoc Networks

Eun-Kyung Ryu[1] and Kee-Young Yoo[2]

[1] *Graduate School of Electrical Engineering and Computer Science, Kyungpook National University*

[2] *School of Computer Science and Engineering, Kyungpook National University*

emails: ekryu@knu.ac.kr, yook@knu.ac.kr

Abstract

The technology of vehicular ad-hoc networks (VANETs) allows vehicles communicate with road-side infrastructure or with nearby vehicles, supporting a wide range of promising vehicular communication applications and services. It relies on the periodic transmission of packets, called beaconing, as single-hop link-layer broadcast to nearby vehicles or road-side units. However, due to the inherent broadcast nature of the wireless channels, beaconing is easily exposed to security attacks, such as spoofing, manipulation, or replaying. In this paper, we propose a new scheme for securing beaconing based on the mechanism of identity-based signature. Our scheme has a number of crucial advantages. It mitigates the requirement of the public key infrastructure (PKI), but without degrading security strength. Consequently, it surpasses existing PKI-based solutions in terms of both the communication and computation overhead associated with certificates. We also show how the scheme can be extended to implement a forward-secure key exchange scheme in VANETs.

Key words: Vehicular ad-hoc networks, VANETs, Security, Efficiency

1 Introduction

The technology of vehicular ad-hoc networks (VANETs) allows vehicles communicate with road-side infrastructure or with nearby vehicles. IEEE 802.11p will standardize the physical and medium access control layers for VANETs. Higher layer is also being standardized and draft standards have been released in the IEEE 1609 set of standards. The VANETs are expected to support a wide range of promising vehicular communication applications and services, relying on the periodic transmission of packets as single-hop link-layer broadcast to nearby vehicles or road-side units [5, 10]. Those packets, called beaconing, contain information like the current location, heading

 ISBN: 978-84-614-6167-7

Payload (m)	Location (x, y, z)	Time (t)	Signature (R, s)	Certificate (pk)
32	11	8	56	125

Figure 1: IEEE 1609.2 Message Format for Secure Beaconing

and speed of the sending vehicle, by which every vehicle is aware of neighboring vehicles within a certain range. The beaconing is also essential for geographic routing and message dissemination in VANET. However, due to the inherent broadcast nature of the wireless channels, beaconing is easily exposed to security attacks, such as spoofing, manipulation, or replaying. Securing beaconing is therefore one of major challenges to the widespread deployment of VANET technology.

A simple approach to address this problem is to use symmetric-based technique for the becon packet authentication, as in [1, 2, 3, 4]. However, it requires time synchronization and allows only delayed beacon verification after key disclosure by the sender. This approach is not applicable to dynamic and time-critical vehicular networks. An alternative and promising approach is to use signature based asymmetric cryptographic mechanism like ECDSA as specified in the IEEE 1609.2 standard draft [7]. In this approach, the underlying strategy is to let vehicles to equip with asymmetric cryptographic key pairs (VK, SK) and certificates issued by a trusted certification authority. Then, all beacons get signed using the vehicle's signature key (SK) and receivers verify them using the verification key VK. Signature and certificate containing VK are attached to the beacon, as shown in Figure 1, requiring additional 181 octets payload (56 octets for the signature and 125 octets for the certificate). This approach not only imposes the public-key infrastructure (PKI) requirement, but also causes a significant overhead of communication bandwidth and processing power associated with signature and certificate. The fact that the problem is much worse in case vehicle density is high limits the use of existing PKI-based schemes [7, 8, 9].

In this paper, we propose an efficient scheme for securing beaconing in VANETs that is based on the mechanism of identity-based signature. The main advantages of our scheme are as follows: It mitigates the requirement of the public key infrastructure, but without degrading security strength in the PKI-based approach. That is, it fully provides the security properties of message authentication, resistance to replay attacks and non-repudiation needed for securing transmitted beacons without using certificates. It significantly reduces the communication overhead associated with certificates, as well as computation cost for signing and verifying beacons. We also show how the scheme can be extended to implement a forward-secure key exchange scheme in VANETs. The rest of this paper is structured as follows. We first describe the communication model we consider and the goals of protocol design in Section 2. We describe our our solution for certificateless secure beaconing in Section 3. The security analysis and performance are discussed in Section 4. We further discuss about the support for forward-secure key exchange in Section 5. Finally we conclude in Section 6.

 ISBN: 978-84-614-6167-7

2 Model and Assumptions

2.1 Network Model

The Model. The network model we consider includes road-side units (RSUs) and on-boad units (OBUs). RSUs are fixed units distributed in the network. RSUs are connected to each other through a high-speed backbone network connecting to the internet. RSUs transmit periodic broadcast packets to nearby vehicles or road-side units, OBUs are embedded in vehicles and can function while moving. By using OBUs, vehicles can communicate with the RSUs or other OBUs. Each OBU is equipped with a global positioning service receiver. The communication between RSUs and OBUs is based on the dedicated short-range communications (DSRC) identified as IEEE 802.11p. All units that expect to use the secure VANET application services must be registered with a Trusted Authority (TA), by which the communication entities obtain cryptographic keys associated with their long-term identities. We assume that the TA is fully trusted by all communication entities in VANET settings. We also assume that each entity is equipped with a tamper-resistant hardware security module to store its security-related materials.

Attacker. We assume that attackers have following abilities. They observe and intercept any transmitted packet they want in VANETs, as usual. From the intercepted packets, they are able to replay the session at some later point in time. They also are capable of adding or in some other way altering transmitted packets on the channel. We assume that attackers are computationally bounded and hence cannot break standard cryptographic algorithms

2.2 Design Goals

The protocol for securing such vehicular communication should have the following properties.

Correctness. The protocol should be correctly verifiable.

Efficiency. The number of operations, the number of message exchanges and the total number of transmitted bits required to execute the protocol should be minimized.

Security. The primary threats against vehicular communication is that an attacker might forge messages or replay the messages sent some time before in order to disrupt the traffic. To prevent the threats, the protocol must fulfill the fundamental security attributes: unforgeability, resistance to replay attacks, message integrity, and non-repudiation. The security attributes mentioned above are briefly described below.

- *Message Authentication:* It is computationally infeasible for an attacker to masquerade an honest sender in creating a valid message that can be accepted by the protocol. It represents that the recipient is able to verify that the contents of the

 ISBN: 978-84-614-6167-7

message sent by the sender have not been tampered with and altered. This property addresses the unauthorized alteration of message, such as such as insertion, deletion, and substitution.

– *Resistance to Replay Attacks:* A replay attack is a form of network attack in which a valid message transmitted by an honest sender is maliciously or fraudulently repeated or delayed to the same or a different recipient. The property of resistance to replay attacks assures that an attacker, obtaining a valid message previously transmitted by an honest sender, is not able to maliciously retransmit it to the same or a different recipient.

– *Non-repudiation:* The communication entities are not able to fraudulently deny the transmission or the content of their previous messages. That is, the recipient has the ability to prove to a third party that the sender has sent the message. Thus when disputes arise the third party can determine which entity was the true source of the message.

3 Our Scheme: Certificateless Secure Beaconing

We use the Schnorr's signature scheme as the underlying cryptographic primitive in our construction. Let $\mathbb{G}$ be an additive cyclic group of prime order q, and let P be a generator of $\mathbb{G}$. Recall that a Schnorr signature of a message m under public key $X \in \mathbb{G}$ is a pair $(R, s) \in \mathbb{G} \times \mathbb{Z}_q$ such that $sP = R + eX \in \mathbb{G}$, where e is the hashed value on $R||m$. The Schnorr's signature scheme generates short signatures and is efficient.

The scheme consists of three phases: system setup, broadcasting signed message and message verification phases. In the system setup, the TA generates system parameters and a master secret key. The TA is also responsible for issuing a public and private key pair for all communication entities, RSUs and OBUs. The key pair is coupled with the identity information of each communication party and TA's master key, using the Schnorr signature scheme. Note that the identity information can be any string including entity's unique serial number, its physical location information and so on. This setup phase is performed before providing any of the VANET services or applications. In broadcasting signed messages, a message sender with a pair of public and private keys constructs a signature σ on a message m of arbitrary length and broadcasts them with the sender's credential cred. Upon arriving the message, the receiver verifies the message and its signature using the sender's credential. More details are given below.

3.1 System Setup

Let $\mathbb{G}$ be an additive cyclic group of prime order q, and let P be a generator of $\mathbb{G}$. Let l be a security parameter. TA generates system-wide parameters as follows. TA chooses a random value $x \in \mathbb{Z}_q$ as the system master key and sets $X = xP$ as a system parameter. It also chooses two secure cryptographic hash functions $H_0, H_1 : \{0,1\}^* \to \mathbb{Z}_q^*$. Then, TA publishes $(\mathbb{G}, q, P, X, H_0, H_1)$ as system-wide parameters and keeps the master key

 ISBN: 978-84-614-6167-7

x secretly. We assume that communication parties in VANETs, OBUs and RSUs, are preloaded with the public parameters.

For each communication party with a unique identity information id, TA generates a pair (Y, k) of public and private keys as follows:

1) Select a random value $y \in \mathbb{Z}_q$.

2) Compute $Y = yP,\ k = (y + e \cdot x) \bmod q$, where $e = H_0(id||Y||T_{exp})$ and T_{exp} is the expiration time of the key pair.

3) The pair (Y, k, T_{exp}) is assigned to each party.

3.2 Broadcasting Signed Messages

Let m be a message of arbitrary length to be signed. The sender S performs the following for generating a signature on m.

1) Select a random value $r \in \mathbb{Z}_q$.

2) Compute $R = rP$ and $s = r + e_1 \cdot k$, where $e_1 = H_1(R||m||T_{cur})$ and T_{cur} represents the sender's current timestamp.

4) Set $\sigma = (R, s)$ as the signature of m.

5) Broadcast the packet $(m, \sigma, \mathsf{cred})$, where the sender's credentials cred includes the tuple of $(id, Y, T_{exp}, T_{cur})$.

$$Sender \rightarrow * : \quad m||\sigma||\mathsf{cred}$$

3.3 Messages Verification

On receiving the broadcast packet $(m||\sigma||\mathsf{cred})$, the packet receiver performs the following.

1) Check if T_{cur} and T_{exp} are valid. If not, reject the packet.

2) Compute $v_1 = sP$ and $v_2 = R + e_1(Y + e_0X)$, where $e_0 = H_0(id||Y||T_{exp})$ and $e_1 = H_1(R||m||T_{cur})$.

3) Accept the message if and only if $v_1 = v_2$.

Correctness. Note that the correctness of the protocol follows because of the relation $sP = R + e_1(Y + e_0X) = H_1(R||m||T_cur)(yP + H_0(id||Y||T_{exp})$.

 ISBN: 978-84-614-6167-7

4 Analysis

4.1 Security

We now show that our new construction, denoted as CL-SB, satisfies the required security properties described in Section 2.2. The security of the CL-SB relies on the discrete logarithm assumption.

Message Authentication. It is not difficult to see that the above CL-SB provides the property of message authentication. This is due to the fact that only legitimate entities who possess the private key corresponding to its identity can create a signature on the transmitted message using its own private key. More specifically, let us consider that an attacker tries to masquerade an honest sender, RSU or OBU, and to send an arbitrary m' to neighboring receivers for fraudulent purpose. In order to do this, the challenge of the attacker is to derive a valid signature value (R', s') of the message m', where $R' = r'P, s' = r' + e_1 k', e_1 = H_1(R'||m'||T_{cur})$ and r' is a random value chosen by the attacker. It however is exactly an instance of underlying primitive, Schonrr signature scheme that is proven to the discrete logarithm problem in the random oracle model in [13]. That is, we can say that the CL-SB is secure against unauthorized insertion, deletion and substitution attack.

Resistance to Replay Attacks. With regards to the security property against replay attacks, an attacker might intercept a valid message previously transmitted by an honest sender, but is not able to maliciously retransmit it the same or different recipients. This is due to the fact that in CL-SB protocol a timestamp is cryptographically combined with a message for detecting message replay. Note that from the use of timestamp the protocol not only enables to avoid the replay attacks but also has the significant advantage of fewer messages, and no requirement to maintain pairwise long-term information, such as sequence numbers or random numbers.

Non-reputation. TAs mentioned before, non-repudiation is necessary to prevent legitimate entities from denying the transmission of their messages. The property of non-repudiation in CL-SB is achieved by requiring all transmitted messages to be cryptographically combined with its sender's private key. That is, the fact that the entity's private key is known only to itself guarantees the property of non-repudiation, as in any other signature-based schemes.

4.2 Performance

The metrics used to evaluate the performance of CL-SB are the computational and the communication costs required for secure beaconing. Here we compare our construction, CL-SB, with the ECDSA based scheme (denoted as ECDSA-SB) in IEEE 1609.2 with regards to computational and protocol overhead for secure beacons. Unlike ECDSA-SB, CL-SB does not assume the use of public-key infrastructure, by which it achieves greater efficiency and easy implementation.

For each scheme we show the computational cost of signing, the computational

Scheme	Sign	Veriy	$\|\text{sig}\|$	$\|\text{cert}(pk)\|$
ECDSA-SB [7]	1 exp + 1 inv	4 exp + 2 inv	56	125
CL-SB	1 exp	3 exp	40	20

Table 1: Comparison of secure beaconing schemes

cost of verifying of a beacon, the size of signature payload and the size of certificate payload required for sending a single secure beacon. For computational cost we only consider the most expensive operations, point multiplication over an elliptic curve and multiplicative inverse operation in $\mathbb{Z}_q$. All sizes are in octets. We denote by "exp" the point scalar multiplication over an elliptic curve and by "inv" the multiplicative inverse operation.

The signing in CL-SB requires one point scalar multiplication, while one point scalar multiplication and one inverse operation in ECDSA-SB. The verification in CL-SB requires only three point multiplication operations, while four point multiplication operations and two inverse operations in ECDSA-SB. Our signatures are 40 octets (320bits), while 56 octets (448bits) in ECDSA-SB. Note that the CL-SB does not require the certificate of public key, by which the additional payload for public key is only 20 octets (160bits) which is significantly shorter than in ECDSA-SB.

5 Supporting Forward Secure Key Exchange

We can easily extend the CLSB protocol to protect transmitted messages by supporting the security service of forward-secure key exchange. For simplicity of notation, let us denote A be a communication initiator and B be a corresponding responder in VANET. The protocol works as follows.

ROUND 1 $A \to B: \ R'_A||\sigma_A||Y_A||t_{e_A}||t_{c_A}$
A chooses two random values a, a' and computes $R_A = aP$, $R'_A = a'P$, $s_A = a + H(R_A||R'_A||t_{c_A}) \cdot k_A$, where t_{c_A} is a A's current timestamp. A sets $\sigma_A = (R_A, s_A)$ as the signature of the keying material R'_A. Then, A sends the tuple $(R'_A, \sigma_A, Y_A, t_{e_A}, t_{c_A})$ to B, where t_{e_A} is the expiration time of A's credentials.

ROUND 2 $B \to A: \ R'_B||\sigma_B||Y_B||t_{e_B}||t_{c_B}$
Similarly, B chooses two random value b, b' and computes $R_B = bP$, $R'_B = b'P$ $\sigma_B = b + H(R_B||R'_B||T_{cur_B}) \cdot k_B$, where and T_{cur_B} is a B's current timestamp. B sets $\sigma_B = (R_B, s_B)$ as the signature of the keying material R'_B. B then sends the tuple $(R'_B, \sigma_B, Y_B, t_{e_B}, t_{c_B})$ to A. From the packet $(R'_B||\sigma_B||Y_B||t_{e_B}||t_{c_B})$, A checks if t_{c_B} and t_{e_B} is valid. Then, A computes $v_B = s_B P$, $v'_B = R_B + H_1(R_B||R'_B||t_{cur_B})(Y_B + H_0(id_B||Y_B||t_{e_B}) \cdot X)$. A accepts R'_B and computes $S_A = a'(R'_B) = a'b'P$ to obtain the shared secret key K_A, if and only if $v_B = v'_B$. The shared secret key is computed by $K_A = \mathsf{kdf}(R'_A, R'_B, S_A)$.

 ISBN: 978-84-614-6167-7

Similarly, B first checks if t_{c_A} and t_{e_A} is valid. Then, B accepts the value R'_A and computes the shared secret $K_B = \mathsf{kdf}(R'_A, R'_B, S_B)$, if and only if $v_A = v'_A$, where $v_A = s_A P$, $v'_A = R_A + H_1(R_A||R'_A||t_{cur_A})(Y_A + H_0(id_A||Y_A||t_{e_A}) \cdot X)$ and $S_B = b'(R'_A) = a'b'P$, as done in A.

6 Concluding Remarks

In this paper, we proposed a new scheme for securing beaconing that is based on the mechanism of identity-based signature. The main advantages of our scheme are as follows: It guarantees security properties of integrity and authentication on transmitted beacons by all beacons get signed using the vehicle's private key and receivers verify them using publicly available information. Unlike the ECDSA based scheme in IEEE 1609.2 it does not assume the use of public-key infrastructure, by which it achieves greater efficiency and easy implementation. We also showed how the scheme can be extended to implement a forward-secure key exchange scheme in VANETs.

References

[1] Ken Laberteaux and Yih-Chun Hu, *Strong VANET Security on a Budget*, Proceeding of 4th Conference on Embedded Security in Cars (ESCAR 2006), (2006).

[2] KenAdrian Perrig, Ran Canetti, Dawn Song, and J. D. Tygar, *Efficient and Secure Source Authentication for Multicast*, Proceeding of NDSS '01, (2001).

[3] Adrian Perrig, Ran Canetti, J.D. Tygar, and Dawn Xiaodong Song, *Efficient Authentication and Signing of Multicast Streams over Lossy Channels*, Proceeding of IEEE Symposium on Security and Privacy, (2000).

[4] Adrian Perrig, Ran Canetti, J.D. Tygar, and Dawn Xiaodong Song, *The TESLA Broadcast Authentication Protocol*, Cryptobytes 5 (2), Summer/Fall 2002.

[5] Elmar Schoch, Frank Kargl,, *On the Efficiency of Secure Beaconing in VANETs*, Proceeding of 3rd ACM Conference on Wireless Network Security (WiSec 2010) (2010).

[6] IEEE 802.11 - Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications, (2007).

[7] IEEE Standard for Information Technology, *IEEE P1609.2 -Standard for Wireless Access in Vehicular Environments (WAVE) - Security Services for Application and Management Services,* (2006).

 ISBN: 978-84-614-6167-7

[8] P. Papadimitratos et al., *Secure Vehicular Communications: Design and Architecture Application and Management Services*, IEEE Comm. Magazine 46(11) (2008).

[9] M. Raya and J.-P. Hubaux,, *The Security of Vehicular Ad Hoc Networks*, Proceeding of SASN 2005 (2005).

[10] F. Kargl, E. Schoch, B. Wiedersheim, and T. Leinm, *Secure and Efficient Beaconing for Vehicular Networks*, Proceeding of 5th ACM VANET 2008 (2008).

[11] P. Papadimitratos, G. Calandriello, A. Lioy, and J.-P. Hubaux,, *Impact of Vehicular Communication Security on Transportation Safety*, Proceeding of MOVE 2008 (2008).

[12] Lei Zhang, Qianhong Wu, Agusti Solanas, and Josep Domingo-Ferrer, Senior Member, *Scalable Robust Authentication Protocol for Secure Vehicular Communications*, IEEE TRANS. ON VEHICULAR TECHNOLOGY 59(4) (2010).

[13] C. Schnorr, *Efficient Signature Generation by Smart Cards*, Journal of Cryptology 4(3) (1991) 161–174.

 ISBN: 978-84-614-6167-7

Proceedings of the 11th International Conference on Computational and Mathematical Methods in Science and Engineering, CMMSE 2011 26–30 June 2011.

On Some Finite Difference Algorithms for Pricing American Options and Their Implementation in Mathematica

A. A. E. F. Saib[1], Y. D. Tangman[1], N. Thakoor[1] and M. Bhuruth[1]

[1] *Department of Mathematics, Faculty of Science, University of Mauritius*

emails: aslam.saib@umail.uom.ac.mu, y.tangman@uom.ac.mu, nawdha.thakoor@umail.uom.ac.mu, mbhuruth@uom.ac.mu

Abstract

The accurate valuation of American options is a difficult problem due to the possibility of early exercise and the performance of a numerical algorithm for solving the American problem is strongly dependent on its capability to accurately locate the optimal exercise boundary. This paper provides some new insights on the performances of three finite difference algorithms for approximating the American option price under the Black-Scholes model. The algorithms we consider are the Han and Wu method, the operator splitting technique of Ikonen and Toivanen and the optimal compact approximation of Tangman, Gopaul and Bhuruth and these methods are implemented in the Mathematica environment. Our choice of software is motivated by Mathematica's capability for supporting functional programming and dynamic interactivity which allows the development of sophisticated codes. Price comparisons over various benchmark problems are carried out and it is shown that the optimal compact approximation produces option prices with higher accuracy and there is not much loss in accuracy when coarser grids are chosen whereas the Han and Wu and Ikonen and Toivanen methods are more sensitive to the grid size chosen. Mathematica codes are provided for all the three methods.

Key words: American Options, Partial Differential Equations, Free Boundary Value Problem, Linear Complementarity Problem, Mathematica

MSC 2000: 35A35; 65M70; 62P05

1 Introduction

American options are among the most traded derivatives in financial markets as they can be optimally exercised at any time up to maturity date. Closed form solutions for American options do not exist and various numerical and analytical techniques have

 ISBN: 978-84-614-6167-7

been proposed for obtaining approximate prices. Analytical approaches to approximating the exact solutions usually yield formulae which are difficult to use in practice which explains the popularity of numerical methods for the valuation problem. One numerical approach which computes the entire price function consists of formulating the valuation problem either as a free-boundary value problem or as a linear complementarity problem (LCP).

For the linear complementarity problem, the PSOR method [4] is known to converge slowly. Another technique for the solution of the LCP is to add a penalty term to the Black-Scholes variational inequality [5, 9] but this technique usually requires the solutions of nonlinear systems. A more efficient technique which avoids the solution of linear complementarity problems is the algorithm proposed by Ikonen and Toivanen [10]. This technique uses an operator splitting (OS) method to decouple the Black-Scholes operator and the constraint for the option value.

The Han and Wu (HW) [7] algorithm is based on transforming the Black-Scholes [2] equation into a heat equation on an infinite domain and then computing the numerical solution on a bounded domain with exact artificial boundary conditions at one end of this truncated domain. At the other end of the domain, the free boundary is located using a simple numerical technique based on properties of the solution to a heat equation.

A different method for the accurate location of the free boundary was proposed in [13]. This algorithm computes the difference between the American and European option prices by using an optimal compact approximation (OCA) on a non uniform grid for the finite difference discretization of the heat equation. The location of the free boundary is determined using the smooth pasting condition. Since the American and European option prices both satisfy the Black-Scholes equation in the continuation region, the discontinuity at the strike price in the payoff function is removed and this enhances the accuracy of the computed solutions.

Our experience with the American option pricing problem has shown that the above three methods (HW, OS and OCA) are among the most efficient for the pricing of American options. We implement these methods in Mathematica and we compare their accuracy against both finite difference algorithms and analytical approximation methods. An outline is as follows. In §2 we recall the American option pricing problem and in §3 we describe the three numerical algorithms. Price comparisons are given in §4 and the Mathematica codes are given in the Appendix.

2 The American Option Valuation Problem

We consider a financial market consisting of a risky asset with price process $\{S_t\}_{t\geq 0}$ and constant volatility $\sigma > 0$ in a risk neutral economy with fixed rate of return $r > 0$. Under the risk neutral measure $\mathbb{Q}$, the dynamics of the Black-Scholes model is given by

$$\frac{dS_t}{S_t} = (r - \delta)dt + \sigma dW_t,$$

where δ denotes the continuous dividend yield and W_t is standard Brownian motion.

 ISBN: 978-84-614-6167-7

Let $V(S,\,t)$ be the price of a European option with strike price K and maturity date T. Then V is the solution of the Black-Scholes equation

$$\frac{\partial V}{\partial t}+\frac{1}{2}\sigma^2S^2\frac{\partial^2V}{\partial S^2}+rS\frac{\partial V}{\partial S}-rV=0, \tag{1}$$

with terminal condition $V(S,\,T)=(S-K)^+=\max(S-K,\,0)$ for a call option and $V(S,\,T)=(K-S)^+$ for a put option. The price $V(S,\,t)$ of a European call can be explicitly calculated and is given by

$$V(S,\,t)=S\Phi\left(d_1\right)-Ke^{-r(T-t)}\Phi(d_2),$$

where

$$d_2=\frac{\log\left(S/K\right)+\left(r-\sigma^2/2\right)\left(T-t\right)}{\sigma\sqrt{T-t}},\quad d_1=d_2+\sigma\sqrt{T-t},$$

and ϕ is the distribution function of the standard normal distribution $Z\sim N(0,\,1)$. The formula for European puts can be obtained using the put-call parity [1, p.123].

The possibility for early exercise means that the American call option price, $V(S,\,t)$ satisfies the constraint $V(S,\,t)\geq(S-K)^+$. The contact point, $S_f(t)$, which varies with t is a point where $V(S,\,t)>(S-K)^+$ for $S<S_f(t)$ and $V(S,\,t)=(S-K)$ for $S\geq S_f(t)$. Let

$$\mathcal{L}_SV=\frac{1}{2}\sigma^2S^2\frac{\partial^2V}{\partial S^2}+rS\frac{\partial V}{\partial S}-rV, \tag{2}$$

Then the free boundary value formulation for computing the American call price requires the solution of [11]

$$\frac{\partial V}{\partial t}(S,\,t)+\mathcal{L}_SV(S,\,t)=0,\quad 0<S<S_f(t),\quad 0\leq t\leq T, \tag{3}$$

$$\begin{aligned}V(S,\,T)&=(S-K)^+,\quad 0\leq S\leq S_f(T),\\ V(S_f(t),\,t)&=(S_f(t)-K)^+,\quad 0\leq t\leq T,\\ \frac{\partial V}{\partial S}(S_f(t),\,t)&=1,\quad 0\leq t\leq T,\\ V(S,\,t)&\to 0,\quad \text{as}\quad S\to 0,\quad 0\leq t\leq T.\end{aligned} \tag{4}$$

The condition (4) is known as the smooth pasting condition and we find that the free boundary divides the strip $(0,\,\infty)\times[0,\,T]$ into two regions, the continuation region $R_1=(0,\,S_f(t))\times[0,\,T]$ in which the Black-Scholes equation (1) holds and $R_2=[S_f(t),\,\infty)\times[0,\,T]$ which represents the exercise region in which $rK<\delta S$. Thus

$$\frac{\partial V}{\partial t}+\mathcal{L}_SV<0. \tag{5}$$

This leads to the LCP formulation for the American call option given by

$$\begin{aligned}\frac{\partial V}{\partial t}+\mathcal{L}_SV&\leq 0,\quad S>0,\,0\leq t\leq T,\\ V(S,\,T)&=(S-K)^+,\quad S>0,\\ V(S,\,t)&\geq(S-K)^+,\\ \left[\frac{\partial V}{\partial t}+\mathcal{L}_SV=0\right]&\vee\left[V(S,\,t)=(S-K)^+\right].\end{aligned}$$

 ISBN: 978-84-614-6167-7

3 Finite Difference Algorithms

We provide brief descriptions of the Han-Wu, Ikonen-Toivanen and Tangman-Gopaul-Bhuruth algorithms. For full details on the numerical techniques, the reader is referred to the original papers [7, 10, 13]. The Appendix lists the Mathematica codes for the three numerical schemes.

3.1 The Han-Wu Algorithm

The Han-Wu algorithm transforms the Black-Scholes equation into a heat equation on an infinite domain. This problem is localized to a finite computational domain and at one end of this domain, an artificial boundary condition is accurately determined. At the other end of the domain, the free boundary is located using a scheme based on properties of the Black-Scholes equation.

We consider the transformations given by $\tau = 2(T-t)/\sigma^2$, $k = 2r/\sigma^2$, $\delta^* = 2\delta/\sigma^2$, $\tau^* = \sigma^2 T/2$, $k' = r^* - \delta^*, S = Ke^x, S_f^*(\tau) = S_f(T - 2\tau/\sigma^2), S_f^*(\tau) = Ke^{x_f(\tau)}$ and $V^*(S, \tau) = V(S, T - 2\tau/\sigma^2)$. Then letting

$$V^*(S, \tau) = Ke^{\alpha x + \beta \tau} u(x, \tau), \tag{6}$$

the problem (3) is transformed to

$$\begin{aligned}
\frac{\partial u}{\partial \tau} &= \frac{\partial^2 u}{\partial x^2}, \quad -\infty < x < x_f(\tau), \quad 0 \le \tau \le \tau^*, && (7)\\
u(x, 0) &= g(x, 0), \qquad -\infty < x \le x_f(0),\\
u(x_f(\tau), \tau) &= g(x_f(\tau), \tau), \quad -\infty < x \le x_f(\tau), \quad 0 \le \tau \le \tau^*,\\
\alpha u(x_f(\tau), \tau) + \frac{\partial u}{\partial x}(x_f(\tau), \tau) &= e^{(1-\alpha)x_f(\tau) - \beta\tau}, \quad 0 \le \tau \le \tau^*, && (8)\\
u(x, \tau) &\to 0 \quad \text{as} \quad x \to -\infty,
\end{aligned}$$

where $g(x, \tau) = e^{-\alpha x - \beta \tau}(e^x - 1)^+$ is the transformed payoff.

The technique for locating the free-boundary and the derivation of the artificial boundary condition are the two main steps in the algorithm. For $a < 0$, let Γ_a given by $\Gamma_a = \{(x, \tau) | x = a, 0 \le \tau \le \tau^*\}$ be the artificial boundary condition. Then the exact boundary condition on Γ_a is given by

$$\frac{\partial u(a, \tau)}{\partial x} = \frac{1}{\sqrt{\pi}} \int_0^\tau \frac{\partial u(a, \lambda)}{\partial \lambda} \frac{d\lambda}{\sqrt{\tau - \lambda}}. \tag{9}$$

On a finite domain, condition (8) is replaced by the artificial boundary condition (9) and the problem solved using the Crank-Nicolson discretisation [7, p. 2090].

3.2 The Optimal Compact Approximation

The singularity at the strike price of the payoff function decreases the accuracy of the computed American option price. In the continuation region, both the European and

 ISBN: 978-84-614-6167-7

American option prices satisfy the Black-Scholes equation and this means that the singularity at the strike price can be removed by considering the difference between the American option price and the European price. This technique is known as the singularity separating method and was developed in [13]. The OCA method uses an optimal compact scheme for the heat equation [12, p.24].

Let u_A denote the transformed American call option price given by (6) and let u_E be the transformed European call price. Since both u_A and u_E satisfy the heat equation in (7), the difference $u_D = u_A - u_E$ also satisfies the same equation. Then u_D is the solution of

$$\begin{aligned} \frac{\partial u_D}{\partial \tau} &= \frac{\partial^2 u_D}{\partial x^2}, \quad -\infty < x < x_f(\tau), \quad 0 \le \tau \le \tau^* \\ u_D(x,\,0) &= 0, \quad -\infty < x < x_f(0), \\ u_D(x_f(\tau),\,\tau) &= g(x_f(\tau),\,\tau) - u_E(x_f(\tau),\,\tau), \quad 0 \le \tau \le \tau^* \\ u_D(x,\,\tau) &\to 0 \quad \text{as} \quad x \to -\infty. \end{aligned}$$

The problem is localized to a finite domain and the use of a non-uniform grid in the OCA algorithm allows choosing a fine grid on the part of the computational domain where the unknown free boundary is located and a coarse grid is chosen on the part extended to incorporate the far field boundary conditions. The algorithm also avoids the implementation of artificial boundary conditions as in the Han and Wu algorithm and instead uses the smooth pasting condition for locating the free boundary. For further details, we refer to [13].

3.3 The Operator Splitting Method

Ikonen and Toivanen [10] proposed an operator splitting technique for solving the linear complementarity problem. The Black-Scholes inequality (5) for the American problem is transformed to

$$\frac{\partial V}{\partial t} + \mathcal{L}_S V + \lambda = 0,$$

where $\lambda \ge 0$ is a penalty term.

The problem is localised on a finite domain $(0,\, S_{\max})$ and central difference approximations are used to discretise the Black-Scholes operator $\mathcal{L}_S V$ in (2) on a uniform mesh with spacing ΔS where $S_{\max} = N\Delta S$. Letting A denote the resulting discretisation matrix and $\zeta = T - t$, we obtain the semidiscrete equation

$$\frac{\partial V}{\partial \zeta} = AV + \lambda, \quad 0 \le t \le T.$$

Let $T = M\Delta\zeta$, $V^j = \left[V_1^j,\, V_2^j,\, \ldots,\, V_{N-1}^j\right]$ where $V_i^j = V\left(i\Delta S,\, j\Delta\zeta\right)$ and let λ^j be the value of λ at time level $j\Delta\zeta$. Then the Crank-Nicolson scheme for the solution of the LCP can be written in the form

$$\left(I - \frac{1}{2}\Delta\zeta A\right) V^{j+1} = \left(I + \frac{1}{2}\Delta\zeta A\right) V^j + \Delta\zeta\lambda^j.$$

 ISBN: 978-84-614-6167-7

In addition we have the constraints

$$\left[V_i^j - (S_i - K)\right] \cdot \lambda_i^j = 0, \quad V_i^j \geq S_i - K, \quad \lambda_i^j \geq 0.$$

The Operator Splitting method solves for the pair $\left(V^{j+1},\, \lambda^{j+1}\right)$ in two fractional steps. The first step solves for $\widehat{V}^{j+1}$ from

$$\left(I - \frac{1}{2}\Delta\zeta A\right)\widehat{V}^{j+1} = \left(I + \frac{1}{2}\Delta\zeta A\right)V^j + \Delta\zeta\lambda^j.$$

Then letting $V_i^0 = (S_i - K)^+$ and denoting by V^0 the vector with components V_i^0 for $i = 1, 2, \ldots, N-1$, the second step computes V^{j+1} and λ^{j+1} using the two-step formula

$$V^{j+1} = \max(V^0, \hat{V}^{j+1} + \Delta\tau\lambda^j),$$
$$\lambda^{j+1} = \lambda^j + \frac{1}{\Delta\tau}(\hat{V}^{j+1} - V^{j+1}).$$

4 Price Comparisons

We compare American option prices computed by different methods that have appeared in the literature. The prices computed by the three finite difference algorithms considered in this paper are compared against both analytical approximation methods and partial differential equations-based methods. All computations have been performed using Mathematica 7 on a Core i7 laptop with 8GB RAM and speed 3.20 GHz.

We first carry out a comparison against the Gauss-Laguerre (GL) method of Frontczak and Schobel [6] which is based on modified Mellin transforms and the lower and upper bound approximation (LUBA) of Broadie and Detemple [3]. Details on specific implementation issues including tuning parameters to obtain the computed results can be found in the original papers of the authors.

The example considered concerns the valuation of an American call option with a maturity of six months ($T = 0.5$). The other parameters are chosen as $K = 100$, $r = 0.03$, $\delta = 0.07$ and $\sigma = 0.2$. Computations for the Han-Wu (HW), operator splitting (OS) and optimal compact approximation (OCA) are performed using $m = 400$ in the space direction and $n = 400$ in the time direction. For the Han-Wu algorithm, the heat equation is localized to the domain $[x_{\min}, x_{\max}]$ where $x_{\min} = -1$ and $x_{\max} = 1$. For the operator splitting method, we have chosen $S_{\max} = 200$ and for the OCA method, we choose $S_{\min} = 10$ and $S_{\max} = 200$. We then construct a non-uniform grid on log price and these details can be found in the commented Mathematica codes.

The prices given by the different methods are shown in Table 1. It is observed that the prices computed by OCA and LUBA for the different asset prices are practically the same as the 'True' values given in the American option pricing literature. The HW and OS produce prices which are less accurate than OCA whereas the semi-analytical GL method is not as accurate as the other methods.

 ISBN: 978-84-614-6167-7

$(T,\ r,\ \delta,\ \sigma,\ K) = (0.5,\ 0.03,\ 0.07, 0.2,\ 100)$						
Asset Price(S)	True Value	GL	LUBA	HW	OS	OCA
80	0.2194	0.2185	0.2195	0.2193	0.2193	0.2194
90	1.3864	1.3851	1.3862	1.3858	1.3858	1.3864
100	4.7825	4.7835	4.7821	4.7816	4.7817	4.7825
110	11.0978	11.1120	11.0976	11.0969	11.0971	11.0977
120	20.0004	20.0000	20.0000	20.0005	20.0000	20.0005

Table 1: American call option prices.

Our second numerical example is designed to study the effect on accuracy on the computed prices by HW, OS and OCA when we vary the number of grid nodes. For this example, we choose the same values as in the first example for the parameters $x_{\max}$, $x_{\min}$ and $S_{\max}$. Comparing the computed prices when the spot price S equals the strike price $K = 100$, we see from Table 2 that the number of grid points has an influence on the accuracy of the HW and OS computed prices. OCA does not produce changes in computed prices when coarser grids are chosen. Table 3 gives further numerical evidence on this important property of OCA.

$(T,\ r,\ \delta,\ \sigma,\ K) = (0.5,\ 0.03,\ 0.07, 0.2,\ 100)$							
Asset	True	HW		OS		OCA	
Price(S)	Value	300×300	200×400	300×300	200×400	300×300	200×400
80	0.2194	0.2193	0.2193	0.2192	0.2193	0.2193	0.2193
90	1.3864	1.3854	1.3859	1.3853	1.3858	1.3863	1.3863
100	4.7825	4.7808	4.7815	4.7809	4.7818	4.7825	4.7825
110	11.0978	11.0963	11.0968	11.0966	11.0973	11.0976	11.0976
120	20.0004	20.0000	20.0005	20.0000	20.0000	20.0006	20.0005

Table 2: American call option prices for different grid size $m \times n$.

$(T,\ r,\ \delta,\ \sigma,\ K) = (0.5,\ 0.03,\ 0.07, 0.2,\ 100)$				
Asset	True	$m \times n$		
Price(S)	Value	200×200	200×100	150×300
80	0.2194	0.2193	0.2193	0.2193
90	1.3864	1.3862	1.3862	1.3862
100	4.7825	4.7823	4.7823	4.7822
110	11.0978	11.0975	11.0975	11.0974
120	20.0004	20.0008	20.0008	20.0010

Table 3: Accuracy of OCA computed American call option prices.

In our third numerical example, we compare option prices computed by HW, OS and OCA with those computed by the finite difference moving boundary method of Muthuraman (MBM) [8], the front fixing method (FF) of Wu and Kwok [14] and the

 ISBN: 978-84-614-6167-7

penalty method (PM) of Nielsen, Skavhaug and Tveito [9]. The problem we consider is a three year American put with $r = 0.08$, $\sigma = 0.2$ and $\delta = 0$. The results for MBM are for a grid size of $m = 500$ and $n = 100$, for both FF and PM the grid is of size $m = 1000$ and $n = 1000$ and for HW, OS and OCA we use $m = 600$ and $n = 600$ with $S_{\max} = 300$, $x_{\min} = -1.2$ and $x_{\max} = 1.2$.

$(T, r, \delta, \sigma, K) = (3, 0.08, 0, 0.2, 100)$							
Asset Price(S)	True Value	MBM	FF	PM	HW	OS	OCA
80	20.0000	20.0000	20.0000	20.0000	20.0000	20.0000	20.0000
90	11.6974	11.6889	11.9029	11.7207	11.6974	11.6972	11.6975
100	6.9320	6.9203	7.2527	6.9573	6.9320	6.9319	6.9321
110	4.1550	4.1427	4.4841	4.1760	4.1548	4.1548	4.1550
120	2.5102	2.4996	2.7760	2.5259	2.5101	2.5101	2.5102

Table 4: American put prices.

Clearly, we can see from Table 4 that HW, OS and OCA are highly accurate for American put options. Among these three methods, OCA is more accurate. For the selected grid size, the MBM, FF and PM all give option prices that are not even accurate to 2 decimal places with the 'True' values for $S > 80$. Moreover, it is well known that penalty methods generally requires the solution of non-linear systems of equations which can prove costly.

5 Conclusion

American options do not admit closed form solutions and numerical methods or semi-analytical methods are necessary for computing approximations to put and call options. This paper studied three finite difference algorithms for solving the American option pricing problem. Prices given by the three methods were compared against prices obtained by both analytical approximation and other finite difference methods. The results indicate that the HW, OS and OCA algorithms perform better in terms of accuracy in comparison to the other finite difference methods that we considered. Among these three methods, we demonstrated that OCA produces more accurate option prices. The merit of this scheme which uses a singularity separating framework lies in the choice of a non-uniform grid which can be chosen to be coarse on the far boundary and the results showed that coarser grid nodes does bring significant decrease in accuracy of the computed American option prices.

 ISBN: 978-84-614-6167-7

A Mathematica Codes for Call Options

We provide the Mathematica codes used to produce the numerical results for call options. Only simple modifications are necessary for producing the results for put options. Figure 1 gives the Mathematica code for the Han and Wu algorithm and Figure 2 gives the code for the operator splitting method. For the optimal compact approximation, the Mathematica codes for producing the tridiagonal matrix and the computation of the European price are given in Figure 3. Finally, Figure 4 lists the Mathematica code for pricing call options using OCA.

```
HanCall[T_, K_, S1_, σ_, r_, δ_, n_, m_, xmin_, xmax_] := Module[{err, k, fb},
(*Transformation*)
   tmax = 0.5 σ^2 T; k = 2 r / σ^2; kδ = 2 (r - δ) / σ^2; ρ = dt / dx^2;
   θ = Sqrt[ρ]; divisor = 4 θ + Sqrt[π] (1 + 2 θ^2);
   alph = 0.5 (kδ - 1); bet = 0.25 (kδ - 1)^2 + k;
(*Grid Specification*)
   dt = tmax / m; dx = Abs[xmax - xmin] / n; vett = Range[0, tmax, dt];
   vetx = Range[xmin, xmax, dx]; vets = K Exp[vetx];
(*Initial condition*)
   payoff = Map[Max[Exp[0.5 (kδ + 1) vetx[[#]]] - Exp[0.5 (kδ - 1) vetx[[#]]], 0] &,
     Range[n + 1]];
(*Calculation of Artificial Boundary Condition*)
   wold = payoff; ϕ = wold[[1]]; s = ConstantArray[0, n + 1];
   s[[1]] = 1 + ρ - ρ θ^2 Sqrt[π] / divisor;
   Allknown = θ wold[[1]] + Sqrt[π] wold[[1]] / 4;
(*Transform the tridiagonal system into a bidiagonal system and locate the
    free boundary condition*)
   news[i_] := Module[{}, s[[i]] = 1 + ρ - ρ^2 / (4 s[[i - 1]])];
   Map[news, Range[2, n + 1]]; hancall[j_] := Module[{},
     g = Exp[bet vett[[j]]] payoff;
     b = ConstantArray[0, n + 1]; y = ConstantArray[0, n + 1];
     w = ConstantArray[0, n + 1];
    b[[1]] = 0.5 ρ (wold[[1]] + wold[[3]]) + (1 - ρ) wold[[2]] + 2 ρ Allknown / divisor;
    y[[1]] = b[[1]]; w[[2]] = (y[[1]] + 0.5 ρ g[[3]]) / s[[1]]; i = 1; wold1 = w[[2]];
     gold1 = g[[2]];
     byw[i_] := Module[{}, b[[i]] = 0.5 ρ (wold[[i]] + wold[[i + 2]])
          + (1 - ρ) wold[[i + 1]];
    y[[i]] = b[[i]] + 0.5 ρ y[[i - 1]] / s[[i - 1]];
    w[[i + 1]] = (y[[i]] + 0.5 ρ g[[i + 2]]) / s[[i]]; z = i;] /; (w[[i]] >= g[[i]]);
    Map[byw, Range[2, n]]; i = z;
(*Set call option values above the free boundary condition as payoff*)
     w[[i + 1 ;; n + 1]] = g[[i + 1 ;; n + 1]];
(*Compute remaining option values*)
     neww[i_] := Module[{}, w[[i + 1]] = (y[[i]] + 0.5 ρ w[[i + 2]]) / s[[i]]];
     Map[neww, Range[i - 1, 1, -1]];
     w[[1]] = 4 (Allknown + 0.5 θ^2 Sqrt[π] w[[2]]) / divisor;
     ϕold = {ϕ, w[[1]]}; ϕ = ϕold // Flatten; Allknown =
      θ (wold[[1]] - Total[(ϕ[[2 ;; j]] - ϕ[[1 ;; j - 1]]) / (Sqrt[j - Range[1, j - 1]]
               + Sqrt[j - Range[2, j]])]) + Sqrt[π] wold[[1]] / 4;
     v = K Exp[-alph vetx - bet vett[[j]]] w; wold = w];
   Map[hancall[#] &, Range[2, m + 1]];
   (*Interpolate the values of S1*)
     Interpolation[Map[{vets[[#]], v[[#]]} &, Range[Length[vets]]]][S1]];
```

Figure 1: Han and Wu Mathematica code for pricing call options.

 ISBN: 978-84-614-6167-7

```
OpSpAmerCall[T_: 0.5, K_: 100, S1_: 80, σ_: 0.2, r_: 0.03, δ_: 0.07,
   n_: 400, m_: 400, smin_: 0, smax_: 200, θ_: 0.5] := Module[{x, t, α, β, γ},
(*Discretising the domain*)
   dx = (smax - smin) / n; dt = T / m;
   x[i_] := smin + i dx; X = Map[x, Range[0, n]]; t[j_] := j dt;
(*Initial condition for a call option*)
   U0 = Map[Max[x[#] - K, 0] &, Range[0, n]];
(*Implementing the left and right boundary conditions*)
   ul = 0; ur = smax - K;
(*Transformations*)
   α = 0.5 σ^2 X^2 / dx^2 - (r - δ) X / (2 dx); β = -σ^2 X^2 / (dx^2) - r;
   γ = 0.5 σ^2 X^2 / (dx^2) + (r - δ) X / (2 dx);
(*Use vectorization to create the coeffcient matrix*)
    λ = ConstantArray[0, n - 1]; iden = SparseArray[{Band[{1, 1}] -> 1}, {n - 1, n - 1}];
    a1 = Drop[Drop[α, 2], -1]; a2 = Drop[Drop[β, 1], -1]; a3 = Drop[Drop[γ, 1], -2];

   A1 = SparseArray[{Band[{2, 1}] -> a1, Band[{1, 1}] -> a2, Band[{1, 2}] -> a3}, {n - 1, n - 1}];
    B = (1 / dt) iden + (1 - θ) A1; A = (-1 / dt) iden + θ A1; uold = U0;
(*Implementation of the Operator splitting method with Crank-Nicolson*)
   OSCNAC[j_] := Module[{}, rhs = -(B.uold[[2 ;; n]] - λ);
     rhs[[1]] = rhs[[1]] - θ α[[2]] ul - (1 - θ) α[[2]] ul;
    rhs[[n - 1]] = rhs[[n - 1]] - θ γ[[n]] ur - (1 - θ) γ[[n]] ur;
    uprime = LinearSolve[A, rhs] // Chop; u = ConstantArray[0, n];
    u[[2 ;; n]] = Map[ Max[uprime[[# - 1]] + dt λ[[# - 1]], U0[[#]]] &, Range[2, n]];
    λ1 = λ + (1 / dt) (uprime - u[[2 ;; n]]); λ = λ1; uold[[2 ;; n]] = u[[2 ;; n]]; j + 1];
(*Evaluation of the option price vectors*)
   Nest[OSCNAC[#] &, 1, m];
   (*Evaluation of the option price when the stock price=S1*)
   Interpolation[
     Map[{X[[#]], Flatten[{ul, u[[2 ;; n]], ur}][[#]]} &, Range[Length[X]]]][S1]];
```

Figure 2: Operator Splitting Mathematica code for call options.

```
In[2]:= (*Define a function FunCoeff to obtain the
     tridiagonal linear system with optimal discretization*)
   FuncCoeff[X_, dt_, n_] := Module[{}, xL1 = X[[2 ;; n]] - X[[1 ;; n - 1]];
   xC1 = X[[3 ;; n + 1]] - X[[1 ;; n - 1]]; xR1 = X[[3 ;; n + 1]] - X[[2 ;; n]];
   bj1 = (1 / 3) (xL1^2 + xL1 xR1 + xR1^2); Dα1 = 1 / 3 - 2 bj1 / (3 xC1 xL1);
   Dxxα1 = 2 / (xC1 xL1); Dβ1 = 1 / 3 - (2 bj1 / (3 xC1)) (-1 / xR1 - 1 / xL1);
   Dxxβ1 = (2 / xC1) (-1 / xR1 - 1 / xL1); Dγ1 = 1 / 3 - 2 bj1 / (3 xC1 xR1);
   Dxxγ1 = 2 / (xC1 xR1); α1 = Dα1 / dt - (1 / 2 - bj1 / (12 dt)) Dxxα1;
   β1 = Dβ1 / dt - (1 / 2 - bj1 / (12 dt)) Dxxβ1; γ1 = Dγ1 / dt - (1 / 2 - bj1 / (12 dt)) Dxxγ1;
   αR1 = Dα1 / dt + (1 / 2 + bj1 / (12 dt)) Dxxα1; βR1 = Dβ1 / dt + (1 / 2 + bj1 / (12 dt)) Dxxβ1;
   γR1 = Dγ1 / dt + (1 / 2 + bj1 / (12 dt)) Dxxγ1; {α1, β1, γ1, αR1, βR1, γR1}];
   (*Define a function bscallsmith to obtain Black-Scholes analytical solution*)
   bscallsmith[S_, T_, r_, δ_, σ_, K_] :=
     Module[{d1, d2}, d1 = (Log[(S + $MachineEpsilon) / K] + (r - δ + 0.5 σ^2) T) /
         (σ Sqrt[T + $MachineEpsilon]);
   d2 = d1 - σ Sqrt[T]; N1 = (1 - 0.5 Erfc[d1 / Sqrt[2]]); N2 = (1 - 0.5 Erfc[d2 / Sqrt[2]]);
   S N1 Exp[-δ T] - K Exp[-r T] N2];
```

Figure 3: Mathematica functions for constructing OCA tridiagonal matrix and computation of European option price.

 ISBN: 978-84-614-6167-7

```
OHCCALL[S1_, m_, n_, T_, σ_, r_, δ_, smax_, K_] := Module[{err, smin, k, fb, α, β, γ},
(*Transformations*)
   tmax = 0.5 σ^2 T; k = 2 r / σ^2; kδ = 2 (r - δ) / σ^2;
   alph = 0.5 (kδ - 1); bet = 0.25 (kδ - 1)^2 + k; smin = 10; ds = (smax - smin) / n;
(*Construction of the non-uniform grid and the computation of the non-uniform
     log transformed x-coordinate*)
   vets = Range[smin, smax, ds];
   vetx = Log[vets / K]; dt = tmax / m; vett = Range[0, tmax, dt];
   wold = ConstantArray[0, n + 1];
(*Use Gauss Elimination to transform the tridiagonal system to a bidiagonal one*)
   {α, β, γ, αR, βR, γR} = FuncCoeff[vetx, dt, n];
    c = ConstantArray[0, n - 1];
   newcβ[i_] := Module[{}, c[[i]] = α[[i]] / β[[i - 1]]; newβ = β[[i]] - c[[i]] γ[[i - 1]];
     β[[i]] = newβ;]; Map[newcβ, Range[2, n - 1]]; fb = ConstantArray[0, m + 1]; fb[[1]] = 1;
   v = ConstantArray[0, n + 1]; w = ConstantArray[0, n + 1];
   b = ConstantArray[0, n - 1]; y = ConstantArray[0, n - 1];
   (*Implementation of the optimal high order compact scheme*)
   Yan[j_] := Module[{},
     (*At each time step,we compute the difference between the transformed payoff and
      the transformed European option*)
     HEurExact = (1 / K) Exp[alph vetx + bet vett[[j]]]
       bscallsmith[vets, vett[[j]] / (0.5 σ^2), r, δ, σ, K]; g1 = Exp[bet vett[[j]]]
       Map[Max[Exp[(alph + 1) vetx[[#]]] - Exp[alph vetx[[#]]], 0] &, Range[n + 1]];
     g = g1 - HEurExact;
     (*locate the free-boundary,compute the Delta for two more nodes and
      extrapolate to obtain an accurate free boundary location*)
     b[[1]] = αR[[1]] wold[[1]] + βR[[1]] wold[[2]] + γR[[1]] wold[[3]]; y[[1]] = b[[1]];
     by[i_] :=
      Module[{}, b[[i]] = αR[[i]] wold[[i]] + βR[[i]] wold[[i + 1]] + γR[[i]] wold[[i + 2]];
       y[[i]] = b[[i]] - c[[i]] y[[i - 1]]]; Map[by, Range[2, n - 1]];
     w[[n]] = (y[[n - 1]] - γ[[n - 1]] g[[n + 1]]) / β[[n - 1]]; i = n; err = g[[i]] - w[[i]];
neww[err2_] := Module[{}, z1 = i - 1; i = z1; w[[i]] =
        (y[[i - 1]] - γ[[i - 1]] g[[i + 1]]) / β[[i - 1]]; g[[i]] - w[[i]]];
NestWhile[neww, err, # > 0 &]; i = i; w[[i + 1]] = g[[i + 1]];
     wk[k_] := Module[{}, w[[k]] = (y[[k - 1]] - γ[[k - 1]] w[[k + 1]]) / β[[k - 1]]];
     Map[wk[#] &, Range[i - 1, i - 2, -1]];
     (*Grid Remanipulation*)
     interpoS = vets[[i - 1 ;; i]];
     val = K Exp[-alph * vetx[[i - 2 ;; i + 1]] - bet * vett[[j]]] *
       (w[[i - 2 ;; i + 1]] + HEurExact[[i - 2 ;; i + 1]]);
     interpoV = (val[[3 ;; 4]] - val[[1 ;; 2]]) / (2 (vets[[i]] - vets[[i - 1]]));
     Off[InterpolatingFunction::dmval];
fb[[j]] = Interpolation[Map[{interpoV[[#]], interpoS[[#]]} &, Range[Length[interpoS]]],
       InterpolationOrder → 1][1];
xf = Log[fb[[j]] / K]; sf = fb[[j]]; Select[vetx, # < xf &];
     i = Length[Select[vetx, # < xf &]];
     (*reconstruct the optimal compact linear equation based on the new grid
      node xf at the critical asset price*)
     {α11, β11, γ11, αR11, βR11, γR11} =
      FuncCoeff[Append[vetx[[i - 1 ;; i]], xf], dt, 2] // Flatten;
     wold[[i + 1]] = Exp[bet vett[[j - 1]]] Max[Exp[(alph + 1) xf] - Exp[alph xf], 0] -
       (1 / K) Exp[alph xf + bet vett[[j - 1]]]
        bscallsmith[sf, vett[[j - 1]] / (0.5 σ^2), r, δ, σ, K]; w[[i + 1]] = Exp[bet vett[[j]]]
        Max[Exp[(alph + 1) xf] - Exp[alph xf], 0] - (1 / K) Exp[alph xf + bet vett[[j]]]
        bscallsmith[sf, vett[[j]] / (0.5 σ^2), r, δ, σ, K];
     b[[i - 1]] = αR11 wold[[i - 1]] + βR11 wold[[i]] + γR11 wold[[i + 1]];
     cc = α11 / β[[i - 2]]; β2 = β11 - cc γ[[i - 2]]; β11 = β2;
     b1 = b[[i - 1]] - cc y[[i - 2]]; b[[i - 1 ]] = b1; y[[i - 1]] = b[[i - 1]];
     w[[i]] = (y[[i - 1]] - γ11 w[[i + 1]]) / β11; w[[i + 1 ;; n + 1]] = g[[i + 1 ;; n + 1]];
     wnew[i_] := Module[{}, z = i; w[[i + 1]] = (y[[i]] - γ[[i]] w[[i + 2]]) / β[[i]]];
     Map[wnew, Range[i - 2, 1, -1]]; v = K Exp[-alph vetx - bet vett[[j]]] (w + HEurExact);
    wold = w;]; Map[Yan, Range[2, m + 1]];
   Interpolation[Map[{vets[[#]], v[[#]]} &, Range[Length[vets]]]][S1]];
```

Figure 4: OCA Mathematica code for call options.

 ISBN: 978-84-614-6167-7

References

[1] T. Björk, *Arbitrage Theory in Continuous Time*, Oxford University Press, 2004.

[2] F. Black and M. S. Scholes, *The pricing of options and corporate liabilities*, Journal of Political Economy **81** (1973) 637–659.

[3] M. Broadie and J. Detemple, *American option valuation: approximations and a comparison of existing methods*, Rev. Finance Stud. **9** (1996) 1211–1250.

[4] C. W. Cryer, *The solution of a quadratic programming problem using systematic overrelaxation*, SIAM Journal on Control and Optimization **9** (1971) 385–392.

[5] P. A. Forsyth and K. R. Vetzal, *Quadratic convergence for valuing American options using a penalty method*, SIAM J. Sci. Comput. **23** (2002) 2095–2122.

[6] R. Frontczak and R. Schöbel, *On modified Mellin transforms, Gauss-Laguerre quadrature, and the valuation of American call options*, Journal of Computational and Applied Mathematics **234** (2010) 1559– 1571.

[7] H. Han and X. Wu, *A fast numerical method for the Black–Scholes equation of American options*, SIAM J. Numer. Anal. **41** (2003) 2081–2095.

[8] K. Muthuraman, *A moving boundary approach to American option pricing*, Journal of Economic Dynamics and Control **32** (2008) 3520–3537.

[9] B. F. Nielsen, O. Skavhaug and A.Tveito, *Penalty and front–fixing methods for the numerical solution of American option problems*, The Journal of Computational Finance **5** (2002) 69–97.

[10] S. Ikonen and J. Toivanen, *Operator splitting method for American option pricing*, Applied Mathematics Letters **17** (2004) 809–814.

[11] U. R. Seydel, *Tools for Computational Finance*, Springer-Verlag, Heidelberg, 2006.

[12] R. Smith, *Optimal and near-optimal advection-diffusion finite-difference schemes Part 2: Unsteadiness and non-uniform grid*, Proc. Roy. Soc. Lond. **456** (2000) 489–502.

[13] D. Y. Tangman, A. Gopaul, and M. Bhuruth, *A fast high–order finite difference algorithm for pricing American options*, Journal of Computational and Applied Mathematics **222** (2007) 17–29.

[14] L. Wu and Y. Kwok, *A front–fixing finite difference method for the valuation of American options*, Journal of Financial Engineering **6** (1997) 83–97.

 ISBN: 978-84-614-6167-7

Proceedings of the 11th International Conference on Computational and Mathematical Methods in Science and Engineering, CMMSE 2011 26–30 June 2011.

Performance evaluation of using Multi-core and GPU to remove noise in images

María. G. Sánchez[1], Vicente Vidal[2], Jordi Bataller[2], Josep Arnal[3] and Juan Seguí[4]

[1] *Departamento de Sistemas y Computación, Instituto Tecnológico de Cd. Guzmán, 49100 Cd. Guzmán, Jal. México*

[2] *Departamento de Sistemas Informáticos y Computación, E.P.S. Gandia, Universidad Politécnica de Valencia, 46730, Grao de Gandia Valencia, Spain*

[3] *Departamento de Ciencia de la Computación e Inteligencia Artificial, Universidad de Alicante, 03071, Alicante, Spain*

[4] *Instituto de Educación Secundaria la Valldigna, Tavernes de la Valldigna, 46760, Valencia, Spain*

emails: msanchez@dsic.upv.es, vvidal@dsic.upv.es, bataller@dsic.upv.es, arnal@dccia.ua.es, juasede@doctor.upv.es

Abstract

In this study, we conducted the parallelization of a digital image to remove impulsive noise with multi-core interface using Open Multi-Processing (OpenMP) and the Graphics Processing Unit (GPU) programming model using Compute Unified Device Architecture (CUDA). Many sequential algorithms to remove noise in digital images, have either an excessive computational cost or are too large when the purpose is real-time processing. We did an analysis of performance using large images in order to identify the amount of pixels to be allocate in Multi-core and GPUs, so that both have work to do. Performance was evaluated in terms of execution time and we did a comparison of the implementation parallelized in multi-core, GPUs and a combination of both. The observed time shows that both devices must have tasks to do, leaving the most of the work to the GPU.

Key words: noise removal, peer group filter, parallel algorithm, graphical processing unit (GPU), open multi-processing (OpenMP).

1 Introduction

Image denoising is still an open problem in the field of image processing, because damaged images can affect the performance and accuracy of some processes. Also, images can be very large and/or require a real-time processing.

 ISBN: 978-84-614-6167-7

Noise removal and the correct perception of a desired color are of paramount importance in emerging applications related to biomedical science, earth science, cultural heritage preservation, video communications, image postprocessing, robotic inspection and surveillance. Impulsive noise is commonly found, caused by the malfunction of sensors and other hardware during the process of image formation, storage or transmission, [12]. This type of noise affects some individual pixels, changing its original value. The most usual model of impulsive noise is the Salt and Pepper noise (or fixed value noise), which considers that the new, wrong, pixel value is an extreme value within the signal range. This is the noise type we consider in this paper.

Many algorithms to reduce impulsive noise in images (known as filters) have been introduced in other papers. Among many others we can cite [1], [2], [3], [4], [5], [6], [7], [8], [9], [10], [11], [12]. Those filters are based on the concept of peer group. A peer group is a set of neighboring pixels of a central one, x_i, beeing similar to it according to an appropriate metric value (this is, the nearest neighbours) [12] [9]. This type of filters have recently shown good results in quality but they do not seem appropiate for real-time processing [1], [8], [9], [10], [12], [13].

In this paper, we introduce a parallel version of peer group based filters in order to keep their good quality results while trying to improve its performance, so to make them usable in real-time processing. We have tested this parallel algorithms developing programs for Graphical Processing Units (GPU) and multi-cores and we did an analysis of the best distribution of pixels in these two devices to take advantage of the hardware.

Graphical processing units are currently a very popular platform for developing parallel applications, considering their availability, price and speed. Therefore, we judge very convenient to develop parallel filter implementations for GPUs. Our implementations are programmed in C, using the CUDA library (Compute Unified Device Architecture), [15].

The paper is organized as follows: Section 2 explains parallel version of the algorithm. Section 3 discusses how the parallel algorithm was implemented for the GPUs and Multi-core. The results of the experimental study are shown in Section 4 and finally Section 5 concludes the paper.

2 Parallel denoising algorithm

Now we will describe our parallel denoising algorithm. Our algorithm uses the *peer group* of a central pixel x_i in a window W according to [12] but using a fuzzy metric instead. The fuzzy distance between the vectors x_i and x_j of the color image is given by the following function:

$$M(x_i, x_j) = \prod_{l=1}^{3} \frac{min\{x_i(l), x_j(l)\} + k}{max\{x_i(l), x_j(l)\} + k}. \quad (1)$$

where (x_i (1), x_i (2), x_i (3)) is the color vector for the pixel x_i in RGB and x_j are the neighbor pixel of x_i. In [6], it was shown that $k = 1024$ is an appropriate setting to

 ISBN: 978-84-614-6167-7

maintain the image quality, and this was therefore the value that we use in the present study. The function to get the peer group is,

$$\mathcal{P}(x_i, d) = \{x_j \in W : M(x_i, x_j) \geq d\} \tag{2}$$

where $0 <= d <= 1$ is the distance threshold. The peer group associated with the central pixel of W is a set formed by the central pixel x_i and its neighbors belonging to W, whose distance from x_i exceeds d.

The algorithm performs two main steps: in the first one (*detection*) the pixels are labeled as either *corrupted* or *uncorrupted*. In the second step (*filtering*) corrupted pixels are corrected. Therefore, the detection and filtering step are described for a single pixel x_i:

- Detection: x_i is declared as corrupted if $\#\mathcal{P}(x_i, d) < (m + 1)$, where m is the voting threshold and $\#A$ the cardinality of set A.
- Filtering: given a pixel x_i previosly marked as corrupted, we replace it by the arithmetic mean of its neighbor pixels in its window W: the well-known arithmetic mean filter (AMF) [1], [2]. This is, the new value for x_{ik} (color component k of x_i) is $\frac{\sum x_{jk}}{(\#W-1)}$ for all $x_j \in W$ with $j \neq i$.

We decomposed the algorithm into two phases –detection and filtering– not only to follow the separation of concerns principle, but also because we use the AMF to replace corrupted pixels. Given that AMF considers only uncorrupted pixels for the mean computation, the filter phase cannot start until the detection phase is done. In consequence, to ensure this synchronization requirement, in our parallel implementation on GPUs we have developed two kernels –a detecting kernel and a filtering kernel– so that the filtering kernel is not launched until the detecting kernel has finished and in multi-core, are two separate functions.

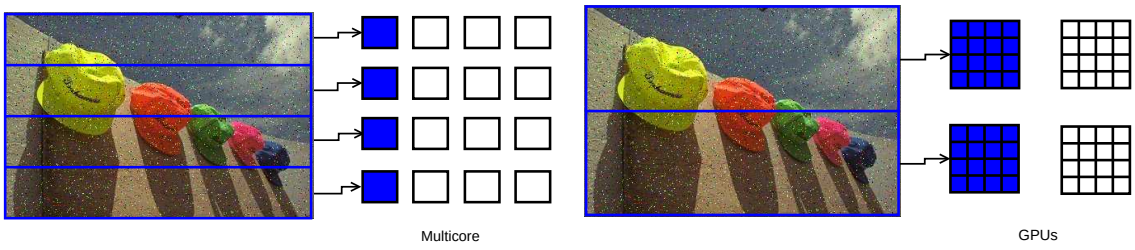
In the implementation on GPU, we use texture to access data from the GPU.

3 Comments on the GPU and multi-core implementation

In the previous section we described the detection and filtering steps of the algorithm. We have done three implementations. The first was in multi-core using OpenMP and the second with CUDA on GPUs. Figure 1 shows an example of these two implementations. The third is a combination of multi-core and GPUs. Figure 2 shows the distribution of an image using the cores and GPUs.

For the distrubution of the pixels of an image in the first implementation, we divided the number of pixels by the number of cores (less than or equal to those available). In the second implementation, the image was divided by a number less than or equal to the GPUs available. In the third implementation, the pixels are distributed in multi-core and GPU running on each of the cores and GPUs specified. The flowchart, Figure 3, shows the elimination of noise with these three implementations.

 ISBN: 978-84-614-6167-7



a b

Figure 1: a)Distributed image in 4 cores b) Distributed image on 2 GPU

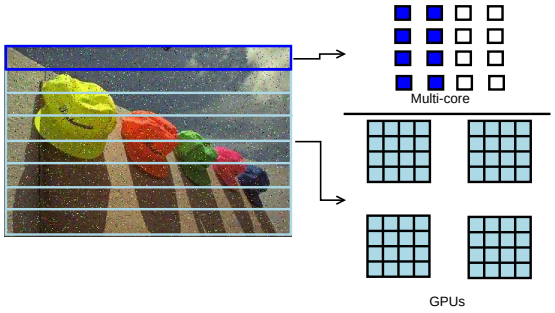


Figure 2: Distributed image on 4 GPUs and 8 cores.

Table 1: Parallelizing image with multi-core. Time in seconds

	core				
image size	1	2	4	8	16
96x64	0.0037	0.0020	0.0012	0.0009	0.0010
192x128	0.0148	0.0076	0.0040	0.0024	0.0024
384x256	0.0601	0.0306	0.0156	0.0087	0.0083
768x512	0.2381	0.1185	0.0604	0.0318	0.0274
1536x1024	0.9367	0.4687	0.2348	0.1256	0.0986
3072x2048	3.7023	1.8814	0.9554	0.5049	0.3817
6144x4096	14.7825	7.4480	3.7615	2.0435	1.4790

4 Experimental Study

This section presents the results of practical experiments conducted on a Mac OS X (Intel Quad-Core Xeon 2 x 2.26 GHz, 8GB of RAM) with four NVIDIA GPU (GeForce GT 120, 512MB of memory) and CUDA toolkit (version 1.1, gcc 4.0). We used the Caps image (Figure 4) from a Kodak image database ([14]).

The first test that we did was to distribute the image only in cores. Table 1 shows the results. As we can see, the shortest time is presented when the image is divided among the 16 cores available, except with small-sized images (96x64 pixels) which have an optimal allocation with 8 cores.

The next test consists of distribute a image among the available GPUs. As shown

 ISBN: 978-84-614-6167-7

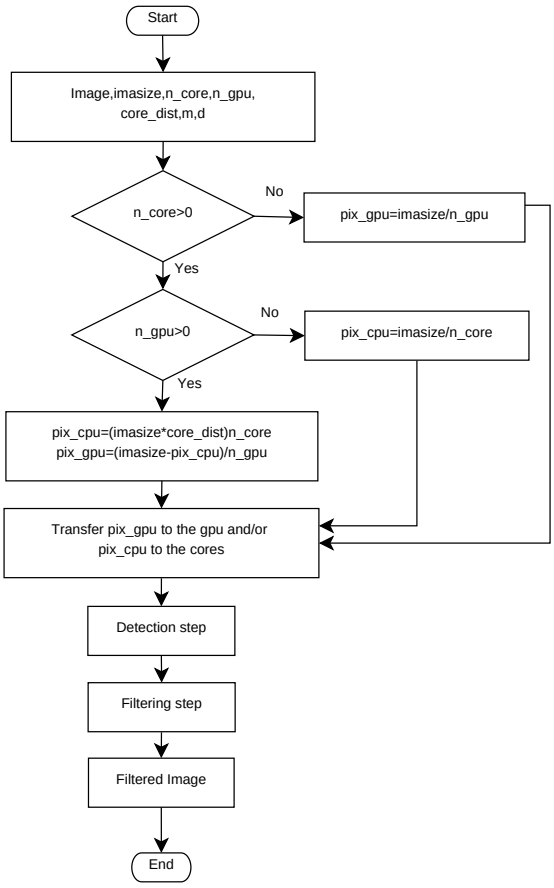


Figure 3: Flowchart describing the algorithm for remove noise image.

in Figure 5, to parallelizing an image smaller than 1536x1024 it works better with only one GPU. The times obtained in one GPU are similar to the ones got by the parallelizing through 16 cores, but if the size of the image increases, then the time difference increases. Distributing a image smaller than 3072x2048 into 4 GPUs smaller , leads to a greater time than dividing the distribution on two GPUs. With this, we see that with a large image, the optimal distribution is obtained by using four GPUs, but not for small images (smaller than 3072x1048).

In the last test conducted, we have distributed the image in both CPU and GPU. Figure 5 shows the results obtained when the image is distributed into 4 GPUs and in all the cores, for different image sizes.

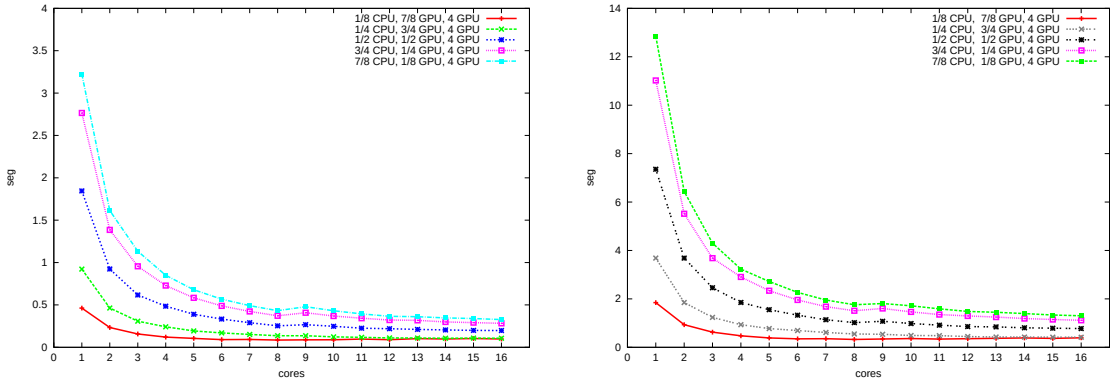
In both cases the behavior is similar. The best time is shown when more load is assigned to the GPU leaving less on the cores. Table 3 shows the performance of time using the available cores and 4 GPUs, and making the comparison using only the 4 GPUs without cores, for a distribution of 1/8 of the image in cores and 7/8 in GPU. As can be seen in the results, the parallelization performed using the harwdare of cores

 ISBN: 978-84-614-6167-7

Figure 4: Caps Image

Table 2: Parallelizing image with GPU. Time in seconds

	GPU		
image size	1	2	4
96x64	0.0009	0.0357	0.1058
192x128	0.0022	0.0363	0.1061
384x256	0.0071	0.0391	0.1074
768x512	0.0267	0.0483	0.1117
1536x1024	0.1047	0.0858	0.1306
3072x2048	0.3415	0.2357	0.2019
6144x4096	1.0681	0.8144	0.4734



a b

Figure 5: a)Caps 3072x2048 b) Caps 6144x4096

and GPUs gives better results than the parallelization performed only in 4 GPUs.

5 Concluding remarks

The availability of inexpensive parallel processing hardware provided by graphical processing units is a clear reason to develop and test programs to solve problems that could benefit from it. The image denoising is a problem that fits well in a parallel scenario because images may be large, the processing is costly, and the image pixels, to an extent, can be handled simultaneously.

Therefore, in this paper we have adapted a denoising algorithm based on the peer

 ISBN: 978-84-614-6167-7

Table 3: Time in seconds for the image parallelized 1/8 in multicore and 7/8 in 4 GPUs.

core	3072x2048	6144x4096
0	0.4734	0.2019
1	0.4625	1.8448
2	0.2338	0.9331
3	0.1568	0.6251
4	0.1201	0.4747
5	0.1053	0.3877
6	0.0911	0.3479
7	0.0926	0.3560
8	0.0851	0.3241
9	0.1942	0.3423
10	0.0888	0.3605
11	0.0973	0.3396
12	0.0900	0.3566
13	0.1023	0.3726
14	0.0979	0.3877
15	0.1035	0.3692
16	0.0980	0.3916

group concept that uses a fuzzy metric with a parallel setting option. We have implemented it to be run on a GPU using the CUDA library and using OpenMP in multi-core.

We conducted experiments to adjust our algorithm, to compare it with the sequential version of the same algorithm, and to compare the quality of the resulting images with other algorithms. We conclude that parallel implemetations of denoising filters on GPUs and multi-core are very advisable, and they open the door to use such algorithms for real-time processing.

For future works, we plan to implement this parallel algorithm to detect edges in an image.

Acknowledgements
This work was funded by the Spanish Ministry of Science and Innovation (Project TIN2008-06570-C04-04) and Ma. Gpe. would also like to acknowledge DGEST ITCG for the scholarship awarded through the PROMEP program (México).

References

[1] J. G. Camarena, V. Gregori, S. Morillas, A.Sapena, *Fast detection and removal of impulsive noise using peer group and fuzzy metrics*, Journal of Visual Communication and Image Representation, 19 (2008) 20-29.

[2] A. Toprak, I. Guler, *Impulse noise reduction in medical images with the use of switch mode fuzzy adaptive median filter*, Fuzzy sets and systems 2006.

[3] S. Schulte., M. Nachtegael, V. De Witte, D. Van der Weken,and E.E. Kerre, *A Fuzzy Impulse Noise Detection and Reduction Method*, IEEE Transaction on Image Processing, Vol 15 No 5, May 2006.

 ISBN: 978-84-614-6167-7

[4] S. Schulte, S. Morillas , V. Gregori, and E.E. Kerre, *A New Fuzzy Color Correlated Impulse Noise Reduction Method*, IEEE Transaction on Image Processing, Vol 15 No 10, October 2007.

[5] S. Schulte, V. De Witte, M. Nachtegael, D. Van der Weken,and E.E. Kerre, *Fuzzy Two-Step Filter for Impulse Noise Reduction From Color Images*, IEEE Transaction on Image Processing, Vol 15 No 11, November 2006.

[6] S. Schulte., V. De Witte, M. Nachtegael, D. Van der Weken,and E.E. Kerre, *Fuzzy random impulse noise reduction method*, Fuzzy sets and systems 2007.

[7] T. Mlange, M. Nachtegael,E. E. Kerre, *Fuzzy Random Impulse Noise Removal From Colour Image Sequences*, IEEE, September 2010.

[8] S. Morillas, V. Gregori, and A. Hervs, *Fuzzy Peer Groups for Reducing Mixed Gaussian-Impulse Noise From Color Images*, IEEE Transaction on Image Processing, Vol 18 No 7, November 2009.

[9] J.G. Camarena, V. Gregori, S. Morillas and A. Sapena, *Some improvements for image filtering using peer group techniques*, Image Vis. Comput., vol. 28, no. 1, pp. 188-201, 2010.

[10] S. Morillas, V. Gregori and G. Peris-Fajarns, *Isolating impulsive noise pixels in color images by peer group techniques*, Comput. Vis. Image Underst., vol. 110, no. 1, pp. 102-116, 2008.

[11] J. G. Camarena, V. Gregori, S. Morillas , A. Sapena, *Two-step fuzzy logic-based method for impulse noise detection in colour images*, Pattern Recognition Letters 31 (2010) 1842-1849.

[12] B. Smolka, *Peer group switching filter for impulse noise reduction in color images*, Pattern Recognition Letters 31 (2010) 484-495

[13] B. Smolka, *Fast detection and impulsive noise remolval in color images*, Real-Time Imaging 11 (2005) 389-402

[14] *http://r0k.us/graphics/kodak/index.html*

[15] *http://www.nvidia.es/page/home.html*

 ISBN: 978-84-614-6167-7

Proceedings of the 11th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2011
26–30 June 2011.

Stability and stabilizability of variational discrete systems

Adina Luminiţa Sasu[1] and Bogdan Sasu[2]

[1] *Department of Mathematics, Faculty of Mathematics and Computer Science, West University of Timişoara*

[2] *Department of Mathematics, Faculty of Mathematics and Computer Science, West University of Timişoara*

emails: sasu@math.uvt.ro, bsasu@math.uvt.ro

Abstract

In this paper we propose a new study concerning the stability of variational discrete systems. We analyze the connections between the exponential stability of a variational discrete system and the stabilizability of an associated control system. Using constructive methods and computational arguments, we deduce that the stabilizability and the ℓ^p input-output stability of the associated control system are necessary and sufficient conditions for the exponential stability of the initial variational system.

Key words: variational discrete system, cocycle, stability, control system, stabilizability
MSC 2000: 93C55; 93C73; 93C25;

1 Introduction

In the last decades, a significant progress was made in the study of the control problems for dynamical systems arising in engineering, fluid mechanics, population dynamics, biology, aerospace and the investigation methods were extended from functional methods and spectral approaches to computational analysis and discrete-time techniques (see [1]–[19] and the references therein). A leading topic was the stabilization problem for control systems, which was treated from various perspectives for a large variety of systems: autonomous, non-autonomous, variational, discrete, with distributed parameters and so forth (see [2], [5], [7]–[16], [18] and the references therein). Moreover, it has been shown that (complete) stabilizability is interrelated with the exact controllability (see [5], [7], [13], [16], [18]). The analysis of the connections between the asymptotic properties of dynamical systems and the stable properties of some associated control systems provided important information with applications in robustness problems (see

 ISBN: 978-84-614-6167-7

[2], [15], [17], [19]). In [2], employing a new technique, relying in the properties of the evolutions semigroup on function spaces, the authors pointed out that for differential systems described by evolution families the exponential stability may be expressed in terms of the stabilizability and detectability of the associated control system. Then, it is natural to study whether this phenomenon occurs also for variational systems and to investigate which is the most general structure of the system, such that these techniques can be implemented. A first attempt in this sense was obtained in [13] for skew-product semiflows, working with integral control systems. The second main step was made in [15] where we have considered the case of difference equations and we have deduced their behavior using input-output techniques in terms of abstract Banach sequence spaces. It worth mentioning that the control properties of difference equations has recently begun to be understood in infinite dimensional spaces (see [3], [6], [16], [18] and the references therein) and therefore their study is of great interest in identifying and clarifying natural phenomena which are modeled in the discrete-time setting. An interesting case in this framework is represented by the variational discrete systems (see [19]). On the one hand this generalizes the case of difference equations and on the other hand leads to a dynamic variational modeling. In the stability theory this aspect is very useful because a variational system is exponentially stable if and only if its discrete-time counterpart is exponentially stable (see [19] and the references therein).

The aim of this paper is to treat a new case of the above described problem and to obtain several interesting connections between stabilizability and exponential stability in the general case of variational discrete systems in infinite-dimensional spaces. Mainly, we consider the variational discrete system

$$(A) \qquad x(\theta)(n+1) = A(\sigma(\theta,n))x(\theta)(n), \quad \theta \in \Theta, n \in \mathbb{N}$$

where σ is a discrete flow on a metric space Θ (see Definition 2.1 below) and $\{A(\theta)\}$ is a family of bounded linear operators on a Banach space X. We associate the discrete control system

$$(A,B) \qquad x(\theta)(n+1) = A(\sigma(\theta,n))x(\theta)(n) + B(\sigma(\theta,n))u(n), \quad \theta \in \Theta, n \in \mathbb{N}$$

where U is a Banach space and $\{B(\theta)\} \subset \mathcal{B}(U,X)$. In what follows we will analyze for the first time the connections between the exponential stability of the system (A) and the stabilizability of the control system (A,B). Our starting point is a recent characterization of the uniform exponential stability of variational discrete systems (see [19]). Throughout this paper we develop a constructive investigation relying on input-output techniques, using direct and computational arguments. Our study will show that the exponential stability of the system (A) is equivalent with two properties of the system (A,B): stabilizability and an ℓ^p input-output stability property.

2 Stability of variational discrete system-preliminaries

Let X be a real or complex Banach space and let $\mathcal{L}(X)$ be the Banach algebra of all bounded linear operators on X. Throughout this paper the norm on X and on $\mathcal{L}(X)$

 ISBN: 978-84-614-6167-7

will be denoted by $||\cdot||$. Let I_d denote the identity operator on X.

Let $\mathbb{Z}$ denote the set of the integers, let $\mathbb{N}$ denote the set of all non-negative integers and let $\mathbb{N}^* = \mathbb{N}\setminus\{0\}$. We denote by $\mathcal{S}(\mathbb{N}, X)$ the linear space of all sequences $s : \mathbb{N} \to X$. For every $p \in [1, \infty)$ let

$$\ell^p(\mathbb{N}, X) = \{s \in \mathcal{S}(\mathbb{N}, X) \ : \ \sum_{k=0}^{\infty} ||s(k)||^p < \infty\}$$

which is a Banach space with respect to the norm $||s||_p = (\sum_{k=0}^{\infty} ||s(k)||^p)^{1/p}$. We consider the space

$$\ell^\infty(\mathbb{N}, X) = \{s \in \mathcal{S}(\mathbb{N}, X) \ : \ \sup_{n\in\mathbb{N}} ||s(n)|| < \infty\}$$

which is a Banach space with respect to the norm $||s||_\infty = \sup_{n\in\mathbb{N}} ||s(n)||$.

Let (Θ, d) be a metric space.

Definition 2.1 A mapping $\sigma : \Theta \times \mathbb{Z} \to \Theta$ is called *a discrete flow* on Θ if $\sigma(\theta, 0) = \theta$ and $\sigma(\theta, m + n) = \sigma(\sigma(\theta, m), n)$, for all $(\theta, m, n) \in \Theta \times \mathbb{Z}^2$.

Let $\{A(\theta)\}_{\theta\in\Theta} \subset \mathcal{L}(X)$. We consider the variational discrete system

$$(A) \qquad x(\theta)(n + 1) = A(\sigma(\theta, n))x(\theta)(n), \quad \forall(\theta, n) \in \Theta \times \mathbb{N},$$

where $x : \Theta \to \mathcal{S}(\mathbb{N}, X)$ and X is the state space.

The discrete cocycle associated with the system (A) is defined by

$$\Phi_A : \Theta \times \mathbb{N} \to \mathcal{L}(X), \quad \Phi_A(\theta, n) = \begin{cases} A(\sigma(\theta, n - 1)) \dots A(\theta) & , \quad n \in \mathbb{N}^* \\ I_d & , \quad n = 0 \end{cases}.$$

Remark 2.1 The discrete cocycle satisfies $\Phi_A(\theta, m+n) = \Phi_A(\sigma(\theta, n), m)\Phi_A(\theta, n)$, for all $(\theta, m, n) \in \Theta \times \mathbb{N}^2$ (*the evolution property*). This shows that the discrete cocycle associated with (A) ensures the propagator property.

Definition 2.2 The system (A) is said to be *uniformly exponentially stable* if there are $K, \nu > 0$ such that $||\Phi_A(\theta, n)|| \leq Ke^{-\nu n}$, for all $(\theta, n) \in \Theta \times \mathbb{N}$.

We associate with the system (A) the input-output system $(S_A) = \{S_\theta\}_{\theta\in\Theta}$, where for every $\theta \in \Theta$

$$(S_\theta) \qquad \begin{cases} x_\theta(n + 1) = A(\sigma(\theta, n))x_\theta(n) + s(n) & , \quad n \in \mathbb{N} \\ x_\theta(0) = 0 \end{cases}$$

with $s \in \mathcal{S}(\mathbb{N}, X)$.

 ISBN: 978-84-614-6167-7

Remark 2.2 For every $(\theta, s) \in \Theta \times \mathcal{S}(\mathbb{N}, X)$, the solution of (S_θ) has the form:

$$x_{\theta,s}(n) = \sum_{k=1}^{n} \Phi_A(\sigma(\theta, k), n-k)s(k-1), \quad \forall n \in \mathbb{N}^*.$$

The input-output stability of the associated control systems provides valuable information concerning the behavior of the initial system as well as facilitates the analysis in the presence of perturbations (see e.g. [14], [19]). A concept of input-output stability strongly related with the exponential stability is given by:

Definition 2.3 Let $p \in [1, \infty]$. The system (S_A) is said to be *completely* $(\ell^p(\mathbb{N}, X), \ell^p(\mathbb{N}, X))$*-stable* if the following properties hold:

(i) for every $s \in \ell^p(\mathbb{N}, X)$ and every $\theta \in \Theta$ the solution $x_{\theta,s} \in \ell^p(\mathbb{N}, X)$;

(ii) there is $L > 0$ such that $||x_{\theta,s}||_p \leq L||s||_p$, for all $(\theta, s) \in \Theta \times \ell^p(\mathbb{N}, X)$.

For the following characterization of the stability of a system (A) in terms of the complete $(\ell^p(\mathbb{N}, X), \ell^p(\mathbb{N}, X))$-stability of the associated control system (S_A) we refer to [19], Corollary 3.15.

Theorem 2.1 *Let $p \in [1, \infty]$. The system (A) is uniformly exponentially stable if and only if the system (S_A) is completely $(\ell^p(\mathbb{N}, X), \ell^p(\mathbb{N}, X))$-stable.*

According to the above result, in stability problems, it makes sense to work with a family of associated input-output operators. Precisely, for every $\theta \in \Theta$ we consider the family of input-output linear operators

$$\Gamma_\theta : \mathcal{S}(\mathbb{N}, X) \to \mathcal{S}(\mathbb{N}, X), \quad \Gamma_\theta(s) = x_{\theta,s}.$$

Then, the solvability of the system (S_A) may be rewritten in a classical form of input-output stability with respect to this family of linear operators.

Definition 2.4 Let $p \in [1, \infty]$. The family $\{\Gamma_\theta\}_{\theta \in \Theta}$ is said to be *completely* $(\ell^p(\mathbb{N}, X), \ell^p(\mathbb{N}, X))$*-stable* if the following properties hold:

(i) for every $s \in \ell^p(\mathbb{N}, X)$ and every $\theta \in \Theta$ we have that $\Gamma_\theta(s) \in \ell^p(\mathbb{N}, X)$;

(ii) there is $L > 0$ such that $||\Gamma_\theta(s)||_p \leq L||s||_p$, for all $(\theta, s) \in \Theta \times \ell^p(\mathbb{N}, X)$.

Remark 2.3 Let $p \in [1, \infty]$. The system (S_A) is completely $(\ell^p(\mathbb{N}, X), \ell^p(\mathbb{N}, X))$-stable if and only if the family $\{\Gamma_\theta\}_{\theta \in \Theta}$ is completely $(\ell^p(\mathbb{N}, X), \ell^p(\mathbb{N}, X))$-stable.

Remark 2.4 According to the condition (i) in Definition 2.4, if the family $\{\Gamma_\theta\}$ is completely $(\ell^p(\mathbb{N}, X), \ell^q(\mathbb{N}, X))$-stable, then the operators $\Gamma_\theta : \ell^p(\mathbb{N}, X) \to \ell^p(\mathbb{N}, X)$ are correctly defined and an easy computation shows that each operator is closed, so this is bounded. The second condition from Definition 2.4 shows that the boundedness holds in a uniform way with respect to $\theta \in \Theta$, i.e. $\sup_{\theta \in \Theta} ||\Gamma_\theta|| < \infty$.

 ISBN: 978-84-614-6167-7

Corollary 2.1 *Let $p \in [1, \infty]$. The system (A) is uniformly exponentially stable if and only if the family $\{\Gamma_\theta\}_{\theta\in\Theta}$ is completely $(\ell^p(\mathbb{N}, X), \ell^p(\mathbb{N}, X))$-stable.*

Proof. This follows from Theorem 2.1 and Remark 2.3. □

3 Stabilizability of variational discrete systems

Notations. If Y, Z are two Banach spaces we denote by $\mathcal{L}(Y, Z)$ the Banach space of all bounded linear operators $T : Y \to Z$. If $Y = Z$ we denote $\mathcal{L}(Y, Y) =: \mathcal{L}(Y)$. If (Θ, d) is a metric space we consider $\ell^\infty(\Theta, \mathcal{L}(Y, Z)) := \{R : \Theta \to \mathcal{L}(Y, Z) \mid \sup_{\theta\in\Theta} ||R(\theta)|| < \infty\}$, which is a Banach space with respect to the norm

$$|||R||| := \sup_{\theta\in\Theta} ||R(\theta)||.$$

Remark 3.1 If Y, Z, U are Banach spaces, $B \in \ell^\infty(\Theta, \mathcal{L}(Y, Z))$ and $F \in \ell^\infty(\Theta, \mathcal{L}(Z, U))$ then the mapping

$$BF : \Theta \to \mathcal{L}(Y, U), \quad (BF)(\theta) = B(\theta)F(\theta)$$

has the property that $BF \in \ell^\infty(\Theta, \mathcal{L}(Y, U))$.

Let (Θ, d) be a metric space and let $\sigma : \Theta \times \mathbb{Z} \to \Theta$ be a discrete flow on Θ. Let X be a Banach space and let $\{A(\theta)\}_{\theta\in\Theta} \subset \mathcal{L}(X)$. We consider the variational discrete time system

$$(A) \qquad x(\theta)(n + 1) = A(\sigma(\theta, n))x(\theta)(n), \quad (\theta, n) \in \Theta \times \mathbb{N}.$$

For every $D \in \ell^\infty(\Theta, \mathcal{L}(X))$ we consider the perturbed system

$$(A + D) \qquad x(\theta)(n + 1) = [A(\sigma(\theta, n)) + D(\sigma(\theta, n))]x(\theta)(n), \quad (\theta, n) \in \Theta \times \mathbb{N}.$$

Remark 3.2 If Φ_A is the discrete cocycle associated to the system (A) and Φ_{A+D} is the discrete cocycle associated to the system $(A + D)$, then the following perturbation formula holds:

$$\Phi_{A+D}(\theta, n) = \Phi_A(\theta, n) + \sum_{k=1}^{n} \Phi_A(\sigma(\theta, k), n - k)D(\sigma(\theta, k - 1))\Phi_{A+D}(\theta, k - 1)$$

for all $(\theta, n) \in \Theta \times \mathbb{N}^*$.

Let U be a Banach space and let $B \in \ell^\infty(\Theta, \mathcal{L}(U, X))$. We consider the discrete control system $(A, B) = (\mathcal{S}_\theta^{A,B})_{\theta\in\Theta}$, where for every $\theta \in \Theta$

$$(\mathcal{S}_\theta^{A,B}) \qquad \begin{cases} x_\theta(n + 1) = A(\sigma(\theta, n))x_\theta(n) + B(\sigma(\theta, n))u(n) & , \quad n \in \mathbb{N} \\ x_\theta(0) = 0 \end{cases}$$

with $u \in \mathcal{S}(\mathbb{N}, U)$.

 ISBN: 978-84-614-6167-7

Definition 3.1 The system (A, B) is said to be *stabilizable* if there is a feedback mapping $F \in \ell^\infty(\Theta,\ \mathcal{L}(X,U))$ such that the perturbed system $(A + BF)$ is uniformly exponentially stable.

Remark 3.3 If the system (A) is uniformly exponentially stable, then the system (A, B) is stabilizable (with the trivial feedback $F \equiv 0$).

The main question is whether the converse implication holds. A first answer is given by:

Example 3.1 Let X be a Banach space, let (Θ, d) be a metric space and let σ be a discrete flow on Θ. If I_d denotes the identity operator on X, then let $A(\theta) = I_d$, for all $\theta \in \Theta$. Let $U = X$ and let $B(\theta) = I_d$, for all $\theta \in \Theta$. If $\delta \in (0, 1)$ and $F(\theta) = -\delta I_d$, for all $\theta \in \Theta$, then we have that

$$||A(\theta) + (BF)(\theta)|| = 1 - \delta < 1, \quad \forall \theta \in \Theta.$$

This implies that the perturbed system $(A + BF)$ is uniformly exponentially stable, so the system (A, B) is stabilizable. For all that, it is obvious that the system (A) is not uniformly exponentially stable.

Then the central question is under what conditions a stabilizable system (A, B) corresponds to a uniformly exponentially stable system (A). The aim of our study is to answer this question. But, first, we need to point out several technical aspects.

Starting from the considerations presented in the previous section, for every $\theta \in \Theta$ we define the linear operator

$$\Gamma_\theta : \mathcal{S}(\mathbb{N}, X) \to \mathcal{S}(\mathbb{N}, X), \quad (\Gamma_\theta s)(n) = \begin{cases} \sum\limits_{k=1}^{n} \Phi_A(\sigma(\theta, k), n-k)s(k-1) & , \quad n \in \mathbb{N}^* \\ 0 & , \quad n = 0 \end{cases}.$$

Remark 3.4 Let $p \in [1, \infty]$. According to Corollary 2.1 we have that the system (A) is uniformly exponentially stable if and only if the family $\{\Gamma_\theta\}_{\theta\in\Theta}$ is completely $(\ell^p(\mathbb{N}, X),\ \ell^p(\mathbb{N}, X))$-stable.

For every $\theta \in \Theta$ we define the linear operator

$$B_\theta : \mathcal{S}(\mathbb{N}, U) \to \mathcal{S}(\mathbb{N}, X), \quad (B_\theta s)(n) = B(\sigma(\theta, n))s(n)$$

which is in fact a multiplication operator over the flow σ.

An natural concept of stability for families of operators (motivated by our previous investigations) is given by:

 ISBN: 978-84-614-6167-7

Definition 3.2 Let $p \in [1, \infty]$ and let V_1, V_2 be two Banach spaces. A family of linear operators $\{D_\theta : \mathcal{S}(\mathbb{N}, V_1) \to \mathcal{S}(\mathbb{N}, V_2)\}_{\theta \in \Theta}$ is said to be *completely* $(\ell^p(\mathbb{N}, V_1), \ell^p(\mathbb{N}, V_2))$-*stable* if the following conditions hold:

(i) for every $(\theta, s) \in \Theta \times \ell^p(\mathbb{N}, V_1)$ we have that $D_\theta(s) \in \ell^p(\mathbb{N}, V_2)$;

(ii) there is $\alpha > 0$ such that

$$||D_\theta(s)||_{\ell^p(\mathbb{N},V_2)} \leq \alpha \, ||s||_{\ell^p(\mathbb{N},V_1)}, \quad \forall (\theta, s) \in \Theta \times \ell^p(\mathbb{N}, V_1).$$

Remark 3.5 We have seen in the previous section that the operators $\{\Gamma_\theta\}$ may be regarded as a completely ℓ^p-stable family in certain conditions (see Corollary 2.1).

Remark 3.6 If $p \in [1, \infty]$, then the family $\{B_\theta\}_\theta$ is completely $(\ell^p(\mathbb{N}, U), \ell^p(\mathbb{N}, X))$-stable. Indeed, this follows by observing that

$$||(B_\theta s)(n)|| \leq ||B(\sigma(\theta, n))|| \, ||s(n)|| \leq |||B||| \, ||s(n)||, \quad \forall n \in \mathbb{N}$$

which implies that

$$||B_\theta s||_p \leq |||B||| \, ||s||_p, \quad \forall (\theta, s) \in \Theta \times \ell^p(\mathbb{N}, U).$$

In what follows, we associate with the control system (A, B) the family of input-output operators $\{I_\theta\}_{\theta \in \Theta}$ defined by

$$I_\theta : \mathcal{S}(\mathbb{N}, U) \to \mathcal{S}(\mathbb{N}, X), \quad I_\theta := \Gamma_\theta B_\theta.$$

Now, we may formulate the main result of this paper:

Theorem 3.1 *Let $p \in [1, \infty]$. A variational discrete system (A) is uniformly exponentially stable if and only if the associated control system (A, B) is stabilizable and the family of input-output operators $\{I_\theta\}_{\theta \in \Theta}$ is completely $(\ell^p(\mathbb{N}, U), \ell^p(\mathbb{N}, X))$-stable.*

Proof. Necessity. If the system (A) is uniformly exponentially stable then the system (A, B) is stabilizable with the null feedback $F = 0$. In addition, from Corollary 2.1 we have that the family $\{\Gamma_\theta\}_{\theta \in \Theta}$ is completely $(\ell^p(\mathbb{N}, X), \ell^p(\mathbb{N}, X))$-stable. Using Remark 3.6 we deduce that the family $\{I_\theta\}_{\theta \in \Theta}$ is completely $(\ell^p(\mathbb{N}, U), \ell^p(\mathbb{N}, X))$-stable.

Sufficiency. If the system (A, B) is stabilizable there is $F \in \ell^\infty(\Theta, \mathcal{L}(X, U))$ such that the system $(A + BF)$ is uniformly exponentially stable. For every $\theta \in \Theta$ we consider the linear operator

$$Q_\theta : \mathcal{S}(\mathbb{N}, X) \to \mathcal{S}(\mathbb{N}, X), \quad (Q_\theta s)(n) = \begin{cases} \sum\limits_{k=1}^{n} \Phi_{A+BF}(\sigma(\theta, k), n-k) s(k-1), & n \in \mathbb{N}^* \\ 0, & n = 0. \end{cases}$$

Since the system $(A + BF)$ is uniformly exponentially stable, according to Corollary 2.1 we have that the family $\{Q_\theta\}_{\theta \in \Theta}$ is completely $(\ell^p(\mathbb{N}, X), \ell^p(\mathbb{N}, X))$-stable.

 ISBN: 978-84-614-6167-7

For every $\theta \in \Theta$ we define the linear operator

$$F_\theta : \mathcal{S}(\mathbb{N}, X) \to \mathcal{S}(\mathbb{N}, U), \quad (F_\theta s)(n) = F(\sigma(\theta, n))s(n).$$

From

$$||(F_\theta s)(n)|| \leq |||F|||\, ||s(n)||, \quad \forall n \in \mathbb{N}, \forall (\theta, s) \in \Theta \times \mathcal{S}(\mathbb{N}, U)$$

we deduce that

$$||F_\theta s||_p \leq |||F|||\, ||s||_p, \quad \forall (\theta, s) \in \Theta \times \ell^p(\mathbb{N}, U),$$

which shows that the family $\{F_\theta\}_{\theta\in\Theta}$ is completely $(\ell^p(\mathbb{N}, X), \ell^p(\mathbb{N}, U))$-stable. Then, using the hypothesis we obtain that the family $\{I_\theta F_\theta Q_\theta\}_{\theta\in\Theta}$ is completely $(\ell^p(\mathbb{N}, X), \ell^p(\mathbb{N}, X))$-stable.

In addition, for every $(\theta, s) \in \Theta \times \mathcal{S}(\mathbb{N}, X)$ we have that

$$(I_\theta F_\theta Q_\theta s)(n) = \sum_{k=1}^{n} \Phi_A(\sigma(\theta, k), n-k)(B_\theta F_\theta Q_\theta s)(k-1) =$$

$$= \sum_{k=1}^{n} \Phi_A(\sigma(\theta, k), n-k)B(\sigma(\theta, k-1))F(\sigma(\theta, k-1))(Q_\theta s)(k-1), \quad \forall n \in \mathbb{N}^*.$$

Since $(Q_\theta s)(0) = 0$, we successively deduce that for every $n \in \mathbb{N}$, $n \geq 2$, we have:

$$(I_\theta F_\theta Q_\theta s)(n) = \sum_{k=2}^{n} \Phi_A(\sigma(\theta, k), n-k)B(\sigma(\theta, k-1))F(\sigma(\theta, k-1))(Q_\theta s)(k-1) =$$

$$= \sum_{j=1}^{n-1} \Phi_A(\sigma(\theta, j+1), n-j-1)B(\sigma(\theta, j))F(\sigma(\theta, j))(Q_\theta s)(j) =$$

$$= \sum_{j=1}^{n-1}\sum_{i=1}^{j} \Phi_A(\sigma(\theta, j+1), n-j-1)B(\sigma(\theta, j))F(\sigma(\theta, j))\Phi_{A+BF}(\sigma(\theta, i), j-i)s(i-1) =$$

$$= \sum_{i=1}^{n-1}\sum_{j=i}^{n-1} \Phi_A(\sigma(\theta, j+1), n-j-1)B(\sigma(\theta, j))F(\sigma(\theta, j))\Phi_{A+BF}(\sigma(\theta, i), j-i)s(i-1) =$$

$$= \sum_{i=1}^{n-1}\sum_{k=1}^{n-i} \Phi_A(\sigma(\theta, k+i), n-i-k)B(\sigma(\theta, i+k-1))F(\sigma(\theta, i+k-1))\cdot$$

$$\cdot\Phi_{A+BF}(\sigma(\theta, i), k-1)s(i-1) =$$

$$= \sum_{i=1}^{n-1}\sum_{k=1}^{n-i} \Phi_A(\sigma(\sigma(\theta, i), k), n-i-k)B(\sigma(\sigma(\theta, i), k-1))F(\sigma(\sigma(\theta, i), k-1))\cdot$$

$$\cdot\Phi_{A+BF}(\sigma(\theta, i), k-1)s(i-1).$$

Using Remark 3.2 we have that

$$(I_\theta F_\theta Q_\theta s)(n) = \sum_{i=1}^{n-1} [\Phi_{A+BF}(\sigma(\theta, i), n-i)s(i-1) - \Phi_A(\sigma(\theta, i), n-i)s(i-1)] =$$

 ISBN: 978-84-614-6167-7

$$= \sum_{i=1}^{n} \Phi_{A+BF}(\sigma(\theta,i), n-i)s(i-1) - \sum_{i=1}^{n} \Phi_A(\sigma(\theta,i), n-i)s(i-1) = (Q_\theta s)(n) - (\Gamma_\theta s)(n).$$

This implies that

$$(\Gamma_\theta s)(n) = (Q_\theta s)(n) - (I_\theta F_\theta Q_\theta s)(n), \quad \forall n \geq 2. \tag{3.1}$$

Since $(\Gamma_\theta s)(0) = 0$ and $(Q_\theta s)(0) = (I_\theta F_\theta Q_\theta s)(0) = 0$ we have that relation (3.1) holds also for $n = 0$. Moreover, for $n = 1$ we have that $(\Gamma_\theta s)(1) = s(0)$, $(Q_\theta s)(1) = s(0)$ and $(I_\theta F_\theta Q_\theta s)(1) = 0$, so relation (3.1) holds for $n = 1$. Then we may conclude that

$$(\Gamma_\theta s)(n) = (Q_\theta s)(n) - (I_\theta F_\theta Q_\theta s)(n), \quad \forall n \in \mathbb{N}$$

which implies that

$$\Gamma_\theta(s) = Q_\theta(s) - I_\theta F_\theta Q_\theta(s), \quad \forall (\theta, s) \in \Theta \times \mathcal{S}(\mathbb{N}, X). \tag{3.2}$$

Since the families $\{Q_\theta\}_{\theta\in\Theta}$ and $\{I_\theta F_\theta Q_\theta\}_{\theta\in\Theta}$ are completely $(\ell^p(\mathbb{N}, X),\ \ell^p(\mathbb{N}, X))$-stable, from relation (3.2) it follows that the family $\{\Gamma_\theta\}_{\theta\in\Theta}$ is completely $(\ell^p(\mathbb{N}, X),$ $\ell^p(\mathbb{N}, X))$-stable. Then, using Remark 3.4 we obtain that the system (A) is uniformly exponentially stable. □

Remark 3.7 A direct computation shows that, for every $\theta \in \Theta$ and $u \in \mathcal{S}(U)$, $x_\theta = I_\theta u$ satisfies the system $(S_\theta^{A,B})$, which reveals the fact that the complete stability of the family $\{I_\theta\}_{\theta\in\Theta}$ represents in fact an input-output stability condition with respect to the system (A, B).

Thus, to conclude our investigation, the above theorem shows that a variational discrete system is exponentially stable if and only if the associated control system is stabilizable and satisfies an ℓ^p input-output stability condition.

Acknowledgements

The work is supported by UEFISCDI-CNCS, Exploratory Research Projects PN II ID 1081/2008 no. 550/2009 and PN II ID 1080/2008 no.508/2009.

References

[1] D. Barcenas, S.-N. Chow, H. Leiva, A. Tineo Moya, *Skew-product semiflows and non-autonomous control systems*, J. Math. Anal. Appl. 381 (2011), 247-262.

[2] S. Clark, Y. Latushkin, S. Montgomery-Smith, T. Randolph, *Stability radius and internal versus external stability in Banach spaces: an evolution semigroup approach*, SIAM J. Control Optim. 38 (2000), 1757-1793.

[3] H. Leiva, J. Uzcategui, *Controllability of linear difference equations in Hilbert spaces and applications*, IMA J. Control Information 25 (2008), 323-340.

[4] H. Leiva, J. Uzcategui, *Exact controllability for semilinear difference equation and application*, J. Differ. Equ. Appl. 14 (2008), 671 - 679.

[5] I. V. Gaishun, *Controllability and stabilizability of discrete systems in a function space on a commutative semigroup*, Differ. Equations 40 (2004), 873–882.

[6] H. R. Henriquez, C. Cuevas, *Approximate controllability of abstract discrete-time systems*, Advances in Difference Equations (2010), Article ID 695290, 1-17.

[7] V. Komornik, *Exact Controllability and Stabilization - The Multiplier Method*, Masson, Paris, 1994.

[8] J.-L. Lions, *Contrôlabilité exacte, perturbations et stabilisation de systèmes distribués*, Tome 1, Masson, Paris, 1988.

[9] R. Medina, *Global stabilization of non-linear discrete-time systems by linear feedback*, IMA J. Control Information 25 (2008), 341-351.

[10] R. Medina, *Stabilization of discrete-time control systems with multiple state delays*, Adv. Difference Equ. (2009), Article ID 240707, 1-13.

[11] R. Medina, *Stabilizability for nonlinear systems of difference equations*, International J. Robust Nonlin. Control, 20 (2010), 1156-1165.

[12] R. Medina, *Non-Exponential Stabilization of Nonlinear Discrete-Time Systems*, J. Differ. Equ. Appl., DOI: 10.1080/10236198.2010. 488224.

[13] M. Megan, A. L. Sasu, B. Sasu, *Stabilizability and controllability of systems associated to linear skew-product semiflows*, Rev. Mat. Complut. 15 (2002), 599-618.

[14] A. L. Sasu, B. Sasu, *A lower bound for the stability radius of time-varying systems*, Proc. Amer. Math. Soc. 132 (2004), 3653–3659.

[15] B. Sasu, A. L. Sasu, *Stability and stabilizability for linear systems of difference equations*, J. Difference Equ. Appl. 10 (2004), 1085-1105.

[16] A. L. Sasu, *Stabilizability and controllability for systems of difference equations*, J. Difference Equ. Appl. 12 (2006), 821-826.

[17] A. L. Sasu, *Exponential dichotomy and dichotomy radius for difference equations*, J. Math. Anal. Appl. 344 (2008), 906-920.

[18] A. L. Sasu, *On exact controllability of variational discrete systems*, Appl. Math. Lett. 23 (2010), 101-104.

[19] B. Sasu, *Stability of difference equations and applications to robustness problems*, Adv. Difference Equ. (2010), Article ID 869608, 1-24.

 ISBN: 978-84-614-6167-7

Proceedings of the 11th International Conference on Computational and Mathematical Methods in Science and Engineering, CMMSE 2011 26–30 June 2011.

Efficient and reliable computation of the solutions of some notable non-linear equations

Javier Segura[1]

[1] *Departmento de Matemáticas, Estadística y Computación, Universidad de Cantabria, Spain.*

emails: javier.segura@unican.es

Abstract

The problem of computing with certainty all the real roots of a function can be solved in an efficient way if the functions satisfy certain functional equations. First, it is well known that if a sequence of functions $\{y_n(x)\}_{n\in\mathbb{N}}$ satisfies a second order difference equation $A_n y_{n+1}(x) + B_n y_n(x) + C_n y_{n-1}(x) = 0$ it is possible to compute the roots of some solutions $y_n(x)$ by solving an exact of approximated eigenvalue problem. Second, for solutions of first order linear difference-differential systems a second order global fixed point method can be obtained from the approximate integration of an associated Riccati equation; this method provides an scheme for computing with certainty all the zeros in any real interval [2]. More recently, fixed point methods with order of convergence four were developed for solutions of second order ODEs $y''(x) + B(x)y'(x) + A(x)y(x) = 0$ [3]. We present several examples of application of the methods and compare their performance. Improvements and extensions of the methods are discussed. In particular, we provide examples showing the validity of the ODE method for computing zeros in the complex plane. Some preliminary examples of a software package based on this methods, currently under construction, will be shown.

Key words: non-linear equations, ODEs, fixed point methods.

Introduction

The study of methods for solving non-linear equations is an important and active field in applied mathematics because of the importance of these topics in approximation theory, the theory of differential equations, and in many fields of physics and engineering.

In particular, some notable non-linear equations appear ubiquitously in many applications. Examples are the non-linear equations defining the nodes of gaussian quadrature (zeros of orthogonal polynomials) or the problem of computing zeros of Bessel functions or related functions.

 ISBN: 978-84-614-6167-7

A satisfactory solution of the problem is found when a method can be designed for computing with certainty all the zeros in a given interval (not losing any zero) and when the method is fast. A very fast method of a high order is of little practical utility if convergence is too local or if it is not clear that the scheme will compute all the zeros.

In many important cases, the function whose zeros are sought satisfy certain functional relations. Many (special) functions satisfy second order difference equations (also called three-term recurrence relations)

$$y_{n+1}(x) + \beta_n(x) y_n(x) + \alpha_n(x) y_{n-1}(x) = 0 \quad (TTRR), \tag{1}$$

or first order difference-differential linear systems

$$\begin{array}{l} y'_n(x) = a_n(x) y_n(x) + d_n(x) y_{n-1}(x) \\ y'_{n-1}(x) = b_n(x) y_{n-1}(x) + e_n(x) y_n(x) \end{array} \quad (DDE), \tag{2}$$

or second order ODEs

$$y''_n(x) + B_n(x) y'_n(x) + A_n(x) y_n(x) = 0 \quad (ODE), \tag{3}$$

or all of them simultaneously.

Next we outline the basic ingredients of three satisfactory methods (in the sense described before) for functions satisfying each of these type of relations.

1 TTRR method (matrix eigenvalue)

Consider recurrence relations of the form

$$a_n y_{n+1}(x) + b_n y_n(x) + c_n y_{n-1}(x) = g(x) y_n(x), \quad n = 0, 1, \dots ,$$

Considering the first N relations:

$$\mathbf{J}_N \mathbf{Y}_N(x) + a_{N-1} y_N(x) \mathbf{e}_N + c_0 y_{-1}(x) \mathbf{e}_1 = g(x) \mathbf{Y}_N(x),$$

$\mathbf{e}_1 = (1, 0, \dots, 0)^T$, $\mathbf{e}_N = (0, \dots, 0, 1)^T$, $\mathbf{Y}_N(x) = (y_0(x), \dots, y_{N-1}(x))^T$

$$\mathbf{J}_N = \begin{pmatrix} b_0 & a_0 & 0 & . & . & 0 \\ c_1 & b_1 & a_1 & 0 & . & 0 \\ 0 & c_2 & b_2 & a_2 & . & 0 \\ . & . & . & . & . & 0 \\ . & . & . & . & . & a_{N-2} \\ 0 & 0 & 0 & . & c_{N-1} & b_{N-1} \end{pmatrix}. \tag{4}$$

If for $x = x_0$ $a_{N-1} y_N(x_0) = c_0 y_{-1}(x_0) = 0$ we have the equation for an eigenvalue problem with eigenvalue $g(x_0)$.

This holds when x_0 is a zero of $y_N(x)$ (or of $y_{-1}(x)$) and, for any reason, $c_0 y_{-1}(x) = 0$ (or $a_{N-1} y_N(x) = 0$). In the case of polynomial solutions $c_0 y_{-1}(x) = 0$ and this gives a method for computing the nodes of gaussian quadrature (the roots of the polynomial y_N); the nodes are exactly the eigenvalues for the case of orthogonal polynomials. This is known as Golub-Welsch algorithm [1]. For minimal solutions of the recurrence, we have $a_{N-1} y_N(x) \approx 0$, and the eigenvalue method is an approximate method for computing the zeros of y_{-1}.

 ISBN: 978-84-614-6167-7

2 DDE method

Now, consider solutions of (2). A satisfactory method for computing the zeros of $y_n(x)$ (and similarly for $y_{n-1}(x)$), can be constructed from the approximate integration of an associated Riccati equation.

Defining

$$H(x) = \text{sign}(d_n(x))(e_n(x)/d_n(x))^{1/2}\frac{y_n(x)}{y_{n-1}(x)} \tag{5}$$

and considering a change of variable $z(x) = \int \sqrt{d(x)e(x)}dx$ one can check that the following Riccati equation is satisfied:

$$\dot{H}(z) = 1 + H(z)^2 - 2\eta(z)H(z) \tag{6}$$

where $\eta(z)$ is a (simple) function of the coefficients in (2).

A second order global method can be obtained from the approximate integration of (6) by neglecting the term with $\eta(z)$. The resulting equation has tangent functions as solutions and inverting this leads to the iteration

$$z_{n+1} = T(z_n),\ T(z) = z - \arctan(H(z)) \tag{7}$$

This iteration has global convergence properties. In addition, in intervals where $\eta(z)$ does not change sign, an scheme to compute all the zeros in succession is made available. For instance, if $\eta(z) < 0$ and $z^{(1)} < z^{(2)}$ are two consecutive zeros of $y_n(z)$. then the starting value $z_0 = z^{(1)} + \pi/2$ provides convergence to $z^{(2)}$.

Variants of this method with order of convergence three can also be constructed under certain monotonicity conditions.

3 ODE method (Sturm method)

Consider now solutions of homogeneous second order ODEs in normal form

$$y''(x) + A(x)y(x) = 0. \tag{8}$$

For equations with first derivative term we can transform to normal form by a suitable change of function.

Now, take $h(x) = y(x)/y'(x)$, which satisfies $h'(x) = 1 + A(x)h(x)^2$. Integrating under the assumption that $A(x)$ is approximately constant and inverting we obtain an iteration

$$x_{n+1} = g(x_n),\ g(x) = x - \frac{1}{\sqrt{A(x)}}\arctan\left(\sqrt{A(x)}h(x)\right) \tag{9}$$

This iteration has fourth order of convergence and has very good global convergence properties (a rare property for high order methods). Furthermore, in intervals where $A(x)$ is monotonic a marching scheme to compute all the zeros becomes available. That scheme can be understood as a consequence of the classical Sturm comparison theorems for second order ODEs.

 ISBN: 978-84-614-6167-7

4 Comparisons between methods

The three methods will be compared in different situations. The eigenvalue method is a good option for the case of classical orthogonal polynomials, but it is of more restricted applicability than the other two methods. As we will see, the ODE method is the fastest not only because it is of order four but also because of its good non-local properties, although in some cases it requires a pre-processing of the equation and an appropriate selection of a change of variable. Many examples of application can be considered. Hypergeometric functions (Gauss and Kummer) are a large set of functions, with important sub-cases (orthogonal polynomials, Bessel functions, Legendre functions) for which the methods hold, but the methods are not limited to this set.

5 Perspectives

Finally, we will show some examples of application of the ODE method to the computation of zeros in the complex plane. It appears that the behavior in $\mathbb{C}$ is also very good [4]; complex oscillation theory will be required in order to understand these good properties.

Acknowledgements

This work was supported by *Ministerio de Ciencia e Innovación*, project MTM2009-11686.

References

[1] G. H. Golub, J.H. Welsch. *Calculation of Gauss quadrature rules.* Math. Comput. **23** (1969) 221–230.

[2] J. Segura. *The zeros of special functions from a fixed point method.* SIAM J. Numer. Anal. **40**(1) (2002) 114–133.

[3] J. Segura. *Reliable computation of the zeros of solutions of second order linear ODEs using a fourth order method.* SIAM J. Numer. Anal. **48**(2) (2010) 452–469.

[4] A. Gil, J. Segura. *Algorithms for computing real and complex zeros of special functions.* In preparation.

 ISBN: 978-84-614-6167-7

Proceedings of the 11th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2011
26–30 June 2011.

On the group generated by the round functions of DESL

Rainer Steinwandt[1] **and Adriana Suárez Corona**[2]

[1] *Department of Mathematical Sciences, Florida Atlantic University*

[2] *Departamento de Matemáticas, Universidad de Oviedo*

emails: rsteinwa@fau.edu, adriana@orion.ciencias.uniovi.es

Abstract

DESL is a lightweight block cipher which is very similar to DES, but unlike DES uses only a single S-box. This work demonstrates that, as is the case for DES, the round functions of DESL generate the alternating group.

Key words: cryptography, block cipher, permutation group

1 Introduction

In the proceedings of the 14th International Workshop on Fast Software Encryption FSE 2007 [7] Leander et al. propose a lightweight cipher which is very similar to the Data Encryption Standard DES [3]. The proposed cipher introduces one radical change, however: all substitution boxes in the DES are replaced with a single new S-box. As detailed by Leander et al., this *DES Lightweight extension* (DESL) has very attractive features in terms of implementability on low-cost platforms. The obvious cryptanalytic question is if these features might have been paid for with a loss of security. In other words, is the security of DESL comparable to that of the original DES? In this contribution we show that the round functions of DESL generate the same permutation group as the round functions of DES, namely the alternating group on 2^{64} points. Our proof strategy is the same as taken by Wernsdorf for DES [9], the core part being to establish 3-transitivity for the group in question. It is not surprising that the replacement of DES's S-boxes in DESL necessitates modifications of Wernsdorf's proof, and one might be tempted to hope that facing only one S-box (instead of several as in DES) simplifies the analysis. For the S-box in question this did not really seem to be the case.

To keep our presentation reasonably self-contained, the next section presents the relevant details on the block cipher DESL.

 ISBN: 978-84-614-6167-7

2 Preliminaries

With the exception of two modifications, DESL is identical to the Data Encryption Standard; in particular, plaintexts and ciphertexts are elements of $\{0,1\}^{64}$ and the key can be taken for an element of $\{0,1\}^{56}$. The first difference between DES and DESL is not relevant for the group-theoretic property we explore: unlike for DES, there is no initial permutation and no final permutation of the data processed in the cipher. The implications of the second modification is less obvious: DESL replaces all eight S-boxes in DES with a single new S-box.

2.1 Description of DESL

Figure 1 illustrates the basic data flow in DESL, and we refer to the DES specification [3] and Leander et al.'s paper [7] for a detailed specification. For our purposes it is enough to be aware of the following:

- There are 16 rounds, each round i implementing a permutation $\pi_i \in S_{2^{64}}$ which depends on a round key $K_i \in \{0,1\}^{48}$. The latter is derived from the secret key $K \in \{0,1\}^{56}$ through a suitable key schedule.
- Each of the 16 rounds involves a round-key dependent function $F'_{K_i}(R_i) = P \circ S \circ \bigoplus \circ E$ where
  - $E : \{0,1\}^{32} \longrightarrow \{0,1\}^{48}$ is an injective map specified in [3].
  - $\bigoplus : \{0,1\}^{48} \longrightarrow \{0,1\}^{48}, x \longmapsto x \oplus K_i$ adds (xor) the round key K_i to the input
  - $S : \{0,1\}^{48} \longrightarrow \{0,1\}^{32}$ splits the input $(a_1, \ldots, a_{48}) \in \{0,1\}^{48}$ into 6-bit blocks and for $j = 1, \ldots, 8$ substitutes $(a_{6j-5}, \ldots, a_{6j}) \in \{0,1\}^6$ with the corresponding 4-bit value obtained from Table 1.

14	5	7	2	11	8	1	15	0	10	9	4	6	13	12	3
5	0	8	15	14	3	2	12	11	7	6	9	13	4	1	10
4	9	2	14	8	7	13	0	10	12	15	1	5	11	3	6
9	6	15	5	3	8	4	11	7	1	12	2	0	14	10	13

Table 1: The substitution function $S : \{0,1\}^6 \longrightarrow \{0,1\}^4$ of DESL is given by this S-box from [7]; $(a_1, \ldots, a_6) \in \{0,1\}^6$ is mapped to the 4-bit binary representation of the table entry in row no. a_1a_6 and column no. $a_2a_3a_4a_5$ (both interpreted as binary representation of a number in $\{0, \ldots, 3\}$ resp. $\{0, \ldots, 15\}$).

  - $P \in S_{2^{32}}$ is a permutation on 32-bit strings as specificed in [3].

 ISBN: 978-84-614-6167-7

- In each round, the 64-bit input is split into a left half $L_i \in \{0,1\}^{32}$ and a right half $R_i \in \{0,1\}^{32}$. Then the value $L'_i := F'_{K_i}(R_i) \oplus L_i$ is computed, where $\oplus$ is addition in $\{0,1\}^{48}$. The output of round $1, \ldots, 15$ is (R_i, L'_i). In the last round there is no swap, i. e., the value (L'_{16}, R_{16}) is output.

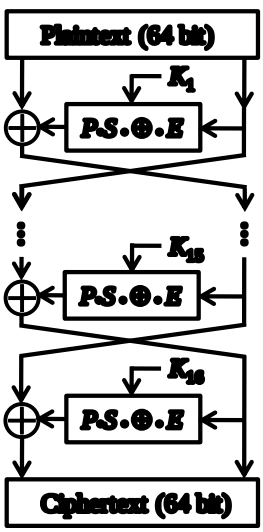


Figure 1: DESL overview

For our discussion we make use of an observation about DES by Davio et al. [4] which has also been exploited in [9]. Namely, we rewrite DESL as shown in Figure 2, i. e., by applying P^{-1} respectively P before the first round respectively after the last round, we combine E and P into a single function EP such that P no longer has to be applied after the application of the S-box. The composition of and E and P is given in Table 2.

3 The group generated by DESL's round functions

In this section we show the main ideas needed for the proof that the round functions of DESL generate the same group as the round functions of DES. The complete proof will be given in the full version of this paper. The main part of the argument is to establish 3-transitivity of the group generated by DESL's round functions.

 ISBN: 978-84-614-6167-7

25	16	7	20	21	29
21	29	12	28	17	1
17	1	15	23	26	5
26	5	18	31	10	2
10	2	8	24	14	32
14	32	27	3	9	19
9	19	13	30	6	22
6	22	11	4	25	16

Table 2: The function $EP : \{0,1\}^{32} \longrightarrow \{0,1\}^{48}$, mapping $(a_1, \ldots, a_{32})$ to $a_{EP(1)}, \ldots, a_{EP(32)}$ where $EP(j)$ is the j-th entry in the table, reading from left to right, top to bottom; e. g., $EP(7) = 21$.

3.1 Some notation

DESL's S-box sends 6-bit strings to 4-bit strings as detailed in Table 1, and the 6-bit inputs to the S-box are obtained by dividing a 48 bit value into eight 6-bit blocks. To refer to the latter, for $a \in \{0,1\}^{48}$ we set $[a]_j := (a_i)_{i=6j-5}^{6j}$ $(j = 1, \ldots, 8)$. Analogously, for $a \in \{0,1\}^{32}$, we write $[a]_j := (a_i)_{i=4j-3}^{4j}$ $(j = 1, \ldots, 8)$ for the selection of 4-bit blocks. It will be clear from the context when we are dealing with 48-bit respecticely 32-bit values.

For ease of readability of the proof, it turns out to be convenient to represent bitstrings by the decimal number they represent in binary; the length of the bitstring will always be clear from the context. Accordingly, we write $A_{2^{64}}$ and $S_{2^{64}}$ for the alternating and symmetric group respectively on $\{0,1\}^{64}$. Given a set of permutations Π, we write $\langle \Pi \rangle$ for the group generated by them. Specifically we are interested in the group G generated by the round functions F_K of DESL, where K ranges over all possible values in $\{0,1\}^{48}$— as in Wernsdorf's analysis of DES in [9], we ignore any restrictions imposed by the key schedule and allow to choose the round keys freely. Using the description and notation from Section 2.1, for a given round key $K \in \{0,1\}^{48}$ we can represent $F_K \in S_{2^{64}}$ as

$$\begin{array}{rccc} F_K : & \{0,1\}^{32} \times \{0,1\}^{32} & \longrightarrow & \{0,1\}^{32} \times \{0,1\}^{32} \\ & (a,b) & \longmapsto & \left(b, ([a]_i \oplus S([K]_i \oplus [EP(b)]_i))_{i=1}^{8}\right) \end{array} .$$

Our goal is to establish that the group $G = \left\langle \{F_K \in S_{2^{64}} | K \in \{0,1\}^{48}\} \right\rangle$ is nothing else but $A_{2^{64}}$.

3.2 Establishing 3-transitivity of G

Verifying transitivity of G is straightforward, and the work of Even and Goldreich [5] ensures that G is contained in the alternating group. In other words, we have

 ISBN: 978-84-614-6167-7

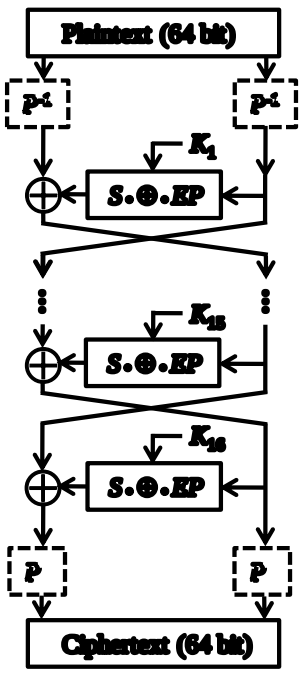


Figure 2: Equivalent description of DESL with the permutation P being applied before the expansion function E.

Lemma 1 *The round functions of DESL generate a subgroup of $A_{2^{64}}$ that acts transitively on $\{0,1\}^{64}$.*

In a first step of the proof one establishes transitivity of $G_0 := \{g \in G|\ g(0) = 0\}$ on $\{0,1\}^{64} \setminus \{(0,\ldots,0)\}$ and transitivity of $G_{0,d} := \{g \in G | g(0) = 0 \text{ and } g(d) = d\}$ on $\{0,1\}^{64} \setminus \{(0,\ldots,0), d\}$ where $d := (\delta_{31,i})_{i=1}^{64}$ has a single non-zero entry at the 31-st position. From [10, Theorem 9.1] we immediately obtain the following.

Lemma 2 *If G_0 is transitive on $\{0,1\}^{64} \setminus \{(0,\ldots,0)\}$ and $G_{0,d}$ is transitive on $\{0,1\}^{64} \setminus \{(0,\ldots,0), d\}$, then G is 3-transitive on $\{0,1\}^{64}$.*

Once it has been established that G is a 3-transitive subgroup of $A_{2^{64}}$, it is not particularly difficult to verify that G is actually equal to the alternating group on 2^{64} points, using results from [1, 2, 8]:

Theorem 1 *The round functions of DESL generate the alternating group, i.e., $G = A_{2^{64}}$.*

 ISBN: 978-84-614-6167-7

4 Conclusion

Unlike DES, the DES Lightweight extension (DESL) uses a single S-box. It turns out that nevertheless the round functions of DESL generate the same permutation group as the round functions of DES, namely the alternating group on 2^{64} points. So from this particular point of view DESL has no disadvantage compared to DES.

Acknowledgments

The second author acknowledges support of Spanish MEC (project MTM2010-18370-C04-01 and FPU grant AP2007-03141, cofinanced by the European Social Fund).

References

[1] Peter J. Cameron. Finite Permutation Groups and Finite Simple Groups. *Bulletin of the London Mathematical Society*, 13:1–22, 1981.

[2] Peter J. Cameron and John J. Cannon. Fast Recognition of Doubly Transitive Groups. *Journal of Symbolic Computation*, 12(4/5):459–474, October 1991.

[3] William M. Daley and Raymond G. Kammer. Data Encryption Standard (DES). Federal Information Processing Standards Publication, U.S. Department of Commerce/National Institute of Standards and Technology, October 1999.

[4] Marc Davio, Yvo Desmedt, Marc Fosséprez, René Govaerts, Jan Hulsbosch, Patrik Neutjens, Philippe Piret, Jean-Jacques Quisquater, Joos Vandewalle, and Pascal Wouters. Analytical Characteristics of the DES. In David Chaum, editor, *Advanves in Cryptology – CRYPTO '83*, pages 171–202. Plenum Press, 1984.

[5] Shimon Even and Oded Goldreich. DES-Like Functions Can Generate the Alternating Group. *IEEE Transactions on Information Theory*, 29(6):863–865, November 1983.

[6] Python Software Foundation. Python Programming Language – Offical Website. `http://www.python.org`, 2010.

[7] Gregor Leander, Christof Paar, Axel Poschmann, and Kai Schramm. New Lightweight DES Variants. In Alex Biryukov, editor, *Fast Software Encryption, 14th International Workshop, FSE 2007*, volume 4593 of *Lecture Notes in Computer Science*, pages 196–210. International Association for Cryptologic Research, Springer, 2007.

[8] Derek J. S. Robinson. *A Course in the Theory of Groups*. Springer, 1982.

 ISBN: 978-84-614-6167-7

[9] Ralph Wernsdorf. The One-Round Functions of the DES Generate the Alternating Group. In Rainer A. Rueppel, editor, *Advances in Cryptology – EUROCRYPT '92*, volume 658 of *Lecture Notes in Computer Science*, pages 99–112. Springer, 1993.

[10] Helmut Wielandt. *Finite Permutation Groups*. Academic Press, 1964.

 ISBN: 978-84-614-6167-7

Proceedings of the 11th International Conference on Computational and Mathematical Methods in Science and Engineering, CMMSE 2011 26–30 June 2011.

High Throughput peptide structure prediction with distributed volunteer computing networks

Timo Strunk[1], Moritz Wolf[1] and Wolfgang Wenzel[1]

[1] *Institute of Nanotechnology, Karlsruhe Institute of Technology*

emails: Timo.Strunk@kit.edu, Moritz.Wolf@kit.edu, Wolfgang.Wenzel@kit.edu

Abstract

Many short peptides are involved in important biological processes in the cell. Recent investigations have focused on the use of artificial peptides as antimicrobial drugs and antibiotics that differentially target bacterial and eucariotic cells. Because the number of possible peptide sequences is very large, functional peptide design necessitates automated synthesis and screening of a large number of peptide sequences. Protein structure prediction methods can aid in this design process by providing structure function relationships for the interaction of the peptide with the membrane. Here we demonstrate our de-novo peptide structure prediction method using massively parallel simulations in a biophysical forcefield to sample a sizable fraction of the peptides conformational space. By Performing simulations with the free-energy forcefield PFF02 on our volunteer computing network POEM@HOME, we were able to select the native conformation as the global minimum of the protein free energy for peptides of β, coiled and extended topologies. Our prediction protocol may be extended for high-throughput screening of large peptide databases for their structural features and by that enable the rapid prototyping of peptides for novel peptide design. *Key words: protein structure, structure prediction, distributed computing, BOINC, volunteer computing*

1 Introduction

The last effective new antibiotic drug was found in the 1970s and the resistance of bacteria has been proliferating ever since. For this reason much life-science research is focused on the development of novel antibiotics. In this context antimicrobial peptides[1, 2] are a promising new system to complement current antibiotics. Because there are plenty naturally occurring amino acids and functional peptides typically contain at least 10 amino acids, the number of possible sequences that must be screened either experimentally or in-silico is rather large. To contribute to the development of structure-function relationships, we therefore require a fast and accurate structure prediction method,

 ISBN: 978-84-614-6167-7

which is the basis of all subsequent analysis of peptide function In this paper we present a high-throughput peptide prediction protocol and validate it by predicting the structures of four experimentally known peptide structures with three different fold motifs to experimental resolution. The calculations have been performed on the distributed volunteer grid POEM@HOME using about 200 parallel simulations for each peptide.

2 Methods

To predict the structure of each peptide, we start with a completely extended conformation, which is subsequently relaxed in the free-energy forcefield in multiple parallel simulations. As previously demonstrated for a set of 32 medium-size proteins[3], the lowest energy conformation is then used as the predictive model.

2.1 Forcefield

PFF02 is an all atom free-energy forcefield, which stabilizes the native protein conformation in the global free-energy minimum.[3, 4, 5] It comprises five terms for Lennard-Jones, Electrostatic, angle-dependent hydrogen bonding and implicit solvation interactions. Furthermore a semi-empirical torsional potential is included.

2.2 Simulation Protocol

Our Monte Carlo simulations sample only the angular degrees of freedom, that is the main-chain and side-chain dihedral angles. New conformations are generated by either randomly perturbing randomly selected angles using angle-changes drawn from a gaussian distribution with a width of ten degrees around the original angle, or selecting angles from a distribution that reflects the dihedral angle distribution of the naturally occurring amino acids in the Ramachandran plot. This distribution is defined by circles comprising the beta sheet region centered at (-125,135) and the right-handed helix region at (-70,-35) both with a radius of 45. Angles are selected equidistributed within these circles.

2.3 POEM@HOME

POEM@HOME is a distributed volunteer computing platform for protein simulation using the BOINC[6] framework. From the website boinc.fzk.de volunteers download the client and continuously receive protein structures simulated on their PCs. Since its start in August 2008, POEM@HOME was well received and is now running with a combined computing power of 30 TFlop/s on average.

Upon submitting a protein sequence to the POEM@HOME server, we generate an extended conformation by chaining the amino acids together using idealized bond-lengths and bond-angles from template amino-acids. We then generate 10.000 parallel

 ISBN: 978-84-614-6167-7

work-units, each of which perform a simulated annealing simulation using a geometric temperature annealing in 1.5M Monte-Carlo steps. The final conformations of the simulations are sorted by energy and monitored for convergence. When a sufficient number of conformations have returned, the best energy structure is chosen as the final prediction. The prediction protocol is illustrated in Fig. 1.

3 Results

To show the validity of our approach, we chose proteins of three different fold motifs for prediction. 1N0A[7] and 1K43[8] with a beta-hairpin fold, 1FUV[9] with a more random coil-like fold and an extended peptide 1N9V[10] (the four letter codes correspond to RCSB pdb ids). Fig. 2 (a) shows 1N9A. As seen in the figure the stretched out global-fold is identified by PFF02 with an all-atom RMSD of 2.6 Å. Among the results one conformation with a RMSD of 1.4 Å could be generated (see Fig. 2b)); the energy of this conformation was found unfavorable however.

The beta-conformation of 1N0A could be reproduced with a RMSD of 1.2 Å. Only a few structures of better RMSD were actually discovered as seen in Fig. 2 d). The only difference observed is a slight tilt in one of the beta-bridges in the native structure, which could not be reproduced in the model (Fig. 2 c)). In contrast 1K43 is a bit longer beta-sheet than 1N0A and could be predicted with a RMSD of 2.4 Å(Fig. 3 a)). In comparison to the experimental structure our prediction extended the sheets a bit longer. The best sampled structure presents an RMSD of 1.8 Å(Fig. 3 b)). The relatively high RMSD for a beta-sheet structure could therefore derive from the fact, that we did not sample the low energy regions completely. The coiled fold of 1FUV could be predicted to a RMSD of 1.8 Å. The loop region deviates the most, as seen in Fig. 3c). The best discovered RMSD structures were only marginally nearer to the experimental structure than the chosen one. The best RMSD structure shows a RMSD of 1.75 Å(Fig. 3 d)).

4 Conclusion

We have implemented a free-energy-based peptide structure prediction protocol and applied it to for peptides of very different structure. We could show that it is possible to reliably predict the native conformation of peptides de-novo in a short time using the volunteer distributed computing resource POEM@HOME. Not only were we able to predict peptides in beta-conformations, but also those that feature only collapsed folds. Furthermore, it was possible to predict an extended peptide conformation. Further tests for a larger set of peptides are presently in progress. The protocol can be easily automated and made available to external users via that interface. The conformations generated may be subsequently used in further analysis to establish structure function relationships that correlate physiochemical properties of the peptide structure with their biological activity.

 ISBN: 978-84-614-6167-7

5 Figures

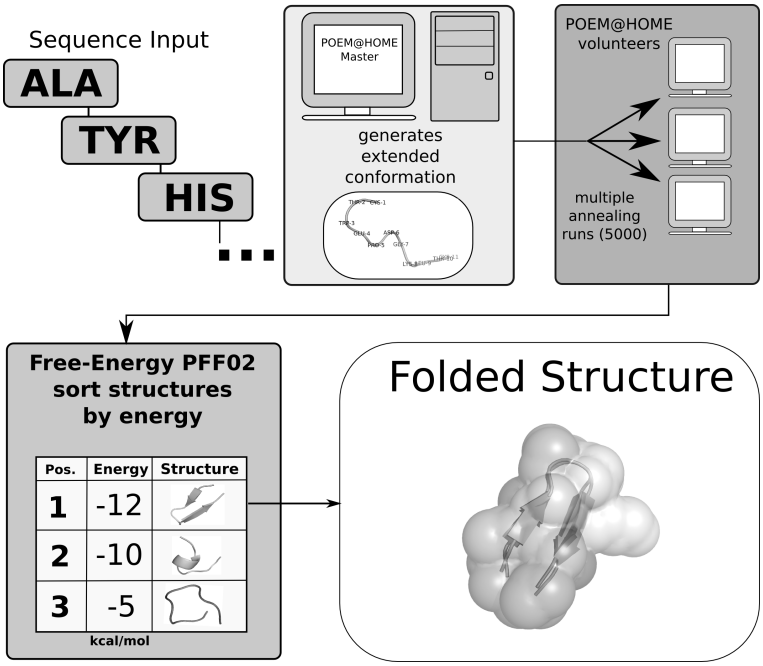


Figure 1: Prediction Algorithm used in this paper. After sequence input, multiple runs are generated from an extended structure and run on the POEM@HOME volunteer PCs. Returning annealed structures are sorted by energy. The best energy conformation is chosen as the prediction.

 ISBN: 978-84-614-6167-7

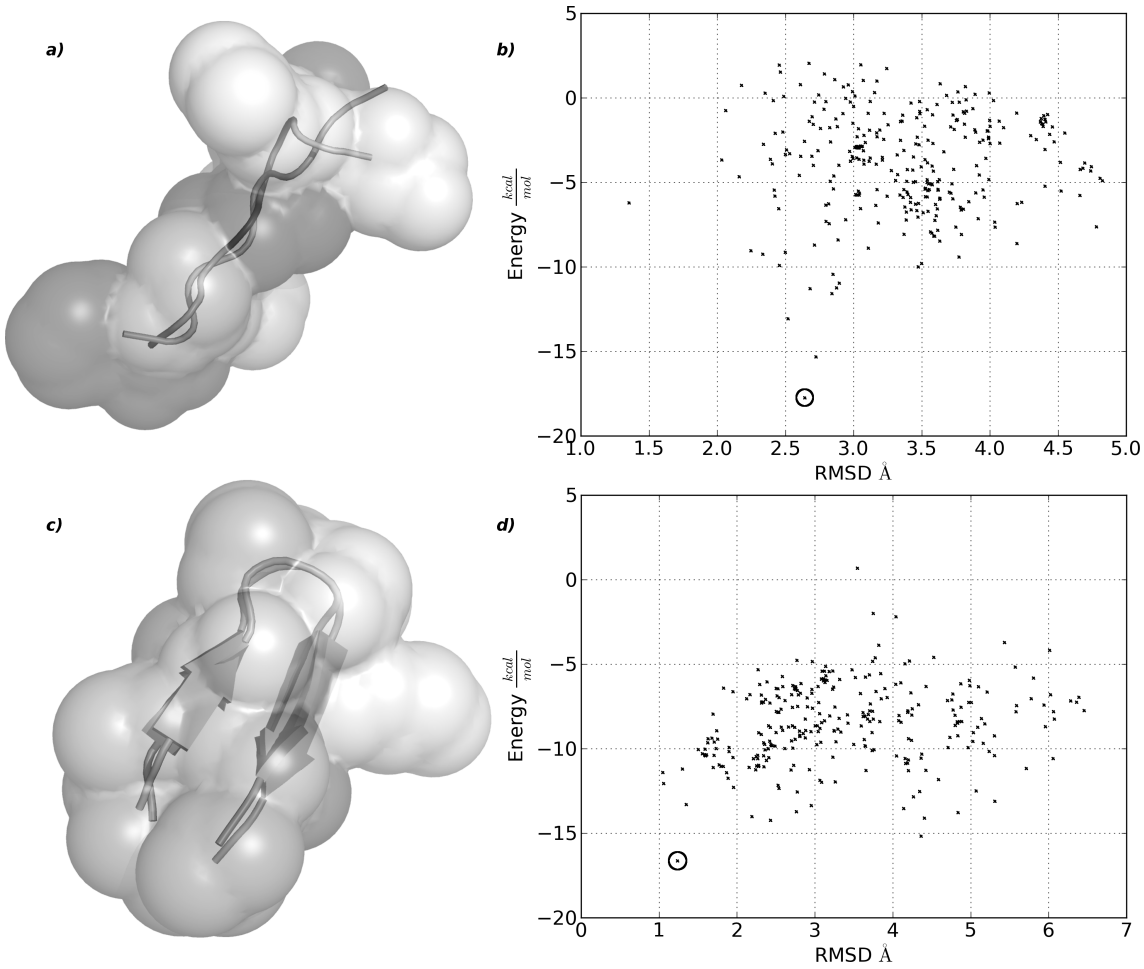


Figure 2: a) Result of the folding simulations of the peptide 1N9V (overlay). The predicted extended conformation could be clearly identified as the best energy structure and corresponds well to the experimental structure. b) RMSD-Energy Plot of all simulations of 1N9V. The chosen prediction has an all-atom RMSD of 2.6 Å. (circle) c) Comparison of the predicted structure of 1N0A with the experimental structure. Especially the loop region is well predicted. d) RMSD vs. Energy plot of all simulated conformations for 1N0A. The best energy conformation is also one of the structures with the smallest RMSD to the experimental structure.(circle)

 ISBN: 978-84-614-6167-7

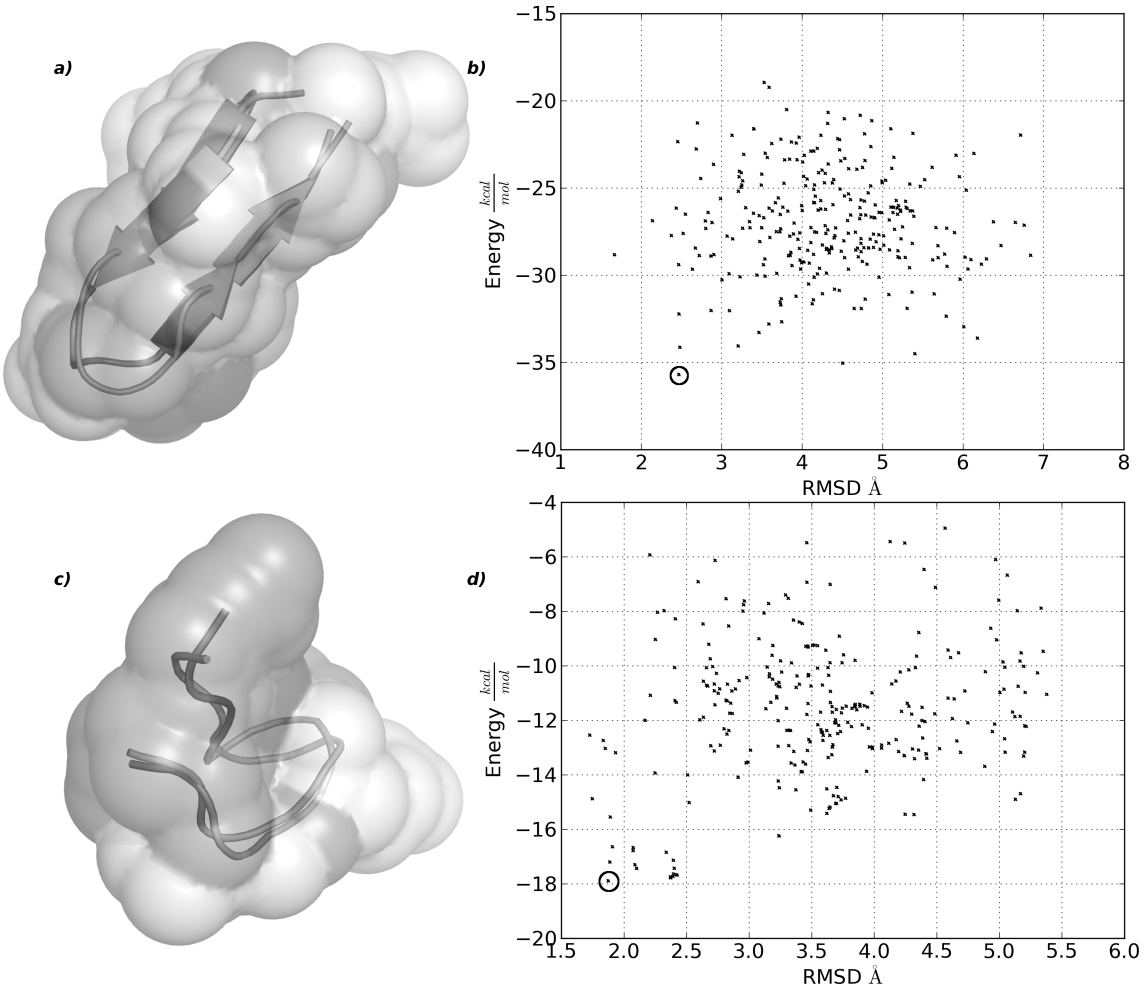


Figure 3: a) Result of the folding simulations of the peptide 1K43. The predicted structure matches the beta-sheet fold of the experimental structure. b) RMSD-Energy Plot of all simulations of 1K43. The chosen prediction has an all-atom RMSD of 2.5 Å(circle). c) Comparison of the predicted structure of 1FUV with the experimental structure. The collapsed fold of 1FUV was identified as native. d) RMSD vs. Energy plot of all simulated conformations for 1FUV. Only a few structures of better RMSD than the chosen one (circle) were identified.

 ISBN: 978-84-614-6167-7

Acknowledgements

We thank all of our POEM@HOME volunteers for their continuous support. Furthermore we thank the Carl-Zeiss Foundation for funding on the project Absolute Quality Control of Protein Structures.

References

[1] K. A. Brogden, Nat Rev Micro **3**, 238 (2005).

[2] M. R. Yeaman and N. Y. Yount, Pharmacological Reviews **55**, 27 (2003).

[3] A. Verma and W. Wenzel, BMC structural biology 12 (2007).

[4] S. M. Gopal and W. Wenzel, Angewandte Chemie **118**, 7890 (2006).

[5] A. Verma and W. Wenzel, Biophysical journal **96**, 3483 (2009).

[6] D. P. Anderson, in *5th IEEE/ACM International Workshop on Grid Computing* (PUBLISHER, ADDRESS, 2004), p. 4–10.

[7] T. Blandl, A. G. Cochran, and N. J. Skelton, Protein Science **12**, 237 (2003).

[8] M. T. Pastor *et al.*, Proceedings of the National Academy of Sciences **99**, 614 (2002).

[9] N. Assa-Munt, X. Jia, P. Laakkonen, and E. Ruoslahti, Biochemistry **40**, 2373 (2001).

[10] G. A. Spyroulias *et al.*, European Journal of Biochemistry **270**, 2163 (2003).

 ISBN: 978-84-614-6167-7

Proceedings of the 11th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2011
26–30 June 2011.

First attempts at modelling sleep

Ilaria Stura[1], Caterina Guiot[2], Lorenzo Priano[2,3] and Ezio Venturino[1]

[1] *Dipartimento di Matematica "Giuseppe Peano", Università di Torino, via Carlo Alberto 10, 10123 Torino, Italy*

[2] *Dept of Neuroscience, University of Torino, Torino, Italy,*

[3] *Dept. Neurology and Neurorehabilitation, IRCCS Ist Auxologico Italiano, Piancavallo (VB), Italy,*

emails: il8aria7stura@alice.it, caterina.guiot@unito.it, lorenzo.priano@unito.it, ezio.venturino@unito.it

Abstract

Sleep is a dynamic process involving different "families" of neurons promoting REM (Rapid Eyes Movements) and NREM (not REM) sleep, which are modulated by circadian and homeostatic needs. NREM sleep is characterized by the occurrence of peculiar transient EEG events, that are expression of synchronous cortical neuron discharges (transient synchronized EEG patterns or briefly TSEP). Our study aims at disclosing and quantifying its hidden dynamics by simulating the TSEP time series. The latter are obtained by adding the simulation of the neuronal spike activity to a Lotka-Volterra model reproducing the REM-nonREM alternation. The frequencies of the neuronal spike activity are randomized around N (up to 4) central values, which propagate to the cortex according to a given refractory period. The refractory period, as well as the system of thresholds accounting for the circadian and homeostatic control, are assumed on euristic bases. The Recurrence Plots of the simulated TSEP time series show a close resemblance with the real ones obtained by elaborating the EEG signals from human sleep registrations.

Key words: Dynamical system, Lotka-Volterra model, recurrence plot, REM sleep, NREM sleep
MSC 2000: AMS codes (92C20, 92C30, 92C50, 92B25, 92D25)

1 Introduction

Sleep is a complex process that can not be simply described as sequences of EEG rhythms detected during polysomnographic recordings. Even if sleep macrostructure, according to Rechtschaffen and Kales (R K) criteria [9], is characterized by a chain

 ISBN: 978-84-614-6167-7

of regular and predictable events (cyclic alternation of NREM and REM sleep, 90-120 minute intervals among REM sleep periods, stage 2 sleep preceding REM sleep, prevalence of REM sleep and stage 2 during the second half of normal sleep), the process does not repeat itself exactly the same each night. Circadian and homeostatic processes induce sleep during nighttime, but several factors may positively or negatively interfere with sleep induction and maintenance (mental state, acute or chronic disease, pain, muscle efforts, drugs, external environment). In this perspective sleep may be considered a dynamic process which has to finely modulate itself in order to obtain the required neurophysiological states at certain times, according to circadian and homeostatic needs, and despite external or internal interfering stimuli. This means that the neurophysiological structures involved in this process should not exhibit a rigidly predetermined behaviour but have to maintain the maximum adaptability. In other words sleep induction and maintenance imply a quote of variability inside fundamentally predetermined neurophysiological processes, so that perturbations below a threshold can be amortized and sleep macrostructure preserved.
Actually, this macrostructure of sleep may be considered the result of finer graduations of transient EEG activities (microstructure of sleep). Among these EEG activities, peculiar transient synchronized EEG patterns (TSEP) are supposed to be the expression of EEG synchronizing mechanisms accompanying the dynamic organization and stabilization of NREM sleep, ensuring flexible adaptation against perturbations. TSEP include: a) high voltage, low frequency component of K-complexes; b) transient delta bursts; c) high voltage, low frequency components of the Cycling Alternating Pattern (CAP) described by Terzano et al [2, 4, 12]. During normal sleep K-complexes, delta bursts and CAP progressively are grouping in recurring clusters, until steady slow wave sleep (SWS), expression of maximal EEG synchrony and deep sleep, is reached.

As it usumally occurs in biological systems, such series are normally non-stationary, requiring non-linear dynamics techniques, as for instance the use of the Recurrence Plot (RP) and the Recurrence Quantitative Analysis (RQA). Recurrence plots (introduced by Eckmann et al. (1987)) can visualize the recurrence of states in a phase space. Usually, a phase space does not have a low enough (two or three) dimension so as to be visually displayed. Higher-dimensional phase spaces can only be visualized by projection onto their two or three-dimensional sub-spaces. However, Eckmann's tool enables us to investigate the m-dimensional phase space trajectory by a two-dimensional representation of its recurrences. Such recurrence of a state at time i and at a different time j is pictured within a two-dimensional squared matrix containing black and white dots, where the black dots denote a recurrence; colour-coded dots may also be used. Here both axes represent time. Such model is called a recurrence plot (RP), Fig. 1.

Such an RP can be mathematically expressed as

$$R(i,j) = H(\epsilon - \|x(i) - x(j)\|), \quad x \in \mathbb{R}^m, \quad i,j = 1,\dots N$$

where N is the number of the state $x(i)$ considered, H is the Heaviside function and ϵ the thresold distance.

If only a time series is available, the phase space can be reconstructed by using a

 ISBN: 978-84-614-6167-7

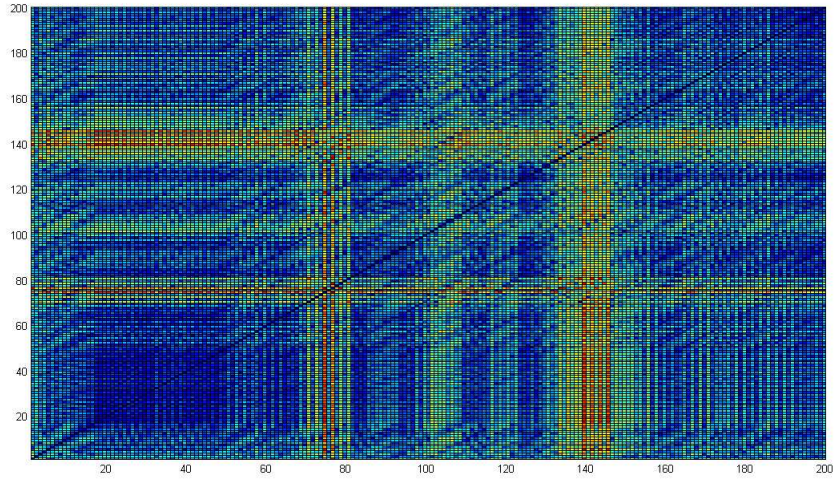


Figure 1: Recurrence Plots of a series of TSEP

time delay embedding

$$x(i) = (u(i), u(i+\tau), \ldots, u(i+(m-1)\tau)$$

where $u(i)$ is the time series, m the embedding dimension and τ the time delay.

The above techniques have already been proposed to analyze EEG signals [1, 10, 7, 13, 6] in both the wake and the sleep states, and have been used to evaluate sleep microstructure, i.e. TSEP time series [8].

Our study aimes at developing a model-based description of TSEP series, which is able to reproduce attractor-driven, hidden periodicity or, conversely, chaotic oscillation patterns in the series of these transient EEG phenomena related to sleep stage transitions and sleep maintenance.

The paper relates our experience in trying to construct a suitable model that provides an adequate simulation of what happens during the wake/sleep cycles. It is organized as follows. At first we present a model based on a dynamical system. In Section 3 we introduce a very simple model, and outline its shortcomings. In Section 4 a more refined model is presented, its results shown and discussed.

2 A model based on differential equations

We tried at first to create a very general model via dynamical systems. Since the REM phase is well characterized by a predator-prey model, we tried to study a system of four ordinary differential equations in which the neuronal populations R (REM ON), I (REM OFF), D (NREM) and N (neutral neurons) were competing with each other. In particular neutral neurons (N) are recruited by the other populations, so that they

 ISBN: 978-84-614-6167-7

can be regarded as prey and the other neuronal populations act like predators. The system is obtained by extending the classical Lotka-Volterra predator-prey system of population theory already used in such context, [3, 5, 11]. It reads

$$\begin{cases} \dot{N} = eN - lND - sNR \\ \dot{D} = kDN - hDR \\ \dot{R} = aRD - bRI + gRN \\ \dot{I} = -mI + cRI \end{cases} \tag{1}$$

where $a, b, c, e, g, h, l, k, m, s$ represent biological parameters, suitably chosen so that solutions oscillate and provide five peaks of REM sleep during the 8 sleeping hours, Fig. 2. Their values are $a = 0.0019$, $b = 0.0019$, $c = 0.0019$, $h = 0.0020$, $k = 0.0010$, $m = 0.0019$, $e = 0.0022$, $l = 0.0013$, $s = 0.0011$, $g = 0.0006$.

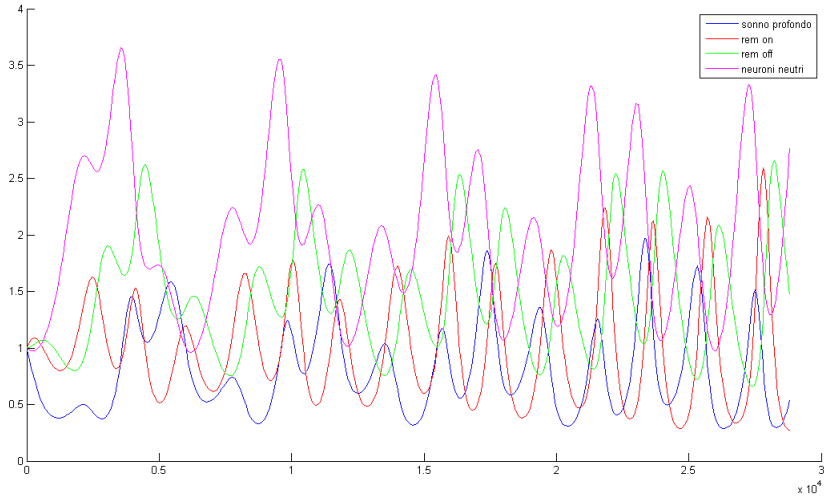


Figure 2: Typical simulation of the dynamic system (1)

The equilibrium points of (1) are the origin $P_1 = (0, 0, 0, 0)$ and the points

$$P_2 = \left(\frac{he}{ks}, 0, \frac{e}{s}, 0\right), \quad P_3 = \left(\frac{he}{hgl - ks}, \frac{-ghe}{a(hgl - ks)}, \frac{ke}{hgl - ks}, 0\right),$$
$$P_4 = \left(\frac{hm}{ck}, \frac{ec - sm}{cl}, \frac{m}{c}, \frac{aD + gN}{b}\right).$$

Note that P_3 is always infeasible, since its second component has always opposite sign with respect to the other ones. Instead P_4 is feasible only if $ec > sm$.

 ISBN: 978-84-614-6167-7

To study their stability we need the system's Jacobian

$$J = \begin{Vmatrix} e - lD - sR & -lN & -sN & 0 \\ kD & kN - hR & -hD & 0 \\ gR & aR & aD - bI + gN & -bR \\ 0 & 0 & cI & -m + cR \end{Vmatrix}$$

From this, now the eigenvalues at each feasible equilibrium could be determined. In view however of the fact that this model does not adequately represent the alternance of REM and deep sleep, which is our first goal, we do not provide its full analysis. In fact, we just limit ourselves to observe that for the chosen parameters, all the feasible equilibria are locally unstable. This can be remarked, by observing that the characteristic equation relative to each of the points P_1, P_2, P_4, when its coefficients are numerically evaluated with the parameter values given above, has at least one variation in the sign of the coefficients. Descartes rule tells then that at least one root with positive real part exists, i.e. one eigenvalue with positive real part. This is enough to establish instability. In fact, we have for P_1 the characteristic equation

$$Z^4 - 0.3 \cdot 10^{-4} Z^3 + 4.18 \cdot 10^{-6} Z^2 = 0$$

for P_2

$$Z^4 - 0.0043 Z^3 + 1.82 \cdot 10^{-5} Z^2 + 1.0882 \cdot 10^{-7} Z = 0$$

and for P_4

$$Z^4 + 4.3368 \cdot 10^{-19} Z^3 + 2.0430 \cdot 10^{-5} Z^2 + 3.0609 \cdot 10^{-8} Z + 5.8681 \cdot 10^{-11} = 0.$$

Thus the system cannot settle to any of these equilibria, and it must then oscillate.

We plan eventually to reexamine a system of this type modelling the behavior of neuronal populations. But for the time being, since this model goes beyond our original intentions, we have chosen to follow another approach, more suitable for our purposes.

3 Simple method

The model that we introduce now is composed only by a number N of frequencies and a barrier. In this model, the onset of TSEP are the moment at which a signal (one of the random frequencies) overshoots a constant barrier. The other constraint is that there is a biological time τ denoting the refractory period. This is the minimum time that needs to elapse between two consecutive signals.

The output of the model is the TSEP, i.e. the set of instants in which the signals overshoot the barrier. The output has been compared with some experimental data. We encountered a few problems, namely

- finding the correct number of frequencies;
- finding the correct value of τ;

 ISBN: 978-84-614-6167-7

- understanding the form of the barrier.

In the implementation of the model, we used some euristics. For τ, we took the value 12.8 ms. The barrier is taken as a constant value. To understand what was the best N for the model, we used a trial and error procedure. A program was written for the implementation of the model. In it, the number of frequencies is required in input.

By running the program several times, we were then able to compare the timing of the steps of the signals (without distinguishing what was the prevailing rate) with real data obtained from hypnogram of healthy patients. But in this way the data do not match at all, see Fig.s 3b, 4b, 5b.

We then tried to select a single frequency, but also in this case the desired result could not be obtained, see Fig.s 3c, 4c, 5c.

Thus this model proves to be too simplistic: the TSEP are too close together and there is no variability in the data. Experimentally, in fact, during the eight hours of sleep there are times when they are very close together and others in which they are more widespread. Since the variability is also due to the presence of REM sleep, in competition with NREM sleep (where one experiences the TSEP), a more complex model taking REM into account is needed.

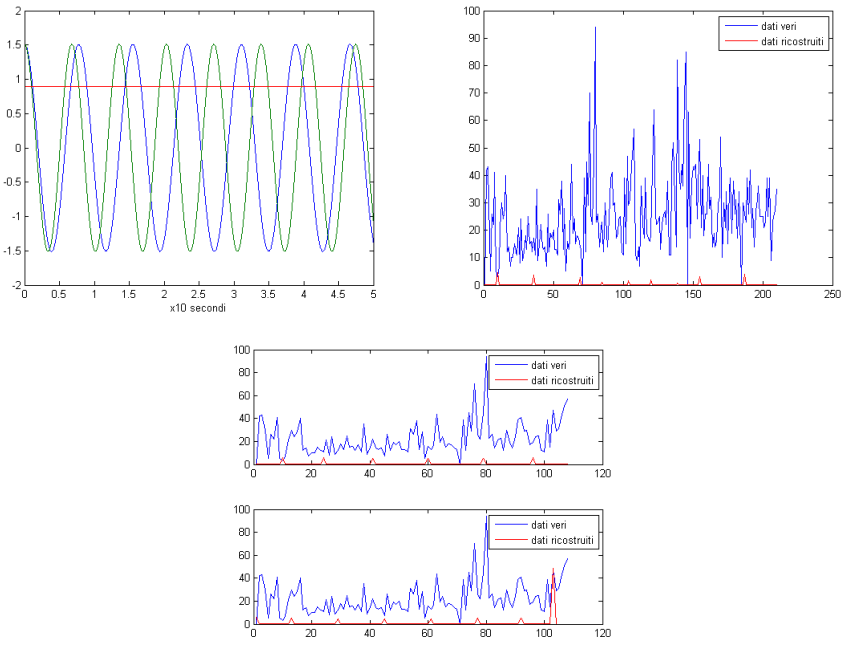


Figure 3: Model with two frequencies: a) the frequencies with the constant barrier, b) comparison between real data and the series created by the passages of all frequencies, c) comparison between real data and the series created by the passages of one single frequency.

 ISBN: 978-84-614-6167-7

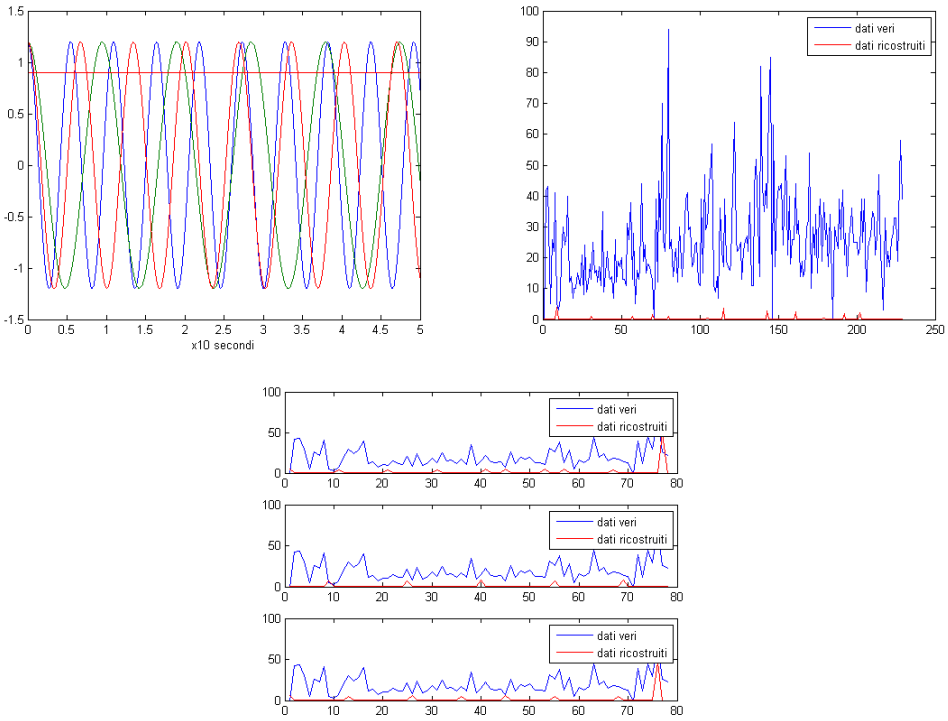


Figure 4: Model with three frequencies: a) the frequencies with the constant barrier, b) comparison between real data and the series created by the passages of all frequencies, c) comparison between real data and the series created by the passages of one single frequency.

 ISBN: 978-84-614-6167-7

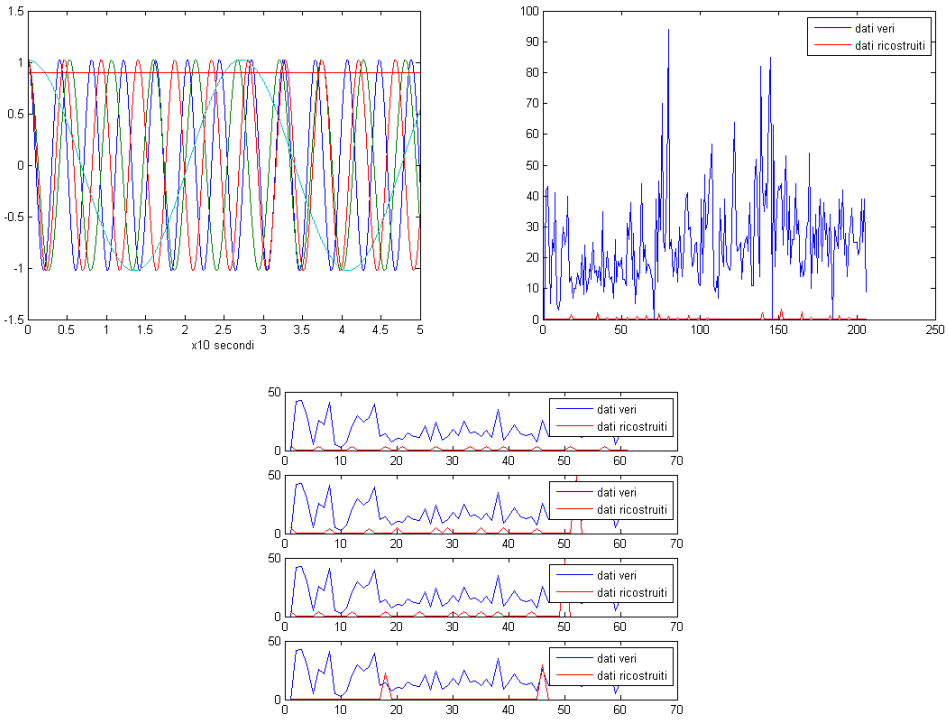


Figure 5: Model with four frequencies: a) the frequencies with the constant barrier, b) comparison between real data and the series created by the passages of all frequencies, c) comparison between real data and the series created by the passages of one single frequency.

 ISBN: 978-84-614-6167-7

4 A more refined model

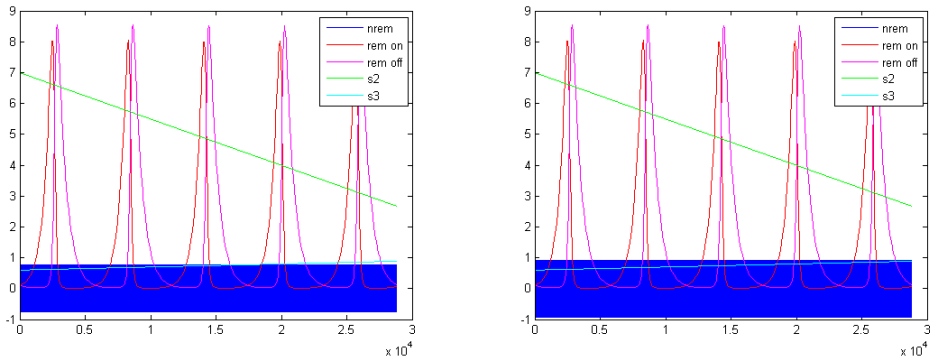


Figure 6: Some implementations of the advanced model.

This model takes into account only 8 hours of sleep. The functions involved are:

- The frequency of NREM sleep, implemented in the same way as in the previous model
- REM ON and REM OFF, according to previous studies, [5], are here modeled as solutions of Lotka-Volterra predator-prey type equations. They can be seen as a signal, the prey, (REM ON) that takes over and grows (at the expense of NREM) until being caught and overtaken by the predator signal (REM OFF) that "turns off" and makes NREM prevail.
- Homeostatic threshold on REM: used to indicate that the "need of sleep" changes overnight. In fact REM sleep is less important at the beginning (the body needs most of NREM, i.e. deep and restful sleep), but as time flows becomes predominant.
- Homeostatic threshold on NREM: in a specular way as for the REM ON barrier, this will indicate how much difficult is to fall in deep sleep as time flows. This is performed mathematically modulating the passing of the TSEP.
- The time τ: in [9, 1], it was shown that the time τ is not constant but variable (2 to 120 sec); in the code it is assumed as a function of time as the result of the modulation of the signals produced by cells population.

4.1 Model implementation

As we did for the parameters of the Lotka-Volterra equations in Section 3, For the REM ON and REM OFF we chose values so that the solutions exhibit five peaks during a typical eight hours sleep, as in reality experimentally verified.

 ISBN: 978-84-614-6167-7

Similarly, to express the barrier of REM ON, we chose suitable the formula $s_2 = 7 - 1.5 \cdot 10^{-4}t$, to model the fact that the importance of REM increase gradually over time. The line of NREM $s_3 = 0.6 + 10^{-5}t$ renders more difficult the switching of signals as the hours of sleep increase, see Fig. 6.

The choice of all these functions has been fixed in all the subsequent implementations of the program. In this next phase we focused on the shape that the function of τ must take, see Fig. 7. We established the fact that it is not a constant. After several tests, we focused on a combination of sines and exponentials that oscillate following the REM sleep behavior. The function of τ is then represented by

$$\tau = a + \sin t + e^{b \sin t} \cdot |\sin \frac{t + c}{d}|$$

where the coefficients $a = 3$, $b = 4.8$, $c = 3600$ and $d = 1800$ are chosen to normalize the function interval of eight hours of sleep.

This mathematical model reflects the trend of pseudo TSEP and their clustering (NREM) and adequately models the result of complex interactions of activities of specific neuronal populations:

- generators of REM sleep (pontine location);
- subcortical neuronal groups that lead, build and maintain the NREM (location of the thalamus, reticular substance, midbrain)

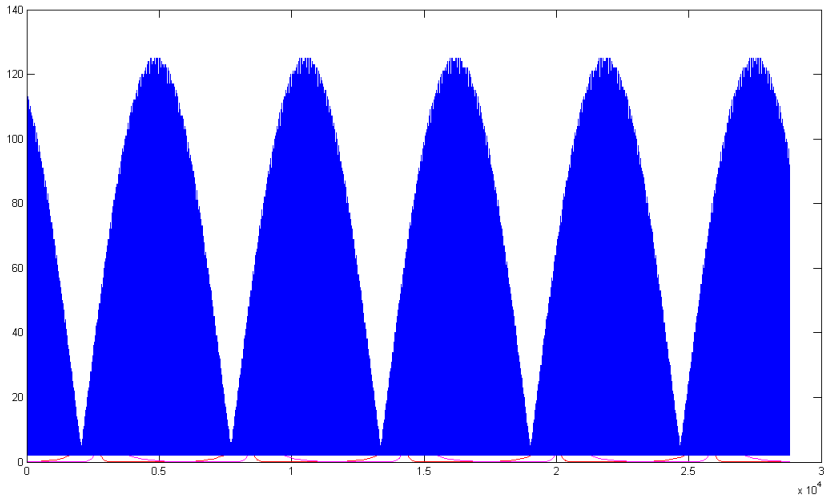


Figure 7: The refractory period τ exhibits large fluctuations in time as euristically assumed by our model

 ISBN: 978-84-614-6167-7

5 Conclusions

Our model-based description of the TSEP series is able to reproduce attractor-driven, pseudoperiodic and conversely, chaotic oscillation patterns in the series of these transient EEG phenomena related to sleep stage transitions and sleep maintenance. It is based upon the concept that TSEP is the resultant of interactions among several neuronal pools activities (inhibitory or excitatory), whose final output is a signal to cortex (generation of TSEP).

Although euristically obtained, the model is an adequate descriptor of the balance between homeostatic needs for NREM sleep and REM sleep pressure, supported by different cortical neuronal populations interactions. Nevertheless it presents some issues regarding the temporal limits of sleep onset and sleep ending. Here a further issue would be the integration of the circadian component into this framework. Moreover, it is still necessary to verify, by means of non-linear analysis techniques, the qualitative and quantitative correspondence between the real dynamic system and the dynamic system created by the model. As a matter of fact there is close resemblance between the recurrence plot obtained by the simulated TSAP time series, Fig. 8a, and the one of a real human subject, Fig. 8b.

Ideas for further developments in future studies are represented by the inclusion of the circadian component, as mentioned, and a refinement of the general pattern of the four neuronal populations, encompassing even the advanced model.

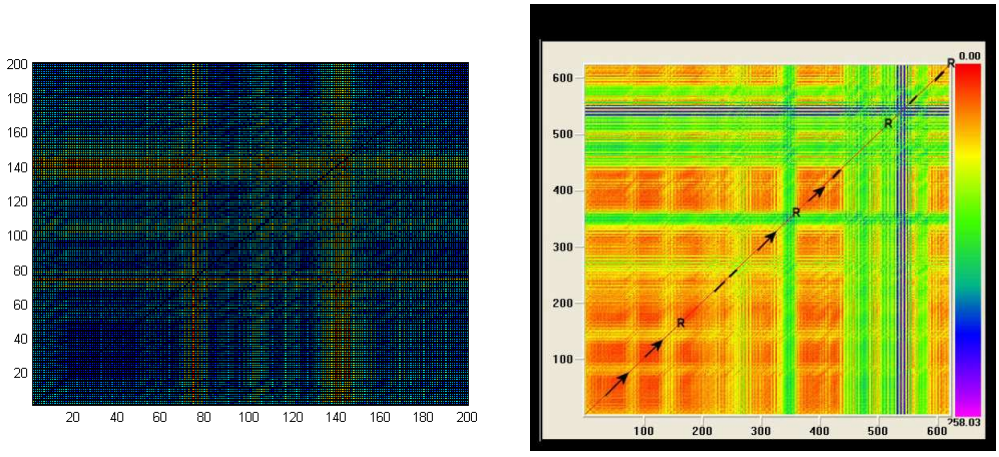


Figure 8: Left: recurrence plot obtained by the simulated TSAP time series; Right: recurrence plot of a real human subject.

References

[1] U. Acharya , O. Faustand, N. Kannathal, T. L. Chua, S. Laxminarayan, *Non-linear analysis of EEG signals at various sleep stages*, Computer Methods and

 ISBN: 978-84-614-6167-7

Programs in Biomedicine **80** (2005) 37–45.

[2] American Sleep Disorders Association (ASDA), *EEG arousals: scoring rules and examples. A preliminary report from the Sleep Disorders Atlas task Force of the American Sleep Disorders Association*, Sleep **15** (1992) 174–184.

[3] F. Ferrillo, S. Donadio, F. De Carli, S. Garbarino, L. Nobili., *A model-based approach to homeostatic and ultradian aspects of nocturnal sleep structure in narcolepsy*, Sleep **30** (2007) 157–165.

[4] P. Halasz, *K-complex, a reactive EEG graphoelement of NREM sleep: an old chap in a new garment*, Sleep Med Rev. **9** (2005) 391–412.

[5] J.A. Hobson, R.W. McCarley, P.W. Wyzinsky, *Sleep cycle oscillation: reciprocal discharge by two brainstem neurons*, Science **189** (1975) 55–58.

[6] J. S. Iwanski, E. Bradley, *Recurrence plots of experimental data: to embed or not to embed?*, Chaos **8** (1998) 861–871.

[7] H. Kantz and T. Schreiber, *Nonlinear Time Series Analysis*, Cambridge University Press, Cambridge, 1997.

[8] L. Priano, F. Saccomandi, A. Mauro, C. Guiot, *Non-linear recurrence analysis of NREM human sleep microstructure discloses deterministic oscillation patterns related to sleep stage transitions and sleep maintenance*, Conf. Proc. IEEE Eng. Med. Biol. Soc. **1** (2010) 4934–4937.

[9] A. Rechtschaffen, A. Kales (Editors), *A Manual of Standardized Terminology, Techniques, and Scoring System for Sleep Stages of Human Subjects*, Washington , USA: Public Health Service, US Government (1968).

[10] I. H. Song, D. S. Lee, S. I. Kim, *Recurrence quantification analysis of sleep electoencephalogram in sleep apnea syndrome in humans*, Neurosci Lett. **386** (2004) 148–153.

[11] Y. Tamakawa, A. Karashima, Y. Koyama, N. Katayama, M. Nakao, *A quartet neural system model orchestrating sleep and wakefulness mechanisms*, J Neurophys. **95** (2005) 2055–2069.

[12] M. G. Terzano, L. Parrino, *Origin and Significance of the Cyclic Alternating Pattern (CAP)*, Sleep Med Rev. **4** (2000) 101–123.

[13] J. P. Zbilut , C. L. Webber Jr, *Embeddings and delays as derived from quantification of recurrence plots*, Physics Letters A **171** (1992) 199–203.

 ISBN: 978-84-614-6167-7

Proceedings of the 11th International Conference on Computational and Mathematical Methods in Science and Engineering, CMMSE 2011 26–30 June 2011.

Hydrogen confined in SWCNTs: Anisotropy effects on ro-vibrational quantum levels.

J. Suarez[1] and F. Huarte-Larrañaga[1]

[1] *Departament de Química Física,, Universitat de Barcelona, 08028 Barcelona, Spain*

emails: jaime.suarez@ub.edu, fermin.huarte@ub.edu

Abstract

The present contribution aims at studying the effects of confinement in the quantum energy levels of H_2 by means of Single Wall Carbon Nanotubes (SWCNTs). 3- and 5-dimensional Cartesian coordinate-based models are employed for the full exact description of rotational, vibrational and translational motion of the H_2 molecule. The Time Dependent Schrödinger Equation (TDSE) is solved using a parallel computational code (GridTDSE) for wave packet propagations.

Key words: Carbon nanotubes, Cartesian coordinates, wave-packet propagations, ro-vibrational spectroscopy

1 Introduction

For the past ten years, the effect of nanoconfinement has generated an increasing interest both at experimental and theoretical levels. Research on Carbon nanotubes (CNTs) has developed rapidly since their discovery by Iijima in 1991 [1], due to their variety of unique mechanical (flexibility, resistance), physical (electronic conductivity, optic absorption) and chemical (catalytic behavior) properties. Amongst the most fascinating properties of CNTs is their capacity for encapsulating molecules and confining them in nearly unidimensional structures. In this sense, CNT structures stand as an appealing and efficient material for hydrogen storage [2]. Consequently, several studies considering the effect of CNT nanoconfinement on molecular and reactive hydrogen systems are presently being developed.

The present contribution aims at exploring how does CNT-confinement affect chemical-physical properties of a H_2 molecular system through the use of "*Cartesian coordinate-based*" time propagation of wave packets. The accurate simulation of the whole system, involving the large number of atoms in the CNT, with a fully quantum formalism is numerically unfeasible. On the other hand, classical formalisms are unable to predict quantum-nature confinement effects on such small-sized systems. An alternative with a reasonable computational cost is to restrict the Quantum formalism to

 ISBN: 978-84-614-6167-7

the description of H_2, modeling the interaction with the nanotube through the use of a semi-empirical potential. Despite this simplification, one has to deal still with a multi-dimensional (> 3) system. Not only this, but the presence of the nanotube breaks the symmetry of the confined system, reducing considerably the suitability of curvilinear coordinates, commonly used in standard quantum dynamics methods.

At this point, the development of a Cartesian coordinates-based code has clear advantages. On the one hand, computational advantages such as the simplicity of the algorithms and the absence of singularities in the Hamiltonian result in a straightforward extrapolation to other molecular confined-systems. The only modifications in the code should be at the level of the potential operator, which is diagonal, without involving new cross-derivatives in the Laplacian operator. On the other hand, the identification of the variables involved in the confinement mechanism is straightforward in a time-dependent cartesian picture. The balancing entry on this method consist of the larger number of dimensions that it usually involves ($3N - 3$, with N the number of atoms), compared with the usual $3N - 6$ of standard curvilinear coordinates. However, in the present case very simple symmetry impositions can help us to work on a reasonable 5-dimensional phase-space, with the additional advantage of an exact treatment of trans-vibro-rotational couplings (Coriolis coupling effect is treated in exact manner, in contrast to the second-order perturbation theory of usual standard approaches [3]).

The cost of such a heavy computational charge can only be assumed by the parallelization of the work over large clusters of computers, following a strategy that increases the memory storage capability and reduces the time of the calculations. The efficiency of the code relies basically on a minimum transfer of information between processors. Such is the strategy of the GridTDSE computational code [4], which has been extensively and successfully used to study molecular bonding as well as reactive systems. We extend its use here to nanoconfined molecular systems.

2 Methodology

GridTDSE code employs Grid methods for direct integration of the Time Dependent Schrödinger Equation (TDSE), which can be written in mass-weighted cartesian coordinates:

$$-\frac{\hbar^2}{2}\nabla^2\Psi(\vec{x},t) + V\Psi(\vec{x},t) = \mathrm{i}\hbar\frac{\partial\Psi(\vec{x},t)}{\partial t}, \tag{1}$$

In Grid methods, the wavefunction $\Psi(\vec{x},t)$ at any time t is discretized over the whole phase space. The accuracy of the calculations relies on the density of the grid of points considered. The action of the Hamiltonian operator over the wave packet turns into a matrix-vector multiplication, which can be calculated with the adequate computational algorithms for linear systems (i.e. the Portable Extensible Toolkit for Scientific Computation (PETSc), which is a variety of MPI-based libraries especially efficient for carrying out parallel vector-vector and matrix-vector operations). The matrix-vector multiplication becomes the most cumbersome task of the program. Although the kinetic part of the Hamiltonian is non-diagonal, computational charge is lightened by

 ISBN: 978-84-614-6167-7

using a Variable Order Finite Difference method (VOFD) [5], which increases the sparsity of the Hamiltonian matrix and reduces the allocated memory. We can additionally increase the sparsity by considering an energy cut-off for the process, excluding grid points with a potential energy above it.

The propagation in time of the wavefunction $\Psi(t) = \mathrm{e}^{-i\hat{H}t/\hbar}\Psi(t=0)$ is extracted by Second Order Difference method (SOD) [5]. Once obtained the wavefunction at any time of the process, all molecular properties, such as power spectra and eigenfunctions, can be readily obtained by Fourier transforming the corresponding time-correlation function. The code might be used as well at a Time Independent scheme to diagonalize the Hamiltonian matrix by means of a Lanczos algorithm, giving more accurate results for the eigenenergies at the same computational effort, although dynamical information would be lost in this scheme.

3 Formulation

The implementation of GridTDSE to study the dynamics of H_2 embedded in a CNT can be achieved by means of a 5-dimensional cartesian system, provided that we consider the CNT as an infinite-length rigid structure. Assigning $\vec{\rho}$ and $\vec{R}$ to the intramolecular H_2 vector and the position of the center of mass (c.m.) of H_2, respectively, three coordinates will be devoted to internal rotation-vibration (ρ_x,ρ_y,ρ_z) and two to c.o.m. translation (R_x,R_y). Here we have made use of the fact that the variation of the potential function along the CNT axis (here z) is negligible. The global Hamiltonian that describes simultaneously the rotation, vibration and translation of H_2 inside the CNT involves a two-term potential function:

$$V(\vec{R},\vec{\rho}) = V_{Morse}(\rho) + U(\vec{R},\vec{\rho}). \tag{2}$$

The first term accounts for the H-H covalent bond, and is modeled employing a Morse potential [6] with the same parameters as in [7]. The second term describes the atom-atom interaction of each hydrogen with the constituents of the rigid nanotube, and consists of a sum of C-H Lennard-Jones potential functions:

$$U(\vec{R},\vec{\rho}) = \sum_{i=1}^{N_C}\sum_{j=1}^{2} 4\epsilon\left[\left(\frac{\sigma}{r_{C_i-H_j}}\right)^{12} - \left(\frac{\sigma}{r_{C_i-H_j}}\right)^{6}\right]. \tag{3}$$

where $r_{C_i-H_j}$ is the C_i-H_j distance obtained from $\vec{R}$ and $\vec{\rho}$. Despite the manifold of different (σ,ϵ) parameters for the C-H interaction present in the bibliography [8, 9, 10], we recall that the scope of the present work is to elucidate a general qualitative picture of the effects of nanoconfinement on H_2 molecules. Therefore, we will not be here very concerned about this choice and we will carry out the calculations taking the values $\sigma = 2.82$ A and $\epsilon = 0.605$ kcal/mol that we already employed in [7].

In previous contributions [8, 11], it was shown that, as a general feature, the variation of the CNT radius resulted in two different types of confining potentials $U(\vec{R},\vec{\rho})$. Narrow-type CNTs exhibit parabolic potentials with the minimum at the center of the

 ISBN: 978-84-614-6167-7

nanotube, while for wider nanotubes the potential minimum is displaced towards the CNT's walls resulting in a ring structure. Due to the different qualitative picture expected in both structures, we have performed our calculations choosing one CNT of each type: CNT with chiral indexes ($n = 8$,$m = 0$) for the parabolic type and a wider ($n = 10$,$m = 0$) CNT for the ring type.

In a 5-dimensional calculation, the identification of the global translational, rotational and vibrational levels is not straightforward. This task can be enlightened by the previous interpretation of a 3-dimensional model, where the c.o.m. of H_2 is constrained at the position of minimum potential energy inside the nanotube. This approximation is especially suitable for the (8,0)-CNT, where this minimum is localized along the axis of the nanotube. Additionally, as a first benchmark for evaluating the effect of the nanotube we have included a simple 3-dimensional model of unconfined (gas phase) H_2. Despite its apparent simplicity, the absence of analytical solutions of the TDSE when a Coriolis coupling term is included is still originating work on the subject [3]. The main approach in the literature is to approximate solutions by the use of second-order perturbation theory. But most of these methods fail to reproduce the exact solutions for strong Coriolis couplings (large-J values). By means of the cartesian GridTDSE method we overcome this limitation treating exactly the Coriolis term, and accurate ro-vibrational energies are obtained with a Lanczos method.

4 Results and Discussion

We start by showing in figure 1 the results of a time-independent Lanczos calculation for the simplified 3-dimensional model in a CNT confinement of chiral indexes (8,0). Comparing to the unconfined energy levels, the primary effect of confinement is to shift the energy levels of the vibrational ground state ($\nu = 0$) by 105 cm^{-1}, while the displacement is 128 cm^{-1} for the first excited vibrational state ($\nu = 1$). This effect can be explained considering that for $\nu = 1$ the vibrational amplitude is higher and consequently the hydrogen atoms get closer to the nanotube's walls, enhancing CNT confinement. Concerning the rotational spectrum, we note as a general pattern that the $(2J+1)$-degeneracy of the unconfined H_2 system (blue broken lines) is broken in $(J+1)$ different lines (red dotted lines). The splitting arises from the cylindrical symmetry imposed by the nanotube, keeping always the same energy for rotational states (J,M) with the same absolute value M (projection on Z of the total angular momentum). On the other hand, we note a very similar pattern in the distribution of the rotational levels from $\nu = 0$ to $\nu = 1$, with smaller gaps for higher ν values as expected from theory (the momentum of inertia gets bigger and this influences $E_J = J(J + 1)/(2I)$). One remarkable fact is that J-manifolds become broader on the excited vibrational state (i.e., 75 cm^{-1} for $\nu = 0$ while 100 cm^{-1} for $\nu = 1$).

In figure 2 we show comparative results of both 3-dimensional (upper figure) and 5-dimensional (lower figure) wave packet propagations, again for the (8,0) CNT. We note that both figures are aligned at the zero point energy (ZPE) in order to better illustrate changes in rotational energies. In the upper figure, we propagate a 3-D

 ISBN: 978-84-614-6167-7

Gaussian wave-packet under a SOD scheme, positioning the initial internuclear vector on several orientations relative to the SWCNT. Taking an initial arbitrary orientation (here called $XYZ-$) we can extract most of the eigenvalues of the confined system. On the other hand, repeating the calculations for specific axis-oriented initial wavefunctions provides useful insight into the nature of the eigenfunctions. In the figure we note for instance that the minimum energy at each J-manifold corresponds to a Z-oriented (i.e. nanotube axis) initial internuclear vector $\vec{\rho}$. On the other hand, due to the cylindrical symmetry, both X and Y orientations yield the same result -for simplicity only X-orientation is displayed-, with a $J-2n$ ($n=0,\ldots$) pattern. We can therefore assign Z-oriented states to $M=0$ states, while X- and Y-oriented states are linear combinations of eigenstates with the same M-parity. The effect of confinement in this CNT can be viewed then as a perturbation of the corresponding spherical harmonics.

The energy spectrum of the 5-dimensional model is displayed in the lower figure. The ground state energy is in this case 273 cm^{-1} higher than the corresponding energy for 3-D. This difference is mainly due to the zero-point translational energy that was hindered in the 3-dimensional model. In the 5D case, translational and rotational energies are of the same order, and couplings between both motions are more efficient. Consequently, the time-independent energy spectrum obtained by a Lanczos method is considerably more dense and the identification of the type of motion in each case is much more difficult than in the 3-dimensional case. However, if we extract the translationally-excited states from the spectrum by time-propagating a wave-packet initially localized at the potential minimum (the center of the nanotube), the structure and nature of each eigenvalue is much more clear. The figure resembles then a lot the 3D pattern above, and the $(2J+1)$ degeneracy of the unconfined molecule is again broken. The M-manifold for each specific J value is slightly wider in the 5D case than in the 3D one. The energies of most Z-oriented ($M=0$) states do not change significantly from 3D to 5D, while the biggest difference is obtained for ($M=J$) states. This can be explained by the fact that high-M states are linear combinations of states oriented on X and Y, which are the axes of the translational coordinates R_x and R_y included in 5D. Not all the lines, however, can be derived from a simple 3-dimensional model. Looking carefully at the $J=3$ ensemble, we note the presence of a new line at 3177 cm^{-1} oriented along X-axis, while there are two new Z-oriented lines at 3370 and 3419 cm^{-1}. These are the result of a strong rotation-translation coupling. In order to confirm that these lines correspond to $J=3$ we have filtered out other J values by introducing angular momentum projection operators P_J in the initial wavefunction, as described in [4], although the results have not been included here.

Results for the ring type potential created by the (10,0)-nanotube will be presented at the CMMSE 2011 Conference.

 ISBN: 978-84-614-6167-7

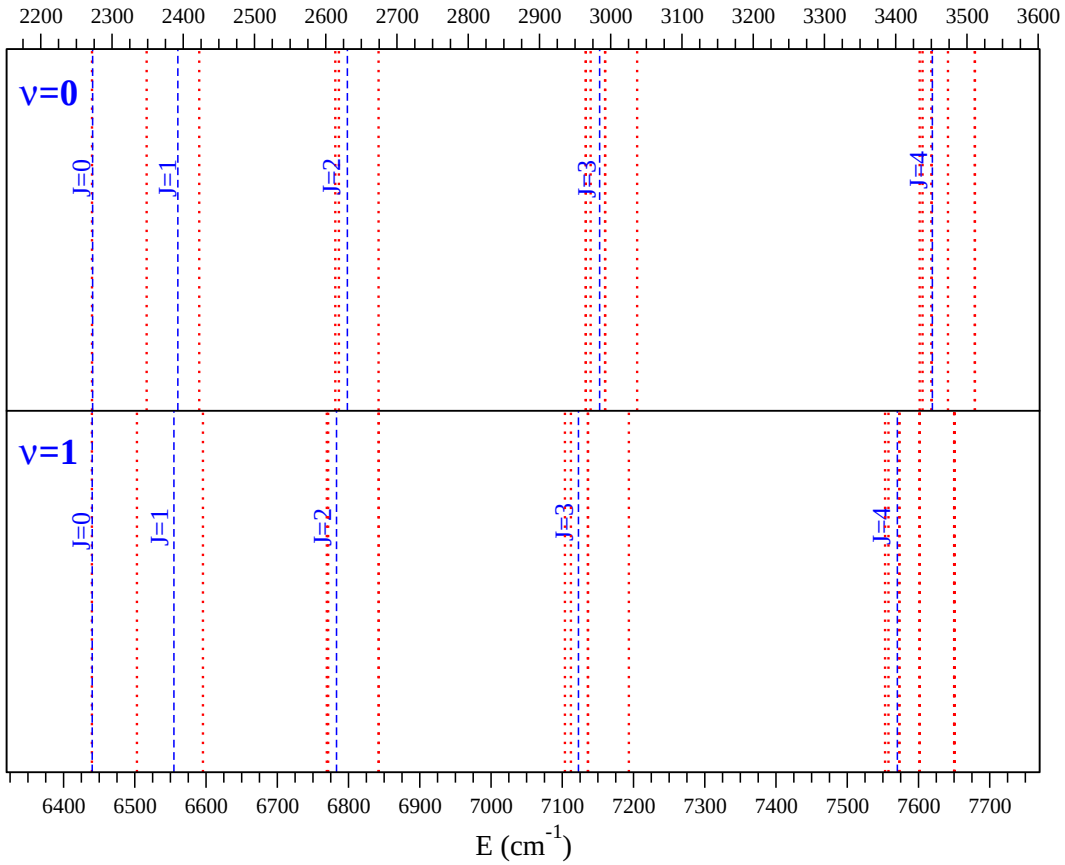


Figure 1: Comparative ro-vibrational energies ($\nu = 0$ up, $\nu = 1$ down) obtained from a 3-dimensional center-of-mass-fixed model ($\vec{R} = \vec{R_0}$) using Lanczos techniques for isolated H_2 (broken line) and H_2 confined in a (8,0)-nanotube. Specific-J values are labeled in the isolated case.

Acknowledgements

This work has been supported by the Spanish MICINN (Project CTQ2009-12215). CESCA is thanked for allocating massive parallel calculations on their supercomputing facilities. The authors would like to acknowledge Prof. Stavros Farantos and Dr. Stamatis Stamatiadis for their help in the development of GridTDSE code.

References

[1] S. Iijima, *Helical microtubules of graphitic carbon*, Nature **354** (1991) 56.

[2] A. Züttel, *Materials for hydrogen storage*, Mater. Today **6** (2003) 24.

[3] D. A. Morales, *Supersymmetric improvement of the Pekeris approximation for the rotating Morse potential*, Chem. Phys. Lett. **394** (2004) 68–75.

 ISBN: 978-84-614-6167-7

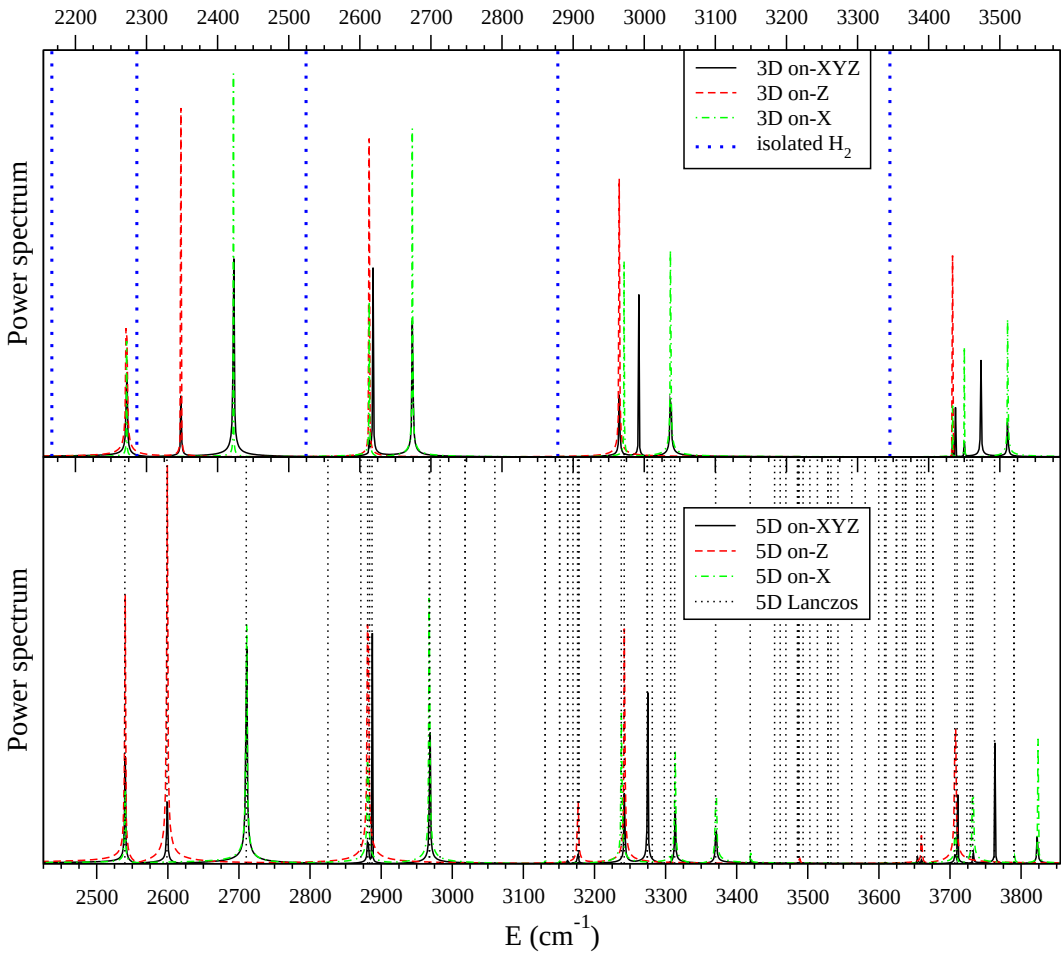


Figure 2: Power spectra for 3-dimensional (upper figure) and 5-dimensional (lower figure) models of H_2 confined in a (8,0)-nanotube obtained by wave-packet propagations.

 ISBN: 978-84-614-6167-7

[4] J. Suarez, S. C. Farantos, S. Stamatiadis, L. Lathouwers, *A method for solving the molecular Schrödinger equation in Cartesian coordinates via angular momentum projection operators*, Comp. Phys. Comm. **180** (2009) 2025–2033.

[5] R. Kosloff, *Quantum molecular dynamics on grids*, Dynamics of Molecules and Chemical Reactions, Marcel Dekker, Inc. (1996) 185.

[6] P. M. Morse, *Diatomic molecules according to the Wave Mechanics. II. Vibrational levels*, Phys. Rev. **34** (1929) 57.

[7] F. Huarte-Larrañaga and M. Albertí, *A molecular dynamics study of the distribution of molecular hydrogen physisorbed on single walled carbon nanotubes*, Chem. Phys. Lett. **445** (2007) 227–232.

[8] T. Lu, E. M. Goldfield and S. K. Gray, *Quantum States of Hydrogen and Its Isotopes Confined in Single-Walled Carbon Nanotubes: Dependence on Interaction Potential and Extreme Two-Dimensional Confinement*, J. Phys. Chem. B **110** (2006) 1742–1751.

[9] Y. Ma, Y. Xia, M. Zhao, R. Wang and L. Mei, *Effective hydrogen storage in single-wall carbon nanotubes*, Phys. Rev. B **63** (2001) 115422.

[10] M. Xu, F. Sebastianelli, B. R. Gibbons, Z. Bacić, R. Lawler and N. J. Turro, *Coupled translation-rotation eigenstates of H_2 in C_{60} and C_{70} on the spectroscopically optimized interaction potential: Effects of cage anisotropy on the energy level structure and assignments*, J. Chem. Phys. **130** (2009) 224306.

[11] T. Yildirim and A. B. Harris, *Quantum dynamics of a hydrogen molecule confined in a cylindrical potential*, Phys. Rev. B **67** (2003) 245413.

 ISBN: 978-84-614-6167-7

Proceedings of the 11th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2011
26–30 June 2011.

Modelling Structure of Colloidal Assemblies: Methodology & Examples

Bosiljka Tadić[1], Milovan Šuvakov[2] and Gregor Trefalt[3]

[1] *Department of Theoretical Physics, Jožef Stefan Institute, Ljubljana, Slovenia*

[2] *Centre for Non-equilibrium Processes, Institute of Physics, Belgrade, Serbia*

[3] *Electronic Ceramics Department, Jožef Stefan Institute, Ljubljana, Slovenia*

emails: bosiljka.tadic@ijs.si, suvakov@gmail.com, gregor.trefalt@ijs.si

Abstract

Large-scale nanoparticle assemblies may appear in different structures, depending on the fabrication method, the parameters and the constraints controlling the self-assembly processes. Consequently, the emergent aggregates may have different physical or chemical properties, closely related with their structure at mesoscopic and global scale. We briefly discuss mathematical methodology to systematically explore structures of nanoparticle assemblies by suitable mapping onto graphs (nanoparticle networks), and present some results of the numerical modeling of the assembly processes with competing attractive–repulsive forces. In particular, we study binary colloidal aggregation, where two types of particles are binding via molecular recognition, and the aggregation of three types of charged particles with different particle sizes. We show how different structures emerge when certain parameters of the self-assembly processes are varied.

Key words: nanoparticle networks, binary colloidal aggregation, charged colloids

1 Introduction

Macroscopic assemblies of nano-particles as bulk nanostructured materials or thin films [1] may exhibit a variety of different structures and with them related collective properties [2]. These are dependent on the assembly processes [3, 4] and the variations of the relevant parameters of the assembly, which affect the interactions among the constitutive nanoparticles [5]. Therefore, the structure of the nanoparticle assembly is of key importance in bottom-up fabrication of the functional nano-material. Two segments along that line, the self-assembly process itself, and the inter-dependences between the emergent spatial arrangement of the nanoparticles with the physical (chemical) properties of the material, require theoretical modelling and numerical analysis [6].

 ISBN: 978-84-614-6167-7

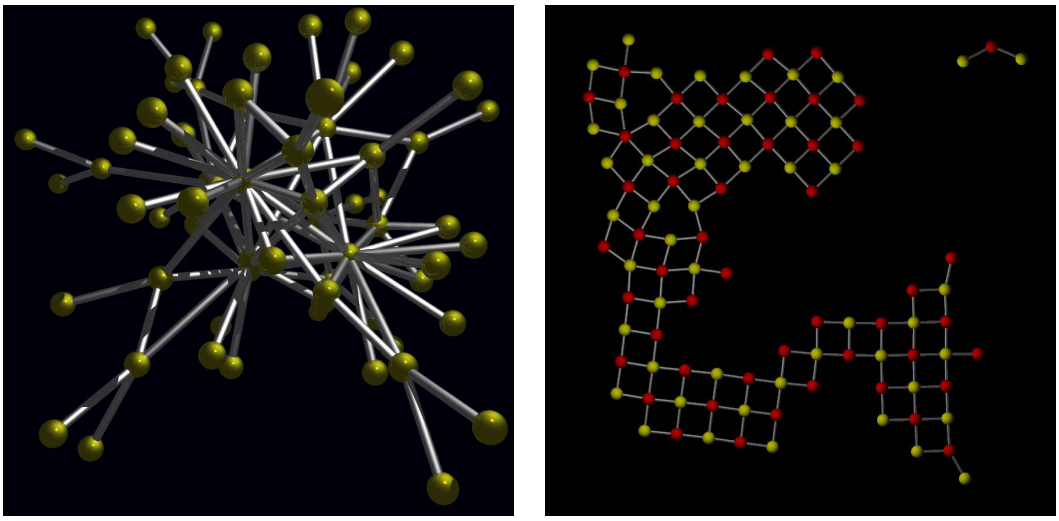


Figure 1: Examples of nanoparticle networks: (left) 3-dimensional arrangement with a scale-free graph structure [6]; (right) Regular 2-dimensional structure obtained in bio-recognition assembly (see below) with equal particle sizes and low coverage.

For the quantitative analysis of complex structures emerging in different assembly processes we have recently introduced a methodology based on mapping of the assemblies onto mathematical graphs (nanoparticle networks) and using the graph theory methods, see Refs. [7, 8, 9, 10]. Two examples of such nanoparticle networks in two- and three-dimensions are shown in Fig. 1. Nodes of these graphs are the nanoparticles, while the links indicate association of the nanoparticle pairs with a certain type of interaction. Gold nanoparticles can be ligated by bio-compatible DNA parts [3], or *functionalized* in another way to promote interaction in a given direction [11], which then affects the assembly process. In the case of charged colloidal particles, as in the case of PMN ceramics $Pb(Mg_{1/3}Nb_{2/3})O_3$ synthesis, ζ-potential can be manipulated by changing the pH of the solvent [13].

The spatial arrangements of nanoparticles of the type in Fig. 1(left) are shown to affect the processes of magnetization reversal [6, 10], which are important in the memory materials (see Fig. 2). Whereas, the planar arrangements of the metallic nanoparticles, as in Fig. 1(right), are relevant for the single-electron tunneling processes in nano-electronics [7, 9]. Both experiments and simulations in [7, 8, 9] show that the current–voltage $I--V$ characteristics strongly depend on the spatial inhomogeneity of the conducting NP films. Specifically, favorable nonlinearity is obtained in the case of stronger inhomogeneity, where the links with large topological betweenness occur. Here in section 2 we will show how several such 2-dimensional structures of the nanoparticles can arise from the self-assembly process which exploits bio-recognition binding and varied parameters. We focus on the binary colloids, where the neutral particles of two different sizes are involved. In section 3 we will also discuss the particle-size effects in the assembly processes of charged colloidal particles in 3-dimensions [13].

 ISBN: 978-84-614-6167-7

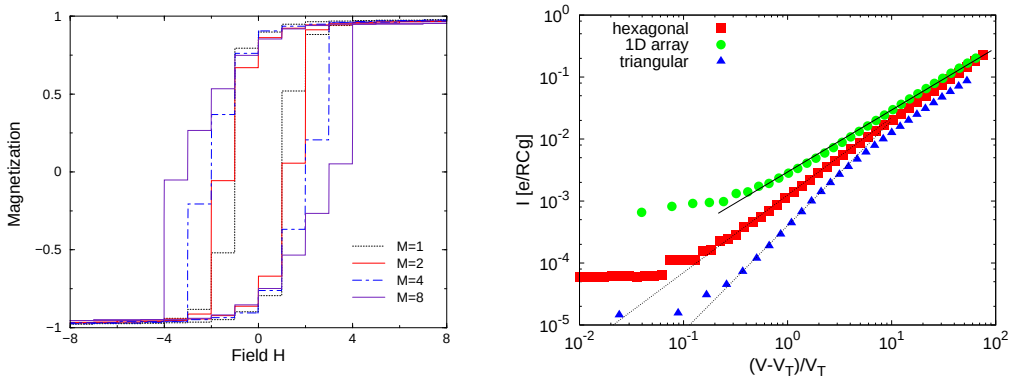


Figure 2: (left) Hysteresis loop in the magnetization-reversal on a scale-free graph, as function of the average node connectivity M [6]. (right) Current-voltage characteristics for single-electron tunneling conduction in nanoparticle films of different geometry.

2 Binary Colloidal Aggregation with Competing Forces

By attaching the biologically compatible strands of the DNA to two different nanoparticles two types of particles, A and B, can be recognized. In addition to carrying different DNA strands, the type A and type B particles in this work are assumed to be of different size, with the radius $R_A = 2.5R_B$. The bio-recognition binding can be represented by an attractive Lennard-Jones potential, with the parameters related with the binding strength (number of DNA base-pairs), the lengths of the attached DNA ligands and the particle radii [9]:

$$V_{A-B}(r) = 4\epsilon \left(\frac{\sigma^{12}}{(r - R_A - R_B)^{12}} - \frac{\sigma^{6}}{(r - R_A - R_B)^{6}} \right). \tag{1}$$

In contrast to A–B interactions, the repulsive potentials apply for the particles of the same type, $V_{X-X}(r) = 4\epsilon' \frac{\sigma'^{12}}{(r-2R_X)^{12}}$ with own parameters ϵ' and σ'^{12}.

Particle diffusion in 2-dimensional plane is simulated by solving the sets of the Langevin equations with these attractive and repulsive pair potentials and in the presence of an external random noise:

$$\nu_i \dot{\mathbf{r}}_i = -\nabla_{\mathbf{r}_i} \sum_j V_{i-j}(|\mathbf{r}_i - \mathbf{r}_j|) + \mathbf{F}_{i,T}, \tag{2}$$

where r_i is the position of ith particle and $\nu_i \equiv 6\pi\eta R_i$ represents the kinetic coefficient different for each particle type, and η is the fluid viscosity coefficient. The stochastic force $F_{i,T}$ originating from the integration over the fluid degrees of freedom, is given by the distribution with the moments $\langle F_{i,T}(t)\rangle = 0$ and $\langle F_{i,T}(t)F_{j,T}(t')\rangle = 6k_BT\nu_i\delta_{i,j}\delta_{t,t'}$.

Details of the numerical implementation can be found in [9]. Here we focus on the structures that emerge after long evolution time (lowest energy) of the assembly in different conditions. The results shown in Fig. 3 correspond to increased particle density, while the relative concentrations of the particle types are kept equal. In the low

 ISBN: 978-84-614-6167-7

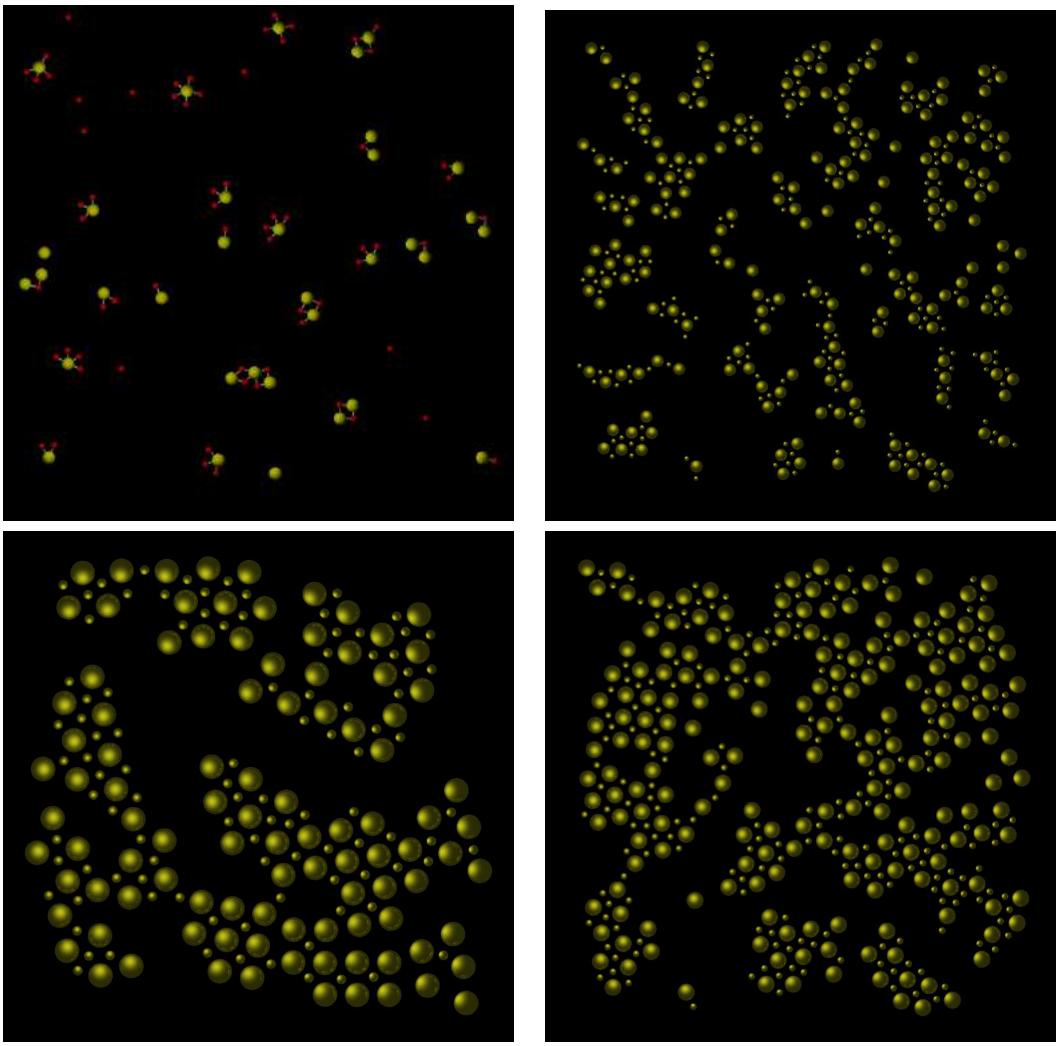


Figure 3: Emergent binary aggregates for equal concentrations and varying density.

particle density, the situation interesting for bio-sensors, a large particle appears to be isolated by six small particles. In the situations with equal concentrations, very small isolated clusters can form, as in Fig. 3 top left, where also the binding links are shown. When the density is increased, but the relative concentrations kept equal, worm-like structures appear and start joining with each other, cf. top right Fig 3. For even larger densities, areas of regular arrangements may occur with voids of different sizes between them, bottom panels in Fig 3. The large particles are cross-linked via small particles, forming a triangular pattern. Whereas, the constraint of equal concentrations prevails small particles to order at a comparable scale. Note that in the case of equal sizes of the particles, the constraint of equal concentrations leads to square lattice arrangements even at very low densities, as in the example shown in Fig. 1 right.

 ISBN: 978-84-614-6167-7

3 Size Effects in the Aggregation of Charged Colloids

As an interesting example of the aggregation of charged colloids, we consider three types of charged colloidal particles, PbO–lead oxide, MHC–magnesium hydroxy carbonate, and Nb_2O_5–niobium oxide, in the relative concentrations and the conditions relevant for the $Pb(Mg_{1/3}Nb_{2/3})O_3$ ceramic synthesis from the aqueous suspension [12, 13]. It has been found [12] that the PMN in pure perovskite phase can be synthesized when pH of the solvent is changed to pH=12.5, compared with the standard pH=11.4, where a disturbing admixture of the pyrochlore phase occurs. With the numerical model of of the system in [13] it has been shown that at pH=12.5 the polarity of the corresponding interactions change such that direct contacts between PbO and Nb_2O_5 particles, and thus the interactions leading to the pyrochlore phase, are prevented. The crucial role in this aggregation process is due to MHC particles [13]. Here we further explore the course of the aggregation at pH=12.5 with the steric effects of the MHC particles when they are prepared in different sizes. In particular, twice smaller and twice larger MCH particles compared with the other two types are considered with the Monte Carlo simulations in 3-dimensions. Snapshots of the emergent clusters are shown in Fig. 4 (MHC are shown by dark/blue color).

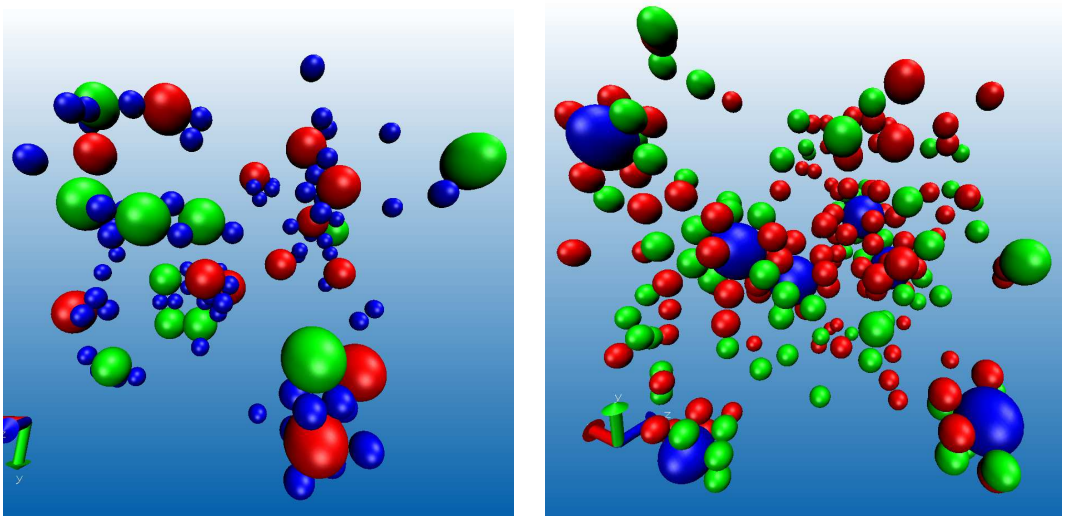


Figure 4: 3-dimensional clusters of ternary aggregates of charged colloidal particles in the model of PMN ceramics assembly at pH=12.5: Case where MHC particles are twice smaller (left) and twice larger (right) compared with PbO and Nb_2O_5 particles.

In addition to the hard-core and the van der Waals potentials, for the colloids of a given size and the type (Hamaker constant), of the key importance are the electrostatic interactions between pairs of charged particles [13]:

$$U_{ij}^{el} = \pi\varepsilon_r\varepsilon_0 \cdot \frac{a_i a_j}{a_i + a_j}\left[(\psi_i + \psi_j)^2 \ln\left(1 + e^{-\kappa h}\right) + (\psi_i - \psi_j)^2 \ln\left(1 - e^{-\kappa h}\right)\right], \qquad (3)$$

where $h = r_{ij} - (a_i + a_j)$ is the surface-to-surface distance between the particles, $\varepsilon_r\varepsilon_0$ is

 ISBN: 978-84-614-6167-7

the dielectric permittivity of the medium(the constant for water $\varepsilon_r = 78.5$ is used). ψ_i is the electrostatic surface potential of the particle i, and κ is the inverse Debye screening length $\kappa = \sqrt{\frac{2N_A e^2 I}{\varepsilon_0 \varepsilon_r kT}}$. Details of the numerical model and the implementation are given in [13]. In view of the above expressions, the steric effects due to different particle sizes is more complex, compared to the case of neutral colloids. We focus on certain features of the aggregation processes when size of the MHC particles is varied.

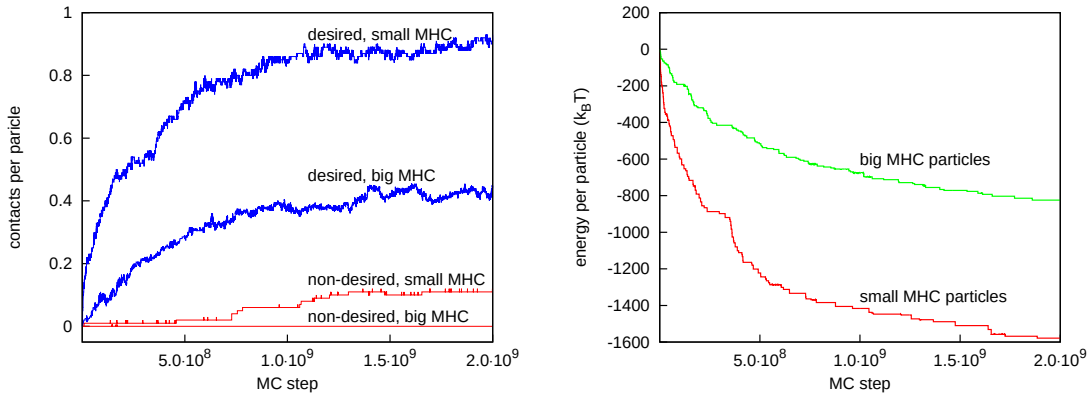


Figure 5: Energy of the assembly (right) and the fraction of "desired" and "non-desired" contacts (left) versus simulation time in the charged particle aggregation corresponding to PMN ceramic synthesis at pH=12.5, when the size of MHC particle size is varied.

In Fig. 5 we show how the number of contacts between different particle types evolves in time and the energy of the assembly. Two curves in each plot are for the situations of smaller/larger MHC particles. As mentioned above, clusters formed at pH=12.5 mostly involve MHC particles preventing direct contact $PbO--Nb_2O_5$. Thus "desired" contacts are $PbO--MHC$ and $MHC--Nb_2O_5$, while all other contact pairs might lead to unwanted chemical reactions or deteriorate the sample quality. Due to fixed concentrations of the three types of particles, the steric effect of smaller/larger MHC particles in preventing such contacts are different. For instance, for the case with smaller MHC particles a small fraction of unwanted contacts remains when the clusters are formed, while the bigger MHC are much more effective, cf. Fig. 5 left. However, from the point of view of the energy, this situation is just the opposite. In this case a large number of particles PbO and Nb_2O_5 remain outside the clusters (cf. Fig. 4).

Acknowledgements

This work has been supported by the Research Agency Program P1-0044 and the Project PR-02485 (Slovenia) and by the National Project MNTRS 141025 (Serbia).

 ISBN: 978-84-614-6167-7

References

[1] P. Moriarty, *Nanostructured materials*, Reports on Progress in Physics **64** (2001) 297-381.

[2] M. P. Pileni, *Nanocrystal self-assemblies: Fabrication and collective properties*, J. Phys. Chem. B **105** (2001) 3358-3371.

[3] C. A. Mirkin, *Programming the assembly of two- and three-dimensional architectures with DNA and nanoscale inorganic building blocks*, Inorg. Chem. **39** (2000) 2258-2272.

[4] A. K. Boal, F. Ilhan, J. E. DeRouchey, T. Thurn-Albrecht, T. P. Russell, and V. M. Rotello, *Self-assembly of nanoparticles into structured spherical and network aggregates*, Nature **404** (2000) 746–748.

[5] C. N. Likos, *Effective interactions in soft condensed matter physics*, Physics Reports **348** (2001) 267–439.

[6] B. Tadić, *From microscopic rules to emergent cooperativity in large-scale patterns*, Ch.12 in *Systems Self-Assembly: Multidisciplinary Snapshots*, eds. N. Krasnogor *et al.*, Elsevier, Amsterdam, 2008.

[7] M. Šuvakov, B. Tadić, *Modelling Collective Charge Transport in Nanoparticle Assemblies*, J. Physics Condensed Matter: Topical Review **22** (2010) 163201-23

[8] M. O. Blunt, M. Šuvakov, F. Pulizzi, C. P. Martin, E. Pauliac-Vaujour, A. Stannard, A. W. Rushforth, B. Tadić, P. Moriarty, *Charge transport in cellular nanoparticle networks*, Nano Letters **7** (2007) 855-860.

[9] M. O. Blunt, A. Stannard, E. Pauliac-Vaujour, C. P. Martin, I. Vancea, M. Šuvakov, U. Thiele, B. Tadić, P. Moriarty, *Patterns and Pathways in Nanoparticle Self-Organization*, chapter in *The Oxford Handbook of Nanoscience and Technology*, volume I, Oxford University Press, 2010.

[10] B. Tadić, K. Malarz, K. Kulakowski, *Magnetization reversal in spin patterns with complex geometry*, Physical Review Letters **94** (2005) 137204.

[11] P. I. Archer, S. A. Santangelo, D. R. Gamelin., *Direct observation of sp-d exchange interactions in colloidal mn2+- and co2+-doped cdse quantum dots.*, Nano Letters **7** (2007) 1037-1049.

[12] G. Trefalt, B. Malič, D. Kuščer, J. Holc, M. Kosec, *Synthesis of $Pb(Mg_{1/3}Nb_{2/3})O_3$ by Self-Assembled Colloidal Aggregation*, J. Am. Ceram. Soc. (2011) DOI:10.1111/j.1551-2916.2011.04443.x.

[13] G. Trefalt, B. Tadić, M. Kosec, *Formation of Colloidal Assemblies in Suspensions for $Pb(Mg_{1/3}Nb_{2/3})O_3$ Synthesis: Monte Carlo Simulation Study*, Soft Matter (2011) DOI:10.1039/C1SM05228D.

 ISBN: 978-84-614-6167-7

Proceedings of the 11th International Conference on Computational and Mathematical Methods in Science and Engineering, CMMSE 2011 26–30 June 2011.

Symmetric Iterative Splitting Method for Non-Autonomous Systems

Gamze Tanoğlu[1] and Sıla Korkut[1]

[1] *Department of Mathematics, Izmir Institute of Technology, Gulbahce Campus, Urla, Izmir,35430, Turkey Izmir Inst*

emails: gamzetanoglu@iyte.edu.tr, silakorkut@iyte.edu.tr

Abstract

The iterative splitting methods have been extensively applied to solve complicated systems of differential equations. In this process we split the complex problem into several sub-problems, each of which can be solved sequentially. In this paper, we develop a symmetric iterative splitting scheme based on the magnus expansion for solving non-autonomous problems. We also study its convergence properties by using the concepts of stability, consistency, and order. Several numerical examples are illustrated to confirm the theoretical results by comparing frequently used methods.

Key words: Iterative scheme, non autonomous system, convergency analysis, magnus series

1 Introduction

The aim of the present paper is to develop and analyze a splitting method for non-autonomous evaluation equation of the form

$$\frac{d}{dt}u(t) = A(t)u(t), \quad t \geq 0 \tag{1}$$

$$u(0) = u_0 \in X \tag{2}$$

on some Banach space X. For solving such non-autonomous system , it is often the case that $A(t) = T + V(t)$, where only the potential operator $V(t)$ is time-dependent and T is the differential operator see [5, 6, 7, 8, 9].

Operator splitting is a widely used procedure in the numerical solution of large systems of partial differential equations. One of the operator splitting methods other than the classical Trotter, Strang splitting is iterative splitting scheme which is based on first splitting the complex problem into simpler differential equations. Then each

 ISBN: 978-84-614-6167-7

sub-equation is combined with the iterative schemes, each of which is efficiently solved with suitable integrators [1, 2, 3, 4].

Some splitting methods have already been used to find numerical solution of the different special non-autonomous system, particularly Hamiltonian ones [15, 17]. It is important to construct such numerical schemes for Hamiltonian dynamics or Schrödinger equations that preserve some important qualitative properties and geometric structure of that solution. In this study, we focus on developing the symmetric iterative scheme. We embed the Magnus expansion [15, 16] which is a popular geometric, an attractive and widely applied method of solving explicitly time-dependent problems, in the solutions of the time dependent split subsystem of the iterative scheme. We consider the time independent split subsystem as an abstract Cauchy problem. Our main focus will be two fold: First, we develop the iterative splitting for non-autonomous problem. Second, its convergence properties are analyzed using the concepts of stability, consistency, and order.

The paper is outlined as follows: In Section 2, the basic idea behind the Magnus method is summarized. In Section 3, the algorithm of the symmetric iterative scheme is presented and its convergence properties are studied. In the last section, several numerical examples are illustrated to confirm our theoretical results and efficiency of the new scheme.

2 Exponential splitting method based on the Magnus expansion

The Magnus integrator was introduced as a tool to solve non-autonomous linear differential equations for linear operators of the form

$$\frac{du}{dt} = A(t)u(t) \ , \tag{3}$$

with solution

$$u(t) = \exp(\Omega(t))u(0) \ . \tag{4}$$

This can be expressed as:

$$u(t) = \mathcal{T}\left(\exp(\int_0^t A(s)\ ds\right) u(0) \ , \tag{5}$$

where the time-ordering operator $\mathcal{T}$ is given in [18].

The Magnus expansion is defined as:

$$\Omega(t) = \sum_{n=1}^{\infty} \Omega_n(t) \ , \tag{6}$$

 ISBN: 978-84-614-6167-7

where the first few terms are [15]:

$$\begin{aligned}
\Omega_1(t) &= \int_0^t dt_1 A_1 \\
\Omega_2(t) &= \frac{1}{2}\int_0^t dt_1 \int_0^{t_1} dt_2 [A_1, A_2] \\
\Omega_3(t) &= \frac{1}{6}\int_0^t dt_1 \int_0^{t_1} dt_2 \int_0^{t_2} dt_3 ([A_1,[A_2,A_3]+[[A_1,A_2],A_3]) \\
&\cdots\cdots \qquad \text{etc.}
\end{aligned} \tag{7}$$

where $A_n = A(t_n)$. In practice, it is more useful to define the nth order Magnus operator

$$\Omega^{[n]}(t) = \Omega(t) + O(t^{n+1}) \tag{8}$$

such that

$$u(t) = \exp\left[\Omega^{[n]}(t)\right] u(0) + O(t^{n+1}). \tag{9}$$

Thus the second-order Magnus operator is

$$\begin{aligned}
\Omega^{[2]}(t) &= \int_0^t dt_1 A(t_1) \\
&= e^{\frac{t(A(t))+(A(0))}{2}} + O(t^3).
\end{aligned} \tag{10}$$

3 Symmetric Iterative Splitting Method and its Convergence Analysis

3.1 Derivation of the Algorithm for Iterative Splitting

Let us consider initial value problem (IVP) given in (1) with the initial condition (2) on the time interval $[0, t_{end}]$ where $t_{end} \in R$. We assume for $A(t)$ a two-term splitting

$$T + V(t).$$

Let us divide the integration interval $[0, t_{end}]$ in n equal parts by the points $t_0, t_1, ..., t_n$, where the length of each interval is $h = t_{j+1} - t_j = t_{end}/n, j = 0, 1..n$. The approximated solution and exact solution at time $t = t_n$ are $U(t_n)$ and $u(t_n)$, respectively.

Our technique are close to that used in [11]. We apply second order iterative process described as below on each subinterval $[t_j, t_{j+1}]$,

$$\dot{u}_1 = Tu_1 + V(t)U(t_j) \quad u_1(t_j) = U(t_j) \tag{11}$$

$$\dot{u}_2 = Tu_1 + V(t)u_2 \quad u_2(t_j) = U(t_j) \tag{12}$$

where $u_2(t_j) = U(t_j)$ denotes the numerical approximation to the true solution $u(t_j)$ at the time $t = t_j$ and $U(t_0) = u_0$. The formal solution of the sub equations given in (11) and (12) on the time interval $[t, t + h]$ can be written by

$$u_i(t+h) = \Phi_i(t+h,t)\, U(t) + \int_t^{t+h} \Phi_i\,(t+h,s) F_i(s)\, ds, \quad i = 1, 2$$

 ISBN: 978-84-614-6167-7

where $F_1 = V(t)U(t)$ and $F_2 = Tu_1(t+h)$. Φ_1 is the fundamental set of solution for sub-equation (11) given by

$$\Phi_1(t,s) \quad = \quad e^{(t-s)T}.$$

Φ_2 is the fundamental set of solution for sub equation (12) given by

$$\Phi_2(t,s) \quad = \quad e^{\frac{h}{2}[V(t+h)+V(t)-V(s+h)-V(s)]}$$

which is the second order approximation of the magnus series given in equation (10). Next we use the trapezoidal rule to approximate the integral

$$\int_t^{t+h} \Phi_i \, F_i \, ds \quad = \quad \frac{h}{2}[F_i(t+h) + \Phi_i(t+h,t)F_i(t)] + O(h^3) \tag{13}$$

Note that $\Phi_i(t+h, t+h) = I$. After combining approximation (13) with the iterative schemes (11), (12) and rearranging expressions, we get the first order approximation

$$u_1(t_n+h) \quad = \quad e^{Th}[U(t_n) + \frac{h}{2}V(t_n)U(t_n)] + \frac{h}{2}V(t_n+h)U(t_n) \tag{14}$$

and the second order approximation,

$$u_2(t_n+h) \quad = \quad e^{\frac{h}{2}[V(t_n+h)+V(t_n)]}[U(t_n) + \frac{h}{2}Tu_1(t_n)] + \frac{h}{2}Tu_1(t_n+h) \tag{15}$$

where $U_{n+1} = u_2(t_n+h)$. Repeat this procedure for next interval until the desired time t_{end} is reached.

Proposition 3.1: New iterative scheme preserve the time-symmetry property.

Proof 3.1: The time- symmetry preservation can be easily seen by interchanging $t_{n+1}, U(t_{n+1}), h$ by $t_n, U(t_n), -h$,respectively.

3.2 Convergency Analysis

In the present section, we analyze the convergence behavior of the symmetric iterative scheme derived in the previous section. We assume that T is unbounded and $V(t)$ is bounded operator. We define a operator norm as $\|.\|_{X\leftarrow X}$ in a (complex) Banach space $(X, \|.\|_{X\leftarrow X})$

In our proofs, we will use the following assumptions:

Hypothesis 1 : Suppose that closed linear operator $A(t) : D \rightarrow X$ where D is dense subset of X and that $A(t)$ is uniformly sectorial for $0 \leq t \leq t_{end}$. Then, there exist constants $a \in \mathbb{R}, 0 < \varphi < \pi/2$,and $M_1 \geq 1$ such that $S_\varphi(a) = \{\lambda \in \mathbb{C} :| arg(a-\lambda) | \leq \varphi\} \cup \{a\}$,

$$\|(\lambda I - A(t))^{-1}\|_{X\leftarrow X} \leq \frac{M_1}{|\ (a-\lambda)\ |} \quad \text{for any } \lambda \in \mathbb{C} \setminus \mathrm{S}_\varphi(\mathrm{a}). \tag{16}$$

 ISBN: 978-84-614-6167-7

Then for fixed $0 \leq s \leq t_{end}$, the analytic semigroup $e^{tA(s)}$satisfy $\| e^{tA(s)} \| \leq Me^{\omega t}$ for some constants $\omega < 0$ and $M \geq 1$. Our general references on semigroups are [12, 13].

Hypothesis 2 : Let $D(T) = D(A(t))$. We note that T is linear closed operator and that generates a strongly continuous semigroup e^{tT} on X. By semi group property, we assume $\| e^{Tt} \| \leq 1$.

Hypothesis 3 : We assume that $V(t)$ is bounded linear operator on X. Then we get $e^{\Omega_V(t)} \leq e^{\|V(t)\|}$ where $\Omega_V(t) \approx \Omega_2(t)$ with the help of the equation (10). As the convergence of Magnus expansion is guaranteed if $\| \Omega(t) \| < \pi$. The details can be found in [14]. There exists inverse of the fundamental set of solution of $e^{\Omega_V(t)}$ if we have $Tr(V(t)) = 0$.

Lemma 4.1 : Let T be an infinitesimal generator of a $\mathcal{C}_0$ semigroup $S(t)$, $t \geq 0$. Let $t_{end} > 0$. If for any $V(t)U \in \mathcal{D}(T)$ satisfying $V(t)U, TV(t)U \in \mathbf{C}^1([0, t_{end}]; X)$ then the solution of problem satisfies $u(t) \in \mathcal{D}(A^2(t))$ for $0 \leq t \leq t_{end}$ whenever $u_0 \in \mathcal{D}(A^2(t))$, and we have

$$\sup_{0 \leq t \leq t_{end}} \|T^i u(t)\| \leq E_i(t_{end}), \quad i = 0, 1, 2 \tag{17}$$

where E_i depends on the specific choice of $t_{end}, T, V(t)U$ and u_0. For the detailed proof see [10].

Hypothesis 4 : We assume that there are non-negative constants $\tilde{C}, R$ with

$$\sup_{0 \leq t \leq t_{end}} \|V(t)\| \leq \tilde{C}$$

$$\|u\| \leq R \quad \text{on} \quad 0 \leq t \leq t_{end}$$

Under these conditions, the following convergence analysis is obtained for the proposed symmetric iterative scheme.

Proposition 4.1 : The symmetric iterative splitting is first order if we consider only one iteration given in (11) with the error bound

$$\|u(h) - U(h)\| \quad \leq \quad Kh^2 \tag{18}$$

Here K only depends on $\tilde{C}, R, E_1(t_{end})$.

Proposition 4.2 : The symmetric iterative splitting is the second order if we consider two iterations given in (12) with the error bound

$$\|u(h) - U(h)\| \quad \leq \quad \tilde{K}h^3 \tag{19}$$

Here $\tilde{K}$ only depends on $\tilde{C}, R, E_1(t_{end})$.

 ISBN: 978-84-614-6167-7

Proposition 4.3 : The symmetric second order iterative splitting scheme is stable on $[0, t_{end}]$ with the bound

$$\| U^n \| \leq e^{t_{end}\tilde{C}} \|u_0\| + h e^{2h\tilde{C}} E_1(t_{end}) \left(\frac{1 - e^{t_{end}\tilde{C}}}{1 - e^{h\tilde{C}}}\right).$$

Proposition 4.4 : The global error of iterative splitting is bounded by

$$\|U^n(h) - u^n(h)\| \quad \leq \quad G\, h^2$$

Here G only depends on $t_{end}, \| u_0 \|, R$ and $\tilde{\mathrm{C}}$.

Proof : Directly from telescoping identity.

4 Numerical Examples

4.1 1. Mathieu equation

We first consider the Mathieu equation,

$$q'' + (\omega^2 - \varepsilon \cos(t))q \quad = \quad 0 \tag{20}$$
$$\tag{21}$$

By redefining the variables as $q(t) = q_1(t)$ and $\dot{q}(t) = q_2(t)$, and $u(t) = (q_1(t), q_2(t))$,then the time dependent oscillator corresponds to

$$A(t) \quad = \quad \begin{pmatrix} 0 & 1 \\ -(\omega^2 - \varepsilon \cos t) & 0 \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ -\omega^2 & 0 \end{pmatrix} + \begin{pmatrix} 0 & 0 \\ \varepsilon \cos t & 0 \end{pmatrix} \equiv T + V(t),$$

We take as initial condition $q(0) = 1.75$ and $\dot{q}(0) = 0$, integrate up to $t = 10$ and measure the average error for different time steps.

h	Iterative Splitting/Order	Lie Trother/Order
0.1	0.0610	0.1015
0.01	0.0066 (0.9658)	0.0105 (0.9853)
0.001	6.6819e-004 (0.9946)	0.0011 (0.9798)

Table 1: Comparison of errors for several h on $[0, 10]$ interval with various methods where $\omega = 0.6$ and $\varepsilon = 0.3$. The expected order is 1

 ISBN: 978-84-614-6167-7

Another comparison is for the second order methods,

h	2nd order Iterative Splitting/ order	Strang Splitting/ order	SWS/ order
0.1	9.8067e-004	0.0011	0.0062
0.01	8.3542e-006 (2.0696)	1.0839e-005 (2.0064)	6.3187e-005 (1.9918)
0.001	8.2197e-008 (2.0070)	1.0801e-007 (2.7672)	6.3309e-007 (1.9992)

Table 2: Comparison of errors for different h on $[0, 10]$ interval with several methods where $\omega = 0.6$ and $\varepsilon = 0.3$. Accepted exact solution is fourth order magnus expansion. The expected order is 2

The numerically observed order in the discrete L^∞ norm is approximately 1 in table 1 which is supported by *proposition 4.1.* In addition, *proposition 4.2* predicts order 2. This number is in perfect agreement with table 2. We also observed in table 2 that second order proposed iterative splitting scheme is more efficient than not only Strang splitting but also symmetrically weighted splitting.

4.2 Schrödinger Equation

Another experiment is time dependent schrodinger equation as following form,

$$i\hbar\frac{\partial\psi(x,t)}{\partial t} = \hat{H}\psi(x,t)$$

where $\psi(x,t)$ denotes the probability amplitude for the particle to be found at position x at time t and $\hat{H}$ is the Hamiltonian operator for a single particle in a potential.

In our study we choose one-dimensional harmonic oscillator in the finite time interval $t \in [0, t_{end}]$ has the form

$$\begin{aligned} i\frac{\partial\psi(x,t)}{\partial t} &= \big(-\frac{1}{2}\frac{\partial^2}{\partial x} + \frac{\omega^2(t)(x^2-1)}{2}\big)\psi(x,t) \\ \psi_0(x) &= \sqrt[4]{\frac{1}{\pi}}\exp\big(-\frac{1}{2}x^2\big) \end{aligned} \tag{22}$$

with $\omega^2(t) = 4 - 3e^{-t}$.
We take into account the system as following form,

$$\begin{pmatrix} \dot{u} \\ \dot{v} \end{pmatrix} = \begin{pmatrix} 0 & A(t,x) \\ -A(t,x) & 0 \end{pmatrix}\begin{pmatrix} u \\ v \end{pmatrix}$$

where $\psi(x,t) = u(x,t) + iv(x,t)$,then consider the splitting methods with ODE system split in the form,T corresponds to spatial derivative $\partial^2\psi(x,t)/\partial x^2$, we use the second order center difference scheme in order to approximate it, thus we get $2N \times 2N$ system.

$$\begin{pmatrix} \dot{u} \\ \dot{v} \end{pmatrix} = \begin{pmatrix} 0 & T \\ -T & 0 \end{pmatrix}\begin{pmatrix} u \\ v \end{pmatrix} + \begin{pmatrix} 0 & V(t,x) \\ -V(t,x) & 0 \end{pmatrix}\begin{pmatrix} u \\ v \end{pmatrix}$$

 ISBN: 978-84-614-6167-7

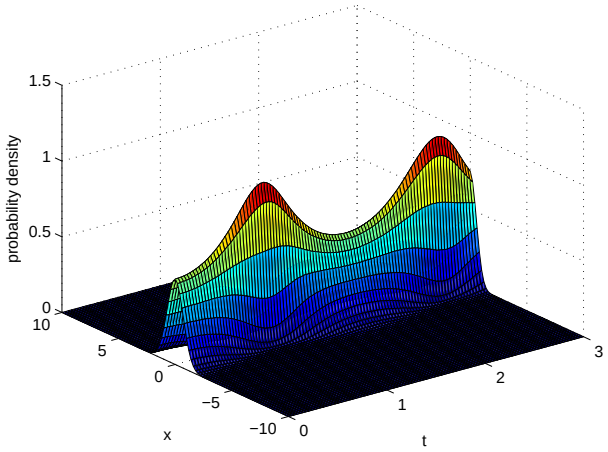


Figure 1: The probability of density,$|\Psi(x,t)|^2$, for the equation in (22).

For exhibited figure, we suppose that the system is defined in the interval $x \in [-10, 10]$, which is split into $M = 100$ parts of length $\Delta x = 0.2$. We integrate the system using proposed method with the time-step size $\Delta t = 0.03$ up to final time $t = 3$.

5 Conclusions and Discussions

We have developed the new symmetric iterative splitting scheme for non-autonomous systems with the help of the Magnus expansion. This new scheme is applicable for obtaining the numerical solution of the non-autonomous systems for example Schrödinger equation in quantum mechanics since it preserves the time symmetry. We also investigated the convergence properties of the new scheme by using the semigroup approaches. We confirm the theoretical results on a test problem. The method also provides the higher order accuracy in approximate solution with increasing number of iteration steps. Finally numerical experiments reveal that our proposed method is efficient and easily adapted to numerically solve for such problems.

References

[1] J. Geiser, *Decomposition methods for differential equations : Theory and application*, *CRC Press, Taylor and Francis Group*, December, (2008).

[2] I.Farago & J.Geiser, *Iterative Operator-Splitting methods for Linear Problems*, *International Journal of Computational Science and Engineering*, 3 (2007), pp. 255-263.

[3] I. Farago, *A modified iterated operator splitting method*, Vol. 32, (8), Applied Mathematical Modelling, August 2008, pp. 1542-1551.

 ISBN: 978-84-614-6167-7

[4] J. Geiser, *Iterative operator-splitting methods with higher-order time integration methods and applications for parabolic partial differential equations*, Journal of Computational and Applied Mathematics, 217, (2008), pp. 227-242.

[5] D. Baye, G. Goldstein and P. Capel. *Fourth-order factorization of the evolution operator for time-dependent potentials.* Phys. Letts, A 317, 337 (2003).

[6] G. Goldstein and D. Baye. *Sixth-order factorization of the evolution operator for time-dependent potentials.* Phys. Rev, E 70, 056703 (2004).

[7] S.A. Chin and C.R. Chen. *Gradient symplectic algorithms for solving the Schrödinger equation with time-dependent potentials.* Journal of Chemical Physics, 117(4), 1409-1415 (2002).

[8] V.C. Aguilera-Navarro, G.A. Estévez and R. Guardiola. *Variational and perturbative schemes for a spiked harmonic oscillator.* J. Math. Phys. **31**, 99 (1990).

[9] S.A. Chin and P. Anisimov. *Gradient Symplectic Algorithms for Solving the Radial Schrödinger Equation.* J. Chem. Phys. 124, 054106, (2006).

[10] M. Bjørhus. *Operator splitting for abstract Cauchy problems.* IMA journal of Numerical Analysis. 18, 419-443, (1998).

[11] S. Blanes and E. Ponsoda. *Exponential integrators for non-homogeneous linear initial and boundary value problems.* Journal of Numerical Analysis ,1-24, (2009)

[12] C. González, A. Ostermann, M. Thalhammer. *A second-order Magnus-type integrator for nonautonomous parabolic problems.* Journal of Computational and Applied Mathematics,189, 142-156 (2006)

[13] A. Bátkai, P. Csomós, B. Farkas, G. Nickel *Operator splitting for non-autonomous evolution equations.* Journal of Functional Analysis, 2163-2190 (2011).

[14] C. Moan and J. Niesen. *Convergence of the Magnus Series.* Found Comput Math, 8, 291?301(2008)

[15] S. Blanes, F. Casas, J.A. Oteo and J. Ros. *The Magnus expansion and some of its applications.* arXiv.org:0810.5488 (2008).

[16] S. Blanes and P.C. Moan. *Fourth- and sixth-order commutator free Magnus integrators for ,linear and nonlinear dynamical systems.* Applied Numerical Mathematics, 56, 1519-1537 (2006).

[17] S. Blanes and P.C. Moan. *Splitting Methods for Non-autonomous Hamiltonian Equations.* Journal of Computational Physics, 170, 205-230 (2001).

[18] F.J. Dyson. *The radiation theorem of Tomonaga.* Swinger and Feynman, Phys. Rev., 75, 486-502 (1976).

 ISBN: 978-84-614-6167-7

Proceedings of the 11th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2011
26–30 June 2011.

Computational Methods for Single Molecule Charge Transport (extended abstract)

Joseph M. Thijssen[1], Christopher J. O. Verzijl[1], Fatemeh Mirjani[1] and Johannes S. Seldenthuis[1]

[1] *Kavli Institute of Nanoscience, Delft University of Technology*

emails: J.M.Thijssen@tudelft.nl, C.J.O.Verzijl@tudelft.nl, F.Mirjani@tudelft.nl, J.S.Seldenthuis@tudelft.nl

Abstract

In this talk, we review several computational methods which are used in the study of charge transport through molecular devices. These methods are either implemented as new software within existing quantum chemistry codes, or based on models which use results of quantum chemical calculations as input for their key parameters. In particular, we describe our implementation of the NEGF method within the Amsterdam density functional theory (ADF) code, which enables us to perform transport calculations for large molecules. Furthermore, we describe calculations on phenomena where vibrational excitations of molecules influence the charge transport. Finally, we comment on methods we developed for analysing transport in the weak-coupling limit which is relevant for large class of experiments. *Key words: Density functional theory, non-equilibrium Green's function methods*

1 Introduction

Molecular electronics is a rapidly developing field with great promise for future application. This promise rests upon two considerations: first, the size of molecules defines a minimum device size for logic applications, and this minimum comes within reach if molecules could be 'wired up' in a controled fashion. The second consideration is the versatility in the behaviour of the 'active element', as there exist very many different molecules with many different properties, which enables us to vary the structure and chemical composition of the molecules in order to have the device exhibit a particular behaviour. For example, the sensitivity of the molecular electronic or nuclear structure on incident light or ambient

 ISBN: 978-84-614-6167-7

temperature enables them to be used as sensors. For a review of the different phenomena and transport mechanisms, see [1]. These phenomena and behaviours may be studied using a variety of theoretical methods, a few of which will be described in this talk, along with examples and results.

Developments in the realization of devices in which charge transport through single molecules can be studied, have caused a rapid expansion and extension of these methods to include the effects of electron-nuclear coupling and correlations, e.g. the exchange correlation responsible for the Kondo effect. Another interesting development in this field is the large body of quantum chemistry and band structure software which has proven very useful in the study of quantum states in molecular and bulk solid systems. The challenge in the field has been and still is to combine theoretical concepts and computational quantum chemistry into tools which have predictive power. There is however in the field of single-molecule electronics an inherent limitation to this predictive power in that the details of the molecule-contact interface are usually not known: standard nanometer probes, such as STM or AFM are not capable of providing this information, as the molecules are usually trapped inside a nano-meter wide gap of metal contacts, preventing a tip from giving us the desired information. Even if the probe itself is used as a contact, which is the case in STM transport, the details of where and how the molecule is attached to the substrate remain unknown. As a consequence, we should sometimes be satisfied with predicting trends rather than reliable a priori predictions for the value of, say, a current at a specified voltage.

2 Coomputational methods

2.1 Nonequilibrium Green's functions

We first review the implementation and the results of a NEGF method which was programmed within the Band module of the Amsterdam Density Functional (ADF) commercial quantum chemistry code. In the NEGF method, the current is calculated from a Green's function which is calculated as $G(E) = 1/(E - H - \Sigma)$, where H is the DFT Hamiltonian and Σ is the so-called self-energy, which is non-hermitian and which accounts for the addition to and removal from electrons of the system. These self-energies contain all the information from the (bulk) contacts. This information is encoded in the Green's function of these contacts, and is calculated in our code from a tight-binding representation of the Hamiltonian, which we fit to the DFT Hamiltonian obtained from a periodic calculation. The Hamiltonian H is restricted to the molecule plus a few layers of contact-metal, and this structure is called the 'extended molecule'.[2]

Some special features of our implementation which distinguish it from the existing ones are

- The possibility to vary the periodicity of the calculation. Band allows for periodicity in

 ISBN: 978-84-614-6167-7

0, 1, 2, or 3 dimensions. Therefore we perform a bulk electronic structure calculation for the contacts, and in the device calculation, involving a molecule, the periodicity is turned off. The same basis set is used in all these calculations.

- The fact that the device calculation is non-periodic allows us to study the effect of a back-gate potential which is not compliant with a periodic unit cell.

- We have included a delicate scheme for fine-tuning the offset between the chemical potential in the bulk contact and the zero-potential level at the molecule.

The code furthermore allows for various LDA and GGA exchange correlation potentials to be used.

In figure 1, the results of a calculation for a porphyrin molecule sandwiched in between two gold electrodes are shown. The picture showing the structure also shows the states corresponding to the HOMO orbital on the molecule. The transmission is shown for this junction in figure 1b.

An example of a trend observed is the dependence of the levels on the distance between the electrodes, which can be varied in a break junction setup. This is currently under investigation.

2.2 Rate equations and vibrational modes

The NEGF/DFT method performs quite well in the case where the orbitals through which the transport takes place are far (in chemical potential) from the chemical potential of the bulk source and drain contact (off-resonant transport), or in the strong coupling limit. In the latter case, the charge on the molecule remains more or less constant and is in general a fractional number of electron charges. In the weak coupling limit, the charge fluctuates, as in that case the charging energy exceeds the coupling energy, which restricts the charge on the molecule to integer values. The standard method for analysing transport in the weak coupling regime is the method of rate equations. This method gives insight in the average occupation on the molecule and in the average current flowing through it. However, it does not predict the line broadenings associated with the transition between one charge state and the next due to quantum fluctuations. Thermal linebroadening of the lineshapes is accounted for, however.

For electron transport, the rate equation method is very straightforward and leads to qualitatively correct results. Richer behaviour is however observed if transitions between different vibrationally excited states of the molecule are included into the picture. The electronic and vibrational excitations are related through the electron-phonon coupling. In the weak coupling limit, the electron-nuclear coupling is responsible for the occurrence of so-called *Frank-Condon* factors in the transition rates between the 'vibronic' states (a vibronic

 ISBN: 978-84-614-6167-7

state denotes the combined electronic and vibrational state of the molecule). In a three-terminal device in the weak coupling limit, where the potential inside the molecule can be shifted up and down by tuning a gate voltage, these factors give rise to vibrational excitation lines that are sometimes visible in the differential conductance (dI/dV) as a function of bias and gate voltage (the bias voltage is the difference between the source-and drain potential).

The vibrations that are observed in this way are normal modes of the molecule. However, we do not observe them for every normal mode – a crucial factor for the visibility of a particular mode is the size of the Frank-Condon factor: a large Frank-Condon factor, sometimes, but not always, corresponding to a large electron-phonon coupling, is necessary. We calculate the Frank-Condon factors for the molecule suspended between the gold contacts and use them in a system of rate equations for which we find the stationary solutions. In this way we are able to obtain a good match with the experimentally observed vibrational spectra. Analysis in this fashion allows us to unambiguously identify a particular type of molecule in the experiment – it is a molecular-transport version of the vibrational fingerprinting technique. In figure 2, the results for a particular example are shown. The current passing through an oligo-phenylene-5 (OPV5) molecule was measured as a function of bias and gate voltage, giving rise to a 'Coulomb diamond', a lozenge-shaped region where the current is suppressed as a result of the fact that there is no discrete level which can be occupied on the molecule and which aligns in energy with the occupied states in the source and with unoccupied states of the drain. In the conducting region just outside of the diamond, we see many lines, corresponding to vibrational modes that are occupied provided the bias is high enough. From these lines the vibrational frequencies can be determined. These frequencies are compared with the ones calculated to show up as visible lines – see figure 2. Note that the molecule has many more vibrations than showing up in the experiment – ony the modes with a significant Frank-Condon factor show up in the transport measurements. An interesting aspect of the comparison in figure 2 is that the calculation in which the molecule is attached to the gold contacts, gives much better agreement with the data than the calculation where that was not the case.

Rate equation techniques can also be used for studying electroluminescence – this is the phenomen where electrons travelling through a molecule, weakly coupled to the electrodes, have two possible orbitals they can occupy on their way. If an electron decays from the higher one of these two the lower, a photon may be emitted. The two electronic orbitals involved both have their own vibrational spectrum, and the optical spectrum of the emited photons carries information about these spectra. Again, we have obtained good agreement with experimental data; see Ref. [5]

 ISBN: 978-84-614-6167-7

3 Weak and intermediate coupling

In the intermediate coupling limit and at low temperatures, the rate equation method is inadequate. A better method for calculating current/voltage characteristics is to use many-body nonequilibrium Green's function techniques. We have developed such a technique [6] which starts with a series of DFT ground calculations for different charge and spin states, and derives a spectrum of the molecule with Coulomb interaction parameters between all possible levels. The *number* of chemical potentials and interaction parameters of the final model is precisely equal to the number of ground state DFT calculations possible for the molecule when the number of electrons and their total spins is varied. For a model system, this method yields very good agreement with density matrix renormalisation group (DMRG) methods applied to interacting spinless and spin-1/2 Hubbard chains. For these model, we used an L(S)DA parametrization presented in refs. [7, 8] and as LDA parametrization based on a Bethe Ansatz solution. We show these results in figure 3 for spinless fermions, which provides a nice system for comparison, as DMRG can be applied to them with realistic computer resources.

The method we use is based on the assumption that the weakly coupled system can be viewed as a system composed of orbitals at chemical potentials ϵ_i with for each of these orbitals an intra-Coulomb interaction between spin-up and -down electrons, and an inter-Coulomb interaction between two different levels. Whether this is a correct representation is somewhat debatable. For example, a system consisting of two quantum dots in series, does not allow for such a description in terms of two levels with Coulomb interactions. How well the eigenvalues of the orbital model with Coulomb interactions represents the original system depends on the values of the hopping parameter and Coulomb interactions in the original model. The picture of molecular orbitals and Coulomb interaction inside and between them is however successful, and widely used in chemistry when interpreting the behaviour of electrons. A further refinement of the method would need more information than ground state energies can give us. In some cases it is possible to obtaine extra information by inspecting the DFT eigenvalues. Consider for example a molecule which consists of two symmetric moieties which are separated by a tunneling barrier. In that case, there is a bonding and an anti-bonding state. The difference in energy between these the Kohn-Sham eigenvalues of these bonding and anti-bonding states enable us to find the electron hopping integral between the two moieties. In that case, the calculation can be redone using a more involved analysis. Our preliminary results in this direction are encouraging.

Of course, the method becomes really useful when it can be implemented in an *ab initio* context. We are currently exploring ways for doing that within constrained DFT (C-DFT). In this technique, part of the system has a fixed charge and spin, and for the Hilbert space with that restriction imposed, a DFT calculation is carried out. The hope is that these extensions will enable us to bridge the gap between the weak strong coupling regime.

 ISBN: 978-84-614-6167-7

4 Conclusion and outlook

The field of molecular electronics is promising for future technology, but poses major theoretical challenges. It is our aim to be able to predict the current-voltage characteristics of a molecular junction. Apart from uncertainties related to the molecule-contact interface, which can only be overcome when the experiments allow for better characterization or when larger series of well reproducible data are supplied, these challenges lie in dealing with the complexity of the problem of charge transport through a molecule. The transition between the weak and strong coupling regime is a major such difficulty, and the inclusion of the coupling between electronic states and vibrational modes is another one. The strong-coupling regime so far is the most successfully treated case. We have developed a new NEGF implementation within the commercial ADF-Band code. We have successfully addressed the problem of theoretically describing vibrational excitations of molecular junctions in the weakly coupled regime. In that case, our methods can be used for 'fingerprinting', i.e. unambiguously identifying a molecule present in the junction. We are in the process of dealing with the intermediate coupling regime, which turns out particularly challenging. Inclusion of vibrational modes into the intermediate coupling regime is a further goal which we shall pursue in the next few years.

Acknowledgements

We appreciate many useful discussions with our colleagues Herre van der Zant (TU Delft), Mark Ratner (Northwestern University), Ferdinand Evers and Peter Schmitteckert (both at KIT Karlsruhe).

We acknowledge financial support from the Stichting FOM (project 86), from the EU FP7 programme under the grant agree- ment SINGLE, and the Stichting Nationale Computerfaciliteiten (National Computing Facilities Foundation, NCF, projects mp-06-111, SH-158-09, SH-180-10) for the use of supercomputer facilities, with financial support from the Nederlandse Organisatie voor Wetenschappelijk Onderzoek (Netherlands Organization for Scientific Research, NWO).

References

[1] J. M. Thijssen and H. S. J. van der Zant, *Charge transport and single-electron effects in nanoscale systems*, Phys. Stat. Sol. (b), **245**, (2008) 1455–1470

[2] Y. Xue and M. A. Ratner, *Microscopic study of electrical transport through individual molecules with metallic contacts. I. Band lineup, voltage drop, and high-field transport*, Phys. Rev. B, **68**, (2004) 115406(18 pages)

[3] J. S. Seldenthuis, H. S. J. van der Zant, M A. Ratner, and J M. Thijssen, *Vibrational Excitations in Weakly Coupled Single-Molecule Junctions: A Computational Analysis* , ACS Nano **2** (2008) 1445–1451

[4] E. A. Osorio, K. ONeill, N. Stuhr-Hansen, O. F. Nielsen, T. Bjrnholm, and H. S. J. van der Zant, *Vibrational Excitations in Weakly Coupled Single-Molecule Junctions: A Computational Analysis* , Adv. Mater. **19** (2007) 281–285

[5] J. S. Seldenthuis, H. S. J. van der Zant, M. A. Ratner, and J. M. Thijssen, *Electroluminescence spectra in weakly coupled single-molecule junctions*, Phys. Rev. B, **81**, (2010) 205430(9 pages)

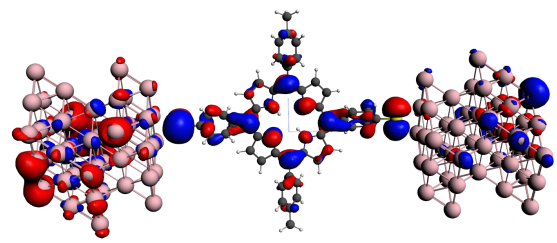
[6] F. Mirjani and J. M. Thijssen, *Density functional theory based many-body analysis of electron transport through molecules* Phys. Rev. B, **83**, (2011) 035415(11 pages)

[7] N. A. Lima, M. F. Silva, L. N. Oliveira and K. Capelle, *Density-Functionals not Based on the Electron Gas: Local-Density Approximation for a Luttinger Liquid* Phys. Rev. Lett., **90**, (2003), 146402(4 pages)

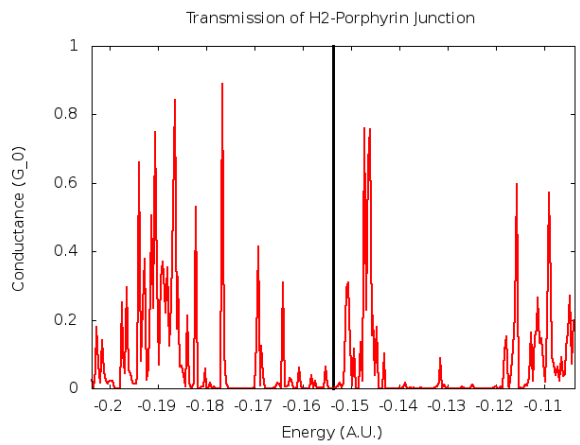
[8] V. V. Franca, D. Vieira and K. Capelle, *Analytical parametrization for the ground-state energy of the one-dimensional Hubbard model* arXiv 1102.5018v1 (2011) 4 pages.

[9] A. Branschädel, G. Schneider and P. Schmitteckert, *Conductance of inhomogeneous systems: Real-time dynamics*, Ann. Physik, **522**, (2010) 657–678

 ISBN: 978-84-614-6167-7



(a)



(b)

Figure 1: (a) The geometry of a gold-porphyrin-gold junction, with a HOMO state visualised which has a weight on the molecule as wall as in the contact. (b) Transmission $T(E)$ for the junction shown in (a). The energy integral over the transmission gives the current in units of the conductance quantum.

 ISBN: 978-84-614-6167-7

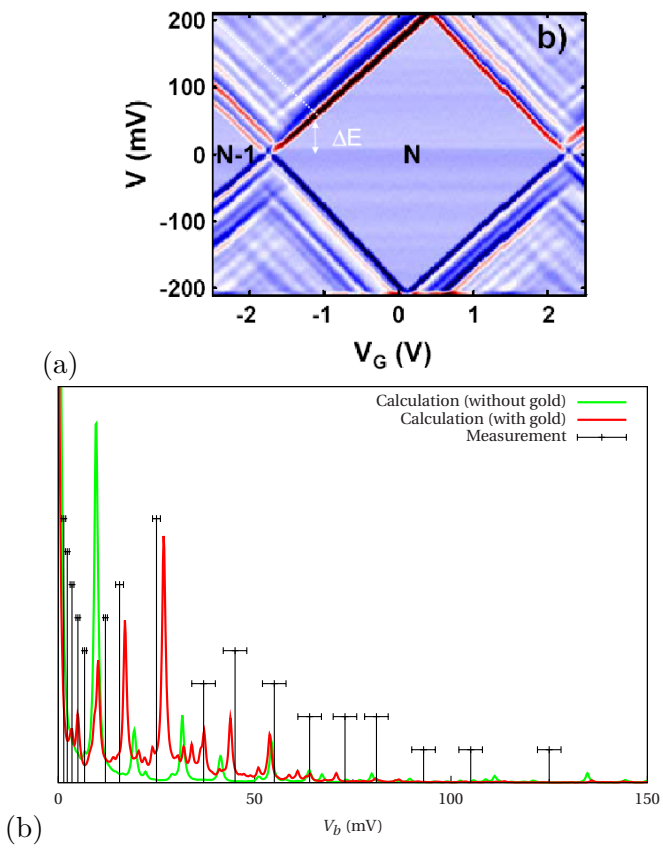


Figure 2: Comparison of the experimental vibrational spectrum of OPV5 (a) with results of rate equations based on vibrational frequencies and Frank-Condon factors calculated using the Amsterdam Density Functional (ADF) quantum chemistry code(b). The experimental plot shows the differential conductance dI/dV as a function of the gate and bias voltage.[4]

 ISBN: 978-84-614-6167-7

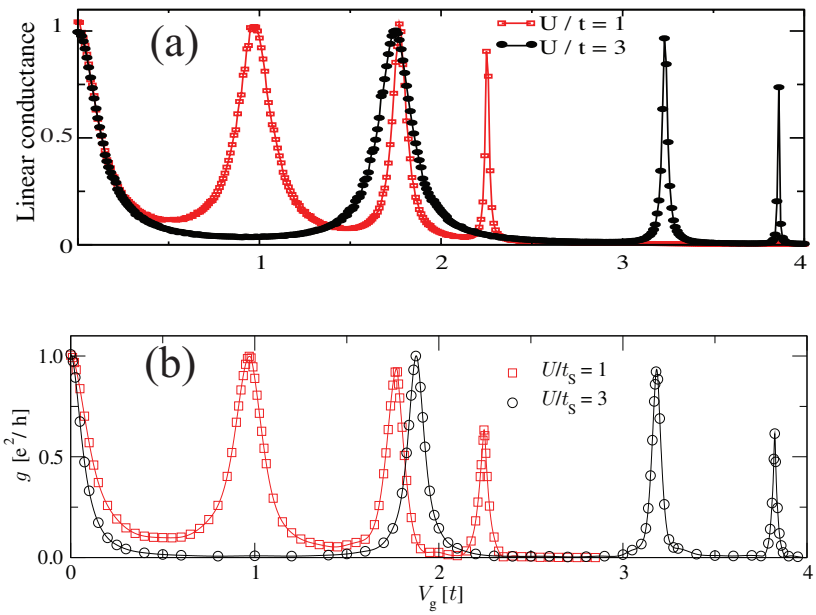


Figure 3: Linear conductance of a chain consisting of seven interacting quantum dots for weak (squares) and strong (circles) interaction. The inter-dot hopping parameter is $t = 0.8$, the coupling to the left and right lead is $t_{\mathrm{L,R}} = 0.5$, and there is an inter-dot Coulomb interaction U indicated in the graphs in units of the inter-dot hopping parameter. The transport is calculated using DMRG (b) (a highly accurate numerical method; graph is taken from ref. [9]) and the method developed in ref. [6] (a).

 ISBN: 978-84-614-6167-7

Proceedings of the 11th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2011
26–30 June 2011.

MutantXL in Focus: *Theoretical Analysis and Run Time Complexity*

Enrico Thomae[1] and Christopher Wolf[1]

[1] *Horst Görtz Institute for IT-security, Ruhr-University of Bochum, Germany*

emails: enrico.thomae@ruhr-uni-bochum.de, chris@Christopher-Wolf.de

Abstract

Solving $\mathcal{M}$ultivariate $\mathcal{Q}$uadratic equations is an established tool in cryptanalysis. In this article, we thoroughly investigate one possible algorithm, named MutantXL. It is a derivate of the *eXtended Linearization* (XL), using special treatment of so called Mutants. Due to the work of Chen, Diem, Moh, and Yang the complexity analysis of XL is well understood.
This paper deals with the question of determining the effect of mutants and achieves a tight complexity analysis of MutantXL. We do so by calculating the number of additional linearly independent equations obtained by mutants. Using the hybrid-strategy, *i.e.* guessing an optimal number of variables beforehand, we show for important parameter sets ($1 \leq m = n \leq 30$) that MutantXL solves at the same degree of regularity as its competitors F_4 and F_5 for many instances. But we also confirm recent results of ePrint 2011/164 that MutantXL is a redundant version of F_4, *i.e.* we show that MutantXL never solves below the degree of regularity of F_4. Thereby we close an important gap in understanding MutantXL.

Key words: Multivariate Cryptography, eXtended Linearization, MutantXL, XL, Cryptanalysis

1 Introduction

Solving general systems of multivariate nonlinear equations is known to be $\mathcal{NP}$-complete [GJ79] and also hard on average. The security of many cryptosystems relies directly or indirectly on this problem. Usually, computing the Gröbner basis of the corresponding ideal is the best choice to solve these kind of equations. The best known and also most efficient algorithms for this task are F_4 and F_5 [Fau99, Fau02]. Note that F_5 is an optimized version of F_4 and hence faster in practical applications.
We investigate another algorithm called "eXtended Linearization" (XL) to solve $\mathcal{MQ}$-systems. XL was first mentioned in the context of cryptography by Courtois *et al.* in [CKPS00]. It could be seen as a generalisation of the "relinearization" technique used

 ISBN: 978-84-614-6167-7

by Kipnis and Shamir to attack HFE at Crypto 1999 [KS99]. Unfortunately the initial papers did not provide a deep analysis of the method and many claims showed to be overly ambiguous. At least since Courtois and Pieprzyk claimed to have broken AES [CP02] using an XL variant called XSL and were disproved by Cid and Leurent [CL05] only a few years later, the community of cryptographers became increasingly reserved against this method. But thanks to Moh [Moh00], Diem [Die04], Yang and Chen [YC04] and others, the MutantXL predecessor XL is understood quit well today.
In section 2 we give a short overview of their results. In section 3 we will extend the analysis to one of the most promising variants of XL, called MutantXL. After theoretically determining the solving degree of MutantXL, a comparison to HybridF$_4$ show that more often than not, it solves at the same degree of regularity for solving the same problem instance.

1.1 Notation

We now give some notations and definitions we will use in the reminder of this article.

$$p^{(k)}(x_1,\ldots,x_n) := \sum_{1\le i\le j\le n} \gamma_{ij}^{(k)} x_i x_j + \sum_{1\le i\le n} \beta_i^{(k)} x_i + \alpha^{(k)}. \tag{1}$$

for $\alpha^{(k)}, \beta_i^{(k)}, \gamma_{ij}^{(k)} \in \mathbb{F}$, $1 \le i \le j \le n$, and $1 \le k \le m$. We denote the coefficients as constant ($\alpha^{(k)}$), linear ($\beta_i^{(k)}$), and quadratic ($\gamma_{ij}^{(k)}$), respectively. In addition, equation $p^{(k)} = 0$ with $p^{(k)}$ defined by (1) are called *inhomogeneous*. Equations with the linear and constant terms being zero (*i.e.* $\alpha^{(k)} = \beta_i^{(k)} = 0$) are called *homogeneous* and simplify to

$$p^{(k)}(x_1,\ldots,x_n) := \sum_{1\le i\le j\le n} \gamma_{ij}^{(k)} x_i x_j. \tag{2}$$

Note that we can write each inhomogeneous polynomial $p^{(k)}$ in n variable as a homogeneous polynomial $\tilde{p}^{(k)}$ in $(n+1)$ variables. As we will see later, this syntactical transformation will not reduce the efficiency of XL or MutantXL for solving the original equation $p^{(k)}$.

Let $P : \mathbb{F}_q^n \to \mathbb{F}_q^m$ be an $\mathcal{MQ}$ system of the form

$$\begin{aligned} p^{(1)}(x_1,\ldots,x_n) &= 0 \\ p^{(2)}(x_1,\ldots,x_n) &= 0 \\ &\vdots \\ p^{(m)}(x_1,\ldots,x_n) &= 0. \end{aligned} \tag{3}$$

Let $\pi^{(k)}$ be the coefficient vector of $p^{(k)}(x_1,\ldots,x_n)$ in lexicographic order, *i.e.*

$$\pi^{(k)} = (\gamma_{11}^{(k)}, \gamma_{12}^{(k)}, \ldots, \gamma_{1n}^{(k)}, \gamma_{22}^{(k)}, \gamma_{23}^{(k)}, \ldots, \gamma_{nn}^{(k)}, \beta_1^{(k)}, \ldots, \beta_n^{(k)}, \alpha^{(k)}).$$

 ISBN: 978-84-614-6167-7

Let Π be the corresponding coefficient matrix

$$\Pi := \begin{pmatrix} \pi^{(1)} \\ \vdots \\ \pi^{(m)} \end{pmatrix}.$$

The naive algorithm is to solve (3) by linearization, *i.e.* to substitute every monomial in $p^{(k)}$ by a new variable and to solve the obtained linear system of equations Π with Gaussian elimination. This will lead to the correct solution if we have $m \approx \frac{n(n+1)}{2} + n$ linearly independent equations, *i.e.* if the dimension of the vector space generated by $\{\pi^{(k)} | 1 \leq k \leq m\}$ is close to the number of monomials. Here "close" means that we can find a univariate equation derived from the original polynomials. If m is not large enough we obtain a (exponential) number of parasitic solutions. The XL algorithm avoids this by producing more *linearly* independent equations through multiplying all equations by all monomials of some degree D. Obviously, these equations are *algebraically* dependent. In addition, this produces new monomials, too, but at some degree D the number of linearly independent equations I will almost be as large as the number of monomials T and we can solve the system by Gaussian elimination. We call this degree D the *saturation degree* of this system P. Note that speaking of the number of linearly independent equations always means the dimension of the vector space generated by a certain set of coefficient vectors $\pi^{(k)}$.

To introduce the XL algorithm formally, we need the following definitions.

Definition 1. *Let* $P^{inh} := \{p^{(k)} \mid 1 \leq k \leq m\}$ *be the set of* inhomogeneous quadratic polynomials p *as defined in (1) and* $P^{hom} := \{p^{(k)} | \ 1 \leq k \leq m\}$ *the set of* homogeneous quadratic polynomials p *defined in (2). We define the set of all monomials of degree* $D \in \mathbb{N}$ *by*

$$Mon_D := \{\prod_{j=1}^{D} x_{i_j} \mid 1 \leq i_1 \leq i_2 \leq \ldots \leq i_D \leq n\} \text{ and } Mon_0 := \{1\}.$$

Multiplying P^{inh} *by all monomials of degree* D *is described by the set*

$$Blow_D^{inh} := \{ab \mid a \in Mon_D \text{ and } b \in P^{inh}\}.$$

The set $Blow_D^{hom}$ *is defined analogous. Now we can write formally what we use as XL algorithm of degree* D*.*

$$XL_D^{inh} := \bigcup_{i=0}^{D} Blow_i^{inh}.$$

For ease of explanation, we mostly use homogeneous equation systems $P^{\mathtt{hom}}$. This is justified as each inhomogeneous system can be transfered into a homogeneous system by introducing one new variable. In particular the number of linearly independent equations produced by $\mathtt{Blow}_D^{\mathtt{hom}}$ equals $\mathtt{XL}_D^{\mathtt{inh}}$ if we replace n by $n+1$. We now define the XL algorithm for a given system $P^{\mathtt{hom}}$.

 ISBN: 978-84-614-6167-7

Definition 2 (XL algorithm)**.** *First we generate* $\texttt{Blow}_D^{hom}$ *and check if the number of linearly independent equations* I *is equal to the number of produced monomials* T *subtracted by* $(D+2)$*. In this case we linearise the system and solve it by Gaussian elimination.*
Notice, if $(T-I) \leq (D+2)$ *we can choose the order of the monomials such that we obtain a univariate equation after linearization. This equation can be solved* e.g. *by Berlekamp's algorithm. Using this value in equations with more monomials, we can successively solve the whole system* P*.*
If $(T-I) > (D+2)$ *we set* $D := D+1$ *and try again.*

Note, if the system is structured as the public key of HFE, XL may still find a univariate polynomial as we now have linear dependencies in the *columns* or *monomials*, not only in the *rows/equations*. Hence, the rank needed on the row-side to solve the matrix equation in Π will drop accordingly.

2 Analysis of XL

The crucial point when using XL is to determine the number I of linearly independent equations produced by $\texttt{Blow}_D^{\texttt{hom}}$. This allows us to calculate the smallest D (saturation degree) such that $(T-I) \leq (D+2)$ holds and therefore determine the complexity of the overall algorithm. Moh [Moh00] was the first to prove (4) for systems of equations fulfilling some special property. For the special case of two equations this property means that they are co-prime. As we investigate random systems of equations, this is true with overwhelming probability (cf. the proof of Lemma 1). So we claim that most of the random systems of equations fulfill this property and thus formula (4) hold. This conjecture is also confirmed by various experiments for D between 0 and 5 over $\mathbb{F}_2$ by Courtois and Patarin [CP03] and over larger fields by Thomae and Wolf [TW10]. Diem [Die04] showed that (4) holds if the *maximal rank conjecture*—also known as *Frï£ ¡bergs's conjecture*—holds for generic systems. For a ground field $\mathbb{F}_q$ and $D+2 < q$, the formulas are independent of the ground field. If $D+2 \geq q$ we have to take the field equations $x^q - x$ into account and formulas become more expensive. For example if $q=2$ the number of monomials of degree D decreases from $\binom{n+D-1}{D}$ to $\binom{n}{D}$ and besides of the trivial dependency $fg = gf$ there is an additional dependency due to $f^2 = f$ for f, g quadratic polynomials (see Yang and Chen [YC04] or Rønjom and Raddum [RR08] for more details). We will only consider the case $(D+2) < q$ in the following. Analyzing the case $(D+2) \geq q$ is analogous but uses a slightly different formula (4).
We can write down the number I of linearly independent equations for most systems (see [Moh00]) by the following equation.

$$\begin{aligned} D = 2k: \quad & I_{\texttt{Blow}_D^{\texttt{hom}},n} := \sum_{i=0}^{k} (-1)^i \binom{m}{i+1} \binom{n+2(k-i)-1}{2(k-i)}, \\ D = 2k+1: \quad & I_{\texttt{Blow}_D^{\texttt{hom}},n} := \sum_{i=0}^{k} (-1)^i \binom{m}{i+1} \binom{n+2(k-i)}{2(k-i)+1}. \end{aligned} \tag{4}$$

 ISBN: 978-84-614-6167-7

Note that $I_{\mathtt{XL}_D^{\mathtt{inh}},n} = I_{\mathtt{Blow}_D^{\mathtt{hom}},n+1}$ holds, *i.e.* moving from inhomogeneous equations to homogenized equations does not affect the running time of XL.
The number of monomials T is $\binom{n+D+1}{D+2}$ and thus we are now able to calculate D such that $(T-I) \leq (D+2)$ holds and XL succeed.
In table 1 we give the number of linearly independent equations I produced by $\mathtt{Blow}_D^{\mathtt{hom}}$ for some specific values of D, always assuming $D+2 < q$ (cf. [Moh00, Die04, YC04, CP03]).

Table 1: Number of linearly independent equations produced by $\mathtt{Blow}_D^{\mathtt{hom}}$, *i.e.* XL with degree D,m equations in P, and n variables. They have been derived both theoretically and empirically.

D	Number of linearly independent equations
0	m
1	mn
2	$m\binom{n+1}{2} - \binom{m}{2}$
3	$m\binom{n+2}{3} - \binom{m}{2}n$
4	$m\binom{n+3}{4} - \binom{m}{2}\binom{n+1}{2} + \binom{m}{3}$
5	$m\binom{n+4}{5} - \binom{m}{2}\binom{n+2}{3} + \binom{m}{3}n$

3 Analysis of MutantXL

One of the most efficient variants of XL is called MutantXL [MMD+08]. It is claimed to be as fast as F_4 in some cases. This was derived from experiments on HFE [BDMM09]. We will now give a theoretical analysis of MutantXL and confirm that it indeed solves at the degree of regularity as F_4 in many cases.
Let I be the number of linearly independent equations produced by $\mathtt{XL}_D^{\mathtt{inh}}$ and $T = \binom{n+D+2}{D+2}$ the number of degree $\leq (D+2)$ monomials. If $(T-I) > (D+2)$ it is highly unlikely that XL finds a univariate polynomial and thus solves the system. As outlined above, XL will continue with $D := D+1$. MutantXL is a step in between. Instead of doing a full extension from D to $D+1$ it uses equations that would be produced by $\mathtt{XL}_{D+k}^{\mathtt{inh}}$ with $k > 0$ as long as they do not introduce new monomials. To this aim we use only polynomials of degree $< D+2$ that are produced in the Gaussian elimination step of $\mathtt{XL}_D^{\mathtt{inh}}$. These polynomials are called *mutants*. For example multiplying these polynomials by all monomials of $\mathtt{Mon}_1$ leads to new equations without generating new monomials. However, this strategy is only useful for *inhomogeneous* equations. In the *homogeneous* case all monomials have same degree and thus mutants simply never occur. Note that this is not true for *homogenised* systems of equations. Here, the mutants are only hidden by the homogenization variable. Hence, as long as the initial system is inhomogeneous, it is not a contradiction to speak of *mutants* and to use (4) for the homogenized system, too.

 ISBN: 978-84-614-6167-7

Definition 3. *Let* $f = \sum_{i=1}^{m} g_{j_i} h^{(i)}$ *with* $h^{(i)} \in P^{\boldsymbol{inh}}$ *and* g_{j_i} *some polynomial of degree* $\leq D$ *be a representation of* f*. This representation is not unique. The set* J *denotes all representations* $(j_1, \ldots, j_m)$*. The* level (lev) *of this representation* $j \in J$ *is defined by*

$$lev\left(\sum_{i=1}^{m} g_{j_i} h^{(i)}\right) := max\left\{deg\left(g_{j_i} h^{(i)}\right) \mid 1 \leq i \leq m\right\}.$$

The level (Lev) *of* g *is defined by the minimum level of all its representations.*

$$Lev\,(g) := min\{lev\left(\sum_{i=1}^{m} g_{j_i} h^{(i)}\right) \mid j \in J\}$$

We call g *a mutant if* $deg(g) < Lev(g)$*.*

The crucial question as always is how many equations produced by mutants are linearly independent to the old ones. We give two upper bounds on this number. We showed experimentally that the smaller bound is tight. We will also give some theoretical explanation on that. To conclude we compare MutantXL to F_4 and show that it solves at the degree of regularity more often than not, but never below.

Remark 1. *Our algorithm does not proceed degree-wise. We first calculate the saturation degree* D *that solves the random system and than calculate* $\boldsymbol{XL}_D^{\boldsymbol{inh}}$*. To implement MutantXL in this scenario in a smart way, we will introduce the term of* trivial mutants*. Using* $\boldsymbol{XL}_D^{\boldsymbol{inh}}$ *all equations produced by* $\boldsymbol{Blow}_d^{\boldsymbol{inh}}$ *with* $d < D$ *are not useful for searching mutants, as their multiples of total degree less than* $(D+2)$ *are already contained in* $\boldsymbol{XL}_D^{\boldsymbol{inh}}$ *and thus are trivially linearly dependent. We call mutants trivial if equations of* $\boldsymbol{Blow}_d^{\boldsymbol{inh}}$ *with* $d < D$ *sum up to equations of degree less than* $(d+2)$*. Let* $g = \sum_i a_i f_i$ *be a* trivial mutant *with* $a_i \in \mathbb{F}_q$*,* $f_i \in \boldsymbol{Blow}_d^{\boldsymbol{inh}}$ *and* $\deg(g) < (d+2)$*. For every* $x \in \boldsymbol{Mon}_{D-d}$ *we obtain* $xg = \sum_i a_i x f_i$ *for* $xf_i \in \boldsymbol{Blow}_D^{\boldsymbol{inh}}$ *and* $\deg(xg) < D+2$*,* i.e. xg *is a mutant of* $\boldsymbol{Blow}_D^{\boldsymbol{inh}}$*. Thus we can reduce the computational workload if we only consider mutants produced by* $\boldsymbol{Blow}_D^{\boldsymbol{inh}}$*,* i.e. *if we only consider* non-trivial mutants *or more precisely only use the Mutant strategy in the final step of XL.*

Let $\mathcal{D}_{\mathtt{Blow}_D^{\mathtt{inh}},n} := I_{\mathtt{XL}_D^{\mathtt{inh}},n} - I_{\mathtt{XL}_{D-1}^{\mathtt{inh}},n}$ denote the difference of the dimensions of the vector spaces generated by $\mathtt{XL}_D^{\mathtt{inh}}$ and $\mathtt{XL}_{D-1}^{\mathtt{inh}}$. According remark 1 all non-trivial mutants are part of the set of $\mathtt{Blow}_D^{\mathtt{inh}}$. To avoid hiding the upper bounds behind formalism, we start with the case $|\mathtt{Mon}_{D+2}| \leq \mathcal{D}_{\mathtt{Blow}_D^{\mathtt{inh}},n} \leq |\mathtt{Mon}_{D+2}| + |\mathtt{Mon}_{D+1}|$ illustrated in figure 1.
We can calculate $\mathcal{D}_{\mathtt{Blow}_D^{\mathtt{inh}},n}$ using (4) by

$$\mathcal{D}_{\mathtt{Blow}_D^{\mathtt{inh}},n} = I_{\mathtt{XL}_D^{\mathtt{inh}},n} - I_{\mathtt{XL}_{D-1}^{\mathtt{inh}},n} = I_{\mathtt{Blow}_D^{\mathtt{hom}},n+1} - I_{\mathtt{Blow}_{D-1}^{\mathtt{hom}},n+1}\,. \tag{5}$$

A first trivial upper bound on the benefit of MutantXL is the dimension of the vector space spanned by all new equations produced by mutants. In the case $k = 1$ (see figure 1) this amounts to $n(\mathcal{D}_{\mathtt{Blow}_D^{\mathtt{inh}},n} - |\mathtt{Mon}_{D+2}|)$ or $n\widetilde{m}$ using the notation of figure 1, as we

 ISBN: 978-84-614-6167-7

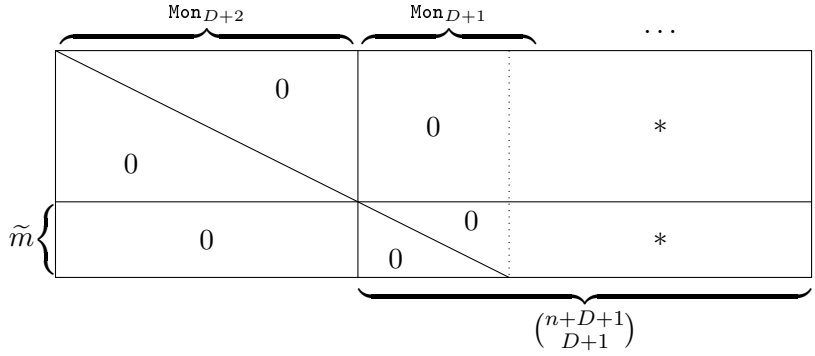


Figure 1: Coefficient Matrix Π of $\mathtt{Blow}_D^{\mathtt{inh}}$ after Gaussian elimination. Here $\tilde{m}$ indicates the number of mutants for the corresponding system P.

multiply all $\tilde{m}$ Mutants by all n Monomials of degree one. Experiments for $2 \leq n \leq 7$ and $n \leq m \leq 9$ over $\mathbb{F}_{2^8}$ show that this trivial bound is far above the correct number of new linearly independent equations. Still, it serves as a useful point of reference to guide further development of the theory.
The second upper bound uses that all $n\tilde{m}$ equations produced by mutants are implicit equations of $\mathtt{XL}_{D+1}^{\mathtt{inh}}$. Exactly $I_{\mathtt{XL}_{D+1}^{\mathtt{inh}},n} - I_{\mathtt{XL}_D^{\mathtt{inh}},n}$ of them are linearly *independent* to $\mathtt{XL}_D^{\mathtt{inh}}$. But they all contain monomials of $\mathtt{Mon}_{D+3}$. Equations produced by mutants have maximal degree $D+2$ and thus first all $|\mathtt{Mon}_{D+3}|$ monomials have to be reduced. Therefore $I_{\mathtt{XL}_{D+1}^{\mathtt{inh}},n} - I_{\mathtt{XL}_D^{\mathtt{inh}},n} - |\mathtt{Mon}_{D+3}|$ is an upper bound on the number of linearly independent equations produced by mutants. Note that this bound was tight in all our experiments for $2 \leq n \leq 7$ and $n \leq m \leq 9$. Note, Experiments with larger n, m were not possible due to the high computational complexity of XL. However, in lemma 1 we give theoretical justification why this bound is tight with overwhelming probability.

To generalise the above example let $k \in \mathbb{N}$:

$$\sum_{j=0}^{k-1} |\mathtt{Mon}_{D+2-j}| \leq \mathcal{D}_{\mathtt{Blow}_D^{\mathtt{inh}},n} \leq \sum_{j=0}^{k} |\mathtt{Mon}_{D+2-j}| \,. \tag{6}$$

The following two upper bounds hold.

Corollary 1. *The maximal number of equations produced by non-trivial mutants is given by*

$$\sum_{i=1}^{k-1} \binom{n+i-1}{i} |\mathtt{Mon}_{D+2-i}| + \binom{n+k-1}{k} \left(\mathcal{D}_{\mathtt{Blow}_D^{\mathtt{inh}},n} - \sum_{i=0}^{k-1} |\mathtt{Mon}_{D+2-i}| \right).$$

 ISBN: 978-84-614-6167-7

Corollary 2. *A nontrivial upper bound on the number of linearly independent equations produced by mutants is given by*

$$\begin{aligned} & \sum_{i=1}^{k} \left(I_{XL^{inh}_{D+i},n} - I_{XL^{inh}_{D+i-1},n} - |Mon_{D+2+i}| \right) \\ = \quad & I_{XL^{inh}_{D+k},n} - I_{XL^{inh}_{D},n} - \sum_{j=1}^{k} |Mon_{D+2+j}| \,. \end{aligned}$$

In figure 2 we calculated k for all values of random systems of quadratic equations with $n \in [1, 30]$ variables and $m \in \left[n, \frac{n(n+1)}{2} \right]$ equations. The restriction on n and m is only in order to draw a picture. Note that for large finite fields the values we chose are of most practical interest.
The rows of figure 2 denote n, the columns denote m and the color denote the value of k, whereby black stands for $k = 1$ and grey stands for $k = 0$. For numerical values of k see table 4. In a nutshell, MutantXL has an advantage over plain XL whenever $k > 0$, i.e. in the black areas of figure 2. In particular $k = 1$ means that MutantXL saturates at one degree less than XL.

We want to add that Mohamed *et al.* confirmed our analysis on ePrint [MM11]. They corrected a typo in the formula of corollary 2 in one of our previous versions. However, for $k \leq 1$ this typo did not change the correctness of the formula above. While we already showed graphically and numerically that $k > 1$ does not appear in practice, we will investigate the question a bit deeper now and give a precise mathematical criterion *when* this claim will hold. Note that the following claim only hold for fields $\mathbb{F}_q$ with $(D + 2) < q$ as otherwise (4) is not longer correct.

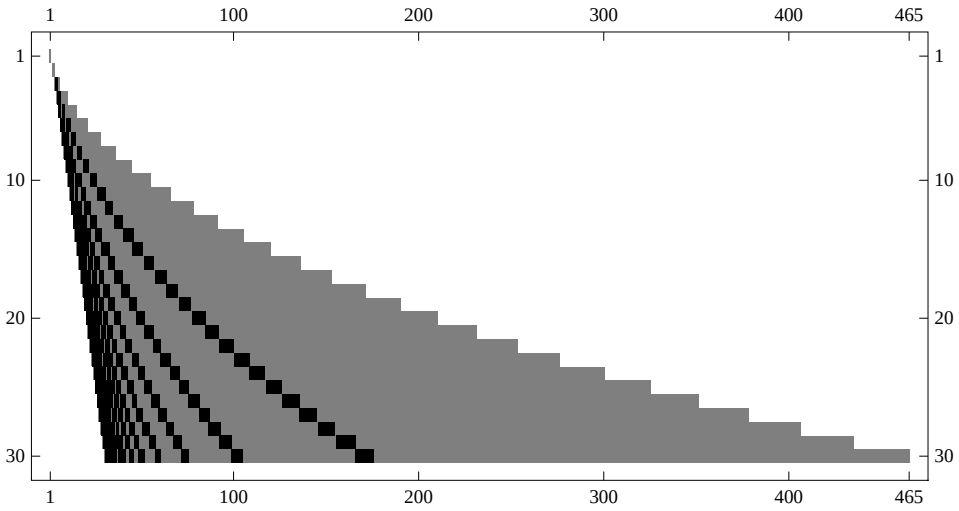


Figure 2: Visualization of mutation variable k for $n = 1 \ldots 30$. Black: $k = 1$, gray: $k = 0$. Y-Axis is number of variables n, x-axis the number of equations m.

 ISBN: 978-84-614-6167-7

Claim 1. *We claim that $k \leq 1$ always holds.*

Proof. First, the following constraint have to be fulfilled, as the problem would be solved with $k = 0$ otherwise.

$$\binom{n+D+2}{D+2} - I_{\mathtt{XL}^{\mathtt{inh}}_D,n} > D+2 \tag{7}$$

To show that $k \leq 1$ hold, we have to proof the following equation.

$$\mathcal{D}_{\mathtt{Blow}^{\mathtt{inh}}_D,n} \leq \binom{n+D+1}{D+2} + \binom{n+D}{D+1}.$$

Using (5) this is equivalent to

$$I_{\mathtt{Blow}^{\mathtt{hom}}_D,n+1} \leq \binom{n+D+1}{D+2} + \binom{n+D}{D+1} + I_{\mathtt{Blow}^{\mathtt{hom}}_{D-1},n+1}.$$

The constraint of equation (7) give us the following.

$$\begin{aligned} I_{\mathtt{Blow}^{\mathtt{hom}}_D,n+1} &< \binom{n+D+2}{D+2} - D - 2 \\ &= \binom{n+D+1}{D+2} + \binom{n+D}{D+1} + \binom{n+D}{D} - D - 2 \end{aligned}$$

The claim follows if $\binom{n+D}{D} - D - 2 \leq I_{\mathtt{Blow}^{\mathtt{hom}}_{D-1},n+1}$ hold. Unfortunately we only could proof this numerically and thus leave it as a claim. Still—for any practical choice of D, n, the claim holds. □

We come back to the example in figure 1 to derive a lower bound. With B_1 we denote the bound given by corollary 1 and B_2 the bound given by corollary 2.

Lemma 1. *If D is even and $k = 1$ then the number of new linearly independent equations produced by mutants is $\min\{B_1, B_2\}$ with probability $1 - negl(n, D)$.*

Proof. As shown in [TW10], we know that new linearly dependent equations are produced block-wise, *i.e.* if we multiply two quadratic polynomials f and g by all monomials of degree two, in general all equations are linearly independent besides one, as $fg = gf$ holds. For m equations there are $\binom{m}{2}$ pairs and thus $\binom{m}{2}$ dependencies. Multiplying by all monomials of degree three only multiplies every dependency by every variable and thus we obtain $\binom{m}{2}n$ dependencies. New dependencies are introduced by multiplying all equations by all monomials of degree four, as for three quadratic polynomials f, g and h the equation $fgh = fhg = ghf$ hold. Thus new linearly dependent equations are only produced by proceeding from odd to even degree D. Multiplying the mutants with degree one monomials we implicitly use equations of the set $\mathtt{Blow}^{\mathtt{inh}}_{D+1}$. As a consequence, for even D no new linear dependencies are introduced. This is also in line with our experiments.
The only way left to produce linearly dependencies is by generating the same equation in the multiplication step, *i.e.* if g_1 and g_2 are mutants and $x_i g_1 = x_j g_2$ holds. This is

 ISBN: 978-84-614-6167-7

implies $x_j \,|\, g_1$ and $x_i \,|\, g_2$. The mutant $g_1 = \sum_i a_i h_i$ with $a_i \in \mathbb{F}_q$ and $h_i \in \mathtt{Mon}_{\leq D-1}$ can be seen as a random dense equation. Hence the probability of x_j being a factor of g_1 is given by

$$\left(\frac{1}{q}\right)^{\binom{n-1+D-1}{D-1}}$$

and thus is negligible in D, n. □

We denote by m the number of quadratic equations, n the number of variables with $n = m$ and r the number of guessed variables *(hybrid strategy)*, *i.e.* $\widetilde{n} = m - r$. First we compare the degree of regularity d_{reg} [BFSY05] with the saturation degree D—or more precisely the comparable degree $(D + 2)$. The degree of regularity in F_4 is the smallest degree such that the dimension of the ideal produced by F_4 is equal to the number of monomials of degree d_{reg} : $\binom{n+d_{reg}-1}{d_{reg}}$. This is similar to the saturation degree for XL where the number of columns T and the number of linearly independent equation I match. Note that we assumed that lemma 1 holds also for odd D and that MutantXL needs $D + 2 = 2^m$ in case $m = n$, *i.e.* if we do not guess some variables. As this is exponential and the corresponding values are both becomes really large and not very instructive, we only indicated the corresponding entries in table 3 with "-" for *"practically not solvable"*.

As Bettale *et al.* we use $\omega = 2$ to calculate the linear part of the algorithm for their hybrid approach and our analysis of MutantXL. We obtain the same results as in [BFP09, table 4] for $m = 20$ and guessing one or two variables over $\mathbb{F}_{2^8}$, see table 5 for the correct value of k. The values in the tables are rounded Log_2 complexities. The exact value for $m = 20$, $r = 1$ and $\mathbb{F}_{2^8}$ is 66.73 respectively 67.79 for $r = 2$. Note that MutantXL has the same optimal complexity for every m, *i.e.* if we guess the correct number of variables.

The complexity of MutantXL is determined by the Gaussian elimination step on all $m\binom{n+D}{D}$ equations produced by XL and the number of mutants—if any. This is captured

$m\backslash r$	0	1	2	3	5
5	6	3	3	2	2
10	11	6	5	4	3
15	16	8	7	6	4
20	21	11	9	8	6
25	26	13	11	10	8
30	31	16	14	12	10

Table 2: Degree of Regularity d_{reg} for F_4

$m\backslash r$	0	1	2	3	5
5	32	4	3	2	2
10	-	9	5	4	3
15	-	14	7	6	4
20	-	19	10	8	6
25	-	24	12	10	8
30	-	29	15	13	10

Table 3: Degree $(D + 2)$ of MutantXL for $k \leq 1$.

 ISBN: 978-84-614-6167-7

$m\backslash r$	0	1	2	3	5
5	0	1	0	0	0
10	0	1	1	1	0
15	0	1	1	1	1
20	0	1	1	1	1
25	0	1	1	1	1
30	0	1	1	1	1

Table 4: k obtained for MutantXL by (6).

$m\backslash r$	0	1	2	3	5
5	20	21	27	32	40
10	41	38	42	46	57
15	62	51	55	59	67
20	83	67	68	72	79
25	103	79	80	84	91
30	123	95	96	96	104

Table 5: Complexity of F_5 over $\mathbb{F}_{2^8}$

$m\backslash r$	0	1	2	3	5
5	41	21	23	29	40
10	-	42	37	41	49
15	-	62	50	54	60
20	-	83	66	66	73
25	-	103	78	78	85
30	-	123	94	94	97

Table 6: Complexity of MutantXL over $\mathbb{F}_{2^8}$.

by $\max\left\{0, \binom{n+D+2}{D+2} - I - D - 2\right\}$. Thus the complexity of table 6 is given by

$$\left(m\binom{n-r+D}{D} + \max\left\{0, \binom{n-r+D+2}{D+2} - I - D - 2\right\}\right)^{\omega} \cdot (2^8)^r.$$

4 Conclusion

In this paper, we have clarified solving $\mathcal{M}$ultivariate $\mathcal{Q}$uadratic equations using MutantXL from a theoretical point of view. To this aim, we have revised some priorly known results in Section 2. From there, we were able to give an upper bound on the number of mutants and also the number of linearly independent equations produced by them in corr. 1+2. We showed tightness of the second bound (for $k \leq 1$) experimentally and in some case also using analytical methods. However, all practically relevant cases are covered by our analysis, so from a cryptographic point of case, we are done. Finally, we have shown that MutantXL can compete with F_4 in terms of the saturation degree. This is not clear from an algorithmical point of view, *e.g.* the memory requirement for MutantXL should be larger as for F_4 as MutantXL to not reduce the basis during the intermediate steps.

Acknowledgements

The authors were funded via an DFG (German Research Foundation) Emmy Noether grant. In addition, the work described in this paper has been supported in part by the European Commission through the ICT programme under contract ICT-2007-216676 ECRYPT II.

 ISBN: 978-84-614-6167-7

References

[ACFP11] Martin Albrecht, Carlos Cid, Jean-Charles Faugère, and Ludovic Perret. On the relation between the mutant strategy and the normal selection strategy in gröbner basis algorithms. Cryptology ePrint Archive, Report 2011/164, 2011. http://eprint.iacr.org/.

[BDMM09] Johannes A. Buchmann, Jintai Ding, Mohamed Saied Emam Mohamed, and Wael Said Abd Elmageed Mohamed. Mutantxl: Solving multivariate polynomial equations for cryptanalysis. In Helena Handschuh, Stefan Lucks, Bart Preneel, and Phillip Rogaway, editors, *Symmetric Cryptography*, number 09031 in Dagstuhl Seminar Proceedings, Dagstuhl, Germany, 2009. Schloss Dagstuhl - Leibniz-Zentrum fuer Informatik, Germany.

[BFP09] Luk Bettale, Jean-Charles Faugère, and Ludovic Perret. Hybrid approach for solving multivariate systems over finite fields. *In Journal of Mathematical Cryptology*, 3:177–197, 2009.

[BFSY05] M. Bardet, J.-C. Faugère, B. Salvy, and B.-Y. Yang. Asymptotic behaviour of the degree of regularity of semi-regular polynomial systems. In *MEGA'05*, 2005. Eighth International Symposium on Effective Methods in Algebraic Geometry, Porto Conte, Alghero, Sardinia (Italy), May 27th – June 1st.

[CKPS00] Nicolas T. Courtois, Alexander Klimov, Jacques Patarin, and Adi Shamir. Efficient algorithms for solving overdefined systems of multivariate polynomial equations. In *Advances in Cryptology — EUROCRYPT 2000*, volume 1807 of *Lecture Notes in Computer Science*, pages 392–407. Bart Preneel, editor, Springer, 2000. Extended Version: http://www.minrank.org/xlfull.pdf.

[CL05] Carlos Cid and GaÃńtan Leurent. An analysis of the xsl algorithm. In *Proceedings of Asiacrypt 2005, LNCS*, volume 3788 of *Lecture Notes in Computer Science*, pages 333–352. Bimal Roy, editor, Springer-Verlag, 2005. ISBN 3-540-30684-6.

[CP02] Nicolas T. Courtois and Josef Pieprzyk. Cryptanalysis of block ciphers with overdefined systems of equations. In *Advances in Cryptology — ASIACRYPT 2002*, volume 2501 of *Lecture Notes in Computer Science*, pages 267–287. Yuliang Zheng, editor, Springer, 2002.

[CP03] Nicolas T. Courtois and Jacques Patarin. About the XL algorithm over GF(2). In *CT-RSA'03: Proceedings of the 2003 RSA conference on The cryptographers' track*, pages 141–157, Berlin, Heidelberg, 2003. Springer-Verlag.

[Die04] Claus Diem. The XL-algorithm and a conjecture from commutative algebra. In *ASIACRYPT*, volume 3329 of *Lecture Notes in Computer Science*. Pil Joong Lee, editor, Springer, 2004. ISBN 3-540-23975-8.

[Fau99] Jean-Charles Faugère. A new efficient algorithm for computing Gröbner bases (F_4). *Journal of Pure anNational Institute of Standards and Technologyd Applied Algebra*, 139:61–88, June 1999.

[Fau02] Jean-Charles Faugère. A new efficient algorithm for computing Gröbner bases without reduction to zero (F_5). In *International Symposium on Symbolic and Algebraic Computation — ISSAC 2002*, pages 75–83. ACM Press, July 2002.

[GJ79] Michael R. Garey and David S. Johnson. *Computers and Intractability — A Guide to the Theory of NP-Completeness*. W.H. Freeman and Company, 1979. ISBN 0-7167-1044-7 or 0-7167-1045-5.

[KS99] Aviad Kipnis and Adi Shamir. Cryptanalysis of the HFE public key cryptosystem. In *Advances in Cryptology — CRYPTO 1999*, volume 1666 of *Lecture Notes in Computer Science*, pages 19–30. Michael Wiener, editor, Springer, 1999. `http://www.minrank.org/hfesubreg.ps` or `http://citeseer.nj.nec.com/kipnis99cryptanalysis.html`.

[MM11] J. Buchmann M.S.E. Mohamed, J. Ding. The complexity analysis of the mutantxl family. Cryptology ePrint Archive, Report 2011/036, 2011. `http://eprint.iacr.org/`.

[MMD+08] Mohamed Saied Mohamed, Wael Said Mohamed, Jintai Ding, Johannes Buchmann, Stefan Tohaneanu, Ralf-Philipp Weinmann, Daniel Carbarcas, and Dieter Schmidt. Mutantxl and mutant grï£¡bner basis algorithm. In *SCC '08: Proceedings of the 1st International Conference on Symbolic Computation and Cryptography*, pages 16–22, 2008.

[Moh00] T. Moh. On the method of "XL" and its inefficiency to TTM, 2000.

[RR08] Sondre Rønjom and Håvard Raddum. On the number of linearly independent equations generated by xl. In *Proceedings of the 5th international conference on Sequences and Their Applications*, SETA '08, pages 239–251, Berlin, Heidelberg, 2008. Springer-Verlag.

[TW10] Enrico Thomae and Christopher Wolf. Unreval xl and its variants. Cryptology ePrint Archive, Report 2010/596, 23rd of November 2010. `http://eprint.iacr.org/`.

[YC04] Bo-Yin Yang and Jiun-Ming Chen. All in the XL family: Theory and practice. In *ICISC 2004*, pages 67–86. Springer, 2004.

 ISBN: 978-84-614-6167-7

Proceedings of the 11th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2011
26–30 June 2011.

Scalable shot boundary detection

Pablo Toharia[1], Oscar D. Robles[1], Jose Luis Bosque[2] and Angel Rodriguez[3]

[1] *Dpto. de ATC y CCIA , Universidad Rey Juan Carlos, Spain*

[2] *Dpto. de Electrónica y Computadores, Universidad de Cantabria, Spain*

[3] *Dpto. de Tecnología Fotónica, Universidad Politécnica de Madrid, Spain*

emails: pablo.toharia@urjc.es, oscardavid.robles@urjc.es, joseluis.bosque@unican.es, arodri@fi.upm.es

Abstract

Shot boundary detection is the first step to be performed on any Content-based Video Retrieval system. The performance of such techniques must be evaluated from a computational point of view since this is a very high demanding task and thus optimization strategies must be seeked. This paper proposes scalable and portable implementations that grant low response times over two alternative parallel architectures: a shared-memory symmetric multiprocessor and a Beowulf cluster. Two alternative parallel programing paradigms have been evaluated, shared-memory and message passing, implementing several strategies for video segmentation and data access and analyzing load balancing issues.

1 Introduction

Automatic techniques for extracting relevant information from raw data must be sought in order to efficiently access multimedia databases. Content-based Multimedia Retrieval (CBMR) systems provide a very useful help to users whose aim is to introduce a query in the system and to retrieve the most similar items in the data sets [6, 7, 10].

When dealing with video data, the first step is to perform a temporal video segmentation in order to isolate the minimum units with semantic meaning: shots. Shot boundary detection (SBD) has two main challenges: to accurately delimit the start and the end of video shots and to process video contents in a more efficient way. It is the unavoidable first step to proceed with new data in a Content-based Video Retrieval process. Depending on the working domain, these techniques can be classified in non-compressed [12] and compressed video shot segmentation [1]. This paper is focused on non-compressed video segmentation since it is an interesting testbench for primitives to be also used in a retrieval stage, as in [9].

 ISBN: 978-84-614-6167-7

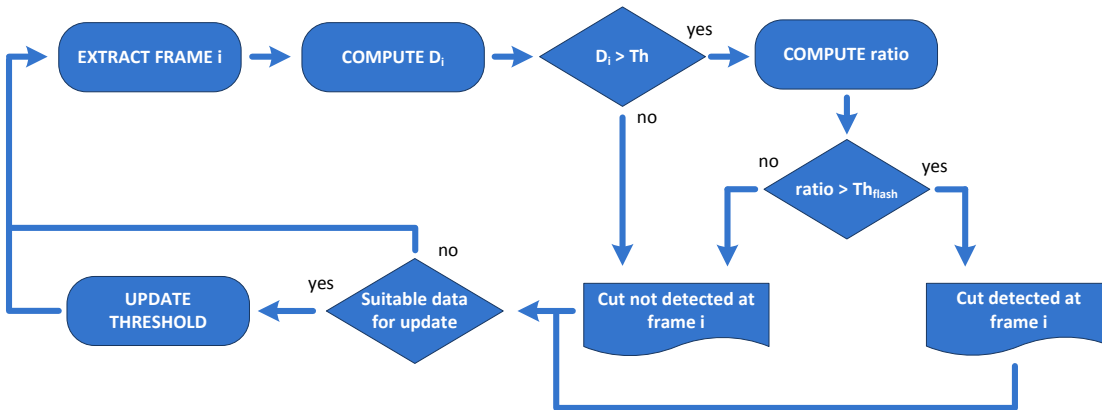


Figure 1: Cut detection algorithm filtering flashes.

Video segmentation is a very high demanding process, so from a computational point of view, efficient and scalable implementations must be sought to grant low response times even when the size of data to be processed noticeably grows up. This paper presents an exhaustive analysis of a typical SBD algorithm, performing experimental tests over suitable architectures, adequate paradigms and efficient implementations. The results allow to extract relevant conclusions about the set of parameters that achieve the best results for the most suitable architecture from the point of view of both performance and scalability. Thus, this work compares the performance achieved by two alternative parallel implementations of a shot boundary detection algorithm using two different parallel programming paradigms: shared-memory communication and distributed memory processing using the message passing paradigm. Taking into account the software solutions, two different parallel architectures are considered to compare both implementations: a shared-memory symmetric multiprocessor (SMP) and a distributed system. Distributed solutions on clusters offer a good cost/performance ratio to solve this problem, given their excellent scalability, fault tolerance and flexibility attributes [3, 2].

2 Shot boundary detection algorithm

The basic idea of shot boundary detection algorithms is to compute differences between consecutive frames or groups of frames. Existing techniques typically differ in the way these differences are computed. Figure 1 depicts a scheme of the whole process. D_i denotes the difference between frame i and the previous one. In the present case, computed D_i difference values are based on several shape and color features, although other possibilities can also be considered [8, 5, 12]. At this point, it must be noticed that the extraction of difference values does not affect the posed parallelization strategy.

A candidate for cut is detected when difference D_i values are higher than a dynamically computed threshold Th [9]. In this way the threshold Th is updated for each processed

 ISBN: 978-84-614-6167-7

Table 1: Execution time and maximum speedup under Amdahl's law for different Zernike polynomials compared with the quantified histogram.

Method	Video size	Dec. time	Seg. time	Total time	Serial frac.	Parallel frac.	S_∞
zer3	5000	62.80	293.20	356	17.64	82.36	5.67
	10000	125.84	586.16	712	17.67	82.33	5.66
	20000	252.34	1172.66	1425	17.71	82.29	5.65
zer5	5000	62.88	587.12	650	9.67	90.33	10.34
	10000	125.70	1174.30	1300	9.67	90.33	10.34
	20000	252.74	2350.26	2603	9.71	90.29	10.30
zer10	5000	63.00	2124.00	2187	2.88	97.12	34.71
	10000	125.84	4249.16	4375	2.88	97.12	34.77
	20000	254.11	8503.89	8758	2.90	97.10	34.47

frame. One of the typical artifacts present in videos is the appearance of flashes that distort normal analysis of video signals, because there is no change in the video content but abrupt changes appear in signal intensity. In order to filter out flashes, a second threshold T_{flash} has been implemented, following the model of Zhang *et al.* [13]. Finally, once comparisons are performed, the threshold Th is recalculated depending on the variance of the sliding window, so that if it varies too much from frame i to next frame $i+1$, as it is the case when, for example, very fast camera movements occur, the value can not be adapted to the new video signal content. Implemented features have been Zernike invariants as shape primitive [11]. Working with video data coded in MPEG implies a first decoding stage on compressed data followed by a second stage where video shot extraction is performed on non-compressed frames with the algorithm described above.

Decoding and segmentation time values are presented in Table 1. It can be observed how the increment of the polynomial degree greatly increases the execution time of the segmentation stage, reaching almost 2 hours and a half for the largest video size (20000 frames) in the case of zer10. Analyzing the execution time involved in each stage, it can be noticed that the shot boundary stage involves much larger computational load than the decoding stage when video size grows even when low order polynomials are considered. This fact justifies the need to search for efficient and scalable solutions to face the shot boundary detection stage. Following Amdahl's law it can be computed the maximum achievable speedup when it is only considered a parallelization of the segmentation stage. However, Gustafson noticed that serial and parallel sections of a program are also dependent on the program size. The bigger is the size of the problem to solve, the greater is the maximum achievable speedup [4]. On the other hand, the different parallel designs proposed have a serial section that overlaps in some way with the parallel section of the problem (the segmentation stage). This overlap is bigger when the number of system nodes increases. This is why the maximum achievable speedup will be higher not only because of the bigger problem size but also because of the system size. All these aspects ensure good scalability properties to the design.

Table 1 shows how maximum speedup is quite sensible to the polynomial degree se-

 ISBN: 978-84-614-6167-7

lected. This is due to the fact that serial fraction (i.e. decoding time) remains constant, while parallel fraction (i.e. segmentation time) grows with the polynomial degree. The decoding stage must be solved sequentially, due to the data dependencies existing among the video frames. Values in Table 1 show the benefits of using a parallel implementation, specially when using high order polynomials. Upon observing the operations involved in the extraction stage, it can be deduced that all processed frames have data dependencies, but only those belonging to shot boundaries are critical to compute the exact points where shots begin and end. This means that an approach based on data decomposition could be feasible if the boundary is fully inside the frame slice assigned to one of the p threads or processes. Therefore, shot extraction stage could be fully parallelized.

The sequential version of this algorithm requires the displacement of a symmetric mask window with W coefficients to detect the presence of a shot boundary in the frame where the mask is placed. It implies that the assignment of frame slices to threads must be done taking into account the extra frames needed to compute mask operations within the slice's first and last frames. Due to this, there will be two small windows overlapping among consecutive slices

3 Scalable implementations

As mentioned before, in this paper we deal with two different parallel architectures (**SMP** and **Cluster**) and with two different parallel programming paradigms. The first solution proposed is based on the shared-memory paradigm. It has been implemented using LinuxThreads. This solution is only implemented for the SMP architecture. The distributed implementation has been programmed using MPI libraries as communication primitives between *master* and *slave* processes. The main advantage of these implementations is that they can also be used on SMP architectures, so it has been tested on both architectures.

Two alternative approaches have been developed: distributed and centralized decoding. They differ in the decoding stage as well as in the way data are accessed. In the case of the cluster architecture, since input data are stored in one of the cluster nodes, one solution could be to share the video file among the network nodes via NFS, AFS or any other network file system, but this choice has been discarded because of the overhead introduced by the simultaneous access of the processing nodes to the selected video file. The most suitable solution is a farm based structure where the *master* distributes the workload among the *slave* processes and collects the partial results processed in each *slave* to obtain the video shots. In this case two alternatives are also proposed: static and dynamic data distribution. These strategies can be tested also in the SMP computer.

Distributed decoding approach (DDA). In this approach, each thread is in charge of data access, video decoding and shot detection tasks, so there is no need of a *master* thread. At the beginning each thread has to perform a positioning stage and once they have reached their previously assigned starting place, they can begin to compute the shot detection algorithm. Taking into account that this algorithm is based on an adaptive

 ISBN: 978-84-614-6167-7

threshold computed over a sliding window, each thread must begin its shot detection process some frames before the corresponding place and has to end some frames after it as well. Finally each thread generates an ordered list of detected cuts including the exact first and last frame numbers of each shot.

Centralized decoding approach (CDA). In this approach a *master* thread[1] is in charge of decoding chunks of video data and passing them right after to the *worker* threads using shared-memory. In this case the idea was to process independently both decoding and segmentation stages since the former has a problematic parallelization and, above all, because decoding time is much lower than segmentation time. This solution avoids the bottleneck that appears when multiple threads are performing simultaneous positioning over the same video data, as it happened with **DDA**.

Static data distribution (SDD). The *master* process does an initial and homogeneous data distribution among the *slave* processes. Data package size is obtained dividing the total number of frames in the video by the number of available processors. Process structure in this case is very simple. The *master* begins a decoding loop, sending a complete data package to each *slave*. *Slave* processes will send the results back to the master after finishing the segmentation stage. The master gathers these partial results and stores them.

Dynamic balanced data distribution (DDD). As previously stated, the workload distribution done by **SDD** approach among the available nodes is homogeneous. This is an optimum distribution to work in a dedicated (without any additional task running) and homogeneous cluster. However, when the available nodes in the system have different computing capabilities (due to a hardware heterogeneity or to the simultaneous execution of different tasks), this distribution produces a serious workload imbalance. It implements a dynamic, global and centralized load balancing algorithm. It divides the video into a certain number of work packages. The number of packages, all with the same fixed size, is greater than the number of available nodes in the system. The *master* sends one work package to each available node and waits for the answers. As soon as slaves finish their assigned work they ask for new work packages, that the *master* will send while there are not assigned pending ones. This approach allows the most powerful nodes to process a higher number of work packages, obtaining a quite similar response time for all nodes, regardless their hardware and current workload. This advantage has a more remarkable impact in very dynamic systems where the local workload of the nodes can vary drastically along the execution of the application.

In comparison with **SDD** approach, **DDD** implementation increases the communication overhead since more messages are exchanged. However, it is a quite reduced increase due to the fact that because messages are quite big, the amount of information they carry is always the same, and communication is limited by the available bandwidth on top of other

[1]From now on there will be no difference on the use of the terms thread and process.

 ISBN: 978-84-614-6167-7

factors. For all these reasons, it would be an optimum solution to choose in running time between **SDD** and **DDD** approaches based on both system's dynamism and heterogeneity.

Several advantages of this implementation can be pointed out, such as that it reduces the memory problems that appear in the previous implementation (**SDD** approach), it favors dynamic load balancing among the processing nodes, specially in heterogeneous and non-dedicated systems and minimizes slaves waiting time. On the other hand, this solution results in a more complex implementation.

4 Experimental results

4.1 Experimental setup

Experiments performed on the shared-memory symmetric multiprocessor, **SMP**, have been tested on a SGI Prism 350 machine, with 16 Intel Itanium processors and 32GB DDRAM main memory and 800 GB from 5 Fibre Channel SGI TP9300 hard disks. This is a CC-NUMA based symmetric multiprocessor in which all processors access the main memory through a high-speed bus The cluster setup, **Cluster**, is made up of 1036 eServer BladeCenter JS20 nodes linked together using a Myrinet network for data communication. In our case, regular tests have been performed using up to 64 nodes due to administration issues. Each of the JS20 nodes has 2 IBM Power 970 2.2 GHz processors and 4 GB of main memory.

Tests have consisted on running the optimized video segmentation with polynomials up to 10^{th} order using videos of different length as input, ranged from 10000 to 80000 frames, and using different number of threads and cluster nodes. In **SMP** several setups with 1, 2, 4, 8 and 16 processors have been tested. In **DDA** approach, all processors run the same code. In **CDA** strategy, each *slave* process is assigned to one of the available processors, plus one processor dedicated to run the *master* code. The only exception is the setup with the maximum number of processing nodes, 16, running the *master* and one slave in one of them and 15 *slaves* processes in the rest of them.

Planned tests allow comparing both architectures, **SMP** vs. **Cluster**, both programming paradigms, shared-memory vs. message-passing, and one of the implemented strategies **CDA** vs. **DDA** only in the **SMP** architecture since **DDA** structure does not fit well in the **Cluster** architecture. The other strategy, **SDD** vs. **DDD**, has been tested in both architectures. The main goals of the experiments are: (1) to validate the viability of a parallel solution for the video segmentation application on different architectures and with several parallel programming paradigms; (2) to compare the performance of two alternative architectures, based on shared-memory and on distributed memory, in order to evaluate which one offers the best figures in this application; (3) to compare two parallel computation paradigms, like shared-memory *vs.* message passing programming and different implementation strategies based on data access and data process; (4) to test dynamic data distribution strategy with load balancing mechanisms.

 ISBN: 978-84-614-6167-7

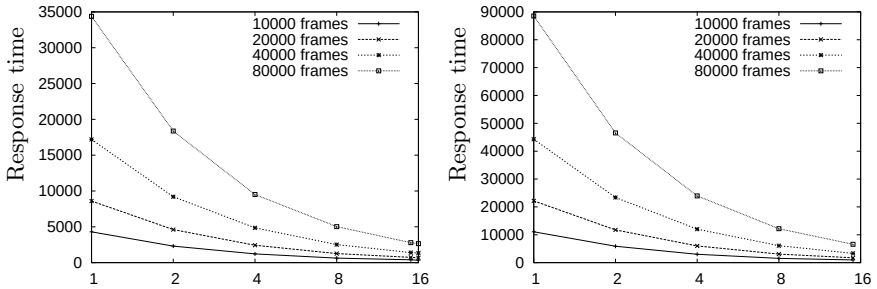


Figure 2: Response time with LinuxThreads on **SMP**: **CDA** vs. **DDA**.

4.2 Comparison among different strategies

4.2.1 Comparison between CDA and DDA in the SMP

Figure 2 presents the evolution of response times for both implementations when the number of processors is increased and several data sizes are considered. Results shown in these figures have been achieved with packages of 157 frames, a value that allows several communication operations among the different processors involved in the experiments. In these cases, figures obtained improve the ones presented as both video and package sizes are increased. All figures show a great reduction of the response times, so it can be deduced that parallel implementations are a good solution to cut down the application response time.

A more detailed analysis of the results achieved by each implementation shows that response time of **CDA** is in all cases at least twice better than **DDA** as can be seen in Fig. 2. Very similar speedup curves are obtained in both cases although response times are quite different. It must be also emphasized that speedup is almost independent of the input data size increment, so it can be concluded that communication overhead is negligible with respect to processing time.

4.2.2 Comparison between SDD and DDD

Performance achieved by both solutions has been tested in both architectures. Results presented correspond to **SMP-Threads-DDA** and **Cluster-MPI-CDA** cases, checking the parallel architecture and the programming paradigm.

SMP

As in the previous case, Figure 3 presents the evolution of the response time for both implementations in **SMP**. Package size for **DDD** is also fixed to 157 frames, while in **SDD** the greatest package size is fixed by the test with the maximum number of nodes (15 in **DDD**). Again, all figures show a great reduction of response times and the speedups are very close to the ideal one, with **SDD** version slightly outperforming **DDD** approach. Table 2 collects experimental results varying the package size in **DDD** for a video size of 80000 frames. With the exception of the setup with 15 processors, the total execution time is

 ISBN: 978-84-614-6167-7

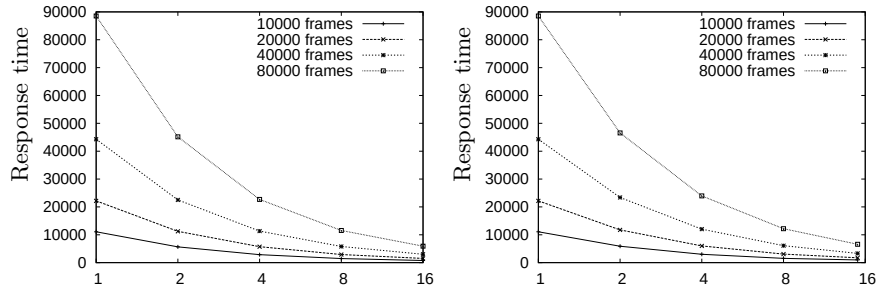


Figure 3: **DDA** Response time with LinuxThreads on **SMP**: **SDD** vs. **DDD**.

Table 2: Execution time (s) for **DDD** compared with **SDD** (80000 frames).

		Packet size					
Method	**No. of processors**	**5000**	**2500**	**1250**	**625**	**313**	**157**
DDD	**2**	44616	44734	44812	45093	45625	46561
	4	22789	22744	22818	22941	23407	23955
	8	11585	11591	11663	11723	11866	12173
	15	11434	8690	7324	6684	6601	6591
SDD	**2**	With max. package size per thread (40000): 45168					
	4	With max. package size per thread (20000): 22687					
	8	With max. package size per thread (10000): 11559					
	16	With max. package size per thread (5000): 5918					

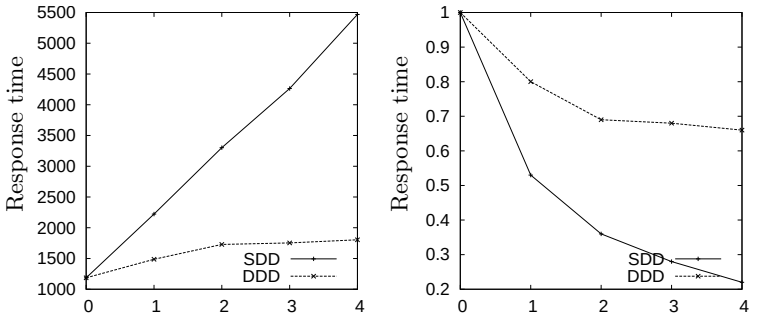
Table 3: Task load distribution when **DDD** introduces load balancing mechanisms.

Initial load	**Processor #**			
(# of tasks)	**[1-4]**	**[5-8]**	**[9-12]**	**[13-16]**
0	8	8	8	8
1	5	9	9	9
2	4	10	9	9
3	3	10	10	9
4	2	10	10	10

reduced while increasing package size, although **SDD** version still beats **DDD** with its maximum package size. **DDD** achieves worse results with 15 processing nodes because the distribution of the workload always involves the processing of the last package in only one of the nodes while the rest are idle, a fact that increases the total execution time.

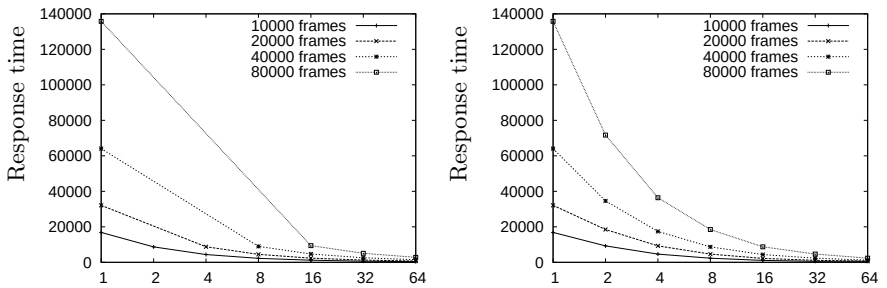
Some of the available processors have been overloaded in order to run load balancing tests. **SMP** has been chosen to perform these tests, although the results obtained can be fully extrapolated since this study does not depend on a particular architecture. Here is presented the performance of the system dealing with 20000 frames and 157 frames per package. Table 3 shows the number of packages processed by each node of the **SMP** architecture when 4 of these processors are executing a limited number of additional tasks, in a range from 0 (no overload at all) to 4 (maximum overload). As it is shown in these values, the workload is distributed among the unloaded processors (from node 5 to 16) depending on the number of packages to distribute from an initial workload of 128 packages. Figure 4 shows the response time and efficiency degradation achieved by **SDD** and **DDD** overloading

 ISBN: 978-84-614-6167-7



(a) **Response time** (b) **Efficiency degradation**

Figure 4: **DDA** response time and efficiency degradation of the LinuxThreads version on **SMP** with load balancing: **SDD** vs. **DDD**.



(a) **SDD** (b) **DDD**

Figure 5: **CDA** Response time with **MPI** on **Cluster**: **SDD** vs. **DDD**.

four nodes in **SMP** as described in Table 3. As it can be noticed in both figures, load balancing improves the performance of the whole system offering good levels of efficiency in the system although there are several nodes dedicated to attend other jobs. However, if the system is homogeneous and not having any external workload, **SDD** will achieve better performance because there is no load balancing overhead. As mentioned above, these experiments can be generalized to other architectures like homogeneous *clusters* with imbalanced nodes or heterogenous *clusters*.

Cluster

Figure 5 presents the evolution of response time for both implementations in **Cluster**. Again, both approaches reach a very important reduction in the response times, obtaining the best values with the largest video sizes and the worst with the smallest ones. For instance, in **DDD**, the increment of the speedup as video size grows can be considered spectacular, reaching the minimum with 10000 frames, surpassing 0.8 with 20000 frames and gaining the maximum with 80000 frames (Fig. 6).

This study has been completed measuring communication times on both implementa-

 ISBN: 978-84-614-6167-7

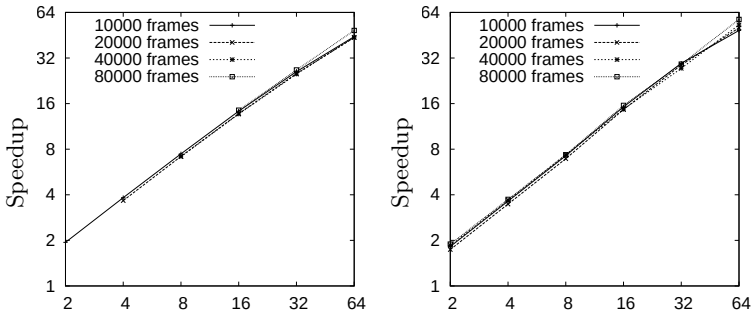


Figure 6: **CDA** Speedup of the **MPI** versions on **Cluster** comparing **SDD** and **DDD**.

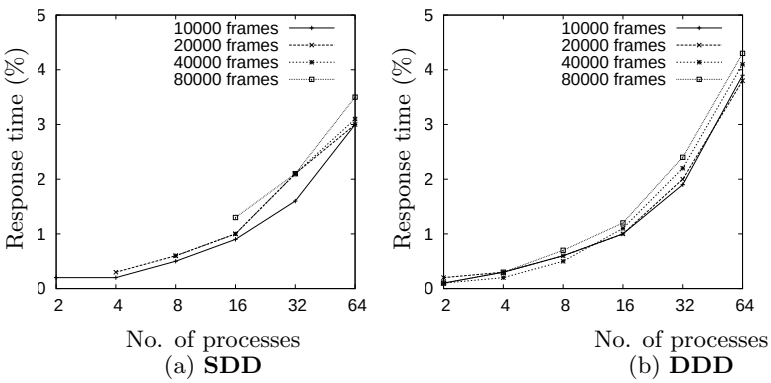


Figure 7: **CDA** Response time percentage dedicated to communication operations of the **MPI** versions on **Cluster** comparing **SDD** and **DDD**.

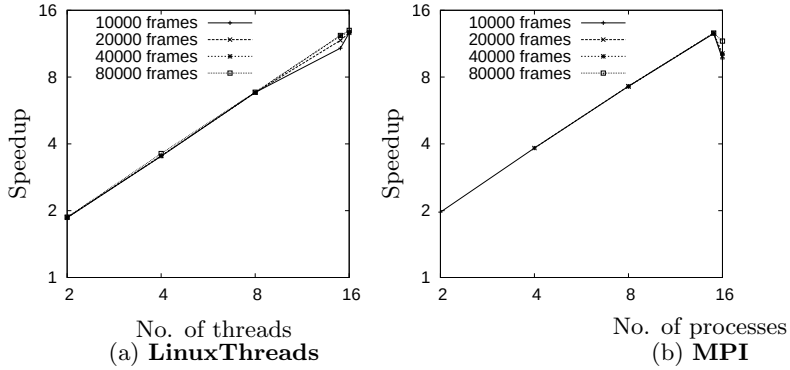
tions. Figure 7 shows the percentage of total execution time devoted to communication data. Communication time only depends on video size, remaining almost constant for all setup tested in each size. On the other hand, Figure 7 shows that the behavior is very similar for all sizes tested. The increment of communication time does not cause the percentage increase shown in Figure 7. This increase is due to a reduction on the processor workload when more processors are involved in the setup.

4.3 Comparison shared-memory vs. message passing

This section presents the results comparing both programming paradigms, shared-memory and message passing, contrasting the branches **SMP-Threads-CDA-DDD** and **SMP-MPI-CDA-DDD**. Figure 8 presents the evolution of speedup for both programming paradigms in **SMP**.

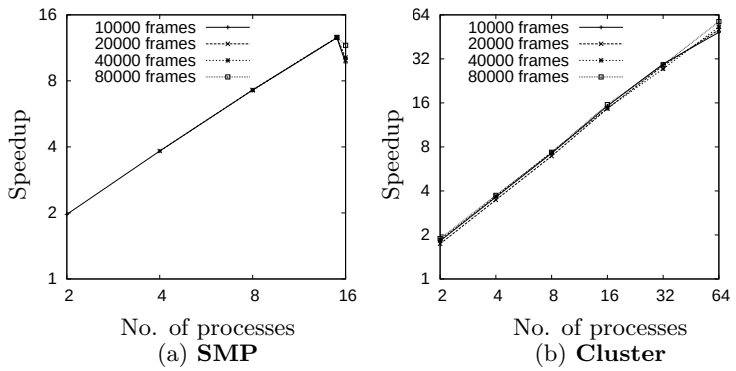
All figures show very outstanding speedup curves. It must be noticed that when the video size is greater than 20000 frames, **LinuxThreads** version improves the results provided by **MPI**. Figure 8(b) clearly shows the degradation that appears when the processor dedicated to run the *master* process in the **MPI** version is shared with an additional 16^{th}

 ISBN: 978-84-614-6167-7



(a) **LinuxThreads** (b) **MPI**

Figure 8: **CDA-DDD** Speedup on **SMP** comparing implementations of shared-memory (**LinuxThreads**) and message-passing (**MPI**) paradigms.



(a) **SMP** (b) **Cluster**

Figure 9: Speedup on both architectures using the same programming paradigm (**MPI**) and the same strategy for distributing the load among the available processors (**CDA**).

slave process, overloading the corresponding **SMP** processor. This means that the load associated to *master* tasks is not negligible, on the contrary, running the main thread of the **LinuxThreads** version in one of the available processors and sharing this processor with one of the launched threads does not diminish speedup figures, as Figure 8(a) shows.

4.4 Comparison between parallel architectures: SMP vs. cluster

Finally, this section presents a comparison between both architectures using the same programming paradigm and the same implemented approach: **SMP-MPI-CDA-DDD** vs. **Cluster-MPI-CDA-DDD**. Figure 9 presents the evolution of the speedup for both architectures in **MPI-CDA-DDD**. Better response times in **Cluster** can be achieved since the number of available processors is greater than in **SMP**, but both architectures can be compared using speedup figures. Again, all figures show a great reduction of response times and very outstanding speedup curves. When the number of slaves in the cluster is increased, the slope increment of the curve is clearly sharper in **Cluster** than in **SMP**, proving a better scalability in the **Cluster** than in **SMP**. A final remark must be made

 ISBN: 978-84-614-6167-7

about Figure 9(a): the above mentioned degradation appearing in **SMP** performance, also shown in Figure 8(b), is due to the overload of the processor that shares one *slave* and the *master* processes.

5 Conclusions and future work

This paper has presented a very exhaustive study among different parallel architectures and parallel programming paradigms of a video shot segmentation algorithm used in a Content-based Multimedia Retrieval System. Programmed implementations attempt to cover all possible parallel programming aspects, just as the different studied paradigms: two LinuxThreads implementations and two MPI versions, tested on a shared-memory symmetric multiprocessor and on a cluster. There are important conclusions that can be extracted from these experiments. Considering the strategy for accessing and segmenting video data, it can be stated that **DDA** has a much simpler implementation since there is one single process involved. Apart from that, **CDA** is clearly better from the point of view of performance and shows a slightly better scalability. Simultaneous access of the threads to the same video data for decoding purposes gives rise to a bottleneck in the I/O subsystem that surpass the penalty introduced when using a single thread dedicated only to decode the video.

The comparison of both distribution strategies, **SDD** vs. **DDD**, when there are no problems with memory accesses, turns out that differences are almost negligible and the simplest one should be chosen (**SDD**). But when video size grows or when dealing with heterogeneous clusters, **DDD** approach is better to introduce load balancing solutions as it has being experimentally shown.

Analyzing both programming paradigms, the performance obtained is very similar in **SMP** with both implementations. LinuxThreads is simpler than MPI from the programmer's point of view, but message passing enjoys portability as an unquestionable advantage over threads, since the same code can be executed in both architectures.

From the experiments it can be deduced that all implementations and architectures tested offer excellent results from a performance point of view, with strong reductions of execution times. Speedup values are almost linear and quite close to the theoretical maximums in all experiments. Shared memory architectures obtain better results with a small number of processors because each one is more powerful than the processing nodes available in the cluster, but are less scalable than clusters.

Beowulf clusters with quite powerful networks, like Myrinet in this case, achieve excellent scalability results. Communication time values remain almost constant when the number of processing nodes is increased, and therefore, the speedup achieved by the parallel system keeps stable when the problem size grows and the number of processing nodes is simultaneously increased.

Future work will include the integration of these systems in a grid infrastructure. Another issue to be considered will be the implementation of the scalable SBD system in other

 ISBN: 978-84-614-6167-7

available many-core architectures, such GPGPUs.

Acknowledgements

This work has been partially funded by the the Spanish Ministry of Education and Science (grants TIN2010-21291-C02-02,CSD2007-00050 and Cajal Blue Brain project).

References

[1] Sameer Antani, Rangachar Kasturi, and Ramesh Jain. A survey on the use of pattern recognition methods for abstraction, indexing and retrieval of images and video. *Pattern Recognition*, 35(4):945–965, April 2002.

[2] Gordon Bell and Jim Gray. What's next in high-performance computing? *Communications of the ACM*, 45(2):91–95, February 2002.

[3] José Luis Bosque, Óscar D. Robles, Luis Pastor, and Angel Rodríguez. Parallel cbir implementations with load balancing algorithms. *Journal of Parallel and Distributed Computing*, 66(8):1062–1075, August 2006.

[4] John L. Gustafson. Reevaluating amdahl's law. *Commun. ACM*, 31(5):532–533, 1988.

[5] Robert A. Joyce and Bede Liu. Temporal segmentation of video using frame and histogram-space. In *Proceedings of the International Conference on Image Processing 2000, ICIP 00*, volume 3, pages 941–944, Vancouver, September 2000. IEEE Computer Society.

[6] Oge Marques and Borivoje Furht. *Content-based Image and Video Retrieval*. Multimedia Systems and Application Series. Kluwer Academic Publishers, Norwell, MA, USA, 2002.

[7] Milan Petkovic and Willem Jonker. *Content-Based Video Retrieval*. Springer, August 2003.

[8] S. V. Porter, M. Mirmehdi, and B. T. Thomas. Video cut detection using frequency domain correlation. In A. Sanfeliu, J. Villanueva, M. Vanrell, R. Alquezar, T. Huang, and J. Serra, editors, *Proceedings of the 15th. Int. Conf. on Pattern Recognition*, volume 3, pages 413–416. IEEE Computer Society, 2000.

[9] Óscar D. Robles, Pablo Toharia, Angel Rodríguez, and Luis Pastor. Towards a content-based video retrieval system using wavelet-based signatures. In M. H. Hamza, editor, *7th IASTED International Conference on Computer Graphics and Imaging - CGIM 2004*, pages 344–349, Kauai, Hawaii, EEUU, August 2004. IASTED, ACTA Press. ISBN: 0-88986-418-7, ISSN:1482-7905.

[10] Nicu Sebe, Michael S. Lew, and Arnold W. M. Smeulders. Video retrieval and summarization. *Computer Vision and Image Understanding*, 92(2–3):141–146, 2003.

[11] Pablo Toharia, Óscar D. Robles, Ángel Rodríguez, and Luis Pastor. A study of zernike invariants for content-based image retrieval. In D. Mery and L. Rueda, editors, *Proceedings of the 2007 IEEE Pacific Rim Symposium on Image Video and Technology, PSIVT2007*, volume 4872 of *Lecture Notes on Computer Science*, pages 944–957, Santiago, Chile, December 2007. IEEE, Springer Verlag.

[12] G. Valencia, J. A. Rodríguez, C. Urdiales, and F. Sandoval. Color-based video segmentation using interlinked irregular pyramids. *Pattern Recognition*, 37(2):377–380, February 2004.

[13] Dong Zhang, Wei Qi, and Hong Jiang Zhang. A new shot boundary detection algorithm. In Heung-Yeung Shum, Mark Liao, and Shih-Fu Chang, editors, *IEEE Pacific Rim Conference on Multimedia*, volume 2195, pages 63–70. IEEE, Springer, October 2001.

 ISBN: 978-84-614-6167-7

Proceedings of the 11th International Conference on Computational and Mathematical Methods in Science and Engineering, CMMSE 2011 26–30 June 2011.

Adaptive artificial boundary conditions for 2D nonlinear Schrödinger equation

Vyacheslav A. Trofimov[1], Anton D. Denisov[1], Zhongyi Huang[2] and Houde Han[2]

[1] *Faculty of Computational Mathematics and Cybernetics, Lomonosov Moscow State University*

[2] *Department of Mathematical Sciences, Tsinghua University*

emails: vatro@cs.msu.su, antondnsv@gmail.com, zhuang@math.tsinghua.edu.cn, hhan@math.tsinghua.edu.cn

Abstract

For linear and nonlinear Shrödinger's equation we propose adaptive non-reflecting boundary conditions based on the analysis of the equation solution near the boundary of the domain. Conservative two-layer finite-difference scheme with two-steps at iteration process is constructed for arbitrary boundary conditions. The efficiency of such scheme is shown for the linear and nonlinear Shrödinger equation.

Key words: conservation laws, finite-difference schemes, adaptive non-reflecting boundary conditions, nonlinear Shrödinger equation, two-steps iterative method.
MSC 2000: AMS codes (optional)

1 Introduction

As it is well-known, one of modern computer simulation problems is a construction of the non-reflecting boundary conditions for equations describing the various physial-chemical process. In particular, developing the such conditions for the wave equation and Shrödinger equation which describes the propagation of electromagnetic waves and energy levels of atoms and molecules is actual at present time (for example [1]-[8]).

In past few years for 1D linear Shrödinger equation have been offered both the exact non-reflecting boundary conditions and the various conditions, which realization leads to appearance of the reflected wave. It should be stressed that for realization of the exact non-reflecting boundary conditions for linear Shrödinger equation the information about its solution on all previous time interval is required. Obviously, it is unacceptable at the numerical solution of the problem. Inaccurate (approximated) non-reflecting boundary conditions don't demand full information about the solution in

 ISBN: 978-84-614-6167-7

time. However, their application leads to occurrence of the reflected wave. In the case of linear propagation of light wave the presence of the reflected wave with amplitude about $3\% - 5\%$ from amplitude of a falling wave can not destroy the solution. At nonlinear propagation of optical beam or pulse the reflected wave with such amplitude leads to essential distortion of the solution because of cross-modulation of falling and reflected waves. Hence, the reflected wave transforms the intensity distribution of falling wave. Therefore, developing the the non-reflecting boundary conditions, possessing small amplitude of the reflected wave (for example, the amplitude is about 0.1% from amplitude of falling wave) and their realization in simple way for computer simulation represents an actual problem essentially in multi-dimensional case.

2 State of problem

The process of laser pulse propagation in a medium with Kerr nonlinearity is described by nonlinear Shrödinger's equation with respect to complex function $A(z, x, y)$:

$$\frac{\partial A}{\partial z} + \imath D_x \frac{\partial^2 A}{\partial A^2} + \imath D_y \frac{\partial^2 A}{\partial y^2} + \imath\gamma |A|^2 A = 0, \;\; z > 0, \; 0 < x < L_x, \; 0 < y < L_y. \qquad (1)$$

The artificial non-reflecting adaptive boundary conditions [8] at the left and right boundaries $\{\Omega_L, \Omega_R\}$ define as:

$$\left(\frac{\partial A}{\partial z} \mp 2D_x \Omega \frac{\partial A}{\partial x} + iD_x \Omega^2 A + \imath\gamma |A|^2 A \right)_{x=0,L_x} = 0, \, \Omega = \{\Omega_{xL}, \Omega_{xR}\},$$

$$\left(\frac{\partial A}{\partial z} \mp 2D_y \Omega \frac{\partial A}{\partial y} + iD_y \Omega^2 A + \imath\gamma |A|^2 A \right)_{y=0,L_y} = 0, \, \Omega = \{\Omega_{yL}, \Omega_{yR}\}. \qquad (2)$$

For the problem (1) there are some invariants (conservative laws). The first invariant can be written in the following manner:

$$I_1(z) = \int_0^{L_x} \int_0^{L_y} |A|^2 dxdy - 2D_x \int_0^{L_z} \int_0^{L_y} \mathrm{Im} \left[A^* \frac{\partial A}{\partial x} - A \frac{\partial A^*}{\partial x} \right]_0^{L_x} dyd\eta -$$
$$- 2D_y \int_0^{L_z} \int_0^{L_x} \mathrm{Im} \left[A^* \frac{\partial A}{\partial y} - A \frac{\partial A^*}{\partial y} \right]_0^{L_y} dxd\eta = const. \qquad (3)$$

 ISBN: 978-84-614-6167-7

The other invariant looks like:

$$I_3(z)=\int\limits_0^{L_x}\int\limits_0^{L_y}\left[\frac{\gamma}{2}|A|^4-D_x\left|\frac{\partial A}{\partial x}\right|^2-D_y\left|\frac{\partial A}{\partial y}\right|^2\right]dxdy+2D_x\int\limits_0^{L_z}\int\limits_0^{L_y}\mathrm{Re}\left[\frac{\partial A}{\partial x}\frac{\partial A^*}{\partial \eta}\Big|_0^{L_x}\right]dyd\eta+$$
$$+2D_y\int\limits_0^{L_z}\int\limits_0^{L_x}\mathrm{Re}\left[\frac{\partial A}{\partial y}\frac{\partial A^*}{\partial \eta}\Big|_0^{L_y}\right]dxd\eta=const. \tag{4}$$

These invariants are used for construction of finite-difference scheme.

3 Construction of conservative finite-difference schemes

To construct a conservative finite-difference scheme for the problem (1), (2) we introduce in the domain $\Omega=(0,L_z)\times(0,L_x)\times(0,L_y)$ the grid $\omega=\omega_z\times\omega_x\times\omega_y$:

$$\omega_z=\{z_k=kh_z,k=0,1,...,N_z,h_z=L_z/N_z\},$$
$$\omega_x=\{x_m=mh_x,m=0,1,...,N_x,h_x=L_x/N_x\},$$
$$\omega_y=\{x_n=nh_y,n=0,1,...,N_y,h_y=L_y/N_y\}.$$

Let us define the grid functions U and introduce the following index-free notation

$$U=U_{m,n}=U(z_k,x_m,y_n),\hat{U}=U(z_{k+1},x_m,y_n),$$
$$\overset{0.5}{U}=0.5(\hat{U}+U),\overset{0.5}{|U|^2}=0.5(|\hat{U}|^2+|U|^2).$$

The second order derivatives in corresponding coordinates are approximated in the following way:

$$\Lambda_{\bar{x}x}U=\frac{U_h(z_k,x_m+h_x,y_n)-2U_h(z_k,x_m,y_n)+U_h(z_k,x_m-h_x,y_n)}{{h_x}^2},$$
$$\Lambda_{\bar{y}y}U=\frac{U_h(z_k,x_m,y_n+h_y)-2U_h(z_k,x_m,y_n)+U_h(z_k,x_m,y_n-h_y)}{{h_y}^2}.$$

Using these notations we write in the internal nodes of mesh the conservative finite-difference scheme with the following iterative method:

$$\frac{\overset{s+1}{\hat{U}}-U}{h_z}+\imath D_x\Lambda_{\bar{x}x}\overset{s+1}{\overset{0.5}{U}}+\imath D_y\Lambda_{\bar{y}y}\overset{s}{\overset{0.5}{U}}+\imath\gamma\overset{s}{\overset{0.5}{|U|^2}}\overset{s}{\overset{0.5}{U}}=0,$$
$$\frac{\overset{s+2}{\hat{U}}-U}{h_z}+\imath D_x\Lambda_{\bar{x}x}\overset{s+1}{\overset{0.5}{U}}+\imath D_y\Lambda_{\bar{y}y}\overset{s+2}{\overset{0.5}{U}}+\imath\gamma\overset{s+1}{\overset{0.5}{|U|^2}}\overset{s+1}{\overset{0.5}{U}}=0. \tag{5}$$

 ISBN: 978-84-614-6167-7

To write the approximation of the boundary conditions (2) let us introduce the additional notations

$$\overset{s+1}{U}_{\bar{z},k} = \frac{\overset{s+1}{\hat{U}}_k - U_k}{h_z}, \overset{s+1}{U}_{x,m} = \frac{\overset{s+1}{0.5}\quad \overset{s+1}{0.5}}{h_x}.$$

Using them we write the following finite-difference equation corresponding to the left boundary condition for differential equation:

$$\overset{s+1}{U}_{\bar{z},0} = 2D\Omega_{xL}\overset{s+1}{0.5}U_{x,0} - iD\Omega_{xL}^2\overset{s+1}{0.5}U_0 - i\gamma|\overset{s}{0.5}U_0|^2\overset{s}{0.5}U_0,\ n = 0,1,...,N_y.$$

The similar condition is written at the right boundary of the domain:

$$\overset{s+1}{U}_{\bar{z},N_x} = -2D\Omega_{xR}\overset{s+1}{0.5}U_{\bar{x},N_x} - iD\Omega_{xR}^2\overset{s+1}{0.5}U_{N_x} - i\gamma|\overset{s}{0.5}U_{N_x}|^2\overset{s}{0.5}U_{N_x},\ n = 0,1,...,N_y.$$

It should be stressed that for brevity we write the finite-difference equations only in x direction.

Coefficients Ω_L and Ω_R is found out numerically. Their values calculate in following way:

$$\Omega_{xL} = -\frac{\psi_1 - \psi_0}{h_x},\ \Omega_{xR} = -\frac{\psi_{N_x} - \psi_{N_x-1}}{h_x},\ n = 0,1,...,N_y,$$

$$\Omega_{yL} = -\frac{\psi_1 - \psi_0}{h_y},\ \Omega_{yR} = -\frac{\psi_{N_y} - \psi_{N_y-1}}{h_y},\ m = 0,1,...,N_x.$$

Above the phase ψ of the complex function U defines by values of its real U_R and image U_L parts and can be calculated as follows: $\psi = 0$ if the inequality $U_R^2 + U_I^2 < 10^{-\delta}$ is valid. Let us underline that the value of δ should be coordinated with data accuracy and calculation accuracy.

The value of the phase ψ for other relations between real and image parts of the complex amplitude is calculated as

$$\begin{aligned}
&for\ \ |U_R| < 10^{-\delta},\ \ \psi = \frac{\pi}{2}\ \ if\ \ U_I > 0,\ \ or\ \ \psi = \frac{3\pi}{2}\ \ if\ \ U_I < 0; \\
&for\ \ U_R > 0,\ \ \psi = arctg\left|\frac{U_I}{U_R}\right|\ \ if\ \ U_I \geq 0,\ \ or\ \ \psi = 2\pi - arctg\left|\frac{U_I}{U_R}\right|\ \ if\ \ U_I < 0; \\
&for\ \ U_R < 0,\ \ \psi = \pi - arctg\left|\frac{U_I}{U_R}\right|\ \ if\ \ U_I \geq 0,\ \ or\ \ \psi = \pi + arctg\left|\frac{U_I}{U_R}\right|\ \ if\ \ U_I < 0;
\end{aligned} \tag{6}$$

It is very important for application that it is necessary to change above the calculation of phase for the case $U_R > 0$ and $U_I \geq 0$ to eliminate an influence of round-off errors. We add to the phase 2π: $\psi \Rightarrow \psi + 2\pi$. Other values of the phase remain the same. Addition of the phase per 2π does not influence on the correctness of calculations. Nevertheless, it allows to eliminate influence of a round-off error.

 ISBN: 978-84-614-6167-7

4 Computer simulation

As example of efficiency of the developing approach we present below the result of computer simulation for a linear ($\gamma = 0$) propagation of input Gaussian laser beam:

$$A|_{z=0}(x,y) = e^{-(x-\theta_x L_x)^2-(y-\theta_y L_y)^2},\ 0 \le x \le L_x\ ,\ 0 \le y \le L_y,\ \theta_x, \theta_y \in [0,1].$$

for $D_x = D_y = 0.25$.

To estimate the error that is caused by artificial boundary conditions we introduce the following norm

$$\xi = \left\| U_L^2 - U^2 \right\|_c = \max_{z,x,y \in \omega} \left| U_L^2 - U^2 \right|,$$

where U_L is also a numerical solution of the Schrödinger equation with zero-value boundary conditions which are valid for domain with increased sizes. Values of both two invariants (3), (4) and error ξ are presented in Table. From the Table we can see that the artificial boundary conditions leads to appearance of error (about 0.05) for the case of symmetric position of the beam center with respect of angular point of the domain boundary.

The intensity distribution at the section $z = 1$ is shown in Figure for other displacement of beam center with respect to the boundary. It is well clear from Figure that the reflected wave is absent.

In the report we discuss the approximation of boundary conditions at angular point of the domain boundary. We consider also the proposed artificial non-reflecting boundary conditions for nonlinear propagation of laser beam. Essentially, that in this case we develop the two-steps iterative method (see (5)) which allows to realize conservative finite-difference schemes for arbitrary boundary conditions.

Table 1. Values of invariants and norm of error for the 2D problem solution.

$\theta_x = \theta_y = 0.9$	$z = 0$	$z = 1$
I_1	1.5315	1.5276
I_3	-0.7150	-0.7325
ξ	0.0000	0.0543

 ISBN: 978-84-614-6167-7

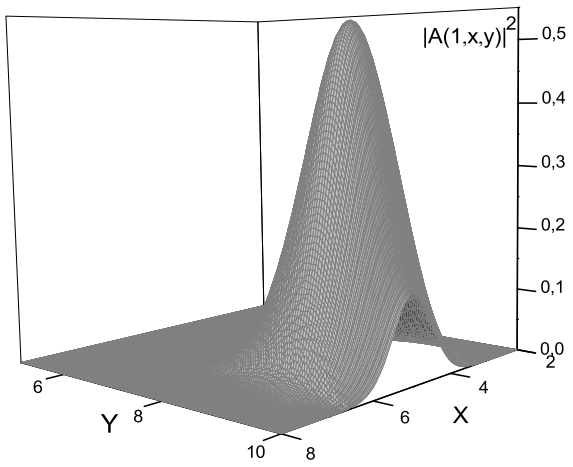


Figure 1: Intensity distribution of laser beam at longitudinal section z=1

Acknowledgements

This paper was partly financially supported by Russian Foundation for Basic Research (grant number 10-07-90004-bel).

References

[1] Alpert B., Greengard L., Hagstram J., *Nonreflecting boundary conditions for the time-dependent wave equation*, J. Comput. Phys. (2002) V. 180. P. 270–296.

[2] Antoine X., Besse C., *Unconditionally stable discretization schemes of non-reflecting boundary conditions for the one-dimensional Schrödinger equation*, J. Comput. Phys. (2003) V. 188. P. 157–175.

[3] Givoli D., *Non-reflecting boundary conditions*, J. Comput. Phys. (1991) V. 94. N1.

[4] Soffer A., Stucchio C., *Absorbing boundaries for the nonlinear Schrödinger equation*, J. Comput. Phys. (2007) V. 225. P. 1218-1232.

[5] Zhenli Xuand, Houde Han., *Absorbing boundary conditions for nonlinear Schrödinger equations*, J. Phys. Rev. (2006) V. 74. 037704.

[6] E. B. Tereshin, V. A. Trofimov and M. V. Fedotov, *Conservative finite-difference scheme for the problem of propagation of a femtosecond pulse in a non-*

 ISBN: 978-84-614-6167-7

linear photonic crystal with nonreflecting boundary conditions, J. Comput. Math. and Math. Phys. (2006) V. 46. N 1. 161–171.

[7] Trofimov V.A., Dogadushkin P.V., *Boundary conditions for the problem of femtosecond pulse propagation in absorption layered structure*, In "Finite Difference Methods: Theory and appli-cations". / Proceedings of Fourth Intern. Conf. / Ed. Farago I., Vabishevich P., Vulkov L. Rouse Univ. "Angel Kanchev". Lozenetz, Bulgaria. (2007) 307–313.

[8] Trofimov V.A., Denisov A.D., *Developing of non-reflecting boundary conditions for nonlinear Schrödinger equation on the base of analysis of solution near the boundary* J. Comput. Math. and Math. Phys. (was sent, 2011) V. 00. N 0. 00–00.

 ISBN: 978-84-614-6167-7

Proceedings of the 11th International Conference on Computational and Mathematical Methods in Science and Engineering, CMMSE 2011 26–30 June 2011.

Effects of the Weight Function Choices on Single-Node Fluctuation Free Integration

Süha Tuna[1] and Metin Demiralp[1]

[1] *Informatics Institute, İstanbul Technical University*

emails: suha.tuna@be.itu.edu.tr, metin.demiralp@be.itu.edu.tr

Abstract

In this paper, a novel integration scheme which uses just a single node will be introduced. To achieve this aim, some weight functions are generated in the integrand of the target integral using the features of fluctuation free integration scheme. Also, a portion of numerical results will be given through graphics to show the efficiency of our approach.

Key words: Fluctuationlessness Theorem, Quadrature, Numerical Integration, Weight Function, Matrix Representations, Orthonormal Basis Sets.

1 Fluctuationlessness Approximation and A Novel Integration Scheme

Fluctuationlessness theorem[1, 2, 3, 4, 5, 6, 7, 8] which was conjectured and proven by M. Demiralp recently, represents an equality between two specific and widely used matrices. The first one of these matrices is the matrix representation of the operator $\widehat{x}$ whose task is to multiply its argument by the universal independent variable x. That is why it may be called "universal matrix" by our group and some others. And the second adverted one is the matrix representation of the $\widehat{f}$ operator which multiplies its argument by the continous function f. The discoursed equality by the way is valid if the space to be worked on is an infinite dimensional space. Then, we should dictate an approximation instead of equality between these two matrices, meaning that if we are working on a subspace instead of original space itself. In other terms, we may state that these two matrices are equal to each other when the fluctuation terms are ignored. As a result of what we have predescribed briefly here, the approximation we mentioned above can be written as follows

$$\mathcal{M}_{\widehat{f}}^{(n)} \approx f\left(\mathcal{M}_{\widehat{x}}^{(n)}\right) \tag{1}$$

 ISBN: 978-84-614-6167-7

where $\mathcal{M}_{\widehat{f}}^{(n)}$ and $\mathcal{M}_{\widehat{x}}^{(n)}$ are the matrix representations of the operators $\widehat{f}$ and $\widehat{x}$ respectively on the n dimensional subspace which can be denoted as $\mathcal{H}_n$ of the infinite dimensional Hilbert space $\mathcal{H}$. The elements of these two matrices are calculated through an inner product defined on $\mathcal{H}_n$, meaning that the functions entering this inner product are square integrable and continous. So the inner product can be expressed as follows

$$(f,g) = \int_0^1 dx w(x) f(x) g(x) \tag{2}$$

The $w(x)$ function appearing in (2) is called the "Weight function" which takes positive real or zero values only at a finite number of points which reside in the unit interval $[\,0,1\,]$. The reason of the selection this unit interval is its convertability meaning that any finite interval can be affine transformed into the interval $[\,0,1\,]$. The other important feature of $w(x)$ which should not be forgotten is the normalization condition which can be represented as

$$\int_0^1 dx w(x) = 1 \tag{3}$$

With the help of the inner product definition in (2), the entries of the matrix representation of $\widehat{x}$ operator and the $\widehat{f}$ operator respectively are calculated as

$$\mathcal{M}_{\widehat{x}}^{(n)} \equiv \begin{bmatrix} X_{11} & X_{12} & \cdots & X_{1n} \\ X_{21} & X_{22} & \cdots & X_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ X_{n1} & X_{n2} & \cdots & X_{nn} \end{bmatrix}; \quad X_{jk} \equiv (u_j, \widehat{x} u_k)\,, \quad 1 \le j,k \le n \tag{4}$$

and

$$\mathcal{M}_{\widehat{f}}^{(n)} \equiv \begin{bmatrix} F_{11} & F_{12} & \cdots & F_{1n} \\ F_{21} & F_{22} & \cdots & F_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ F_{n1} & F_{n2} & \cdots & F_{nn} \end{bmatrix}; \quad F_{jk} \equiv \left(u_j, \widehat{f} u_k\right), \quad 1 \le j,k \le n \tag{5}$$

where the u_i's $(i = 1, 2, ..., n)$ are the orthonormal basis functions[1, 2, 3] spanning $\mathcal{H}_n$ which the very first element of this set, meaning u_1, is the constant function 1. Also, one can easily observe that the matrix $\mathcal{M}_{\widehat{x}}^{(n)}$ has a symmetric structure guaranteeing its all eigenvalues, which can be proven to be distinct and located inside the interval $[\,0,1\,]$, are real.

The approximation in (1) and its components standing in the formulations (2), (4) and (5) can be used to approximate the exact result of a definite integral. A definite integral of the continous function $f(x)$ over the unit interval can be formulated in the most general case as follows

$$\mathcal{I} \equiv \int_0^1 dx f(x) \tag{6}$$

 ISBN: 978-84-614-6167-7

Our purpose is to evaluate this integral numerically and in a desired precision using a quadrature structure which will be called fluctuation free integration. Generally, the entries of the matrix representations of the $\widehat{x}$ and $\widehat{f}$ operators which reside at the intersection of ith row jth column can be obtained by premultiplying corresponding matrix with ${\mathbf{e}_i^{(n)}}^T$ and postmultiplying with $\mathbf{e}_j^{(n)}$ whose formulations are given as follows

$$\begin{aligned} {\mathbf{e}_i^{(n)}}^T \mathcal{M}_{\widehat{f}}^{(n)} \mathbf{e}_j^{(n)} &= \int_0^1 dx u_i(x) x u_j(x), \quad 1 \leq i,j \leq n \\ {\mathbf{e}_i^{(n)}}^T \mathcal{M}_{\widehat{x}}^{(n)} \mathbf{e}_j^{(n)} &= \int_0^1 dx u_i(x) f(x) u_j(x), \quad 1 \leq i,j \leq n \end{aligned} \tag{7}$$

where the $\mathbf{e}_j^{(n)}$ is the n dimensional jth unit vector whose only nonzero element, that is 1, resides at the jth position. By using these entities in (7) and the reality that $u_1(x)$ is 1, we can write down the following formulation using the fluctuationlessness approximation[1, 2, 3].

$$\mathcal{I} \equiv \int_0^1 dx u_1(x) f(x) u_1(x) = \left(u_1, \widehat{f} u_1\right) = \left(u_1, f\left(\widehat{x}\right) u_1\right) \approx {\mathbf{e}_1^{(n)}}^T f\left(\mathcal{M}_{\widehat{x}}^{(n)}\right) \mathbf{e}_1^{(n)} \tag{8}$$

Since the $\mathcal{M}_{\widehat{x}}^{(n)}$ matrix is symmetric, we can represent its spectral decomposition as follows

$$\mathcal{M}_{\widehat{x}}^{(n)} = \sum_{i=1}^{n} \lambda_i \boldsymbol{\xi}_i \boldsymbol{\xi}_i^T \tag{9}$$

where the λ_i's are the eigenvalues and the $\boldsymbol{\xi}_i$'s are the corresponding eigenvectors. Using the spectral decomposition of the matrix representation of $\widehat{x}$ operator in equation (9), we can obtain the image of this matrix under the integrand function, $f(x)$, as follows

$$f\left(\mathcal{M}_{\widehat{x}}^{(n)}\right) = \sum_{i=1}^{n} f(\lambda_i) \boldsymbol{\xi}_i \boldsymbol{\xi}_i^T \tag{10}$$

If we embed this equality into the right hand side of the formulae in (8), we can derive a quadrature-like formulae to approximate the given definite integral in (6).

$$\mathcal{I} \approx \sum_{i=1}^{n} f\left(\lambda_i\right) \left({\mathbf{e}_i^{(n)}}^T \boldsymbol{\xi}_i\right)^2 \tag{11}$$

The expression obtained in (11) is called fluctuation free integration formula[1, 2, 3] for univariate functions.

2 Single-Node Fluctuation Free Integration Using Subspace Construction

Consider the definite integral for the univariate function $f(x)$ in (6). If we apply the fluctuation free integration scheme using n nodes to this integral we get

 ISBN: 978-84-614-6167-7

$$\mathcal{I} \approx w_1 f(x_1) + \cdots + w_n f(x_n) \tag{12}$$

where w_i's refer the squares of the first elements of the eigenvectors, $\boldsymbol{\xi}_i$'s, and x_i's implies the eigenvalues, λ_i's, appearing in the spectral decomposition in (9). If we could handle just one single node instead of the multi-node case above, the approximation to the target integral in (6) and the accompanying conditions, also called the "moment conditions" for the fluctuation free integration (or Gauss quadrature[9]) would have been expressed by the formulations

$$\mathcal{I} \approx w_1 f(x_1) \tag{13}$$

and

$$w_1 = \int_0^1 dx w(x) = \mu_0, \quad w_1 x_1 = \int_0^1 dx w(x) x = \mu_1 \tag{14}$$

respectively. At that point we need to construct a weight function which takes zero values only at a finite number of points and can take only positive values at the rest of the interval and satisfies the moment conditions in (14). For this purpose, if we do a little manipulation in (6) we can obtain the new-shaped integral

$$\mathcal{I} = \int_0^1 dx s(x) \left\{ \frac{f(x)}{s(x)} \right\} \tag{15}$$

which satisfies the following homogenous condition which comes from the moment conditions in (14) when $s(x)$ is replaced by $w(x)$.

$$\int_0^1 dx s(x)(x - x_1) = 0 \tag{16}$$

We can construct a subspace using the homogenous condition (16) but the important property that we desire is to produce positive functions to utilize as a weight function existing in this subspace.

To achieve this goal, we start dealing to produce $s(x)$ by engaging its Taylor expansion. For symmetry reasons, this expansion will be held at the vicinity of $1/2$, that is the center point of the unit interval $[\,0, 1\,]$. So, the expansion can be expressed as

$$s(x) = \sum_{j=0}^{\infty} s_j \left(x - \frac{1}{2} \right)^j \tag{17}$$

where the s_j's come from the Taylor expansion of the relevant function. By the way, the homogenous condition can be decomposed as

$$\int_0^1 dx s(x) \left(x - \frac{1}{2} \right) - \left(x_1 - \frac{1}{2} \right) \int_0^1 dx s(x) = 0 \tag{18}$$

to embed the Taylor expansion in (17) in a more convenient way. If the integrals encountered in the expression reached by arranging the terms of (18) are evaluated analytically the following equation is obtained.

 ISBN: 978-84-614-6167-7

$$\sum_{j=0}^{\infty} s_j \left[\frac{1-(-1)^j}{(j+2)2^{j+2}} - \left(x_1 - \frac{1}{2}\right) \frac{1+(-1)^j}{(j+1)2^{j+1}} \right] = 0 \tag{19}$$

Using this equality we can separate the first term of the summation, so we can express s_0 in terms of the other s_j's as follows

$$s_0 = -\sum_{j=1}^{\infty} s_j \left[\frac{1+(-1)^j}{(j+1)2^{j+1}} - \frac{1}{x_1 - \frac{1}{2}} \frac{1-(-1)^j}{(j+2)2^{j+2}} \right] \tag{20}$$

With the help of the properties discovered above, $s(x)$ function can be reexpressed as follows

$$s(x) = s_0 + \sum_{j=1}^{\infty} s_j \left(x - \frac{1}{2}\right)^j = \sum_{j=1}^{\infty} s_j p_j(x, x_1) \tag{21}$$

where the p_j coefficients[4, 5] are indicated evidently

$$p_j(x, x_1) = \left(x - \frac{1}{2}\right)^j + \frac{1}{\left(x_1 - \frac{1}{2}\right)} \frac{1-(-1)^j}{(j+2)2^{j+2}} - \frac{1+(-1)^j}{(j+1)2^{j+1}}; \quad j = 1, 2, \ldots \tag{22}$$

Since the number of p_j's are infinite, there is no way to investigate the positivity on the unit interval $[\,0, 1\,]$ for all of them. For this reason we will handle the very first of them, also the one which has the simplest to calculate structure, p_1. If we substitute 1 for j in (22),

$$p_1(x, x_1) = \left(x - \frac{1}{2}\right) + \frac{1}{\left(x_1 - \frac{1}{2}\right)} \frac{1}{12} \tag{23}$$

is obtained. This is a polynomial of first degree and has the dependency to x_1[4, 5]. Since this polynomial will be considered as a weight function on the unit interval, it must satisfy the normalization condition under integration

$$\int_0^1 dx w_1(x, x_1) = 1 \tag{24}$$

With the existence of the condition (24), the weight function[4, 5] which will be utilized is committed as

$$w_1(x, x_1) = 12 \left(x_1 - \frac{1}{2}\right) p_1(x, x_1) = 1 + 12 \left(x_1 - \frac{1}{2}\right) \left(x - \frac{1}{2}\right) \tag{25}$$

Using the weight function obtained in (25), we can realize an approximation to the target integral in (6) except the x_1 points that makes p_1 as zero. This fact entails a lack of continuity of the integrand in (15) as it can be easily noticed. By the way, we can say that the values of p_1 must be less then 0 and greater than 1 for all x_1's in

 ISBN: 978-84-614-6167-7

an appropriate open interval. With a simple inequality analysis for the root of p_1, the interval from which x_1 values should be chosen is determined as

$$\frac{1}{3} < x_1 < \frac{2}{3} \tag{26}$$

which dictates us a restriction that the x_1 points should be taken from the middle one third of the unit interval. If the x_1 value is considered using the inequality in (26) the approximation to the integral in (6) can be realized as follows

$$\mathcal{I} \approx \frac{1}{w_1(x_1)} f(x_1) \tag{27}$$

Although this approximation formula is valid for only x_1 values residing in the interval $(1/3, 2/3)$, it is possible to make this interval broader using an appropriate transformation. For example, if the transformation $x \to 1 - x^{k+1}$ is applied to the process above from the beginning, the value space of the node x_1 can be extended to the interval $(e^{-2}, 1 - e^{-2})$ while k goes to infinity[4, 5]. The details of this circumstance will not be given here, since this fact is out of the scope of our paper.

3 The Generation and The Utilization of Non-Polynomial Weight Functions

In the previous section, we have dealt with the weight function which has the polynomial nature. Beyond this situation, generating a weight function of non-polynomial structure is also possible. The only thing we need is to specify the main structure of the weight function and decide the x_1 values according to that new structure constructed by using the homogenous condition in (16). Then, with the help of the x_1 value obtained, we can approximate to the target integral by using the formulae in (27). So, we consider the structure of the weight function as

$$s(x, \tau) = e^{\tau\left(x - \frac{1}{2}\right)} \tag{28}$$

where τ is a real constant multiplied with the factor $(x - 1/2)$ for symmetry reasons. If this weight function is embedded into the homogenous condition in (16)

$$\left(x_1 - \frac{1}{2}\right) \int_0^1 dx s(x, \tau) - \int_0^1 dx s(x, \tau) \left(x - \frac{1}{2}\right) = 0 \tag{29}$$

By evaluating the above integrals analytically and making some reorganizations

$$x_1 = \frac{1}{2} - \frac{1}{\tau} + \frac{1}{2} \coth \frac{\tau}{2} \tag{30}$$

is obtained. We can produce more weight functions and nodes by changing the structure of the weight function s.

 ISBN: 978-84-614-6167-7

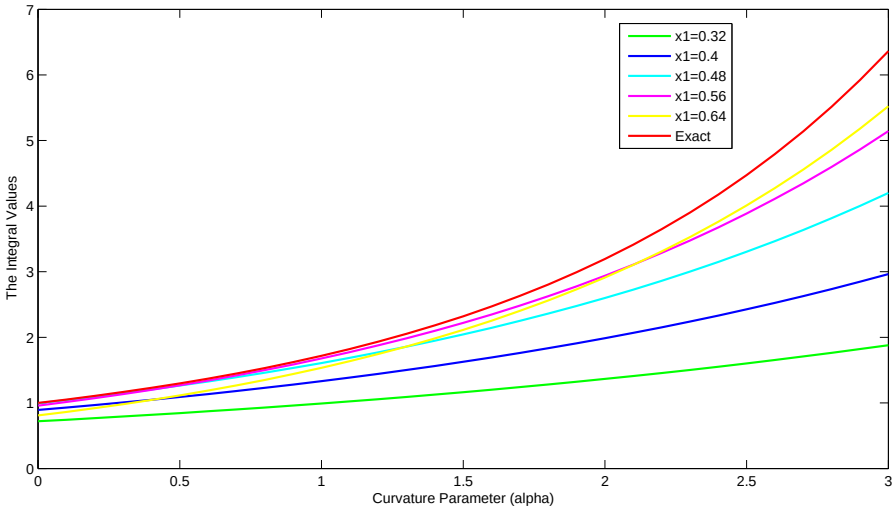


Figure 1: Exact and approximation results for the integral of $e^{\alpha x}$ function using polynomial weight

4 Numerical Implementations

In this section, we will give some graphics reflecting the numerical calculations[10] and accuracy for the integral values of the elementary functions such as $e^{\alpha x}$, $e^{-\alpha x}$ and $\sin(\alpha x)$ for different positive and real α values obtained by our method. One can easily verify from the structures of the functions above that α value is the parameter which determines the curvature of the exponential functions and oscillation for the trigonometric functions.

In Figure 1, the approximate values evaluated using single node fluctuation free integration and the exact result for the integral of $e^{\alpha x}$ on the unit interval are plotted for different x_1 values to be seen together. It is obvious that the best approximation to the exact value of the integral is obtained by using the greatest x_1. But on the other hand, if we look at Figure 2 which has been plotted for $e^{-\alpha x}$ with the same α and x_1 values, one can easily notice that the best approximation is provided by taking x_1 as its smallest value. These results are encountered because of the monotonicity of the functions under consideration. Thus, it can be noticed that, the x_1 values which are closer to the right bound of the interval express the integral value of monotonously increasing functions. On the other hand, the x_1 values which are closer to the left border of the interval works better to approximate the integral of monotonously decreasing functions using polynomial weight function structure.

In Figure 3, Figure 4 and Figure 5 we realized the calculations for an exponential function having less curvature, again an exponential function with high curvature and a trigonometric function with the help of the exponentially structured weight function, respectively. It can be noticed from the Figure 3 that, using exponential weight function

 ISBN: 978-84-614-6167-7

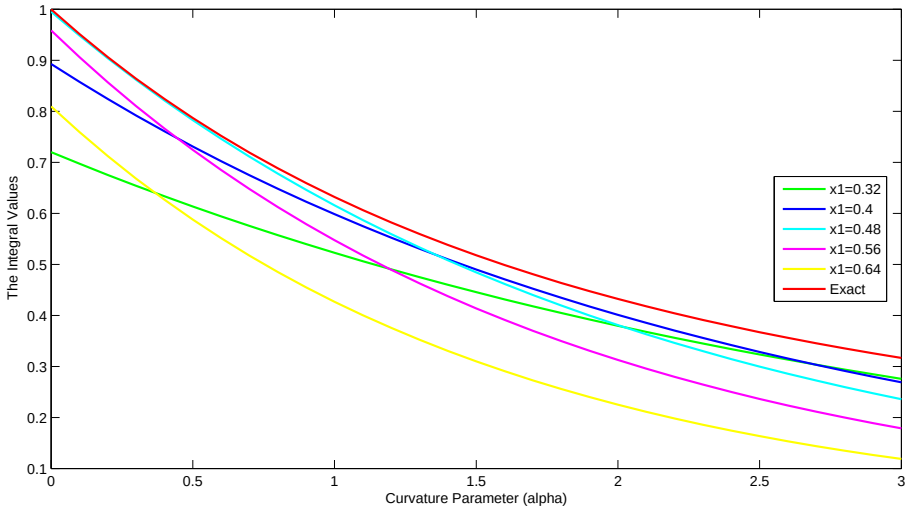


Figure 2: Exact and approximation results for the integral of $e^{-\alpha x}$ function using polynomial weight

resulted better approximations for the exponential integrands according to the values shown in Figure 1. If the curvature level of the exponential function, meaning α, increases we can overcome this situation by increasing the curvature of the weight function, meaning τ. In this manner, we can produce more precise approximations. Of course the exponentially structured weight function is not a cure-for-all phenomena for single node fluctuation free integration. This fact can be seen in Figure 5 evidently. For example, to approximate a trigonometric function, we can derive a convenient formula using a trigonometric type weight function. Although we mentioned about this case, the implementations will be discussed and possible pitfalls will be mentioned during the presentation.

5 Concluded Remarks

The outcomes of what we have encountered during the development of the theory and the calculations can be itemized as follows.

- If the function under consideration is flat, meaning almost constant, then choosing the node x_1 in the vicinity of $\frac{1}{2}$ gives better approximation.
- If the function is monotonously increasing, then choosing x_1 close to the right border of the interval will increase the approximation quality.
- If the integrand of the target integral has a monotonous decreasing structure then working with the nodal points which are close to the left border of the interval will cause better approximations.

 ISBN: 978-84-614-6167-7

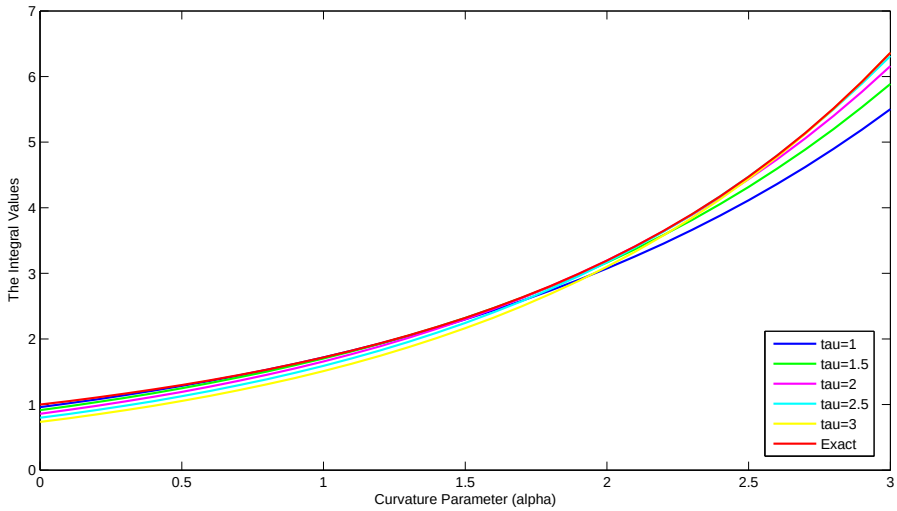


Figure 3: Exact and approximation results for the integral of $e^{\alpha x}$ function using exponential weight

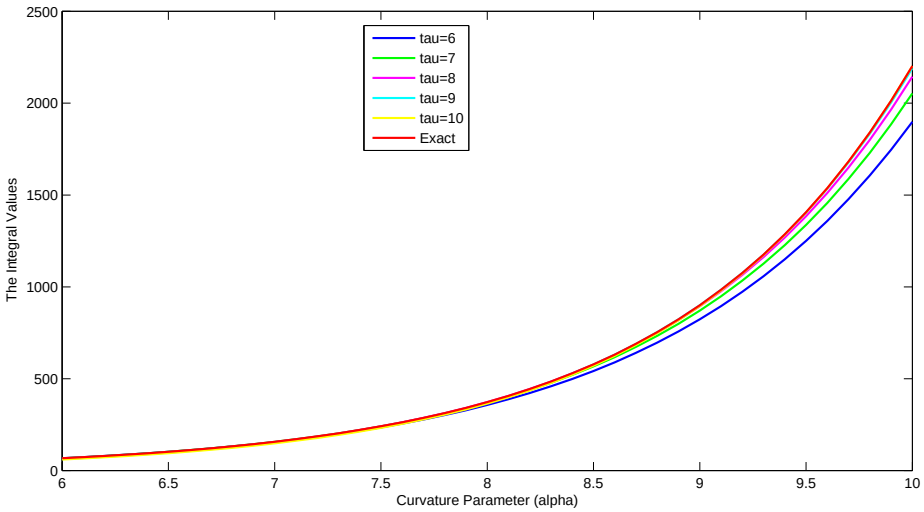


Figure 4: Exact and approximation results for the integral of $e^{\alpha x}$ function using exponential weight

 ISBN: 978-84-614-6167-7

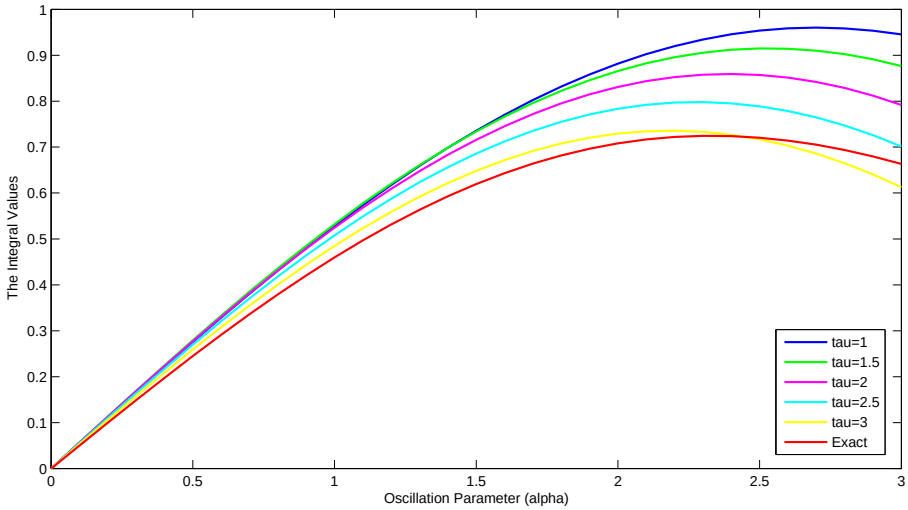


Figure 5: Exact and approximation results for the integral of $\sin(\alpha x)$ function using exponential weight

- If the analytic structure of the function is already known, we can compare the results and find the best approximation. To find the best nodal location is a function dependent issue since it changes from function to function.
- If the $s(x)$ function is structured according to the function to be integrated then the approximation for the target integral will be more precise.

Acknowledgements

The first author thanks to İstanbul Technical University for its support and the second author is grateful to the Turkish Academy of Sciences for its support and motivation.

References

[1] M. Demiralp *Fluctuationlessness Theorem to Approximate Univariate Function's Matrix Representations* WSEAS Transactions on Mathematics **8**, Issue 6 (2009), 258–267

[2] M. Demiralp *A New Fluctuation Expansion Based Method for the Univariate Numerical Integration under Gaussian Weights* WSEAS Transaction on Mathematics **68**, (2005)

[3] M. Demiralp, *No Fluctuation Approximation in any Desired Precision for Univariate Matrix Representations* J. Math. Chem. **47** (2010) 99–110.

 ISBN: 978-84-614-6167-7

[4] A. Kurşunlu, M. Demiralp, *A Fluctuation Removal Based Univariate Integration Over Prescribed Nodes - Certain Important Aspects of One Node Fluctuation Free Integration*, J. Math. Chem. **49** (2011) 407–427.

[5] A. Kurşunlu, M. Demiralp, *A Fluctuation Removal Based Univariate Integration Over Prescribed Nodes - One Node Fluctuation Free Integration Under Higher Order Set Of Conditions With An Extension To Multinode Case*, J. Math. Chem. **49** (2011) 428–443.

[6] B. Kalay, M. Demiralp *Smoothening the Integrands to Increase the Quality of Fluctuationlessness Approximation in Numerical Integration* Proceedingd of the 1st WSEAS Int. Conf. on Multivariate Anal. and Its Appl. in Science and Engineering (2008), 204–208

[7] S. Tuna, B. Tunga, N. A. Baykara, M. Demiralp *Fluctuation Free Matrix Representation Based Univariate Integration in Hybrid High Dimensional Model Representation (HHDMR) Over Plain and Factorized HDMR* WSEAS Transactions on Mathematics **8**, Issue 6 (2009), 225–230

[8] N. Altay, M. Demiralp *Numerical Solutions of Ordinary Differential Equations by Fluctuationlessness Theorem* J. Math. Chem. **47**, Number 4 (2010), 1323–1343

[9] F. B. Hildebrand *Introduction to Numerical Analysis* ISBN-13: 978-0-486-65363-1 (1987)

[10] B. Hunt, R. Limpson, J. Rosenberg with K. Coombes, J. Osborn and G. Stuck *A Guide to MATLAB for Beginners and Experienced Users, Cambridge University Press*, (2000).

 ISBN: 978-84-614-6167-7

Proceedings of the 11th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2011
26–30 June 2011.

Bilevel E Cost-Time-P Programming Problems

Oana Ruxandra Tuns (Bode)[1]

[1] *Faculty of Mathematics and Computer Science, Babeş-Bolyai University*

emails: oana.tuns@ubbcluj.ro

Abstract

The present paper reflects a problem which is rooted in a mathematical model corresponding to some particular portfolio problems. The studied problem belongs to the class of bilevel optimization problems, in which the objective function of the lower level problem is a cost-time type vector function. The problem restraints contain the condition that the variable have boolean values. Based on the particularities of the problem we solve it by solving two pseudo-boolean optimization problems: the first problem being knapsack type; the second one has an objective function which is a cost-time type vector function and has knapsack type restraints.

Key words: bilevel optimization problem, portfolio optimization
MSC 2000: 90C29, 90C90, 90B50, 91B06

1 Introduction

Nowadays, many economic problems can be solved by using mathematical tools. In this way, financial optimization became one of the most interesting areas in investment processes. The literature on financial optimization models is rooted in the work of Markowitz [12]. His main contribution was the introduction of the concept of efficient or optimal portfolio, that would provide the maximum expected return for a certain level of risk each investor is willing to take. Since the publishing of Markowitz's work, a lot of effort has been made to apply this approach to a larger scale of real life situations and to make investment decision more measurable.

We mention that a financial investment is represented by any capital investment in order to gain a profit. Whenever an investor decides to invest, he/she thinks in terms of "is it worth it?" and "when should I stop?". The answer to the first question can be different for each investor: how much the maximum return will be, in how much time the invested amount will be recovered, etc.. The answer to the second question is related to the goals of each investor. This means that the investor has to establish each time both realistic and achievable goals. The time associated with each goal is one of the factors that affects the investment strategy. By grouping the goals based on

 ISBN: 978-84-614-6167-7

time periods, three types of investment can be determined: short-term investment (less than one year), medium-term investment (one to five years) and long-term investment (five years and beyond).

Since investment can be seen as a link between present and future, time represents an important factor in the investment process (time is immanent for any investment). In addition, the risk (seen as a result of the uncertainties) is inherent for any investment. Therefore, time periods and uncertainties play important roles in the investment process. From a mathematical perspective, multilevel programming models must be applied in order to emphasize both these aspects and resolve the investment problems.

The present paper proposes an approach to resolve a concrete economic problem faced by an investor (a company), regarding it's decision on a medium-term investment. An economic problem has been formulated and the author proposes a method for solving this concrete problem.

2 Our Particular Portfolio Problem

The main objective of a firm S, which owns n branches S_1,...,S_n, is to obtain a largest return by investing in different stock portfolios available on the capital market. We mention that the stocks enclosed in a portfolio have the same quotation on the market. We define as *different stock portfolios* each portfolio available on the capital market that has a different market quotation in time. For some stock portfolios, the quotation has an ascending trend. For other stock portfolios the price increases or decreases frequently and significantly, even if the long-term trend is known as being ascending, descending or constant.

Let P_1, ..., P_m, $m > n$, be the stock portfolios in which the firm S will invest. For each stock portfolio P_i, $i \in \{1, \ldots, m\}$, the firm S has historical data based on which it can predict the expected return for a certain level of risk undertaken for a period of time T. It can also predict the expected return for a certain level of risk undertaken for shorter periods of time, subdivisions T_h, $h \in \{1, \ldots s\}$, of T.

The firm S can make transactions with the stock portfolios in two different ways: directly, through its n branches, and indirectly, through p companies, $m - n \leq p \leq (m-n)s$, denoted by C_1,... C_p, within a group of companies, denoted by C, specialized in financial investment services.

The problem that firm S needs to solve is how to choose n stock portfolios in which to invest directly (one portfolio through each branch) so that it maximizes its return, while the biggest risk for which it obtains the maximum return shall not exceed a defined value e. We mention that the firm's return is equal to the sum of direct return, achieved through the investment made through its branches, and indirectly return, achieved through investment made by specialized investment companies. In the second case, the company engaged by the firm to invest on its behalf will get its own return from the transactions made and, based on an agreed share will yield a part of the return to the firm.

Let a_{ij} be the expected return that the firm S could obtain if they trade the stock

 ISBN: 978-84-614-6167-7

portfolio P_i, $i \in \{1, \ldots, m\}$ through its branch S_j, $j \in \{1, \ldots, n\}$. Let r_{ij} be the risk taken by the firm S, through its branch S_j, in the above case.

The investment company C must invest in all stock portfolios that were not chosen by the firm S to invest directly. The investment company considers the medium period of time T divided in s subperiods T_1,..., T_s. It is known the expected return c_{ikh} that the company C_k, $k \in \{1, \ldots, p\}$ can obtain if it will trade the stock portfolio P_i, in the subperiod of time T_h, $h \in \{1, \ldots, s\}$; also, it is known the risk d_{ikh} taken by the company C_k, if it will trade the stock portfolio P_i, in the subperiod of time T_h. Let $q_{ik}\%$ be the share (percentage) that the firm S gains from the total return obtained by the company C_k, if it will trade on market the stock portfolio P_i.

The decision on what company C_k will trade the stock portfolio P_i will be made by the management of the investment company C, and not by the firm S.

The main objective of the investment company C is to obtain a biggest net return, the condition being that each stock portfolio, that was not chosen by the firm S, must be traded at least once. The same stock portfolio can be traded in several of the s subperiods of time but only once in the same subperiod.

The return of the investment company C is obtained by summing up the net returns obtained from the transactions made by each company C_k, $k \in \{1, \ldots, p\}$. The return of one company C_k is obtained by summing up the returns obtained in each subperiod of time T_h, $h \in \{1, \ldots, s\}$ and decreasing the share agreed to be yielded to firm S. The management of the investment company C requires that the biggest risk taken for the stock portfolios traded in a period of time T_h, $h \in \{1, \ldots, s\}$, shall not exceed a defined value e_h, shall be as small as possible and impacts the smallest number of stock portfolios. They also request that each company C_k must trade at least one stock portfolio in at least one subperiod of time.

The mathematical model for our portfolio problem is a bilevel type problem.

3 Brief Background of Bilevel Programming Problems

Nowadays, multilevel programming, and subsequently bilevel programming, has become an important area in optimization. These types of problems are strongly motivated by real world applications in economics, medicine, engineering etc.

In mathematical terms, the bilevel programming problem is an optimization problem where a subset of the variables is constrained to be an optimal solution of a given optimization problem parameterized by the remaining variables.

Let be $\Omega \subseteq \mathbf{R}^n \times \mathbf{R}^m$ and let $f : \Omega \to \mathbf{R}$, $g : \Omega \to \mathbf{R}$ be given functions. We set

$$\Omega_1 = \{x \in \mathbf{R}^n \,|\, \exists\, y \in \mathbf{R}^m \text{ such that } (x, y) \in \Omega\},$$

and, for each $x \in \Omega_1$, we denote by

$$\Omega_x = \{y \in \mathbf{R}^m \,|\, (x, y) \in \Omega\}.$$

 ISBN: 978-84-614-6167-7

We can mathematically formulate the bilevel programming problem as:

$$(BP) \quad \begin{cases} f(x,y) \ \rightarrow \ \max \\ \text{such that} \\ (x,y) \in \Omega, \\ y \in S(x), \end{cases} \tag{1}$$

where $S(x)$ denotes the set of optimal solutions of the mathematical programming problem parameterized in x, i.e.,

$$S(x) \ = \ \text{argmax}\{g(x,y) \,|\, y \in \Omega_x\}. \tag{2}$$

The original formulation for the bilevel programming problem appeared in 1973, in a paper authored by J. Bracken and J. McGill ([2]). But, the bilevel programming problem has its origins in the work of H.F. von Stackelberg ([14] or [13]). He used for the first time an hierarchical model to describe real market situations. This model reflects that there can be market situations when different decision makers are not able to make their decisions independently, being forced to act according to a certain hierarchy. The simple case of such a situation is the one when there are only two decision makers on the market: the leader (upper level) and the follower (lower level) and, more over, the lower level actions depend on upper level decisions.

Although the bilevel programming problem was first introduced by J. Bracken and J. McGill, it was W. Candler and R. Norton ([4]) that first used, in 1977, the term of optimization problem on more levels and, subsequently, bilevel optimization problem, for this type of optimization problems.

References [16], [6], [5] are useful paper studies concerning bilevel programming. In [17] one can find a detailed bibliography of works in the field of bilevel and multilevel programming problems. Also, reference [1] provides general aspects of bilevel optimization problems (the real objective functions are replaced with multiobjective functions).

In time, multilevel optimization problems were used for modeling many types of concrete problems, of which we recall: the network design problem ([3]), optimal pricing problem ([11]), the optimal signal setting problem ([10]) and train set organization ([8]).

Among other works in the field of bi and multilevel programming we cite: L. Vicente, G. Savard and J. Júdice [18], who studied the linear bilevel programming problems, D. Duca and L. Lupşa [7], who studied transport bilevel problems, and St. Kosuch, P. Le Bodic, J. Leung and A. Lisser [9], who treated the stochastic aspect.

4 The Mathematical Model for our Portfolio Problem

Let us denote by

$$I = \{1, \ldots, m\}, \ J = \{1, \ldots, n\}, \ K = \{1, \ldots, p\}, \ H = \{1, \ldots, s\}.$$

Let us denote by x_{ij}, $i \in I$, $j \in J$ the binary variable having the significance $x_{ij} = 1$, if the firm S trades, through its branch S_j the stock portfolio P_i, and $x_{ij} = 0$,

 ISBN: 978-84-614-6167-7

otherwise. Let us denote by y_{ikh}, $i \in I$, $k \in K$, $h \in H$, the binary variable having the significance $y_{ikh} = 1$, if the company C_k trades the stock portfolio P_i in the subperiod of time T_h, and $y_{ikh} = 0$, otherwise.

Any direct investment of the firm S will be expressed by a matrix $X = [x_{ij}] \in \{0,1\}^{m\times n}$, and any indirect investment of the firm S will be expressed by a matrix $Y = [y_{ikh}] \in \{0,1\}^{m\times p\times s}$. Therefore, an investment of the firm S will be a pair (X, Y), where X represents a direct investment of the firm and Y represents an indirect investment of the firm.

The return of the firm S is given by the function $f : \{0,1\}^{m\times n} \times \{0,1\}^{m\times p\times s} \to \mathbf{R}$,

$$f(X,Y) = \sum_{i\in I}\sum_{j\in J} a_{ij}x_{ij} + \sum_{i\in I}\sum_{k\in K}(q_{ik}\sum_{h\in H} c_{ikh}y_{ikh}), \tag{3}$$

for all $(X = [x_{ij}], Y = [y_{ikh}]) \in \{0,1\}^{m\times n} \times \{0,1\}^{m\times p\times s}$.

The biggest risk taken by a direct investment $X = [x_{ij}]$ is equal to:

$$\max\{r_{ij}x_{ij} \,|\, i \in I, j \in J\}.$$

Therefore, the condition that the risk taken by the firm S shall not exceed the defined value e enforces the restraint

$$\max\{r_{ij}x_{ij} \,|\, i \in I, j \in J\} \le e. \tag{4}$$

The condition that each branch must trade one stock portfolio enforces the following restraint

$$\sum_{i\in I} x_{ij} = 1, \text{ for each } j \in J, \tag{5}$$

and the condition that the firm S must invest in all m stock portfolios on the market either directly or through the investment company C, enforces the following restraint

$$\sum_{j\in J} x_{ij} + \sum_{k\in K} \operatorname{sgn}(\sum_{h\in H} y_{ikh}) \ge 1, \text{ for each } i \in I. \tag{6}$$

The biggest risk of the investment company C, taken in the subperiod of time T_h, $h \in H$, is equal to

$$\max\{d_{ikh}y_{ikh} \,|\, i \in I,\, k \in K\}.$$

Therefore, the condition that the risk taken by the company C, shall not exceed the defined value e_h, in the subperiod of time T_h, enforces the restraint

$$\max\{d_{ikh}\operatorname{sgn} y_{ikh} \,|\, i \in I, k \in K\} \le e_h, \text{ for each } h \in H. \tag{7}$$

In terms of the investment company C, the return is given by the function $g_1 : \{0,1\}^{m\times p\times s} \to \mathbf{R}$,

$$g_1(Y) = \sum_{i\in I}\sum_{k\in K}(1 - q_{ik})(\sum_{h\in H} c_{ikh}y_{ikh}), \tag{8}$$

for all $Y = [y_{ikh}] \in \{0,1\}^{m\times p\times s}$.

 ISBN: 978-84-614-6167-7

The function $g_2 : \{0,1\}^{m\times p\times s} \to \mathbf{R}$,

$$g_2(Y) = \max\{d_{ikh}\text{sgn}y_{ikh} \,|\, i \in I, k \in K, h \in H\}, \tag{9}$$

for all $Y = [y_{ikh}] \in \{0,1\}^{m\times p\times s}$, gives us the maximum risk that the investment company C is willing to take. Thus, the vector function $g = (g_1, g_2)$ is the objective function of the investment company C.

Due to the fact that each company C_k trades at least once a stock portfolio and in the same subperiod of time only one company from the group must trade one stock portfolio. Therefore, the following restraints are imposed:

$$\sum_{i\in I}\sum_{h\in H} y_{ikh} \geq 1, \text{ for each } k \in K, \tag{10}$$

$$\sum_{k\in K} y_{ikh} \leq 1, \text{ for each } h \in H,\ i \in I. \tag{11}$$

Let us denote

$$\Lambda = \{X = [x_{ij}] \in \{0,1\}^{m\times n} | \sum_{i\in I} x_{ij} = 1, \forall\, j \in J, \sum_{j\in J} x_{ij} \leq 1, \forall\, i \in I\}. \tag{12}$$

The set Λ represents the set of all possible direct investments of the firm S. Indeed, for each portfolio P_i, the firm S can invest in it by at most one of its branches. So, $\sum_{i\in I} x_{ij} \leq 1, \forall\, i \in I$. On the other hand, each branch of the firm S must choose only one portfolio to invest in it. So, $\sum_{i\in I} x_{ij} = 1, \forall\, j \in J$.

For the mathematical model of our portfolio problem we will use the max-min-p points, presented in the paper [15].

4.1 Max-min-p points

Let $A \subseteq \mathbf{N}^n$ be a non empty set and $\varphi = (\varphi_1, \varphi_2) : \mathbb{N}^n \to \mathbf{N}^2$ a given function, where φ_2 is of time type, i.e. there is a vector $\tau = (\tau_1, \ldots, \tau_n) \in \mathbb{N}^n$, such that $\varphi_2(x) = \max\{\tau_j \cdot \text{sgn}x_j | j \in \{1, \ldots, n\}\}$, for all $x \in \mathbb{N}^n$.

As the set $TM = \{\tau_j \,|\, j \in J\}$ is a finite set, we can number its elements. If $\text{card}\, TM = q$, and we denote by z_i, $i \in \{1, ..., q\}$, its elements, then

$$TM = \{z_1, ..., z_q\}. \tag{13}$$

Let be

$$L_k = \{i \in \{1, \ldots n\} \,|\, \tau_i = z_k\}, \text{for every } k \in \{1, ..., q\}, \tag{14}$$

Let $A \subseteq \mathbb{N}^n$ a finite set.

A point $a^0 \in A$ is said to be a $\max - \min -$p point of A with respect to the function φ_2 if for every $a \in A$, $a \neq a^0$, the following conditions are satisfied:

H1:

$$\varphi_1(a^0) \geq \varphi_1(a); \tag{15}$$

 ISBN: 978-84-614-6167-7

H2: If

$$\varphi_1(a^0) = \varphi_1(a), \tag{16}$$

then

$$\varphi_2(a^0) \leq \varphi_2(a); \tag{17}$$

H3: If

$$\varphi_1(a^0) = \varphi_1(a) \text{ and } \varphi_2(a^0) = \varphi_2(a) \, = \, z_s, \tag{18}$$

then or

$$\sum_{i \in L_k} a_i^0 \, = \, \sum_{i \in L_k} a_i \quad \text{for all} \quad k \in \{s, ..., p\}. \tag{19}$$

or there is a natural number $r \in \{s, ..., p\}$ such that

$$\sum_{i \in L_k} a_i^0 \, = \, \sum_{i \in L_k} a_i \quad \text{for all} \quad k \in \{s, ..., r-1\}$$

$$\text{and} \tag{20}$$

$$\sum_{i \in L_r} a_i^0 \, < \, \sum_{i \in L_r} a_i.$$

We will denote the set of all points of A which are max-min-p points by

$$arg - \max - \min - \mathrm{p}\{\varphi_2(x) | x \in A\}.$$

Recalling our problem, for each $X \in \Lambda$, we set

$$\begin{aligned} U^X \; = \; & \left\{ Y \in \{0,1\}^{m \times p \times s} | \operatorname{sgn} \left(\sum_{i \in I} \sum_{h \in H} y_{ikh} \right) = 1, \forall \, k \in K, \right. \\ & \operatorname{sgn} \left(\sum_{k \in K} \operatorname{sgn} (\sum_{h \in H} y_{ikh}) \right) = 1 - \sum_{j \in J} x_{ij}, \forall i \in I, \\ & \left. \sum_{k \in K} y_{ikh} \, \leq \, 1, \; \forall \, h \in H, \; i \in I \right\}. \end{aligned} \tag{21}$$

The set U^X is the set of all possible investment solutions for the investment company C, not taking into consideration the portfolios already chosen by the firm S to invest directly. If the investment company C_k chooses to invest in the stock portfolio P_i from the short listed ones, in the subperiod of time T_h, the value of y_{ikh} is 1, otherwise is 0. $Y^X \in U^X$ is a possible max-min-p point, condition being that each company of the group C must invest in only one portfolio that is still available on the market and has not been previously chosen by firm S.

For each $X \in \Lambda$, let us denote by U^{*X} the set

$$U^{*X} \, = \, arg - \max - \min - \mathrm{p} \, \{\varphi_2(Y) | Y \in A_0(X)\},$$

 ISBN: 978-84-614-6167-7

where

$$A_0(X) = \{Y \in U^X \mid \max\{d_{ikh} sgn\,(y_{ikh}) | i \in I, k \in K\} \leq r_h, \forall h \in H\}$$

and the function $\varphi_2 : U^X \to \mathbb{R}$ is given by

$$\varphi_2(Y) \;=\; \max\{d_{ikh}\mathrm{sgn}(y_{ikh}) \,|\, i \in I, k \in K, h \in H\},\; \forall\, Y \in A_0(X).$$

Recalling our problem, the objective function of the firm S is to maximize its final return: a direct return gained through investment made by its own branches and an indirect return gained through investment made by the company specialized in investments. By using these notations, the mathematical model for our portfolio problem is

$$(EBCT)\begin{cases} f(X,Y) \;\to\; \max \\ X = [x_{ij}] \in \Lambda, \\ \max\{r_{ij}x_{ij} \,|\, i \in I,\, j \in J\} \;\leq\; e, \\ Y \in U^{*X}. \end{cases}$$

We will call the problem $(EBCT)$ - *assignment bilevel cost-time type problem or E bilevel cost-time type problem.*

5 A Method for Solving (EBCT) Problems

The particularity of the restraints allows us to give a finite algorithm for solving the (EBCT) problem. The base is that if $X \in \Lambda$, then there are exactly n lines $i_1, \ldots, i_n$ so that

$$\sum_{j\in J} x_{i_t,j} = 1,\; \forall\, t \in J \text{ and } \sum_{j\in J} x_{ij} = 0,\; \forall\, i \in I \setminus \{i_t | t \in J\}.$$

In economic terms, there are exactly n stock portfolios so that for each portfolio there is one branch of the firm S to invest in it and, this branch will not invest in the other $m - n$ portfolios available.

If, in the problem $(\mathrm{P2}_X)$, we consider X as a parameter, there is the possibility to split the set Λ in a finite number, less or equal to C_m^n of subsets. For this purpose we introduce the set:

$$V = \{v = (v_1, ..., v_m) \in \{0,1\}^m \,|\, v_1 + \ldots + v_m = n\}.$$

For each $v = (v_1, ..., v_m) \in V$, we set

$$W^v = \{i \in I | v_i = 1\},$$

$$\Lambda^v \;=\; \{X = [x_{ij}] \in \Lambda \,|\, \sum_{j\in J} x_{ij} = v_i,\; \forall\, i \in I\},$$

and

$$\begin{aligned} U^v \;=\; & \{Y = [y_{ikh}] \in \{0,1\}^{m\times p\times s} \,|\, \textstyle\sum_{k\in K}\sum_{h\in H} y_{ikh} \geq 1, \forall i \in I \setminus W^v, \\ & \textstyle\sum_{i\in I\setminus W^v}\sum_{h\in H} y_{ikh} \geq 1, \forall k \in K,\; \sum_{k\in K} y_{ikh} = 1, \forall i \in I \setminus W^v, h \in H, \\ & y_{ikh} = 0, \forall i \in W^v, k \in K, h \in H\}. \end{aligned}$$

 ISBN: 978-84-614-6167-7

Then

$$\Lambda^{v'} \bigcap \Lambda^{v''} = \emptyset,\ \forall\, v', v'' \in V,\ v' \neq v''\ \text{and}\ \bigcup_{v \in V} \Lambda^v = \Lambda.$$

In what follows, for every $v \in V$, we consider the problems

$$(P_1^v) \quad \begin{cases} f_1^v(X) = \sum\limits_{i \in W^v} \sum\limits_{j \in J} a_{ij} x_{ij} \ \to\ \max \\ \\ X = [x_{ij}] \in \Lambda^v, \\ \\ \max\{r_{ij} x_{ij} \,|\, i \in I,\ j \in J\} \ \leq\ e, \end{cases}$$

and

$$(P_3^v) \quad \begin{cases} g^v(Y) = \left(\begin{array}{l} \sum\limits_{i \in I \setminus W^v} \sum\limits_{k \in K} (1 - q_{ik})(\sum\limits_{h \in H} c_{ikh} y_{ikh}) \\ \\ \max\{d_{ikh} y_{ikh} \,|\, i \in I, k \in K,\ h \in H\} \end{array} \right) \ \to\ \text{lex} - \max - \min \\ \\ Y = [y_{ikh}] \in U^v, \\ \\ \max\{d_{ikh} y_{ikh} \,|\, i \in I, k \in K\} \leq e_h,\ \forall h \in H. \end{cases}$$

We denote by $\mathcal{X}^v$ the set of optimal solutions of the problem (P_3^v) and by F_1^v the corresponding value of it. By $\mathcal{V}^v$, we denote the set of the lex-max-min solutions of the problem (P_3^v) and by (G_1^v, G_2^v) the corresponding vector value of it.

Theorem 1. If (X^0, Y^0) is an optimal solution of the problem (EBCT), then there is $v^0 = (v_1^0, \ldots, v_m^0) \in V$ so that X^0 is an optimal solution of the problem $(\mathrm{P}_1^{v^0})$ and Y^0 is an lexicographic max-min solution of the problem $(\mathrm{P}_3^{v^0})$.

The proof of the theorem is not difficult. We have to take $v_i^0 = \sum_{j \in J} x_{ij}^0$, for each $i \in I$.

Now, let be $F : V \to \mathbf{R}$,

$$F(v) = F_1^v \ +\ \max \left\{ \sum_{i \in I \setminus W^v} \sum_{k \in K} q_{ik} (\sum_{h \in H} c_{ikh} y_{ikh}) \mid y \in \mathcal{Y}^v \right\}$$

and let us consider the problem.

$$(EBVT) \quad \begin{cases} \text{find } v^0 \in V \text{ such that} \\ F(v^0) \ =\ \max\{F(v) \,|\, v \in V,\ \mathcal{X}^v \neq \emptyset,\ \mathcal{Y}^v \neq \emptyset\}. \end{cases}$$

Theorem 2. If the function g is injective and v^0 is an optimal solution of the problem (EBVT), then (X^0, Y^0) is an optimal solution of the problem (EBCT) for each $X^0 \in \mathcal{X}^{v^0}$, $Y^0 \in \mathcal{Y}^{v^0}$.

Theorem 1, Theorem 2 and the results of the paper [15] allow us to reduce the solving of the problem (EBCT) by solving C_m^n couples of classical E-type problems.

 ISBN: 978-84-614-6167-7

Example 1. Let be $m = 4$, $n = 2$, $p = 2$, $s = 1$, $e = 3$, $e_1 = e_2 = 3$,

$$A = \begin{bmatrix} 1 & 2 \\ 0 & 3 \\ 2 & 1 \\ 1 & 2 \end{bmatrix}, \; C = \begin{bmatrix} 2 & 3 \\ 1 & 1 \\ 2 & 3 \\ 3 & 1 \end{bmatrix}, \; R = \begin{bmatrix} 3 & 5 \\ 1 & 2 \\ 6 & 4 \\ 1 & 3 \end{bmatrix}, \; D = \begin{bmatrix} 3 & 2 \\ 1 & 2 \\ 2 & 2 \\ 4 & 4 \end{bmatrix}.$$

The set V is

$$V = \{v^1 = (1,1,0,0),\, v^2 = (1,0,1,0),\, v^3 = (1,0,0,1),\, v^4 = (0,1,1,0),$$

$$v^5 = (0,1,0,1),\, v^6 = (0,0,1,1)\}.$$

It is easy to see that v^1, v^2 and v^4 can not generate admissible solutions for the problem (P_3^v), whereas v^2, v^4 and v^6 can not generate admissible solutions for the problem (P_1^v). Solving the problems (P_1^v), (P_3^v), $v \in V$, we obtain

$$\max\{F(v)|v \in V\} \; = \; 9 \; = \; F(v^5).$$

As the function g is injective, the optimal solution of the problem (EBCT) from this example is

$$(X^0 = \begin{pmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 0 \\ 0 & 1 \end{pmatrix}, Y^0 = \begin{pmatrix} 0 & 1 \\ 0 & 0 \\ 1 & 0 \\ 0 & 0 \end{pmatrix}),$$

in view of the Theorem 2.

Acknowledgements

The author wish to thanks for the financial support provided from program: Investing in people! Ph.D. scholarship, Project co-financed by the Sectoral Operational Program for Human Resources Development 2007 - 2013 Priority Axis 1. "Education and training in support for growth and development of a knowledge based society" Key area of intervention 1.5: Doctoral and post-doctoral programs in support of research. Contract POSDRU/88/1.5/S/60185 - "Innovative Doctoral Studies in a Knowledge Based Society", Babes-Bolyai University, Cluj-Napoca, Romania.

References

[1] M.J., Alves, St. Dempe, and J. J. Júdice *Computing the Pareto frontier of a bi-objective bilevel linear problem using a multiobjective mixed-integer programming algorithm*, http://AlvesDempeJudice2010.pdf.

[2] J. Bracken and J. McGill, *Mathematical programs with optimization problems in the constraints*, Operations Res. **21** (1973) 37–44.

 ISBN: 978-84-614-6167-7

[3] L. Brotcorne, M. Labbé, P. Marcotte and G. Savard *Joint design and pricing on a network* Operations Res. **56** (2008) 1104-1115.

[4] W. Candler, and R. Norton, *Multilevel programming*, Tech. Rep. 20, World Bank Development Research Center, Washington D.C., 1977.

[5] B. Colson, P. Marcotte and G. Savard, *Bilevel programming: A survey*, 4OR, **3** (2005) 87-107.

[6] St. Dempe, *Foundations of Bilevel Programming*, Kluwer Academic Publishers, Dordrecht, 2002.

[7] D. Duca and L. Lupşa, *Bilevel transportation problems*, Rev. Anal. Numér. Théor Approx. **30(1)** (2001) 25–34.

[8] Y. Gao, J. Lu, G. Zhang, and S. Gao *A Bilevel Model for Railway Train Set Organizing Optimization*
htpp://Lu_Jie.pdf
ISKE2007-Lu_Jie.pdf, www.atlantis-press.com.

[9] St. Kosuch, P. Le Bodic, J. Leung and A. Lisser *On a Stochastic Bilevel Programming Problem with Knapsack Constraints*
http://www.kosuch.eu/stefanie/veroeffentlichungen/INOC09_abstract.pdf.

[10] Y. Li and S. Chen *Optimal Traffic Signal Control for an $M \times N$ traffic Network*, Journal of Industrial and Management Optimization **4** (2008) 661–672.

[11] P. A. Lotito, E. M. Mancinelli, J. P. Quadrat and L. Wynter, *A global descent heuristic for bilevel network pricing* (2004) http://www-rocq.inria.fr/metalau/quadrat/bpalgocomplet.pdf.

[12] H.M. Markowitz, *Portfolio Selection*, The Journal of Finance. **8** (1952) 77–91.

[13] H.F. von Stackelberg, *The theory of the market economy*, Oxford University Press, Oxford, 1952.

[14] H.F. von Stackelberg, *Marktform und Gleichgewicht*, Springer-Verlag, Berlin, 1934.

[15] O. R. Tuns (Bode), Max-min-p points, Annals of the Tiberiu Popoviciu Seminar of Functional Equations, Approximation and Convexity, **9** (2011), in print.

[16] L. N. Vicente, Bilevel Programming: Introduction, History, and Overwiew,BP
http://www.mat.uc.pt/ lnv/papers/EoO.pdf.

[17] L. N. Vicente and P. H. Calamai, *Bilevel and multilevel programming : a bibiliography review*, J. Global Optim., **5** (1994) 291–306.

 ISBN: 978-84-614-6167-7

[18] L. Vicente, G. Savard and J. Júdice *The discrete linear bilevel programming problem*, JOTA **89** (1996) 597–614.

 ISBN: 978-84-614-6167-7

Proceedings of the 11th International Conference on Computational and Mathematical Methods in Science and Engineering, CMMSE 2011 26–30 June 2011.

Increasing the Parallelism of Distributed Crowd Simulations on Multi-core Processors

Guillermo Vigueras[1], Juan M. Orduña[1] and Miguel Lozano[1]

[1] *Departamento de Informática, University of Valencia. Spain*

emails: guillermo.vigueras@uv.es, juan.orduna@uv.es, miguel.lozano@uv.es

Abstract

The simulation of large crowds of autonomous agents with realistic behavior requires an efficient use of the increasing number of cores existing in current multi-core processors, in order to achieve scalable simulations. In this paper, we propose an implementation of a distributed action server for crowd simulation based on the RCU synchronization method. In this way, distributed architectures for crowd simulations can also fully exploit the inherent parallelism in multi-core processors, increasing their throughput and scalability. We have compared the proposed implementation with a parallel implementation based on Mutex, a traditional locking synchronization method for solving race conditions among threads in parallel applications. The performance evaluation results show that the use of *RCU* significantly increases the system throughput, supporting a higher number of agents while providing the same latency levels. The reason for that behavior is the bottleneck that the sequential access to the data structures locked by a Mutex represents for a parallel application. Since the RCU method allows read accesses in parallel with write accesses to these data structures, it significantly reduces these bottlenecks. Also, these results represent the first evaluation of the *RCU* method in a real and complex parallel application with large data structures, since the *RCU* method has only been evaluated using small benchmarks until now.

Key words: distributed crowd simulation, multi-core programming

1 Introduction

The increasing number of cores existing in current multi-core processors provides these systems with computing capabilities that can be exploited by distributed applications. The current technology trend of integrating more cores in a single processor has generated an interest for researchers in developing software capable of efficiently exploiting the computational capabilities of these systems. One of the distributed applications that can benefit from multi-core processors is crowd simulation. Crowd simulation can be considered as a special case of Virtual Environments where the avatars are intelligent

 ISBN: 978-84-614-6167-7

agents instead of user-driven entities. Each of these agent-based entities can have its own goals, knowledge and behavior [16]. In recent years, crowd simulation has become an essential tool for many virtual environment applications in education, training, and entertainment [18, 1, 12]. These applications require both rendering visually plausible images of the virtual world and managing the behavior of complex autonomous agents. The sum of these requirements results in a computational cost that exponentially increases with the numbers of agents in the system, requiring a scalable design that can support huge amounts of agents (of different orders of magnitude) by simply adding more hardware.

Different distributed schemes have been presented last years for improving the scalability of crowd simulations [13, 20]. In previous works we proposed a distributed system architecture for crowd simulation that can take advantage of the underlying distributed computer system [9, 23, 22]. That architecture consists of a distributed system where some of the computing nodes contain a distributed Action Server controlling the simulation. The rest of the computers host a set of agents implemented as threads of a single process. That architecture was shown efficiently enough to support simulations up to tens of thousands of complex agents with plausible graphic quality. However, this distributed scheme can be still improved by fully exploiting the potential of new multicore architectures, thus increasing the system throughput.

Several researchers have already studied the capabilities of multi-core architectures for crowd simulations. In this sense, an approach has been presented for PLAYSTATION3 that distributes the load among the PS3-Cell elements[15]. Another work uses graphics hardware to simulate crowds of thousands individuals using models designed for gaseous phenomena [2]. Other authors started to use GPU in an animation context (particle engine) [14, 8], and there are also some proposals for running simple stochastic agent simulations on GPUs [10, 13]. However, these proposals are far from displaying complex behaviors at interactive rates. Finally, other proposals provide different interactive and complex crowd systems [17, 21], but they are not scalable with the number of agents.

In this paper, we show that the distributed architecture previously proposed for large scale crowd simulations [23] can also benefit from multi-core processors. Concretely, we propose an efficient parallel implementation of a distributed Action Server for crowd simulation. This implementation is based on *RCU*, a *concurrently-readable* synchronization method that allows to significantly improve the scalability of the Action Server with the number of cores and threads. We have compared the proposed implementation with a parallel implementation based on Mutex, a traditional locking synchronization method for solving race conditions among threads in parallel applications. The performance evaluation results show that the use of *RCU* significantly increases the system throughput, supporting a higher number of agents while providing the same latency levels. Also, these results represent the first evaluation of the *RCU* method in a real and complex parallel application with large data structures, since the *RCU* method has only been evaluated using small benchmarks until now.

The rest of the paper is organized as follows: Section 2 describes the parallel implementations proposed for the distributed Action Server for Crowd Simulation.

 ISBN: 978-84-614-6167-7

Section 3 shows the performance evaluation of the parallel implementations. Finally, Section 4 shows some concluding remarks and future work to be done.

2 Parallel implementations of the Action Server

As multicore processors become mainstream, multithreaded applications will become more common, increasing the need for efficient programming models. Efficiently programming a multicore processor is relatively easy when the program input is a static structure without data dependencies that can be partitioned among the execution threads. The OpenMP API for example, permits the parallelization of a program by means of directives, partitioning the workload among different execution threads. However, problems arise if dynamic data structures are not safely managed in a multithreaded program. Furthermore, an efficient thread coordination of concurrent accesses to shared data structures is needed in order to obtain a good speed-up with the number of cores. Traditional locking methods requires expensive atomic operations, such as compare-and-swap (CAS), even when locks are uncontended. Locking is also susceptible to priority inversion, convoying, deadlock, and blocking due to thread failures. Therefore, many researchers recommend avoiding a locking-based synchronization methods. Some proposals use non-blocking (or lock-free) synchronization methods in multithreaded applications, obtaining good results [19, 11]. Other works have studied the impact of replacing a locking-based synchronization by Software Transactional Memory (STM) in a multi-player game server [24, 5]. Nevertheless, these studies reveal that regardless the granularity of memory transactions used in STM, the performance obtained with STM is even worse than the one obtained with a locking-based synchronization method.

A major challenge for lockless synchronization methods is handling the *read/reclaim* races that arise in dynamic data structures. Figure 1 illustrates this problem: thread *T1* removes node N from a list while thread *T2* is referencing it. N's memory must be reclaimed to allow reuse, because otherwise memory exhaustion could block all threads. However, such reuse is unsafe while *T2* continues referencing N. For languages like C, where memory must be explicitly reclaimed (e.g. via `free()`),programmers must combine a *memory reclamation scheme* with their lockless data structures to resolve these races.

This section describes two parallel implementations of an Action Server for crowd simulation [23]. The first implementation is based on *Mutex*, the thread synchronization method provided by the POSIX threads API. This locking method is used to protect the dynamic data structures shared among the threads during the simulation, but it can severely limit the parallelism achieved with the number of cores, significantly reducing the corresponding speed-up. For that reason, another implementation based on a lock-free data structure, the *Read-Copy Update* (*RCU*) method [11], is proposed. *RCU* is a *concurrently-readable* synchronization method that can be used to improve the scalability of the Action Server with the number of cores and threads.

The Action Server (AS) is the key module of a distributed architecture for crowd

 ISBN: 978-84-614-6167-7

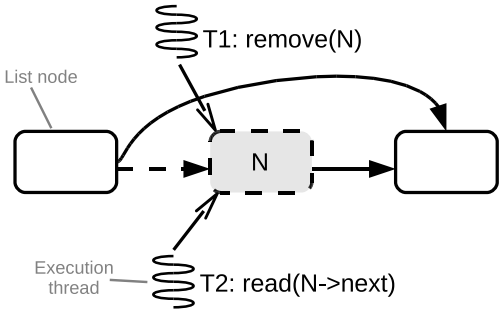


Figure 1: Read/reclaim race among threads that access a shared dynamic data structure.

simulations [23]. An Action Server [23] can be viewed as a partial world manager, since it controls and properly modifies the information in a region of the whole simulation space. Thus, it can be considered as the system core. Each AS process contains three basic elements: the Interface module, the Crowd AS Control (CASC) module and the Semantic Data Base (SDB). Figure 2 illustrates a detailed scheme of an AS.

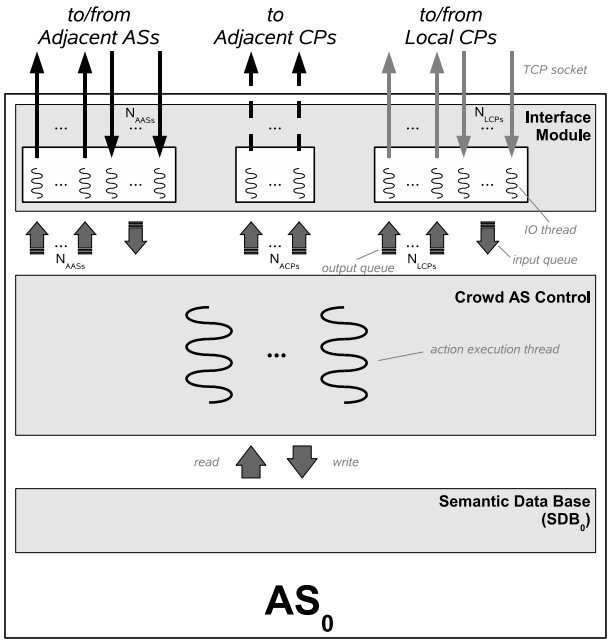


Figure 2: Internal structure of the initial Action Server.

The main module is the Crowd AS Control (CASC) module, which is responsible for executing the crowd actions. This module contains a configurable number of threads for executing actions (action execution threads in Figure 2). For an action execution thread (AE thread), all messages sent to or received from other ASs and CPs are exchanged asynchronously (the details are hidden by the Interface module, see below).

 ISBN: 978-84-614-6167-7

This means that the AE threads only may have to wait when accessing shared data structures. Thus, experimental tests have shown that having more AE threads than cores allows each AS to take advantage of several cores.

Most action requests from agents are executed from start to end by the AE thread, that extracts the requests from the corresponding input queue and process them. These requests consist of collision tests, in order to check if the new position computed by each agent when it moves coincides with the position of other agent or object. The Interface module hides all the details of the message exchanges. This module provides the Crowd AS Control module with the abstraction of asynchronous messages. Two separate input queues exist, one for messages coming from local CPs (action requests) and the other one for messages coming from adjacent ASs (responses to requests issued because of local border actions or requests for remote border actions from adjacent ASs). Having two separate input queues is an efficient way of giving a higher priority to messages from adjacent ASs. The reason for improving the priority of these messages is that the border actions are the ones whose processing takes longer, and we should reduce as much as possible their response time to provide realistic interactive effects.

2.1 Mutex-based implementation of the AS

Messages are processed as soon as they arrive in the mutex-based implementation of the AS. For that reason, the Interface module in the CASC contains one IO thread dedicated to getting incoming messages from each TCP socket. There are no input threads associated with sockets connecting one AS to their adjacent CPs, because CPs only send messages to their local AS. In the same way, there is one IO thread and one output queue per TCP socket, so that messages are sent as soon as the corresponding TCP socket is ready for writing. All IO queues are implemented using the queue data structure provided by C++ *standard template library*. Each queue is protected using one Mutex per queue. In this way, each Mutex synchronizes the concurrent accesses of IO threads and AE threads.

The Mutex-based implementation of the AS also uses a grid-based data structure to perform the collision check procedure within the CASC module, this data structure is denoted as *collision grid*. Figure 3 illustrates the implementation details of the Mutex-based *collision grid*. The top part of the figure shows the geometric space partitioned using a grid with 16 grid cells. Four agents, represented as numbered circles, are allocated within the grid at given positions. The bottom part of Figure 3 shows that the *collision grid* is implemented as a linear array. Each element of this array contains a mutex and a pointer to a dynamic data structure, implemented as a linked list, that contains the agents positions. The mutex avoids the corruption of the dynamic data structure when reader and writer threads concurrently access it during the collision check. Each thread performs the collision check for an agent. In order to achieve this goal, the thread computes the mapping of the agent's position into the *collision grid* by means of the hashing method. The mutex located at the position returned by the hashing method is locked, and the dynamic data structure is queried in case of a read access, or updated in case of a write access. This implementation uses an array of

 ISBN: 978-84-614-6167-7

mutexes instead of a single mutex for the *collision grid*, in order to provides a certain level of concurrency among threads.

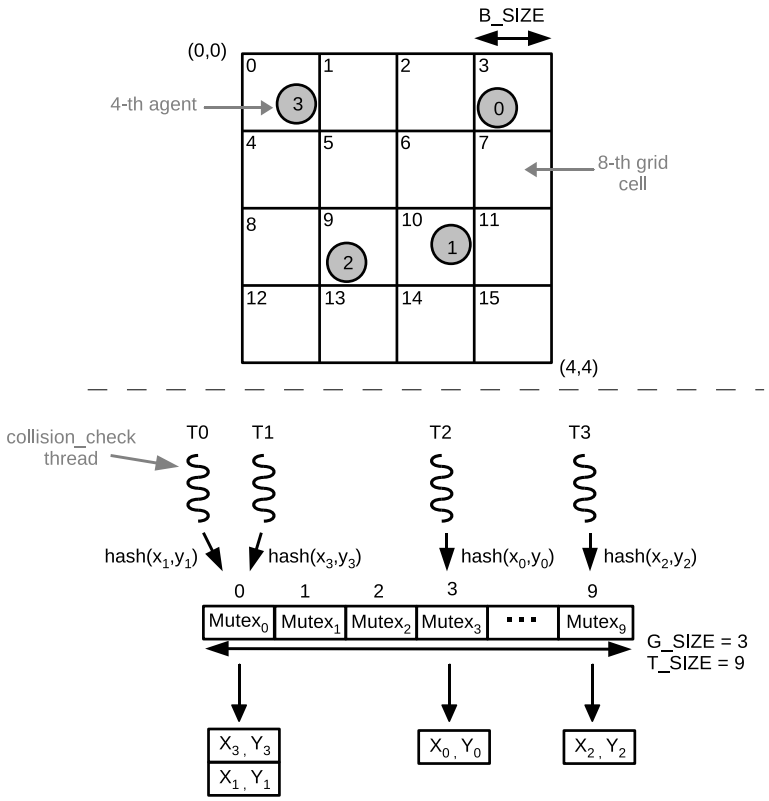


Figure 3: Diagram of the CPU collision checking using Mutex.

2.2 RCU-based implementation of the Action Server

RCU is a synchronization mechanism that was added to the Linux kernel during the development of version 2.5. Recently, it has been released for user-space access [3]. However, the benefits provided by *RCU* have not been checked in real and complex problems with large data structures. The idea behind *RCU* is to split data structures updates into *removal* and *reclamation* phases. The removal phase removes references to data items within a data structure (possibly by replacing them with references to new versions of these data items), and it can concurrently run with reader threads (readers). These concurrent executions are safe, because the semantics of modern CPUs guarantee that readers will see either the old or the new version of the data structure, rather than a partially updated reference. The reclamation phase performs the work of claiming (e.g., freeing) the data items removed from the data structure during the removal phase. Since claiming data items can disrupt any reader that is concurrently referencing these data items, the reclamation phase must not start until readers no longer hold references to these data items. Splitting the update into removal and reclamation phases permits the

 ISBN: 978-84-614-6167-7

updater thread to perform the removal phase immediately, and defer the reclamation phase until all active readers during the removal phase have finished their task (in order to do so, the updater thread can either block himself or register a callback that is invoked after the completion of the readers). Only the readers that are active during the removal phase need to be considered, because any reader starting after the removal phase will be unable to gain a reference to the removed data items.

Different reclamation schemes can be used to implement *RCU*. We have implemented a *RCU* version using the *Quiescent State Based Reclamation* (*QSBR*) scheme, because it provides concurrent reads with the lowest overhead [6]. However, the application should be modified in order to explicitly manage reclamations [3]. *QSBR* uses the concept of a *grace period*. A *grace period* is a time interval *[a,b]* such that, after time b, all nodes removed before time a may safely be reclaimed. *QSBR* uses quiescent states to detect grace periods. A quiescent state for thread T is a state in which T holds no references to shared nodes. Hence, a grace period for *QSBR* is any interval of time during which all threads pass through at least one quiescent state. Figure 4 illustrates the relationship between quiescent states and grace periods in *QSBR*. Thread *T1* goes through quiescent states at times t_1 and t_5 , *T2* at times t_2 and t_4 , and *T3* at time t_3 . Hence, a grace period is any time interval containing either $[t_1, t_3]$ or $[t_3, t_5]$.

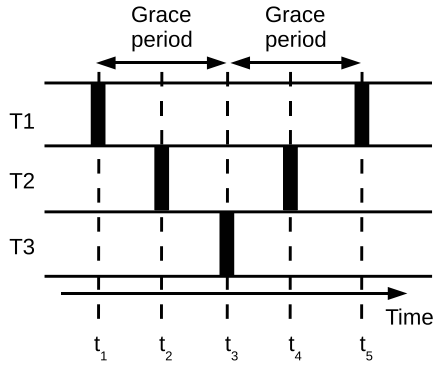


Figure 4: Illustration of QSBR. Black boxes represent quiescent states.

Figure 5 shows an example of the management of a linked-list in a multithreaded environment by using *RCU* API calls [11]. The figure shows the changes suffered by a linked list containing three elements (A, B and C) while an updater thread deletes element B. This element is deleted using the *RCU* API call *list_del_rcu()*. This function removes a list element, but it allows some concurrent readers to continue seeing the removed element. Looking at Figure 5, it can be seen that after execute *list_del_rcu()*, the element B has been removed from the list. Since readers do not synchronize directly with updaters, readers might be concurrently scanning this list. These concurrent readers might or might not see the element that has been recently removed, depending on timing. However, readers that were delayed (e.g., due to interrupts) just after fetching a pointer to the recently removed element might see the old version of the list

 ISBN: 978-84-614-6167-7

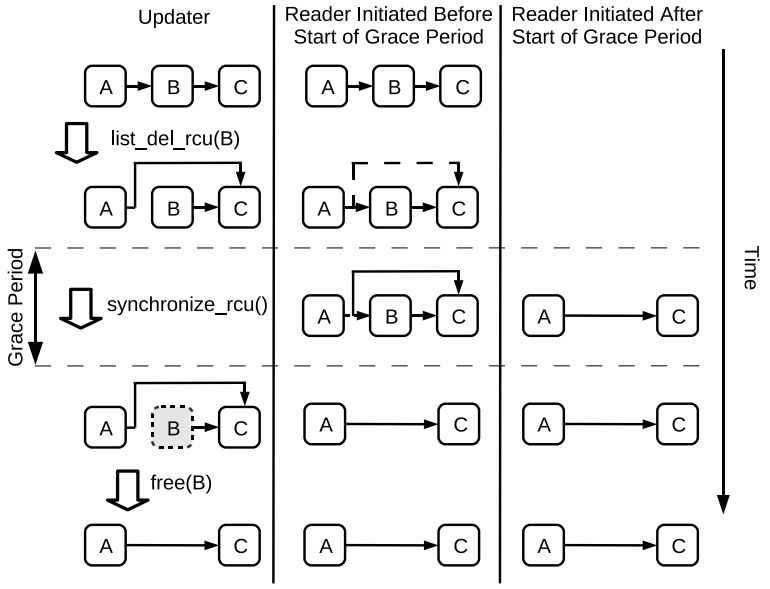


Figure 5: Deletion of one element in a RCU linked-list.

for quite some time after the removal. Therefore, during the grace period there are two versions of the list, one with element B and one without it. For that reason, the freeing of element B is postponed. Readers are not allowed to maintain references to element B after exiting from their RCU read-side critical sections. Therefore, when all readers have exited their critical sections, then no more readers can be referencing element B, as indicated by the grey filling color and dashed frame of element B in the *Updater* column in Figure 5. When no more readers hold references to element B, then the *synchronize_rcu()* function is completed. At this point, the list is back to a single version and element B may safely be freed.

The CPU implementation based on mutex described in Figures 2 and 3, has been adapted in order to support lock-free data structures along with the RCU synchronization method. In this sense, the IO queues in the interface module have been implemented as FIFO lockless linked lists. In the same way, the linear array representing the *collision grid* is modified, removing the mutex from each element of the *collision grid*. As a consequence, lock-free linked lists containing the agents positions are obtained. Read and update accesses to all lockless linked lists are performed by threads through a user-space RCU API [3]. The QSBR version of RCU implemented by this API allows the definition of quiescent states in order to safely update the linked lists contained in the interface module and in the *collision grid*.

3 Performance Evaluation

This section shows the performance evaluation of the implementations described in the previous section. We have performed different measurements on different real systems

 ISBN: 978-84-614-6167-7

using these implementations. Like other distributed systems, the most important performance measurements in DVE systems are latency and throughput [4]. Since we are focusing on the system scalability, we have performed simulations with different number of agents and we have measured the response time provided to the agents. In this way, we can study the maximum number of agents that the system can support while providing a response time below a given threshold value. In order to define an acceptable behavior for the system, we have considered 250 ms. as the threshold value, since it is considered as the limit for providing realistic effects to users in DVEs [7].

We have performed crowd simulations with wandering agents, because all their actions (movement requests) should be verified by an AS. Each simulation consists of the crowd moving within the virtual world following k-length random paths. For all the populations tested, the average response time has become stable within the first minute of execution time. Therefore, we have used one minute simulations for all the configurations tested.

The computer platform used in the experiments has been a cluster of computers, where one machine hosts the Action Server and up to 16 machines host 16 clients processes (i.e. one client per machine). The machine hosting the server is a 16 core machine integrating 8 AMD Opteron processors (2 cores @ 1 GHz per processor), with 32.5 GB of RAM and the operating system was Linux 2.6.18-92. We have used the POSIX API in order to obtain different configurations with an increasing number of cores. This API allows to set the affinity of the execution threads, limiting the cores set in which the threads are executed. Four configurations containing 2, 4, 8 and 16 cores were used in order to check the scalability with the number of cores of the parallel implementations. The machines hosting the clients are based on AMD Opteron (2 x 1.56 Ghz processors) with 3.84GB of RAM, executing Linux 2.6.9-1 operating system. The interconnection network in the cluster was a Gigabit Ethernet network.

Figure 6 shows the response times provided by the implementation based on Mutex (Mutex implementation) . On the X-axis, this figure shows the number of thousands of agents in the system when the Action Server uses 16 cores. The Y-axis shows both the average and the maximum response times provided to the agents during the simulation. Each point in this figure has been computed as the average value of thirty different simulations. The label "AVG RT" corresponds to the plot showing the average response time, while the label "MAX RT." corresponds to the maximum response time provided to an agent. Figure 6 shows that the average maximum number of agents supported by the Mutex-based implementation is 7500 (for a simulation with 7500 agents the average response time does no exceed the maximum allowable time of 250 ms.). Regarding the maximum response time, the number of agents supported is around 7000. On the other hand, Figure 7 shows that the average maximum number of agents supported by the RCU-based implementation is 24000. Regarding the maximum response time, the number of agents supported is around 23500. Comparing Figures 6 and 7, it can be clearly seen the significantly higher number of agents supported by the RCU-based implementation when using 16 cores. In addition, it can be see the steeper slope of the plots for the Mutex-based implementation within the range of 7500 agents up to 9000 agents. The reason for this behavior is the higher overhead introduced by the

 ISBN: 978-84-614-6167-7

Mutex synchronization method with respect to the RCU method. Although they are not shown here due to space limitations, we obtained similar results fro configurations with lower number of cores.

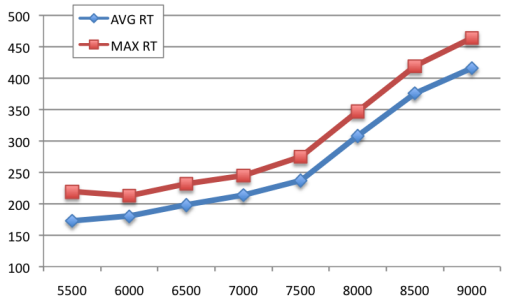


Figure 6: Server response time for the Mutex implementation using 16 cores

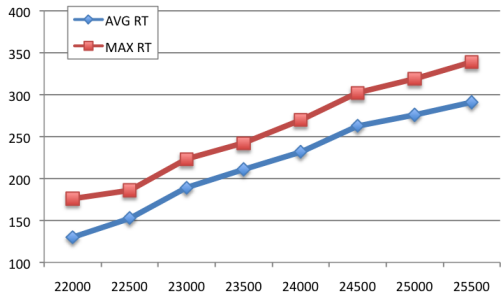


Figure 7: Server response time for the RCU implementation using 16 cores

Figure 8 shows the maximum throughput (the maximum number of agents supported by the simulation system while not exceeding an average server response time of 250 ms.) provided by the considered implementations when the number of available cores is increased. It can be seen that the number of supported agents does not significantly increases for the Mutex-based implementation with the number of cores, ranging from 4000 agents when using 2 cores to 7500 agents when using 16 cores. The reason for that behavior is the bottleneck that the sequential access to the data structures locked by a Mutex represents for a parallel application. Since the RCU allows read accesses in parallel with write accesses to these data structures, it significantly reduces these bottlenecks, providing a plot with a significantly higher slope than the Mutex-based implementation. Thus, the number of supported agents ranges from 4000 when using 2 cores to 24000 when using 16 cores. These results show the benefits that the RCU method can provide to parallel applications with respect to traditional locking synchronization methods like Mutex.

4 Conclusions

In this paper, we have proposed an implementation of a distributed action server for crowd simulation based on the RCU synchronization method. We have compared the proposed implementation with a parallel implementation based on Mutex, a traditional locking synchronization method for solving race conditions among threads in parallel applications. The performance evaluation results show that the use of *RCU* significantly increases the system throughput with respect to the implementation based on Mutex, supporting a higher number of agents while providing the same latency levels. The reason for that behavior is the bottleneck that the sequential access to the data structures locked by a Mutex represents for a parallel application. Since the RCU method allows read accesses in parallel with write accesses to these data structures, it

 ISBN: 978-84-614-6167-7

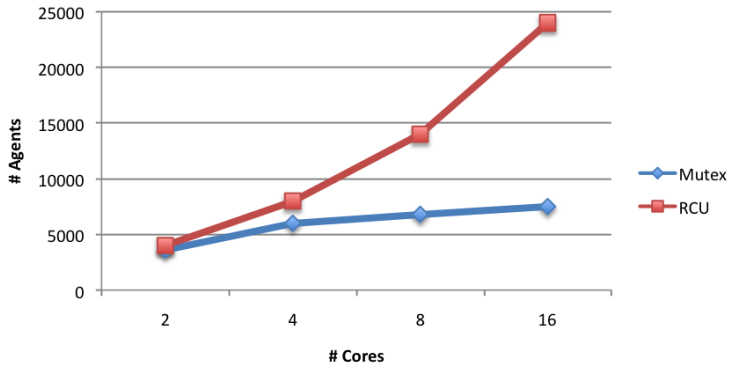


Figure 8: Throughput of the Mutex and RCU implementations of the AS.

significantly reduces these bottlenecks. Also, these results represent the first evaluation of the *RCU* method in a real and complex parallel application with large data structures, since the *RCU* method has only been evaluated using small benchmarks until now.

Acknowledgements

This work has been jointly supported by the Spanish MICINN and the European Commission FEDER funds under grants Consolider-Ingenio CSD2006-00046, TIN2009-14475-C04-04, and TIN2010-12011-E.

References

[1] D. Chen, G. K. Theodoropoulos, S. J. Turner, W. Cai, R. Minson, and Y. Zhang. Large scale agent-based simulation on the grid. *Future Generation Computer Systems*, 24(7):658 – 671, 2008.

[2] N. Courty and S. R. Musse. Simulation of large crowds in emergency situations including gaseous phenomena. In *CGI '05: Proceedings of the Computer Graphics International 2005*, pages 206–212. IEEE Computer Society, 2005.

[3] M. Desnoyers. Userspace rcu library : What linear multiprocessor scalability means for your application. In *Linux Plumbers Conference*, 2009.

[4] J. Duato, S. Yalamanchili, and L. Ni. *Interconnection Networks: An Engineering Approach*. IEEE Computer Society Press, 1997.

[5] V. Gajinov, F. Zyulkyarov, O. S. Unsal, A. Cristal, E. Ayguade, T. Harris, and M. Valero. Quaketm: parallelizing a complex sequential application using transactional memory. In *Proceedings of the 23rd international conference on Supercomputing*, ICS '09, pages 126–135, New York, NY, USA, 2009. ACM.

[6] T. E. Hart, P. E. McKenney, and A. D. Brown. Making lockless synchronization fast: Performance implications of memory reclamation. In *20th IEEE International Parallel and Distributed Processing Symposium*, Rhodes, Greece, April 2006.

 ISBN: 978-84-614-6167-7

[7] T. Henderson and S. Bhatti. Networked games: a qos-sensitive application for qos-insensitive users? In *Proceedings of the ACM SIGCOMM 2003*, pages 141–147. ACM Press / ACM SIGCOMM, 2003.

[8] L. Latta. Building a million particle system. In *In Proc. of Game Developers Conference(GDC-04)*, 2004.

[9] M. Lozano, P. Morillo, J. M. Orduña, V. Cavero, and G. Vigueras. A new system architecture for crowd simulation. *J. Netw. Comput. Appl.*, 32(2):474–482, 2009.

[10] M. Lysenko and R. M. D'Souza. A framework for megascale agent based model simulations on graphics processing units. *Journal of Artificial Societies and Social Simulation*, 11(4):10, 2008.

[11] P. E. McKenney and J. D. Slingwine. Read-copy update: Using execution history to solve concurrency problems. In *Parallel and Distributed Computing and Systems*, pages 509–518, Las Vegas, NV, October 1998.

[12] D. L. París and A. Brazalez. A new autonomous agent approach for the simulation of pedestrians in urban environments. *Integr. Comput.-Aided Eng.*, 16(4):283–297, 2009.

[13] K. S. Perumalla and B. G. Aaby. Data parallel execution challenges and runtime performance of agent simulations on gpus. In *SpringSim '08: Proceedings of the 2008 Spring simulation multiconference*, pages 116–123, New York, NY, USA, 2008. ACM.

[14] K. Peter, S. Mark, and W. Rudiger. Uberflow: a gpu-based particle engine. In *HWWS '04: Proceedings of the ACM SIGGRAPH/EUROGRAPHICS conference on Graphics hardware*. ACM, 2004.

[15] C. Reynolds. Big fast crowds on ps3. In *Proceedings of the ACM SIGGRAPH symposium on Videogames*, pages 113–121, New York, NY, USA, 2006. ACM.

[16] C. W. Reynolds. Flocks, herds and schools: A distributed behavioral model. In *SIGGRAPH '87: Proceedings of the 14th annual conference on Computer graphics and interactive techniques*, pages 25–34, New York, NY, USA, 1987. ACM.

[17] W. Shao and D. Terzopoulos. Autonomous pedestrians. In *SCA '05: Proceedings of the 2005 ACM SIGGRAPH/Eurographics symposium on Computer animation*, pages 19–28, New York, NY, USA, 2005. ACM.

[18] A. Shendarkar, K. Vasudevan, S. Lee, and Y.-J. Son. Crowd simulation for emergency response using bdi agent based on virtual reality. In *WSC '06: Proceedings of the 38th conference on Winter simulation*, pages 545–553, 2006.

[19] H. Sundell and P. Tsigas. Lock-free deques and doubly linked lists. *J. Parallel Distrib. Comput.*, 68(7):1008–1020, 2008.

[20] H. Tianfield, J. Tian, and X. Yao. On the architectures of complex multi-agent systems. In *Proc. of the IEEE/WIC International Conference on Web Intelligence / Intelligent Agent Technology,*, pages 195–206. IEEE Press, 2003.

[21] A. Treuille, S. Cooper, and Z. Popovic. Continuum crowds. In *SIGGRAPH '06: ACM SIGGRAPH 2006 Papers*, pages 1160–1168. ACM, 2006.

[22] G. Vigueras, M. Lozano, J. Orduña, and Y. Chrysanthou. A distributed visual client for large-scale crowd simulations. In *International Conference on Computational and Mathematical Methods in Science and Engineering, (CMMSE 2010)*, pages 999–1010, 2010.

[23] G. Vigueras, M. Lozano, C. Perez, and J. Ordua. A scalable architecture for crowd simulation: Implementing a parallel action server. In *Proceedings of the 37th International Conference on Parallel Processing (ICPP-08)*, pages 430–437, Sept. 2008.

[24] F. Zyulkyarov, V. Gajinov, O. S. Unsal, A. Cristal, E. Ayguadé, T. Harris, and M. Valero. Atomic quake: using transactional memory in an interactive multiplayer game server. In *Proceedings of the 14th ACM SIGPLAN symposium on Principles and practice of parallel programming*, PPoPP '09, pages 25–34, New York, NY, USA, 2009. ACM.

 ISBN: 978-84-614-6167-7

Proceedings of the 11th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2011
26–30 June 2011.

A Multiple Prior Monte Carlo Method for the Backward Heat Diffusion Problem

Antoine E. Zambelli[1]

[1] *Department of Mathematics, University of California, Berkeley*

emails: antoine_elabdouni@berkeley.edu

Abstract

We consider the nonlinear inverse problem of reconstructing the heat conductivity of a cooling fin, modeled by a 2-dimensional steady-state equation with Robin boundary conditions. The Metropolis Hastings Markov Chain Monte Carlo algorithm is studied and implemented, as well as the notion of priors. By analyzing the results using certain trial conductivities, we formulate several distinct priors to aid in obtaining the solution. These priors are associated with different identifiable parts of the reconstruction, such as areas with vanishing, constant, or varying slopes. Although more research is required for some non-constant conductivities, we believe that using several priors simultaneously could help in solving the problem. *Key words: Inverse Problems, Heat Diffusion, Monte Carlo, Prior.*

1 Introduction

In this problem, we attempt to reconstruct the *conductivity* K in a steady state heat equation of the cooling fin on a CPU. The heat is dissipated both by conduction along the fin and by convection with the air, which gives rise to our equation (with H for convection, K for conductivity, δ for thickness and u for temperature):

$$u_{xx} + u_{yy} = \frac{2H}{K\delta}u \tag{1}$$

The CPU is connected to the cooling fin along the bottom half of the left edge of the fin. We use the Robin Boundary Conditions (detailed in [1]):

$$Ku_{normal} = Hu \tag{2}$$

 ISBN: 978-84-614-6167-7

Our data in this problem is the set of boundary points of the solution to (1), which we compute using a standard finite difference scheme for an $n \times m$ mesh (here 10×10 or 20×20). We denote the correct value of K by K_{correct} and the data by d. In order to reconstruct K_{correct}, we will take a guess K', solve the forward problem using K' and compare those boundary points to d by implementing the Metropolis-Hastings Markov Chain Monte Carlo algorithm (or MHMCMC). Priors will need to be established to aid in the reconstruction, as comparing the boundary points alone is insufficient.

2 MHMCMC

Markov Chains produce a probability distribution of possible solutions (in this case conductivities) that are most likely given the observed data (the probablility of reaching the next step in the chain is entirely determined by the current step). The algorithm is as follows ([2]). Given K_n, K_{n+1} can be found using the following:

1. Generate a candidate state K' from K_n with some distribution $g(K'|K_n)$. We can pick any $g(K'|K_n)$ so long as it satisfies
   (a) $g(K'|K_n) = 0 \Rightarrow g(K_n|K') = 0$
   (b) $g(K'|K_n)$ is the transition matrix of Markov Chain on the state space containing K_n, K'.
2. With probability
$$\alpha(K'|K_n) \equiv min\left\{1, \frac{Pr(K'|d)g(K_n|K')}{Pr(K_n|d)g(K'|K_n)}\right\} \tag{3}$$
set $K_{n+1} = K'$, otherwise set $K_{n+1} = K_n$ (ie. accept or reject K'). Proceed to the next iteration.

Using the probability distributions of our example, (3) becomes

$$\alpha(K'|K_n) \equiv min\left\{1, e^{\frac{-1}{2\sigma^2}\sum_{i,j=1}^{n,m}\left[\left(d_{ij}-d'_{ij}\right)^2-\left(d_{ij}-d_{n_{ij}}\right)^2\right]}\right\} \tag{4}$$

(where d' and d_n denote the set of measured boundary points using K' and K_n respectively, and $\sigma = 0.1$)
To simplify (4), collect the constants and seperate the terms relating to K' and K_n:

$$\frac{-1}{2\sigma^2}\sum_{i,j=1}^{n,m}\left[\left(d_{ij}-d'_{ij}\right)^2-\left(d_{ij}-d_{n_{ij}}\right)^2\right] \tag{5}$$

 ISBN: 978-84-614-6167-7

$$= \frac{-1}{2}\sum_{i,j=1}^{n,m}\left[\left(\frac{d_{ij}-d'_{ij}}{\sigma}\right)^2-\left(\frac{d_{ij}-d_{n_{ij}}}{\sigma}\right)^2\right] \quad (6)$$

$$= \frac{-1}{2}\left[D'-D_n\right]=f_n-f' \quad (7)$$

Now, (4) reads

$$\alpha(K'|K_n)\equiv min\left\{1,e^{f_n-f'}\right\} \quad (8)$$

We now examine the means by which we generate a guess K'. If the problem consists of reconstructing a constant conductivity, we can implement a uniform change, for every iteration we take a random number ω between -0.005 and 0.005 and add it to every entry in K_n to obtain K' (we initialize K_0 to a matrix of 1s). The algorithm is highly efficient, and the reconstructed value will consistently converge to that of the solution to within ω.
In order to approximate a nonconstant K_{correct}, the obvious choice is a pointwise change, at each iteration we add ω to a random entry of K_n, thus genrating K'. Unfortunately, systematic errors occur at the boundary points of our reconstruction (they tend to rarely change from their initial position).
In order to sidestep this, we use a gridwise change; change a square of the mesh (chosen at random as well) by adding ω to the four corners of said square. While this fixes the boundary problem, another major issue which arises from a non-uniform change is that the reconstruction will be marred with "spikes", which we must iron out.

3 The Smoothness Prior

To aid in ironing out the wrinkles in the reconstruction we use "priors". Priors generally require some knowledge of the quantity we wish to find, and will add a term to (8). Naturally, the more unassuming the prior, the more applicable the algorithm. This applicability will be tested as often as possible throughout these tests. The first prior compares the sum of the differences between adjacent points of K_{correct} to those of K' (keping the spikes in check), and is given by

$$T'=\sum_{j=1}^{n}\sum_{i=2}^{m}\left(K'(i,j)-K'(i-1,j)\right)^2+\sum_{i=1}^{m}\sum_{j=2}^{n}\left(K'(i,j)-K'(i,j-1)\right)^2 \quad (9)$$

$$T_n=\sum_{j=1}^{n}\sum_{i=2}^{m}\left(K_n(i,j)-K_n(i-1,j)\right)^2+\sum_{i=1}^{m}\sum_{j=2}^{n}\left(K_n(i,j)-K_n(i,j-1)\right)^2 \quad (10)$$

and modifying (8), we obtain

 ISBN: 978-84-614-6167-7

$$\alpha_c(K'|K_n) \equiv min\left\{1, e^{f_n - f' - \lambda(T' - T_n)}\right\} \tag{11}$$

So the guess is most likely if K_n and K' are similarily smooth (ie, $T' \approx T_n$), if an iteration gives a K' that is noticeably less smooth than the last accepted iteration, we are less likely to accept it.

As an initial test for the smoothness prior developed above, we attempt the gridwise change on a constant conductivity ($K_{\text{correct}} = 1.68$, using $\lambda = 100$). While we can still see the problem at the boundary points, they are limited to being a noticeable nuisance as oppposed to adamantly ruining an otherwise accurate reconstruction (whose mean comes to within $\approx 5\omega$ of 1.68). The next step is therefore to test the algorithm on a non-constant conductivity.

3.1 Results of the Smoothness Prior on the Tilted Plane and Gaussian Well

As a simple nonconstant trial, we look at a tilted plane with constant slope, given by

$$K_{\text{correct}}(i,j) = \frac{i+j}{20} + 1; \tag{12}$$

Once again, we take K_0 to be a matrix of all 1s and $\lambda = 100$. The boundary points again have trouble increasing from 1 to the desired values, and in so doing lower the mean value of the reconstruction; though we still consistentily get to within about 5ω of the solution (in 100000 iterations).

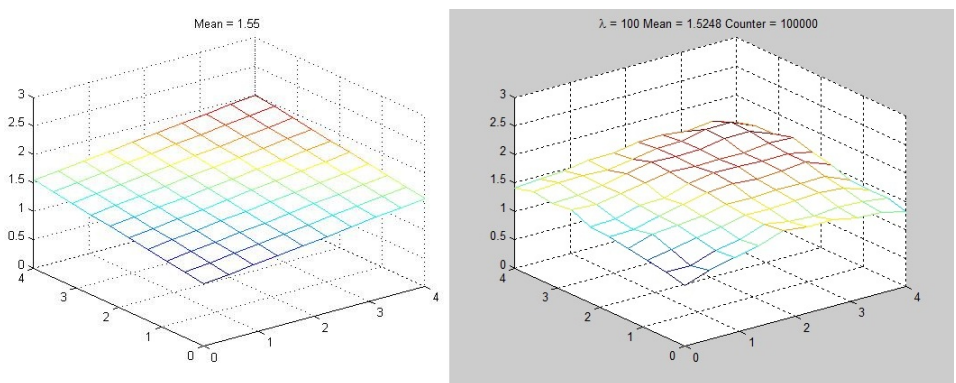


Figure 1: The 10×10 tilted plane we wish to reconstruct, and a reconstruction using the smoothness prior.

We now attempt to reconstruct a more complicated conductivity: a Gaussian well.

 ISBN: 978-84-614-6167-7

The Gaussian well is the first real challenge that the algorithm will face, and will be the main focus of the rest of the paper as it contains different regions which require different priors. It is given by the following equation

$$K_{\text{correct}}(i,j) = \sum_{i=1}^{m}\sum_{j=1}^{n}\left(\frac{2}{1+50e^{-\frac{[(x(i)-2)^2+(y(j)-2)^2]}{0.2}}}\right) \tag{13}$$

This conductivity represents a much more significant challenge, with both flat regions, and regions with steep slopes. After several trials, the optimal λs were found to be between 1 and 10, though obataining a specific value for which the reconstruction is best is impossible due to the high innaccuracy of the algorithm when faced with this well. There is an evident

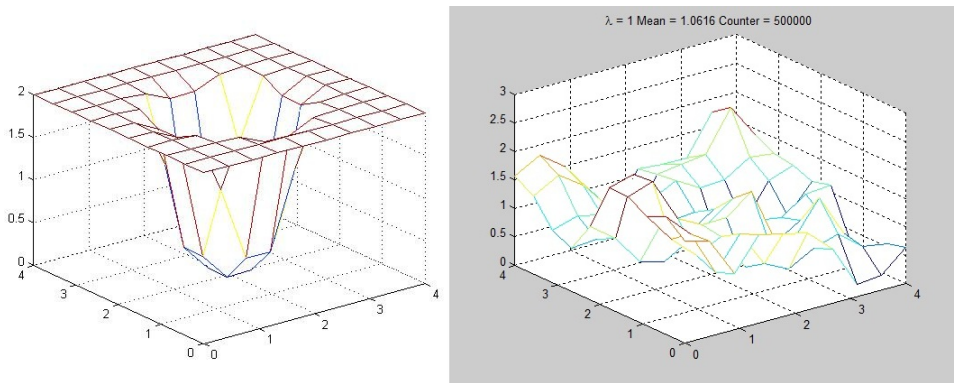


Figure 2: The 10×10 Gaussian well we wish to reconstruct, and a reconstruction using the smoothness prior.

need, at this point, for much more precision. We turn once again to priors, this time developing one that will look at the slopes of the reconstruction.

4 The Slope Prior

One of the main concerns in implementing a new prior is the generality mentioned earlier. In theory, one could use a prior that only accepts Gaussian wells of the form we have here, but that code would not be very versatile. We therefore try to keep our slope prior as general as possible. In keeping with this, we look at the ratios of adjacent slopes, both in the x and y directions, as follows:

$$S'_x(i,j) = K'(i+1,j) - K'(i,j) \tag{14}$$

$$S'_y(i,j) = K'(i,j+1) - K'(i,j) \tag{15}$$

 ISBN: 978-84-614-6167-7

and define

$$P'_x = \sum_{j=1}^{n} \sum_{i=1}^{m-3} \left| \frac{S'_x(i,j) + \varepsilon_0}{S'_x(i+1,j) + \varepsilon_0} - \frac{S'_x(i+1,j) + \varepsilon_0}{S'_x(i+2,j) + \varepsilon_0} \right| \tag{16}$$

$$P'_y = \sum_{j=1}^{n-3} \sum_{i=1}^{m} \left| \frac{S'_y(i,j) + \varepsilon_0}{S'_y(i,j+1) + \varepsilon_0} - \frac{S'_y(i,j+1) + \varepsilon_0}{S'_y(i,j+2) + \varepsilon_0} \right| \tag{17}$$

(where ε_0 is 0.00005)
The generality of these prior terms comes from the fact that they go to 0 so long as the conductivity doesn't change its mind. It is equally "happy" with a constant slope as it is with slopes that, say, double, at each grid point. It should be noted that the formulas above break down for regions where we have very small slopes adjacent to large ones, where one ratio goes to 0 while the other grows very large. Nevertheless, we now set

$$\alpha_s(K'|K_n) \equiv min\left\{1, e^{f_n - f' - \mu\left(P'_x + P'_y\right)}\right\} \tag{18}$$

With this new prior, we define

$$\alpha = \max\{\alpha_c, \alpha_s\} \tag{19}$$

and use that in the acceptance step of the MHMCMC algorithm.

4.1 Results of the Smoothness and Slope Priors

Again, as a first test of the algorithm, we test it on the tilted plane. The reconstructions reach the same precision in 100000 iterations as we had with only the smoothness prior, so we have not yet implemented anything that is too problem-specific to the Gaussian well.

The initial result of the test on the well is arguably substantially better, but still rather imprecise. In an attempt to see more clearly, we make the mesh finer (20×20). In addition, we set K_0 to be a matrix of all 2s. The results of the combined slope and smoothness priors are below.
As we can see, a substantial improvement has been made over the attempt in section (3.1). We now consistently obtain somewhat of a bowl shape. In comparing the solution we wish to achieve and the reconstruction we have, one notices that the major problem areas are the outer regions, where the conductivity is nearly constant. As previously stated, equations (16) and (17) break down when the slopes are vanishing, so it is reasonable to assume that with this alone, the reconstruction will not improve as substantially as we need it to. As before, we implement another prior to aid us.

 ISBN: 978-84-614-6167-7

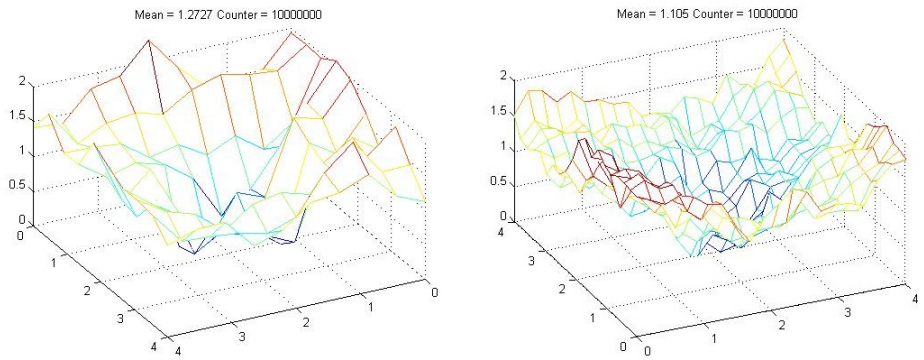


Figure 3: Results on the 10×10 and 20×20 Gaussian wells with parameters: $\lambda = 5,\ \mu = 10$ and $\lambda = 10,\ \mu = 7.5$, respectively.

5 Smoothness, Flatness, and Slope Priors

To help reconstruct the outermost regions of the well, we need a prior that will go to 0 for regions that have vanishing slope. The most obvious choice is therefore to use what we computed for the smoothness prior

$$T' = \sum_{j=1}^{n}\sum_{i=2}^{m}\big(K'(i,j) - K'(i-1,j)\big)^2 + \sum_{i=1}^{m}\sum_{j=2}^{n}\big(K'(i,j) - K'(i,j-1)\big)^2 \tag{20}$$

and set

$$\alpha_f(K'|K_n) \equiv min\left\{1, e^{f_n - f' - W(T')}\right\} \tag{21}$$

using

$$\alpha = \max\{\alpha_c, \alpha_s, \alpha_f\} \tag{22}$$

in the MHMCMC algorithm. Again, the worry that adding a new prior would undermine the generality of the algorithm can be eased by noting that we are simply accounting for a problematic case not treated by the slope prior, though we still test this prior on the tilted plane.

5.1 Results of the Combined Priors

The 20×20 tilted plane is given by

$$K_{\text{correct}}(i,j) = \frac{i+j}{40} + 1 \tag{23}$$

 ISBN: 978-84-614-6167-7

Running the MHMCMC algorithm with all three priors yields fairly accurate reconstructions, that miss the solution by $\approx 6\omega$. One should again note the presence of the familiar (though no less troublesome) boundary points.

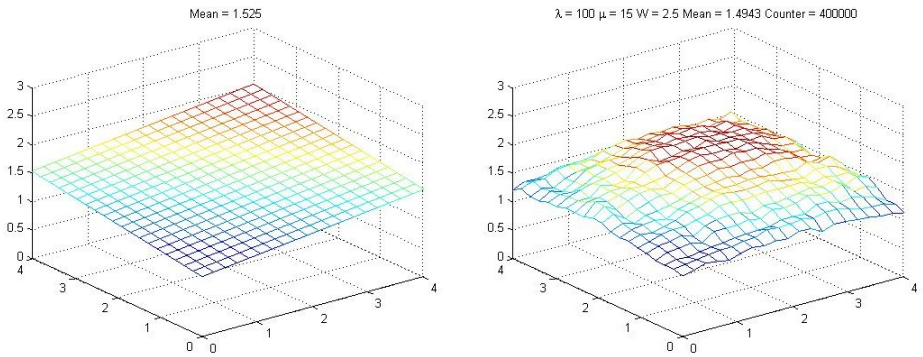


Figure 4: The 20×20 tilted plane, and a reconstruction using all three priors.

We now try once again to reconstruct the Gaussian well. The results of the added prior are apparent, and regions with vanishing slope are treated much more accurately than before. Perhaps the most successful reconstruction thus far is the following (though many more possible combinations of λ, μ and W must be explored).

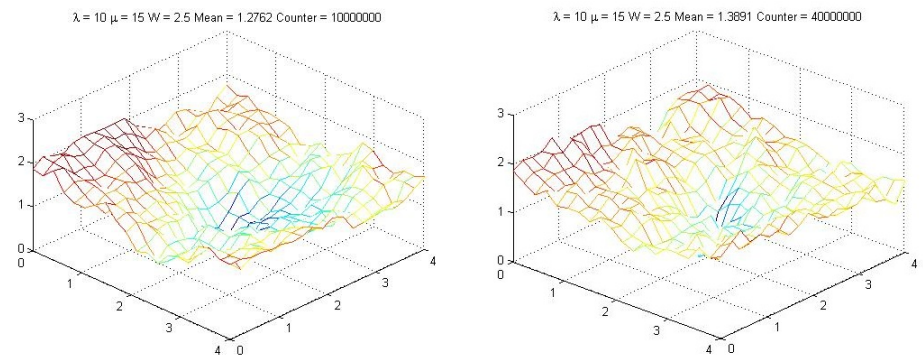


Figure 5: Results on the 20×20 Gaussian well using all three priors, at 10 and 40 million interations.

An obvious flaw in these reconstructions happens to be the width of the well, the algorithm is still capable of reconstructing the center of the well, and its depth, but it is often much narrower than in the actual solution. It would seem the algorithm has trouble starting to drop off from the vanishing slope region into the varying one. This exposes an inherent

 ISBN: 978-84-614-6167-7

problem with the patchwork we have taken thus far: getting the seams to match up nicely.

6 Conclusion

As we have seen, reconstructions of the heat conductivity greatly benefit from added priors. There is certainly much work left to be done, and a very careful analysis of the seams at which the various priors trade off is in order. However, we believe that in testing the algorithm against other complex nonconstant conductivities, which is the next step we plan to take, it is possible to complete the aforementioned analysis of the seams and reconstruct complex quantities via this patchwork method.

Acknowledgements

I was introduced to this problem at a National Science Foundation REU program at George Mason University, and would like to thank both of those institutions for the oppurtunity that gave me. I would also like to thank Professors Timothy Sauer and Harbir Lamba at GMU, who got me started on this project while I was there and helped me decipher the MHMCMC algorithm.

References

[1] T. Sauer, *Numerical Analysis*, Pearson Addison-Wesley, 2006.

[2] C. Fox, G. K. Nicholls and S. M. Tan, *Inverse Problems, Physics 707, The University of Auckland*, ch 7.

 ISBN: 978-84-614-6167-7

Volume IV

 ISBN: 978-84-614-6167-7

 ISBN: 978-84-614-6167-7

Contents:

Volume I

 ISBN: 978-84-614-6167-7

 ISBN: 978-84-614-6167-7

 ISBN: 978-84-614-6167-7

Contents:

Volume II

 ISBN: 978-84-614-6167-7

 ISBN: 978-84-614-6167-7

 ISBN: 978-84-614-6167-7

Contents:

Volume III

 ISBN: 978-84-614-6167-7

 ISBN: 978-84-614-6167-7

 ISBN: 978-84-614-6167-7

Contents:
Volume IV

 ISBN: 978-84-614-6167-7

 ISBN: 978-84-614-6167-7

 ISBN: 978-84-614-6167-7

Contents:
Abstracts & Late Papers

 ISBN: 978-84-614-6167-7

 ISBN: 978-84-614-6167-7

Proceedings of the 11th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE2011
Benidorm, Spain, 26-30 June 2011

Improved refined model of subannual nonuniform axial rotation of the Earth

Akulenko L.D. [(1)], Barkin M.Yu. [(2)], Markov Yu.G. [(2)], V.V. Perepelkin [(2)]

[(1)] Institute of problems of mechanics, [(2)] Moscow Aviation Institute - Technical University

By methods of celestial mechanics we have developed and refined the previously constructed a mathematical model of non-uniformity of the axial rotation of the Earth taking into account the secondary terms in the expansion of the luni-solar gravitational tidal torque and amended to the effect of perturbations from zonal tides.

The results of numerical simulation and the expansion of the rotational motion of the Earth's on components in accordance with the constructed model are given. The basic features of the model and examples of the construction of forecasts compared with observations and measurements of the International Earth Rotation Service (IERS) have been obtained.

Recently, high-precision measurements of tidal oscillations of the axial rotation of the deformable Earth. It is known [1-7] that in the tidal variations of the Earth rotation are observed the main components of tides (annual, semi-annual, monthly, fortnightly) and many different combination of harmonics of short-period tides. To study the variations in the rate of axial rotation of the Earth

 ISBN: 978-84-614-6167-7

it is introduced change (variation) of duration of the day - (lenght of the day changes) - $d(t)$ [1, 6]

$$d(t) = D(t) - D_0; \quad D(t) = \frac{r_0}{r(t)} \cdot D_0 .$$

Here $r(t)$ is the velocity of axial rotation of the Earth; $r_0 = 7.292115 \times 10^{-5}$ rad/s; D_0 is the duration of standard day (in the scientific literature is taken as the value of the standard unit of time of day, consisting of 86400 seconds for the atomic time scale TAI); $D(t)$ – duration of the day, meaning the duration in seconds TAI, corresponding to the rotation of the Earth on 360°, or to increasing time of 24 hours in UT1.

The value of $r(t)$ can be derived from the published values of the $l.o.d.(t)$ and has the form

$$r(t) \cong [1 + l.o.d.(t)/D_0] r_0 .$$

We use the classical dynamical equations of the Euler-Liouville problem with variable inertia tensor [5-7], which are represented in a certain form

$$\begin{gathered} \frac{dJ\boldsymbol{\omega}}{dt} + \boldsymbol{\omega} \times J\boldsymbol{\omega} = \mathbf{M}; \quad \boldsymbol{\omega} = (p,q,r)^T; \quad J = J^* + \delta J; \quad J^* = const; \\ J^* = diag(A^*, B^*, C^*); \quad \delta J = \delta J(t); \quad \|\delta J\| << \|J^*\|; \\ \mathbf{M} = \mathbf{M}_K + \mathbf{M}^S + \mathbf{M}^L . \end{gathered} \tag{1}$$

Here $\boldsymbol{\omega}$ is the angular velocity vector in the linked-Earth coordinate system (reference system) whose axes approximately coincide with the principal central axes of inertia $J*$ of "frozen" figure of the Earth, taking into account "the equatorial bulge" [5, 6].

 ISBN: 978-84-614-6167-7

Chosen coordinate system $x^{\alpha} = (x^1 x^2 x^3)$ rotates together with the Earth, and the axis x^3 indicates the direction closest to the direction of the instantaneous axis of proper rotation of the Earth, and the axis x^1 determines the position of longitude so that the longitude of the Greenwich meridian was approximately equal to zero. This coordinate system is derived from a non-rotating geocentric reference system by a spatial rotation, which takes into account the motion of the Earth's axis in space, and at its own rotation (the coordinate system qualitatively and quantitatively consistent with the ITRF). It is believed that small variations in the inertia tensor δJ may contain various harmonic components caused by the gravitational perturbing influence of the regular diurnal tides from the Sun and the Moon, and possibly others (annual, semi-annual, monthly, fortnightly, etc.). Additional perturbing terms are obtained by differentiating the angular momentum of a deformable Earth. They attributed to the vector $\mathbf{M}_K$ very complex structure, which is included additively in $\mathbf{M}$. Vectors $\mathbf{M}^{S,L}$ are the gravitational tidal disturbing moments from the Sun and Moon, respectively [6]. For example, the expression of components M_r^S have the following structure:

$$M_r^S = 3\omega_0^2 \left\{ \left[B^* + \delta B - \left(A^* + \delta A \right) \right] \gamma_p \gamma_q + \delta J_{pq} \left(\gamma_p^2 - \gamma_q^2 \right) + \delta J_{qr} \gamma_p \gamma_r - \delta J_{pr} \gamma_p \gamma_r \right\}, \qquad (2)$$

where ω_0 is a frequency of orbital motion; γ_p, γ_q, γ_r are direction cosines of the radius-vector of the Sun in connected reference frame; A^*, B^*, C^* are effective principal moments of inertia taking into account the deformation of the "frozen" Earth. They can be calculated with reasonable accuracy. Coefficients δA, δB, δJ_{pq}, δJ_{qr}, δJ_{pr} due to diurnal and semidiurnal tidal gravitational influence of the Moon and Sun. They are not amenable to direct measurement. For them, can be obtained indirect estimates based on measurements of the

 ISBN: 978-84-614-6167-7

characteristics of the process.

In order to improve the accuracy of interpolation and prediction of Earth's rotation irregularity at short time intervals seems appropriate to take into account the expansion of the lunisolar gravitational tidal time following the third harmonic in the expression $\cos\theta\sin\theta$:

$$\cos\theta\sin\theta = b\left(\theta^0,\psi^0\right)\cos\nu + d\cos 3\nu + \ldots; \qquad (3)$$
$$0.4 \le b \le \frac{4}{3}\pi^{-1}; \quad |d| << 1.$$

Here θ is an angle of nutation, ψ is an angle of precession, ν is a thrue anomaly of orbit.

An analysis of observations and measurements IERS amplitudes of these terms of higher degree of smallness and components due to perturbations of zonal tides, are of the same order. Thus, to refine the basic model of the Earth's rotation irregularity is also necessary to consider amendments to the tidal perturbations with small amplitudes. For this, a rezidium $\Delta d(t)$ is a oscillating change of the length of day $l.o.d.(t)$, caused by the tidal perturbations of the inertia tensor of the deformable Earth.

Averaging over the fast variable φ (φ is an angle of own rotation of the Earth) of expressions of the components of moumentums $M_r^{S,L}$ [5 - 7] let us to determine coefficients $\chi_{1r}^{S,L}$, $\chi_{2r}^{S,L}$ by corresponding terms in (2), which have form

 ISBN: 978-84-614-6167-7

$$\chi_{1r}^{S,L} = \frac{1}{2}\left\langle \frac{\delta B - \delta A}{C^*}\sin 2\varphi \right\rangle_\varphi - \left\langle \frac{\delta J_{pq}}{C^*}\cos 2\varphi \right\rangle_\varphi ; \quad (4)$$

$$\chi_{2r}^{S,L} = \frac{1}{2}\left\langle \frac{\delta J_{qr}}{C^*}\sin \varphi \right\rangle_\varphi - \left\langle \frac{\delta J_{pr}}{C^*}\cos \varphi \right\rangle_\varphi .$$

They are periodic functions of main frequencies ϑ_j of luni-solar tidal influences, as well as other tidal factors. For example

$$\chi_{1r}^{S,L} = b_{10}^{S,L} + \sum_j b_{1j}^{S,L}\cos(2\pi\vartheta_j\tau + \beta_{1j}^{S,L}). \quad (5)$$

Coefficients $\chi_{1r}^{S,L}$, $\chi_{2r}^{S,L}$ contain constant components (with coefficients $b_{i0}^{S,L}$), appropriate base model, as well as variables that caused by other tidal factors. The coefficients $b_{ij}^{S,L}$, $\beta_{ij}^{S,L}$ in expressions of $\chi_{1r}^{S,L}$, $\chi_{2r}^{S,L}$ of type (5) to be determined on the basis of observational data. In the above expression an argument τ means the time measured in years.

By integrating the third equation of system (1) for the component of the axial rotation of the Earth $r(t)$, we get (taking into account the changing of the tidal coefficients) the structure of variations of length of day:

$$l.o.d.(\tau) = d_1(\tau) + d_2(\tau) + \Delta d(\tau);$$

$$d_1(\tau) = a_0 + \sum_{i=1}^{4} a_{i0}\sin(2\pi\nu_i\tau + \alpha_i); \quad d_2(\tau) = \sum_{i=5}^{6} a_{i0}\sin(2\pi\nu_i\tau + \alpha_i); \quad (6)$$

$$\Delta d(\tau) = \Delta_1 d(\tau) + \Delta_2 d(\tau) = -\chi_{3r}\left(a_0 + \sum_{i=1}^{4} a_{i0}\sin(2\pi\nu_i\tau + \alpha_i) \right) +$$

$$+ \frac{1}{1+\chi_{3r}} \sum_{ij} \int a_{ij}\cos(2\pi\vartheta_j\tau + \beta_{ij})\cos(2\pi\nu_i\tau + \alpha_i)d\tau.$$

 ISBN: 978-84-614-6167-7

Here $\nu_1 = 1,\ \nu_2 = 2,\ \nu_3 = 13.28,\ \nu_4 = 26.68,\ \nu_5 = 3,\ \nu_6 = 40$ are frequencies, caused by lunar-solar perturbations; ϑ_j are frequencies of the lunar-solar tidal influences and other factors that determine variations of variations of the inertia tensor (assuming that the set of frequencies ϑ_j can be empirically adjusted during the numerical simulation); α_i are the phases of the oscillations; unknown a_{ij} are values to be calculated using the method of least squares from measurements of IERS. These factors are uniquely related to the unknown coefficients contained in the expression of gravitational tidal torque (2) taking into account the representations of the type (5).

We present graphical results of the expansion of the rotational motion of the Earth on the components $d_1(\tau),\ d_2(\tau),\ \Delta d(\tau)$ in accordance with the constructed model. Fig. 1 shows the interpolation of the basic model $d_1(\tau)$ (reviewed in [5 - 7]) in 2008 compared with observations and measurements IERS (Here and in the graphs τ is a time measured in days, contrasting line - theoretical curve, the usual line - observation data). The discrepancy between the basic model and observational data can be represented as two terms, due to the presence of unaccounted frequencies in the expansion of the gravitational tidal moment - $d_2(\tau)$ and disturbances of the zonal tidal potential of the Earth - $\Delta d(\tau)$, described by models (8), (9), respectively. Interpolation of these processes are shown in Figures 1 and 2 in comparison with carrying out the splitting observations IERS. Standard deviations of the main 9-parametric and extended models respectively:

$$\sigma_1 = 1.6 \times 10^{-4};\quad \sigma = 0.7 \times 10^{-4}.$$

 ISBN: 978-84-614-6167-7

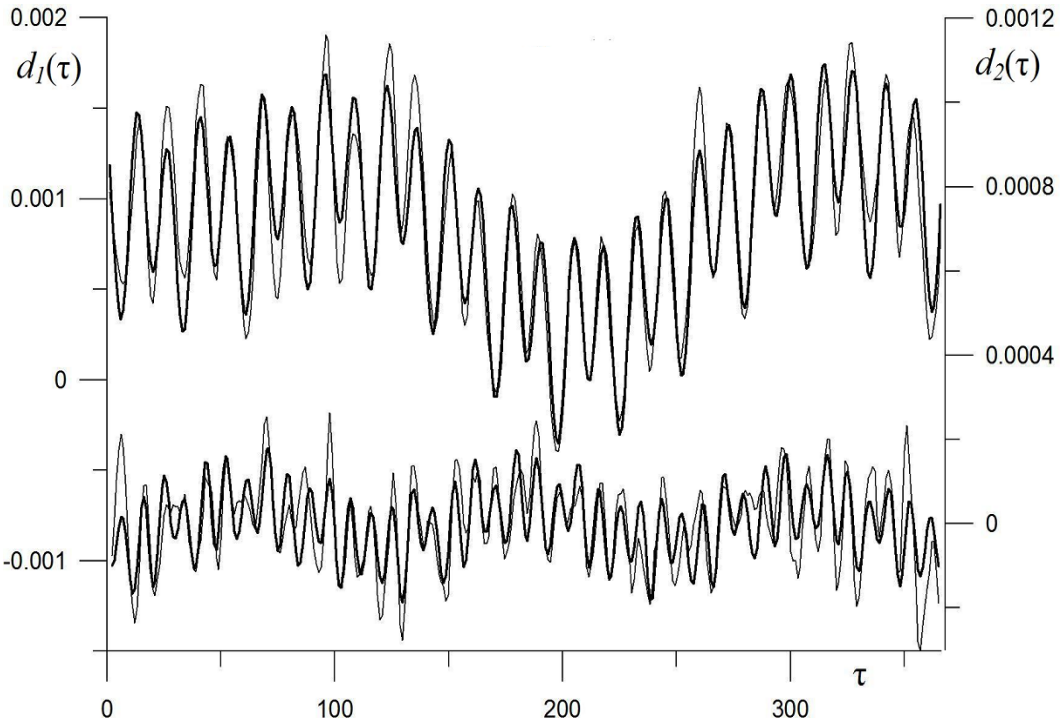
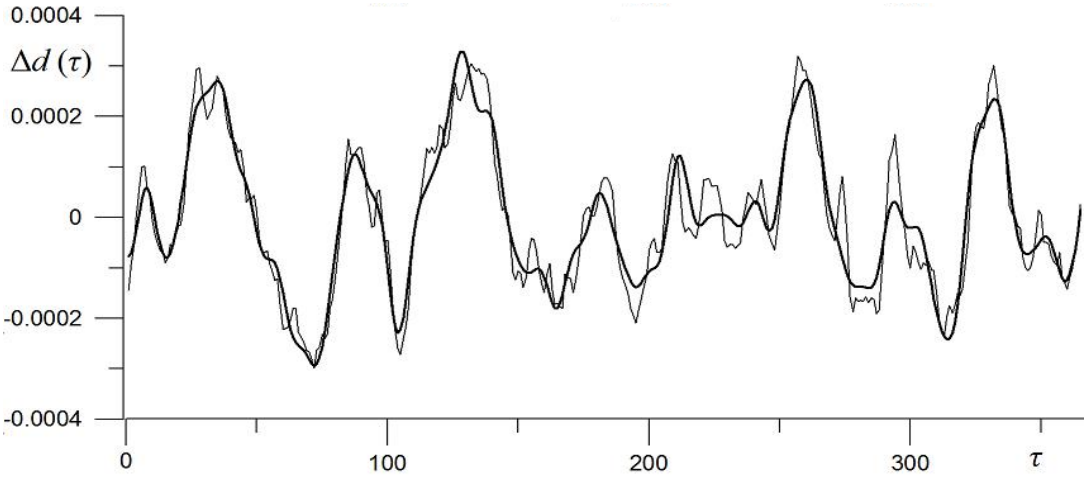


Fig. 1. An interpolation of variations of length of day in 2008, made with the basic model $d_1(\tau)$ compared with observational data, ICRE (upper graph) and interpolation of component $d_2(\tau)$ in comparison with the oscillations $d_2(\tau)$, extracted from observational data IERS (bottom graph).



 ISBN: 978-84-614-6167-7

Fig. 2 Interpolation of rezidiuma $\Delta d(\tau)$ (smooth line) in comparison with the oscillations of residuum $\Delta d(\tau)$, extracted from observational data IERS (zigzag line).

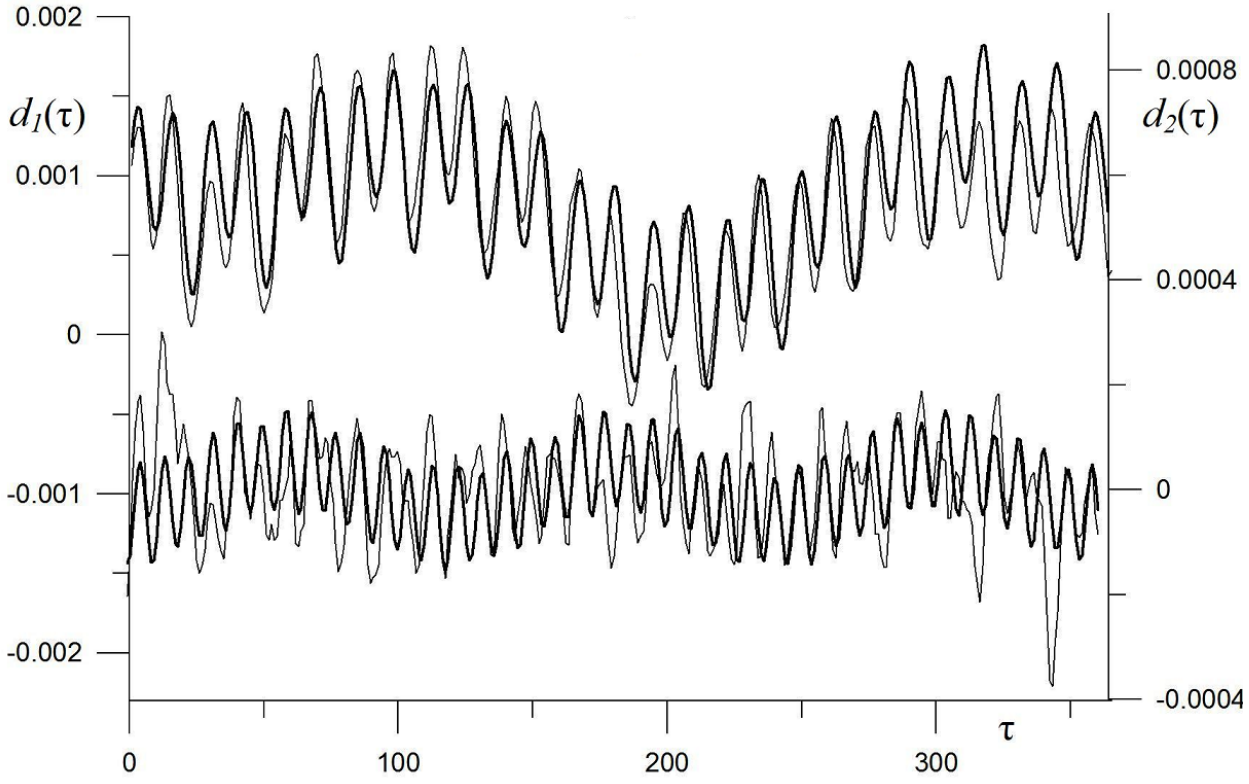


Fig. 3. Forecast of variations of the length of day for the year 2009, executed by the basic model $d_1(\tau)$ compared with observational data, IERS (upper graph) and the forecast component $d_2(\tau)$ in comparison with the oscillations $d_2(\tau)$, extracted from observational data IERS (lower graph).

 ISBN: 978-84-614-6167-7

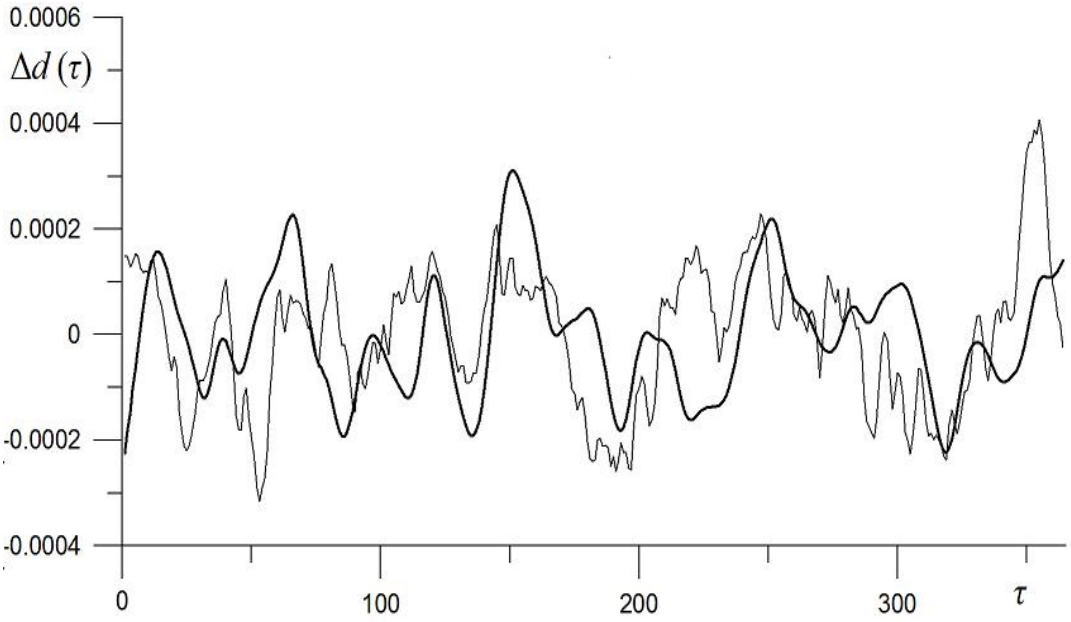


Fig. 4 Forecast of rezidium $\Delta d(\tau)$ (smooth line) for the year 2009 in comparison with the oscillations of rezidium $\Delta d(\tau)$, extracted from observational data IERS (zigzag line).

Forecast for 2009, made as a result of interpolation, the model constructed in the previous time intervals is presented in Fig. 3, 4. It corresponds qualitatively to the observational data and can be used for the analysis of geophysical processes of global character.

Analysis of the results of numerical simulation that forecasts constituents $d_2(\tau)$, $\Delta_1 d(\tau)$, $\Delta_2 d(\tau)$, clarifying the basic model $d_1(\tau)$, are different depending on the length of the projected interval reliability and accuracy characteristics. Therefore, to improve the forecast accuracy is required to assess the need to incorporate those or other terms of the model (depending on the length of the interval prediction) and to choose their optimal interpolation ranges. In particular, the reliability of forecasts of high-frequency components of the components $d_1(\tau)$ and $d_2(\tau)$ unlike from $\Delta_1 d(\tau)$ and $\Delta_2 d(\tau)$ decreases markedly with an increase in the projected range and affect the results. In the case

 ISBN: 978-84-614-6167-7

of short-prediction standard deviations in Fig. 1-4 projections of uneven rotation of the earth according to the basic 9-parametric model $d_1(\tau)$ and developed extended $l.o.d.(\tau)$ model 15-day interval are, respectively:

$$\sigma_1 = 2.9\times10^{-4}; \quad \sigma = 2.8\times10^{-4}.$$

Based on the above calculations we can conclude about the feasibility of using the extended model (6) in predicting short-and medium-sized intervals of time - within a few months.

This work was supported by the Russian Foundation for Basic Research (10-02-00595).

References

1. IERS Annual Reports, 2000-2002 (Frankfurt am Mein: BKG. 2001-2003).
2. Mank H. Macdonald, G. The Earth's rotation. Moscow: Mir, 1964, 384 pp.
3. Oduan C., Guinot B. The measurement of time. Fundamentals of GPS. MA: Cam-2002, 399s.
4. Capitane N., Guinot B., McCarthy D.D. Astron. and Astrophys. 2000, № 355, p. 398.
5. Akulenko L.D., Markov Yu.G., Perepelkin V.V Dokl. Academy of Sciences, 2007, T. 417, № 4, P. 483-488.
6. Akulenko L.D., Markov Yu.G., Perepelkin V.V Space and rocket science, 2009, Vol. 2 (55), P. 121-129.
7. Akulenko L.D., Markov Yu.G., Perepelkin V.V Space research, 2009, v. 4

 ISBN: 978-84-614-6167-7

Proceedings of the 11th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE2011
Benidorm, Spain, 26-30 June 2011

Simulation of non-lineal ordinary differential equations using the electric analogy and the code Pspice

Alhama, I., Alhama F. and Soto Meca, A.

[1]*Simulation by networks research groups*
Technical University of Cartagena (UPCT)
Campus Muralla del Mar, ETSII, 30202, Cartagena, Spain

emails: iam3@alu.upct.es, paco.alhama@upct.es,
antonio.soto@upct.es

Abstract

Linear or non-linear ordinary differential equations reproduce the behaviour of many physical events in science and engineering. The object of this work is to present an alternative solution to these equations based on the network method. The electrical model of the equations is composed of a principal network, which implements a balance between the addends of the differential equations, and auxiliary networks to implement the derivative terms. Non-lineal terms of the differential equations are implemented by a controlled source, a kind of device whose operation is quite intuitive. To illustrate the reliability and power of the method applications are shown.

Key words: Non-linear differential equation, electrical analogy, network method

1. Introduction

The study of the analogy between different processes governed by the same mathematical model (a set of differential equations) is an important goal since it relates phenomena mathematically equivalent. The use of electrical analogy to describe physical processes – even though it has been used in many areas of Science [1,2], especially in heat conduction [3] – remains under-exploited from the point of view of obtaining solutions to non-lineal problems, particularly.

 ISBN: 978-84-614-6167-7

Instead, its use has been relegated to that of an academic subject: description of the physical process in terms of electrical circuits, this is, as an alternative way of describing a problem [4]. In such exercises, once the electrical model is designed, it is forgotten and the solution is reached using classical analytical or numerical methods.

The electric analogy proposed, based on the network simulation method [5], has been demonstrated to be an efficient numerical method that provides reliable solutions for many kinds of problems in fields such as magnetohydrodynamic flow [6], Burger equation modelling [7], transport through membranes [8], heat transfer [9] and fluid flow and solute transport [10] among others. One of the main advantages of using this method is that, if the models are correctly designed, their simulation in suitable software provides (almost) the exact solution of the problem due to the powerful mathematics algorithms implemented in the circuit simulation codes.

In this work, we study the solution of lineal and non-lineal ordinary differential equations, such as those of harmonic and anharmonic oscillators and harmonic oscillators with damping. No restrictions are assumed as regards the order and degree of the equation as well or the kind of non-linearity involved.

The proposed analogy is based on the following [5]. On the one hand, the addends of the differential equation are considered as currents (branches in the main circuit) that enter (or leave) the only node of this main circuit, according to their sign; the unknown variable of the equation is the voltage at that node. The first derivative term (dy/dt), one of the branches of the main circuit, is simply the current flowing through a capacitor according to the constitutive equation ic=C(dVc/dt). The successive derivatives are obtained by auxiliary circuits formed by a capacitor whose capacitance is the coefficients of the term and a special kind of device, named controlled source, contained in the libraries of the software. Once obtained, these derivative terms are transported to the main circuit and, again, implemented by controlled current-sources that balance other terms in the common node according to their sign. The terms of the differential equation that depend on the unknown variable and/or its powers (integer or fractional), which must be balanced at the node of the main circuit, are also implemented by controlled current sources. Finally, the independent term (or constant) is implemented by a constant current source.

The model is completed by fixing the initial voltages at the capacitors which are defined by the initial conditions. Once the model is designed no mathematical manipulation is needed; the code Pspice [11] does this work with its powerful computational algorithms.

 ISBN: 978-84-614-6167-7

To show the proposed procedure, a detailed explanation of the design of network models is made and illustrative problems are presented.

2. Mathematical and network models

The differential equations to be solved contain derivative terms of any order and any degree, as well as terms which are arbitrary functions of the dependent variables and an independent term. Both derivative terms and terms which function of the dependent variable can be the power of real numbers. The only restriction is that the equations must contain one dependent and one independent variable. So, the terms that form the equations are proportional to the following expressions:

$y´,\ y´´, y´´´ \ldots, (y´)^q, (y´´)^q, (y´´)^q \ldots, y^q$ and a_o

where $y´=dy/dt$, $y´´ = d^2y/dt^2 \ldots$, q is a real number and a_o a constant. Examples of differential equations that can be solved by the method are:

$y´´ - c_1(y´)^2 - c_2\, y = 0$
$y´´´ + c_1(y´´)^2 - c_2 y^{1/2} - a_o = 0$
$y´ + c_1 \sin^2(y) - c_2 y + a_o = 0$
$(y´´´)^{1/3} + c_1(y´´)^2 - c_2\, y = 0$

with c_1 and c_2 constants. Initial conditions must be added to complete the mathematical model.

2.1 Auxiliary circuits

Before describing the design of the network main loop for a given problem, let us first look at the design of the auxiliary networks that implement the derivative terms of any order and any degree. Pspice, or any bother code for circuit simulation, contains a group of ideal controlled sources capable of assuming any kind of non-linearity; these, suitably connected with capacitors, provide the auxiliary circuits that implement any derivative term.

Four differrent sources can be used, Figure 1: E is a voltage source whose output is defined (by programming) as an arbitrary function of the voltage at any node (or voltages of any nodes) of the network, while H is a voltage source whose output is proportional to the current of a time-independent voltage source. The other two current-sources, G and F, have similar meanings.

Now, if we call V_j the voltage at node j, the auxiliary network of Figure 2 (a) formed by a capacitor (of capacitance $C_a=a_1$) and a voltage-controlled voltage-source (whose output voltage is $v_{E1} = V_j$, the same as the input voltage) is able to

 ISBN: 978-84-614-6167-7

provide the value $a_1(dV_j/dt)$ since the current through C_a is defined as $i_{Ca} = C_a(dV_j/dt) = a_1(dV_j/dt)$. A new auxiliary loop, formed by the current-controlled voltage-source H_1 and the resistor R_1 (of resistance a_1), provides the first derivative of V_j.

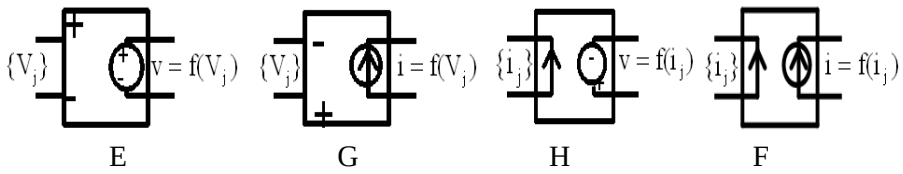


Figure 1 Controlled sources: E: voltage-controlled voltage-source, G: voltage-controlled current-source H: current-controlled voltage-source, and F: current-controlled current-source

The output of H_1 is a voltage whose value is the input current $i_{Vzero,1}$, i.e. the current of the ammeter $V_{zero,1}$ which, in turn, is the current of the capacitor C_a; consequently, the voltage through R_1 is $(1/a_1)a_1(dV_j/dt)=dV_j/dt$, the first derivative function of V_j. Resistor $R_{\infty,1}$ is included to satisfy the continuity criteria required by Pspice. Also, the use of V_{zero} as ammeter is prescribed by the requirements of Pspice: the input current of the controlled sources of type H must be specified as a current coming from a constant voltage source.

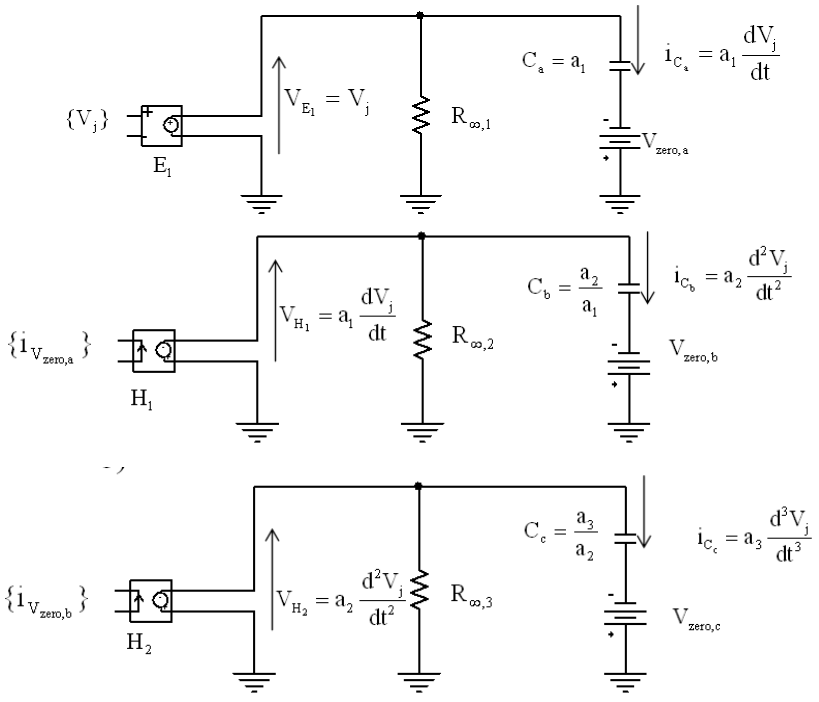


Figure 2 Auxiliary networks to implement the first derivative (a), the second derivative (b) and the third derivative (c)

 ISBN: 978-84-614-6167-7

In the same way, the second derivative of V_j, $a_2(d^2V_j/dt^2)$, is provided by the auxiliary network of Figure 2 (b). The output of the current-controlled voltage-source H_b, $v_{Hb} = i(V_{zero,1}) = a_1(dV_j/dt)$, defines the current through the capacitor C_b (of capacitance a_2/a_1) as $i_{Cb} = C_b(a_1d^2V_j/dt^2) = a_2(d^2V_j/dt^2)$. In addition, H_2 and R_2 (of resistance a_2/a_1) provide the second derivative of V_j (the voltage through the resistor R_2. The following derivative terms are implemented in the same way; Figure 2 (c) shows the network of the third derivative term. Quantities between brackets always denote the control variables that determine the input of the sources.

2.2 Main circuit loop

The main network is now formed by as many branches in parallel as terms of the differential equations. Each branch, in turn, drives a current (whose value is that of the term) that comes in or out of the common node, according to the sign of the term in the equation. The term related to the first derivative term, if it exists, is implemented by a capacitor, while the rest of the derivative terms are implemented by voltage-controlled current-sources that read from their respective auxiliary circuits.

When the derivative term has a degree different from unity (a real number), it is possible to define by programming the control function that provides the output current of the source. The rest of the terms of the differential equation are also implemented by voltage-controlled current-sources defined, also, by programming their output currents.

Any kind of non-linearity, such as arbitrary dependencies on the dependent variable is defined by software. Finally, the independent term is simply implemented by a constant source. A resistor of very high value that does not influence the solution is also located in parallel in the main circuit to satisfy the continuity requirements imposed by Pspice.

Whatever be the initial conditions, they are implemented in the model by giving initial voltages at the capacitors. The solution y(t) is read at the only node of the main circuit (as a consequence of the balance between the currents of the branches, Kirchhoff´s law) when the differential equation contains the term dy/dt, or in a node of the auxiliary circuit when the term does not exist – this will become clearer in the applications. As an example, Figure 3 shows the network model of the non-lineal equation

$$y'' + 0.1y' + 20\sin(y) = 0$$

under the initial condition

 ISBN: 978-84-614-6167-7

t=0, y=3, y´= 5

The main network contains three branches according to the three terms. The voltage at the common node (node 'A') is the solution y(t) of the equation. The first branch, with a capacitor of capacitance C_1 =0.1, drives the current $C_1(dy/dt) = 0.1y´$. This current, read as $i_{Vzero,a}$ in the ammeter $V_{zero,a}$, controls the current-controlled voltage-source H_1 whose output is a voltage of value 0.1y´; this leads that the current in the capacitor C_2 (of capacitance C_2 = 10) to have the value $C_b(d(0.1y´)/dt) = y´´$. This current converts itself in a voltage of the same value at the node C of the auxiliary circuit formed by H_2 and the resistor R_1 of value unity.

Now, we are ready to design the main loop, one of the branches of which (formed by C_1 and $V_{zero,a}$) has already been defined. The second branch, the voltage-controlled current-source G_1 controlled by $V_C = V_{(R1)}$, drives the current $i_{G1} = y´´$, while the third branch, a voltage-controlled current-source controlled by $V_A = y$, drives the current $y_{G2} = 10\sin(y)$. The initial condition is implemented by an initial voltage of 3 V at C_1 and 5 V at C_2.

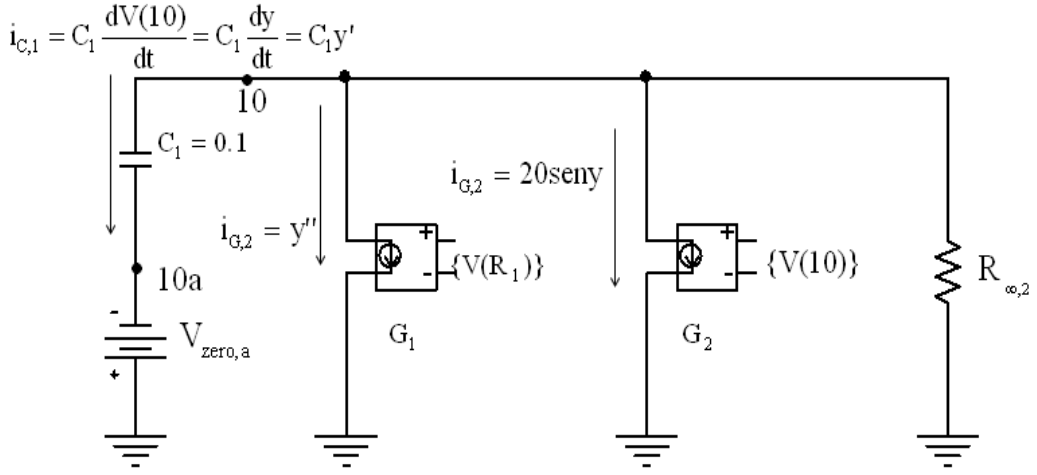


Figure 3 Network model of the differential equation y´ + 0.1y´´ + 20sin(y) = 0.
a): main loop, b): auxiliary circuits

It is important to mention that this is not the only way to design the network model of the above equation. For example, we may implement the branch of the term y´´ by a current-controlled current-source that directly reads the $i_{Vzero,b}$ and provides an output current of this value; this substitution deletes the auxiliary circuit formed by H_2 and R_1. However we have preferred to implement the currents of the branches of the main loop – the first derivative – by the same kind of device (a current source device, type G).

 ISBN: 978-84-614-6167-7

Using the programming rules of Pspice, the text file of the model – written to help readers – is the following (A is the node 10, B the node 20 and C the node 30):

```
C1 10 10a 0.1 IC=3
Vnulaa 0 10a 0
G1 10 0 VALUE = {V(30,0)}
G2 10 0 VALUE = {20*sin(V(10,0))}
Rinf2 10 0 1E10
H1 0 20 Vnulaa 1
Rinf1 20 0 1E10
C2 20 20a 1 IC=5
Vnulab 0 20a 0
H2 0 30 Vnulab 1
R1 30 0 1
.TRAN 0 2.5 UIC
.PROBE
.END
```

This text file requires very few programming rules. The last sentences, .TRAN 0 2.5 UIC and .PROBE, are used to specify the simulated time window [0-2.5s] – UIC means 'using initial conditions' – and the start up of Pspice graph ambient once the simulation has finished. Figures 4, provided by Pspice, shows the functions y(t), y´(t) and y´´(t). The units of the vertical axis, volts (V) or Amperes (A), must be converted into those related with the physical significances of the related functions.

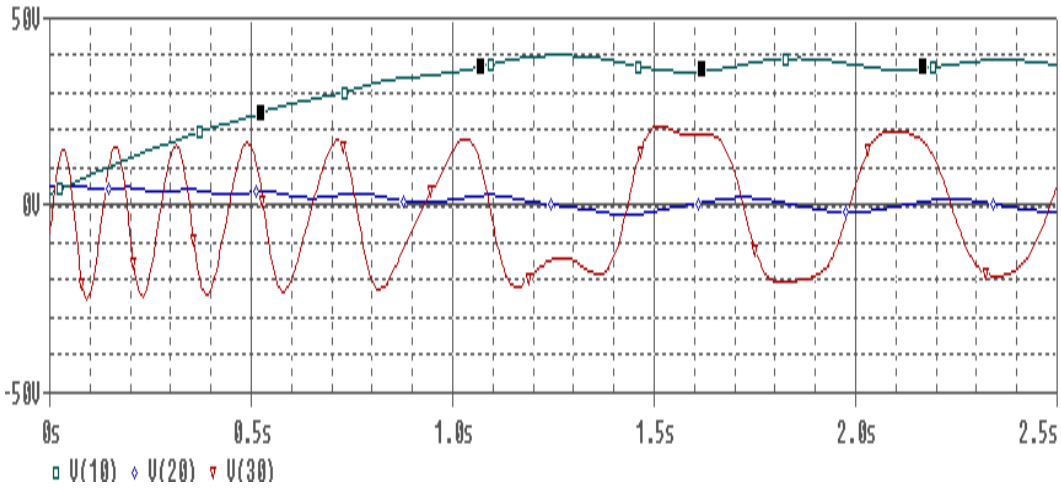


Figure 4 Numerical solution in Pspice output. y(t)=V(10); y´(t)=V(20a); y´´(t)=V(30a)

 ISBN: 978-84-614-6167-7

3. Applications

3.1. The skydiver equation

This equation applies to a person who jumps from an airplane to enjoy the sensation of free fall before opening a parachute. The net force acting on the person is his weight plus the drag force caused by the friction of air which is proportional to the square of the velocity. Assuming that y increases downwards, the momentum equation can be written in the form

$$d^2y/dt^2 = g - a_o(dy/dt)^2$$

with g the gravitational acceleration and a_o a constant that depends on the mass of the person, on the density of the air, on the front cross-sectional area and on the drag coefficient (a_o=0.003). The simulation in Pspice, with the initial conditions t=0, y=y´=0, provides the solution for the location, velocity and acceleration of the skydiver; Figure 5 shows these unknowns in the output graph ambient of Pspice . The transient period lasts 20 s approximately and the final steady velocity is 57.15 m/s. Computing time in a PC is of the order of 0.1 s.

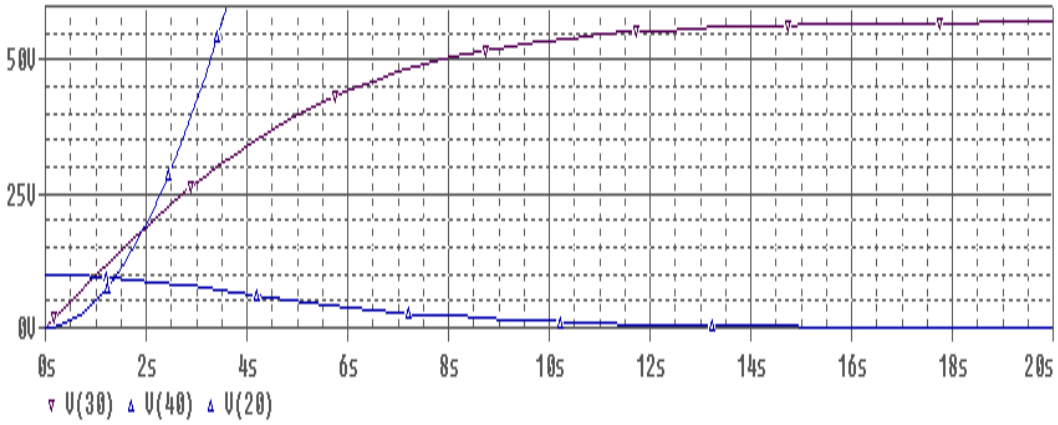


Figure 5 Numerical solution of the skydiver equation. Location: V(node A) or V(20); Velocity: V(node C) or V(30); Acceleration: V(node E) or V(40)

3.2. The anharmonic oscillator (pendulum equation)

The nonlinear differential equation

$$d^2y/dt^2 = - a_o \sin y$$

is that of a pendulum called anharmonic oscillator. This equation does not have an analytical solution that can be expressed by a finite number of terms. Note that the main loop only contains two branches corresponding to the two terms of the

 ISBN: 978-84-614-6167-7

differential equations. Figure 6 (a) shows the solution y(t) for the initial condition t=0, y=0, y´=1 with $a_o = 10^3$. The anharmonic effect is better appreciated comparing the Fast Fourier Transform of y(t) and a pure harmonic sine function of the same frequency; this comparison, which is also provided by Pspice, is shown in Figure 6 (b). In this detailed figure, the bandwidth of y(t) is appreciably larger than that of the sine function.

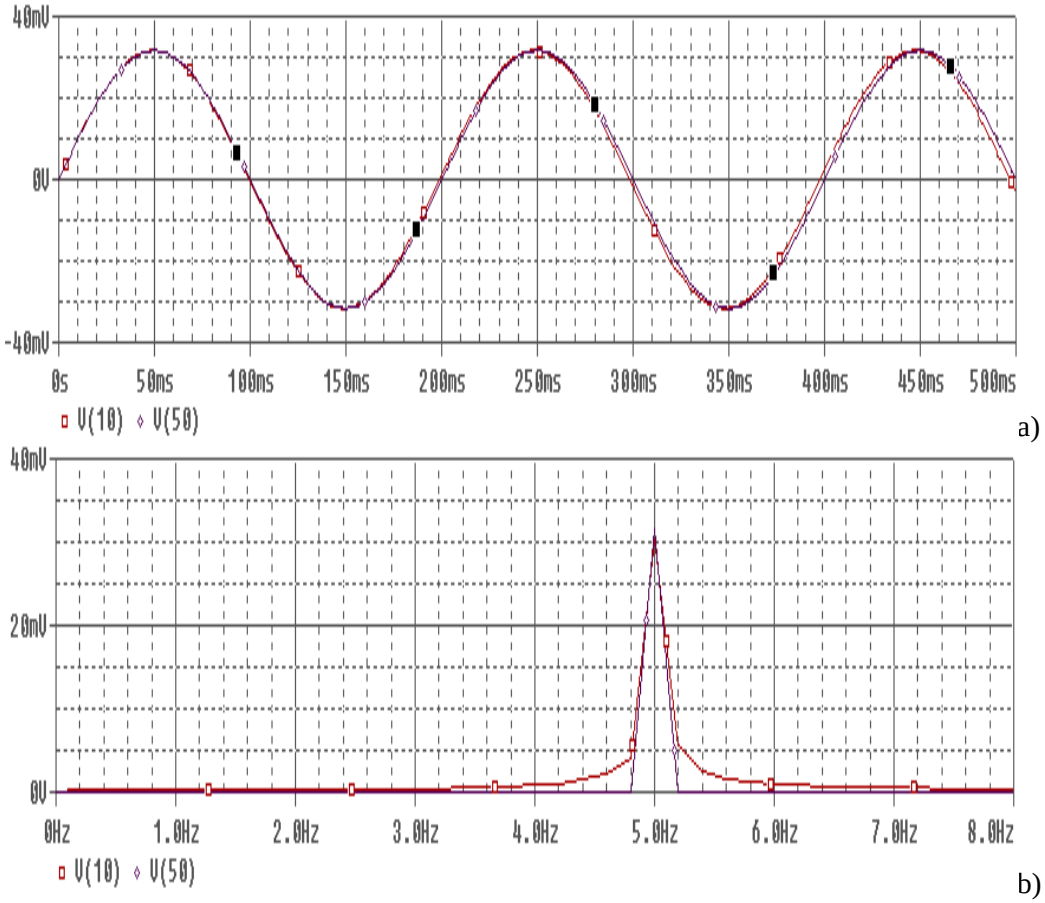


Figure 6 Numerical solution of the anharmonic oscillator. a) y(t): V(10) and pure harmonic function: V(50); b) FFT comparison of y(t) with a pure harmonic function

3.3. The damped oscillator

The equation

$$d^2y/dt^2 = -\alpha\, y - \beta\, (dy/dt)$$

where α and β are constants applied to a damped oscillator which moves not large enough with relatively small velocities to prevent turbulence in the surrounding

 ISBN: 978-84-614-6167-7

fluid. This is an oscillator in which the force producing the oscillation obeys Hooke's law and in which the oscillating body experiences fluid friction proportional to the first power of the speed. The solution y(t) for α = 4 and 400 and β=0.5, and initial conditions t=0, y=y´=1, is depicted in Figure 7 (a) and (b), respectively.

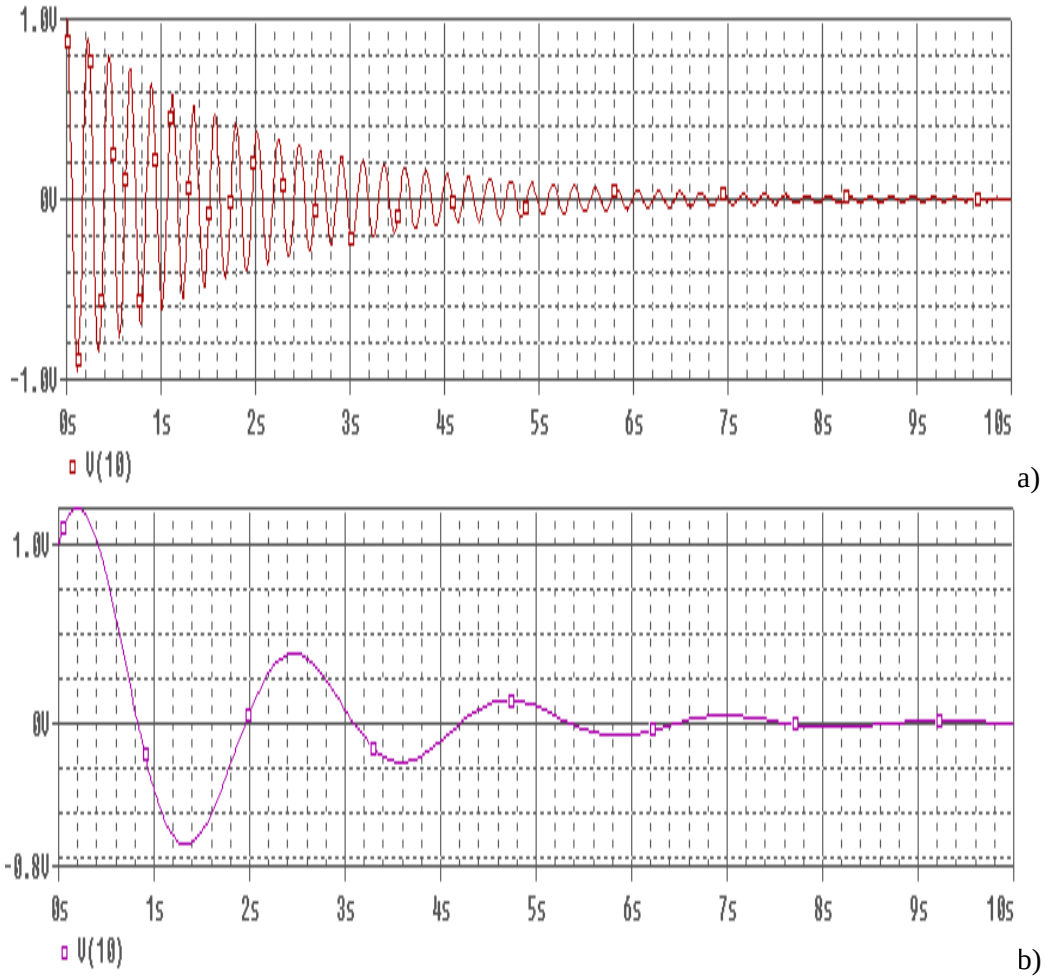


Figure 7 Numerical solution of the damping oscillator.
a): α=400, β=0.5; b): α=4, β=0.5

4. Conclusions

The applications proposed in this work demonstrate that the method based on the electrical analogy and the code Pspice to numerically solve ordinary differential equations is easy to understand for first year students of technical and scientific degree courses since it uses simple rules. This analogy allows us to implement the

 ISBN: 978-84-614-6167-7

successive linear derivative terms of the ordinary equations by means of auxiliary networks simply formed by capacitors and controlled sources; the latter implements the successive derivatives as well as the non-linear terms of the equation. With the proposed method, the user does not need to know the numerical algorithms required to solve the equations or any other mathematical manipulation since this work is made by the code Pspice; in addition, the design of the circuit, whatever the order and degree of the equations, is relatively straightforward thanks to the very few rules required, since the model only implements three types of electric devices: capacitors, resistors and controlled sources specified by software.

5. References

[1] J. F. LÓPEZ SÁNCHEZ, ALHAMA, F. AND GONZÁLEZ-FERNÁNDEZ, C.F. *Introduction and performance of species in a diffusive Lotka-Volterra system with time dependent coefficients.* Ecological Modelling 183, 1-9, 2005.

[2] A. SOTO MECA, F. ALHAMA AND C.F. GONZÁLEZ-FERNÁNDEZ. *An efficient model for solving density driven groundwater flow problems based on the network simulation method*. J. Hidrology 339, 39-53, 2007.

[3] CAMPO, A. AND ALHAMA, F. *The R-C analogy provides a versartile computational tool for unsteady, unidirectional heat conduction in regular solid bodies cooled by adjoining fluids*. Int. J. Mechanical Engineering Education 31 (3) 233-244, 2003.

[4] A.F. MILLS. *Heat and mass transfer*. Richard D. Irwin Inc., Chicago, 1995.

[5] C.F. GONZÁLEZ-FERNÁNDEZ. *Heat Transfer and the Network Simulation Method*. Horno J. Ed. Research Signpost, Kerala. 2002.

[6] O. ANWAR BÉG, J. ZUECO AND H.S. TAKHAR. *Unsteady magnetohydrodynamic Hartmann Couette flow and heat transfer in a Darcian channel with Hall current, ion slip, viscous and Joule heating effects: Network numerical solutions*. Comm. Nonlinear Science Numerical Simulation, 14 (4), 1082-1097, 2009.

[7] J. ZUECO. *A network thermodynamic method for the numerical solution of Burgers' equation*. Mathematical Computer Modelling, 47, 401-410, 2008.

[8] C. F. GONZÁLEZ-FERNÁNDEZ, M.T. GARCÍA HERNÁNDEZ AND J. HORNO, J. *Computer simulation of a square scheme with reversible and irreversible charge transfer by the network method*. J. Electroanal. Chem., 395, pp.39-44, 1995.

[9] J. A. MORENO NICOLÁS, F.C. GÓMEZ DE LEÓN AND F. ALHAMA. *Solution of temperature fields in hydrodynamics bearing by the numerical network model*. Tribology International 40, 139-145, 2007.

 ISBN: 978-84-614-6167-7

[10] A. SOTO MECA, F. ALHAMA AND C.F. GONZÁLEZ-FERNÁNDEZ. *An efficient model for solving density driven groundwater flow problems based on the network simulation method*. J. Hidrology 339, 39-53, 2007.

[11] *PSPICE*, versión 6.0: Microsim Corporation, 20 Fairbanks, Irvine, California 92718, 1994.

 ISBN: 978-84-614-6167-7

Proceedings of the 11th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE2011
Benidorm, Spain, 26-30 June 2011

Design for an asymmetrical cyclic neutron activation process for determining fluorite grade in fluorspar concentrate

T. Alonso-Sánchez [1], M.A. Rey-Ronco[2] and M.P. Castro-García[2]

[1] *Departamento de Energía, Universidad de Oviedo*

[2] *Departamento de Explotación y Prospección de Minas, Universidad de Oviedo*

emails: tjalonso@uniovi.es, rey@uniovi.es, castromaria@uniovi.es

Abstract

This paper presents a design for a neutron activation process used for samples of fluorspar concentrate, in which the duration of the first cycle is different from the rest, and subsequent activation cycles take place once radioactive decay has partially reduced the radioactive concentration of the element of interest. We have called this type of cycle "asymmetrical."
This method is compared with the conventional procedure, in which the activation and reading cycles are of equal duration, and the advantages of the new procedure are examined.

Keywords: neutron activation, cyclic, symmetrical, asymmetrical

1 Introduction

Nuclear activation analysis is considered as a method for qualitative determination of elements based on the measurement of characteristic radiation from radionuclides formed directly or indirectly by neutron irradiation of the material [1]. A normal activation measurement is subdivided into two phases: (1) the irradiation of a suitable sample, (2) the counting of the induced activity.

In order to carry out a neutron activation analysis, one must have a sample, a neutron source, a gamma ray detector, and a deep understanding of the reactions that take place when the material is exposed to neutrons.

The different neutron activation methods are classified according to when the gamma rays are detected:

 ISBN: 978-84-614-6167-7

- Prompt Gamma-Ray Neutron Activation Analysis (PGNAA), in which the measurement is taken during radiation, or
- Delayed Gamma-Ray Neutron Activation Analysis (DGNAA), in which the measurements are taken based on the radioactive element's decay emissions.

In general, DGNAA is used in the majority of applications [2-6]. In this case, during the sample's activation phase an element was exposed to neutrons for a period of time t, which produced a new element. If this product is radioactive, its concentration increases over time according to the exponential function $g(t) = k \cdot \left(1 - e^{-\lambda \cdot t}\right)$. This function is shown in Figure 1a. In the equation $\lambda = \frac{\ln(2)}{T_{1/2}}$, where $T_{1/2}$ is the half-life of the product of the reaction, and k is a parameter that depends on several factors, among them the concentration in the sample of the element that produced the radioactive reaction, the type and intensity of the neutron source, the cross-section of the reaction of interest, or the spatial arrangement of the components. When τ reaches a large enough value, $e^{-\lambda\tau} \rightarrow 0$, the concentration of the radioactive product stabilizes, in a state called "saturation".

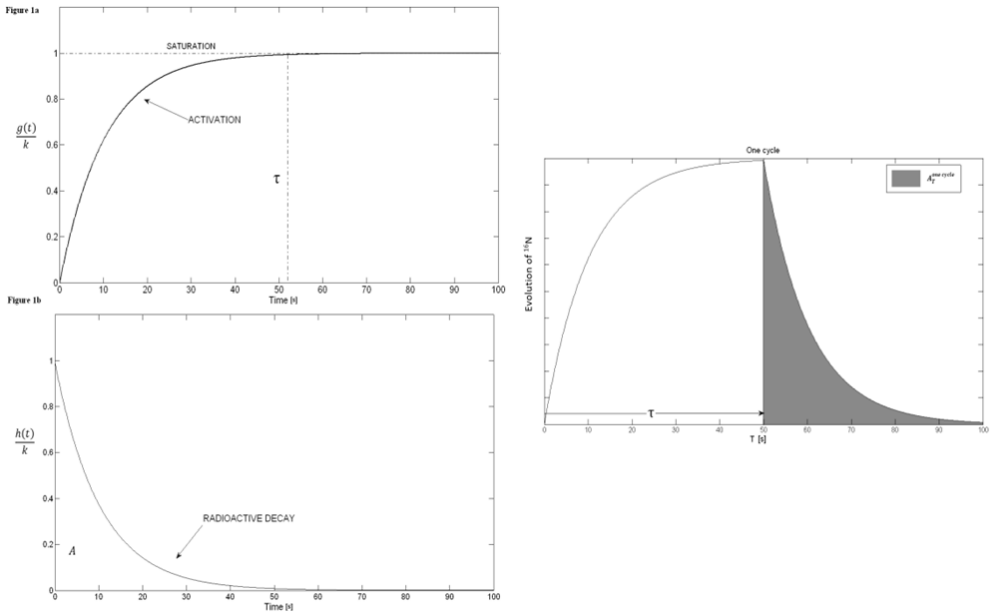


Figure 1. Radioactive activation and decay in one activation cycle. (Note that the y-axis expresses $\frac{g(t)}{k}$ and $\frac{h(t)}{k}$).

Later on, after the total activation time τ, the sample is removed from the neutron source either right away, or else after a short delay due to the time needed to transfer the sample. The spectrum of the gamma rays emitted during the decay of the nuclear reaction are recorded over a period of time called

 ISBN: 978-84-614-6167-7

"counting time." The spectrum indicates the intensity of the gamma ray radiation recorded at any given moment. The spectrum is analyzed based on two factors: on the one had the magnitude of the energy emitted indicates the nature of the radioactive product, and on the other, the intensity of the radiation for a given energy reading (expressed in counts per second "cps") indicates the concentration of the radioactive element that was produced. The concentration of the radioactive product decays with time, expressed by the equation $h(t) = k \cdot \left(1 - e^{-\lambda \cdot \tau}\right) \cdot e^{-\lambda \cdot (t-\tau)}$ (Figure 1b).

The detector records cumulative gamma ray spectra over a given period of time. For any given energy interval, this cumulative reading is equal to the value of the integral of the above function during this period. (Area A of Figure 1).

This research team has experimentally demonstrated the viability of using a single-cycle neutron activation method for analyzing the fluorite content in a sample of fluorspar concentrate [7-8]. When the sample is irradiated with neutrons from an isotopic Americium-Beryllium neutron source, hundreds of nuclear reactions are produced. The spectrum of the gamma rays emitted by the radiation's products is recorded throughout the counting time with a NaI(Tl) detector. The most important reaction in this study is $^{19}F(n,\alpha)N^{16}$. ^{16}N is a radioactive element which is only brought into existence by nuclear reactions, and whose most important characteristics are:

- a half-life, $T_{1/2}$, of 7.13 s [9],
- the emission of high-energy gamma rays of around 7,000keV, which do not interfere with the energy emitted by other radioactive elements produced in the sample during neutron bombardment, [10],
- a highly effective cross-section or probability of occurrence in the energy range of the isotopic source used (between 3 and 10MeV),

The team demonstrated that the fluorine (F) present in the sample came only from fluorite, and that all the ^{16}N that was produced came solely from the above reaction, which means that there is a direct relationship between the fluorite grade in the mineral concentrate and the ^{16}N found by the detector during the decay phase at an energy interval of around 7,000 keV. The name of this reading is $A_T^{one\ cycle}$, which refers to the area recorded by the detector in the energy interval corresponding to ^{16}N, during an activation and counting process equal to T, with just one activation. It was found that for a given total experiment time T, the maximum value of $A_T^{one\ cycle}$ occurs when the activation and counting times are equal to $T/2$.

While a very high correlation coefficient was found between a fluorspar sample's fluorite content and the area of $A_T^{one\ cycle}$, the amount of ^{16}N produced during irradiation in an activation cycle is scant, which means that the number of counts recorded by the detector is low in this energy interval. In

 ISBN: 978-84-614-6167-7

hopes of improving the results' resolution, we are trying to increase the intensity of the counts recorded by the detector, by using the neutron source itself.

Because of this, we decided to try increasing the area's value by using the cyclic activation technique. The principle of the cyclic process has been described by various authors [11-12]. In this method, the sample is first irradiated for a short period of time, and then, after a short delay from the end of irradiation, the radiation emitted by the sample is counted, after which the sample is irradiated again and the entire process is repeated for a number of cycles *n* [13].

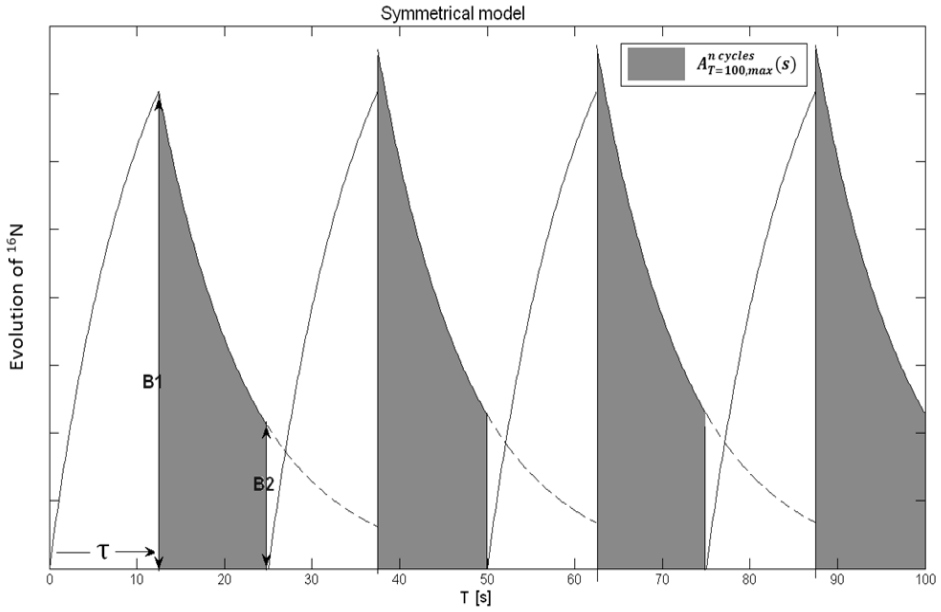


Figure 2. Symmetrical cyclic activation.

In this case, if in the first activation 70% of the concentration of ^{16}N at saturation is reached (Figure 2), then the amount of ^{16}N present at the outset of the second activation cycle will be around 2%, as shown in Figure 2. This moment marks the beginning of a new activation phase, in which new ^{16}N from the second activation begins to form. At the end of the second activation, the amount of the ^{16}N left over from the first activation is around 6% of saturation, such that the second decay begins with a concentration of around 76% of saturation. The amount of ^{16}N from the second activation is essentially the same as the amount present at the end of the first activation, and the third and subsequent cycles are likewise similar to the second. This type of cycle has been referred to as "symmetrical cyclic activation."

For each total process duration T, the following parameters specific to the cyclic activation were defined: activation time and counting time for each phase in a cycle, along with the number of cycles *n*. In addition, optimum

 ISBN: 978-84-614-6167-7

values were chosen in order to maximize the detector response. Here the detector response is referred to as $A_T^{n\ cycles}(s)$.

Based on this work, the optimum number of cycles, n_{op}, was determined for various experimental times, in order to maximize the detector response for radiation from ^{16}N [14]. Furthermore, it was concluded that the activation and counting times in each cycle should be equal, and should have a value of $\tau_{op} = \frac{T}{2 \cdot n}$.

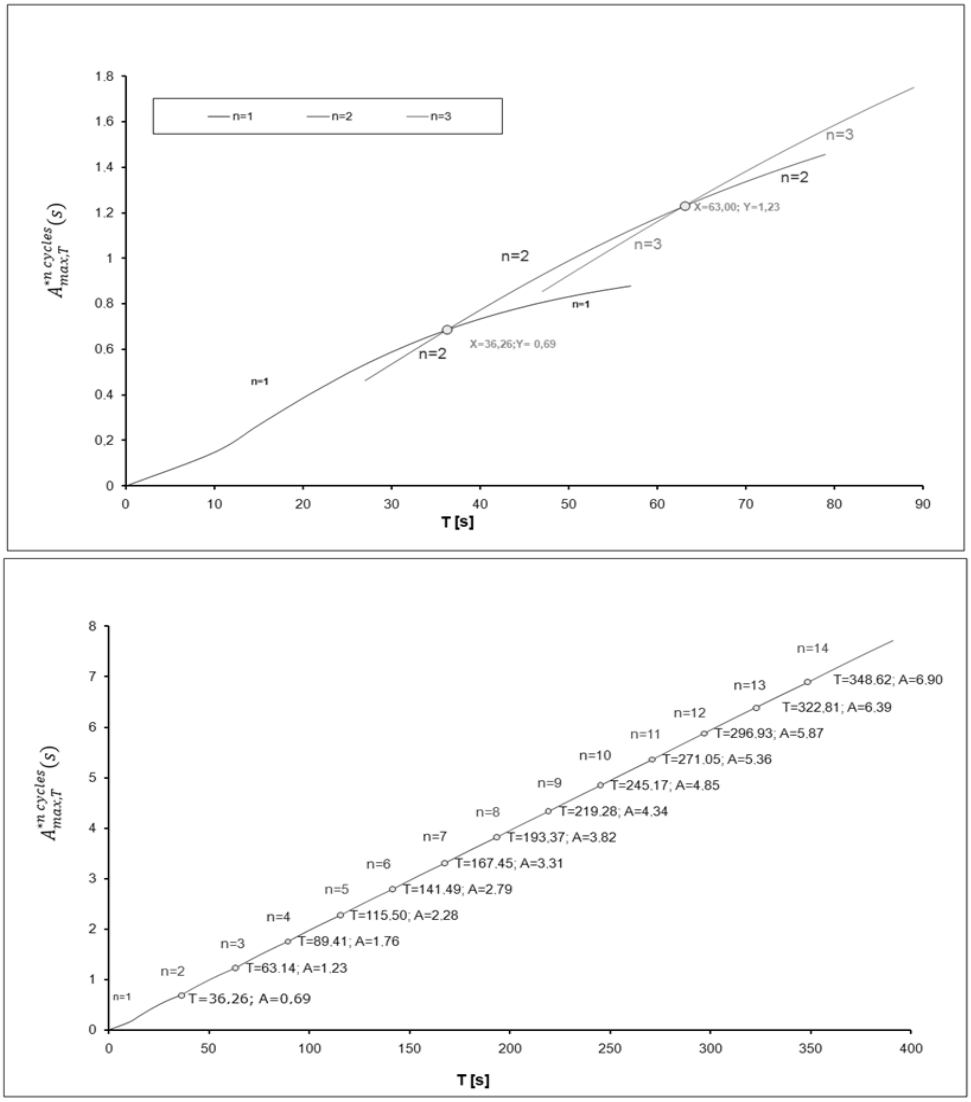


Figure 3. Relationship between experiment time *T*, number of cycles *n* with respect to $A_{max,T}^{*n\ cycles}$, following M. A. Rey-Ronco [14].

The results of these trials are shown in Figure 3. The optimum number of cycles for an experimental time of T=200s is eight, and the duration of each phase is $\tau_{op} = \frac{T}{2 \cdot n} = \frac{200}{2 \cdot 8} = 12.5s$.

 ISBN: 978-84-614-6167-7

The maximum area for this situation is given by substituting T=200 in the following equation, which corresponds to the line in Figure 3:

$$A_{max,T}^{*n\ cycles} = 0.0199 \cdot T - 0.0221 \qquad [1]$$

Which gives us $A_{max,T}^{*n\ cycles} = 3.9579$.

2 Analysis of the asymmetrical cyclic process

The next step is to examine another type of cycle, which we have called "asymmetrical," and analyze whether it is better at detecting fluorite than the symmetrical cycle. This process consists of a first activation cycle up to the concentration of ^{16}N (B_1), followed by a decay cycle down to a concentration of ^{16}N (B_2), which is still significantly higher, and then repeating the activation and decay cycles within the limits of B_1 and B_2, as can be seen in Figure 4. The goal is to increase the number of cycles and record higher values for the area under the decay curve, while disregarding the tail of the curve. The number of cycles, the activation time and the counting time should be optimized in order to obtain the maximum area $A_T^{n\ cycles}(a)$ for experiment time T. It can then be determined whether there is indeed an improvement over the symmetrical cyclic activation $A_T^{n\ cycles}(s)$.

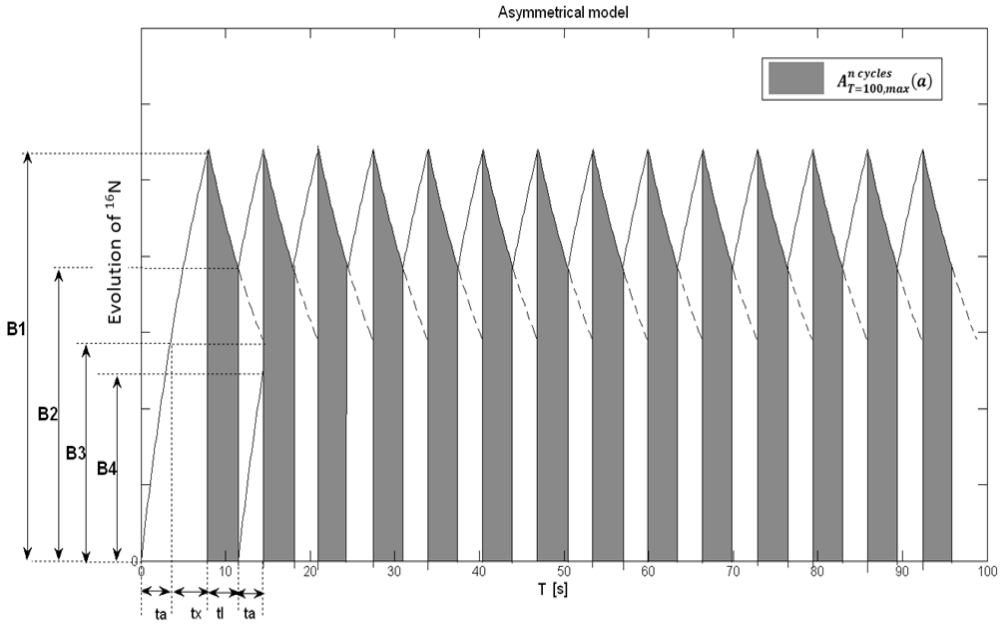


Figure 4. Representation of the activation-decay cycles in the asymmetrical model. The grey sections represent the area of $A_T^{*n\ cycles}(s) = A_T^{n\ cycles}(s) \cdot \frac{\lambda}{k}$ for experiment time T=100s.

Figure 4 shows the variation in the concentration of a radioactive element over time during an asymmetrical cyclic activation process, as defined above.

According to the initial hypotheses:

 ISBN: 978-84-614-6167-7

$$B_1 = k \cdot (1 - e^{-\lambda \cdot (ta+tx)}) \quad [2]$$

$$B_2 = B_1 \cdot e^{-\lambda \cdot tl} \quad [3]$$

$$B_3 = k \cdot (1 - e^{-\lambda \cdot ta}) \quad [4]$$

$$B_4 = B_1 \cdot e^{-\lambda \cdot (tl+ta)} \quad [5]$$

In Figure 4 the parameter B_1 represents the maximum concentration of ^{16}N, and the area marked $A_T^{*n\ cycles}(a)$ in grey represents the measurement or detector response of gamma rays for the total experiment time T and for *n* cycles. Based on Figure 4, it is clear that:

$$B_1 = B_3 + B_4 \quad [6]$$

The area of each activation and counting cycle is expressed as:

$$A_T^{n\ cycles}(a) = n \cdot \int_0^{t_l} k \cdot \left(1 - e^{-\lambda \cdot (t_a+t_x)}\right) \cdot e^{-\lambda \cdot t} dt \quad [7]$$

Which can be reduced to:

$$A_T^{n\ cycles}(a) = n \cdot \frac{k}{\lambda} \cdot \left(1 - e^{-\lambda \cdot (t_a+t_x)}\right) \cdot [1 - e^{-\lambda \cdot t_l}] \quad [8]$$

By making $A_T^{*n\ cycles}(a) = A_T^{n\ cycles}(a) \cdot \frac{\lambda}{k}$ in the above expression, the parameters λ and *k* can be avoided, as they are constant factors that depend on each individual sample, or on the decay characteristics of ^{16}N, or on the instruments being used, and they have no bearing on the maximization of the area:

$$A_T^{*n\ cycles}(a) = n \cdot \left(1 - e^{-\lambda \cdot (t_a+t_x)}\right) \cdot [1 - e^{-\lambda \cdot t_l}] \quad [9]$$

2.1 Maximization of the area function $A_T^{*n\ cycles}(a)$

In order to optimize the asymmetrical cyclic activation process, one must determine the number of cycles, *n*, and the times t_a, t_x, and t_l for a given experiment time T which will maximize the function $A_T^{*n\ cycles}(a)$.

2.1.1 Finding a relationship between variables

By substituting expressions [2], [4] y [5] for the terms in Equation [6], we get:

$$k \cdot (1 - e^{-\lambda \cdot (t_a+t_x)}) = k \cdot \left(1 - e^{-\lambda \cdot t_a}\right) + k \cdot \left(1 - e^{-\lambda \cdot (t_a+t_x)}\right) \cdot e^{-\lambda \cdot (t_l+t_a)} \quad [10]$$

From which we can deduce that:

 ISBN: 978-84-614-6167-7

$$e^{-\lambda\cdot(t_a+t_x)} = \frac{k\cdot\left(1-e^{-\lambda\cdot t_a}\right)-k+k\cdot e^{-\lambda\cdot(t_l+t_a)}}{-k+k\cdot e^{-\lambda\cdot(t_l+t_a)}} \quad [11]$$

And,

$$t_x = \frac{\lambda\cdot t_a + ln\left(\frac{e^{-\lambda\cdot t_a}-e^{-\lambda\cdot(t_a+t_l)}}{1-e^{-\lambda\cdot(t_a+t_l)}}\right)}{\lambda} \quad [12]$$

Substituting the values from Equation [11] in Expression [9] and simplifying results in the following expression:

$$A_T^{*n\ cycles}(a) = n\cdot\frac{\left(1-e^{-\lambda\cdot t_a}\right)\cdot\left(1-e^{-\lambda\cdot t_l}\right)}{\left(1-e^{-\lambda\cdot(t_a+t_l)}\right)} \quad [13]$$

Furthermore, it holds that:

$$T = n\cdot(t_l+t_a)+t_x \quad [14]$$

Whereby, if we substitute [12] in [13], we are left with:

$$T = n\cdot(t_l+t_a)+\frac{\lambda\cdot t_a + ln\left(\frac{e^{-\lambda\cdot t_a}-e^{-\lambda\cdot(t_a+t_l)}}{1-e^{-\lambda\cdot(t_a+t_l)}}\right)}{\lambda} \quad [15]$$

2.1.2 Procedure using an algorithm

An algorithm created using Matlab was used to determine, for a given experiment time T and a number of cycles *n*, the value of $A_T^{*n\ cycles}(a)$, and the parameters $t_a,\ t_l, t_x$, corresponding to the number of cycles that would maximize $A_T^{*n\ cycles}(a)$. Figure 5 shows the flow chart for this algorithm.

Below is an example of the results of this algorithm for the total experiment time T=200s and n=7 (Table 1).

Table 1. Example of the results of the algorithm

T [s]	n	$A_T^{*n\ cycles}(a)$
200	7	4.159974

It was found that for smaller n values, $A_T^{*n\ cycles}(a)$ is lower. This effect was observed for all experiment times. As *n* increases, $A_T^{*n\ cycles}(a)$ increases until reaching an asymptote for n= ∞.

 ISBN: 978-84-614-6167-7

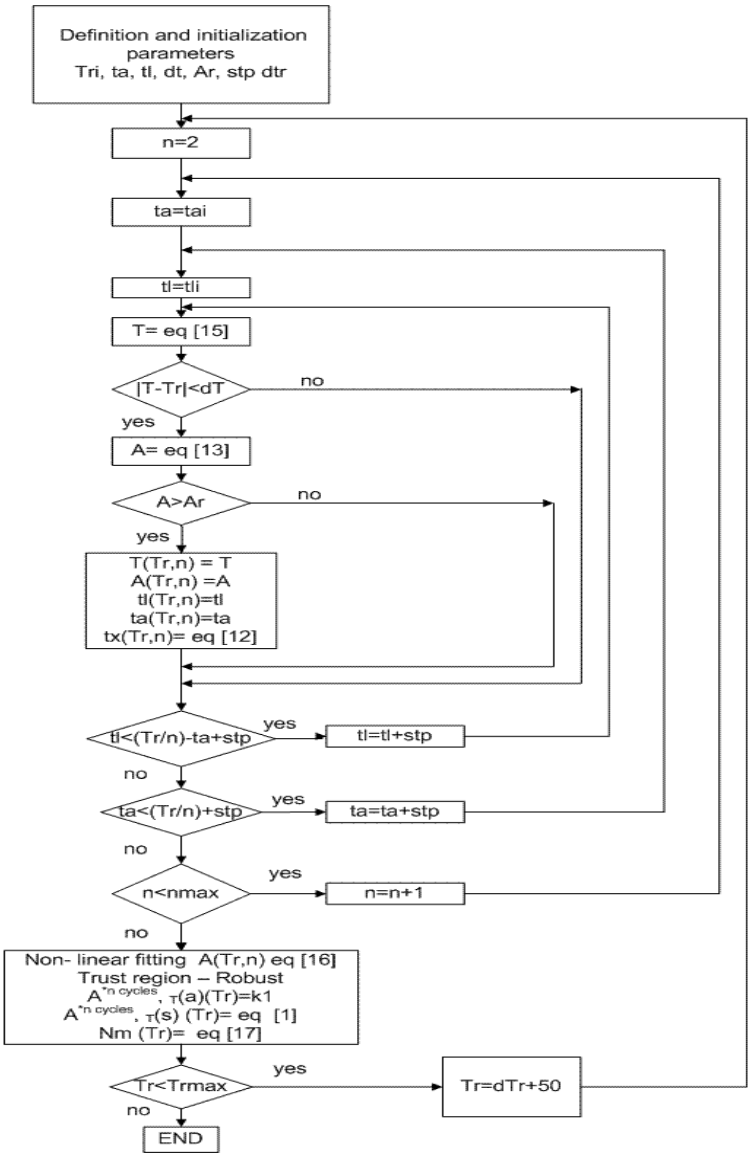


Figure 5. Flow chart.

Subsequently, the results for $A_T^{*n\ cycles}(a) - n$ for T=200 were adjusted to the following curve:

$$A_T^{*n\ cycles}(a) = k1 \cdot (1 - e^{-k2 \cdot n}) \qquad [16]$$

where k1 and k2 are the adjustment coefficients. k1 is the value of the asymptote, and represents the maximum value that could be obtained with asymmetrical cyclic activation.

 ISBN: 978-84-614-6167-7

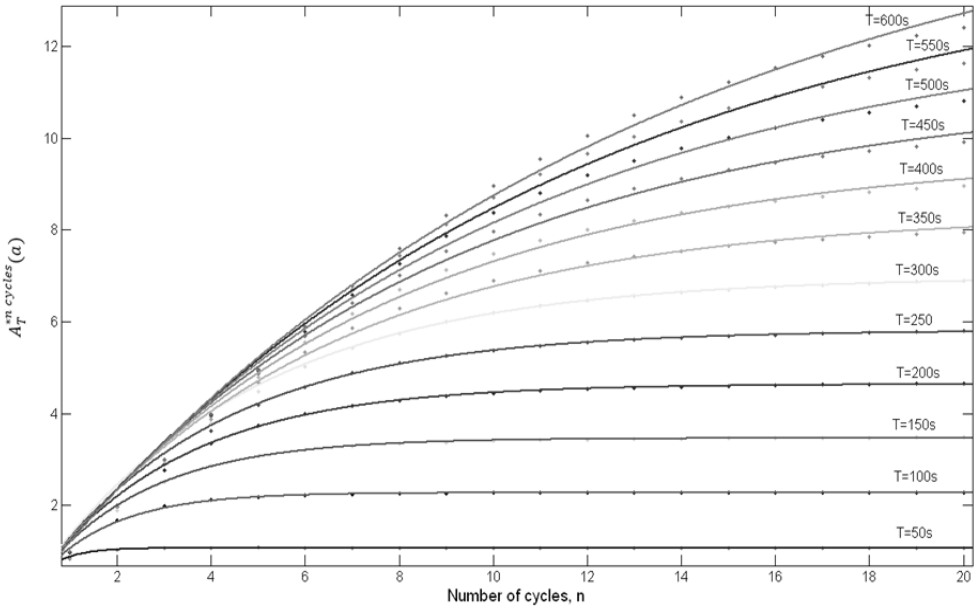


Figure 6. curves for the experiment times 50 to 600s.

The results of the above example are as follows (with 95% confidence bounds):

k1= 4.646 (4.623, 4.669)

k2= 0.3225 (0.3152, 0.3298)

Goodness of fit:

SSE: 0.01992

R_square:0.999

Adjusted R-square: 0.9989

RMSE: 0.03326

This curve is shown in Figure 6, alongside the curves for the experiment times 50 to 600s.

3 Comparison of the symmetrical and asymmetrical cycle processes

This section compares the maximum decay area of the symmetrical activation method, $A_{T,max}^{*n\ cycles}(s)$with that of the asymmetrical method, $A_{T,max}^{*n\ cycles}(a)$, as well as the two methods' activation parameters. Figure 6 shows the number of cycles needed for the asymmetrical cycle to surpass the symmetrical cycle, a cross-over point that we have called n_{lim}. This value was calculated with the equation:

$$n_{lim} = \frac{1}{k_2} \cdot \ln\left(1 - \frac{0.0199 \cdot T - 0.0221}{k_1}\right) \qquad [17]$$

 ISBN: 978-84-614-6167-7

These values are shown in Table 2. This means that, for example, for experiment time T=300, the maximum area that could be obtained by the asymmetrical method is 6.98, compared to 5.9479 for the symmetrical method, which constitutes at 17.4% improvement. Nine cycles are needed for the asymmetrical method to surpass the symmetrical method, with t_l=16.71, t_a=16.43 and t_x=1.

Table 2. Results

T [s]	Symmetrical cycle			Asymmetrical cycle					% improvement
	$A_{T,max}^{*n\ cycles}(s)$	τ	n	$A_{T,max}^{*n\ cycles}(a)$	t_l	t_a	t_x	n_{lim}	
50	0.9729	12.5	2	1.075	12.47	11.3	2.558	2	10.5
100	1.9679	12.5	4	2.28	16.72	16.04	1.82	3	15.9
150	2.9629	12.5	6	3.468	14.93	14.66	2.149	5	17.0
200	3.9579	12.5	8	4.646	16.61	16.43	1.859	6	17.4
250	4.9529	12.5	10	5.816	15.65	15.36	2.017	8	17.4
300	5.9479	12.5	12	6.98	16.71	16.43	1.838	9	17.4
350	6.9429	12.5	14	8.353	16.02	15.63	1.948	11	20.3
400	7.9379	13.33	15	9.69	15.5	15.12	2.039	13	22.1
450	8.9329	13.24	17	11.1	16.04	15.97	1.96	14	24.3
500	9.9279	13.16	19	12.61	16.61	16.6	1.866	15	27.0
550	10.9229	13.1	21	14.24	16.12	16.12	1.949	17	30.4
600	11.9179	13.04	23	16.01	15.16	14.74	2.098	20	34.3

Figure 7 shows the variation of $A_{T,max}^{*n\ cycles}(a)$and $A_{T,max}^{*n\ cycles}(s)$ with experiment time. It can be seen how the improvement increases alongside experiment time.

 ISBN: 978-84-614-6167-7

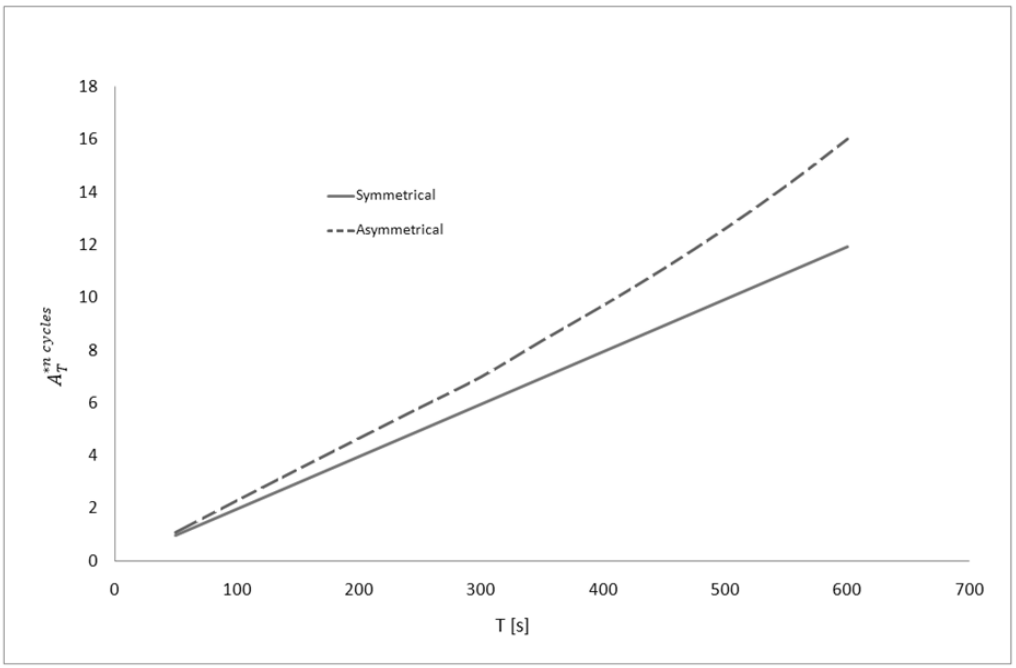


Figure 7. Comparison between $A_{T,max}^{*n\ cycles}(s)$ and $A_{T,max}^{*n\ cycles}(a)$ for different T.

4 Results and discussion

This paper discusses a new step forward in neutron activation processes for determining fluorite grade in samples of fluorspar concentrate.

This new procedure can be used to more precisely determine fluorite grades, though a neutron activation process with a 1Ci Americium-Beryllium source, which has been shown to be inadequate for single-cycle activation processes.

Better results can be obtained through the use of cyclic activation. For total experiment times of up to 200s, both symmetrical cyclic activation and the method proposed in this paper provide similar results, though the latter is a slight improvement over the former. However, for longer experiment times, for example around 600s, the asymmetrical method constitutes almost a 35% improvement over the conventional method. The asymmetrical process's improvement with respect to the symmetrical process tends to increase alongside the total experiment time.

Any system that is set up for symmetrical cyclic activation can easily accommodate the asymmetrical method by simply modifying the software used for transferring the samples.

Unlike the symmetrical method, the asymmetrical method's activation and counting times are not equal, although the difference between them is never greater than one second.

 ISBN: 978-84-614-6167-7

The total duration of an activation-counting cycle is considerably higher in the asymmetrical process than in the symmetrical process.

The number of cycles n_{lim} in the asymmetrical process is less than or equal to that of the symmetrical process.

Time t_x (the extra activation time in the first cycle of the asymmetrical process) is around two seconds, regardless of the experiment time.

5 Acknowledgements

The authors would like to thank Mineral Products and Derivatives Company SA (Minersa), the Government of the Principality of Asturias, and the research grant scheme of the University of Oviedo for their collaboration and financial support throughout this study.

6 References

[1] R.H. Filby. *Isotopic and nuclear analytical techniques in biological systems: A critical study. Neutron activation analysis*. Pure and Applied Chemistry. **67** (11) (1995) 1929-1941.

[2] Y. Maki, T. Nojiri, B.A. Masilungan. *The determination of fluorine by cyclic activation analysis method using 241Am-Be neutron source*. Radioisotopes **23** (1974) 149-154.

[3] S.E.B. Nonie, K. Randle, A.G. Blackband. *Development of a computer-based system for fast neutron cyclic activation analysis*. Journal of Radioanaltical and Nuclear Chemistry. **167** (1993) 89-95.

[4] I.W. Croudace, K. Randle. *A rapid and non-destructive method of fluorine determination using fast-neutron activation analysis*. Chemical Geology. **67** (1988) 165-170.

[5] I.W. Croudace. *Fluorine abundances of twenty nine geological and other reference samples using fast-neutron activation analysis*. Geostandards Newsletter. **127** (2) (1993) 127-218.

[6] A. Chatt, K.N. DeSilva, J.Holzbecher, D.C. Stuart, R.E. Tout, D.E. Ryan. *Cyclic neutron activation analysis of biological and metallurgical samples.* Canadian Journal of Chemistry. **59** (1981) 1660-1664.

[7] M.A. Rey-Ronco., 2007. *Desarrollo de un método rápido basado en técnicas de activación neutrónica para la determinación del contenido en flúor de muestras de mineral de fluorita*. Universidad de Oviedo. Doctoral thesis. Avalaible from <http://www.tesisenred.net/TDR-0802107-133201/index_cs.html.>

[8] T. Alonso-Sánchez**,** M.A. Rey-Roncob**,** M.P. Castro-García. *A neutron activation technique for the analysis for fluorine in fluorspar samples*. International Journal of Mineral Processing. **94** (2010) 1-13.

 ISBN: 978-84-614-6167-7

[9] A. Sonzogni 2011. Nuclear Structure & Decay Data (NuDat). Available from <www.nndc.bnl.gov>.

[10] M. Herman. Experimental Nuclear Reaction Data (EXFOR / CSISRS), 2007. Available from <www.nndc.bnl.gov>.

[11] N.M. Spyrou. *Cyclic activation analysis. A review*. Journal of Radioanalytical Chemistry. **61** (1981) 211-242.

[12] W.W. Givens, W.R Mills, R.L. Caldwell. *Cyclic activation analysis*. Nuclear Instruments and Methods. **80** (1970) 95-103.

[13] X. Hou. *Cyclic activation analysis*. Encyclopedia of Analytical Chemistry. R.A.Meyers. (2000) 12447-12459.

[14] M. A. Rey-Ronco, T. Alonso-Sánchez, M. P. Castro-García. *Mathematical study to improve the sensitivity in the neutron activation analysis of fluorspar*. Journal of Mathematical Chemistry. **48** (2010) 165-174.

 ISBN: 978-84-614-6167-7

Proceedings of the 11th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE2011
26-29 June 2011.

On the Numerical Solution of Fractional Schrödinger Differential Equations

Allaberen Ashyralyev[1,2] and Betul Hicdurmaz[3]
[1] *Department of Mathematics, Fatih University*
[2]*International Turkmen-Turkish University, Ashgabat, Turkmenistan and*
[3] *Department of Mathematics, Gebze Institute of Technology*

emails: aashyr@fatih.edu.tr, betulhicdurmaz@gmail.com

Abstract

The stable difference scheme of multidimensional fractional Schrödinger differential equation is presented. Stability estimates for the difference scheme of the fractional Schrödinger equation is obtained. A procedure of modified Gauss elimination method is used for solving this difference scheme in the case of one dimensional fractional Schrödinger differential equation with dependent coefficients and Neumann condition.

Key words: difference scheme, fractional Schrödinger differential equation
MSC2000: 65M06, 35R11

1. Introduction

It is known that various problems in quantum mechanics and other areas of physics lead to partial fractional differential equations. Methods of solutions of problems for fractional differential equations have been studied by many researchers (see [1]-[5], [11] and [15]). This type of equations can be solved by classical methods. However these classical methods can be used only in the case when the differential equation has constant coefficients. It is well known that the most useful method for solving partial differential equations with dependent coefficients in t and in space variables is finite difference method.

 ISBN: 978-84-614-6167-7

The role played by stability inequalities in the study of Schrödinger equation is well-known (see [6]-[8] and [13]). In the present paper, the mixed boundary value problem for the multidimensional fractional Schrödinger equation

$$\begin{cases} i\frac{\partial u}{\partial t}+\int\limits_0^t \alpha(s)D_s^{1/2}u(s,x)ds-\sum\limits_{r=1}^m (a_r(x)u_{x_r})_{x_r}+\delta u(t,x)=f(t,x), \\ \\ 0<t<1, x=(x_1,\cdots,x_m)\in\Omega, \\ \\ u(0,x)=0,\ x\in\overline{\Omega}, \\ \\ \left.\frac{\partial u(t,x)}{\partial \vec{n}}\right|_{x\in S}=0,\ 0\le t\le 1,\ x\in S, S=\partial\overline{\Omega} \end{cases} \tag{1}$$

is considered under the condition

$$\left|\alpha(y)\right|\le \frac{M_1}{y^{1/2}}.$$

Here $D_t^{1/2}=D_{0+}^{1/2}$ is the standard Riemann-Lioville's derivative of order ½ and $\vec{n}$ denotes the unit normal vector to the boundary $\partial\Omega$. Ω is the unit open cube in m-dimensional Euclidean space R^m : $\Omega=\{x=(x_1,...,x_m):\ 0<x_j<1,\ 1\le j\le m\}$ with boundary S, $\overline{\Omega}=\Omega\cup S$. $a_r(x),\ (x\in\Omega)$ and $f(t,x)\ (t\in(0,1),\ x\in\Omega)$ are given smooth functions and $a_r(x)\ge a>0$.

The difference scheme of the problem is presented. The stability estimates for the solution of this difference scheme is established. A procedure of modified Gauss elimination method is used for solving this difference scheme in the case of one-dimensional fractional Schrödinger differential equation with Neumann condition and dependent coefficient in space variable.

2. The Difference Scheme and Stability Estimates

The discretization of problem (1) is carried out in two steps. In the first step, let us define the grid sets

$$\overline{\Omega}_h=\{x=x_r=(h_1r_1,\ldots,h_mr_m),\ r=(r_1,\ldots,r_m),$$

 ISBN: 978-84-614-6167-7

$$0 \le r_j \le N_j,\ h_j N_j = 1,\ j = 1,\ldots,m,\Big\},$$

$$\Omega_h = \overline{\Omega}_h \cap \Omega,\ S_h = \overline{\Omega}_h \cap S.$$

We introduce the Banach spaces $L_{2h} = L_2(\overline{\Omega}_h)$ and $W_{2h}^2 = W_2^2(\overline{\Omega}_h)$ of the grid functions $\varphi^h(x) = \{\varphi(h_1 r_1,..., h_m r_m)\}$ defined on $\overline{\Omega}_h$, equipped with the norms

$$\|\varphi^h\|_{L_{2h}} = \left(\sum_{x\in\overline{\Omega}_h} \left|\varphi^h(x)\right|^2 h_1 \cdots h_m\right)^{1/2}$$

and

$$\left\|\varphi^h\right\|_{W_{2h}} = \left\|\varphi^h\right\|_{L_{2h}} + \left(\sum_{x\in\overline{\Omega}_h}\sum_{r=1}^{m}\left|\left(\varphi^h\right)_{x_r}\right|^2 h_1 \cdots h_m\right)^{1/2}$$

$$+\left(\sum_{x\in\overline{\Omega}_h}\sum_{r=1}^{m}\left|\left(\varphi^h\right)_{x_r \bar{x}_r, j_r}\right|^2 h_1 \cdots h_m\right)^{1/2}$$

respectively. To the differential operator A generated by problem (1), we assign the difference operator A_h^x by the formula

$$A_h^x u_x^h = -\sum_{r=1}^{m}\left(a_r(x) u_{\bar{x}_r}^h\right)_{x_r, j_r} + \delta u^h \tag{2}$$

acting in the space of grid functions $u^h(x)$, satisfying the conditions $D_h^n u^h(x) = 0$ for all $x \in S_h$. Here, D_h^n is the approximation of the operator $\frac{\partial}{\partial \vec{n}}$. It is known that A_h^x is a self-adjoint positive definite operator in $L_2(\overline{\Omega}_h)$. With the help of A_h^x, we arrive at the initial value problem

$$\begin{cases} \frac{du^h(t,x)}{dt} + \int_0^t D_s^{1/2} u^h(s,x)\alpha(s)ds + A_h^x u^h(t,x) = f^h(t,x),\ 0 < t < 1,\ x \in \overline{\Omega}_h, \\ u^h(0,x) = 0,\ x \in \overline{\Omega}_h. \end{cases} \tag{3}$$

In the second step, we replace problem (3) by the following first order of accuracy difference scheme.

 ISBN: 978-84-614-6167-7

$$\begin{cases} i\frac{u_k^h(x)-u_{k-1}^h(x)}{\tau}+A_h^x u_k^h(x)+\sum\limits_{l=1}^{k}\frac{1}{\sqrt{\pi}}\sum\limits_{m=1}^{l}\frac{\Gamma(l-m+1/2)(u_m^h(x)-u_{m-1}^h(x))}{(l-m)!}\alpha_l\tau^{1/2}=f_k^h(x), \\ f_k^h(x)=f(t_k,x),\ t_k=k\tau,\ 1\le k\le N,\ N\tau=1,\ x\in\overline{\Omega}_h, \\ u_0^h(x)=0,\ x\in\overline{\Omega}_h. \end{cases} \tag{4}$$

Theorem 2.1. Let τ and $|h|$ be sufficiently small numbers. Then, the solution of difference scheme (4) satisfy the following estimate:

$$\max_{1\le k\le N}\left\|u_k^h\right\|_{W_{2h}^2}+\max_{1\le k\le N}\left\|\frac{u_k^h-u_{k-1}^h}{\tau}\right\|_{L_{2h}}\le M_1\left[\left\|f_1^h\right\|_{L_{2h}}+\max_{1\le k\le N-1}\left\|\tau^{-1}(f_{k+1}^h-f_k^h)\right\|_{L_{2h}}\right].$$

Here M_1 does not depend on τ, h and $f_k^h, 1\le k\le N$.

The proof of Theorem 2.1 is based on the self-adjointness, positive definiteness and symmetry properties of the operator A^x defined by formula (2) and the following theorem on coercivity inequality for the solution of the elliptic problem in L_{2h}.

Theorem 2.2. For the solutions of the elliptic difference problem

$$\begin{cases} A_h^x u^h(x)=w^h(x),\ x\in\Omega_h, \\ u^h(x)=0,\ x\in S_h \end{cases}$$

the following coercivity inequality holds (see [20]):

$$\sum_{r=1}^{m}\left\|u_{x_r\bar{x}_r,j_r}^h\right\|_{L_{2h}}\le M\|w^h\|_{L_{2h}}.$$

3. Numerical Results

For the numerical result, the mixed problem

 ISBN: 978-84-614-6167-7

$$\begin{cases} i\frac{\partial u(t,x)}{\partial t}-\dfrac{\partial}{\partial x}\left(x\frac{\partial u(t,x)}{\partial x}\right)+u(t,x)+\int\limits_0^t D_s^{1/2}u(s,x)\frac{1}{\sqrt{s}}\,ds=f(t,x), \\ f(t,x)=\cos(\pi x)\left(2ti+t^2\pi^2x+t^2+\dfrac{4t^2}{3\sqrt{\pi}}\right)+\sin(\pi x)t^2\pi, \\ 0<t\le 1, 0<x<1; \\ u(0,x)=0,\ \ 0\le x\le 1, \\ u_x(t,0)=u_x(t,1)=0,\ 0\le t\le 1 \end{cases} \tag{5}$$

for one-dimensional Schrödinger differential equation is considered. Applying difference scheme (4), we obtain the following difference scheme

$$\begin{cases} i\frac{u_n^k-u_n^{k-1}}{\tau}-\frac{1}{h}\left(x_{n+1}\dfrac{u_{n+1}^k-u_n^k}{h}-x_n\dfrac{u_n^k-u_{n-1}^k}{h}\right)+u_n^k \\ +\sum\limits_{l=1}^{k-1}\frac{1}{\sqrt{\pi}}\sum\limits_{m=1}^{l}\frac{\Gamma(l-m-1/2+1)}{(l-m)!}\left(\frac{u_n^m-u_n^{m-1}}{\tau^{1/2}}\right)\frac{1}{\sqrt{l}}\tau^{1/2}=f(t_k,x_n), \\ f(t_k,x_n)=\cos(\pi x_n)\left(2t_ki+{t_k}^2\left(\pi^2x_n+1+\dfrac{4}{3\sqrt{\pi}}\right)\right)+\sin(\pi x_n){t_k}^2\pi, \\ 1\le k\le N, 1\le n\le M-1; \\ u_n^0=0,\ 0\le n\le M, \\ u_1^k-u_0^k=0,\ \ u_M^k-u_{M-1}^k=0,\ 0\le k\le N. \end{cases} \tag{6}$$

So, we have the $(N+1)\times(M+1)$ system of linear equations. This system can be written in the matrix form

 ISBN: 978-84-614-6167-7

$$\begin{cases} A\ U_{n+1} + B\ U_n + C\ U_{n-1} = D\varphi_n,\ 1 \le n \le M-1, \\ U_1 - U_0 = \vec{0},\quad U_M - U_{M-1} = \vec{0}, \end{cases}$$

where

$$\varphi_n = \begin{bmatrix} \varphi_n^0 \\ \varphi_n^1 \\ \cdots \\ \varphi_n^N \end{bmatrix}_{(N+1)\times 1}$$

$$A_n = \begin{bmatrix} 0 & 0 & \ldots & 0 & 0 \\ 0 & a_n & \ldots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \ldots & a_n & 0 \\ 0 & 0 & \ldots & 0 & a_n \end{bmatrix}_{N+1\times N+1},$$

$$B_n = \begin{bmatrix} 1 & 0 & 0 & \ldots & 0 & 0 \\ b_1 - i/\tau & (d_{2,2})_n & 0 & \ldots & 0 & 0 \\ b_2 & (d_{3,2})_n & (d_{3,3})_n & \ldots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ b_{N-1} & (d_{N,2})_n & (d_{N,3})_n & \ldots & (d_{N,N})_n & 0 \\ b_N & (d_{N+1,2})_n & (d_{N+1,3})_n & \ldots & (d_{N+1,N})_n & (d_{N+1,N+1})_n \end{bmatrix}_{N+1\times N+1}$$

and

 ISBN: 978-84-614-6167-7

$$C_n = \begin{bmatrix} 0 & 0 & \dots & 0 & 0 \\ 0 & c_n & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & c_n & 0 \\ 0 & 0 & \dots & 0 & c_n \end{bmatrix}_{N+1\times N+1},$$

$$D = I_{N+1},$$

$$U_s = \begin{bmatrix} U_s^0 \\ U_s^1 \\ \dots \\ U_s^{N-1} \\ U_s^N \end{bmatrix}_{N\times 1}, \quad s = n-1,\ n, n+1.$$

Here,

$$a_n = -\frac{1}{2h} - \frac{x_n}{h^2},$$

$$c_n = \frac{1}{2h} - \frac{x_n}{h^2},$$

$$b_i = -\sum_{m=1}^{i} \frac{\Gamma(m-1/2)}{\sqrt{\pi}(m-1)!\sqrt{m}},\ 1 \le i \le N,$$

$$(d_{i,j})_n = \begin{cases} \sum_{m=1}^{i} \frac{\Gamma(m-j+1/2)}{\sqrt{\pi}(m-j)!\sqrt{m}} - \sum_{m=1}^{i} \frac{\Gamma(m-j-1/2)}{\sqrt{\pi}(m-j-1)!\sqrt{m}} + \frac{i}{\tau} + \frac{2x_n}{h^2} + 1, & \text{where } i = j, \\ \sum_{m=1}^{i} \frac{\Gamma(m-j+1/2)}{\sqrt{\pi}(m-j)!\sqrt{m}} - \sum_{m=1}^{i} \frac{\Gamma(m-j-1/2)}{\sqrt{\pi}(m-j-1)!\sqrt{m}} - \frac{i}{\tau}, & \text{where } i = j+1, \\ \sum_{m=1}^{i} \frac{\Gamma(m-j+1/2)}{\sqrt{\pi}(m-j)!\sqrt{m}} - \sum_{m=1}^{i} \frac{\Gamma(m-j-1/2)}{\sqrt{\pi}(m-j-1)!\sqrt{m}}, & \text{elsewhere,} \end{cases}$$

 ISBN: 978-84-614-6167-7

for $2 \le i \le N+1,\ 2 \le j \le N+1.$

$$\varphi_n^k = \cos(\pi x_n)\left(2t_k i + {t_k}^2 \pi^2 x_n + {t_k}^2 + \frac{4{t_k}^2}{3\sqrt{\pi}}\right) + \sin(\pi x_n){t_k}^2 \pi,$$

for $1 \le k \le N,\ 1 \le n \le M-1.$

To solve this difference equation we have applied a procedure of modified Gauss elimination method. Hence, we seek a solution of the matrix equation in the following form

$$U_n = \alpha_{n+1} U_{n+1} + \beta_{n+1},\ n = M-1, \cdots, 2, 1, 0,\ U_M = (I - \alpha_M)^{-1} \beta_M, \tag{7}$$

where $\alpha_j\ (j = 1, \cdots, M-1)$ are $(N+1)\times(N+1)$ square matrices and $\beta_j\ (j = 1, \cdots, M-1)$ are $(N+1)\times 1$ column matrices which are defined by the following formulas

$$\alpha_{n+1} = -\left(B + C\alpha_n\right)^{-1} A, \tag{8}$$

$$\beta_{n+1} = \left(B + C\alpha_n\right)^{-1}\left(D\varphi_n - C\beta_n\right), n = 1, 2, 3, \cdots, M-1. \tag{9}$$

Now, we need to find α_1 and β_1. We can find them from $U_0 = \alpha_1 U_1 + \beta_1$ in the following form

$$\alpha_1 = \begin{bmatrix} 1 & 0 & 0 & \dots & 0 & 0 \\ 0 & 1 & 0 & \dots & 0 & 0 \\ 0 & 0 & 1 & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & 1 & 0 \\ 0 & 0 & 0 & \dots & 0 & 1 \end{bmatrix}_{N+1\times N+1}, \beta_1 = \begin{bmatrix} 0 \\ 0 \\ \dots \\ 0 \\ 0 \end{bmatrix}_{N+1\times 1}.$$

Thus, using formulas (7), (8) and (9) we can compute $U_n,\quad 1 \le n \le M-1.$

In order to get the solution of (7) we use MATLAB programs. The numerical solutions are recorded for different values of N, M and u_n^k represents the numerical solution of the difference scheme at (t_k, x_n). For their comparison, the errors are computed by

 ISBN: 978-84-614-6167-7

$$E = \max_{\substack{1\le k\le N\\ 1\le n\le M}} \left| u(t_k, x_n) - u_n^k \right|.$$

Table 1 and Table 2 give the error analysis between the exact solution and the solution derived by difference scheme for different values of N and M.

Table 1: Comparison of the errors of the first order of accuracy difference scheme with Neumann condition for $N = M = 15,$ $N = M = 30$ and $N = M = 60$.

Method	$N = M = 15$	$N = M = 30$	$N = M = 60$
E_M^N	0.1203	0.0596	0.0297

Table 2: Comparison of the errors of the first order of accuracy difference scheme with Neumann condition for $N = 10, M = 80,$ $N = 20, M = 80$ and $N = 40, M = 80$.

Method	$N = 10$ $M = 80$	$N = 20$ $M = 80$	$N = 40$ $M = 80$
E_M^N	0.1036	0.0457	0.0240

Now, the exact solution for problem (5) and the numerical solution obtained by using first order of accuracy difference scheme (6) for $N = M = 80$ are shown in figures (1) and (2) respectively as an example.

 ISBN: 978-84-614-6167-7

Figure 1: Exact solution of problem (5) for N=M=80.

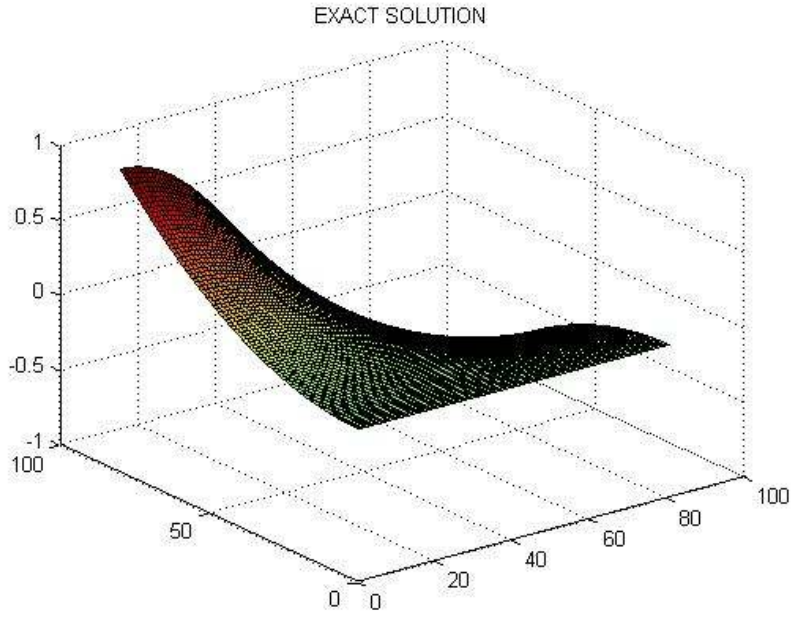
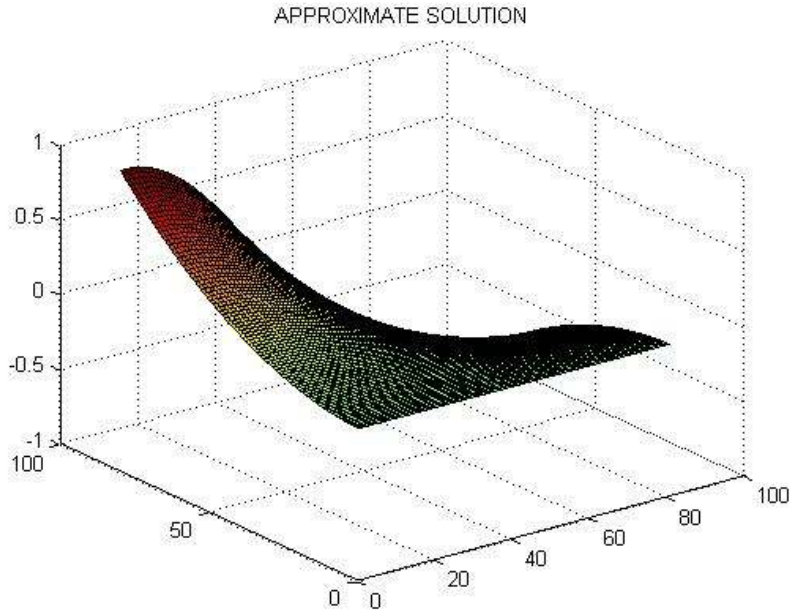


Figure 2: Approximate solution of problem (5) for N=M=80.



 ISBN: 978-84-614-6167-7

4. References

[1] F. B. ADDA, J. CRESSON, *Fractional differential equations and the Schrödinger equation*, Acta Math. Acad. Sci. Hung., **161** (2005) 323-345.

[2] A. ASHYRALYEV, *A note on fractional derivatives and fractional powers of operators,* Journal of Mathematical Analysis and Applications, **357**:1 (2009) 232-236.

[3] A. ASHYRALYEV, F. DAL, Z. PINAR, *On the numerical solution of fractional hyperbolic partial differential equations*, Mathematical Problems in Engineering, **2009** (2009) Article ID 730465.

[4] A. ASHYRALYEV, F. DAL, Z. PINAR, *A note on the fractional hyperbolic differential and difference equations, Applied* Mathematics and Computation, **217**:9 (2011) 4654-4664.

[5] A. ASHYRALYEV, A. SIRMA, *Nonlocal Boundary Value Problems for the Schrödinger Equation,* Computers and Mathematics with Applications*,* **55**:3 (2008) 392-407.

[6] A. ASHYRALYEV, A. SIRMA, *A note on the numerical solution of the semilinear Schrödinger equation,* Nonlinear Analysis: Theory, Methods and Applications, **71**:12 (2009) e2507-e2516.

[7] A. ASHYRALYEV, A. SIRMA, *Modified Crank-Nicholson Difference Schemes for Nonlocal Boundary Value Problem for the Schrödinger Equation,* Discrete Dynamics in Nature and Society, (2009) Article ID 584718.

[8] A. ASHYRALYEV, P. E. SOBOLEVSKII, *New Difference Schemes for Partial Differential Equations*, Operator Theory: Advances and Applications, **148**, Birkhauser, Basel, Boston, Berlin, 2004.

[9] D.G. GORDEZIANI, G.A. AVALISHVILI, *Time-nonlocal problems for Schrödinger type equations: I. Problems in abstract spaces*, Differential Equations*,* **41**:5 (2005) 703-711.

[10] H. HAN, J. JIN, X. WU, *A finite difference method for the one-dimensional Schrödinger equation on unbounded domain,* Computers and Mathematics with Applications*,* **50** (2005) 1345-1362.

[11] A. A. KILBAS, H. M. SRIVASTAVA, J. J. TRUJILLO, *Theory and applications of fractional differential equations*, North-Holland Math. Studies, Elsevier, Amsterdam, 2006.

[12] N. LASKIN, *Fractional Schrödinger equation*, Phys. Rev. E, **66** (2002) 056108.

 ISBN: 978-84-614-6167-7

[13]J. L. LAVOIE, T. J. OSLER, R. TREMBLAY, *Fractional derivatives and special functions,* SIAM Review, **18**:2 (1976) 240-268.

[14]M. NABER, *Time fractional Schrodinger equation,* Journal of Mathematical Physics, **45**:8, (2004) 18. Available at: http://arxiv.org/abs/math-ph/0410028.

[15]I. PODLUBNY, *Fractional Differential Equations*, Academic Press, New York, 1999.

[16]P. E. SOBOLEVSKII, *Difference Methods for the Approximate Solution of Differential Equations,* Izdat. Voronezh. Gosud. Univ., Voronezh, 1975. (Russian)

[17]V. E. TARASOV, *Fractional derivative as fractional power of derivative*, International Journal of Mathematics, **18** (2007) 281-299.

[18]P. XANTHOPOULOS, G.E. ZOURARIS, *A linearly implicit finite difference method for a Klein-Gordon-Schrödinger system modeling electron-ion plasma waves*, Discrete and Continuous Dynamical Systems Series, **B 10** (2008) 239-263.

 ISBN: 978-84-614-6167-7

Proceedings of the 11th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE2011
Benidorm, Spain, 26-30 June 2011

Nanoscale DGMOS modeling

M.Bella[1], S. Latreche[1], and S.Labiod[1]

[1]*Laboratoire Hyperfréquence & Semi-conducteur (LHS),*
Département d'Electronique, Faculté des Sciences de l'Ingénieur,
Université Mentouri, Constantine, 25000, Algérie

emails: latreche.saida@gmail.com, bella.mourad@yahoo.fr,

Abstract

The double gate MOSFET (Double Gate Metal Oxide Semiconductor Field Effect Transistor) devices have recently been of great interest; particularly for the investigation of sub-50nm field effect transistor [1, 2].
The quantum effects are dominant [3]. The analyse of their electrical performances versus the technological parameters is so important.
We present in this work, an efficient numerical model of DGMOS based on a quantum mechanical description of the carriers' concentration [4, 5]. It focuses on an accurate and self consistent treatment of quantum mechanical effects and provides a self-consistent solution of the Schrödinger and Poisson equations.
Moreover, physical phenomena such as carrier quantization (confinement carrier) or quasi- ballistic transport are considered [5], since Double Gate structure will be precisely used to design very integrated DGMOS (nanometric channel and ultra thin silicon films).
Time-independent, effective mass Schrödinger equation has been self-consistely solved with Poisson equation in fully-depleted (100) oriented silicon of DGMOS device with symmetrical gate and n channel type.
The computation of the potential and the carrier concentration has been conducted using a self numerical program based on a finite difference scheme with a uniform mesh. The numerical method considered to solve the PDE equation considered is Newton- Raphson one.
In this work, Double Gate architecture considered is presented on figure 1 with a channel length between 5 and 25nm.

 ISBN: 978-84-614-6167-7

The drain current in dependence on the bias voltages, canal length as well as the work function has been carefully investigated (figures 2 and 3). *Figure 2 represent I_{DS} (V_{GS}) characteristics for different channel lengths. It can be observed that when channel length (L_G) decrease, drain current (I_{DS}) increase; as well as lead to a decrease of the threshold voltage dramatically. In order to have an acceptable threshold voltage, we study the influence of the gate work function (figure 3) and conclude that for a nanometric length of channel, and in order to optimise the threshold voltage, the solution is to choose a gate metal with a high value of work function.*

In order to validate the obtained results, we compare our model with Sentaurus numerical simulation (ISE-TCAD software). We have considered a DG MOSFET with the following parameters: doping level $N_A = 10^{10} cm^{-3}$; doping source/drain $N^+ = 10^{20} cm^{-3}$; silicon thickness T_{Si}= 1.5 nm, oxide thickness T_{OX}=1.5 nm ; channel length L_G = 10 nm; source/drain length L_{SD}= 5 nm (figure 4).
The comparison between the modelled and simulated characteristics gives good agreement for V_{GS} down to 0.4 V. these devices are operated at V_{GS} lower than 0.4 V (linear regime), and therefore , the model is valid for the regimes of practical interest.

Key words: DG MOSFET, Self-consistent, Schrödinger equation, Poisson equation, quantum effects

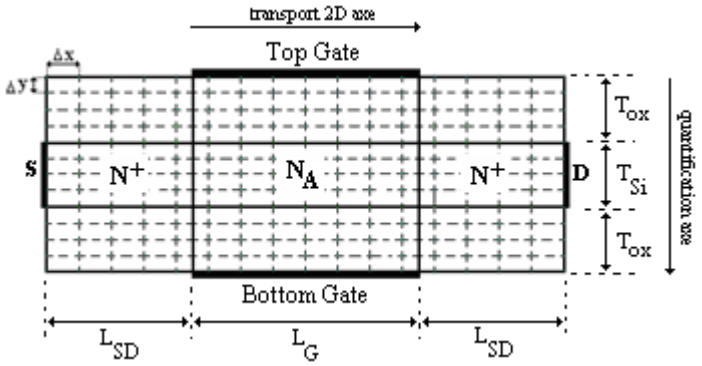


Figure 2: Structure and Mesh of double gate MOS transistor

 ISBN: 978-84-614-6167-7

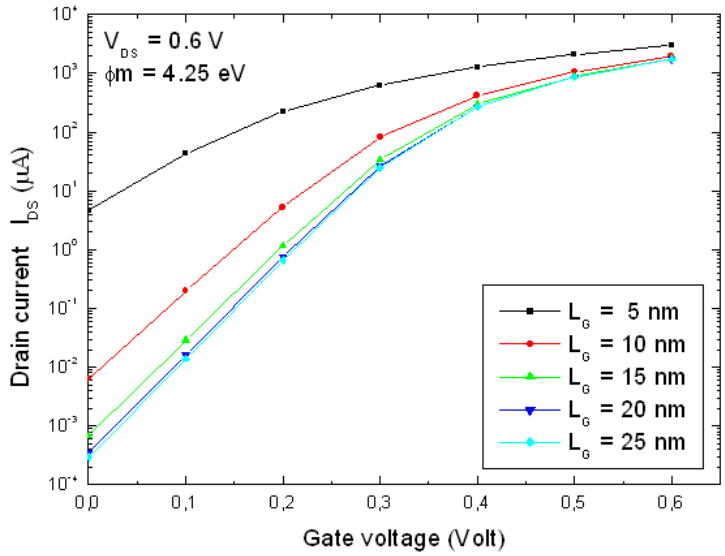


Figure 2: Influence of the channel length on the current drain

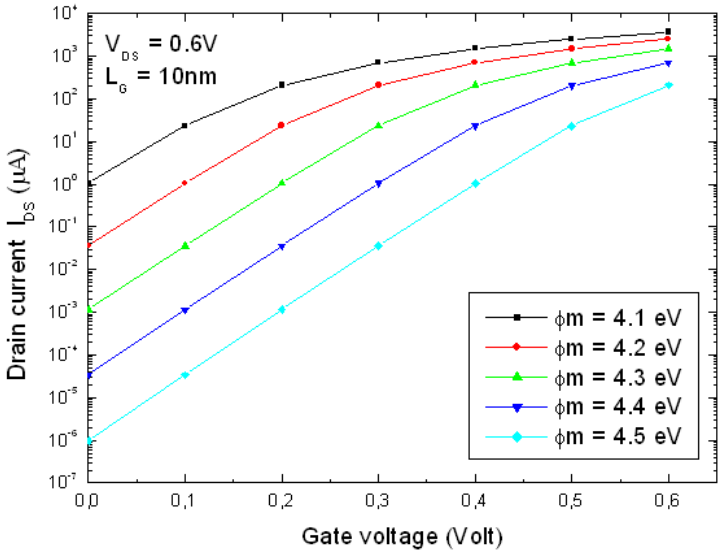


Figure 3: Influence of the gate work function on the current drain for different gate voltages

 ISBN: 978-84-614-6167-7

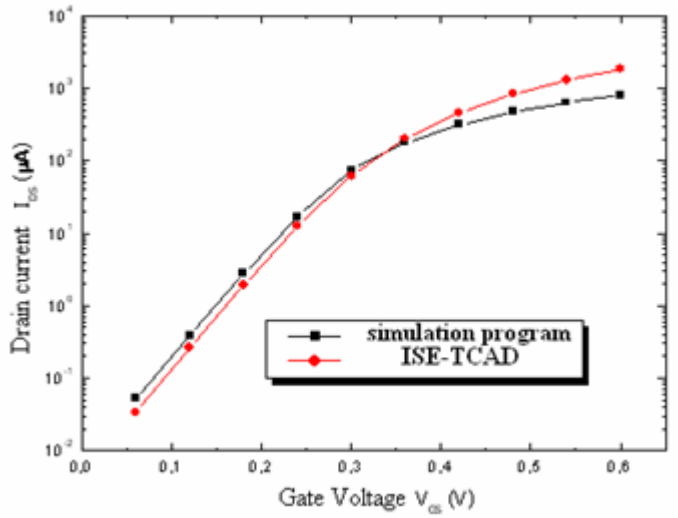


Figure4: Comparison of the output characteristics between simulation program and ISE-TCAD

References

[1] A.SVIDENKO AND M.P.ANANTIAM. *Role of scattering in nanotransistors.* IEEE. Trans.Election Devices, vol 50, pp 1459-1466, 2003.

[2] D.J.FRANK, R.H.DENNARD, E.NOWAK, P.M.SOLOMON, Y.TOUR AND M.S.P.WONG, *Device scaling limits of Si MOSFETs and their application dependencies.* Proc. IEEE, vol. 89, pp 259-288, 2001.

[3] J.WANG AND M.LURDSTION. *Does source- to- drain tunnelling limit the ultimate scaling of MOSFET*, in IEDM Tech. Dig.,2002.

[4] Z.REN, R.VENUGOPAL, S.GOASGUEN, MEMBRE, IEEE, S.DATTA, FELLOW, IEEE, AND M.S.LUNDSTROM, FELLOW, IEEE. *nano MOS 2.5: A Two-Dimensional Simulator for Quantum transport in Double Gate MOSFETs.* Proc. IEEE, vol. 50, NO. 9, September 2003.

[5] H.IWATA, T.MATSUDA AND T.OHZONE. *Multiband simulation of quantum transport in nanoscale double-gate MOSFETs.* Solid-State electronics, 53: 1130-1134, 2009.

 ISBN: 978-84-614-6167-7

Proceedings of the 11th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE2011
Benidorm, Spain, 26-30 June 2011

Large Scale Calculations with the deMon2k code

Patrizia Calaminici

Departamento de Química, CINVESTAV, Avenida Instituto Politecnico Nacional, 2508, A.P. 14-740, Mexico D.F. 07000, Mexico

email: pcalamin@cinvestav.mx

Abstract

State of-the-art density functional theory calculations of large systems such as fullerenes containing several hundred atoms and mordenite type zeolite clusters models containing more than 400 atoms will be presented. The calculations have been performed using the linear combination of Gaussian-type orbitals density functional theory (LCGTO-DFT) approach as implemented in the deMon2k code. For the calculations all-electron basis sets in combination with local and generalized gradient approximations were employed. All fullerenes structures were fully optimized without symmetry constrains. For the mordenite models, in order to constrain the structure of the cluster to that of the solid, the coordinates of the terminal hydrogen atoms, positioned along the Si-O bonds have been fixed during the optimization procedure whereas and all other atomic coordinates have been relaxed. For both cluster models applications, the analysis of the obtained structure and of the calculated relevant energy properties will presented and discussed. The obtained results are compared with available experimental and theoretical data. These studies demonstrate the capability of DFT calculations for energy and structure computations of large cluster structures.

 ISBN: 978-84-614-6167-7

Key words: deMon2k, Density Functional Theory, Giant Fullerenes, Mordenite Type Zeolites, Binding Energy, Confinement Effects.
MSC2000: AMS Codes (optional)

1 Introduction

1.1Fullerenes

Fullerenes are carbon clusters formed by the closing of a graphitic sheet with the needed curvature supplied by intersecting, among a given number of graphitic hexagons, of twelve pentagons [1,2]. These carbon aggregates have been experimentally known for more than twenty years [3] and, consequently, a large number of works, experimental as well as theoretical has focused on this subject. One main reason for the large interest in the study of fullerenes is certainly to be found in their particularly appealing geometrical form. The best known fullerene is the so-called buckminsterfullerene that contains sixty carbon atoms (C_{60}) and it is composed of twelve pentagonal carbon rings located around the vertices of an icosahedron and twenty hexagonal carbon rings at the centers of icosahedral faces [3]. Larger fullerenes that have an icosahedral symmetry can be constructed, as well. These clusters, known as *giant fullerenes*, can be thought of as cut-out pieces of graphene plane that are folded into final shape (icosahedron). This kind of procedure generates twelve pentagonal carbon rings situated around vertices of an icosahedron, while all other carbon rings are hexagonal. Giant or large fullerenes have been the subject of different theoretical studies in the last years. These studies were focused either to understand if the shape of these clusters is spherical or faceted, to calculate their response properties or to test new algorithms developed for the investigation of large systems. So far, all previous first-principle type theoretical studies of large fullerenes have been performed at the Hartree-Fock level of theory using symmetry restrictions and relative small basis sets or analytic density-functional theory. Recently we presented the first all-electron density functional theory based study on the large fullerenes C_{180}, C_{240}, C_{320} and C_{540} [4]. The originality of our work is that the structures of these clusters were for the first time in the literature fully optimized without any symmetry constrain. This work provides important insights about the structural changes, the evolution of the bond length and the binding energy with increasing the fullerene size.

 ISBN: 978-84-614-6167-7

1.1Mordenites Models

Diffusion, adsorption and reactivity of molecules within micro- or mesoporous materials are specifically related with the physico-chemical properties of the material structures. The interactions of the host molecules with the material surfaces depend on the volume, shape and topology of the cavities, which generate particular organisations of these molecules [5]. The inter-relationship between the porous materials and the host molecules has been referred to as “confinement” and attributed a large role in the selectivity and catalytic activity of zeolite materials, in particular, in acid-catalysed reactions [6-8]. Computer modelling based on *ab-initio* techniques have been used in the recent years to provide a better understanding of the physico-chemical processes involved in the protonation of reactants. They lead to somewhat different conclusions, according to the methods applied and the reactions studied. Mordenites are natural and synthetic zeolites with Si/Al ratios of 4.3-6.0 in the former case and 5.0 to 12.0 in the latter. Synthetic mordenites are used for acidic catalysis. Mordenites catalysts are synthesized in the Na-form followed by a mild treatment with NH_4Cl which leads to H-exchanged forms. The mordenite structure can be described as composed of edge-sharing five-membered (5-m) rings of tetrahedra forming chains along the *c* crystallographic axis. Their architectures comprise mono-directional large accessible 12-m ring channels of TO_4 tetrahedra where T stands for either Si or Al and small 8-m ring channels, which are interconnected through 8-m ring tubes. The topological symmetry of mordenites is orthorhombic with space group Cmcm having in the unit cell four symmetrically independent tetrahedral sites, usually called T1, T2, T3 and T4, being T1 and T2 sites connecting four different rings while the T3 and T4 sites constitute the 4-m rings of the zeolite framework. A lot of work has been performed over the years to solve several problems about the real symmetry of the solid due to the presence of Al tetrahedra and extra-framework cations, framework defects, etc... . Having different Si/Al ratios, natural and synthetic zeolites have thus different Al distribution patterns and slightly different T-O bond lengths. Our strategy, in this study, i.e. the determination of the structure and intrinsic properties of the catalyst, has been to use clusters containing 120 tetrahedra, sufficiently large to enclose the main 12m-rings and the side pocket 8m-rings. These models include thus 2 unit-cells along *a* and *b* whereas the 2x*c* dimension has been cut at the middle of the second 12m-ring channel. In our study the original models have been cut from a solid with the adequate Al distribution and terminated with hydrogens. Exhaustive studies of nitriles adsorption in mordenite zeolites led to the conclusion that small nitriles, such as acetonitrile, are able to penetrate into small cavities and are more

 ISBN: 978-84-614-6167-7

strongly interacting with their Br?nsted sites. Results obtained with modern computational methods demonstrate that the concept of confinement is now widely invoked but not yet clearly quantified, differing with the level of accuracy and approximation. Such a complicated study necessitates first to set up an accurate and also efficient methodology and answer the fundamental questions: (i) is the BrØnsted acid strength different in small and large cavities and what is the role of the local site geometry on this property?; (ii)
how dependent on the cavity size are the stabilization of the guest molecule through long range dispersion and electrostatic polarization due to the solid framework? Recently, we have focused our attention on these questions, strting with the study of the structure and the intrinsic properties of mordenite (MOR) cationic sites, in particular sodium binding energies and acid strengths. For this purpose we proposed firstly a new methodological approach [9], based on very large model clusters, which can be eventually embedded in a classical environment, as an alternative to the periodic representation of a zeolitic solid, when the experimental Si/Al ratio is not very large, i.e. when the solid contains an Al distribution which can hardly be represented as periodic. More recently we performed an *ab-initio* DFT based study on the interaction of single guest CO and CH_3CN molecules with hydroxyls located in the main channels and in the side-pockets [10]. Since zeolite micropores induce substantial van der Waals interactions with guest molecules, our results are analyzed with respect to these long-range effects and in relation with the most recent experimental conclusions.

1 Computational Details

All calculations were performed using the density functional theory (DFT) deMon2k program [11]. The exchange-correlation potential was numerically integrated on an adaptive grid [12]. The grid accuracy was set to 10^{-5} in all calculations. The Coulomb energy was calculated by the variational fitting procedure proposed by Dunlap, Connolly and Sabin [13,14]. The structure optimizations were performed in the local density approximation (LDA) with the exchange-correlation functional of Vosko, Wilk and Nusair (VWN) [15]. DFT optimized double zeta plus valence polarization (DZVP) all-electron basis sets optimized for local functionals [16] were employed. For the structure optimization a quasi-Newton method in internal redundant coordinates with analytic energy gradients was used [17]. The geometry optimizations were performed using the parallel version of the deMon2k code [18]. The convergence was based on the Cartesian gradient and displacement vectors with a threshold of 10^{-4} and 10^{-3} a.u., respectively. The binding energies were calculated with the functional proposed

 ISBN: 978-84-614-6167-7

by Perdew, Burke and Ernzerhof functional (PBEPBE) [19]. For the geometry optimization of the studied structures the parallel version of the deMon2k code [20,21] was used. The calculations were performed on 8 or 6 Intel Xeon CPUs with 2.4 Ghz. These nodes were connected by a Myrinet network.

1Results and Discussion

1.1Large fullerenes

In Figure 1 the all-electron optimized structures of C_{180}, C_{240}, C_{320} and C_{540} are depicted. As it can be seen from this figure our first-principle based calculations predict that the optimized structures of the largest fullerenes, C_{240}, C_{320} and C_{540} are faceted. Moreover, even for the smallest carbon cluster here studied, C_{180}, there is clear evidence that the faceted shape is preferred over a spherical shape if first-principle all-electron optimization without any symmetry restriction is performed (see Fig. 1).

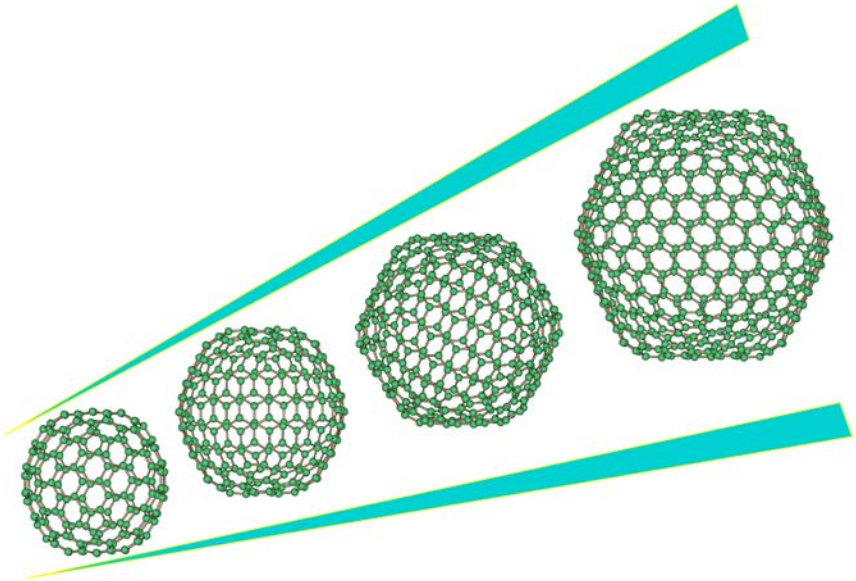


Figure 1: Optimized structures of the large fullerenes C_{180}, C_{240}, C_{320} and C_{540} .

In order to gain more insight into the structural changes of these systems as the number of carbon atoms increases we performed a detailed analysis of the bond

 ISBN: 978-84-614-6167-7

length evolution which has shown two important facts. First the number of different lengths increases and, second, the longest bond shortness with the increasing of the number of carbon atoms. With the aim to guide future desirable experiments on large fullerenes and to gain more informations about their stability we have also explored the behaviour of the binding energy of the studied fullerenes with the increase of the cluster size. The results of the obtained binding energy show that this property increases monotonically as the number of carbon atoms increases meaning that the large fullerenes here studied from C_{180} to C_{540} became more and more stable.

3.2 Mordenite type zeolite models

Figure 2 illustrated the structure of the mordenite models we have investigated. The four tetrahedral sites present in this zeolite (T1, T2, T3 and T4) are indicated by arrows (see Fig. 2).

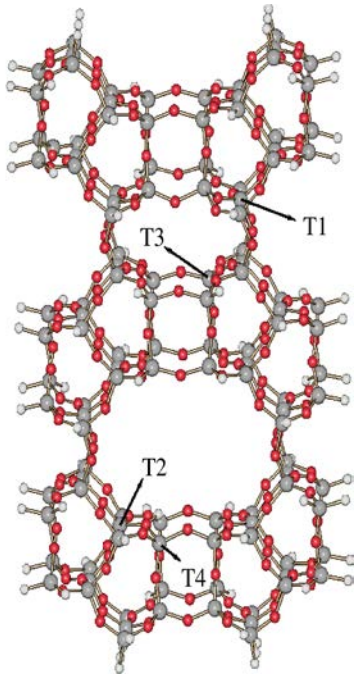


Figure 1: Mordenite model containing 120 tetrahedra. Atom legend: H atoms are in white, Si atoms are in gray, O atoms are in red. The four tetrahedral sites present in this zeolite (T1, T2, T3 and T4) are indicated by arrows.

 ISBN: 978-84-614-6167-7

From the analysis of the obtained results on the study of the mordenite models, the following conclusions can be drawn: (i) a good agreement with experimental bond length and bond angle data reported for synthetic Na-mordenites has been obtained; (ii) the binding energies the binding energies of the Na cation follows the same ordering as the populations of Al T sites, as derived from X-Ray measurements on synthetic Na-mordenites; (iii) the compensation cation at the T1 site is found more stable in the side pocket than in the main channel; (iv) the calculated proton affinities at T1, T3 and T4 sites are equivalent, indicating that these BrØnsted sites have similar acid strengths. In order to evaluate any difference in local acidity, the presence of the associated base (CO, CH_3CN, NH_3, etc...) has to be taken into account. Our recent work shows that, at zero coverage (1 molecule of CO or CH_3CN per site, per cavity), the host - guest energetic interactions do not depend on the cavity volumes, leading to similar adsorption energies for main channel and side-pocket sites. This unexpected result is in agreement with recent experimental studies which show that confinement is more related with higher concentration of host molecules in small rather than large cavities at high pressure.

1Acknowledgments

Support from the French-Mexican Post Graduate Program, CONACYT (project n. 130726), and from the ICyTDF (project n. PICCO-10-47) is gratefully acknowledged.

References

[1]K.M. KADISH, R.S. RUOFF, IN FULLERENES: CHEMISTRY, PHYSICS, AND TECHNOLOGY, JOHN WILEY & SONS INC., NEW YORK, 2007

[2]W. ANDREONI, IN THE PHYSICS OF FULLERENE-BASED AND FULLERENE-RELATED MATERIALS, KLUWER ACADEMIC PUBLISHERS, DORDRECHT, THE NETHERLANDS, 2007

[3]H.W. KROTO, J.R. HEATH, S.C. O'BRIEN, R.F. CURL, R.E. SMALLEY, NATURE (LONDON) **318** 162 (1985)

[4]P. CALAMINICI, G. GEUDTNER, A.M. KÖSTER, *J. OF CHEM. THEORY AND COMP.* **5** 29 (2009)

[5]E.G. DEROUANE, J.M. ANDRE, A.A. LUCAS, *J. CATAL.* **110** 58 (1988)

 ISBN: 978-84-614-6167-7

[6]F. THIBAULT-STARZYK, A. TRAVERT, J.C. SAUSSEY, *TOP. CATAL.* **6** 11 (1998)
[7]R. ANQUETIL, J.C. SAUSSEY, J.C. LAVALLEY, *PHYS. CHEM. CHEM. PHYS.* **1** 555 (1999)
[8]K. SMIRNOV, F. THIBAULT-STARZYK, *J. PHYS. CHEM. B* **103** 8595 (1999)
[9]V.D. DOMINGUEZ-SORIA, P. CALAMINICI, A. GOURSOT, J. CHEM. PHYS. **127** 154710 (2007)
[10] V.D. DOMINGUEZ-SORIA, P. CALAMINICI, A. GOURSOT, J. PHYS. CHEM. C **115** 6508 (2011)
[11] KÖSTER A.M., CALAMINICI P., CASIDA M.E., FLORES-MORENO R., GEUDTNER G., GOURSOT A., HEINE T., IPATOV A., JANETZKO F., MARTIN DEL CAMPO J., PATCHKOVSKII S., REVELES J.U., SALAHUB D.R., VELA A., THE DEMON DEVELOPERS CINVESTAV, MEXICO, (2006)
[12] A.M. KÖSTER, R. FLORES--MORENO, J.U. REVELES, J. CHEM. PHYS. **121** 681 (2004)
[13] B.I. DUNLAP, J.W.D. CONNOLLY, J.R. SABIN, J. CHEM. PHYS. **71** 4993 (1979)
[14] W. MINTMIRE, B.I. DUNLAP, PHYS. REV. A **25** 88 (1982)
[15] S.H. VOSKO, L. WILK, M. NUSAIR, CAN. J. PHYS. **58** 1200 (1980)
[16] N. GODBOUT, D.R. SALAHUB, J. ANDZELM, E. WIMMER, CAN. J. PHYS. **70** 560 (1992)
[17] J.U. REVELES, A.M. KÖSTER, J. COMPUT. CHEM. **25** 1109 (2004)
[18] G. GEUDTNER, F. JANETZKO, A. M. KÖSTER, A. VELA, P. CALAMINICI, J. COMP. CHEM., **27**, 483 (2006)
[19] J.P. PERDEW, K. BURKE, M. ERNZERHOF, PHYS. REV. LETT. **77** 3865 (1996)
[20] P. CALAMINICI, V.D. DOMINGUEZ-SORIA, G. GEUDTNER, E. HERNANDEZ-MARIN, A.M. KÖSTER, *THEOR. CHEM. ACC.* **115**, 221 (2006)
[21] G. GEUDTNER, F. JANETZKO, A.M. KÖSTER, A. VELA, P. CALAMINICI, *J. COMP. CHEM.* **27**, 483 (2006)

 ISBN: 978-84-614-6167-7

Proceedings of the 11th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE2011
Alicante, Spain, 26-30 June 2011

Estimation and analysis of lead times of metallic components in the aerospace industry through a Cox model

F. J. de Cos Juez[1], F. Sánchez Lasheras[2], A. Suárez Sánchez[3], P. Riesgo Fernández[3] and P.J. García Nieto[4]

[1] *Project Engineering Area, University of Oviedo*
[2] *Department of Construction and Manufacturing Engineering, University of Oviedo*
[3] *Department of Business Administration, University of Oviedo*
[4] *Department of Mathematics, University of Oviedo*

e-mails: fjcos@uniovi.es, sanchezfernando@uniovi.es, suarezana@uniovi.es, priesgo@uniovi.es, pjgarcia@uniovi.es

Abstract

The aim of the present paper is the analysis of the factors that have influence over the lead time of batches of metallic components of aerospace engines. The approach used in this article employs Cox models, which are a well-recognised statistical technique for exploring the relationship between a time variable, in this case the lead time, usually called survival variable, and several explanatory variables (covariates). A model that estimates the lead time of different components has been developed using some sample batches, and its validity is checked with a different sample of similar components.

Key words: aerospace industry; Cox model; lead time; supply chain management; survival analysis.

1. Introduction

Due to the continuous and rapid changes in the aerospace industry, companies are constantly seeking better ways to manage complexity, cut costs, and boost productivity. This market is controlled by a small number of prime contractors and a large and increasing number of small and medium-sized specialized

 ISBN: 978-84-614-6167-7

companies which work as sub-contractors for one or more of the prime contractors.

In recent years some techniques such as lean manufacturing and six sigma [1] have become widespread among the companies in the field of aerospace, both in Europe and the United States. One of the key points of this strategy is that materials and work-in-progress are delivered just before they are needed, and finished goods are produced just before being sent to customers. This means a high degree of control over the production system; this relies on accurate reporting of stock levels, which is handled using computers. It has the advantage of reducing the space required to hold stock, and the costs of financing it, but increases the costs due to the risk of lack of goods for manufacturing.

In other words when an engine or an aircraft is on the assembly line and for some reason there is not enough of one of the components in stock, the assembly line must be stopped and not only is that operation of the assembly line affected, but also all the operations of the line upstream and downstream, including mechanical tests and inspections, implying a cost much higher than the cost of a security stock of the component that is missing.

The aim of the present article is the analysis of the manufacturing times required for different metallic components of aerospace engines in order to define which factors are the most important for batches being delivered on time. The analysis has been performed using a Cox model. The purpose of the model is to assess the importance of different manufacturing parameters in the total lead time. The validation of the model obtained was performed with data belonging to another sample of batches. The model obtained is considered as a worthy tool for the adjustment and revision of manufacturing schedules.

Although up to now many different parametric models and Artificial Intelligence techniques have been applied to the estimation of the price of aerospace components or even to the simulation of the behaviour of the supply chain management [2], as far as the authors know this is the first time that a survival analysis by a Cox model has been applied to the study of the lead times of metallic components in the aerospace industry.

2. Problem statement

For the present study 524 batches of aerospace engine pieces were considered. The total manufacturing cost of the batches studied was 5.5 M€. This is the total added value without including the raw material. These batches were studied from the beginning of their manufacture up to their delivery to the customer. In the present research the event time of interest is the lead time. The lead time is the period of time between the initiation of any process of production and the

 ISBN: 978-84-614-6167-7

completion of that process. In non-continuous manufacturing processes where machines do not work 24 hours a day, 7 days a week, and manufacturing queues are frequent due to bottlenecks, the lead time is not simply the sum of the times of all the manufacturing operations. In this article the lead time will be represented by letter t where t_i represents the lead time of the i-th batch. The list of variables that from an expert point of view have certain influence over the lead time are called covariates. A list of all the covariates with their meanings is included in Table 1.

Table 1. Covariates of the study.

Code	Description
BASI	Batch size (units)
CNCM	CNC machine (time in minutes)
GRMA	Grinding machine (time in minutes)
HETR	Heat treatments (time in minutes)
HOLA	Horizontal lathe (time in minutes)
ISNU	Individual serial number (yes / no)
ISTT	Inspection tests time (time in minutes)
MAFC	Manufacturing's forecasted cost (in €)
MIMA	Milling machine (time in minutes)
RMAC	Raw material cost (in €)
SUTR	Surface treatments (time in minutes)
VELA	Vertical lathe (time in minutes)

The variables CNCM, GRMA, HOLA, MIMA and VELA represent the amount of time in minutes that each piece of the batch spends in the manufacturing operations performed by the kind of machine specified (i.e., CNC machine, grinding machine, heat treatments oven, etc.). In those batches where a piece undergoes more than one operation with the same kind of machine (i.e., lathe, grinding machine, etc.) their times are added. When a piece does not require the use of any of the machine categories, the value of the correspondent variable is zero. The variables HETR and SUTR represent the amount of time that is spent on operations of heat and surface treatments. As pieces are not usually treated individually, in this case the time considered is the total amount of time that is required for processing the batch in either surface or heat treatments divided by the number of loads that have been necessary for its process. The variable ISTT represents the number of minutes that are necessary for the inspection of each piece. If there is not only a final inspection but also intermediate inspections, all times are added. While for each manufacturing operation, times are strictly controlled by electronic clocks, inspection times are manually controlled by an

 ISBN: 978-84-614-6167-7

inspector. The accuracy of this variable is lower than the accuracy of the rest of variables of the model but it has been taken into account on the advice of experts because in certain circumstances inspection times may affect the lead time of the components. There are two variables that represent different kinds of costs: MAFC and RMAC. MAFC is the manufacturing's forecasted cost expressed in euros. This variable is not always linearly related to the selling price. The raw material cost is the price of the raw material necessary for the manufacturing of each piece but does not include any of the non-recurrent costs of the product such as fixtures or tools. The variable BASI represents the batch size. The variable ISNU is a categorical binary variable that indicates if the pieces of the batch must be marked with an individual serial number or not. This variable has been introduced in the model mainly to distinguish between those pieces that are considered critical for safety reasons and are serialized, and those ones that are not.

3. Mathematical models

3.1. Batch lead time analysis through a Cox survival model

Survival analysis examines and simulates the time it takes for events to occur [3]. In the present article the survival time means the time that a batch took, from the beginning of its manufacture to the moment it was finished and ready to be dispatched to the customer. There are well known methods for estimating unconditional survival distributions. The most interesting survival modelling examines the relationship between survival and some predictors, usually termed covariates [4] in the survival-analysis literature. The Cox regression model is a standard tool in survival analysis for studying the dependence of a hazard rate on covariates and time.

A parametric model based on the exponential distribution may be written as [5]

$$\log h_i(t) = \alpha + \beta_1 \cdot x_{i1} + \beta_2 \cdot x_{i2} + ... + \beta_k \cdot x_{ik} \quad (1)$$

$$h_i(t) = e^{(\alpha + \beta_1 \cdot x_{i1} + \beta_2 \cdot x_{i2} + ... + \beta_k \cdot x_{ik})} \quad (2)$$

That is, as a linear model for the log-hazard or as a multiplicative model for the hazard. Here, i represents a subscript for observation while each one of the x_{ij} terms represents the covariates. The constant α in this model is a kind of baseline log-hazard, since

$$\log h_i(t) = \alpha \ \ (\text{or } h_i(t) = e^{\alpha}) \quad (3)$$

when all the x_{ij} are equal to zero.

The Cox model leaves the baseline hazard function unspecified:

$$\alpha(t) = \log h_0(t) \quad (4)$$

$$\log h_i(t) = \alpha(t) + \beta_1 \cdot x_{i1} + \beta_2 \cdot x_{i2} + ... + \beta_k \cdot x_{ik} \quad (5)$$

 ISBN: 978-84-614-6167-7

or, again equivalently,

$$h_i(t) = h_0(t) \cdot e^{(\beta_1 \cdot x_{i1} + \beta_2 \cdot x_{i2} + ... + \beta_k \cdot x_{ik})} \tag{6}$$

This model is semi-parametric because while the baseline hazard can take any form, the covariates enter in the model linearly. Consequently, the Cox model is a proportional-hazards model [6]. Remarkably, even though the baseline hazard is unspecified, the Cox model can still be estimated by the method of partial likelihood, developed by Cox [7] in the same paper in which he introduced the Cox model. Although the resulting estimates are not as efficient as maximum-likelihood estimates for a correctly specified parametric hazard regression model, not having to make arbitrary, and possibly incorrect, assumptions about the form of the baseline hazard is a compensating virtue of Cox's specification. Having fit the model, it is possible to extract an estimate of the baseline hazard.

3.2. Schoenfeld residuals for model diagnostics

In order to determine whether a fitted Cox regression model properly describes the data used for its calculus, three kinds of diagnostics can be considered: violation of the assumption of proportional hazards, influential data and nonlinearity in the relationship between the log-hazard and the covariates. All these diagnostics can be performed using residuals [8].

There are several graphical methods available for assessing how good the fit of a proportional hazards model is. As it is well known, in the proportional hazards model, the usual concept of residuals is not applicable [9]. In the present article, the Schoenfeld residuals will be employed [10]. The Schoenfeld residual is defined as the covariate value for the process that failed minus its expected value. This residuals method was proposed by Schoenfeld [11] and modified by Grambsch and Thernau [12]. The original Schoenfeld residuals are defined for each batch and each covariate, and are based on the first derivate of the log-likelihood function:

$$\frac{\partial l(b)}{\partial b_u} = \sum_{i=1}^{k} (x_{u(i)} - A_{ui(b)}) = 0 \qquad u = 1,2,..., p \tag{7}$$

where

$$A_{ui}(b) = \frac{\sum_{l \in R(t_i)} x_{ul} \cdot e^{b' \cdot x_l}}{\sum_{l \in R(t_i)} e^{b' \cdot x_l}} \tag{8}$$

A Schoenfeld residual for the *j*-th covariate of the *i*-th batch with the observed lead time t_i is:

 ISBN: 978-84-614-6167-7

$$R_{ji} = \delta_i \left[x_{ji} - \frac{\sum_{l \in R(t_i)} x_{jl} \cdot e^{\hat{b}' \cdot x_l}}{\sum_{l \in R(t_i)} e^{\hat{b}' \cdot x_l}} \right] \text{ with } j = 1,2,...,p \text{ and } i = 1,2,...,n \tag{9}$$

where $\hat{b}$ is the maximum partial likelihood estimator of b. The Schoenfeld residuals are defined only at uncensored survival times. For censored observations, they are set as missing. Since $\hat{b}$ is the solution of Eq. (7), the sum of the Schoenfeld residuals has a mean equal to zero. It can also be shown that these residuals are not correlated with one other.

Grambsch and Thernau [12] suggested that Schoenfeld residuals be weighted by the inverse of the estimated covariance matrix of $R_i = (R_{1i},...,R_{pi})'$ denoted by $\hat{V}(R_i)$. The weighted Schoenfeld residuals are better diagnostics of the power, and are used more often than the non-weighted residuals in assessing the proportional hazards assumption. To simplify the computations, Grambsch and Therneau [12] proposed the following approximation:

$$\left[\hat{V}(R_i)\right]^{-1} \approx r\hat{V}(\hat{b}) \tag{10}$$

where r is the number of events or the number of observed uncensored manufacturing times and $\hat{V}(\hat{b})$ is the estimated covariance matrix of $\hat{b}$. With this approximation, the weighted Schoenfeld residuals can approximated by:

$$R_i^* = r \cdot \hat{V}(\hat{b}) \cdot R_i \tag{11}$$

The graph of deviance and Schoenfeld residuals against manufacturing times are used to check the adequacy of the proportional hazards model. Since the Schoenfeld residuals are, in principle, independent of time, the presence of certain patterns in the graph may indicate departures from the proportional hazards assumption.

4. The lead time estimation model

The lead times of all the batches considered for the present study were all comprised between 20 and 200 days with an average value of 81.70 days. A Cox proportional hazards model was calculated (model Cox524). It was fitted with the covariates detailed in Table 1. The results of the Cox proportional hazards regression model are listed in Table 2. The meanings of the columns of Table 2 are as follows:

- coef: the coefficient of each parameter for the adjusted Cox model.
- exp(coef): the value that results from using the coefficient as the exponent of number e. It represents risks variation.
- se(coef): the standard deviation of each coefficient.

 ISBN: 978-84-614-6167-7

- *z*: the ratio of each regression coefficient to its standard error, a Wald statistic which is asymptotically standard normal under the hypothesis that the corresponding β_i is zero.
- *p*: is the *p*-value of each covariate.
- exp(-coef): the value that results from using the coefficient multiplied by -1 as the exponent of number e.

The parameter exp(coef) in Table 2 represents the changes in the lead time of the batches when there is a change of one unit in the value of the covariate. Thus for example, for variable BASI, holding the other covariates constant, an additional unit in the batch increases the hazard of delay by 0.02%. This means that although a batch with more pieces is more likely to be delayed, the increase in the likelihood of this is not significant.

Another example of the interpretation of parameter exp(coef) in Table 2 is that keeping constant the other covariates, an additional minute of time in a vertical lathe (VELA) reduces the hazard of delay by a factor of 0.9444 on average, that is, by 5.56%. Similarly, one minute less increases the hazard of delay by $\frac{1}{0.944} = 1.0589$. Although this is the exact meaning from a mathematical point of view, a more sensible interpretation of the meaning which takes into account the framework in which the survival analysis theory is being applied, lead us to conclude that as vertical lathes are not machines frequently used for the manufacturing of the analysed components, those pieces that need them are more likely to be finished on time (less average lead time) because the likelihood of finding bottlenecks in this operation during manufacturing is minimum. The interpretation that a reduction of one minute in the machining time increases the risk of delay does not make sense, as it could lead to the conclusion that a reduction of the machining time of a batch not only does not help to improve lead times but also increases the likelihood of a batch delay. This statement is not true, and the result must be interpreted simply as it has been mentioned above: those pieces that need more lathe machining are less likely to suffer from delays.

For the rest of the variables that represent machining times (CNCM, GRMA, HETR, HOLA, MIMA, SUTR), the time spent on heat treatments is the one that has a greater influence over the likelihood of delay. The other two variables of a higher impact over the lead times are the time spent on operations performed by grinding machines (GRMA) and the time necessary for the surface treatments of the piece (SUTR).

The information obtained from the analysis of the results of variable ISTT informs us that the inspection time is a variable with little influence over the delay risk. This result seems to be related to a lack of bottlenecks in the inspection area

 ISBN: 978-84-614-6167-7

together with a higher flexibility in the shift schedules. As each inspector has a basic set of inspection tools, if necessary some of them can change their working hours in order to reinforce the personnel available during certain shifts, and even work over the weekends.

Table 2. Parameters coef, exp(coef), se(coef), *z*, *p* and exp(-coef) of the Cox524 model.

Variable	coef	exp(coef)	se(coef)	*z*	*p*	exp(-coef)
BASI	0.0002	1.0002	0.3833	2.3081	0.0088	0.9998
CNCM	0.0339	1.0345	0.0690	2.4793	0.0109	0.9667
GRMA	0.0901	1.0943	1.2881	2.7647	0.0148	0.9138
HETR	0.0113	1.1200	0.0343	2.7975	0.0196	0.9888
HOLA	0.0316	1.0321	0.0523	2.4027	0.0279	0.9689
ISNU	-0.1458	0.8643	1.0935	2.7475	0.0376	1.1570
ISTT	0.0023	1.0023	0.6573	2.3115	0.0654	0.9977
MAFC	-0.0536	0.9478	1.1109	1.6107	0.0758	1.0551
MIMA	0.0522	1.0536	1.0935	2.4867	0.0826	0.9491
RMAC	-0.0092	0.9124	0.8533	1.3579	0.0977	1.009
SUTR	0.0881	1.0921	0.0523	2.5639	0.1074	0.9157
VELA	-0.5721	0.9444	0.0357	1.0996	0.1106	1.7720

Finally, the results obtained for the variables that represent the manufacturing's forecasted cost (MAFC) and the raw material cost (RMAC) suggest that the most expensive pieces (a higher cost of the raw material and more minutes of machining) have a lower likelihood of suffering from delays. This fact seems to be linked with the fact that in a factory which does not use finite capacity scheduling for the production planning, the planning of the most expensive goods is controlled in more detail than the planning of those with low value.

As it was explained before, the column marked z in the output records the ratio of each regression coefficient to its standard error, a Wald statistic which is asymptotically standard normal under the hypothesis that the corresponding β_i is zero [13]. In addition, low *p*-values mean covariates statistically significant.

 ISBN: 978-84-614-6167-7

Table 3. Results of the likelihood ratio test, Wald test and score (log rank) test of the Cox524 model.

		p-value
Likelihood ratio test	32.6342	0.00004
Wald test	31.3453	0.00003
Score (log rank) test	32.8787	0.00002

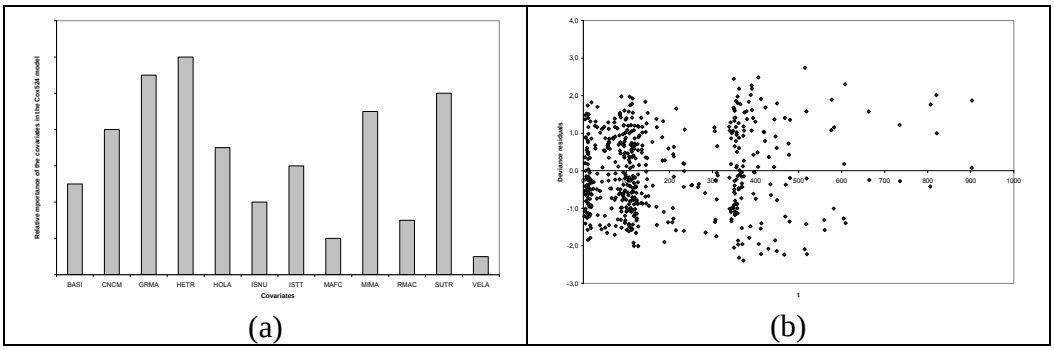


Fig. 1. (a) Relative importance of the covariates in the Cox model, and (b) deviance residuals from the fitted Cox proportional hazards model on the data.

Table 3 shows the results of the likelihood ratio test, Wald test and score (log rank) test of the Cox524 model. The likelihood ratio, Wald, and score chi-square statistics are asymptotically equivalent tests of the omnibus null hypothesis where all the β_i are zero. The *p*-value for all three overall tests is all significant, indicating that the model is significant. Therefore, this model allows one to determine the relative importance of the covariates over the lead time. This relative importance is shown in Fig. 1 (a).

The Schoenfeld residual will be calculated with a suitable transformation of time based on the Kaplan-Meier estimates of the survival function [14]. Fig. 1 (b) shows the scaled Schoenfeld residuals against transformed time. As it can be observed residuals are distributed symmetrically around zero and do not present any pattern. Larger positive values of *t* are associated in general with larger *t* values, but in general the value of the deviance residuals does not seem to be associated with the value of *t*. Therefore, the deviance residuals suggest that the proportional hazards model provides a reasonable fitting to the data.

 ISBN: 978-84-614-6167-7

5. Model validation

In order to validate the built Cox model, another sample of 110 batches from a different factory and from a different period of time but of the same kind of components was studied by means of a Cox model. The validation of the model Cox524 was carried out in two ways, using methodologies previously applied by other researchers [15]:

- A new Cox model (Cox110) was performed for these batches. The value of all the parameters of the covariates was obtained. The importance in the classification of the covariates was calculated and compared with the classification obtained for the previous Cox model (Cox524). The results are presented in Fig. 2.
- The previous Cox model (Cox524) was applied to the data and the differences in the results are shown in Fig. 2.

A comparison of the importance of variables in Fig. 2 shows that in general, for different sample sizes, the main variables that have an influence over the lead time were the same, most of them having the same position in relative importance.

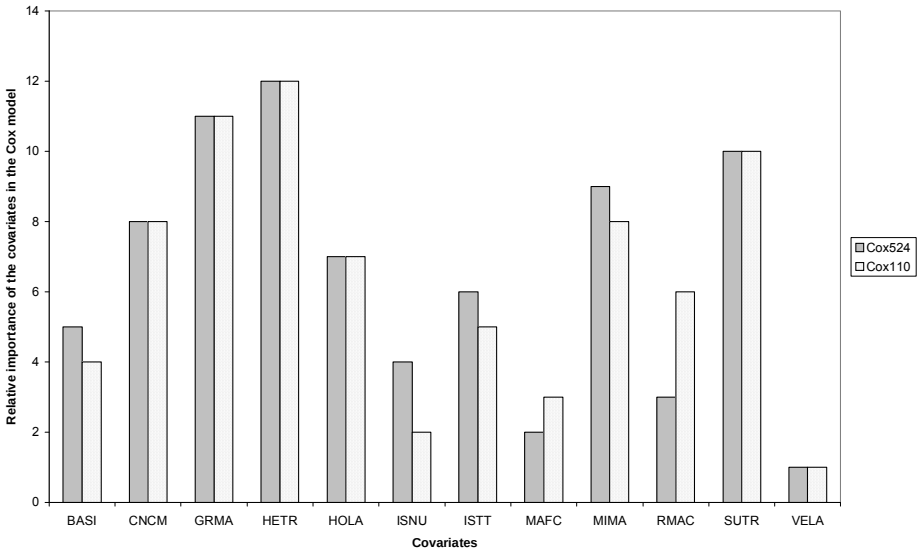


Fig. 2. Relative importance of the covariates in Cox model 524 and 110.

Although the results shown in Fig. 3 demonstrate that both models accurately represent the real lead times of the sample of 110 batches, the calculation of the mean absolute percentage error (*MAPE*) using the formula:

$$MAPE = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{A_t - F_t}{A_t}\right| \tag{12}$$

 ISBN: 978-84-614-6167-7

where A_t represents the actual value of the parameter and F_t is the forecasted value, shows that while the MAPE value for the Cox524 model is 0.1294 the same parameter for the Cox110 model is 0.0482. This demonstrates that the results of the model specifically trained for the batches of the sample are better. These results lead us to the conclusion that the model Cox524 achieves good predictive results and can be used in the future with different batches in order to predict the time necessary for its manufacturing (see Fig. 3).

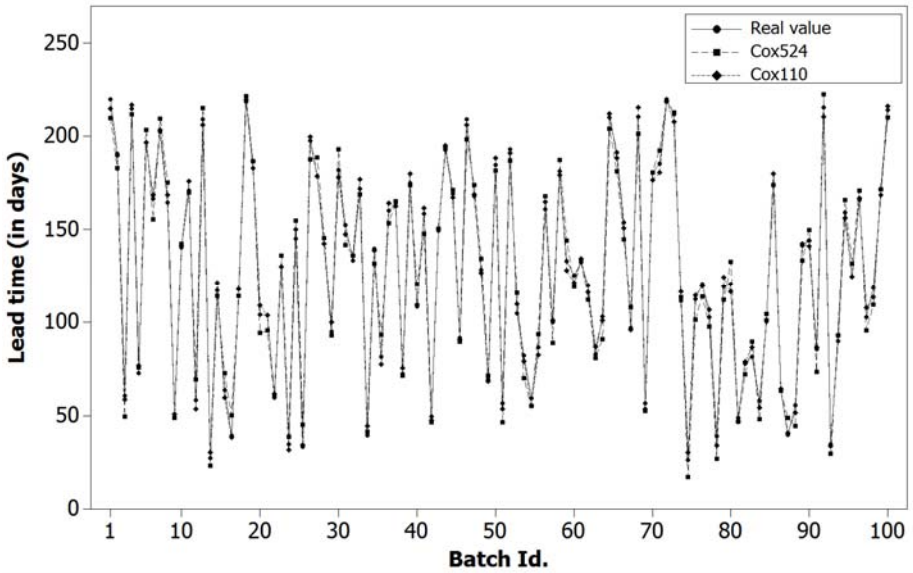


Fig. 3. Real lead times and forecasted values using Cox110 and Cox 524 for the sample of 110 batches.

6. Summary and conclusions

As a conclusion of this research it can be stated that the analysis using a survival Cox model is of great interest in order to determine which parameters have an influence over the final lead time of a batch, taking into account its manufacturing characteristics and raw material. The use of two samples, from the same kind of component and belonging to different factories and different times, allows the researchers to affirm that these results are valid and that the covariates and their patterns of relationship are maintained through time regardless of the amount and type of the rest of batches that are being manufactured at any given moment.

Further researches were performed and the results compared with the planned lead times that would presumably have been applied to the batches studied. The results obtained were coherent but in order to avoid making this paper longer are not exposed here. Finally, it must be remarked that a predictive model has been developed. This model can be used as a complementary tool by controllers and

 ISBN: 978-84-614-6167-7

production managers in order to either generate production schedules or predict the lead time of a certain component.

7. References

[1] J. DE COS, F. SANCHEZ, F. ORTEGA AND V. MONTEQUIN, *Rapid cost estimation of metallic components for the aerospace industry*, Int. J. Prod. Econ. **112** (1) (2008) 470-82.

[2] J. TANNOCK, B. CAO, R. FARR AND M. BYRNE, *Data-driven simulation of the supply-chain—Insights from the aerospace sector*, Int. J. Prod. Econ. **110** (1-2) (2008) 70-84.

[3] D.W. HOSMER AND S. LEMESHOW, *Applied Survival Analysis; Regression Modeling of Time to Event Data*, John Wiley & Sons, New York, 1999.

[4] D.G. KLEINBAUM, *Survival Analysis: A self-Learning Text*, Springer-Verlag, New York, 1996.

[5] D. COLLET, *Modelling survival data in medical research*, Chapman and Hall, London, 1995.

[6] C. HILL, *Asymptotic Relative Efficiency of Survival Tests with Covariates*, Biometrika **68** (3) (1981) 699-702.

[7] R. COX, *Regression Models and Life-Tables*, J. R. Stat. Soc. **B34** (1972) 187-220.

[8] T.M. THERNEAU AND P.M. GRAMBSCH, *Modeling Survival Data: Extending the Cox Model*, Springer, New York, 2000.

[9] C.P. FARRINGTON, *Residuals for Proportional Hazards Models with Interval-Censored Survival Data*, Biometrics **56** (2) 2000 473-82.

[10] A. WINNETT AND P. SASIENI, *Miscellanea. A note on scaled Schoenfeld residuals for the proportional hazards model*, Biometrika **88** (2) (2001) 565-71.

[11] D. SCHOENFELD, *Partial Residuals for the Proportional Hazards Regression Model*, Biometrika **69** (1) (1982) 239-41.

[12] P.M. GRAMBSCH AND T.M. THERNEAU, *Proportional hazards tests and diagnostics based on weighted residuals*, Biometrika **81**(3) (1994) 515-26.

[13] C. GOURIEROUX, A. HOLLY AND A. MONFORT, *Likelihood Ratio Test, Wald Test, and Kuhn-Tucker Test in Linear Models with Inequality Constraints on the Regression Parameters*, Econometrica **50**(1) (1982) 63-80.

[14] D.W. HOSMER, S. LEMESHOW AND S. MAY, *Applied Survival Analysis: Regression Modeling of Time to Event Data*, John Wiley & Sons, New Jersey, 2008.

[15] D. HALES, J. ROUCHIER AND B. EDMONDS, *Model-to-Model Analysis*, J. Artif. Soc. Soc. Simul. **6** (4) (2003).

 ISBN: 978-84-614-6167-7

Proceedings of the 11th International Conference on Computational and Mathematical Methods in Science and Engineering, CMMSE2011 Benidorm, Spain, 26-30 June 2011

A Metadata Management Implementation for a Symmetric Distributed File System

Antonio F. Díaz, Mancia Anguita, Erik Nieto, Hugo E. Camacho and Julio Ortega

Departamento de Arquitectura y Tecnología de Computadores, Universidad de Granada

emails: afdiaz@ugr.es, manguita@ugr.es, enieto@ugr.es, hcamacho@atc.ugr.es, julio@ugr.es

Abstract

AbFS is a symmetric distributed file system that makes it possible to share the inexpensive devices attached to the commodity computers of a cluster. The implementation of AbFS metadata management avoids the problems posed by hash-based and table-based approaches by combining hashing, tables, hierarchical structures and caches. The approach mixes attributes and namespace in the same structure. Along with the description of the proposed implementation for metadata management, this work provides experimental results to evaluate its performance. These results show that the implementation proposed offers good performance and scalability.

Key words: File system, metadata, storage system, metadata management scalability.

Introduction

The availability of interconnected computers, with multicore and multithreaded architectures and local storage, to build high performance platforms makes it possible to take advantage of this high amount of distributed storage devices through distributed file systems, such as PVFS [1], Lustre [2] and Ceph [3]. Following this researching line, we have developed AbFS (Abierto File System), a distributed file system that allows inexpensive DAS (Direct Attachment Storage) of the commodity computers (PCs or servers) of a cluster to be shared by all of these computers. This distributed file system offers a single-disk image of these DAS resources.

 ISBN: 978-84-614-6167-7

There are cluster networks (system area networks) with similar performance to the dedicated networks of commercial storage systems (Store Area Networks or SANs). Both networks can reach a throughput of several Gbps, such as Fibre Channel (http://www.fibrechannel.org/roadmaps) or 10 Gigabit Ethernet, and ten of Gbps, such as Infiniband (http://www.infinibandta.org/). Moreover, there are network technologies, such as Infiniband, that are used as both system area and store area network. Thus, a system, such as AbFS, which aggregates inexpensive commodity DAS of a cluster through a system area network, could potentially provide a cost-effective high-performance storage service compared to SAN file systems (shared disk file systems), such as IBM GPFS [4], RedHat GFS [5], SGI CXFS, PolyServe, Oracle OCFS. Every computer in the platform can be client, data server and metadata server (and metadata storage) at the same time, so AbFS is a symmetric file system. Moreover, AbFS can export SAN storage as other distributed file systems, such as Sun Lustre [2], do.

To reach high throughput, even with hundreds of nodes, the file system has to provide high performance and scalability not only for data requests but also for metadata requests. Metadata requests could account for more than 50% of all the requests produced in the file system operation ([6], [7]), making performance of metadata accesses of critical importance. With this figures, a centralized metadata management cannot scale, especially in big systems.

Metadata management provides namespace, file attributes and data block addresses but also synchronizes concurrent updates, enforces access control and maintains consistency between user data and file metadata.

This work presents the implementation of the metadata management in AbFS and the evaluation of its performance. This implementation has been done at kernel level and avoids the problems reported in previous works about hash-based or table-based metadata distribution ([8], [9]). Moreover, it does not use a database to store metadata, as, for example, PVFS does. The performance evaluation tests show experimental results obtained in the framework of a complete file system prototype.

The remainder of this paper is organized as follows: Section II deals with the AbFS metadata implementation and includes comparisons with other implementations; Section III presents the experimental results obtained for the first prototype of AbFS; and Section IV summarizes the conclusions of the paper and future work.

Metadata management implementation

Although AbFS offers a symmetrical server model, it can be configured so that there are client nodes that do not share their resources (Fig. 1).

A block is the logic unit in AbFS. All the components of the file system (inodes, files, etc.) are composed of blocks of 4 KB. A block is identified by 64 bits: 12 bits to identify the volume and 52 bits to identify the block within the volume. A volume is a partition in a disk.

 ISBN: 978-84-614-6167-7

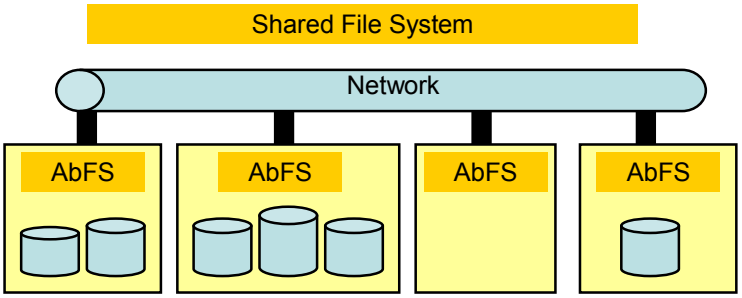


Fig. 1. AbFS system

AbFS uses three structures to manage metadata: the volume table, the delegation table, and the *inode* structure (which mixes inode table hierarchy and dentry list hierarchy). Fig. 2 shows a delegation table of 32,767 entries, and the inode subtables of the inode structure. Each inode in an inode structure occupies 512 bytes. Volume and delegation tables are replicated in each client.

A metadata server can have several volumes (Fig. 1). The volume table contains the relation between volumes and the corresponding metadata servers.

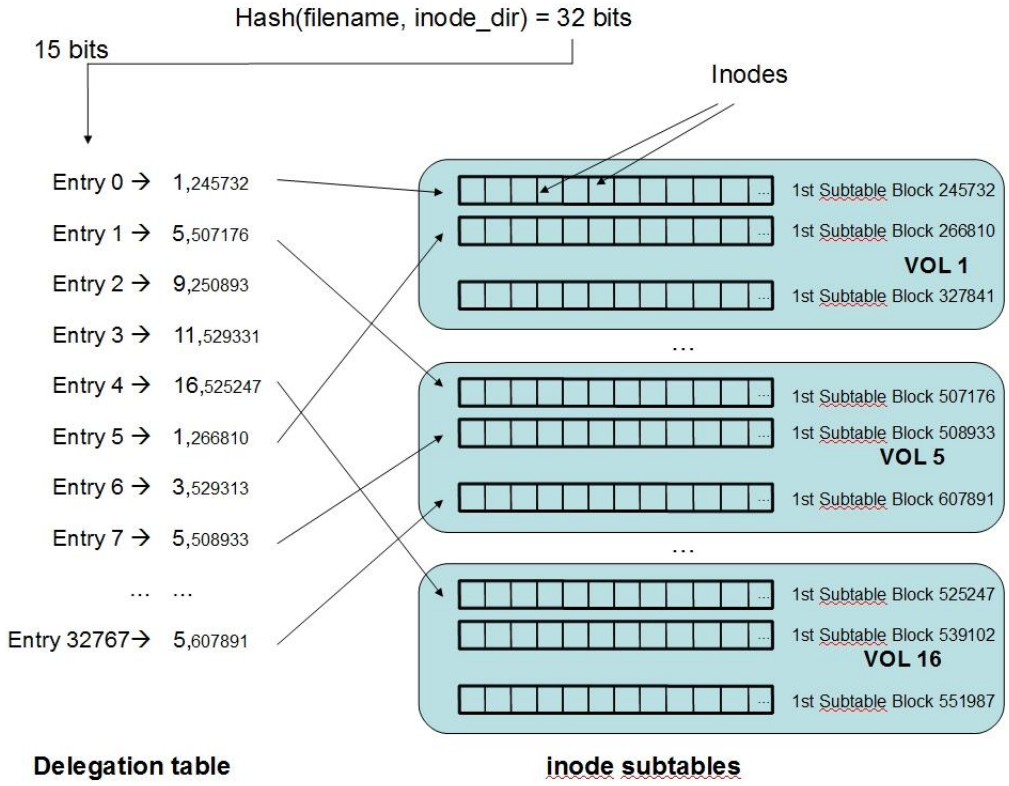


Fig. 2. Delegation table and inode subtables

 ISBN: 978-84-614-6167-7

A number of inode identifies a file or directory in AbFS. Some bits of this number are used to distribute files (directories are treated like files) among volumes and metadata servers. A hash function generates this inode number. The inode number is composed of (Fig.3): N bits (delegation bits) identify an entry in the delegation table (hash table, N is 15 bits in Fig. 2), the rest of bits (3-N2 bits) are used to identify the inode in this input. N is a maximum value, the delegation can use less bits if the number of volumes is low; i.e. an extendible hashing [10] is used to allow low cost delegation table resizing. The size of the delegation table will depend on the system size (number of volumes). In Fig. 2 there are 16 volumes and 32 K entries in the delegation table. GPFS [4] and the metadata management of [9] use extendible hashing to manage very large directories.

Fig. 3. Inode number.

Each entry in the delegation table points to a volume and an inode subtable in this volume (i.e. to the first block of the inode subtable). Each volume has several subtables. The number of subtables of a volume (i.e. the number of delegation table entries of a volume) depends on the volume size. AbFS assigns randomly the delegation entries among the volumes taking into account their sizes. AbFS assumes that a node with higher storage capacity is generally correlated with higher performance servers in heterogeneous clusters. The hash/table-based metadata implementation of [11] distributes files among metadata servers and that of [12] distributes subdirectories among metadata servers. AbFS distributes files among volumes instead of among metadata servers.

When a new volume is added some inodes migrate to other volume to balance metadata management, but the delegation table avoids the redistribution of all the inodes, as the tables used in [11] and [12] do when a server is added. Some entries of the delegation table are assigned to the new volumes, so just the inodes of these entries have to migrate. When a volume is removed, in order to restore balance, the system assigns its entries in the delegation table to other volumes and moves the inodes of the removed volume to these new assigned volumes.

Only when the inode number of a file or a directory is unknown, i. e. a lookup operation is executed, the hash function obtains the inode number of the file/directory from the inode number of its parent directory and the file/directory name (Fig. 2). In [11] and [12] the file pathname is the hash input. AbFS does not use the file pathname as hash input and maintains the inode number of a directory unchanged; so, when a directory is renamed, data redistribution is avoided. When a directory is renamed, the new entry in the inode block for this directory will point to the first inode while there are references to the original inode.

In order to obtain a quick access, the inodes in a subtable are indexed in a structure with up to three levels. For additional improvements, which can be useful for a large amount of files, AbFS can use more than one level of hash to locate the inode subtable.

 ISBN: 978-84-614-6167-7

In traditional file system dentries and inodes are store separately on disk, although, frequently, a request to access a dentry is followed by a request to access its attribute. To increase performance AbFS mixes inodes and dentries in the same structure. The directory partitioning implementation in [9] also mixes dentries and inodes in the same structure.

Test Results

The results have been obtained with twelve nodes, each one with 2 Quad-Core Intel Xeon E5450 of 3 GHz ~16GB RAM, connected through IPoIB stack using the Mellanox MT25418 Infiniband switch.

The metadata performance of AbFS was measured by using mdtest, a benchmarking tool that measures the performance of most common metadata operations like directory and file creation and the "stat" operation. The mdtest tool uses MPI to coordinate and synchronize processes between nodes.

Fig. 3 shows directory and file creation performance with 12 nodes, 6 nodes of them are (store and metadata) servers, and all the 12 nodes can be clients. The "stat" performance for the same configuration can be seen in Fig. 4. As it can be seen in both Figs. 3 and 4, performance scales linearly until the 12 clients are used, i.e., until there are 12 processes. With more processes performance does not scale. This can be observed more clearly in Fig. 5 and 6, where it is shown the performance for directory and file creation (Fig. 5) and directory and file "stat" (Fig. 6) with 12 nodes, which 6 of them working as servers and the other 6 as clients. The degradation occurs whenever there is more than one process per node contending for resources.

AbFS metadata performance clearly outperforms the performance figures of Lustre (1.6 to 2.0) ([13], [14]), PVFS [15] and Ceph [3] found in the bibliography. Tests in [14] were run in 70 client nodes, each one with 4 Quad-Core AMD Opteron 8380 ~16GB RAM and a metadata server Sun Fire X4540 with 2 Quad-Core AMD Opteron 2356 ~64GB RAM. The maximum directory and file creation performance reached with Lustre 2.0 using mdtest is of 7,517.27 ops./sec and 900.76 ops./sec respectively. The maximum directory and file "stat" performance is of 25,778.78 ops./sec and 25,016.13 ops./sec respectively. This Lustre version does not distribute metadata management as AbFS does, so the use of just one metadata server limits its performance. The forthcoming Lustre 3 will distribute metadata management among several metadata servers.

Conclusion

This work presents the performance results for the implementation of metadata management in AbFS. The tests with 12 nodes, 6 servers and 12 clients, show high performance figures for metadata management operations: more than 65,000 ops./sec. for directory creation, more than 90,000 ops./sec. for file creation, more than 5,500,000 ops./sec. for directory stat, and more than 5,800,000 ops./sec. for file "stat". We are planning to test AbFS performance with a larger number of

 ISBN: 978-84-614-6167-7

nodes and to compare it with other distributed file systems, such as Ceph and Lustre, in the same platform.

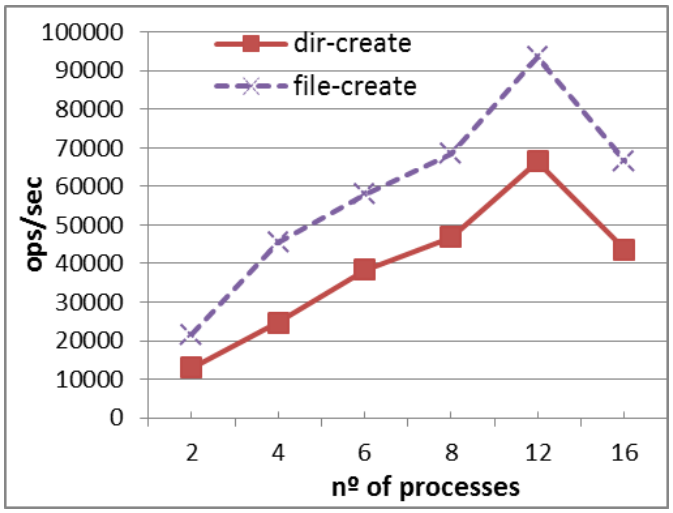


Fig. 3. Create performance (operations per second) for directories and files with 12 nodes: 6 servers, 12 clients

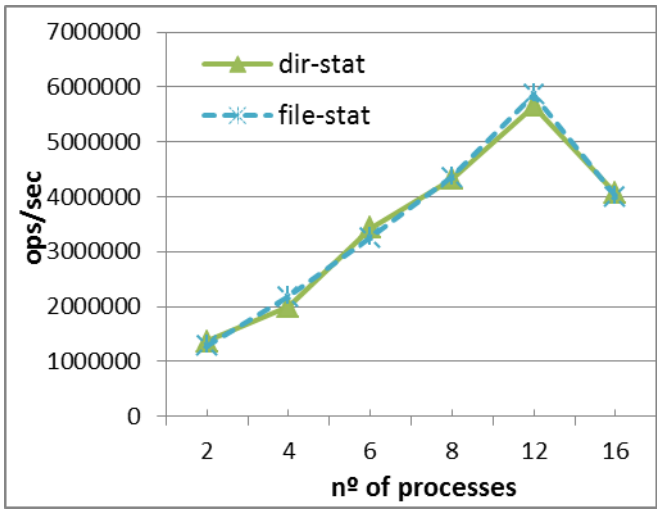


Fig. 4. "Stat" performance (operations per second) for directories and files with 12 nodes: 6 servers, 12 clients

 ISBN: 978-84-614-6167-7

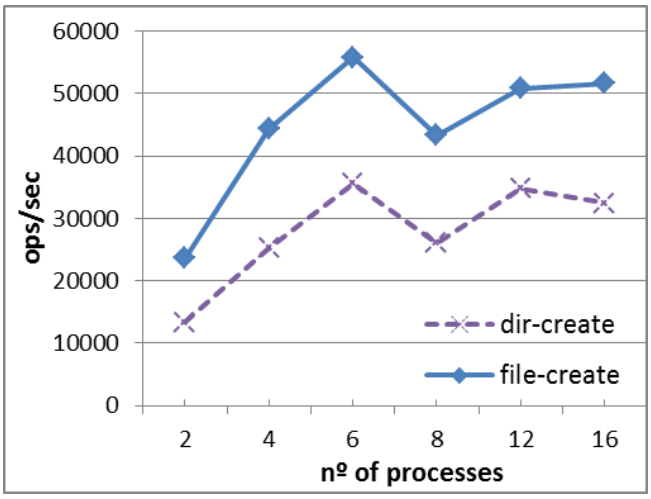


Fig. 5. Create performance (operations per second) for directories and files with 12 nodes: 6 servers plus 6 clients

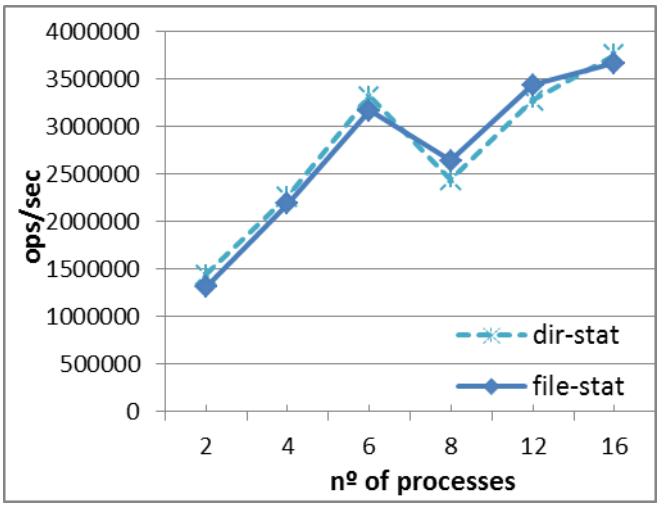


Fig. 6. "Stat" performance (operations per second) for directories and files with 12 nodes: 6 servers plus 6 clients

Acknowledgments

The authors would like to thank FCSCL (Fundación Centro de Supercomputación de Castilla y León) for giving access to a cluster of its supercomputer Calendula. This work was partially funded by the project SAF2010-20558.

 ISBN: 978-84-614-6167-7

References

[1] P. H. CARNS, W. B. LIGON III, R. B. ROSS and R. THAKUR, *PVFS: A Parallel File System for Linux Clusters*, Proceedings of the 4th Annual Linux Showcase and Conference (2000) 317-327

[2] P. J. BRAAM, *The Lustre Storage Architecture* (2002)

[3] S. A. WEIL, S. A. BRANDT, E. L. MILLER, D. D. E. LONG and C. MALTZAHN, *Ceph: a scalable, high-performance distributed file system*, OSDI '06: Proceedings of the 7th symposium on Operating systems design and implementation (2006) 307-320

[4] F. SCHMUCK and R. HASKIN, *GPFS: A Shared-Disk File System for Large Computing Clusters*, FAST '02: Proceedings of the 1st USENIX Conference on File and Storage Technologies (2002) 19

[5] S. R. SOLTIS, T. M. RUWART and M. T. O'KEEFE, *The Global File System*, Proceedings of the Fifth NASA Goddard Conference on Mass Storage Systems and Technologies (1996) 319-342

[6] J. K. OUSTERHOUT, H. DA COSTA, D. HARRISON, J. A. KUNZE, M. KUPFER and J. G. THOMPSON, *A trace-driven analysis of the UNIX 4.2 BSD file system*, SOSP '85: Proceedings of the tenth ACM symposium on Operating systems principles (1985) 15-24

[7] D. ROSELLI, J. R. LORCH and T. E. ANDERSON, *A comparison of file system workloads*, ATEC '00: Proceedings of the annual conference on USENIX Annual Technical Conference (2000) 4-4

[8] Y. ZHU, H. JIANG, J. WANG and F. XIAN, *HBA: Distributed Metadata Management for Large Cluster-Based Storage Systems*, Parallel and Distributed Systems, IEEE Transactions on **19** (2008) 750-763

[9] J. XING, J. XIONG, N. SUN and J. MA, *Adaptive and scalable metadata management to support a trillion files*, SC '09: Proceedings of the Conference on High Performance Computing Networking, Storage and Analysis (2009) 1-11

[10] R. FAGIN, J. NIEVERGELT, N. PIPPENGER and H. R. STRONG, *Extendible hashing---a fast access method for dynamic files*, ACM Trans.Database Syst. **4** (1979) 315-344

 ISBN: 978-84-614-6167-7

[11] S. A. BRANDT, E. L. MILLER, D. D. E. LONG and LAN XUE, *Efficient metadata management in large distributed storage systems*, Proceedings 20th IEEE/11th NASA Goddard Conference on Mass Storage Systems and Technologies, MSST 2003 (2003) 290-298

[12] M. XIONG, H. JIN and S. WU, *FDSSS: An Efficient Metadata Management Scheme in Large Scale Data Environment*, Fifth International Conference on Grid and Cooperative Computing Workshops, GCCW '06 (2006) 71-77

[13] W. TUREK and P. CALLEJA,
High Performance, Open Source, Dell Lustre Storage System, White Paper. Dell - University of Cambridge (2010)

[14] P. KONDEKAR, *MDS Performance Analysis*, Sun Microsystems (2009)

[15] J. M. KUNKEL and T. LUDWIG, *Performance Evaluation of the PVFS2 Architecture*, 15th EUROMICRO International Conference on Parallel, Distributed and Network-Based Processing, PDP '07. (2007) 509-516

 ISBN: 978-84-614-6167-7

Proceedings of the 11th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE2011
Benidorm, Spain, 26-30 June 2011

An Owner-based Cache Coherent Protocol for distributed file systems

Antonio F. Díaz, Mancia Anguita, Hugo E. Camacho, Erik Nieto and Julio Ortega

Departamento de Arquitectura y Tecnología de Computadores, Universidad de Granada

emails: afdiaz@ugr.es, manguita@ugr.es, enieto@ugr.es, hcamacho@atc.ugr.es, julio@ugr.es

.

Key words: distributed file system, client data cache, storage system

Extended Abstract

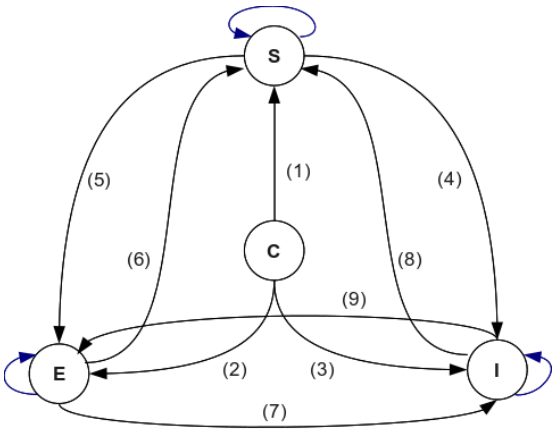
Midrange and high-end systems use I/O systems based on Storage Area Network (SAN) to achieve high performance I/O. Nevertheless, due to the increased deployment of clusters, the use of the disks in the cluster nodes to improve the performance/cost ratio has been proposed as an adequate approach for storage requirements that could avoid the cost of a SAN. AbFS is a DAS-based (Direct Attachment Storage) distributed file system that allows taking advantage of the commodity disks that already exist as an integral part of each node in a typical low-cost cluster.

AbFS uses client caches to improve client data and metadata read/write performance. There are applications that clearly do not benefit from client caches; the Google File System ([1]), for example, does not use client cache because it offers little benefit to most of its working set. But there are also applications that can benefit from client caches, some of them because the accesses to files exhibit temporal and/or spatial locality. Moreover, many large-scale applications, such as satellite image processing, engineering simulations, bioinformatics, etc., need to process with several parallel programs lots of data that they must load from or store in files ([2], [3], [4]). These programs communicate through these files. The execution time of some of them will improve if the input file data are in the local file system cache. They can be executed one after another or in parallel, each one

 ISBN: 978-84-614-6167-7

in a different core of a processor. Additionally, client cache can increase parallel program performance in multicore processors without changing the code or adding extra synchronization and communication among the cores because, for instance, (a) client cache can be used as a buffer that combines writes from different cores or from the same core before sending them to the servers or (b) file accesses from a core could also bring to client cache the data that other cores are going to read (by using read-ahead or pre-fetching). Multicore processing nodes is going mainstream, so it is important that applications be able to make effective use of the source of parallelism that these processors have.

AbFS uses device buffer cache in Linux to implement data client cache creating a virtual device. It also uses the Linux metadata caches, inode cache and dentry cache, to implement metadata caches. Thus, no additional structure or layers are needed to implement caches. The data cache takes advantages of the read-ahead implemented in VFS (Virtual File System), which stores data in the device buffer cache.



Nº	State	Event	Comment	Next state
1	(C)	Int. lookup or dir-create	Shared inode	(S)
2	(C)	Int. file create	Node gains inode ownership	(E)
3	(C)	Int. lookup or file opening of a file already opened for writing in other node	Node does not gain inode ownership	(I)
4	(S)	Ext. invalidation, or int. lookup or file opening of a file and other node has the ownership	Node does not gain inode ownership	(I)
5	(S)	Int. file opening for writing (1st node)	Node gains inode ownership	(E)
6	(E)	Int. close file for writing	Inode shared and writing to server	(S)
7	(E)	Server denies ownership	Exception	(I)
8	(I)	Int. lookup or open shared file for reading	Inode shared	(S)
9	(I)	Int. file opening for writing (1st time)	Node gains inode ownership	(E)

Fig. 1. Resume of the owner-based cache coherent protocol in a client (int.= internal event, ext.=external event)

 ISBN: 978-84-614-6167-7

Moreover, client caches improve performance because they reduce network accesses to the servers and consequently server bottleneck. Client caches in AbFS reduce both the bottleneck of the servers and the bottleneck of their own caches also when one node has opened a file for writing (this node is the so-called owner node). This is so because when other clients write in this file their writes are combined in the owner's cache and because other clients will read the file from the owner's cache instead of from the servers. In this situation, the cache is similar in performance to a home-based cooperative cache ([5]). Notice that while other network file systems, such as NFS, just use client-server communications, AbFS can also use client-client and server-server communications.

Fig. 1 resumes the cache coherent protocol in a client. Although the blocks of device buffer caches are of 4 KB, the block in the cache coherent protocol is a file (inode). An inode in a client or server cache can be in one of these states:

(C)lear: the inode does not exist in cache (it may exist in disk). It is the initial state of all inodes

(E)xclusive: the data and metadata of this inode are valid just here, in the cache of the owner node. The other nodes do not have valid copies of them. The owner node is the first node that opens the inode for writing.

(S)hared: the data and metadata of the inode are valid here and they may be valid in the cache of other nodes.

(I)nvalid: the data and metadata of the inode are invalid.

Fig. 2 shows the advantage of using data client cache. The read and write time are reduced in 97% approximately when client cache is used. The performance

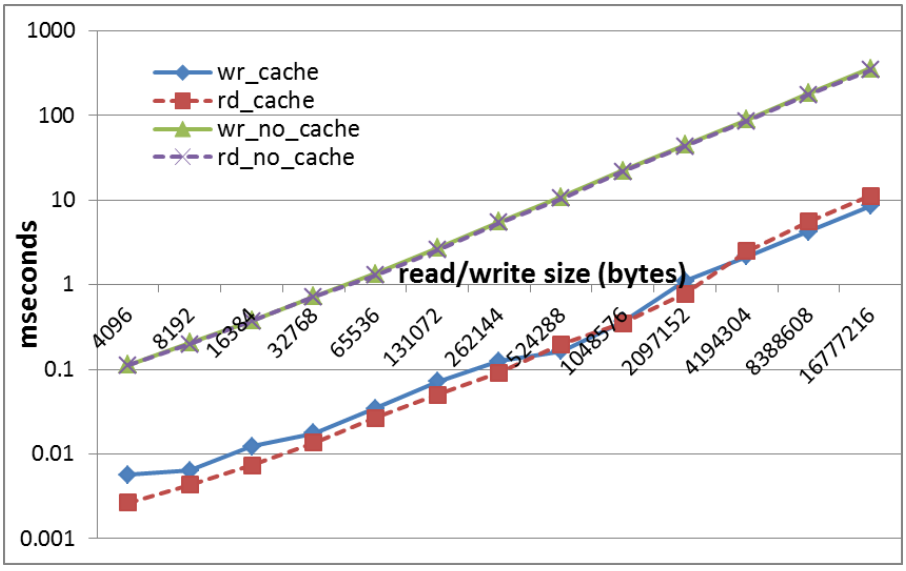


Fig. 2. Cache performance. wr_no_cache and rd_no_cache are writes and reads respectively without client cache. wr_cache and rd_cache are write hits and read hits respectively in client cache

 ISBN: 978-84-614-6167-7

increases more than with another cache that we implemented in PVFS from scratch [6].

Acknowledgment

The authors would like to thank FCSCL (Fundación Centro de Supercomputación de Castilla y León) for giving access to a cluster of its supercomputer Calendula. This work was partially funded by the project SAF2010-20558.

References

[1] S. GHEMAWAT, H. GOBIOFF and S. LEUNG, *The Google file system*, SOSP '03: Proceedings of the nineteenth ACM symposium on Operating systems principles (2003) 29-43

[2] Y. ZHU, H. JIANG, X. QIN and S. D, *A case study of parallel I/O for biological sequence search on Linux clusters*, Proceedings 2003 IEEE International Conference on Cluster Computing, (2003) 308-315

[3] A. CHING, A. CHOUDHARY, W. LIAO, R. ROSS and W. GROPP, *Evaluating structured I/O methods for parallel file systems*, Int.J.High Perform.Comput.Netw. **2** (2004) 133-145

[4] A. I. DELIS and E. N. MATHIOUDAKIS, *A finite volume method parallelization for the simulation of free surface shallow water flows*, Math.Comput.Simul. **79** (2009) 3339-3359

[5] I. HWANG, S. MAENG and J. CHO, *Home-based Cooperative Cache for parallel I/O applications*, Future Generation Computer Systems **22** (2006/4) 633-642

[6] H. E. CAMACHO, E. NIETO, M. ANGUITA, A. F. DÍAZ and J. ORTEGA, *Client cache for PVFS2*, in IEEE Proceedings of the 2010 International Conference on Parallel, Distributed and Grid Computing, (2010) 38-43.

 ISBN: 978-84-614-6167-7

Proceedings of the 11th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE2011
Benidorm, Spain, 26-30 June 2011

Application of Mathieu functions for the study of non-slanted reflection gratings

L. Alberto Estepa[1], Cristian Neipp[1], Jorge Francés[1], Manuel Pérez-Molina[1], Elena Fernández[2], Augusto Beléndez[1]

[1] *Departamento de Física, Ingeniería de Sistemas y Teoría de la Señal Universidad de Alicante,*
[2] *Departamento de Óptica, Farmacología y Anatomía, Universidad de Alicante*

Correspondence author's email: cristian@dfists.ua.es

Abstract

In this work we present an analysis of non-slanted reflection gratings by using exact solutions of the second order differential equation, derived from Maxwell equations, in terms of Mathieu functions. The results obtained by using this method will be compared to those obtained by using the well known Kogelnik's Coupled Wave Theory which predicts with great accuracy the response of the efficiency of the zeros and first order for volume phase gratings, for both reflection and transmission gratings.
.
Key words: diffraction grating, Mathieu equation, volume holograms

1. Introduction

The study of the interaction of electromagnetic radiation with diffractive elements has received much attention in the literature [1-7]. In particular, several theoretical models have been proposed to accurately describe the behaviour of diffraction gratings of different kind. The attention posed on these structures is in part due to the fact that a sinusoidal diffraction grating is the simplest periodic structure that can be recorded on a photosensitive material. Therefore the basic problem in volume holography theory is to describe accurately the properties of this kind of structures. A usual way to calculate the efficiencies of the different

 ISBN: 978-84-614-6167-7

orders that propagate in the volume grating is to solve Maxwell equations for the case of an incident plane wave on a medium where the relative dielectric permittivity varies. Although the idea seems clear and precise, in the literature there are a great number of models that allow solving the problem.

One of the most predictive and popular theories to calculate the efficiency of the orders that propagate inside a diffraction grating is the Kogelnik's Coupled Wave Theory [1]. This theory has the advantage over other theories in that, in spite of being mathematically simple, it predicts very accurately the response of the efficiency of the zero and first order for volume phase gratings. Nonetheless, the accuracy decreases when either the thickness is low or when over-modulated patterns (high refractive index modulations) are recorded in the hologram. In these cases, the coupled wave theory (CW) allowing for more than two orders or the rigorous coupled wave theory (RCW) [4-5] which doesn't disregard second derivatives in the coupled wave equations as does CW, are needed.

Although exact predictions can be obtained by using the RCW it is still interesting to work with analytical expressions in order to calculate the efficiency of the different orders that propagate inside the hologram. Analytical expressions give a deeper understanding of the physical processes than numerical solutions do. In addition, by direct inspection of the analytical expressions a clearer interpretation of how the different parameters influence in the efficiency of the different orders is got. In this work the efficiencies of the zero and first order are obtained by solving the second order differential equation, from Maxwell equations, applied to a non-slanted reflection in terms of Mathieu functions. The results obtained by using this method will be compared to those obtained by Kogelnik's coupled wave theory showing good agreement.

2. Theory

Consider a plane electromagnetic wave incident onto a periodic non-magnetic medium, which dielectric constant varies in form:

$$\varepsilon_r = \varepsilon_{r0} + \varepsilon_{r1}\cos(Kz) \tag{1}$$

The treatment is done only for TE polarization, but can be extended to TM polarization. In this case the function E (z) for the electric field inside the medium verifies the following differential equation:

$$\frac{d^2E}{dz^2} + \left[k_0{}^2\left(\varepsilon_{r0} + \varepsilon_{r1}\cos(Kz)\right) - K_x{}^2\right]E = 0 \tag{2}$$

Where:

$$k_0 = \frac{2\pi}{\lambda} \tag{3}$$

Being λ the wavelength in vacuum.

 ISBN: 978-84-614-6167-7

If θ_1 is the angle of incidence and θ_2 is the angle between the wave vector and the normal to the substrate of refractive index n_2, then the following parameters can be defined:

$$K_x = n_1(\omega/c)\sin\theta_1 = n_2(\omega/c)\sin\theta_2 \tag{4}$$

$$q_1 = n_1(\omega/c)\cos\theta_1 \tag{5}$$

$$q_2 = n_2(\omega/c)\cos\theta_2 \tag{6}$$

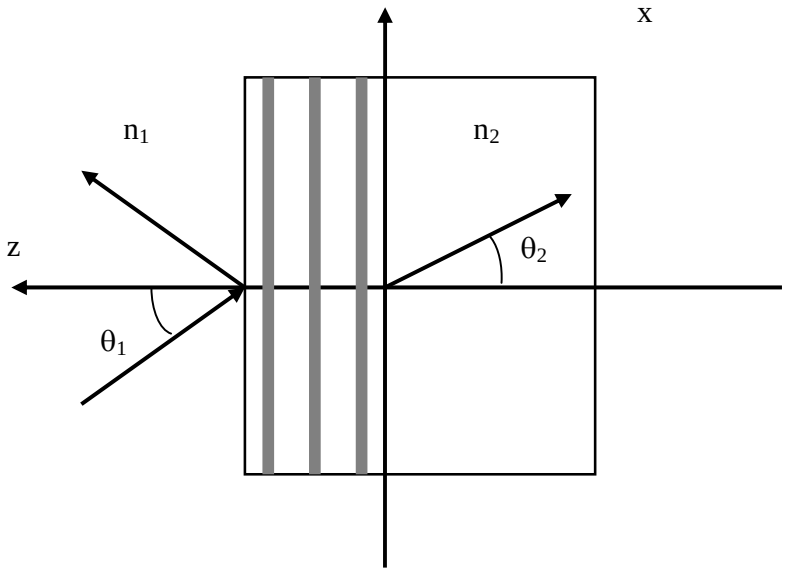


Figure 1.- Nonslanted reflection grating

The electric field in the first medium can be expressed as the superposition of an incident wave and a reflected wave of the form:

$$E^I(z) = \exp(jq_1z) + r_s\exp(-jq_1z) \tag{7}$$

While in the second medium only the transmitted wave exist:

$$E^{II}(z) = t_s\exp(jq_1z) \tag{8}$$

Now suppose that f(z) is a solution of the differential equation with a unit amplitude transmitted wave:

$$E^{II}(z) = \exp(jq_1z) \tag{9}$$

with initial conditions:

$$E(0) = 1 \tag{10}$$

 ISBN: 978-84-614-6167-7

$$\frac{dE}{dz}(0) = jq_2 \tag{11}$$

The boundary conditions in z = d imply:

$$E(d) = \exp(jq_1 d) + r_s \exp(-jq_1 d) \tag{12}$$

$$\frac{dE}{dz}(d) = jq_1 \exp(jq_1 d) - jq_1 r_s \exp(-jq_1 d) \tag{13}$$

Given the linearity of Maxwell's equations we have:

$$t_s f(d) = \exp(jq_1 d) + r_s \exp(-jq_1 d) \tag{14}$$

$$t_s \frac{df}{dz}(d) = jq_1 \exp(jq_1 d) - jq_1 r_s \exp(-jq_1 d) \tag{15}$$

From equations (14) and (15) we can obtain the amplitudes of the reflected and transmitted waves based on the solution f(z) with initial conditions (10) and (11):

$$t_s = \frac{2q_1 \exp(jq_1 d)}{-j\frac{df}{dz}(d) + q_1 f(d)} \tag{16}$$

$$r_s = \frac{\frac{df}{dz}(d) - jq_1 f(d)}{\frac{df}{dz}(d) + jq_1 f(d)} \exp(2jq_1 d) \tag{17}$$

It is now necessary to obtain a solution of the differential equation with initial conditions (10) and (11).

In this case the differential equation can be solved in terms of Mathieu functions. If we call

$$a = \frac{4(-K_x^{\,2} + k_0^{\,2}\varepsilon_{r0})}{K^2} \tag{18}$$

and

$$q = \frac{-2k_0^{\,2}\varepsilon_{r1}}{K^2} \tag{19}$$

The function f(z), solution of the differential equation (2) with initial conditions (10-11) has the form:

$$f(z) = \frac{sm(a,q,Kz/2)(2jq_2 cm(a,q,0) - K \cdot cmp(a,q,0))}{K(cmp(a,q,0)sm(a,q,0) - cm(a,q,0))} + \\ + \frac{cm(a,q,Kz/2)(-2jq_2 sm(a,q,0) - K \cdot smp(a,q,0))}{K(cmp(a,q,0)sm(a,q,0) - cm(a,q,0))}$$

And its derivative:

 ISBN: 978-84-614-6167-7

$$f'(z) = \frac{cmp(a,q,Kz/2)(-2jq_2 sm(a,q,0) + K \cdot smp(a,q,0))}{2cmp(a,q,0)sm(a,q,0) - 2smp(a,q,0)cm(a,q,0)} + \\ + \frac{smp(a,q,Kz/2)(2jq_2 sm(a,q,0) - K \cdot cmp(a,q,0))}{2cmp(a,q,0)sm(a,q,0) - 2smp(a,q,0)cm(a,q,0)}$$

Where cm (a, q, z) is the even Mathieu function sm (a, q, z) the odd Mathieu function, cmp (a, q, z) and smp (a, q, z) the corresponding derivatives.

3. Results and discussion

To validate the theoretical model previously developed we will now conduct a comparison between the results obtained using the model (classical differential theory, TDC) with those obtained by the coupled wave theory of Kogelnik (TK). A simulation for a non-slanted reflection grating with a grating period of 0.22 μm is presented, the average refractive index was supposed to be n_0 = 1.63 and an index modulation of 0,015, the incident wavelength was assumed to be of 633 nm. Figure 2 shows the diffraction efficiency as a function of the angle for a grating of thickness d = 22 μm. As shown in the figure the degree of agreement between the two theories is quite good indicating the validity of the model.

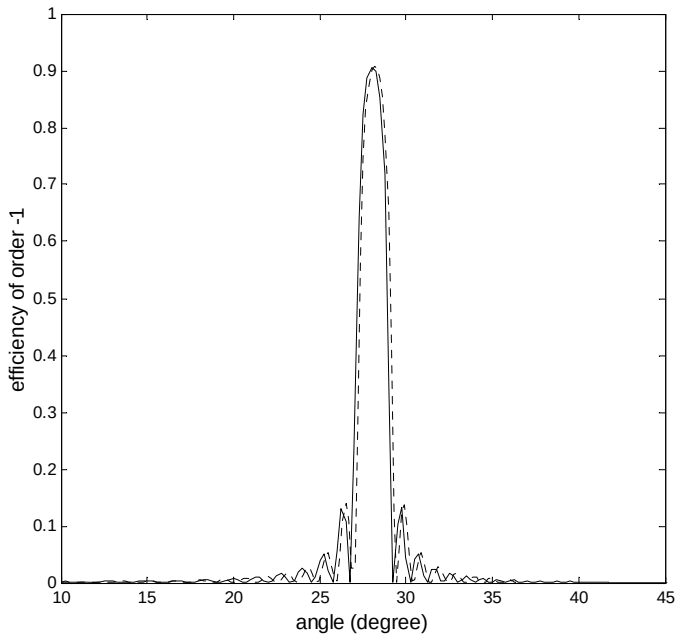


Figure 2.- Comparison of Kogelnik's Coupled Wave Theory with the method proposed in this work. Dotted line: Kogelnik's Theory; Continuous line: method of this work.

 ISBN: 978-84-614-6167-7

4. Conclusions

A solution of the second order differential equation obtained from Maxwell equations for TE describing a non-slanted reflection grating in terms of Mathieu functions is presented. The model is rigorous in the sense that no approximations are made. The results obtained by this method were compared to those obtained by using Koglenik's coupled wave theory showing a good agreement between both simulations, and thus validating the model proposed.

Acknowlegement

This work was supported by the "Ministerio de Ciencia e Innovación" of Spain under projects FIS2008-05856-C02-01 and FIS2008-05856-C02-02, and by the "Generalitat Valenciana" of Spain under project PROMETEO/2011/021.

References

[1] H. Kogelnik, *Coupled wave theory for thick hologram gratings*, Bell Syst. Techn. J. **48** (1969) 2909-2947,.

[2] L. Solymar and D. J. Cooke, Volume *Holography and Volume Gratings*, Academic, London, 1981.

[3] R. R. A. Syms, *Practical Volume Holography*, Clarendon Press, Oxford, 1990.

[4] M. G. Moharam and T. K. Gaylord, *Rigurous coupled-wave analysis of planar-grating diffraction*, J. Opt. Soc. Am. **71** (1981) 811-818.

[5] M. G. Moharam, E. B. Grann, D. A. Pommet and T. K. Gaylord, *Formulation for stable and efficienct implementation of the rigurous coupled-wave analysis of binary gratings*, J. Opt. Soc. Am. A **12** (1995) 1068-1076.

[6] N. Y. Chang and C. J. Juo, *Algorithm based on rigorous coupled-wave analysis for diffractive optical element design*, J. Opt. Soc. Am . A **18** (2001) 2491-2501.

[7] P. Dansas and N. Paraire, *Fast modelling of photonic bandgap structures by use of a diffraction-grating approach*, J. Opt. Soc. Am. A **15** (1998) 1586-1598.

 ISBN: 978-84-614-6167-7

Proceedings of the 11th International Conference on Computational and Mathematical Methods in Science and Engineering, CMMSE2011 Benidorm, Spain, 26-30 June 2011

A Novel Multi-Step Method for the Solution of Nonlinear Ordinary differential equations Using Bézier curves

A. Fallah, M.M. Aghdam and P. Haghi

Department of Mechanical Engineering, Amirkabir University of Technology

emails: ali.fallah.66@gmail.com, Aghdam@aut.ac.ir

Abstract

A new multi-step method for the solution of linear and nonlinear ordinary differential equations (ODEs) is developed based on the Bézier curves. The method is extended to solve system of ODEs. A general relation is also derived for *m*-th order multi-step formula. It is shown that the method is convergent and stable using conventional convergence and stability analyses. Using various examples, performance of the presented technique in terms of accuracy and stability is investigated. It is shown that the method provide the same order of accuracy as the well-known Adams-Moulton technique for very small step sizes. However, for relatively large values of the step size, results revealed that the presented method provide more accurate predictions at the same computational cost in comparison with Adams-Moulton. Furthermore, from stability point of view, it is also demonstrated that initiation of instability behavior occurs at smaller step sizes for Adams-Moulton method in comparison with the presented multi-step method.

Key words: multi-Step Method, Bézier Curves, Bernstein basis polynomials, nonlinear ordinary differential equations.

1. Introduction

Bernstein polynomials and B-splines are useful piecewise polynomials to represent a great variety of functions. They can be differentiated and integrated without difficulty and construct spline curves to approximate functions with desired accuracy. There has been widespread interest in the use of these polynomials to obtain solutions for various types of differential equations [1-6]. For instance, N.Caglar and H.Caglar [1] applied third-degree B-splines to singular

 ISBN: 978-84-614-6167-7

boundary value problems. In this paper, a spline method for solving singular boundary value problems is outlined based on the collocation approach [1]. Jator and Sinkala [2] applied B-splines together with collocation method to obtain solution for linear boundary value problems. In this research, solution of the boundary value problems for *d*-th order linear boundary value problem by a B-spline collocation method using *k*-th order B-splines [2]. Khalifa, Raslan, and Alzubaidi [3] developed a collocation method using cubic B-splines finite element for the numerical solution of modified regularized long wave (MRLV) equation.
H.Caglar, Ozer, and N.Caglar [4] applied B-splines to one dimensional heat equation. In this study, the boundary value problem for the one-dimensional heat equation with a nonlocal initial condition is examined by using the third degree B-splines functions [4]. A.Zerarka and B.Nine [5] developed a non-variational Galerkin-B-spline method to obtain solution for circular thin flexible plates within the context of Von Ka`rma`n equations. Main result shows that the low orders of approximation have been sufficient to build the solutions from the basis functions in a high quality [5]. Bhatti and Brackenb [6] introduced an algorithm for approximating solutions to differential equations in a modified new Bernstein polynomials basis. The algorithm expands the desired solution in terms of a set of continuous polynomials over a closed interval and then makes use of the Galerkin method to determine the expansion coefficients to construct a solution [6].
In all of these studies, the desired solution is considered as a set of piecewise or continuous polynomials over a closed interval followed by using general idea of weighted residual methods such as Galerkin or collocation to determine the expansion coefficients. On the other hand, multi-step methods convert the ordinary differential equation to equivalent integral equation. Then, the integral equation is approximated by interpolating curves that composed from some previous point. Among well-known multi-step methods, one can refer to the Adam–Moulton method in which the Newton backward difference method is used to approximate this integral equation. However, the same as other multi-step techniques for solution of differential equations the Adams-Moulton also suffers from instability when the step sizes are large.
In this paper, a new multi-step technique is developed to obtain solutions for linear and nonlinear ordinary differential equations based on the Bézier curves. The presented method shows better stability behavior in comparison with Adams-Moulton method as it postpones initiation of instability to higher values of step size. This leads to significant reduction of computational time. Furthermore, presentation of a general multi-step relation for *m*-th order formula can be considered as the other advantage of the method. Results revealed that while the method offers similar order of accuracy with Adams-Moulton technique for very small step sizes, the method is more stable and accurate for relatively larger values of step size.

2. Bézier curves

Bézier curves were widely publicized in 1962 by French engineer Pierre Bézier, who used the curves mainly to design various parts of automobile bodies.

 ISBN: 978-84-614-6167-7

Simultaneously, Paul de Casteljau also presented de Casteljau's algorithm, which is a numerically stable method to evaluate Bézier curves in 1959. The *n*th-degree Bézier curve for (*n*+1) given points $P_0, P_1, \ldots P_n$ can be written as:

$$B(u) = \sum_{i=0}^{n} B_{i,n} P_i \quad , \quad u \in [0,1] \tag{1}$$

in which the points P_i are called control points for the Bézier curves and $B_{i,n}$ are known as *Bernstein basis polynomials* of degree *n* defined as:

$$B_{i,n}(u) = \binom{n}{i} (1-u)^{n-i} u^i \quad , i = 0,1, \ldots, n \tag{2}$$

For example, Fig 1 shows cubic Bernstein basis polynomials for *n*=3.

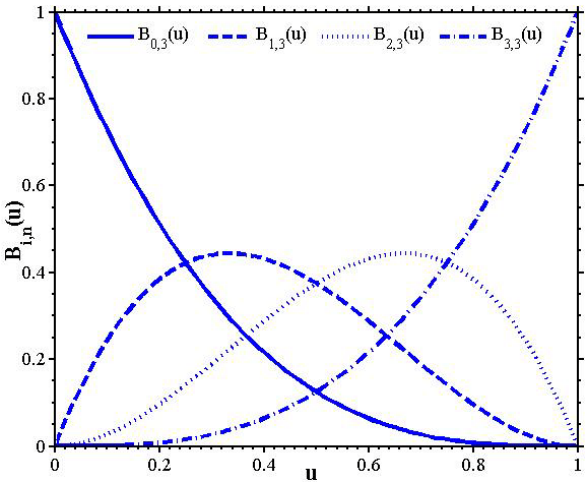


Fig 1. Cubic Bernstein basis polynomial

The location of control points in the Cartesian coordinate system can be shown as:

$$P_i = \begin{Bmatrix} x_i \\ y_i \end{Bmatrix} \tag{3}$$

Using (3), one can rewrite (1) in the parametric form of:

$$x(u) = \sum_{i=0}^{n} B_{i,n} x_i \tag{4.1}$$

$$y(u) = \sum_{i=0}^{n} B_{i,n} y_i \tag{4.2}$$

Bézier curves are widely used in computer graphics and computer-aided design mainly to provide smooth curves. It should be noted that similar to the least-square curves, Bézier curves are not really interpolating curves, as they do not normally pass through all of the control points. However, they have the important property of staying within the polygon determined by the given points as shown in Fig 2. More details of Bézier curves and their properties can be found elsewhere [7-9].

 ISBN: 978-84-614-6167-7

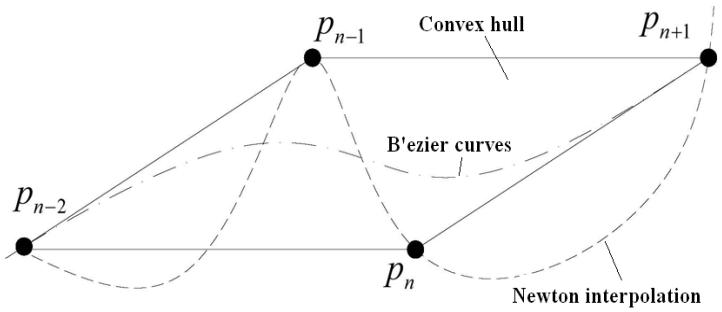


Fig 2. Comparing Newton interpolation with B´ezier curves and Convex hull

3. Technique

Consider a general linear or nonlinear first order ordinary differential equation as:

$$\frac{dy}{dx} = f(x,y), \quad x\epsilon[a,b] \ , y(a) = y_0 \tag{5}$$

The integral form of (5) can be expressed as:

$$y(x) = y_0 + \int_a^x f(x,y)dx \tag{6}$$

In multistep techniques, numerical solution is obtained by using a discretized version of (6) as:

$$y_{n+1} = y_n + \int_{x_n}^{x_{n+1}} f(x,y)dx \tag{7}$$

Various multistep methods use different ways to approximate the integral term in (7). For instance, Adam–Moulton method employs an approximation of *f(x,y)* in (7) by using the Newton backward difference method to obtain a multistep formulation.

In the present study, Bezier curves of any order are used to approximate *f(x,y)*. To approximate *f(x,y)* by *m*th-degree Bézier curves, control points should be defined as:

$$P_{n-j} = \begin{Bmatrix} x_{n-j} \\ f_{n-j} \end{Bmatrix} = \begin{Bmatrix} x_n - jh \\ f(x_{n-j}, y_{n-j}) \end{Bmatrix}, j = n-m \dots n \tag{8}$$

Where *h* is the step size. Using (4), one can write the new parametric form of *m*th-degree Bézier curves as:

$$x(u) = \sum_{i=0}^{m} B_{i,m} x_{n-m+i} = \emptyset(u) \tag{9.1}$$

$$f(x,y) = \sum_{i=0}^{m} B_{i,m} f_{n-m+i} = \varphi(u) \tag{9.2}$$

 ISBN: 978-84-614-6167-7

Finally, after approximation of the integral term in (7) for the presented study, the new form of (7) can be obtained as:

$$y_{n+1} = y_n + \int_{x_n}^{x_{n+1}} f(x,y)dx \cong y_n + \int_{\emptyset^{-1}(x_n)}^{\emptyset^{-1}(x_{n+1})} \varphi(u)\emptyset'(u)du \quad (10)$$

Using the general formula (10) together with various orders of Bézier curves in (9) leads to different multistep formulas. For example, assuming quadratic Bézier curve leads to three-step formula. Results for various multistep formulas related to degree of Bézier curves; i.e. 1, 2, 3 to *m*, are as follow:

Two-step formula:

$$y_{n+1} = y_n + \frac{h}{2}(3f_n - f_{n-1}) \quad (11)$$

Three-step formula:

$$y_{n+1} = y_n + \frac{h}{12}(19f_n - 8f_{n-1} + f_{n-2}) \quad (12)$$

Four-step formula:

$$y_{n+1} = y_n + \frac{h}{108}(175f_n - 81f_{n-1} + 15f_{n-2} - f_{n-3}) \quad (13)$$

And the general *m*+1-step formula:

$$y_{n+1} = y_n + mh\int_1^{\frac{m+1}{m}} \left(\sum_{i=0}^{m} \binom{n}{i} (1-u)^{m-i}u^i f_{n-m+i}\right)du \quad (14)$$

4. Stability and convergence analyses

A multi-step formula such as *m*+1-step can be written in general form as:

$$a_{m+1}y_{n+1} + a_m y_n + \cdots + a_0 y_{n-(m+1)} = h\left[b_{m+1}f_{n+1} + \cdots + b_0 f_{n-(m+1)}\right] \quad (15)$$

Associated with (15) are also two *characteristics* polynomials of the multi-step method as:

$$p(z) = a_{m+1}z^{m+1} + a_m z^m + \cdots + a_1 z + a_0 \quad (16)$$

$$q(z) = b_{m+1}z^{m+1} + b_m z^m + \cdots + b_1 z + b_0 \quad (17)$$

It should be noted that (14) is an explicit formula and therefore, b_{m+1}=0.

Definition 1 [10]. The multi-step formula in general form of (15) is stable if all roots of *p(z)* lie in the disk $|z|<1$ and if each root of modulus one is simple.

From (14) and (15), it can be concluded that for the *m*+1-step formula

$$p(z) = z^{m+1} - z^m \quad (18)$$

Solution to (18) for different *z* leads to:

$$z_i = \begin{cases} 0 & i = 1 \ldots m \\ 1 & i = m+1 \end{cases} \quad (19)$$

Where z_i are roots of (18). Then with regard to Definition 1, the *m*+1-step formula (14) is stable.

Definition 2 [10]. The multi-step formula (15) is consistent if $p(1)=0$ and $p'(1)=q(1)$

 ISBN: 978-84-614-6167-7

Considering (18) to determine *p'(1)* and *q(1)*, one can easily conclude for the *m*+1-step formula (14) that

$$\begin{cases} p(1) = 0 \\ p'(1) = 1 \end{cases} \quad (20)$$

Therefore, in order for *m*+1-step formula (14) to be consistent, the value of *q(1)* must be equal to one. Using general *m*+1-step formula (14) in conjunction with (15), one can determine coefficients of *q(z)* in (17) as:

$$b_i = \binom{m}{i}\left(\sum_{j=0}^{m-i}\binom{m-i}{j}(-1)^j\left(\frac{m}{i+j+1}\right)\left[\left(\frac{m+1}{m}\right)^{i+j+1}-1\right]\right), i = 0 \dots m \quad (21)$$

Substitution of (21) in (17) one can derive *q(1)* as:

$$q(1) = \sum_{i=0}^{m} b_i = \sum_{i=0}^{m}\left(\binom{m}{i}\left(\sum_{j=0}^{m-i}\binom{m-i}{j}(-1)^j\left(\frac{m}{i+j+1}\right)\left[\left(\frac{m+1}{m}\right)^{i+j+1} - 1\right]\right)\right) \quad (22)$$

After some mathematical simplification, equation (22) can be rewritten as:

$$\sum_{i=0}^{m} b_i = \sum_{i=0}^{m} c_i \quad (23)$$

Where

$$c_i = \binom{m}{m-i}(-1)^{m-i}\left(\frac{m}{m+1-i}\right)\left[\left(\frac{m+1}{m}\right)^{m+1-i}-1\right](1+(-1))^{m-i} \quad (24)$$

It can easily be found that the last parameter in (24) is zero for *i=0…m-1*. However, for *i=m* the value of c_i in (24) becomes one. Therefore, considering (22) and (24) results in:

$$q(1) = \sum_{i=0}^{m} b_i = \sum_{i=0}^{m} c_i = c_m = 1 \quad (25)$$

Consequently, based on (20), (25), and Definition 2, it can be concluded that the general *m*+1-step formula (14) is consistent.

Theorem 1 [10]. For the multi-step formula (15) to be convergent, it is necessary and sufficient that the formula (15) to be stable and consistent.

Since it was proved that the general *m*+1-step formula (14) is both stable and consistent, it can be concluded that the formula (14) is also convergent.

5. Results and discussion

 ISBN: 978-84-614-6167-7

In this section, efficiency of the presented multi-step method in terms of accuracy, convergence and computational cost is studied by various numerical examples. The predictions of the present work are compared and validated with analytical solutions where available. Furthermore, results of the well-known Adams-Moulton multi-step method are also included in the figures mainly to compare stability of the presented method.

In all examples, three figures are presented for three different levels of the step size *h*. These step sized are carefully selected to show three different stages. At the first stage, both presented and Adams-Moulton numeric methods are stable and accurate while in the second stage instability initiates in the Adams-Moulton technique and finally presented method shows unstable behavior at the third stage, too. Furthermore, in order to examine accuracy of the method, error percentage for both numeric techniques is reported at a sample point in all figures in which errors in brackets are related to Adams-Moulton while values in parenthesis are related to error of the present work.

Four-step formula (13) for presented numeric method is used to find predictions of this method in all examples. In order to present comparison between presented and the Adams-Moulton numeric methods, the four-step formula of the Adams-Moulton [9] is also used to obtain predictions in all examples.

In using these two methods, a special procedure must be employed to start the methods since initially only y_0 is known. Of course, a Runge-Kutta method is ideal for obtaining y_1,y_2,y_3. In this research, Ode45, that is a single-step solver for the solution of ordinary differential equations in MATLAB software, is used. Ode45 is based on an explicit fourth or fifth order Runge-Kutta formula, the Dormand-Prince pair [11].

5.1. Example 1

The first example is a linear ordinary differential equation:

$$y' = e^x - 3y\,, y(0) = 1 \tag{26}$$

With the exact solution as:

$$y = \frac{e^{-3x}}{4}(e^{2x} + 3) \tag{27}$$

Figs 3.a-c show predictions of the four-step formula (13) for various step sizes *h*. Included in the figures are also exact solution (27) and Adams-Moulton technique. It can be concluded from Fig 3.a that for small step size of *h=0.01* both numeric methods are accurate and stable. As the step size increases to *h=0.15* instability behavior can be seen for the Adams-Moulton method predictions while the presented method is still stable, see Fig 3.b. Finally, initiation of the instability can be observed for the presented method for higher step size of *h=0.6* in Fig 3.c. Furthermore, error percentage for both numeric methods is reported at *x=5.4* in all figures which also reveal for small step size of *h=0.01* error percent in sample point has very small value of 0.019% for both methods. When the step size increases to *h=0.15* error percentage in the sample point for the Adams-Moulton grow rapidly to 9.23% while this value for the presented method reaches to

 ISBN: 978-84-614-6167-7

0.15%. For higher step size of *h=0.6*, the error percentage in the sample point for the Adams-Moulton reaches to 75.48% while for the presented approach becomes 13%.

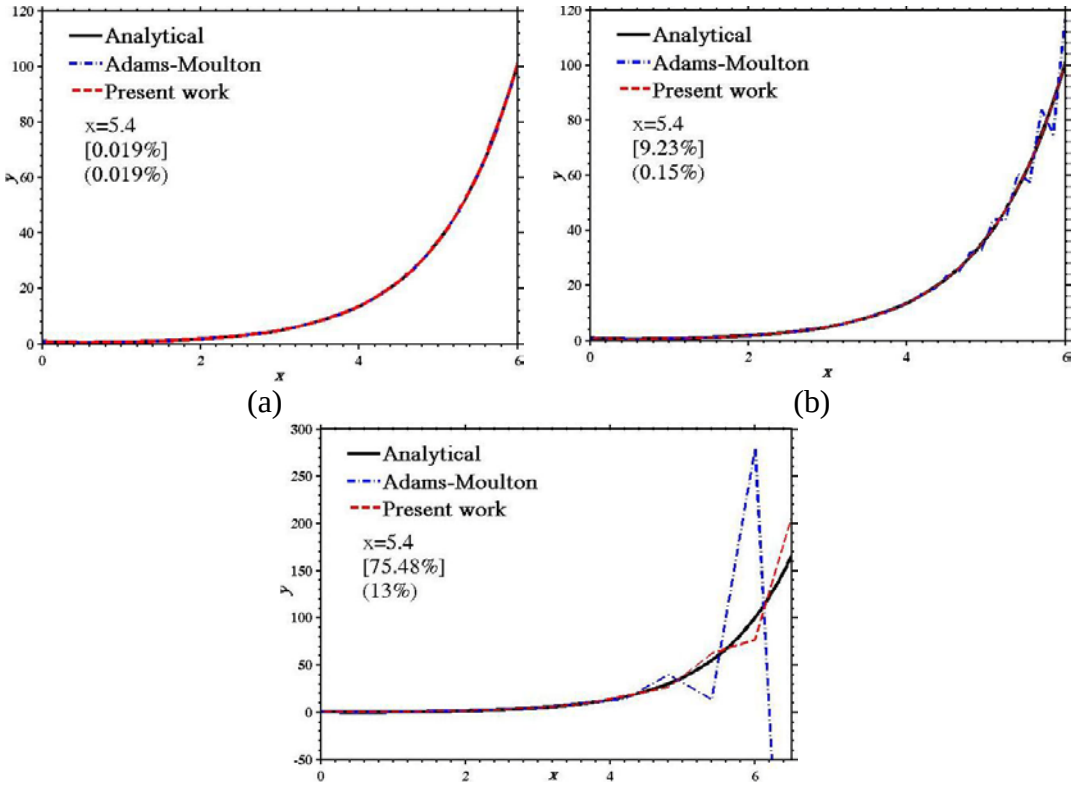


Fig 3. Results for Example 1. (a) *h*=0.1 ,(b) *h* =0.15,(c) *h* =0.6

5.2. Example 2

Second example is a nonlinear ordinary differential equation:

$$y' = \sin(x-y)\ , y(0) = 1 \tag{28}$$

With exact solution of:

$$y = x - 2\tan^{-1}\left(\frac{x-0.7066}{x+1.2934}\right) \tag{29}$$

Results of the presented method, Adams-Moulton and exact solution (29) for (30) are shown in Figs 4.a-c for various step size *h*. Fig 4.a depicts predictions for relatively small step size of *h=1* which indicates good stability and agreement with exact solution for both numeric techniques. After increasing the step size to *h=1.5*, initiation of instability can be observed for the Adams-Moulton method predictions while the presented method is still stable, see Fig 4.b. Finally, Fig 4.c shows that the unstable behavior for predictions of the presented method also initiates at greater step size of *h=2.4*.

Again, error percentage for both numeric methods is reported at a sample point of *x=12* for all cases in Figs 4.a-c. For instance, in Fig 4.a with *h=1*error percentage in the sample point is 0.28% and 0.04% for Adams-Moulton and presented

 ISBN: 978-84-614-6167-7

method, respectively. Error percentage increases to 2.92% and 0.06% for Adams-Moulton and presented method, respectively by increasing the step size to *h=1.5*. Finally, when the step size reaches to *h=2.4* in Fig 4.c, the error percentage in the same sample point grows rapidly to 32.17% for the Adams-Moulton method while for the presented approach becomes 3.56%.

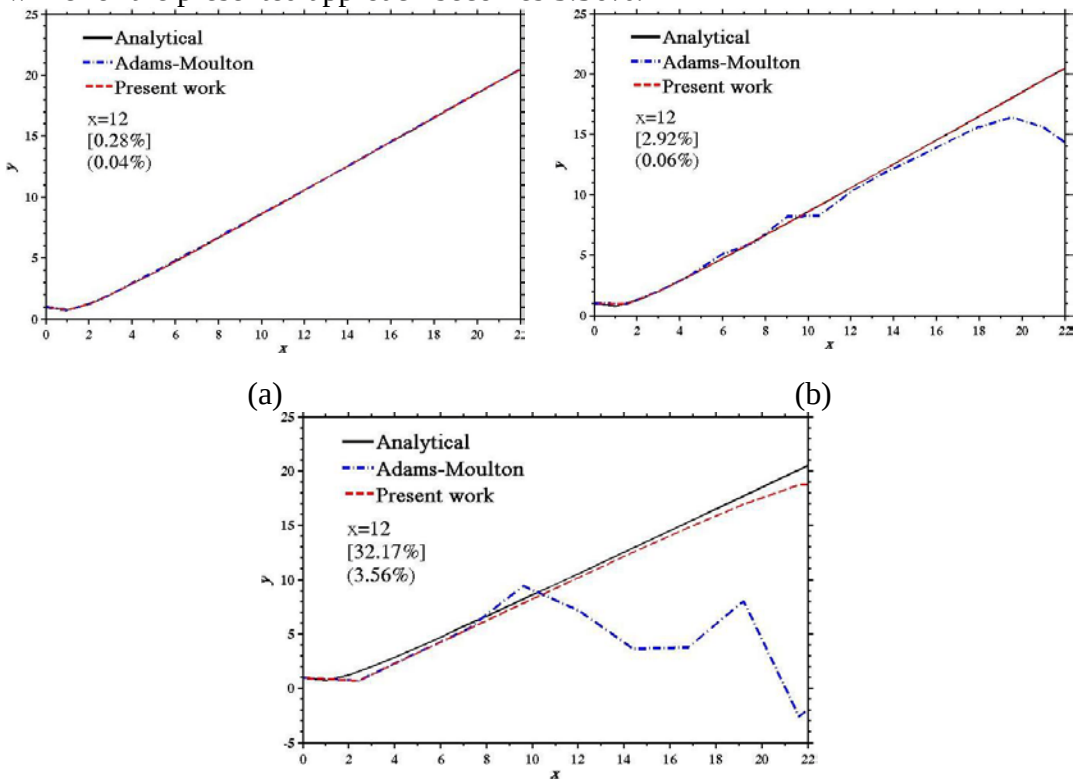


Fig 4. Results for Example 2. (a) *h*=1 ,(b) *h*=1.5 ,(c) *h*=2.4

5.3 Example 3

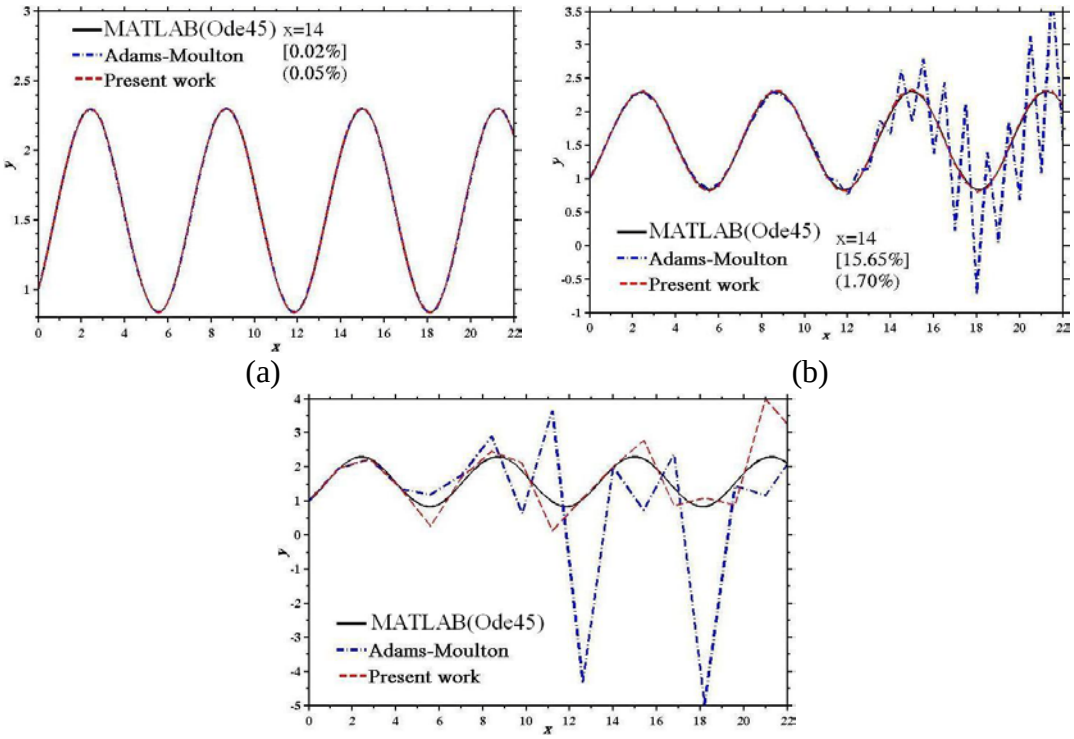
The next example is also a nonlinear ordinary differential equation:

$$y' = sin(x) + cos(y); \qquad y(0) = 1 \tag{30}$$

Results of the presented method, Adams-Moulton and MATLAB Ode45 for (30) for various step sizes *h* are shown in Figs 5.a-c. Again, for small step size of *h*=0.1 both numeric methods are accurate and stable as shown in Fig 5.a. For larger step size of *h*=0.5 instability behavior initiates for Adams-Moulton results while the presented method is still stable, see Fig 5.b. Finally, when the step size reaches to *h*=1.4 instability occurs in the results of the presented method, see Fig5.c. Furthermore, error percentage for both numeric techniques is reported at *x=14* in Figs 5.a-b which reveal that in small step size of *h*=0.1 error percentage of the Adams-Moulton (0.02%) in the sample point is less than presented method (0.05%). However, for *h*=0.5 error percentage in the same sample point for the Adams-Moulton method (15.65%) is about ten times of the presented method

 ISBN: 978-84-614-6167-7

(1.70%). It should be noted that for h=1.4 error is not reported in Fig 5.c as the Adams-Moulton method becomes seriously unstable. It is also worth mentioning that exact solution for (30) is not available and therefore, results are compared with the results obtained by ODE45 in MATLAB and all error percentage are determined based on MATLAB predictions.



(a) (b)

(c)

Fig 5. Results for Example 3. (a) h=0.1 ,(b) h=0.5 ,(c) h=1.4

5.4 Example 4

The final example is a system of two coupled nonlinear first-order differential equations:

$$\begin{cases} \dfrac{dy_1}{dx} = sin(y_1 + y_2), & y_1(0) = 1 \\ \dfrac{dy_2}{dx} = cos(y_1 + y_2), & y_2(0) = 0 \end{cases} \tag{31}$$

Figs 6.a-c, show results of the presented, Adams-Moulton and MATLAB Ode45 for system of equations (31) with different step size *h.* Fig 6.a shows results for small step size of h=0.1 where both methods are accurate and stable. Fig 6.b presents predictions for higher step size of h=0.25 where instability can be observed in Adams-Moulton predictions while the presented method is still stable. Finally, instability also occurs in the results of the presented method when the step size reaches to h=0.7, see Fig 6.c. Furthermore, error percentages for both numeric techniques are reported at $x=14$ for y_1 and y_2 in Figs 6.a-b. As shown in

 ISBN: 978-84-614-6167-7

Fig 6.a, in small step size of h=0.1 error percentage of the Adams-Moulton in the sample point is 0.003% for both y_1 and y_2 while for the presented method is 0.016% and 0.02% for y_1 and y_2, respectively. It is noticeable that for h=0.1 error percentage of Adams-Moulton method for both y_1 and y_2 is less than presented method. However, in the case of h=0.25 error percentage in the same sample point for the Adams-Moulton method is 0.61% and 0.96% for y_1 and y_2, while for the presented method is 0.04% and 0.05% for y_1 and y_2, respectively. It should be noted that accumulated error in the Adams-Moulton method increases to 60.18 % and 98.9% for y_1 and y_2 for h=0.7 while error percentage of the presented method is 2.18% and 1.28% for y_1 and y_2 in Fig 6.c. Moreover, since exact solution for (31) is not available, results are compared with the results obtained by ODE45 command in MATLAB and error percentage is determined based on the MATLAB predictions.

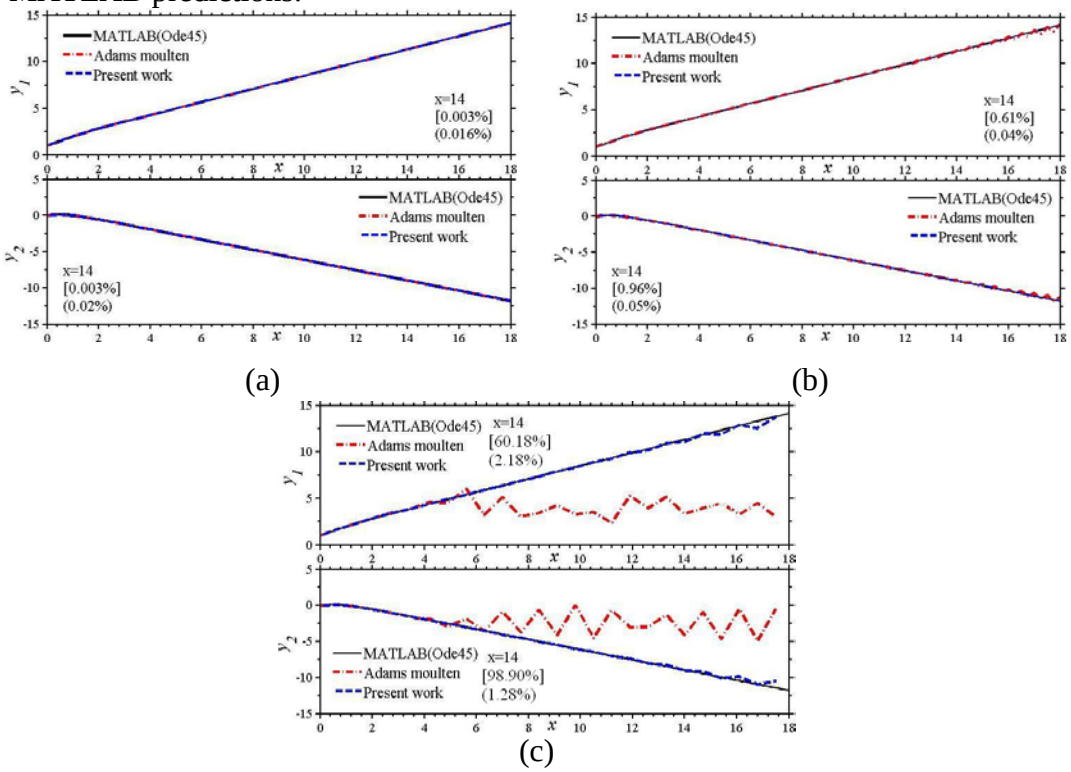


Fig 6. Results for Example 4. (a) h=0.1 ,(b) h=0.25 ,(c) h=0.7

6. Conclusion

A new multi-step method is presented for the solution of single or system of linear and nonlinear ODEs based on the Bézier curves. The presented method shows better stability behavior in comparison with the well-known Adams-Moulton method, which yields to significant reduction of computational cost. The other advantage of the method is presentation of a general multi-step relation for m-th order formula. It is shown that the method is convergent and stable using conventional convergence and stability analyses. Both accuracy and stability of

 ISBN: 978-84-614-6167-7

the method are studied using several linear and nonlinear examples. Results of the presented method together with Adams-Moulton for various linear and nonlinear ODEs are presented for three different step sizes. For very small step sizes, both methods show similar order of accuracy. As the step size increases, initiation of instability can be observed in Adams-Moulton predictions while results of presented method are still stable. Finally, instable behavior initiates in the results of the presented method for relatively high step sizes.

7. References

[1] N. Caglar AND H. Caglar, *B-spline solution of singular boundary value problems*, Appl. Math. Comput. 182 (2006) 1509–1513.

[2] S. Jator AND Z. Sinkala, *A high order B-spline collocation method for linear boundary value problems*, Appl. Math.Comput. 191 (2007) 100–116.

[3] A.K. Khalifa, K.R. Raslan AND H.M. Alzubaidi, *A collocation method with cubic B-splines for solving the MRLW equation*, Comput. Appl. Math. 212 (2008) 406 – 418.

[4] H. Caglar, M. Ozer AND N. Caglar, *The numerical solution of the one-dimensional heat equation by using third degree B-spline functions*, Chaos. Solit. Fract. 38 (2008) 1197–1201.

[5] A. Zerarka AND B. Nine, *Solutions of the Von Ka`rma`n equations via the non-variational Galerkin-B-spline approach*, Commun. Nonlinear. Sci. Numer. Simul.13 (2008)2320–2327.

[6] M.I. Bhattia AND P. Brackenb, *Solutions of differential equations in a Bernstein polynomial basis*, Comput. Appl. Math. 205 (2007)272 – 280.

[7] C.F. Gerald AND P.O. Wheatley, *Applied Numerical Analysis*, sixth ed., Addison-Wesley, 1999.

[8] G.G. Lorentz, *Bernstein Polynomials*, second ed., Chelsea Publishing Co., 1986.

[9] G. Farin, Curves and surfaces for CAGD. A practical guide. Fifth ed., Elsevier LTd, 2002.

[10] D. Kincaid AND W. Cheney, *Numerical Analysis. Mathematics of scientific computing*, first ed., Brooks/Cole Publishing Company, Pacific Grove, CA, 1990.

[11] MATLAB Reference Guide, Version 7.0, the Math Works Inc, 2004.

 ISBN: 978-84-614-6167-7

Proceedings of the 11th International Conference on Computational and Mathematical Methods in Science and Engineering, CMMSE2011 Benidorm, Spain, 26-30 June 2011

Numerical Prediction of Velocity, Pressure and Shear Rate Distributions in Stenosed Channels

Carla S. Fernandes[1], Ricardo P. Dias[2,3] and Rui Lima[3,4]

[1] *Departamento de Matemática, Escola Superior de Tecnologia e Gestão, Instituto Politécnico de Bragança*

[2] *Departamento de Tecnologia Química e Biológica, Escola Superior de Tecnologia e Gestão, Instituto Politécnico de Bragança*

[3] *CEFT – Centro de Estudos de Fenómenos de Transporte, Faculdade de Engenharia da Universidade do Porto*

[4] *Departamento de Tecnologia Mecânica, Escola Superior de Tecnologia e Gestão, Instituto Politécnico de Bragança*

emails: cveiga@ipb.pt, ricardod@ipb.pt, ruimec@ipb.pt

Abstract

Wall shear rates and pressure developed in circulatory system play an important role on the development of some clinical problems such as atherosclerosis and thrombosis. In the present work, blood flow behaviour was numerically studied in simplified domains and several relevant local properties were determined. The stenosis degree was varied in the distinct studied channels and blood rheology was described by three different models – constant viscosity, power-law model and Carreau model. Pressure attains maximum values in the wall of the atheroma and shear rates achieved maximum values in the top of the atheroma. It was also observed that, with the studied flows, the predictions for velocity and shear rate using non-Newtonian models were very similar. This observation can be explained by the magnitude of the obtained shear rates.

Key words: atheroma, velocity, shear rate, pressure, CFD

 ISBN: 978-84-614-6167-7

1. Introduction

Arthrosclerosis means literally “arteries hardening”, however it is a generic term that refers to three patterns of vascular diseases which have the hardening and loss of elasticity of the arteries walls as a common factor [1]. The dominant pattern is atherosclerosis, characterized by the formation of atheromas - fibrous plaques that generally exhibit a centre rich in lipids.

The initial lesion of an atheroma formation can trigger due to the turbulence of the flow. Most of the times, the formation of an atheroma is accomplished by a thrombus formation. It is thought that the location of higher pressures and velocities promote the endothelium lesion and hence the formation of a thrombus, which normally conduce to a thromboembolism due to the high speeds and pressures [1].

2. Numerical Simulations

The governative equations for the isothermal laminar incompressible blood flow were solved by the finite-element software POLYFLOW®. Since experimental study of blood flow in the circulatory system is not an easy task, the numerical results can be useful in order to understand the blood behaviour in stenosed arteries/vessels.

In the calculations, blood was considered both Newtonian and non-Newtonian fluid, its rheology being described, in the second case, by the power-law and Carreau models which can be mathematically expressed, respectively, by:

$$\eta = K\dot{\gamma}^{n-1}, \tag{1}$$

$$\eta = \eta_\infty + \left(\eta_0 - \eta_\infty\right)\left[1+\left(\lambda\dot{\gamma}\right)^2\right]^{(n-1)/2} \tag{2}$$

where η is the apparent viscosity, K the consistency index, n the flow index behaviour, $\dot{\gamma}$ the shear rate, η_0 the viscosity for lower shear rates, η_∞ the viscosity for higher shear rates and λ the natural time. For blood, the values of all these rheological parameters were presented in Tab.1.

Table 1: Rheological properties of blood [2].

Rheological model	η **(Pas)**	K **(Pas^n)**	n **(-)**	λ **(s)**	η_∞ **(Pas)**	η_0 **(Pas)**
Newtonian	0.00345	-	-	-	-	-
Power-law model	-	0.035	0.6	-	-	-
Carreau model	-	-	0.3568	3.313	0.00345	0.056

 ISBN: 978-84-614-6167-7

The simulations were carried out in 3D geometries representing cylindrical stenosed channels, the atheroma being constructed resorting to a semi-sphere, Fig. 1.

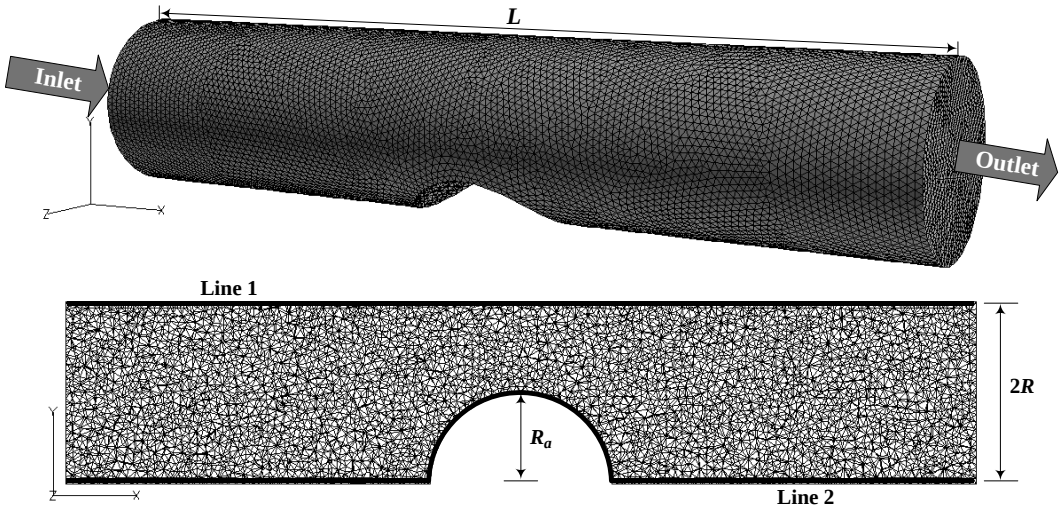


Figure 1: Representation of the computational domain and used mesh.

Three channels presenting different stenosis degrees were studied. All the channels had the same length (L = 30 mm) and radius (R = 3 mm) and the radius of the atheroma varied as presented in Tab. 2. Stenosis degree was defined as the ratio between the diameter of the channel and the radius of the atheroma ($2R/R_a$).

Table 2: Geometrical properties of the studied channels.

Channel	Stenosis degree (%)	Atheroma radius (mm)
C1	20	1.2
C2	30	1.8
C3	50	3.0

The discretization of the geometrical domain was made using an unstructured non-uniform mesh (Fig. 1) and the size of the elements was fixed after a grid independence test in which the size of the elements is successively reduced and the velocity results, obtained with the different meshes, were compared. The results were considered to be independent when a difference below 1% was achieved [3, 4].

The boundary conditions were established in order to reproduce the experimental work developed by Johnston et al. [2] to study the rheology of blood in arteries.

 ISBN: 978-84-614-6167-7

In the inlet ($x = 0$), a constant flow rate - $M_v = 1.27\times10^{-7}$ m^3s^{-1} - was imposed and non-slip at the walls was admitted.

The equations solved were the conservation of mass and momentum equations for laminar incompressible blood flow. Since this is a non-linear problem, it was necessary to use an iterative method to solve the referred equations. In order to evaluate the convergence of this process, a test based on the relative error in the velocity field was performed. For the velocity field, the modification on each node between two consecutive iterations is compared to the value of the velocity at the current iteration. In the present work, the convergence value was set to 10^{-4}, since this value is appropriate for the studied problem [4-8].

In order to verify the reliability and exactness of the computational fluid dynamics (CFD) calculations, two tests were performed: one involving local properties and other with global properties of the flow. First, velocity profiles were compared with the analytical solution for the fully developed flow of a power-law fluid in a cylindrical duct [9]:

$$v(r) = \frac{3n+1}{n}\left[1-\left(\frac{r}{R}\right)^{(n+1)/n}\right]u \tag{3}$$

where u is the average velocity and is given by:

$$u = \frac{M_v}{\pi R^2}\,. \tag{4}$$

In Fig. 2 it is possible to observe the good agreement between the numerical velocities and Eq. (3) for both Newtonian ($n = 1$) and power-law fluid ($n = 0.6$) (mean deviation of 0.28 % and 1.27% for the Newtonian and power-law fluid, respectively). As expected, the maximum deviations were observed near the wall, since the velocities in this region were close to zero.

 ISBN: 978-84-614-6167-7

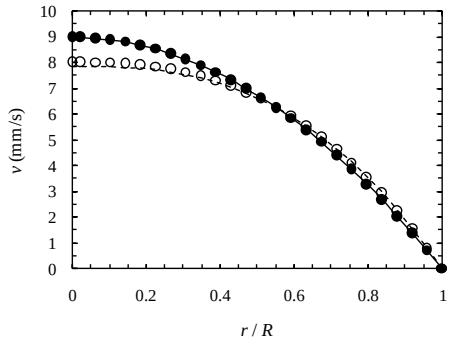


Figure 2: Velocity profiles for fully developed flow in the region before the atheroma in channel C3 for different rheological models. (●) Newtonian; (○) Power-law model; (—) Eq.(3) with $n = 1$; (- - -)Eq.(3) with $n = 0.6$.

To estimate the pressure drop, it is usual to use correlations between Fanning friction factor, f, and Reynolds number, Re. However, when the viscosity of fluid is not constant and difficult to predict, Reynolds number in the relation fRe can be replaced by a generalized Reynolds number, Re_g, the referred relation being:

$$f = a\,\mathrm{Re}_g^{-1} \Leftrightarrow a = f\,\mathrm{Re}_g \tag{5}$$

for fully developed laminar flows. In Eq. (5) a is a constant dependent of the geometry and f is given by:

$$f = \frac{\Delta P D_H}{2L\rho u^2} \tag{6}$$

with ΔP the pressure drop in a channel with length L, ρ the density of the fluid and D_H the hydraulic diameter ($D_H = 2R$ for a cylindrical duct).

The use of Re_g instead of Re allows the calculation of a single friction curve for both Newtonian and non-Newtonian fluids. For flows of power-law fluids in ducts with constant arbitrary section, the following expression for Re_g have been proposed by Delpace and Leuliet [10]:

$$\mathrm{Re}_g = \frac{\rho u^{2-n} D_H^n}{K\left\{\left(24n+\xi\right)/\left(\left(24+\xi\right)n\right)\right\}^n \xi^{n-1}} \tag{7}$$

ξ being a geometrical parameter that assumes the value 8 for cylindrical ducts.

 ISBN: 978-84-614-6167-7

Substituting Eqs. (7) and (6) in Eq. (5), constant *a* can be expressed as function of rheological and geometrical parameters as follows:

$$a = \frac{D_H^{n+1}\Delta P}{2u^n KL\left\{\left(24n+\xi\right)/\left(\left(24+\xi\right)n\right)\right\}^n \xi^{n-1}}. \tag{8}$$

In order to calculate the constant *a* for the studied channels, the pressure drop in small cylinders of 1 mm length were used and constant *a* was estimated resorting to Eq. (8). In the region before the atheroma the average values of *a* were 16.053 and 15.931 for Newtonian (n = 1 and $K = \eta$ = 0.00345 Pas) and power-law fluid (n = 0.6 and K = 0.035 Pas$^{0.6}$), respectively. Comparing these results with the one predicted analytically (16) it is possible to conclude, once again, that the numerical model used in the present work describes well the studied flow, since the mean deviations were 0.329% and 0.767% for the Newtonian and power-law fluid, respectively.

3. Results and Discussion

In the present work, velocity, pressure and shear rate profiles in stenosed channels were analyzed in order to understand the blood flow when this pathology appears.

In Fig. 3 it is possible to observe that the atheroma leads to a distortion of the velocity profile developed in a cylindrical duct.

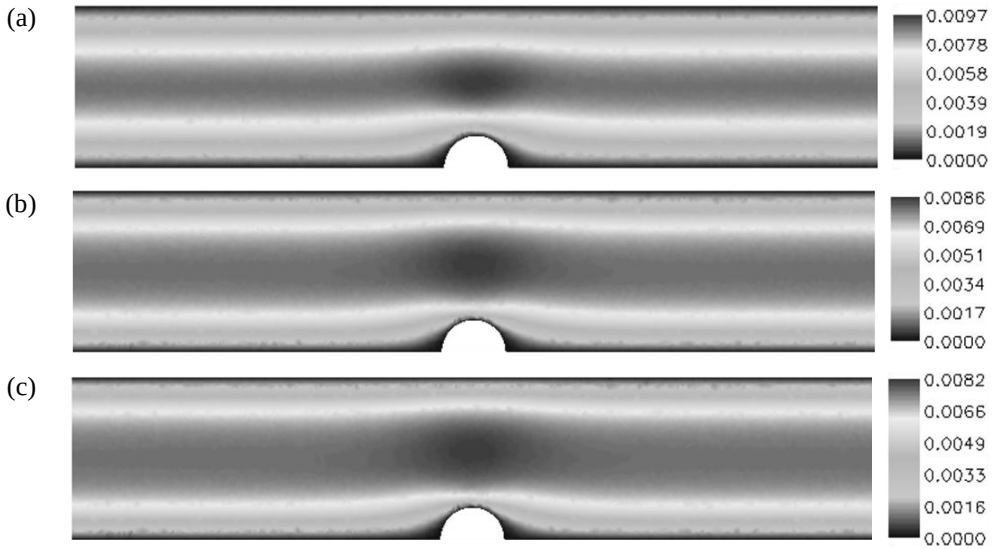


Figure 3: Velocity profiles (ms^{-1}) in the plane z = 0 (see Fig. 1) for channel C1 fluid and distinct rheological models. (a) Newtonian; (b) Power-law model; (c) Carreau model.

 ISBN: 978-84-614-6167-7

The influence of the non-Newtonian behaviour of blood is very small in the velocities developed in the studied channels, since the profiles present very similar aspect and the values of velocity were also very close when considering the different constitutive equations (Fig. 3). However, a lower difference was verified between the velocities obtained for the two non-Newtonian models for the three channels.

Quantitatively, the effect of the atheroma can be evaluated by the increase of the maximum velocity obtained for the stenosed channels compared to the one predicted analytically for a flow of a power-law fluid in a cylindrical duct, which can be determined by (Eq. (3) with $r = 0$):

$$v_{max} = \frac{3n+1}{n+1} u \,. \tag{9}$$

From the values reported in Tab. 3 it is possible to observe that the influence of the atheroma increases with the increase of the stenosis degree, as expected. In the referred table, the values of v_{max} were the ones obtained numerically and I represents the increase of v_{max} due to the presence of the atheroma (v_{max} for the channel without atheroma was calculated by Eq. (9)).

Table 3: Increase of maximum velocity due to the presence of the atheroma.

	C1		**C2**		**C3**	
	v_{max} (ms^{-1})	I (%)	v_{max} (ms^{-1})	I (%)	v_{max} (ms^{-1})	I (%)
Newtonian	0.0097	7.977	0.0108	20.2216	0.0151	68.0876
Power-law	0.0086	9.4080	0.0096	22.1299	0.0136	73.017

As observed for the velocity profiles, the shear rate distribution in the different channels and for the different rheological models were qualitatively the same. In Fig. 4 it is possible to observe that shear rate achieve its maximum in the top of the atheroma.

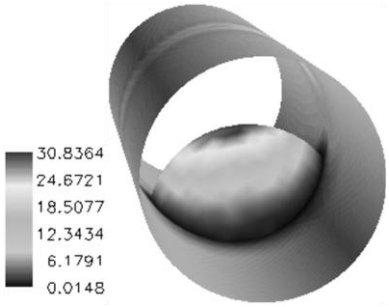
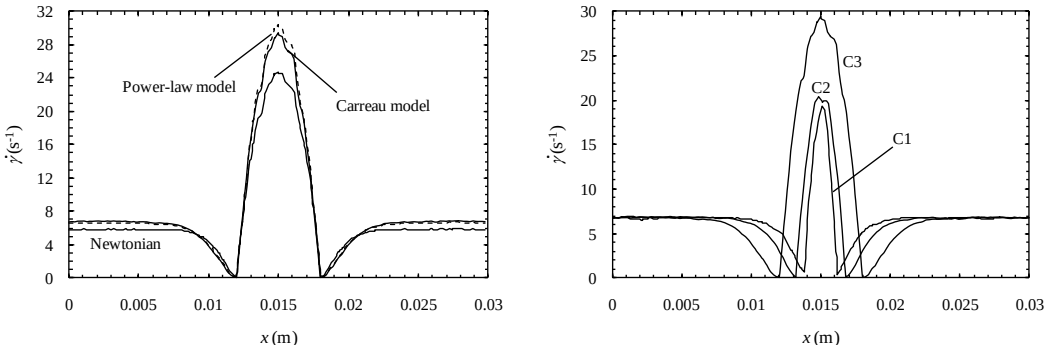


Figure 4: Shear rate in the wall of the channel C3 and Carreau model.

 ISBN: 978-84-614-6167-7

Two lines were considered in a more detailed analysis of shear rate - Lines 1 and 2 represented in Fig.1. This way it was possible to study the impact of the atheroma and rheological properties of the blood in the shear rate along the wall of the channels.

Like observed in the velocity field, the results obtained for the non-Newtonian models where very close (Figs. 5(a) and 6 (a)). The proximity of these results can be explained by the linear behaviour between shear rate and viscosity predicted by the referred models in the range in which this study was performed.



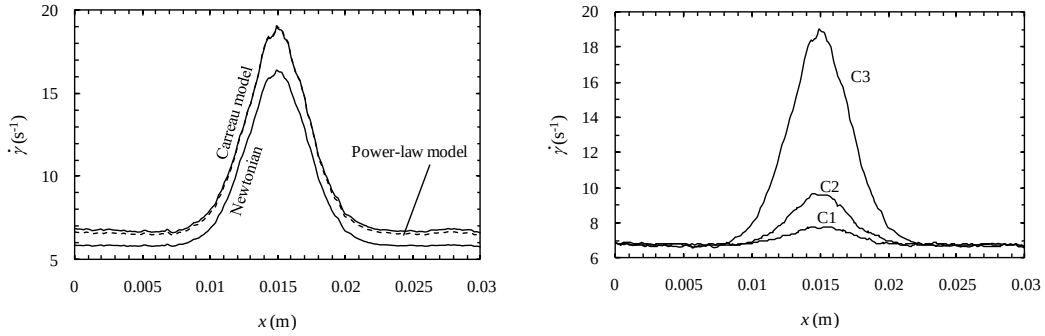
(a) (b)

Figure 5: Shear rate along Line 2 (Fig. 1). (a) Channel C3 and different rheological models. (b) Carreau model and distinct channels.

In Fig. 5(a) the behaviour of the shear rate along Line 2 (Fig. 1) can be clearly observed. Shear rate remains constant in the beginning of the channel for about 4 mm before the atheroma and then decrease until a value close to 0 s^{-1} in the base of the atheroma. Along the wall of the atheroma, shear rate exhibit a parabolic profile the maximum being reached in the top of the atheroma. As expected, the maximum shear rate increase with the increase of stenosis degree, Fig. 5(b).

The impact of the existence of the atheroma along Line 2 is also felt in Line 1 (Fig. 1), as can be observed in Fig. 6.

 ISBN: 978-84-614-6167-7



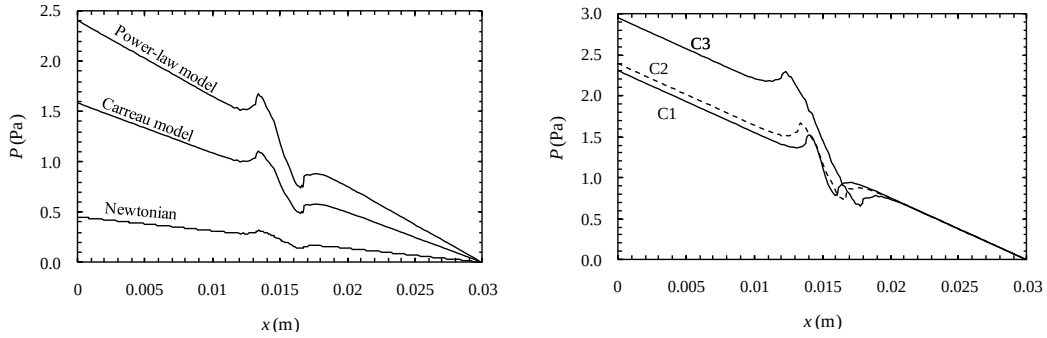
(a) (b)

Figure 6: Shear rate along Line 1 (Fig. 1). (a) Channel C3 for the different rheological models. (b) Carreau model and distinct channels.

Like it was observed for the velocity and shear rate profiles, the pressure fields were qualitatively the same for the distinct channels and rheological models, Fig 7.

Figure 7: Pressure distribution (Pa) in the plane $z = 0$ for channel C3 and Carreau model.

The pressure profile along Line 2 (Fig. 1) is dependent of the used rheological model and the pressure drop along the atheroma is much lower when the Newtonian behaviour is considered (Fig. 8(a)).



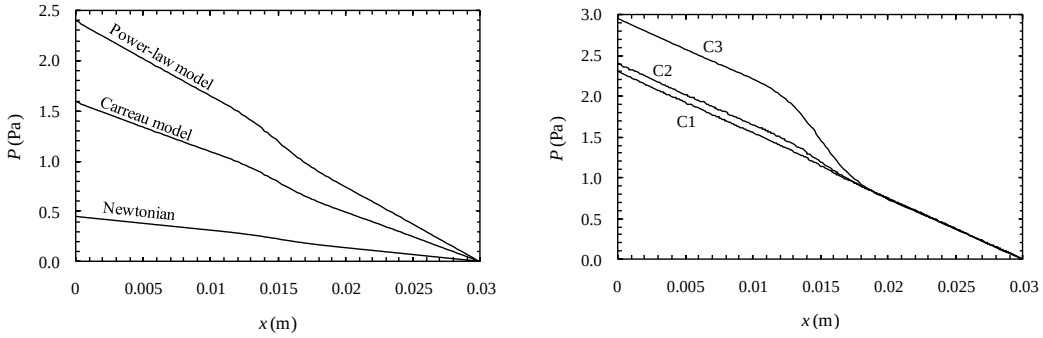
(a) (b)

Figure 8: Pressure along Line 2 (Fig. 1). (a) Channel C2 and distinct rheological models. (b) Power-law model and distinct channels.

 ISBN: 978-84-614-6167-7

In Figs. 8(a) and (b) it can also be observed that the pressure reaches a maximum in the wall of the atheroma.

In the opposite side of the base of the atheroma, Line 1 (Fig.1), the presence of this obstruction is also felt, as can be observed in Fig. 9. Since the influence of the atheroma is much lower along the referred line, the pressure for Newtonian fluid exhibits almost a linear behaviour, as the one existent in a cylindrical duct.



(a) (b)

Figure 9: Pressure along Line 1 (Fig. 1). (a) Channel C2 and distinct rheological models. (b) Power-law model and distinct channels.

4. Concluding Remarks

In the present work, the influence of rheological properties of blood and stenosis degree in the properties of the laminar blood flow in stenosed channels has been studied using the commercial finite element code POLYFLOW®. The governative equations were solved using different constitutive models – Newtonian fluid, power-law model and Carreau model. Additionally, different computational domains were also analysed.

The impact of the different rheological models in the velocity profiles were analysed and it was observed that velocities obtained for the Newtonain fluid were slightly different from the ones predicted when the blood was considered a non-Newtonian fluid, this tendency being also observed for the shear rate. Concerning the pressure profiles, different results were obtained for the distinct constitutive equations.

The analysis of pressure and shear rate developed in the wall of the channels shown that he maximum shear rates were achieved in the top of the atheroma and pressure reaches a maximum in the wall of the atheroma.

 ISBN: 978-84-614-6167-7

When the walls of the atheromas are submitted to large pressures, it is possible to generate an endothelium disruption and consequently lead to the formation of a new thrombus. The numerical results revealed that pressure achieves a maximum in the walls of the atheromas, fact that can explain the referred clinical problem.

5. Acknowledgements

The authors acknowledge the financial support provided by: PTDC/SAU-BEB/108728/2008 and PTDC/SAU-BEB/105650/2008 from the FCT (Science and Technology Foundation) and COMPETE, Portugal.

6. References

[1] S.L. ROBBINS, R.S. COTRAN, V. KUMAR, T. COLLINS, *Fundamentos de Robbins – Patologia estrutural e functional*, Rio de Janeiro, 2000.

[2] B.M. JOHNSTON, P.R. JOHNSTON, S. CORNEY, D. KILPATRICK, *Non-Newtonian blood flow in human right coronary arteries: steady state simulations*, J Biomech **37** (2004) 709-720.

[3] H.M. METWALLY, R.M. MANGLICK, *Enhanced heat transfer due to curvature-induced lateral vortices in laminar flows in sinusoidal corrugated-plate channels*, IntJHeatMassTransf **47** (2004) 2283-2292.

[4] C.S. FERNANDES, R.P. DIAS, J.M. NÓBREGA, J.M. MAIA, *Laminar flow in chevron-type plate heat exchangers: CFD analysis of tortuosity, shape factor and friction factor*, ChemEngProc **46** (2007) 825-833.

[5] R.P. DIAS, C.S. FERNANDES, J.A. TEIXEIRA, M. MOTA, A. YELSHIN, *Starch analysis using hydrodynamic chromatography with a mixed-bed particle column*, Carbohydrate Polymers **74** (2008) 852-857.

[6] C.S. FERNANDES, R. DIAS, J.M. NÓBREGA, I.M. AFONSO, L.F. MELO, J. M. MAIA, *Simulation of stirred yoghurt processing during cooling in plate heat exchangers*, JFoodEng **76** (2005) 433-439.

[7] C.S. FERNANDES, R.P. DIAS, J.M. NÓBREGA, J.M. MAIA, *Friction factors of power-law fluids in chevron-type plate heat exchangers*, JFoodEng **89** (2008) 441-447.

[8] R. LIMA, C. S. FERNANDES, R. DIAS, T. ISHIKAWA, Y. IMAI, T. YAMAGUCHI, *Microscale flow dynamics of red blood cells in microchannels: an experimental and numerical analysis*, Computational Vision and Medical Image Processing: Recent Trends, Computational Methods in Applied Sciences **19**,Springer, London, 2011.

[9] R. B. BIRD, R. C. ARMSTRONG, O. HASSAGER, *Dynamics of Polymeric Liquids*, John Wiley & Sons, New York, 1987.

[10] F. DELPLACE, J. C. LEULIET, *Generalized Reynolds number for the flow of power law fluids in cylindrical ducts of arbitrary cross-section*, ChemEngJournal **56** (1995) 33-37.

 ISBN: 978-84-614-6167-7

Proceedings of the 11th International Conference on Computational and Mathematical Methods in Science and Engineering, CMMSE2011 Benidorm, Spain, 26-30 June 2011

Multiscale computational modeling of polymer biodegradation

Luca Formaggia[1], Alfonso Gautieri[2], Azzurra Porpora[1], Alberto Redaelli[2], Simone Vesentini[2], and Paolo Zunino[1]

[1] MOX - *Department of Mathematics, Politecnico di Milano, Milan, 20133 Italy*

[2] *Bioengineering Department of Politecnico di Milano, Milan, 20133 Italy*

emails: Alfonso.Gautieri@polimi.it, Luca.Formaggia@polimi.it Azzurra.Porpora@polimi.it, Alberto.Redaelli@polimi.it, Simone.Vesentini@polimi.it, Paolo Zunino@polimi.it

Abstract

Degradable materials find a wide variety of applications in the biomedical field. Degradation, which takes place at molecular scale, propagates through the space/time scales and affects the global characteristics of the degradable device, such as drug release and the overall mechanical behavior. In this work, a bottom-up multiscale analysis is used to model the degradation mechanism taking place in poly(lactic acid) (PLA) matrices. The macroscale model is based on diffusion-reaction equations for hydrolytic polymer degradation whereas the microscale model is based on atomistic simulations to predict the water diffusion as a function of the polymer matrix degradation and swelling degree. The proposed multiscale analysis is capable to predict the time evolution time of several properties of the degradable matrix.

Key words: biomaterials, drug delivery, hydrolytic degradation, multi-scale modeling.

[1] Introduction

Biodegradable polymeric materials offer tremendous potential for the development of implantable devices and systems for treating disease. Currently, biodegradable polymers are used in diverse applications ranging from drug delivery devices [1], absorbable sutures [2], orthopedic implants [3], scaffolds for

 ISBN: 978-84-614-6167-7

tissue engineered constructs [4], medicated and biodegradable stents [5-7]. If applications involve either negligible or well-known design requirements, the design of these classes of implants are greatly facilitated. However, if the requirements are more complex, as in implants with complex geometries or under conditions that influence the course of degradation and erosion, the design process is usually hindered by the lack of rational models of biodegradable material behavior [5]. Thus, device designers must rely on a combination of intuition and trial-and-error approaches that are time consuming and often fail due to two major reasons: (i) the lack of models able to describe the evolution of the material as it degrades and erodes, and (ii) the difficulty to collect reliable experimental data quantifying and characterizing this behaviour.
Theoretical models to predict polymer degradation and erosion would seem to be important tools for a number of different applications. If drug elution is to be part of the therapy, drug delivery profiles should be programmable at the design stage. For load bearing implants, mechanical properties and structural integrity of the implant as well as their evolution should be accounted for. Because the implant is ultimately absorbed, structural breakdown and loss of function must be predicted and carefully designed.

The prevailing mechanism of biological degradation for synthetic biodegradable aliphatic polyesters (the most commonly employed biodegradable polymers in the medical such as polyglycolic acid and polylactic acid [8]) is scission of the hydrolytically unstable backbone chain by passive hydrolysis. Erosion is the process of dissolution or wearing away of degradation byproducts, resulting in mass loss from the polymer bulk. Two main modes of erosion can be systematized from widely established empirical evidence [9]. If degradation is fast, the diffusing water is absorbed quickly by hydrolysis and is hindered from penetrating deep into the polymer bulk. In this case, degradation and consequently erosion are restricted to the surface of the polymer, a phenomenon referred to as heterogeneous or surface erosion [10]. On the other hand, if degradation is slower than the rate of diffusion of water through the polymer, the reaction takes place through its entire swollen bulk, a behavior which has been termed homogeneous or bulk erosion [10]. Nevertheless, surface or bulk erosion modes are two extremes and the erosion of a polymer usually shows characteristics of both.

The authors have introduced a general class of mixture models to study water uptake, degradation, erosion, and drug release from degradable polymeric matrices [11]. The model is comprehensive starting from individual polymer scission reaction all the way up to the macro-scale diffusion, allows for the systematic characterization of the mass loss during the erosion process, and unifies both bulk and surface extremes of the erosion mode spectrum. The model unifies the behaviors of surface and bulk erosion and the nondimensionalization of the governing equations allowed the identification of the Thiele modulus, the ratio between characteristic timescales of diffusion and reaction, which is a key

 ISBN: 978-84-614-6167-7

parameter in conferring the shift between bulk or surface erosion behavior to the solution of the equations. This approach would have immediate direct impact in the design of biodegradable implants if these phenomenologically derived general constitutive behaviors were specifically characterized in regard to particular polymeric systems of interest. One possible strategy would be to perform an unprecedented series of experiments with the goal of characterizing the diffusivity of each constituent (water, drug, monomers, oligomers, etc) in a changing media (as the network degrades and erodes). To overcome this unfeasible plan, we have developed a coupling between the macroscale of the biodegradable polymer bulk, which is governed by the reaction diffusion system, and the microscale of chemical reactions and molecular diffusion, which characterizes locally the diffusion of constituents in the polymer bulk accordingly to its changing microstructure with the aid of atomistic simulations. Atomistic or molecular dynamics simulations cannot yet provide further insight into the rates of reactions taking place, but on the other hand, have been able to characterize the diffusion coefficient of molecules in a polymeric network with success (cf. [12-15]). With these new tools, we can provide a well defined data for the macroscale behavior of the polymeric matrix in terms of local diffusivities of mixture constituents.

[2] Microscale model

To study how water transport properties of PLA matrices change during the degradation process, we generate several atomistic models of PLA, with increasing level of water content (modeling the swelling process) and with decreasing chain length (resulting from the hydrolysis process). We consider a total of 30 different molecular models of PLA matrices, characterized by different degree of polymerization (monodisperse systems with 600, 300, 150, 75, 30 and 1 monomers per chain, respectively) and different degree of swelling (with 2%, 20%, 40%, 60% and 80% of water). Additionally, we study a system containing pure water for validating reasons. In order to obtain water and polymer diffusivity by means of an atomistic model of the polymer matrix, we select an ensemble of M (water or polymer) molecules in the model and we compute their mean square displacement MSD(t). Manipulating Einstein's formula one easily obtain that for sufficiently large times $\log(6D)+\log(t) = \log(MSD(t))$. Then, the realm of normal diffusion (also known as Fickian diffusion) is reached when $\log(MSD(t))$ is a linear function of time with unit slope. The validity of this fundamental property is equivalent to say that the application of Fick's law is correct.

[3] Macroscale model

We describe a polydisperse polymeric network as a collection of different linear chains of repeating units. Each chain is characterized by its degree of polymerization, defined as the number of repeating units. Diffusion driven by negative density gradients is the driving force for mass transport. An open system

 ISBN: 978-84-614-6167-7

is considered as water can penetrates into the polymer matrix from the outside aqueous environment and polymeric mass is lost to the exterior. As a result of that, we look at the polymer bulk as a mixture of water and polymer chains characterize by the partial density of water and each polymer fraction. The mass balances for each individual constituent yield the system of reaction-diffusion equations constituting the mathematical model. For the present model, hydrolysis is the mechanism that drives polymer degradation. The system of equations (for which we remand to [11]) is capable to address the following phenomena: (i) following in vivo implantation or in vitro submersion, water uptake occurs into the initially dry and non-degraded polymeric network through the mechanism of diffusion; (ii) as water becomes readily available in the vicinity of the chemical bonds, the likelihood of hydrolytic scission increases, leading to an overall molecular weight reduction; (iii) smaller chains are progressively produced, which are more eager to dissolve and diffuse through the network leading to polymer erosion; and finally, (iv) drug release from the polymeric matrix, whose material properties change due to degradation and erosion. Constitutive relationships for the diffusivities of each constituent and for the reaction rates must be specified. The mechanisms of diffusion and reaction are the only physical mechanisms that need constitutive specification and once known, the model is closed and can be solved. Diffusion depends on the nature of the constituent in question and on the local characteristics of the mixture on which is diffusing. a simplified characterization of such models will be provided by atomistic simulations of molecular diffusion. Because of the considerable computational cost of this multiscale approach, we limit ourselves to consider a static coupling strategy to feed the macroscale model with data provided by microscale sumulations. One advantage of the static coupling strategy consists in the fact that different calls to the microscale model are independent and thus can be performed in parallel on CPU or GPU clusters. By this way, the computational time needed for the miscroscale simulations is considerably reduced [16].

[4] References

1. LANGER, R., DRUG DELIVERY AND TARGETING. NATURE, 1998. 392(6679): P. 5-10.
2. LAUFMAN, H. AND T. RUBEL, SYNTHETIC ABSORABLE SUTURES. SURG GYNECOL OBSTET, 1977. 145(4): P. 597-608.
3. PIETRZAK, W.S., D.R. SARVER, AND M.L. VERSTYNEN, BIOABSORBABLE POLYMER SCIENCE FOR THE PRACTICING SURGEON. JOURNAL OF CRANIOFACIAL SURGERY, 1997. 8(2): P. 87-91.
4. AGRAWAL, C.M. AND R.B. RAY, BIODEGRADABLE POLYMERIC SCAFFOLDS FOR MUSCULOSKELETAL TISSUE ENGINEERING. JOURNAL OF BIOMEDICAL MATERIALS RESEARCH, 2001. 55(2): P. 141-150.
5. SOARES, J.S., BIOABSORBABLE POLYMERIC DRUG-ELUTING ENDOVASCULAR STENTS A CLINICAL REVIEW. MINERVA BIOTECNOLOGICA, 2009. 21(4): P. 217-230.

 ISBN: 978-84-614-6167-7

6. FORMAGGIA L. ET AL., MODELING EROSION CONTROLLED DRUG RELEASE AND TRANSPORT PHENOMENA IN THE ARTERIAL TISSUE, MATH. MOD. METH. APPL. SCI. 2010. 20(10): PP. 1759-1786.
7. ZUNINO P. , ET AL., NUMERICAL SIMULATION OF DRUG ELUTING CORONARY STENTS: MECHANICS, FLUID DYNAMICS AND DRUG RELEASE, COMPUT. METHODS APPL. MECH. ENGRG., 2009. 198(45-46): PP. 3633-3644
8. HAYASHI, T., BIODEGRADABLE POLYMERS FOR BIOMEDICAL USES. PROGRESS IN POLYMER SCIENCE, 1994. 19(4): P. 663-702.
9. VON BURKERSRODA, F., L. SCHEDL, AND A. GOPFERICH, WHY DEGRADABLE POLYMERS UNDERGO SURFACE EROSION OR BULK EROSION. BIOMATERIALS, 2002. 23(21): P. 4221-4231.
10. TAMADA, J.A. AND R. LANGER, EROSION KINETICS OF HYDROLYTICALLY DEGRADABLE POLYMERS. PROCEEDINGS OF THE NATIONAL ACADEMY OF SCIENCES OF THE UNITED STATES OF AMERICA, 1993. 90(2): P. 552-556.
11. SOARES, J.S. AND P. ZUNINO, A MIXTURE MODEL FOR WATER UPTAKE, DEGRADATION, EROSION AND DRUG RELEASE FROM POLYDISPERSE POLYMERIC NETWORKS. BIOMATERIALS, 2010. 31(11): P. 3032-3042.
12. HOFMANN, D., ET AL., MOLECULAR SIMULATION OF SMALL MOLECULE DIFFUSION AND SOLUTION IN DENSE AMORPHOUS POLYSILOXANES AND POLYIMIDES. COMPUTATIONAL AND THEORETICAL POLYMER SCIENCE, 2000. 10(5): P. 419-436.
13. TOCCI, E., ET AL., A MOLECULAR SIMULATION STUDY ON GAS DIFFUSION IN A DENSE POLY(ETHER-ETHER-KETONE) MEMBRANE. POLYMER, 2001. 42(2): P. 521-533.
14. GAUTIERI, A., ET AL., COMPUTER-AIDED MOLECULAR MODELING AND EXPERIMENTAL VALIDATION OF WATER PERMEABILITY PROPERTIES BIOSYNTHETIC MATERIALS. JOURNAL OF COMPUTATIONAL AND THEORETICAL NANOSCIENCE, 2010. 7: P. 1-7.
15. IONITA, M., ET AL., DIFFUSION OF SMALL MOLECULES IN BIOARTIFICIAL MEMBRANES FOR CLINICAL USE: MOLECULAR MODELLING AND LABORATORY INVESTIGATION. DESALINATION, 2006. 200(1-3): P. 157-159.
16. ZUNINO P. ET AL., MULTISCALE COMPUTATIONAL ANALYSIS OF DEGRADABLE POLYMERS MODELLING OF PHYSIOLOGCAL FLOWS, D. AMBROSI, A. QUARTERONI, G. ROZZA (EDITORS) SPRINGER SERIES IN MODELING, SIMULATION AND APPLICATIONS (MS&A) TO APPEAR, 2011

 ISBN: 978-84-614-6167-7

Proceedings of the 11th International Conference on Computational and Mathematical Methods in Science and Engineering, CMMSE2011 Benidorm, Spain, 26-30 June 2011

Surface Integral Modelling of Plasmonic and High Permittivity Nanostructures

Benjamin Gallinet, Andreas M. Kern and Olivier J. F. Martin

Nanophotonics and Metrology Laboratory, Swiss Federal Institute of Technology Lausanne (EPFL), Switzerland
www.nanophotonics.ch

emails: benjamin.gallinet@epfl.ch, olivier.martin@epfl.ch

Abstract

A surface integral formulation for modelling the optical properties of plasmonic and high permittivity nanostructures is presented. Applications include the engineering of fluorescence or Raman scattering enhancement, photonic crystals, plasmonic nanostructures and metamaterials.

Key words: surface integral equations, method of moments, plasmons, nanoantenna, Fano resonances

Among the wide variety of computational methods for electromagnetic scattering, the surface integral equation (SIE) method combines the advantages of finite elements (flexible discretization, accuracy of the model) with those of integral methods (excitation field and open boundary conditions explicitly included in the equations). The SIE method is widely used in microwave studies but rarely in optics. We have developed a SIE formulation based on method of moments for the simulations of three-dimensional plasmonic and high permittivity nanostructures [1,2]. The occurring integrals and their singularity can be solved quasi-analytically, enabling the accurate determination of rapidly varying fields, even arbitrarily close to a scatterer.

As the surface of an object scales with only the second power of its lateral dimension, this approach bears advantages for especially large problems but also for rounded or irregular objects as surface discretization techniques proves extremely flexible. This allows for the investigation of the effect of fabrication accuracy and material inhomogeneity on the optical response of the

 ISBN: 978-84-614-6167-7

nanostructures. For certain applications, such as plasmon-enhanced fluorescence or Raman scattering, choosing a realistic simulation geometry closely resembling the actual nanostructure is imperative. For instance, the optical response of two plasmonic nanoantennae, one with an idealized geometry and the other realistically shaped, has been investigated using the SIE method (Figure 1). While the far-field shows a similar response for both geometries, the near-field properties of the two structures are distinctly different [3]. In another application, the electric field enhancement in a random array of vertically aligned, multi-walled silver-coated carbon nanotubes has been studied in the context of surface-enhanced Raman scattering (SERS) [4], revealing different types of "hot-spot" associated with the metal inhomogeneities.

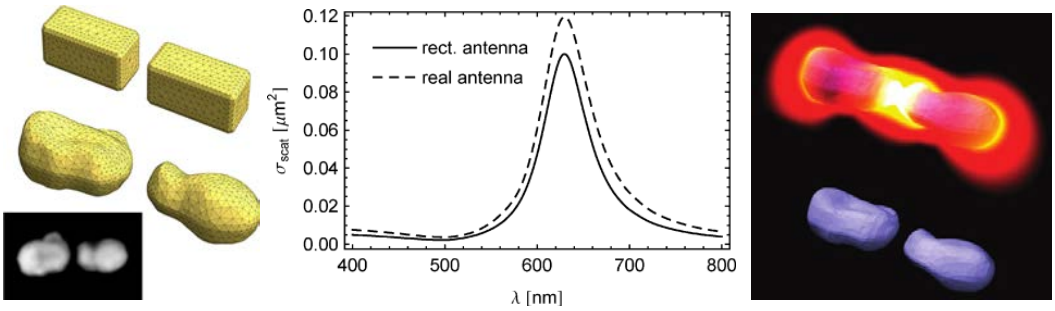


Figure 1: (a) Comparison between (top) an idealized plasmonic nanoantenna and (bottom) a realistic nanoantenna derived from a SEM image [3].

The SIE method has further been developed for periodic nanostructures [5]. This requires the evaluation of the periodic Green's function that can be efficiently performed with Ewald's method, and the explicit implementation of periodic boundary conditions at the edges of the unit cell. A large variety of geometries can be simulated this way (Figure 2), including photonic crystals, metamaterials, double periodic structures on subtrates or plasmonic nanostructures [5,6].

Both the optical near-field and far-field response of these systems can be calculated accurately. This fact becomes of particular importance in the study of experimentally relevant nanophotonic systems, such as those that exhibit asymmetric Fano resonances, which are characterized by a very strong field enhancement in the vicinity of the nanostructure and sharp variations of the optical spectrum in the far-field. Fano resonances are currently the subject of considerable research efforts in photonics and plasmonics [7]. We have recently derived an electromagnetic theory of Fano resonances in plasmonic nanostructures and metamaterials [8]. The influence of electromagnetic interactions onto the resonance line shape is revealed by our analytical theory and

 ISBN: 978-84-614-6167-7

verified by numerical calculations with the SIE method for a broad variety of plasmonic nanostructures.

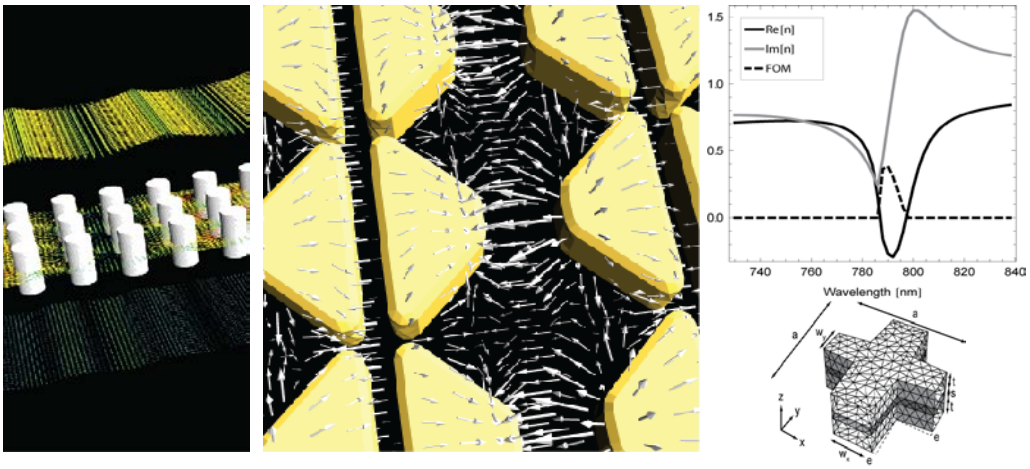


Figure 2: Applications of surface integral method for periodic nanostructures include photonic crystals, plasmonics and metamaterials.

References

[1] A.M. KERN AND O.J.F. MARTIN, *Surface Integral Formulation for 3D Simulations of Plasmonic and High Permittivity Nanostructures*, J. Opt. Soc. Am. A **26** (2009) 732-739.

[2] A.M. KERN AND O.J.F. MARTIN, *Modeling Near-Field Properties of Plasmonic Nanoparticles: a Surface Integral Approach*, Proc. of SPIE **7395** (2009) 739518.

[3] A.M. KERN AND O.J.F. MARTIN, *Excitation and Reemission of Molecules near Realistic Plasmonic Nanostructures*, Nano Lett. **11** (2011) 482-487.

[4] P. DAWSON, J.A. DUENAS, M.G. BOYLE, M.D. DOHERTY, S.E.J. BELL, A.M. KERN, O.J.F. MARTIN, A.-S. TEH, K.B.K. TEO, AND W.I. MILNE, *Combined Antenna and Localized Plasmon Resonance in Raman Scattering from Random Arrays of Silver-Coated, Vertically Aligned Multiwalled Carbon Nanotubes*, Nano Lett. **11** (2011) 365-371.

[5] B. GALLINET, A.M. KERN AND O.J.F. MARTIN, *Accurate and Versatile Modeling of Electromagnetic Scattering on Periodic Nanostructures with a Surface Integral Approach*, J. Opt. Soc. Am. A **27** (2010) 732-739.

[6] B. GALLINET, AND O.J.F. MARTIN, *Scattering on Plasmonic Nanostructures Arrays Modeled with a Surface Integral Formulation*, Photonics Nanostruct. **8** (2010) 278-284.

[7] B. LUK'YANCHUK *ET AL.*, *The Fano Resonance in Plasmonic Nanostructures and Metamaterials*, Nature Mat. **9** (2010), 707-715.

 ISBN: 978-84-614-6167-7

[8] B. GALLINET, AND O.J.F. MARTIN, *Ab Initio Theory of Fano Resonances in Plasmonic Nanostructures and Metamaterials*, Phys. Rev. B. (2011), in press.
[9] B. GALLINET, AND O.J.F. MARTIN, manuscript in preparation.

 ISBN: 978-84-614-6167-7

Proceedings of the 11th International Conference on Computational and Mathematical Methods in Science and Engineering, CMMSE2011
Benidorm, Spain, 26-30 June 2011

Comparison of two methods for defining geometric properties of surfaces measured with laser scanner for automatic geometry extraction in urban areas

Miguel García[1], Francisco Ruiz-Lopez[2], José Herráez[2], Eloina Coll[2], Jose Carlos Martínez-Llario[2]

[1] *Departamento de Bomberos, Prevención e Intervención en Emergencias de Valencia*
[2] *Instituto de Restauración Del Patrimonio, Universitat Politècnica de València*

emgarga0@aaa.upv.es, fraruilo@gmail.com, jherraez@cgf.upv.es, ecoll@cgf.upv.es, jomarlla@cgf.upv.es

Abstract

Along this paper we are going to highlight some aspects about two different methods to obtain some geometric parameters of surfaces measured with laser scanner, by calculating such parameters using the plane adjust and the inertia moment assignation. The possible different shapes of such point clouds have its own response to the geometric parameters, so we can distinguish between shapes using those parameters.

Key words: automatic, extraction, laser scanner, geometry, method, plane, inertia, moment

1. Introduction

Actually, in the process of 3D modelling of buildings, the longest task is to determine and create the break lines of each shape. Usually, within the software suite of the laser scanner there is a tool which implements some algorithms and techniques to obtain such lines, however, the quality of those automatic process usually is not as good as required, so finally it has to be done manually by expert technicians. Furthermore, the laser scanner point clouds have its own accidental errors and accumulate the errors from the registration process, so usually, we cannot create a line perfectly straight to define a corner using those automatic

 ISBN: 978-84-614-6167-7

tools. The final users prefer to "draw" the corner break line themselves, so we will classify which points are break lines and give them an attribute (for example a RGB color) in order to distinguish from the non break line points. This way, the user will notice fast and easy which ones are the candidate points and it will facilitate their task a lot. Actually there are some interesting papers about object recognition techniques[1], for different kinds of objects like feet [1] or the same way than us, "simple" surfaces [2] and [3]. The main difference between these techniques and our technique is the mathematical principle in which is based on.

2. Process

We start from a text file, which has 7 columns (X, Y, Z, Intensity, Red, Green, Blue)[2] of each point measured and registered, but the order within this file is chaotic because they are grouped based on its source scan station and it makes that these points are consecutively stored in the file. So we have to reorder them.
To put the points in order and optimize the processing time we will rearrange the point clouds in cubes, sequentially, to make closer points between them be stored in closer rows, so the search algorithms which are needed find the points faster and reduces the computational charge. There are also other available methods, like the octree-based 3D-grid used in [4].

2.1 Adjusting planes

To adjust a plane firstly we create the matrix equation system for n points where the (a^{-1}, b^{-1}, c^{-1}) is the perpendicular vector for the near points within the cube.
These sets are all the points contained into a cube. Sequentially the algorithm searches inside the cube and tries to adjust a plane using Least Squares (1), storing the parameters of the calculated plane and also the statistics.

$$x = (A^t A)^{-1} A^t K \tag{1}$$

Once the matrix equation system is solved we get the parameters which define the plane (a, b, c), and also we can get the variance estimator after the plane adjustment.

2.2 Calculating the inertia moments

The second way consists in obtaining the matrix of the inertia moments (2) for each point based on the neighbour points. In first place we will obtain the location of the mass centre and the inertia matrix. Then, we will diagonalize it, obtaining the eigenvalues and eigenvectors. The eigenvalues are the solutions to the third degree equation (3). The three solutions λ_1, λ_2, λ_3 are (4).

[1] We can get a global overview of such techniques in the R. J. Campbell and P. J. Flynn paper [5].
[2] These Received intensity and RGB colors are provided by the laser scanner, which usually has calibration errors and therefore, we cannot use them in the break line discrimination process.

 ISBN: 978-84-614-6167-7

$$I = \begin{pmatrix} I_{xx} & I_{xy} & I_{xz} \\ I_{yx} & I_{yy} & I_{yz} \\ I_{zx} & I_{zy} & I_{zz} \end{pmatrix} \tag{2}$$

$$a\lambda^3 + b\lambda^2 + c\lambda + d = 0 \tag{3}$$

Where a, b, c and d in (3) are obtained from I, and λ is any of the eigenvalues.

$$\begin{aligned}
\lambda_1 &= -\frac{b}{3a} - \frac{2^{\frac{1}{3}}(-b^2+3ac)}{3a\left(-2b^3+9abc-27a^2d+\sqrt{4(-b^2+3ac)^3+(-2b^3+9abc-27a^2d)^2}\right)^{\frac{1}{3}}} + \\
&\frac{\left(-2b^3+9abc-27a^2d+\sqrt{4(-b^2+3ac)^3+(-2b^3+9abc-27a^2d)^2}\right)^{\frac{1}{3}}}{32^{\frac{1}{3}}a} \\
\lambda_2 &= -\frac{b}{3a} + \frac{(1+i\sqrt{3})(-b^2+3ac)}{32^{\frac{2}{3}}a\left(-2b^3+9abc-27a^2d+\sqrt{4(-b^2+3ac)^3+(-2b^3+9abc-27a^2d)^2}\right)^{\frac{1}{3}}} - \\
&\frac{(1-i\sqrt{3})\left(-2b^3+9abc-27a^2d+\sqrt{4(-b^2+3ac)^3+(-2b^3+9abc-27a^2d)^2}\right)^{\frac{1}{3}}}{62^{\frac{1}{3}}a} \\
\lambda_3 &= -\frac{b}{3a} + \frac{(1-i\sqrt{3})(-b^2+3ac)}{32^{\frac{2}{3}}a\left(-2b^3+9abc-27a^2d+\sqrt{4(-b^2+3ac)^3+(-2b^3+9abc-27a^2d)^2}\right)^{\frac{1}{3}}} - \\
&\frac{(1+i\sqrt{3})\left(-2b^3+9abc-27a^2d+\sqrt{4(-b^2+3ac)^3+(-2b^3+9abc-27a^2d)^2}\right)^{\frac{1}{3}}}{62^{\frac{1}{3}}a}
\end{aligned} \tag{4}$$

Those three values (λ_1, λ_2, λ_3) after the diagonalization represent the inertia moments in the main axis directions, in a diagonal matrix with null cross moments. This diagonal is also ordered from highest to lowest moment, so we store these moments as another descriptor parameter for the geometry of the point cloud.

3. Conclusions

To determine if the point belongs to a plane wall or a corner we can use two strategies. The first one is to compare the laser scanner precision and the variance estimator after the plane adjustment in Least Squares. Establishing a threshold value to determine if the plane does not fit enough to the point set (the variance is bigger than the expected) and therefore the point cloud belongs to a corner. For a set of 49 points and a laser scanner with 6mm of precision if we are measuring a wall (plane) we expect σ values smaller than 0.0007mm.

The second one is analyzing the dimensions of the neighbour set points, projecting its components on the eigenvector axis, obtaining coordinate increments (d_x, d_y, d_z) in the particular system defined by the eigenvectors

 ISBN: 978-84-614-6167-7

calculated. Once determined the spatial extension of our data set we make the inertia moments independent of the size of the point cloud, by normalizing the module of the moment using the dimensions and the number of points in the sample (5).

$$I_{i\,Norm} = \frac{I_i}{d_x\, d_y\, d_z\, n} \tag{5}$$

Experimental tests point out it is not recommended to calculate these parameters with samples smaller than 5x5cm in X and Y because the reliability declines.
As a result of the nature of the inertia moment we will use the minimum moment (the perpendicular to the plane one) to determine, if it is bigger than the expected, that this points set belongs to a corner, if this value is small then it will be a plane.
Future investigation branches are use moments of superior order to recognize more complicated geometries.

4. Acknowledgements

This project is part of the research project "MOCAIDE", Creation and cartographic feeding of spatial data infrastructures at the Local Administration trough a data model integrating cadastre, urban planning and historic heritage, with reference CSO2008-04808 and financed by the CICYT and European funds.

5. References

[1] RAINER GRIMMER, BJÖRN ESKOFIER, HEIKO SCHLARB, JOACHIM HORNEGGER. *Comparison and classification of 3D objects surface point clouds on the example of feet*. Machine Vision and Applications 22 (2011) 235-243.

[2] PINGBO TANG, DANIEL HUBER, BURCU AKINCI, ROBERT LIPMAN, ALAN LYTLE. *Automatic reconstruction of as-built building information models from laser-scanned point clouds: A review of related techniques.* Automation in Construction 19 (2010) 829-843.

[3] MIGUEL VIEIRA, KENJI SHIMADA. *Surface mesh segmentation and smooth surface extraction trough region growing*. Computer Aided Geometric Design. 22 (2005) 771-772.

[4] H. WOO, E. KANG, SEMYUNG WANG, KWAN H. LEE. *A new segmentation method for point cloud data*. International Journal of Machine Tools & Manufacture. 42 (2002) 167-178.

[5] RICHARD J. CAMPBELL, PATRICK J. FLYNN. *A Survey Of Free-Form Object Representation and Recognition Techniques.* Computer Vision and Image Understanding 81 (2001) 166-210.

 ISBN: 978-84-614-6167-7

Proceedings of the 11th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE2011
Benidorm, Spain, 26-30 June 2011

Solving anisotropic elliptic and parabolic equations by a meshless method. Simulation of the electrical conductivity of a tissue.

M. Lucía Gavete [(1)], Francisco Vicente[(2)], Luis Gavete[(2)],

Francisco Ureña[(3)] , Juan José Benito[(4)]

[(1)]*Universidad Rey Juan Carlos. Madrid. Spain, lucia.gavete @urjc.es*
[(2)]*Universidad Politécnica de Madrid. Spain, lu.gavete @upm.es*
[(3)]*Universidad Castilla La Mancha. Spain, Francisco.urena @uclm.es*
[(4)]*U.N.E.D., Madrid. Spain, jbenito@ind.uned.es*

emails:, lucia.gavete@urjc.es, franvipa@hotmail.com,
lu.gavete@upm.es, francisco.urena@uclm.es,
jbenito@ind.uned.es

Abstract

In this paper the extension of the generalized finite difference (GFD) method to the solution of anisotropic elliptical and parabolic partial differential equations is given in the case of considering two space dimensions. The explicit finite difference formulae and the criterion of stability for the anisotropic parabolic equations are given. This has been expressed in function of the coefficients of the star equation for irregular clouds of nodes. Various anisotropic cases of simulation of cardiac tissue, including elliptic and parabolic equations have been solved and the results show the accuracy of the method.

Key words: generalized finite difference, anisotropic, simulation, tissue
MSC2000: AMS Codes (optional)

1. Introduction

 ISBN: 978-84-614-6167-7

An evolution of the method of finite differences has been the development of generalized finite difference (GFD) method that can be applied to irregular grids or clouds of points. Lizska and Orkisz [1,2] proposed a generalized finite difference method on irregular grids. Their solution was obtained using moving least squares (MLS) approximation.

Benito, Ureña and Gavete have made interesting contributions to the development of this method. The paper [3] shows explicit formulae for Generalized Finite Difference Method (GFDM) using irregular grids and the influence of parameters that define them. In the paper [4] a procedure is given that can easily assure the quality of numerical results by obtaining the residual at each point. Also, in [4], the GFD method is compared with another meshless method the, so-called, element free Galerkin method (EFG). The possibility of employing the GFD method over adaptive clouds of points progressively increasing the number of nodes is studied in [5,6]. In [7] the extension of the generalized finite difference method to the explicit solution of parabolic and hyperbolic equations is given.

In this paper the GFDM is applied for modelling electrical anisotropy of cardiac tissues. Two different anisotropic equations elliptic and parabolic are solved and the results have been compared with analytical solutions. Section 2 introduces the GFDM. In section 3 a case of anisotropic elliptic equation is solved. In Section 4 is studied the stability for the anisotropic parabolic equation. In Section 5 a case of anisotropic parabolic equation is solved to illustrate the application of the numerical explicit generalized finite difference scheme. Finally, in Section 6 some conclusions are given.

2. The Generalized Finite Difference Method

The intention is to obtain explicit linear expressions for the approximation of partial derivatives in the points of the domain to be introduced in a partial differential equation defined as:

$$L_2[U] = f \text{ in } \Omega \tag{1}$$

with the boundary conditions:

$$L_1[U] = g \text{ in } \Gamma \tag{2}$$

where $\Omega \subset R^2$ with boundary Γ, L_2 and L_1 are linear partial differential second and first order operators respectively, f and g are two known functions.

Firstly, an irregular cloud of points is generated in $\Omega \cup \Gamma$. On defining the central node with a set of nodes surrounding that node, the star then refers to a group of

 ISBN: 978-84-614-6167-7

established nodes in relation to a central node. Each node in the domain is assigned an associated star.
As star selection criterium we follow the denominated cross criterium: for example, in 2-D case the area around the central nodal point, 0, is divided into four sectors corresponding to quadrants of the cartesian coordinates system originating at the central node (see Fig. 1). Each of its semi axes is assigned to one of these quadrants. In each sector two or more nodes are selected, the closest to the origin. If this is not possible, e.g., at the boundary, missing nodes can be supplemented to provide the total number of nodes necessary in each star.

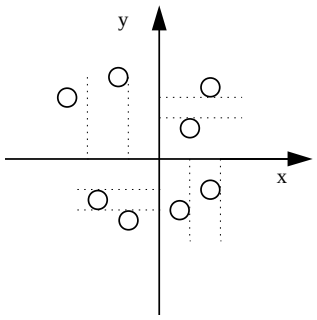


Figure 1. The four quadrants criterium, using 2 nodes in each quadrant.

If U_0 is the value of the function at the central node of the star and U_i are the function values at the rest of nodes, with i = 1,..., N, then, according to the Taylor series expansion in 2-D

$$U_i = U_0 + h_i \frac{\partial U_0}{\partial x} + k_i \frac{\partial U_0}{\partial y} + \frac{1}{2}\left(h_i^2 \frac{\partial^2 U_0}{\partial x^2} + k_i^2 \frac{\partial^2 U_0}{\partial y^2} + 2h_i k_i \frac{\partial^2 U_0}{\partial x \partial y} \right) + .. \qquad (3)$$

where (x_0, y_0) are the coordinates of the central node, (x_i, y_i) are the coordinates of the i^{th} node in the star, and $h_i = x_i - x_0$, $k_i = y_i - y_0$.

If in equation (3) are ignored the terms over the second order, an approximation of second order for the U_i function is obtained. This is indicated as u_i. It is then possible to define the function $B_5(u)$ as

$$B_5(u) = \sum_{i=1}^{N} \left[\left(\begin{array}{l} u_0 - u_i + h_i \dfrac{\partial u_0}{\partial x} + k_i \dfrac{\partial u_0}{\partial y} + \dfrac{h_i^2}{2} \dfrac{\partial^2 u_0}{\partial x^2} \\ + \dfrac{k_i^2}{2} \dfrac{\partial^2 u_0}{\partial y^2} + h_i k_i \dfrac{\partial^2 u_0}{\partial x \partial y} \end{array} \right) w(h_i, k_i) \right]^2 \qquad (4)$$

where $w(h_i, k_i)$ are the weighting functions.

 ISBN: 978-84-614-6167-7

If the norm (4) is minimized with respect to the partial derivatives, the following linear equation system is obtained:

$$A_5 D_{u5} = b_5 \tag{5}$$

The matrix $\mathbf{A_5}$ is of 5x5, and the vector $\mathbf{D_{u5}}$ is given, by:

$$\mathbf{D_{u5}} = \left\{ \frac{\partial u_0}{\partial x}, \frac{\partial u_0}{\partial y}, \frac{\partial^2 u_0}{\partial x^2}, \frac{\partial^2 u_0}{\partial y^2}, \frac{\partial^2 u_0}{\partial x \partial y} \right\}^T \tag{6}$$

From the previously obtained matrix equation (5) and by the fact that the matrix of coefficients $\mathbf{A_5}$ is symmetrical, it is possible to use the Cholesky method to solve the system. The aim is to obtain the decomposition in upper and lower triangular matrices:

$$\mathbf{A_5 = L_5 L_5^T} \tag{7}$$

The coefficients of the matrix $\mathbf{L_5}$, are denoted by $l(i,j)$.

On solving the system (5), the following explicit difference formulae are obtained

$$\mathbf{D_u}(k) = \frac{1}{l(k,k)} \left(-u_0 \sum_{i=1}^{N} M(k,i) c_i + \sum_{j=1}^{N} u_j \left(\sum_{i=1}^{5} M(k,i) d_{ji} \right) \right) (k = 1,...,5) \tag{8}$$

$$M(i,j) = (-1)^{1-\delta_{ij}} \frac{1}{l(i,i)} \sum_{k=j}^{i-1} l(i,k) M(k,j) \quad \text{with } j < i \ \ (i, j = 1\ldots,5)$$

$$M(i,j) = \frac{1}{l(i,i)} \quad \text{with } j < i \ \ (i, j = 1\ldots,5)$$

$$M(i,j) = 0 \quad \text{with } j < i \ \ (i, j = 1\ldots,5)$$

with δ_{ij} the Kronecker delta function, and

$$c_i = \sum_{j=1}^{N} d_{ji}, d_{j1} = h_j w^2, d_{j2} = k_j w^2, d_{j3} = \frac{h_j^2}{2} w^2, d_{j4} = \frac{k_j^2}{2} w^2, d_{j5} = h_j k_j w^2$$

where

$$w^2 = \left(w(h_i, k_i) \right)^2 \tag{9}$$

On including the explicit expressions for the values of the partial derivatives (8) in the equations (1) and (2), the star equation is obtained:

$$L_2(u) = -m_0 u_0 + \sum_{j=1}^{N} m_j u_j = f_0 \tag{10}$$

 ISBN: 978-84-614-6167-7

If the partial differential equations coefficients are constant and f = 0, then:

$$u_0 = \frac{1}{m_0}\sum_{i=1}^{N} m_i u_i, \qquad \sum_{i=1}^{N} m_i = m_0 \tag{11}$$

The application of the above procedure to each one of the nodes of the mesh, gives us a system of linear equations. On solving this system of equations we are provided with approximated values of the function in the nodes of the domain.

3. Numerical results anisotropic elliptic equation

A dominant operator in the monodomain and bidomain equations is the anisotropic Laplacian operator. To test the GFD method, solutions to the following anisotropic Laplace equation

$$\alpha\frac{\partial^2 U}{\partial x^2} + \beta\frac{\partial^2 U}{\partial x \partial y} + \gamma\frac{\partial^2 U}{\partial y^2} = 0 \tag{12}$$

were considered with anisotropic conductivity tensor, on an annular domain with interior radius 5 mm and exterior radius 10 mm. An analytic solution to this equation on an infinite domain can be derived [8]. For the purpose of testing, according [8] the solution was modified by the addition of constants to represent approximately physiological ranges of transmembrane potentials in a cardiac tissue. The solution was

$$U(x,y) = 5e^{0.3x}e^{-\lambda_1 0.3y}\cos(\lambda_2 0.3y) + k \tag{13}$$
$$\lambda_1 = \frac{\beta}{2\gamma} \quad , \quad \lambda_2 = \frac{\sqrt{\alpha\gamma - (\beta/2)^2}}{\gamma}$$

being k a constant and $\alpha, \beta/2, \gamma$ the conductivities of the anisotropic tensor. Two different cases were considered corresponding both to conductivities of 0.5 and 0.1 in the fiber and cross-fiber directions respectively. The difference between the two cases is the fiber angle considered: first case with a fiber angle of 45º, and second case with a fiber angle of 60º. In each case we calculate two different models with 160 and 576 nodes.

The iso-potentials of the analytic solutions within the annular region for the two anisotropic cases (45º and 60º) with 576 nodes are shown in Fig. 2 and 3. Neumann homogeneous boundary conditions have been used.

 ISBN: 978-84-614-6167-7

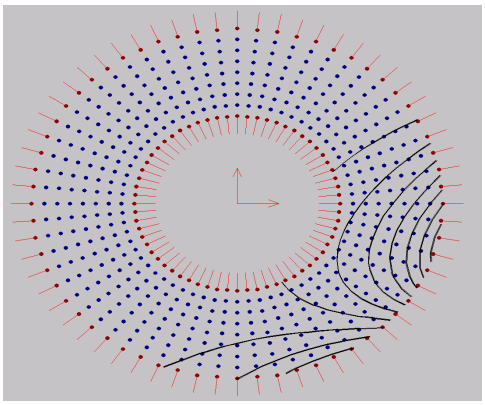


Figure 2: Iso-potentials for the 45º case.

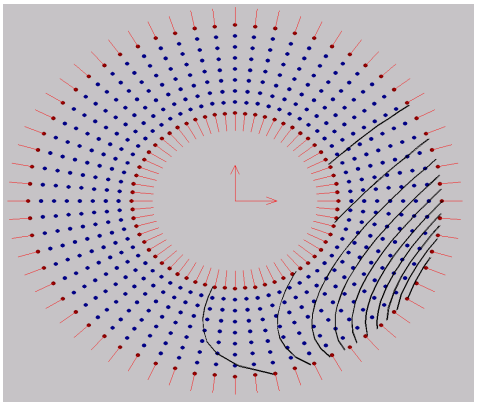


Figure 3: Iso-potentials for the 60º case.

The global relative error was calculated at all the internal points using

$$\%\text{Error}_f = \frac{100}{|f|_{max}} \sqrt{\frac{1}{N_N} \sum_{i=1}^{N_N} (f_i^{(e)} - f_i^{(n)})^2} \tag{14}$$

where f can be u, $\partial u/\partial x$,$\partial u/\partial y$, *t*he superscripts (e) and (n) refer to the exact and numerical solutions, respectively, and N_N is the total number of nodes. In the following figures 4 and 5 are shown the errors obtained for the first derivatives with the models of 160 and 576 nodes.

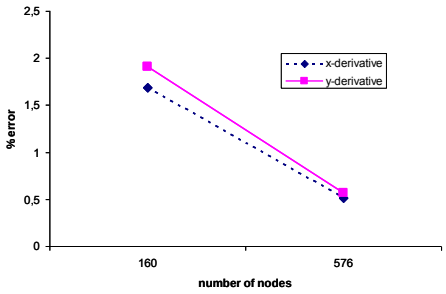


Figure 4: % Error for the 45º case.

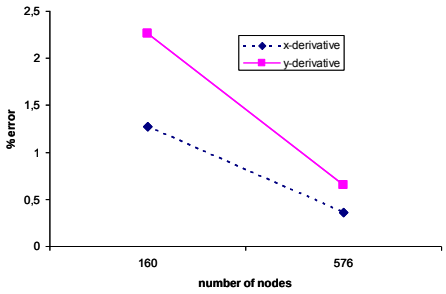


Figure 5: % Error for the 60º case.

 ISBN: 978-84-614-6167-7

4. Anisotropic parabolic equation

The equation considered in this case is:

$$\frac{\partial U}{\partial t}=\alpha\frac{\partial^2 U}{\partial x^2}+\beta\frac{\partial^2 U}{\partial x\partial y}+\gamma\frac{\partial^2 U}{\partial y^2} \tag{15}$$

with initial condition

$$U\left(\bar{x},0\right)=f\left(\bar{x}\right) \tag{16}$$

and boundary conditions

$$U=g\left(\bar{x},t\right)\quad \text{in}\quad \Gamma \tag{17}$$

being $f\left(\bar{x}\right), g\left(\bar{x},t\right)$ two known functions and $\bar{x}=x,y$.

The first derivative of U with respect to time is approached using the explicit method by the forward difference formula

$$\frac{\partial U}{\partial t}=\frac{u_0^{n+1}-u_0^n}{\Delta t} \tag{18}$$

and the spatial derivatives

$$\frac{\partial^2 U}{\partial x^2}=-a_0u_0^n+\sum_{j=1}^{N}a_ju_j^n;\frac{\partial^2 U}{\partial x\partial y}=-b_0u_0^n+\sum_{j=1}^{N}b_ju_j^n;\frac{\partial^2 U}{\partial y^2}=-c_0u_0^n+\sum_{j=1}^{N}c_ju_j^n \tag{19}$$

Substituting (18) and (19) into (15), we obtain

$$\frac{u_0^{n+1}-u_0^n}{\Delta t}=-m_0u_0^n+\sum_{j=1}^{N}m_ju_j^n \tag{20}$$

where

$$m_0=\alpha a_0+\beta b_0+\gamma c_0;m_j=\alpha a_j+\beta b_j+\gamma c_j$$

with

$$m_0=\sum_{j=1}^{N}m_j \tag{21}$$

The equation (20) can be written as

$$u_0^{n+1}=u_0^n\left(1-m_0\Delta t\right)+\Delta t\left(\sum_{i=1}^{N}m_iu_i^n\right) \tag{22}$$

 ISBN: 978-84-614-6167-7

The expression (22) relates the value of the function at the central node of the star, in time $n+1$, with the values of the functions in the nodes of the star, for a time n, multiplied by specific coefficients. This then indicates that by way of the so-called "explicit method", the value of the function in a time $n+1$ is a weighting sum of the values of the function in the star for a time n.

For the stability analysis a harmonic decomposition is made of the approximate solution at grid points at a given time level. Then, by following the von Neumann idea for stability analysis, we can write that the finite difference approximation in the central node at time n, may be expressed as

$$u_0^n = \xi^n e^{i\bar{\nu}^T \bar{x}_0} \tag{23}$$

and the finite difference approximation in the other nodes of the star

$$u_j^n = \xi^n e^{i\bar{\nu}^T \bar{x}_0} \tag{24}$$

where $\bar{\nu}$ is the column vector of the wave number, $\bar{x}$ is the vector of coordinates of central node of star and $\bar{x}_j$ is the vector of coordinates of the other nodes of star, being:

$$\bar{x}_j = \bar{x}_0 + \bar{h}_j \tag{25}$$

then $\bar{h}_j$ are the relative coordinates between the nodes of star and the central node.

On the other hand, ξ is called the amplification factor and it is in general a complex constant. If this amplification factor has a modulus greater than unity $(|\xi| > 1)$ the method is unstable.

Substituting (23) and (24) into (22), we obtain

$$\xi^{n+1} e^{i\{\nu_k\}^T\{x_0\}} = \xi^n e^{i\{\nu_k\}^T\{x_0\}}(1 - \Delta t m_0) + \Delta t \sum_{j=1}^{N} m_j \xi^n e^{i\{\nu_k\}^T\{x_j\}} \tag{26}$$

Using (21), cancellation of $\xi^n e^{i\bar{\nu}^T \bar{x}_0}$,leads to

$$\xi = (1 - \Delta t m_0) + \Delta t \sum_{j=1}^{N} m_j e^{i\bar{\nu}^T \bar{h}_j} \tag{27}$$

The complex ξ, is

$$\xi = 1 - \Delta t \sum_{j=1}^{N} m_j (1 - \cos(\bar{\nu}^T \bar{h}_j)) + \Delta t \sum_{j=1}^{N} m_j \sin(\bar{\nu}^T \bar{h}_j) \tag{28}$$

 ISBN: 978-84-614-6167-7

The real part of complex ξ

$$\begin{aligned} &-1<1-\Delta t\sum_{j=1}^{N} m_j(1-\cos(\bar{v}^T\bar{h}_j))<1 \\ &\Leftrightarrow 0<2\Delta t\sum_{j=1}^{N} m_j \sin^2\left(\frac{\bar{v}^T\bar{h}_j}{2}\right)<2 \\ &\Leftrightarrow 0<\Delta t<\frac{1}{\sum_{j=1}^{N} m_j \sin^2\left(\frac{\bar{v}^T\bar{h}_j}{2}\right)} \Rightarrow 0<\Delta t<\frac{1}{|m_0|} \end{aligned} \tag{29}$$

The imaginary part of complex ξ

$$\begin{aligned} &0<\left|\Delta t\sum_{j=1}^{N} m_j \sin(\bar{v}^T\bar{h}_j)\right|<1\Leftrightarrow \\ &0<\Delta t\left|\sum_{j=1}^{N} m_j \sin(\bar{v}^T\bar{h}_j)\right|<1\Rightarrow 0<\Delta t<\frac{1}{|m_0|} \end{aligned} \tag{30}$$

If we consider now the condition for stability

$$\left\|\xi\right\|^2 \le 1 \tag{31}$$

$$\begin{aligned} &\left[1-\Delta t\sum_{j=1}^{N} m_j(1-\cos(\bar{v}^T\bar{h}_j))\right]^2+\left[\Delta t\sum_{j=1}^{N} m_j \sin(\bar{v}^T\bar{h}_j)\right]^2 \le 1\Leftrightarrow \\ &\Leftrightarrow\left[\Delta t\sum_{j=1}^{N} m_j \sin(\bar{v}^T\bar{h}_j)\right]^2 \le \\ &\Delta t\sum_{j=1}^{N} m_j(1-\cos(\bar{v}^T\bar{h}_j))\left[2-\Delta t\sum_{j=1}^{N} m_j(1-\cos(\bar{v}^T\bar{h}_j))\right] \\ &\Leftrightarrow 0<\Delta t\left[\sum_{j=1}^{N} m_j \sin(\bar{v}^T\bar{h}_j)\right]^2 \le 2\sum_{j=1}^{N} m_j \sin^2\left(\frac{\bar{v}^T\bar{h}_j}{2}\right)\left[\frac{1}{2}\right] \\ &\Rightarrow 0<\Delta t\left[m_0\right]^2 \le |m_0| \Rightarrow 0<\Delta t<\frac{1}{|m_0|} \end{aligned} \tag{32}$$

 ISBN: 978-84-614-6167-7

The stability of a linear difference scheme comes very close to insuring that computations with the proposed scheme are practical. More precisely, what we are assured is that the stable linear generalized finite difference equations have a unique solution and that the growth of roundoff errors is bounded. In this case, the scheme has convergent approximate solutions.

5. Numerical results anisotropic parabolic equation

A dominant operator in the monodomain and bidomain equations is the anisotropic coefficients parabolic operator. To test the GFD method, solutions to the following anisotropic parabolic equation:

$$\frac{\partial U}{\partial t} = \alpha \frac{\partial^2 U}{\partial x^2} + \beta \frac{\partial^2 U}{\partial x \partial y} + \gamma \frac{\partial^2 U}{\partial y^2} \tag{33}$$

were considered with anisotropic conductivity tensor, on an irregular domain as given in Fig. 6. An analytic solution to this equation is given in this paper. The analytical solution is

$$\begin{aligned} U(x,y,t) &= e^{\frac{-A\pi^2 t}{4}} \operatorname{sen} \frac{\pi}{2}(\sqrt{\alpha}x + \sqrt{\beta}x + \sqrt{\beta}y + \sqrt{\gamma}y) \\ A &= \alpha^2 + \beta^2 + \gamma^2 + \alpha\beta + \beta\gamma + (2\alpha + \beta)\sqrt{\alpha\beta} + \beta\sqrt{\alpha\gamma} + (\beta + 2\gamma)\sqrt{\beta\gamma} \end{aligned} \tag{34}$$

being $\alpha, \beta/2, \gamma$ the conductivities of the anisotropic tensor.

We consider the case of the irregular model with 289 nodes given in Fig. 6, corresponding to anisotropic material with conductivities of 0.5 and 0.1 in the fiber and cross-fiber directions respectively with a fiber angle of 45º. Dirichlet boundary conditions have been used. The iso-potentials of the analytic solutions within the domain after 1000 time steps are shown also in Fig. 6.

In Fig. 7 we can see the % error of the function *u* according with formula (14) versus the number of time steps. The time step has been taken according with the stability condition given previously in (32).

 ISBN: 978-84-614-6167-7

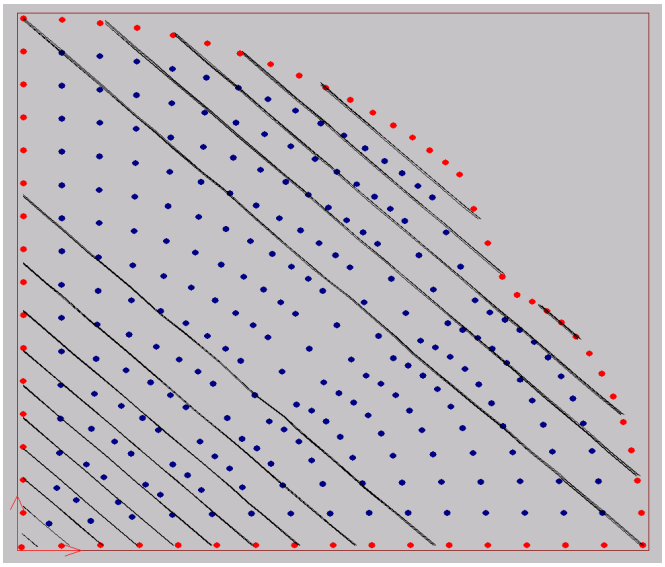


Figure 6: Iso-potentials of *u* for the model with a fiber angle of 45º after 1000 time steps.

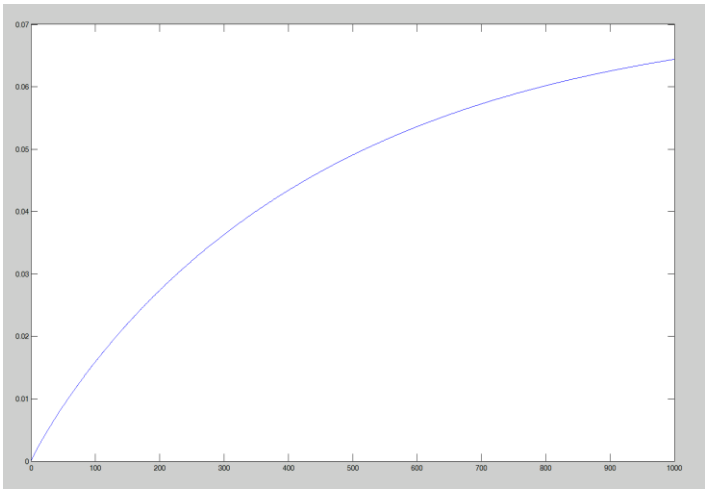


Figure 7: % Error of the function *u* versus the number of time steps (1000 time steps).

6. Conclusions

The dominant operators in the monodomain and bidomain equations are the corresponding to Laplace and Parabolic equations with anisotropic conductivity tensor. In this paper we have presented a generalized finite difference method for solving elliptic and parabolic equations in anisotropic medium. A new case of analytical solution for homogeneous parabolic anisotropic equation has been presented. Anisotropic elliptic and parabolic equations have been solved. For parabolic equations an explicit generalized finite difference method has been

 ISBN: 978-84-614-6167-7

applied. The results obtained show the high accuracy and flexibility of the method.

Acknowledgement

The authors acknowledge the support from Ministerio de Ciencia e Inovación of Spain, project CGL2008-01757/CLI.

References

[1] T. LISZKA, *An interpolation method for an irregular net of nodes*, Int. J. Numerical Methods in Engineering, 20(1984) 1599-1612.

[2] T. LISZKA, J. ORKISZ, *The finite difference method at arbitrary irregular grids and its application in applied mechanics*, Computer and Structures, 11(1980) 83-95.

[3] J. J. BENITO, F. UREÑA, L. GAVETE, *Influence of several factors in the generalized finite difference method*, Applied Mathematical Modelling, 25 **(12)** (2001)1039-1053.

[4] L. GAVETE, M.L. GAVETE, J.J. BENITO, *Improvements of generalized finite difference method and comparison with other meshless method*, Applied Mathematical Modelling, 27,10,(2003), 831-847.

[5] J.J. BENITO, F. UREÑA, L. GAVETE, R. ALVAREZ, *An h-adaptive method in the generalized finite differences*, Comput. Methods Appl. Mech. Engrg. 192, (2003), 735-739.

[6] F. UREÑA, J.J. BENITO, L. GAVETE, R. ALVAREZ, *Computational error approximation and h-adaptive algorithm for the 3-D generalized finite difference method*, International Journal for Computational Methods in Engineeering Science and Mechanics, 6(2005), 31-39.

[7] 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.

[8] M. L. TREW, B. H. SMAIL, D. P. BULLIVANT, P.J. HUNTER, A.J. PULLAN, *A generalized finite difference method for modelling cardiac electrical activation on arbitrary irregular computational meshes*, Mathematical Biosciences **198** (2005) 169-189.

 ISBN: 978-84-614-6167-7

Proceedings of the 11th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE2011
Benidorm, Spain, 26-30 June 2011

Computational Nanoscience: From Schrödinger's Equation to Maxwell's Equations

Stephen K. Gray
Center for Nanoscale Materials, Argonne National Laboratory,
Argonne, Illinois 60439

email: gray@anl.gov

The huge variety and complexity of problems that arise in nanoscience lead to a plethora of computational methods being employed to model these problems. At the most fundamental level, microscopic solutions to the time-independent and time-dependent Schrödinger equation are employed to get an idea of the dynamics of electrons and nuclei, although the large number of atoms in nanosystems generally precludes a full quantum mechanical treatment. At the continuum level, microscopic features are averaged away and computational fluid dynamics and electrodynamics are used to describe macroscopic fields interacting with various nanostructures. Sometimes it suffices for one level of theory to simply feed into another, e.g. quantum mechanics can be used to estimate potentials of interaction that are employed in larger scale classical molecular dynamics simulations. However it is also the case that there can be problems that are fundamentally multiscale in character, requiring some parts of the system to be treated at one level of theory and other parts at a different level. In this talk I give examples of all these scenarios that I draw on from my own research on nanoconfined chemistry and nanophotonics. These examples show the challenging character of modeling nanoscience problems and also illustrate some of the fascinating physical phenomena that can emerge.

As a first example, I discuss how nanoscale confinement can lead to interesting chemical and transport effects using numerical solution of the time-dependent and time-independent Schrödinger equations [1-3]. The gas-phase chemical reaction $D + H_2 \rightarrow HD + H$ is perhaps the simplest of all chemical reactions and plays a fundamental role chemical rate theory. What happens if the reactants are confined to move within a carbon nanotube

 ISBN: 978-84-614-6167-7

(CNT)? Employing a quantum based potential model for the D-H_2 system, and an empirical model for carbon-hydrogen interactions, the time-dependent Schrödinger equation was solved to obtain reaction probabilities. A mixture of discrete grids and basis functions are used to represent the wave function, and a symplectic integrator adapted for solving the time-dependent Schrödinger equation is employed. It is found that confinement can significantly enhance reaction probabilities and lower energetic thresholds, thus suggesting that CNTs can act as nanocatalysts [1]. It is also the case that CNTs and nanoporous carbon structures in general might allow one to separate isotopes of molecular hydrogen [2, 3]. The heavier mass of D_2 relative to H_2 leads to a much higher density of states associated with D_2 confined to CNTs than H_2. By determining the allowed quantum mechanical energy levels using a Lanczos iterative solution of the relevant time-independent Schrödinger equation, partition functions are constructed that allow estimates of both equilibrium and kinetic properties, affirming the possible utility of CNTs as vehicles for isotope separation. The effect, however, is significant only at relatively low (< 100K) temperatures.

In my second example, I consider light interacting with thin silver films that have periodic arrays of slits etched in them [3]. This work involves numerically solving the time-dependent form of Maxwell's equations with a system defined by materials with various frequency-dependent, complex dielectric constants. It is averaging over many quantum mechanical energy levels that lead to the dielectric constant. (In practice, empirically inferred bulk dielectric constant values are frequently used.) The base simulation method employed is the finite-difference time-domain (FDTD) method, which involves representing the electric and magnetic fields on interwoven grids and a leapfrog discrete time propagation. Nanostructures such as the silver slit arrays are relevant to chemical sensing and a figure of merit (FOM) relevant to the performance of such a sensor is the near field electric field intensity. The general problem is to vary a variety of parameters characterizing the system (e.g., film thickness, nature of substrate, slit width) and find those values that maximize the FOM. Since the FDTD calculations can be very time consuming, the Gaussian process model was used, wherein a "small" number of evaluations of the expensive "function" (an FDTD simulation) is used to construct a cheap surrogate function (a Gaussian stochastic process) and the FOM was optimized. The role of surface plasmon excitations, which manifest themselves as evanescent surface waves in the simulations, is discussed.

My third and final example concerns light interactions with hybrid systems composed of metallic nanostructures and semiconductor quantum dots (QDs) such as those that can be synthesized from Cadmium and Selenium. QDs are large (nanometer scale) relative to molecules, but have

 ISBN: 978-84-614-6167-7

discrete electronic energy levels that can be adjusted depending on how they are synthesized. The interaction of QDs with, say, metal nanoparticles (MNPs) while the system is exposed to light can lead to a variety of fascinating phenomena. There are several ways of describing such problems [5, 6]. At the simplest level, the QD is assumed to be absorbing the light of one frequency and then emitting (fluorescing) it at another frequency. The ordinary FDTD method can be used for such a calculation, with the QD being replaced by a radiating dipole. This approach is used to study a QD placed near a small array of gold bipyramidal MNPs [5]. In particular, the QD can excite "dark" plasmons, i.e. surface plasmons that cannot be excited with ordinary light. It is also possible for QDs to have the opposite effect, i.e. to lead to a transparency or a spectral region where the MNP system does not absorb or scatter light [6]. For this to occur it is necessary to allow the QD to absorb light energy, something that the simple dipole model does not include. This is accomplished by describing the region of space occupied by the QD as a polarizable medium and again carrying out ordinary FDTD calculations. Figure 1 illustrates a result when a QD is placed between two ellipsoidal silver nanoparticles. Of course a more correct description of the QD is as a microscopic object satisfying quantum mechanical equations of motion and interacting with the electromagnetic fields. One way of accomplishing this is to combine Maxwell's equations with the quantum mechanical equation of motion for the density matrix. This approach leads to the Maxwell-Bloch equations, wherein the density matrix equation involves a time-dependent, semiclassical dipole-electric field coupling term, with the electric field coming from Maxwell's equations. Maxwell's equations, in turn, have an additional current density term arising from the density matrix. The Maxwell-Bloch equations are implemented within the FDTD framework and

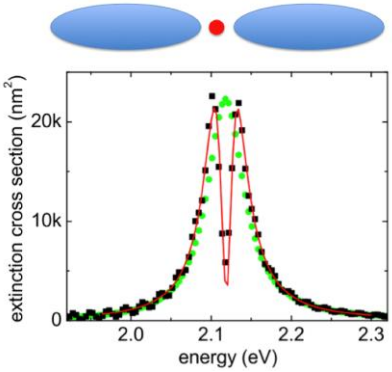


Figure 1. Schematic picture of a quantum dot placed between two silver MNPs and the calculated extinction cross section as a function of photon energy. The MNPs are 40nm long and 10nm wide, and the QD is taken to be spherical with a diameter of 4nm. Unconnected light green symbols correspond to the extinction in the absence of the quantum dot, dark black symbols correspond to the full system, and the red curve is the result of a two-oscillator analytical model developed in [6] to analyze the numerical results.

 ISBN: 978-84-614-6167-7

the QD-induced transparency is shown to also occur in the low-field limit, validating the previous calculations. However, as the field strength is increased into the strong field (Rabi flopping) regime, the transparency effect is quenched [7].

Use of the Center for Nanoscale Materials was supported by the U. S. Department of Energy, Office of Science, Office of Basic Energy Sciences, under Contract No. DE-AC02-06CH11357.

References

[1] T. LU, E. M. GOLDFIELD AND S. K. GRAY, J. Phys. Chem. C **112** (2008) 2654-2659.
[2] T. LU, E. M. GOLDFIELD AND S. K. GRAY, J. Phys. Chem. B **110** (2006) 1742-1751.
[3] M. HANKEL, H. ZHANG, T. X. NGUYEN, S. K. BHATIA, S. K. GRAY AND S. C. SMITH, . Phys. Chem. Chem. Phys. **13** (2011) 7834-7844.
[4] R. L. MILLER, M. J. DAVIS AND S. K. GRAY, J. Phys. Chem. C **114** (2010) 20741-20748.
[5] M. LU, T. W. LEE, S. K. GRAY, P. GUYOT-SIONEST, Phys. Rev. Lett. **102** (2009) 107401 (4 pages).
[6] X. WU, S. K. GRAY AND S. K. GRAY, Opt. Exp. **18** (2006) 23633-23645.
[7] L. CAO, N. F. SHERER AND S. K. GRAY, in preparation.

 ISBN: 978-84-614-6167-7

Proceedings of the 11th International Conference on Computational and Mathematical Methods in Science and Engineering, CMMSE2011 Benidorm, Spain, 26-30 June 2011

A New Predicting Method for Long-Term Photovoltaic Storage Using Rescaled Range Analysis

Samia Harrouni[1]

[1] *Instrumentation Laboratory, Faculty of Electronics and Computer, University of Science and Technology H. Boumediene (USTHB), P.O. Box 32, El-Alia, 16111, Algiers, Algeria*

email: sharrouni@yahoo.fr

Abstract

A predicting approach of long-term storage capacity for autonomous PV installations has been developed using the rescaled range analysis. The method consists mainly in establishing a mathematical law between the $(R/S)_\tau$ ratio and the time period τ. The method has been tested over one year for a PV system located in Tahifet at the huge desert of Algeria. Data used are converted solar energy which are not stationary. The experimental results show that even if the condition of stationarity is not satisfied, the rescaled range is well described by a power function of the time, this is possible by introducing a new exponent *E*. Using the power law the PV storage capacity is predicted for periods ranging from 1 to 5 years.

Key words: solar irradiation, photovoltaic storage, fractal, Hurst, rescaled range analysis, prediction

1. Introduction

The use of solar energy especially the photovoltaic one is of great interest to many applications, but the discontinuous supply of energy is in general not tolerable. In fact, a fundamental characteristic of a photovoltaic system is that power is produced only while sunlight is available.

For systems in which the photovoltaic is the sole generation source, it is necessary to employ energy storage systems to provide a reliable energy source to consumers. In any photovoltaic system that includes batteries, the batteries

 ISBN: 978-84-614-6167-7

become a central component of the overall system which significantly affect the cost, maintenance requirements, reliability, and design of the photovoltaic system.

Because of large impact of the storage system in a stand-alone photovoltaic system, the storage sizing is one of the important questions investigated to improve the efficiency of the operation of photovoltaic systems and reduce their costs.

Several approaches have been established in order to find the best way for sizing the storage capacity for autonomous PV systems [1-3]. In this paper a new approach for predicting the long-term photovoltaic storage is presented. The approach is based on the statistical method: the rescaled range analysis (R/S analysis).

2. Rescaled range analysis and Hurst exponent

R/S analysis was established by Hurst in 1951 and the same was further developed by Mandelbrot and Van Ness (1968). Hurst was a hydrologist and worked on the Nile River Dam project for about 40 years during the early years of the last century. He spent a lifetime studying the Nile and the problems related to water storage. He invented a new statistical method - the rescaled range analysis (R/S analysis) - which he described in detail in an interesting book [4].

He tried to find out the ideal features for reservoir design. An ideal reservoir should discharge certain amount of water every year and should never overflow. However, the inflow of the reservoir varies due to changes in the climatic conditions. If the inflow of the reservoir is too low then releasing fixed amount of water will make reservoir dry. Thus, he was confounded with the problem of fixing the water discharge policy, such that the reservoir will never be emptied nor it will overflow [5]. In developing such a model, Hurst studied the inflow of water from rainfall. He measured how reservoir level rises and falls around its average and recorded range of the variations.

The description of the rescaled range statistic given in the following borrows heavily from Jens Feder's book Fractals [6].

In any given year, *t*, the ideal reservoir will accept the influx $\varepsilon(t)$ from the Lake (Lake Albert taken as an example by Hurst), and a volume per year (discharge), $\langle\epsilon\rangle_\tau$, will be released from the reservoir. The average influx over the time period of τ years is:

 ISBN: 978-84-614-6167-7

$$\langle \epsilon \rangle_\tau = \frac{1}{\tau} \sum_{t=1}^{\tau} \varepsilon(t) \tag{1}$$

Let $X(t)$ be the accumulated departure of the influx $\varepsilon(t)$ from the mean $\langle \epsilon \rangle_\tau$:

$$X(t,\tau) = \sum_{u=1}^{t} \{\varepsilon(u) - \langle \epsilon \rangle_\tau\}, \quad t = 1, 2, \dots \tau \tag{2}$$

The difference between the maximum and the minimum accumulated influx X over the time period τ is the range $R(\tau)$. This range represents the storage capacity required to maintain the mean discharge throughout the period. Thus, for an ideal reservoir, R represents the difference between the maximum and minimum amounts of water contained in the reservoir. The expression of R is:

$$R(\tau) = \max_{1 \le t \le \tau} X(t,\tau) - \min_{1 \le t \le \tau} X(t,\tau) \tag{3}$$

This is illustrated in figure 1 where the range R for the lake Albert is calculated for the first 30 years.

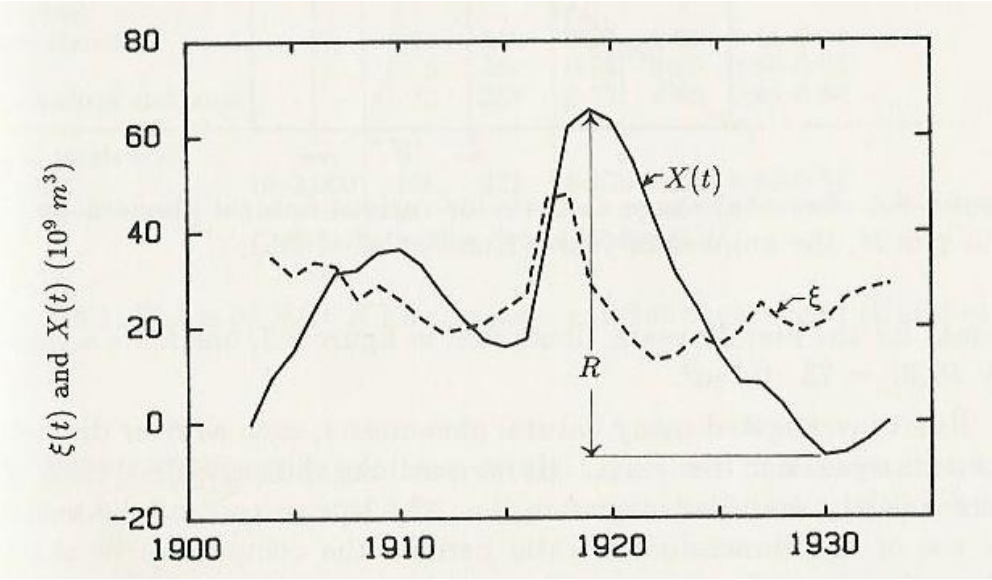


Fig. 1. Lake Albert accumulated departures from the mean discharge $X(t)$ for the first 30 years. The range is indicated by *R*. [6]

 ISBN: 978-84-614-6167-7

Clearly, we note that the range R depends on the time period τ considered and it increases with increasing τ.

In order to compare observed ranges of various phenomena, Hurst used the rescaled range R/S which is a dimensionless ratio, it is obtained by dividing the range $R(\tau)$ by the standard deviation $S(\tau)$ defined by:

$$S = \sqrt{\frac{1}{\tau}\sum_{t=1}^{\tau}\{\varepsilon(t) - \langle\epsilon\rangle_{\tau}\}^2} \tag{4}$$

For large τ the mean value of $(R/S)_{\tau}$ ratio is a power function of τ:

$$(R/S)_{\tau} = (a\tau)^{H} \tag{5}$$

where a is a constant and H is the Hurst exponent. H can be estimated as the slop of the log/log plot of $(R/S)_{\tau}$ vs. τ by using least-squares estimation:

$$\log(R/S)_{\tau} = H\,\log\tau + constant \tag{6}$$

The Hurst exponent taking values from 0 to 1 allows the measure of the time series persistence. When H = 1/2 the process does not possess long memory (has independent increments), a value $1/2 < H < 1$ means that the process is persistent (positive dependent), $0 < H < 1/2$ means antipersistency (negative dependent) of the process.

3. Application of the R/S analysis to sizing and predicting a PV storage

3.1. Application to sizing

The storage sizing method proposed in this paper is based on the R/S analysis described above. In this method the PV storage (batteries) is assimilated to the water reservoir studied by Hurst. Hence, the determination of the ideal PV storage capacity size requires the estimation of the energy deficit which represents the difference between solar radiation input and energy demand on long-term period.

For a given day, *d*, the PV installation will accept a global irradiation *I(d)*, the PV generator will then convert this energy in $E_g(d)$ according to equation below:

 ISBN: 978-84-614-6167-7

$$\boldsymbol{E_g(d) = \eta(d) I(d) S} \tag{7}$$

In this relation, S is the PV generator area, $\eta(d)$ is the daily efficiency of the PV generator.

The energy supplied by the installation to the load is assumed to be the mean converted energy $\langle E_g \rangle_\tau$ for the studied period τ. It is expressed as:

$$\langle E_g \rangle_\tau = \frac{1}{\tau} \sum_{t=1}^{\tau} E_g(t) \tag{8}$$

The accumulated difference between the converted energy and the energy demand for τ days is expressed as:

$$D(t,\tau) = \sum_{u=1}^{t} \{E_g(u) - \langle E_g \rangle_\tau\} \tag{9}$$

The difference between the maximum and the minimum accumulated energy $D(t,\tau)$ is the range $R(\tau)$ which can be identified to the maximum size of PV energy storage.

$$R(\tau) = \max_{1 \le t \le \tau} D(t,\tau) - \min_{1 \le t \le \tau} D(t,\tau) \tag{10}$$

3.2. Application to predicting

According to the Hurst empirical law equation (Eq.5) we can take the logarithm of both the sides of the equation, we then obtain:

$$\log(R/S)_\tau = H \log\tau + H \log a \tag{11}$$

Using the least-squares estimation to fit Eq. (11) we obtain the estimation values of H and a. Knowing these parameters, one can predict the possible future value of the adjusted range $(R/S)_n$ for any period n.

To estimate the range $(R)_n$ the adjusted range must be multiplied by the standard deviation $(S)_n$. This later is taken to be equal to the greatest value of S over the period τ.

 ISBN: 978-84-614-6167-7

4. Implementation

4.1. Data description

The data used to implement this method are daily global irradiation $I(d)$ recorded during the year 1992 on a tilted surface (10°) in Tahifet located in the south of Algeria (latitude = 22° 53' north, longitude = 6° east and altitude = 1400m).

The PV system used in our experimentation is a stand-alone 720 Wc photovoltaic power installation operating during the studied year at Tahifet. The PV generator is composed of 16 monocrystalline silicon modules with a total area of 6m².

Figure 2 shows the annual evolution of the monthly mean of daily converted energy E_g, This later is calculated according to Eq.(7).

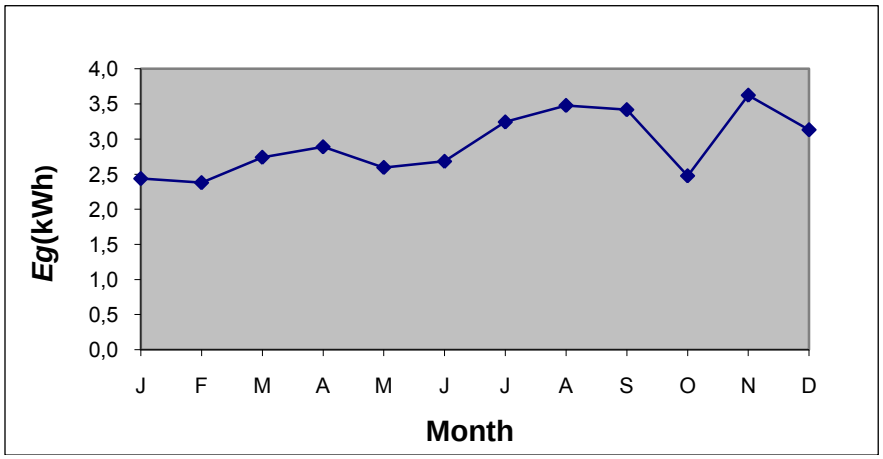


Fig. 2. Monthly mean of daily converted energy.

4.2. R/S analysis

The converted energy series $(E_g(1), E_g(2), \ldots, E_g(N))$, ($N$ =366) is divided into m consecutive non-overlapping subseries of length $\tau = N/m$: $E_g\big((k-1)\tau+1\big)$, $E_g\big((k-1)\tau+2\big), \ldots, E_g\big((k\tau)\big)$, k = 1, 2, …, m. The values of τ in our study range from 10 to $N/2$.

For each subseries the mean:

$$\langle E_g \rangle_{k,\tau} = \frac{1}{\tau}\sum_{t=1}^{\tau} E_g\big((k-1)\tau + t\big) \tag{12}$$

 ISBN: 978-84-614-6167-7

and the standard deviation:

$$S_{k,\tau} = \sqrt{\frac{1}{\tau}\sum_{t=1}^{\tau}\left\{E_g\left((k-1)\tau + t\right) - \langle E_g\rangle_{k,\tau}\right\}^2} \tag{13}$$

are calculated. Then the energy deficit resulted from accumulating the difference between the converted energy and the mean which represents the load consumption is determined:

$$\left(D_{k,\tau}\right)_t = \sum_{u=1}^{t}\left\{E_g\left((k-1)\tau + t\right) - \langle E_g\rangle_{k,\tau}\right\}, t = 1, 2, \ldots\ \tau \tag{14}$$

and the range, $R_{k,\tau}$, for the subseries is calculated

$$R_{k,\tau} = \max_{1\le t\le\tau}\left(D_{k,\tau}\right)_t - \min_{1\le t\le\tau}\left(D_{k,\tau}\right)_t \tag{15}$$

We finally obtain the rescaled range $(R/S)_\tau$ by averaging $R_{k,\tau}/S_{k,\tau}$ over the m subseries:

$$(R/S)_\tau = \frac{1}{m}\sum_{k=1}^{m}\left(R_{k,\tau}/S_{k,\tau}\right) \tag{16}$$

5. Results and discussion

The assessment of our PV storage sizing is carried out using experimental data resulting from monitoring the PV system with a data acquisition. Thus, for values of τ ranging from 10 to $N/2$, where N is the series size, the accumulated energy $D(t,\tau)$ which must be stored in the battery is first computed according to Eq.(14). The resulting curve for the first 366 days is shown in figure 3.

 ISBN: 978-84-614-6167-7

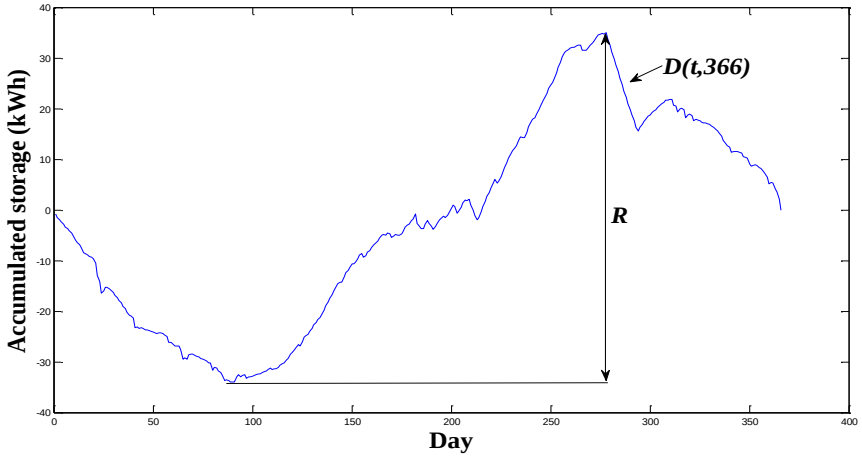


Fig. 3. Accumulated storage $D(t,\tau)$ for the first 366 days. The range is indicated by R.

Then, for each τ the range is estimated for the different subseries. The mean values of R for some periods τ are given in Table 1.

τ	R(kWh)
10	1.42
20	2.84
30	4.17
40	5.37
50	7.09
60	9.03
70	9.75
80	12.34
90	11.52
100	14.62
110	16.07
120	17.05
130	18.17
140	20.47
150	26.12
160	28.11
170	31.03
180	34.11

Table 1. The range R for different values of τ

 ISBN: 978-84-614-6167-7

One can clearly note that the range depends on the time period τ, R increases with increasing τ.

In order to perform the prediction of the PV storage the $(R/S)_\tau$ versus τ plot is developed, this is shown in Figure 4. It is apparent from the figure that Most of $(R/S)_\tau$ points seems to lie along straight line, the rescaled range, is then a power function of τ.

A trend line is fitted amongst the logτ versus log$(R/S)_\tau$ plot and the equation of the trend line is computed. The slope coefficient of the equation gives the estimation of H which is found equal to 0.97, the value of a parameter is then 0.28.

According to the R/S analysis theory the slope of regression line is identified to the Hurst exponent which is an index of the long memory of the time series. For self-affine time series H is also an index of their roughness which is the role of the fractal dimension.

For a curve, fractal dimension is between 1 and 2, it approaches 2 if it is extremely irregular and tends towards 1 if it is more regular. Since the Hurst exponent H is related to the fractal dimension D by the relation: $D = 2 - H$, we deduced that low values of H means the curve is smooth and the high values indicate the irregularity of the curve.

Observing the converted energy series used in our study, we found that the corresponding curve is irregular, consequently, the corresponding fractal dimension must be high and the Hurst exponent low.

To confirm this, the fractal dimension is calculated for these data using a method we already developed [7-8], we found D=1.80, as a result H=0.2. This later is very far from the slope of the R/S plot regression line obtained (0.97). This is can be explained by the non-stationarity of the studied series, since the stationarity of the series is an essential prerequisite for the rescaled range analysis. It is then obvious that the slope of the fitted line found is different from the Hurst parameter.

Let's recall that the purpose of this work is the PV storage prediction and not the Hurst exponent estimation. Therefore, and to avoid any confusion the slope of the line we found is noted E instead of H.

Hence, the rescaled range, R/S, for the solar energy converted series studied is described by the following relation:

 ISBN: 978-84-614-6167-7

$$(R/S)_\tau = (0.28\,\tau)^{0.97}$$

Table 2 illustrates some predicted values of the adjusted range (R/S) for different time periods n (1 to 5 years) and the corresponding range R which is identified to the maximum size of the PV storage. Note that the medium life duration of PV batteries is 5 years.

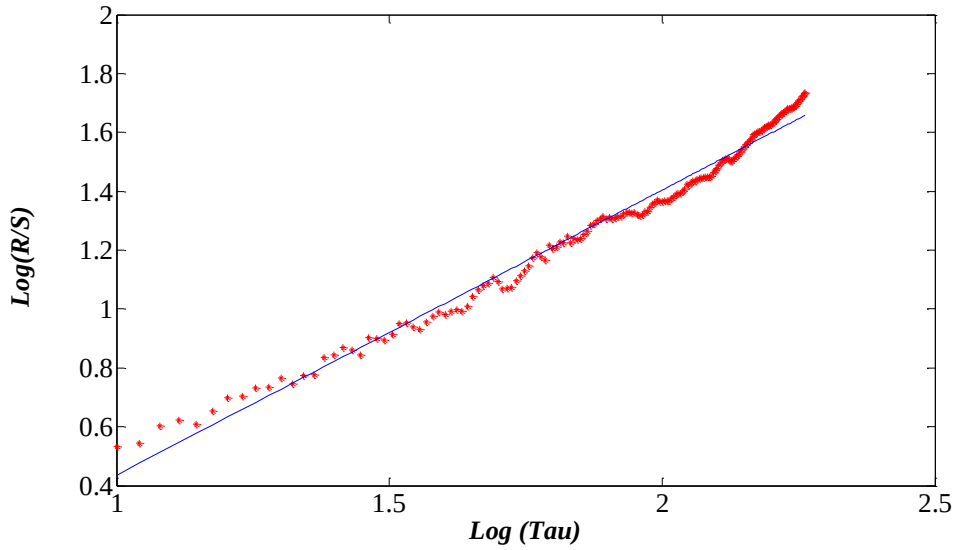


Fig. 4. Log R/S versus Log (τ) Plots

n (years)	R/S	R(kWh)
1	89.19	58.87
2	174.48	115,16
3	258.44	170.57
4	341.55	225.42
5	424.26	280.01

Table 2. Predicted values of the R/S and the corresponding R for a period ranging from 1 to 5 years

 ISBN: 978-84-614-6167-7

6. Conclusion

In this paper, a method for predicting a photovoltaic storage at long-term is presented. It was shown that the rescaled range analysis applied to the solar converted energy data provides an alternative way to size and predict the PV storage. Results presented here confirm that although the data studied are non-stationary the rescaled range is well described by a power function of the time. A new exponent *E* different from the Hurst exponent has been introduced. Results presented here confirm that the PV storage can be predicted using the power law. The method is still under investigation and further experimentation is needed to validate the preliminary results presented in this paper.

References

[1] R. N. CHAPMAN, *Development of sizing nomograms for stand-alone photovoltaic / storage systems*, Solar Energy. **43** (1989) 71-76.

[2] A. HADJ ARAB, B. AIT IDRISSS, R. AMIMEUR AND E. LORENZO, *Photovoltaic systems sizing for Algeria*, Solar Energy. **54** (1995) 99-104.

[3] A. MAAFI ET C. DELORME, *Modélisation à long terme et optimisation du stock d'énergie des installations solaires autonomes*, *J. Phys. III France*. **6** (1996) 511-527.

[4] H. E. HURST, R.P. BLACK AND Y. M. SIMAIKA, *Long-term storage: An experimental study*, Constable, London, 1965.

[5] S. K. MITRA, *Trends in Stock Prices and Range to Standard Deviation Ratio*, I.J. Busi. Mana. **6** (2011) 223-234.

[6] J. FEDER, *Fractals*, Plenum Press, 1988.

[7] S. HARROUNI, *Fractal Classification of typical meteorological days from global solar irradiance: Application to five sites of different climates*, Modeling solar radiation at the earth's surface, Berlin, 2008.

[8] S. HARROUNI AND A. GUESSOUM, *Using fractal dimension to quantify long-range persistence in global solar radiation*, Chaos, Solitons & Fractals. **41** (2009) 1520-1530.

 ISBN: 978-84-614-6167-7

Proceedings of the 11th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE2011
Benidorm, Spain, 26-30 June 2011

Secure Universal Protocol for E-Assessment

Andrea Huszti [1], Zita Kovács [2]

[1]*University of Debrecen, Faculty of Informatics and Hungarian Academy of Sciences and University of Debrecen*
[2]*University of Debrecen, Faculty of Informatics*

emails: huszti.andrea@inf.unideb.hu, kovacs.zita@inf.unideb.hu

Abstract

We propose a universal protocol that possesses all necessary security requirements, such that eligibility, secrecy, anonymity, individual and global verifiability and can be applied for e-exam, e-pollster, e-auction, e-tender and e-vote, as well. Our solution employs blind signature and secret sharing schemes, digital envelope technique and provides receipt on a bullten board.

Key words: e-vote, e-auction, e-exam, e-pollster, e-tender
MSC2000: AMS Codes (optional)

1. Introduction

Nowadays more and more people use computers to manage their everyday life. We submit our tax report, we give our opinion to on-line questions. There are students, who attend to on-line courses, hence receive lessons, tests via Internet.

There are several electronic systems that require people to fill out a simple or sometimes rather complex questioner and later all the answers are evaluated. We assume that these questioners might contain multiple choice tests and write in questions, as well. In case of write in questions human correctors are needed for evaluation. We will call these systems e-assessment systems. E-assessment systems bring up serious security issues. Security requirements for different assessment systems are sometimes similar. We propose a universal protocol that possesses all necessary security requirements and can be applied for e-exam, e-pollster, e-auction, e-tender and e-vote, as well.

 ISBN: 978-84-614-6167-7

2. Security requirements

Considering e-assessment systems in general security issues should be considered carefully. There are certain requirements that any kind of e-assessment system should possess, such that *eligibility, individual and global verifiability.* Most of them require *anonymity* of participants, or *secrecy* of answers.

One of the most important requirements is *anonymity*. In case of e-assessment systems anonymity of individuals not only protects them from unpleasantness, but makes the final result more authentic. There are situations, when anonymity after the evaluation phase should be revoked**.** Anonymity revocation is necessary for e-exam, e-auction and e-tender systems. In case of e-pollster the company that ordered the survey might be interested in the list of respondents in order to check the pollster company, whether it really asked the individuals, but answers should not be linked to them.

Another important requirement is in e-assessment schemes that only eligible individuals are allowed to participate. The systems should possess anonymity and *eligibility* at the same time. Hence after verifying eligibility of a participant, he is anonymous during the whole process, but we should know that each anonymous participant is eligible and does not frame another person.

Individual and global verifiability for an e-assessment system is also essential. A participant is able to verify that his opinion is counted properly in the final result. Global verifiability means that any observer can verify that the assessment is processed well.

Necessary requirements are *data integrity* of questions, answers and the final result and *secrecy* of partial results in case of e-vote, e-tender systems, questions and answers should be secret for e-exams.

In all cases the protocol should protect against *framing attacks*, when a participant answers the questions instead of another one.

3. Our solution

The proposed scheme meets the security requirements listed above, applying several cryptographic tools. Anonymity of participants is provided by blind signatures [1] and a trusted anonymous server that exchanges IP addresses of senders. Our system gives the possibility of anonymity revocation with the help of servers that securely share the corresponding secret key.

 ISBN: 978-84-614-6167-7

We employ digital envelope technique to provide confidentiality of answers. We apply secret sharing technique to store secret decryption key for the envelope. Electronic bulletin board is applied for individual and global verifiability and also for announcing the final result.

E-voting, e-auction schemes in the literature [2,3] usually assume that voters or bidders have certificates. Our goal is to provide a design that can be implemented easily and respondents do not have to possess certificates

The proposed scheme consists of four stages: *registration, operational, processing, announcement stages*. During registration stage Registry verifies eligibility of an individual, if a participant is eligible, than receives a pseudonym. In the operational stage an individual answers the questions or sends the next bid, this can happen several times. All the messages that are sent by individuals are time-stamped. Always the last message before the deadline is counted. In the processing stage evaluation of answers happens either automatically or by a corrector. After evaluation the final result is made public in the announcement stage and in case of e-exam, e-auction and e-tender schemes anonymity is revoked.

The proposed scheme is a generalization of a system that is implemented in frame of project GOP-1.1.2-07/1-2008-0001. The first author is supported by TÁMOP 4.2.1-08/1-2008-003 project. The project is implemented through the New Hungary Development Plan co-financed by the European Social Fund, and the European Regional Development Fund and also by the Hungarian National Foundation for Scientific Research Grant No. K75566. The authors are partially supported by the project GOP-1.1.2-07/1-2008-0001.

4. References

[1] D. CHAUM, *Blind signatures for untraceable payments*, Advances in Cryptology: Proceedings of Crypto, (1983), 199-204

[2] A. HUSZTI, *A Secure Electronic Voting Scheme*, Periodica Polytechnica Electrical Engineering **51/3.-4**., (2007), 141-146.

[3] NGUYEN HOANG ANH, *Survey on e-auction protocols*, MTAT.07.006 Research Seminar in Cryptography, (2008)

 ISBN: 978-84-614-6167-7

Proceedings of the 11th International Conference on Computational and Mathematical Methods in Science and Engineering, CMMSE2011 Benidorm, Spain, 26-30 June 2011

Mobility Management Scheme for the integration of Internet of Things in the HIMALIS ID/Locator Split Future Internet Architecture avoiding the Identity Attack

Antonio J. Jara[1] and Antonio F. G. Skarmeta[1]

[1] *Dept. Information and Communications Engineering, Computer Science Faculty, University of Murcia, Spain*

emails: jara@um.es, skarmeta@um.es

Abstract

Future Internet is characterized by its capacity to support the mobility of the entities connected to the network. Mobility offers, from the services and user experience point of view, the capacity for extending and adapting the network to changes of location and infrastructure, and also offers, from the network point of view, the increase of the fault tolerance capacity, connectivity, dependability and scalability. The mobility support for the current Internet has presented limitations mainly caused by the role of IP address as both node ID for session determination in the application/transport layer, and Locator in the network layer. For that reason, the proposals for the Future Internet are focused on the ID/Locator split architectures. The presented work is focused on analyze the security challenges for one of these new architectures (HIMALIS), since ID/Locator management messages are potentially dangerous and present several vulnerabilities. This analysis also considers the particularities from the Internet of Things, since it is a pillar of the Future Internet. Finally, this works proposes a secure and scalable mobility management scheme considering the requirements and constrains from the Internet of Things, the proposed scheme is based on Return Routability and ECC-based asymmetric cryptography, in order to support scalable inter-domain authentication and secure location update and binding transfer for the mobility process avoiding the identity attack.

Key words: Security; mobility; Authentication protocol; ID/Locator Split; Internet of Things.

 ISBN: 978-84-614-6167-7

1. Introduction

There has been a tremendous increase of the use of Internet, from the 360 million of users in 2000 to 1.6 billion of Internet users and 4 billion of mobile users with over 570 million Internet-enabled handheld devices, and this is only the beginning, since it is estimated that the extension of the Internet to smart things, will reach by 2020 between 50 to 100 billion devices connected to Internet, such as described by Sundmaeker [1], defining the so-called Internet of Things, described by Atzori [2].

This grows leads to serious problems for scalability, manageability, addressing/identity, and robustness. In addition, the openness and ubiquity features of the current Internet present problems to offer a suitable support for security, privacy, and secure mobility. Therefore new redesign of the Internet architecture and definition of new protocols are required to solve the mentioned problems for the Future Internet of Things. For this purpose, several projects from industrial and international collaboration are being carried out to define the Future Internet architecture which solves the limitations of the current Internet architecture [3]. The architecture considered under this paper takes into account the possibility of that one of the edge network is based on Internet of Things, i.e. an IP-Based Wireless Sensor Networks, such as the defined by 6LoWPAN [4] under the frame of the Internet Engineering Task Force (IETF). 6LoWPAN extends Wireless Sensor Networks (WSN) to Internet, adding to IEEE 802.15.4 a layer to support IPv6.

An architecture for the Future Internet of Things needs to support security, mobility and interoperability for the heterogeneity of the network providing a scalable universal integration among the different technologies from the current deployments and solutions, such as sensors, RFID tags and legacy technologies in a common domain. In addition, other issues need to be considered such as a resolution infrastructure which supports distributed look-up and discovery for resources and services, context-awareness, reliability, energy management, self-management, self-configuration and self-healing properties [5],

This paper focuses on satisfying the security and mobility requirements. For that purpose, this presents a novel scheme to support authentication and allows to extend the trust domain for mobile devices. Previously to this work, it was proposed a suitable security stack based on Elliptic Curve, which has been optimized for embedded IoT devices [6]. Over the mentioned stack has been defined a scalable mechanism to extend the trust domain to entities and devices which are changing its position in the network.

Mobility support for small and smart devices is one of the most important issues in the Future Internet of Things, since it is utilized for realizing many applications. In addition, mobile communication increases the fault tolerance

 ISBN: 978-84-614-6167-7

capacity of the network, increases the connectivity between nodes and clusters, allows to extend and adapt network to changes of their location and environment infrastructure. The mentioned features are requirements to satisfy the dependability and scalability of the Future Internet [7]. The mobility support in the proposed architecture is based on ID/Locator split, since the main currently limitations to support mobility are caused by the role of IP address as both ID for session determination in the application/transport layer, and Locator in the network layer [7]. For that reason, in order to solve the mentioned mobility problem, it is considered split between ID and Locator, for the Future Internet architecture. This issue is being extensively discussed in the Internet Engineering Task Force (IETF) and ITU-T, in order to design efficient solutions for mobility, multi-homing, routing, security as well as for integrating heterogeneous network layer protocols. The mobility supports has been already addressed for the Future Internet in [7] and [8]. These approaches do not consider any security issues for the mobility management, and several vulnerabilities have been found for the management of the location, for example a malicious host might be able to establish false updates of the location, thereby preventing some packets from reaching their intended destination, diverting some traffic to the intruder, or flooding third parties with unwanted traffic.

For that reason, this paper presents the security analysis, and extends the mentioned proposal with a novel secure and scalable mobility management scheme, where the requirements and constrains from the Internet of Things are being considered, in order to support edge networks based on IP-based Wireless Sensor Networks (6LoWPAN).

Finally, for the validation of the proposed mobility and authentication scheme, this has been verified with the tools OFMC, and ATSE from the Automated Validation of Internet Security Protocols and Applications (AVISPA) framework, which is described deeply in [9], and the visual tool Security Protocol Animator for AVISPA (SPAN).

2. Future Internet Architecture

The architecture considered, over which is evaluated and defined our scheme is based on a ID/Locator split. This defines two different values for IDs and Locators. Thereby, this allows to the network layer changes its Locator when the device move i.e. changes its position in the network without requires the upper layers change the ID. Thereby, this ensures that the established communication sessions associate with the ID are not interrupted because mobility, see [7]. The basic entities for the ID/Locator management, such as, for example, found also in the Mobile Oriented for Future Internet Architecture (MOFI) in [8], where a mapping agent exists between Name, Device ID and Locator. The reference architecture considered for this work is the HIMALIS (Heterogeneity Inclusion and Mobility Adaptation through Locator ID Separation) architecture.

 ISBN: 978-84-614-6167-7

The diagram of the architecture is shown in the Fig. 1.

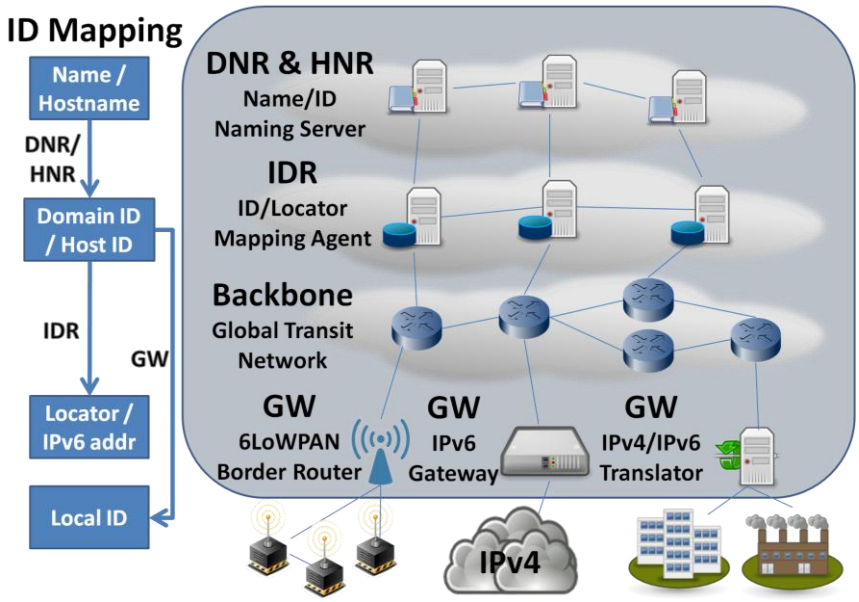


Fig. 1. Future Internet architecture (HIMALIS).

The components are:

2.1 Signalling Control Network

1) Domain Name Registries (DNR) and Hostname Registries (HNR): These offer a mapping between the hostname and domain name to the Device ID (ID), such as presented in the Fig. 1.

- Domain name represents "whose it is", are usually denoted by Uniform Resource Locator (URL) such as the used for the WEB, e.g. lab.cs.um.es. The location of the domain manager is carried out by the DNR, which locates to the HNR where is managed the domain.
- Device name/Hostname represents "who it is", are usually denoted by variable-length string e.g. Uniform Resource Name (URN), or a human readable and remembered names such as a Network Access Identifier (NAI), e.g. sensor1@lab.cs.um.es, which represents a sensor1 inside the domain presented previously.
- Device ID/Host ID represents "what it is", ID is used as control information in communication protocols and packet headers to identify sessions, packets and a communication end points usually for sensors a globally unique ID of 128-bits, it can be based on the identifiers defined such as Overlay Routable Cryptographic Hash Identifiers (ORCHID), or Host Identity Protocol (HIP) for

 ISBN: 978-84-614-6167-7

improving the security. For example HIP uses a public key and its hash value as ID.

2) ID registries (IDR): This offers a mapping between the Device ID (ID) and Locator (LOC), which is used to represent the location of an object in the network, it is usually used the IPv6 address e.g. HIP uses IPv6 address as Locators. The mapping relation is presented in the Fig. 1.

3) Mapping local level Device ID/Link ID (Gateway): This translates between the Device ID and Local ID (LID), which is used to identify a link connection / interface. LID is needed for data packet delivery between gateway and the end node of the edge network, when the end device is not supporting a global Locator, such as for example IEEE 802.15.4 networks without 6LoWPAN support, where Local ID is defined by short 16-bits address in network layer, or another kind of local identification. In these cases the Locator can be addressed to the GW.

2.2 Global Transit Network /Core IPv6 Backbone

This is a collection of networks which interconnect to the public organizations, research centres, and end users through Internet Service Providers around the entire world. This is composed by routers, backbones, servers and agents of some of the entities mentioned from the signalling control network.

2.3 Edge networks

The edge networks provide network access to the end systems such as hosts and clients. Example of these networks can be any of the current industrial networks such as Control Area Network (CAN), vehicular network and hospitals. These networks are connected to the Global Transit Network through one or more gateways such as IPv4 to IPv6 translators for networks which are not adapted, IPv6 gateways and 6LoWPAN Border Routers.

3. Analysis of Vulnerabilities and Threats

This section describes some of the major threats found in the management of the mobility and architectures presented in the Section 2. In order to understand the vulnerabilities, it is initially described how the original mobility management over the architecture from the Fig. 1. is carried out.

The messages defined are considering the guidelines defined in [7], but it has been extended since originally are not in this level of detail presented. The messages are:

1, 2- Router Solicitation (RS) and Router Advertisement (RA) in order to join to the gateway from the Neighbour Discovery protocol.

 ISBN: 978-84-614-6167-7

3- ID Update (IDU). MN requests to the F-GW that it updates its ID with the new locator to the other parts of the architecture such as Home IDR (H-IDR), Home Gateway (H-GW) and Corresponding Gateway (C-GW).
4- Location Update. F-GW sends this message to its IDR (F-IDR), in order to carry out the related operations to update the MN ID with its new locator.
5, 6, 7- Device ID Lookup of the IDR for the MN previously to the movement.
8, 9- Home Registration, F-IDR contacts with H-IDR to update its mapping table with the new locator and transfer the binding. Message 9 carries out the mapping table update.
10, 11- Location Update confirmation to the F-IDR and the F-GW.
12- ID Update ACK (IDA), F-GW confirms to MN, that the update of its ID with the new locator is being processed.
13- Binding Request. At the same time, when are sent the ACKs from 10, 11 and 12, H-IDR continues with the process. This message requests to the H-GW to transfer the binding to the gateway in the foreign network (F-GW).
14- De-Registration: H-GW removes from its table of nodes to the MN, since it has left its network.
15- Transfer the binding. H-GW transfers the messages pending to be delivered and other information to the F-GW.
16, 17- Location Update. H-GW sends this messages to the corresponding gateway (C-GW) and (C-IDR) to inform about the new locator.
18, 19- Location Update ACK, C-GW and C-IDR confirm to H-GW the location update.

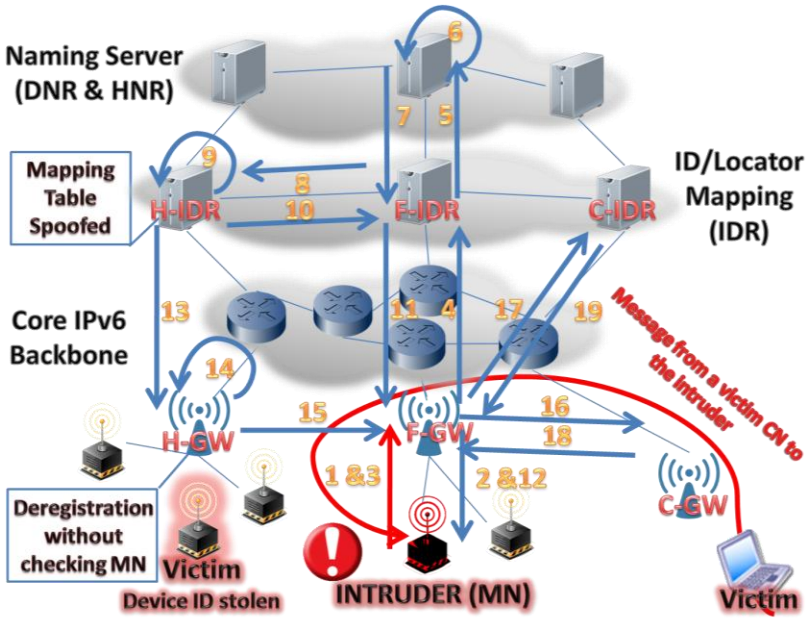


Fig. 2. Attack with a false IDU, when MN is an intruder.

The main security vulnerabilities appear with the management of the messages which make changes in the mapping tables from the different architecture entities,

 ISBN: 978-84-614-6167-7

such as Location Update, ID Update, and Home Registration. The goal of an intruder could be to corrupt the Correspondent Node's (CN) binding cache to provoke wrong delivery of packets, which can cause Denial-of-Service (DoS) and compromises the privacy and integrity of the communications. The intruder could also exploit features of the mentioned messages for manage location to exhaust the resources of the parties from the architecture, etc. These vulnerabilities need to be solved to ensure the safe status of the network and satisfy the security requirements from the applications. Specifically, the attacks considered is versus the identity property, since Locator is volatile, the unique real property from a device is its ID. Therefore, the most obvious danger in the mobility management is the steals of the Device ID. For example, an intruder could claim the Locator of a victim node with a determined Device ID, and then steal traffic destined to the victim node. This follows the similar conception of address stealing from MIPv6.

The basic approach to steal Device ID in the presented architecture, where the location management messages are not authenticated, is with the ID Update (IDU) message, which starts to the process. This attack is presented in the Fig. 2., where in the message 3 the intruder sends a spoofed IDU to the gateway identifying itself with a stolen Device ID. This vulnerability is presented since MN authenticity is not verified. Therefore, this attack is proposed to be solved authenticating the MN with its H-GW through its IDR, in order to verify the veracity of the indicated Device ID, which is indicated in the IDU message. The solution to this is presented with the proposed scheme.

A second approach to steal the Device ID, could be based on false Location Update message, which are originally sent between F-GW and C-GW (see message 16 in Fig. 2). An intruder could sham to be a GW, which is able to inform to another gateway (victim GW) about a new Locator for the Device ID stolen to a victim node. This attack is based on the communication between gateways, sending a false Location Update with the information of the node which is being its Device ID stolen. This vulnerability arises because the GW, which receives the false Location Update, does not check the authenticity of the indicated new mapping Device ID/Locator with the H-IDR. The solution proposed to solve this attack lays in that C-GW (victim GW) checks the received Device ID/Locator mapping by the F-GW, in the Location Update, with the defined in the H-IDR from the original owner of the Device ID. This real Locator for the indicated Device ID is recovered through the "Locator discovery" functionality, which offers the architecture through the IDR, HNR and DNR entities, for Locator, and Device ID/Device name respectively. “Locator Discovery” technique is used in proposed secure and scalable mobility management scheme presented, and it is feasible to make it secure with the protocol presented in the Section 4, and analyzed with AVISPA tool.

 ISBN: 978-84-614-6167-7

4. Secure and Scalable Mobility Management Scheme

Our proposal is a scheme to allow the authentication of the mobile nodes in the visited networks, in order to reach a secure mobility management of the mobile nodes in the architecture, and the trust extension for the optimization of the binding transfer avoiding Triangular Routing, and make this more scalable.

The main goals pursued by the proposed scheme are:

1- Verify that a visited domain can be trusted.
2- Authenticate to the Mobile Node.
3- Extend the trust domain to the visited domain.

These goals are reached with a scheme based on:

1- Lightweight cryptographic suite based on ECC.
2- DoS countermeasure and nodes authentication with Return Routability.
3- Trust extension based on Diffie-Hellman, and a technique to establish keys similar to the tickets used in Kerberos, where a new session key is sent encrypted with a key which is only by the end node (client) and provider.

This can find a wide range of solutions for key exchange and security for the Internet domain, but they are not designed to consider the resource constrained networks of smart things, which will build the Future Internet of Things. We have collected the most relevant features of them, and adapted to the architecture considered in the Section 2, and the 6LoWPAN mobility management requirements.

The security protocols make an effort to reduce the cryptographic cost of the required public-key-based key exchanges and signatures with ECC; this reduction is highly desirable for IoT domain. Some of the protocols, which already have defined its version based on ECC, are Transport Layer Security (TLS) in RFC5246, Internet Key Exchange (IKE) and IKEv2 in RFC 5903, and Host Identity Protocol (HIP) Diet Exchange (DEX), described by Moskowitz in [10]. In addition, ECC cryptographic primitives have been optimized for 6LoWPAN devices, and in particular for the operations used in the proposed protocol.

The proposed protocol includes a Return Routability mechanism such as Denial-of-Service attacks countermeasure, since this mechanism delay the establishment of the Binding Update Confirmation and establishment of the new tunnels (trust extension) until that the Home Gateway (H-GW) and Mobile Node (MN) are verified. This Return Routability countermeasure is used in protocols such as MIPv6, DTLS, IKEv2, HIP, and Diet HIP. The effectiveness of this countermeasure depends on the routing topology defined in the global transit network; in our solution the messages used for Return Routability are the same as required for the confirmation of the update, therefore it is not really meaning a significant overload.

 ISBN: 978-84-614-6167-7

Additionally, the protocol proposed provides end-to-end security, which is required to make scalable the integration of the Internet of Things in the Future Internet, support the secure individual communication and authentication in machine to machine communications (M2M), which are defined and considered within the Internet of Things domain. For that reason, we also define the trust extension with two new tunnels to allow the trustable communication, on the one hand, between the Mobile Node and the visited gateway (F-GW), and on the other hand, between the previous gateway (H-GW) and the new gateway (F-GW), to avoid the inefficient route through the IDRs. Thereby, this allows a safe and direct communication for the communications during the "roaming" of the mobile node.

Finally, the establishment and authentication of the Mobile Node through the visited gateway (F-GW) is carried out with a technique, similar to the defined in Kerberos, where the H-GW sends a key encrypted end-to-end, which is forwarded by intermediate nodes which cannot understand it. See message 8 in the Fig. 3.

Previously to describe the protocol proposed and presented in the Fig. 3, the secure elements (shared keys) considered and established during the process are:

- SK0: is the shared key between the MN and the Home GW (GWA), which was established in the bootstrapping phase.
- SK1: is the new shared key established between the Home GW (GWA) and the new visited GW (GWB), which is established with Diffie-Hellman, through the trust chain defined via the IDRs.
- SK2: is the new shared key established between the Mobile Node and the new visited GW (GWB), which is established with a technique based on Kerberos, since ESK2 is SK2 only deciphered with SK0 by the MN.
- TunA, TunB and TunC: are the safe channels (tunnels) defined among the IDRs and the GWs; they build the denominated trust chain. They were also established in the bootstrapping phase.

The messages exchanged are:

1- MN to GWB - IDU.{S}_SK0. This sends the IDU message with the fields related with new Locator, Device ID, etc. and the signature (S) based on SK0 for authentication with the GWA (which follows the role of Home Agent in other mobility protocols)
2- Generation of the Part B of SK1 (PB), for the Diffie-Hellman Key exchange.
3- GWB to IDRB - {PB.IDU.Sciphered}_TunA
IDRB to IDRA - {PB.IDU.Sciphered}_TunB
IDRA to GWA - {PB.IDU.Sciphered}_TunC
This is the exchange of PB to the GWA, where Sciphered is S ciphered with SK0, i.e. {S}_SK0.
4- Verification of the authenticity of the MN by the GWA by way of the signature S.
5- Generation of the Part A of SK1 (PA), for the Diffie-Hellman Key exchange.

 ISBN: 978-84-614-6167-7

6- Return Routability Process
This sends the new timestamp (T1), PA, SK2 and ESK2, through the trust chain (TunA, TunB and TunC).
GWA to IDRA - {T1.PA.SK2.ESK2}_TunC
IDRA to IDRB - {T1.PA.SK2.ESK2}_TunB
IDRB to GWB - {T1.PA.SK2.ESK2}_TunA
This also sends the same message directly.
GWA to GWB - {T1.PA.SK2.ESK2}_SK1.

7- GWB verifies with the messages from step 6 the MN authenticity, and establish the new tunnel with SK1 between GWA and GWB, which is used to transfer the binding, pending messages and future packets.

8- GWB to MN - {{T1}_SK2.ESK2}. GWB sends the SK2 encapsulated with ESK2 (Kerberos technique which has permitted to send SK2 encrypted end-to-end, being forwarded by the intermediate nodes of the trust chain which cannot understand it). In addition, the timestamp T1 is sent encrypted using SK2 in order to avoid reply attack. Finally The MN verifies with this message that the GWA has authenticated to it, and establish the new tunnel with GWA based on SK2.

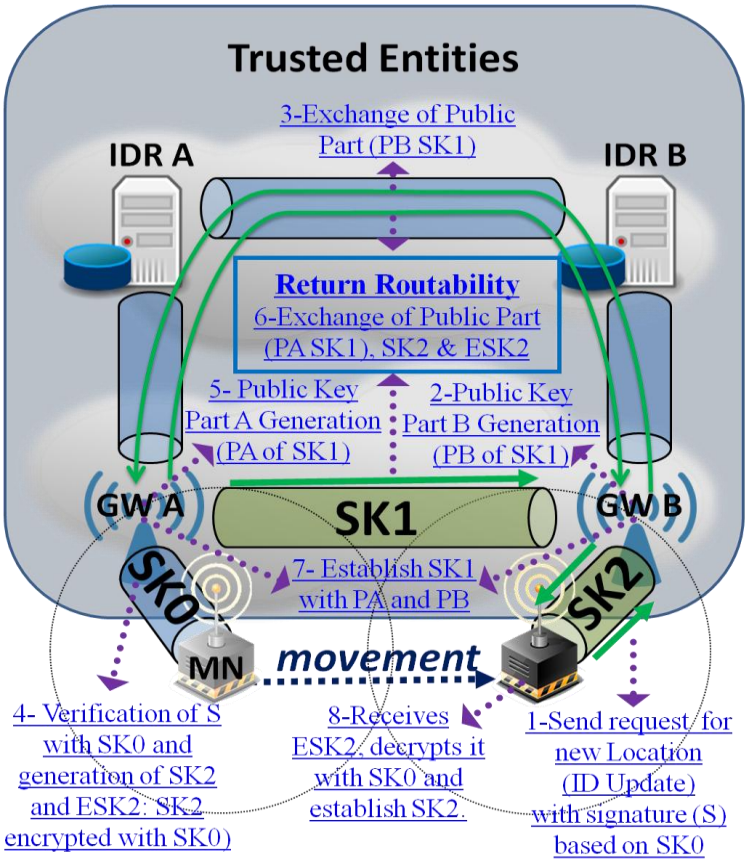


Fig. 3. Kernel of the secure mobility management scheme.

 ISBN: 978-84-614-6167-7

5. Evaluation and Validation

The MN intruder attack is solved with the authentication of the MN with its H-GW through its IDR, in order to verify the veracity of the indicated Device ID indicated in the IDU message. This verification is carried out with the protocol presented in the Section 4, where the extension of the Return Routability (RR) process for the presented architecture, allows the verification of the MN is trustworthy by the F-GW. In this approach, the vulnerabilities found in other previous applications of RR such as in MIPv6, where the communication between Home Agent and Correspondent Node is vulnerable, are not found in our proposal, since the message sent by the alternative path is encrypted with the exchanged key through the safe path, converting this initially unsafe path, also in trustable.

Finally, the proposed scheme and the MN and GW intruder attacks have been validated with the AVISPA tool, it has been validated on the one hand the basic protocol, and on the other hand, the basic protocol together with sessions where the GW is an intruder, and the MN is an intruder. The security issues considered are:

1- Secrecy of the keys: SK0, TunA, TunB, and TunC.
2- Authentication of the MN with the GWA through the signature (S), based on SK0.
3- Finally, it has been verified the well forming of the new trust links between the MN and the new GW (SK2) and the previous and new gateway (SK1).

The results are presented in the Table 1:

Version	***Tool***	***Description***	***Result***
Basic session	OFMC	Visited Nodes: 50 nodes Depth: 10 plies	SAFE
Basic session	ATSE	Analyzed: 214 states Reachable: 55 states	SAFE & goal as specified
Basic session + MN Intruder	OFMC	Visited Nodes:23246 nodes Depth: 15 plies	SAFE
Basic session + MN Intruder	ATSE	Analyzed: 2022179 states Reachable: 120601 states	SAFE & goal as specified
Basic session + GW Intruder	OFMC	VisitedNodes: 80573 nodes Depth: 15 plies	SAFE
Basic session + GW Intruder	ATSE	Analyzed: 504369 states Reachable: 66219 states	SAFE & goal as specified

Table I. AVISPA Tool Results

 ISBN: 978-84-614-6167-7

6. Conclusions and Future Works

The scheme presented in this paper, proposes a solution to solve the threats and vulnerabilities found in the HIMALIS architecture. The scheme proposed offers a safe and scalable scheme to support mobility, and extend the trust in the architecture with the determination of shared keys between entities which require a direct exchange of information. This is based on Diffie-Hellman key exchange, Return Routability, and other techniques such as the tickets from Kerberos to establish a key between different domains.

7. Acknowledgments

This work has been supported by the Autonomous Region of Murcia (CARM, Spain), with funds for Science and Technology Program (PCTRM 07/10), through the research project Hydrological Modelling in Semiarid Zones, the grants from the Foundation Séneca (04552/GERM/06), and the Spanish Ministry of Science and Education with the FPU program grant (AP2009-3981).

8. References

[1] Sundmaeker, H.; Guillemin, P.; Friess, P.; Woelfflé, S. (2010). Vision and Challenges for Realising the Internet of Things. European cluster CERP-IoT, European Union, ISBN: 978-92-79-15088-3.

[2] Atzori, L.; Iera, A.; Morabito. G. (2010). The Internet of Things: A survey. Computer Networks. Vol. 54, No. 15, pp. 2787-2805.

[3] Papadimitriou, D.; Tschofenig, H.; Rosas, A.; Zahariadis, S.; et al. (2010). Fundamental Limitations of Current Internet and the path to Future Internet, European Commission, FIArch Group, Ver. 1.9.

[4] Kushalnagar, N.; Montenegro, G.; Hui, J.; Culler, D. (2007). Transmission of IPv6 Packets over IEEE 802.15.4 Networks, RFC 4944.

[5] Zorzi, M.; Gluhak, A.; Lange, S.; Bassi, A. (2010). From today's INTRAnet of things to a future INTERnet of things: a wireless- and mobility-related view", Wireless Communications, IEEE, 17(6), pp.44-51.

[6] Ayuso, J.; Marin, L.; Jara, A.; Skarmeta, A.F.G. (2010). Optimization of Public Key Cryptography for 16-bits Devices based on 6LoWPAN, 1st Int. Workshop on Security of the Internet of Things, (Tokyo), Japan.

[7] Kafle, V.P.; Otsuki, H.; Inoue, M. (2010). An ID/locator split architecture for future networks, Comm. Magazine, IEEE, vol. 48, No. 2, pp. 138-144.

[8] Heeyoung Jung; Seok-Joo Koh. (2010). Mobile-Oriented Future Internet (MOFI): Architecture and Protocols (Release 1.2), Electronics Telecommunications Research Institute (ETRI).

[9] Viganò, L. (2005). Automated Security Protocol Analysis with the AVISPA Tool. XXI Mathematical Foundations of Programming Semantics (MFPS'05), ENTCS 155:61--86, Elsevier.

[10] Moskowitz, R. (2010). HIP Diet EXchange (DEX), IETF draft-moskowitz-hip-rg-dex-02 (work in progress).

Proceedings of the 11th International Conference on Computational and Mathematical Methods in Science and Engineering, CMMSE2011 Benidorm, Spain, 26-30 June 2011

Theoretical study and formation predication of ultra-cold alkali dimer CsFr

I. Jendoubi[1], H. Berriche[1, 2] and H. Ben Ouada[1]

[1] *Laboratoire de Physique et Chimie des Interfaces, Département de Physique. Faculté des Sciences de Monastir, Avenue de l'Environnement, 5019 Monastir, Tunisia*

[2] *Physics Department, College of Science, King Khalid University, P. O. Box 9004, Abha, Saudi Arabia.*

emails: hamid.berriche@fsm.rnu.tn, hamidberriche@yahoo.fr

Abstract

The adiabatic potential energy curves, the spectroscopic constants (R_e, D_e, T_e, ω_e, $\omega_e\chi_e$ and B_e) and the permanent and transition dipole moments of the $^1\Sigma^+$ lowest electronic states of the CsFr molecule dissociating into Cs (6s, 6p, 5d, 7s) and Fr (7s, 7p, 6d, 8s, 8p, 7d) have been performed. We have used an *ab initio* approach based on pseudopotential, parameterized l-dependent polarization potentials and full configuration interaction calculations. Our spectroscopic constants of the ground state are in good agreement with the available theoretical works.

Key words: Pseudopotential, Potential energy, Spectroscopic constants, Dipole moments.

MSC2000: AMS Codes (optional)

1. Introduction

The new development of cold and ultra-cold atomic trapping techniques has increased the interest in experimental and theoretical investigation on homonuclear and heteronuclear alkali molecular systems [1]. Such an important theoretical and experimental effort is motivated by the possible applications such as controlling of ultracold chemical reactions [2-4], ultracold molecular collision dynamics [5-8], quantum computing [9, 10] and experimental preparation of few-

 ISBN: 978-84-614-6167-7

body quantum effects [11] (such as Efimov states) where the aim is to prepare molecules in definite quantum states with respect to the center of mass, electronic, rotational and vibrational motions [12]. In fact, it is suspected that the long-range dipole-dipole interaction will introduce new important physical phenomena. Furthermore, the growing development of the laser cooling and optical trapping technique demands an accurate knowledge of the spectroscopic proprieties of the alkali dimers.

Several theoretical studies of the electronic structure for some of these systems have been performed in our group for LiCs [13], NaCs [14] and LiNa [15, 16]. Other authors were also interested in studying alkali dimers such as CsLi [17], KCs [17] and NaCs [17-20] and NaK [21]. Recently, an extensive ab initio study on heteronuclear mixed alkali pairs LiNa, LiK, LiRb, LiCs, NaK, NaRb, NaCs, KRb, KCs and RbCs has been performed by Aymar and Dulieu [19, 22] in order to produce accurate potential energy and permanent dipole moment for the ground state and lower triplet states. Aymar et al. [23] have investigated the possibility of creating cold CsFr molecule through the photoassociation of cold atoms. In this study we have calculated the adiabatic potential energy curves of the first 10 electronic states of $^1\Sigma^+$ symmetry dissociating into Fr (7s, 7p, 6d, 8s, 8p, 7d) and Cs (6s, 6p, 5d, 7s) for a dense and large grid of internuclear distances ranging from 6.6 to 85 a. u . Their spectroscopic constants have been extracted and compared with the theoretical works when available.

This paper is organized as follows. In section 2, we present a brief summary of the *ab initio* approach. Section 3 is devoted to the presentation of our results. Finally, we summarize our conclusion in section 4.

2. Method of calculations

Cs atom is treated through a one electron non-empirical pseudopotential proposed by Barthelat and Durand. [24-25]. The Fr ionic core is described through a semi-local *l*-dependent pseudopotential (PP) [26-27] taken from the recent work of Aymar et al. [23]. In addition, Core-polarization and core-valence interactions are considered using l-dependent core-polarization potentials following the formalism of Foucrault et al. [28]. For Cs atom, we use a 6s/4p/4d Gaussian basis set taken from Ref. 29; while for the Fr atom we used a 9s/9p/9d Gaussian basis set [23]. These authors have used two basis sets (named A and B), two dipole polarizabilities $\alpha_d^{M^+}$ and two sets of cutoff radii (ρ_l) parameters introduced in the effective core-polarization potential for Fr. In our calculation we have chosen to use the basis *A* as it gives better agreement with the experimental atomic limits for the Francium atom. The used core dipole polarizability for Cs^+ and Fr^+ ions are, respectively, 15.117 and 20.380 a_0^3. The cutoff radii for the lowest one-electron valence s, p and d orbitals are,

 ISBN: 978-84-614-6167-7

respectively, 2.690, 1.850 and 2.810 a. u. for Cs [29] and 3.16372, 3.045 and 3.1343 a. u. for Fr [23]. In order to test the quality of the used basis sets, we have performed a one-electron SCF atomic calculation for Cs and Fr atoms. A good agreement between our theoretical atomic energy levels and the experimental ones is observed. Such accuracy will be transmitted to the molecular calculations. Having only two valence electrons, the CsFr diatomic alkali molecule is studied by full CI calculations. The CI calculation is performed using the CIPCI algorithm of the Laboratoire de Physique et Chimie Quantique of Toulouse. The potential energy curves of the CsFr molecule are performed for a large and dense grid of interatomic distances varying from 6.6 to 85 a. u.

3. Results

The first ten $^1\Sigma^+$ potential energy curves of CsFr dissociating into Cs (6s, 6p, 5d, 7s) and Fr(7s, 7p, 6d, 8s, 8p, 7d) have been calculated for a large and dense grid of intermolecular distance ranging from 6.6 to 85 a. u. They are displayed in Fig.1 and the spectroscopic constants (R_e, D_e, T_e, ω_e, $\omega_e x_e$ and B_e) of the first eight states are presented in table 1 and compared with the available theoretical studies.

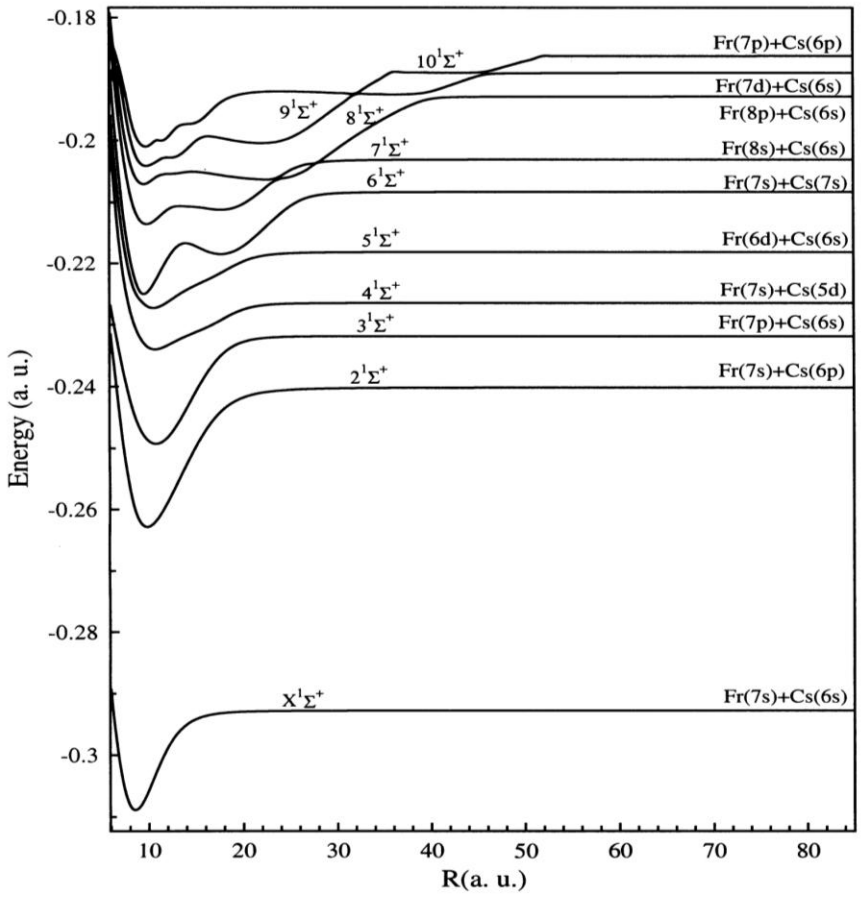
Table 1: Spectroscopic constants of the $^1\Sigma^+$ states of CsFr molecule

State	R_e (Å)	D_e (cm^{-1})	T_e (cm^{-1})	ω_e (cm^{-1})	$\omega_e x_e$ (cm^{-1})	B_e (cm^{-1})	Ref.
X $^1\Sigma^+$	4.523	3537		37.33	0.098	0.0098	This work
	4.523	3553					[23]
2 $^1\Sigma^+$	5.238	4971	10116	28.71	0.041	0.0073	This work
3 $^1\Sigma^+$	5.735	3813	13087	21.82	0.031	0.0061	This work
4 $^1\Sigma^+$	5.677	1651	16438	17.98	0.049	0.0062	This work
5 $^1\Sigma^+$	5.582	2010	17906	19.64	0.048	0.0064	This work
6 $^1\Sigma^+$							
1st min	5.047	3672	18404	38.61	0.101	0.0079	Thiswork
2nd min	9.375	2237	19839				
(7) $^1\Sigma^+$							
1st min	5.232	2309	20916	25.66	0.071	0.0073	This work
2nd min	9.391	1790	21435				
8 $^1\Sigma^+$							
1st min	5.058	3147	22347	26.64	0.056	0.0079	This work
2nd min	11.947	2991	22503				

 ISBN: 978-84-614-6167-7

Figure 1 presents the potential energy curves of the first ten $^1\Sigma^+$ states. An important general behaviour is observed for the higher excited states. It corresponds to the existence of undulations, which leads to double and, sometimes, to triple wells. We observe the existence of avoided crossings, which can be explained by the ionic interaction between Cs^+Fr^- and Cs^-Fr^+ arrangements. This phenomenon can be observed distinctly in the the permanent dipole moment [31]. We remark that the dipole moment of these states, one after the other, behaves as a +R function and then drops to zero at particular distances corresponding to the avoided crossings between the two neighbour electronic states. If these curves are combined, they produce piecewise the whole R function due to the ionic character of these states. The discontinuities between the consecutive parts are due to the avoided crossings. The dipole moment of CsFr reveals an electron transfer yielding a Cs^+Fr^- instead of Cs^-Fr^+ ionic character as it is expected and as observed for all other mixed alkali pairs where the electron is transferred towards the lighter species. For the $7^1\Sigma^+$, $9^1\Sigma^+$ and $10^1\Sigma^+$ states the same linear behaviour is observed but with a -R function [31]. This is related to the second ionic character of these states related to Cs^-Fr^+ arrangement.

Figure 1: Adiabatic potential energy curves of the ten lowest $^1\Sigma^+$ states of CsFr.



 ISBN: 978-84-614-6167-7

4. Conclusion

The adiabatic potential energy curves of the lowest electronic states of $^1\Sigma^+$ symmetry and their spectroscopic constants for the CsFr molecule dissociating into Fr(7s, 7p, 6d, 8s, 8p, 7d) and Cs (6s, 6p, 5d, 7s) have been performed using *ab initio* approach based on the pseudopotentials technique, parameterized l-dependent polarization potentials and full configuration interaction calculations. A good agreement between our spectroscopic constants R_e and D_e for the ground state and those of the theoretical work of Aymar et al. [23] is observed. The spectroscopic properties of the higher excited $^1\Sigma^+$ states are reported here for the first time. A complete study of this molecule including adiabatic and diabatic potential energy and dipole function for all symmetries is in progress and will be published in details [31].

Acknowledgment

We gratefully acknowledge support of this work by King Abdul Aziz City for Science and Technology (KACST) through the Long-Term Comprehensive National Plan for Science, Technology and Innovation program under Project No. 08-NAN148-7.

References

[1] TOPICAL ISSUE ON ULTRACOLD POLAR MOLECULES: FORMATION AND COLLISIONS, Eur. Phys. J. D **31** (2004).
[2] N. BALAKRISHNAN AND A. DALGARNO, Chem. Phys. Lett. **341** (2001) 652.
[3] E. BODO, F. A. GIANTURCO AND A DALGARNO, J. Chem. Phys. 116 (2002) 9222.
[4] T. ROM, T. BEST, O. MANDEL, A. WIDERA, M. GREINER, T. W. HANSCH AND I. BLOCH, Phys. Rev. Lett. **93** (2004) 073002.
[5] M. KAJITA, Eur. Phys. D **23** (2003) 337; Eur. Phys. D **31** (2004) 39.
[6] R. V. KERMS, Int. Rev. Phys. Chem. **24 (**2005) 99.
[7] A. V. AVDEENKOV, M. KAJITA AND J. L. BOHN, Phys. Rev. A. **73** (2006) 022707.
[8] R. V. KERMS, Phys. Rev. Lett. **96** (2006) 123202.
[9] D. DEMILLE, Phys. Rev. Lett. **88** (2002) 067901.
[10] S. F. YELIN, K. KIRBY AND R. COTE, PREPRINT QUANT-PH/0602030.
[11] T. KRAEMER, M. MARK, P. WALBURGER, C. CHIN, B. ENGESSER, A. D. LANGE, K. PILCH, A. JAAKKOLA, H. C. NAGERL, R. GRIMM, Nature. **440** (2002) 315.
[12] S. D. KRAFT, P. STAANUM, J. LANGE, LVOGEL, R. WESTER AND M. WEIDEMULLER, J. Phys. B: At. Mol. Opt. Phys. **39** S993–1000.
[13] N. MABROUK, H. BERRICHE AND F. X. GADEA, AIP Conf. Proc, **2** (2007) 23.
[14] N. MABROUK AND H. BERRICHE. AIP Conf. Proc, **1148** (2009) 326.

 ISBN: 978-84-614-6167-7

[15] N. MABROUK, W. ZRAFI, H. BERRICHE AND H. BEN OUADA, Lect. Ser. Comp. Comput. Sc. **7** (2006) 1451.
[16] N. MABROUK AND H. BERRICHE, J. Phys. B: At. Mol. Opt. Phys. **41** (2008) 155101.
[17] M. KOREK, A. R. ALLOUCHE, K. FAKHREDDINE, AND A. CHAALAN, Can. J. Phys. **78** (2000) 977.
[18] M. KOREK, S. BLEIK AND A.R. ALLOCHE, J. Chem. Phys. **126** (2007) 124313.
[19] M. AYMAR AND O. DULIEU, J. Chem. Phys, **122** (2005) 204302.
[20] G. IGEL-MANN, U. WEDIG, P. FUENTEALBA, AND H. STOLL, J. Chem. Phys. **84** (1986) 5007.
[21] S. MAGNIER, M. AUBERT-FRECON, PH. MILLIE, J. Mol. Spect. **200** (2000) 96.
[22] M. AYMAR AND O. DULIEU, J. Chem. Phys. **125** (2006) 047101.
[23] M. AYMAR, O. DULIEU AND F. SPIEGELMAN, J. Phys. B: At. Mol. Opt. Phys. **39** (2006) 905-927.
[24] J C BARTHELAT AND P. DURAND, Theor. Chim. Acta. **38** (1978) 283.
[25] J C BARTHELAT AND P. DURAND, *Gazz. Chimi. Ital.* **108** (1975) 225.
[26] P. DURAND AND J. C. BARTHELAT, Chem. Phys. Lett. **27** (1974) 191.
[27] P. DURAND AND J. C. BARTHELAT, Theor. Chim. Acta. **38** (1975) 283.
[28] M. FOUCRAULT, PH. MILLIÉ AND J. P. DAUDEY, J. Chem. Phys. **96** (1992) 1257.
[29] D PAVOLINI, T GUSTAVSSON, F SPIEGELMAN AND J-P DAUDEY, J. Phys. B: At. Mol. Opt. Phys. 22 (1989) 1721.
[30] YU RALCHENKO , F C JOU D E KELLEHER , A E KARMIDA , L MUSGROVE, J READER, W L WIESSE AND K OLSEN, 2007 *NIST Atomic Spectra Database* (version 3.1.2) (Online) available: http:// physics.nist.gov/asd3 (2007, June 26) (Gaithersburg, MD, National Institute of Standards and Technology).
[31] I. Jendoubi and H. Berriche (unpublished)

 ISBN: 978-84-614-6167-7

Proceedings of the 11*th International Conference*
on Computational and Mathematical Methods
in Science and Engineering, CMMSE2011
Benidorm, Spain, 26-30 June 2011

Hub Detour Routing in Future Mobile Social Networks

Sangsu Jung, Boram Jin, and Okyu Kwon
Division of Fusion and Convergence of Mathematical Sciences,
National Institute for Mathematical Sciences, S. Korea

emails: {ssjung, boramjin, okw}@nims.re.kr

Abstract

Challenges in large-scale future mobile social networks (MSNs) are congestion mitigation and robust message-passing. As the scale of a social community increases, a mobile network based on the community forms a scale-free network (SFN). In a SFN with proximity or density based routing, a hub of high betweenness centrality is frequently utilized in forwarding messages of other nodes. Thus, traffic congestion occurs in the vicinity of a hub. Furthermore, a hub failure critically affects network-wide performance. With the enhancement of an ant colony algorithm, we develop a hub detour routing scheme for future MSNs. Our scheme autonomously distributes loads and minimizes the impact of a specific node in message-passing. Mathematical analyses describe the properties of our scheme. In addition, simulation results exhibit autonomous and robust behaviors of our scheme compared with a conventional ant colony routing scheme.

Key words: mobile social network, scale-free network, hub, routing, ant colony algorithm

1. Introduction

A mobile social network (MSN) [1] is an emerging networking structure boosted by increases in hands-on communication devices such as smart phones and tablet PCs. In this network, mobile users with the same interest form a social community on a social network service (SNS) such as Foursquare [2], Gowalla [3], or MyCubee [4]. A conventional online social network (OSN) [5] (e.g. Bebo [6], MySpace [7], or Facebook [8]) is based on wired connections with little mobility. In an OSN, the roles of users in message-passing are insignificant because communications between users are from data stored in servers of service

 ISBN: 978-84-614-6167-7

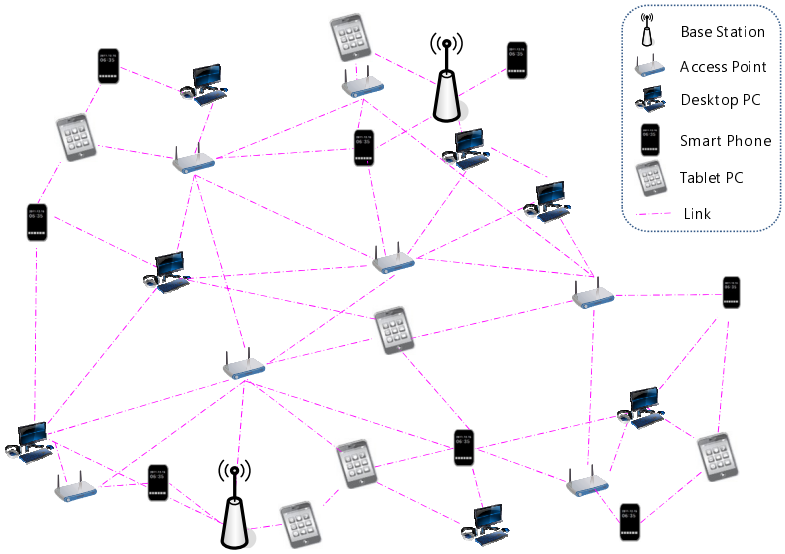


Figure 1: An example of a future heterogeneous multi-hop MSN architecture.

providers. On the other hand, the users in a MSN distinctively characterize communications with contents in their mobile devices. Currently, data transfers rely on one-hop communications via cellular networks, wireless networks, and direct communications by Bluetooth between mobile devices. However, we expect that MSNs will be based on heterogeneous multi-hop networks incorporating ad-hoc networks of mobile devices in the near future as shown in Figure 1. Because this framework can save the resources of cellular and wireless networks occupied by smart phone and tablet PC users, it has an advantage in mitigation of data explosion [9].

According to [10], as the size of a social network increases, the network is characterized as a scale-free network (SFN). In a SFN, there exists a hub, which has relatively many connections with other nodes and plays a significant role to forward the packets of other nodes. The works of [11] and [12] report that routes determined by greedy routing based on node density are heuristically the shortest paths. That is, in a SFN with proximity or density based routing, a hub is most frequently chosen for relaying messages of other nodes. Thus, a hub encounters traffic congestion. Further, a hub is exposed to malicious attacks so that it has risk to become a means to propagate viruses or worms to an entire network.

MSNs are constructed in a decentralized manner. Previously, an ant colony algorithm is proposed to design autonomous routing for distributed ad-hoc networks [13]. Because it requires no centralized controller, it is suitable for a dynamic network topology. However, a traditional ant colony algorithm greedily

 ISBN: 978-84-614-6167-7

finds the shortest paths so that it dominantly utilizes a specific node, which limits achievable total network throughput. In this paper, we propose an enhanced ant colony algorithm which achieves an autonomous traffic control for MSNs or SFNs in a distributed manner. The contributions of our work are as follows: *we provide a new pheromone assignment function which can control the amount of pheromones in a specific node or hub*; *we mathematically analyze the property of our scheme*; *we observe the behaviors of our scheme and compare it with a conventional ant colony routing scheme through simulations*.

2. Related Work

Many ant colony based routing schemes such as AntHocNet [14], ARA [15], DAR (Distributed Ant Routing) [16], or HOPNET [17] have been proposed for wireless ad hoc networks. In these schemes, each link has a pheromone trail by ants (packets) and routing paths are established through links of high pheromone concentration. Because these schemes explore paths in a stochastic manner with no centralized controller, these can support mobility and minimize routing control overheads. However, these schemes focus on relatively small-scale networks so that these have limitations for applying into large-scale MSNs or SFNs. Conventional ant colony routing schemes consist of pheromone concentration and evaporation. After successful packet delivery, the value of pheromone increases. To reflect the dynamic nature of a network, the value of pheromone decreases by a constant amount through the evaporation process as time goes on. In these procedures, the shortest paths are heuristically found and these paths are frequently utilized. That is, there is little consideration for load balancing of traffic loads. When a network size grows, there exists a hub which has many connections with other nodes, especially in MSNs or SFNs. As reported in [18], the frequent usage of a hub generates a congested hot spot or vulnerable region.

On the other hand, a MSN is a relatively new concept of network architecture. Thus, recent trends in this area are extending works in wireless ad networks, delay tolerant networks (DTNs), pocket switched networks (PSNs), and other opportunistic networks. In [19] for PSNs, a routing decision is based on two social metrics such as centrality and community by using real human mobility traces. Whereas in [20], a routing metric considers ego-centric centrality and social similarity. Similar to density based routing, messages are forwarded to the node with higher centrality. For DTNs, [21] proposes an agenda based routing protocol. A majority of DTN applications [22][23] are based on epidemic routing schemes, which significantly degrade performance in high dense networks. In these approaches, there is no reflection for the features of SFNs in MSNs so that it is required to incorporate a hub based networking structure for protocol design.

 ISBN: 978-84-614-6167-7

3. Hub Detour Routing

We propose a hub detour routing scheme (HDR) for MSNs or SFNs by inspiration of an ant colony algorithm. In the initial phase, a packet as an ant moves to its destination by randomly selecting a path. After a packet passes through a node, the node increases its pheromone value. As this operation is iteratively cumulated by other packets in nodes, a stable and efficient path for each destination is established. Distinctively from previous works, our scheme allows no dominant usage of a specific node in a natural manner. In order to determine a pheromone of each node, let us consider a system model of a MSN with a connected graph $G=(V,E)$ where V and E are the set of vertices (nodes) and the set of directed edges (links), respectively. We assume that each node v contains multiple pheromones $P_{vd}(t)$ corresponding to destinations $d \in D$ at time t. To control the amount of pheromone in each node, we design the pheromone function as follows:

$$P_{vd}(t+1) = e^{-KP_{vd}(t)} P_{vd}(t) + \Delta \sum_{l=0}^{\infty} \delta(t-t_l), \qquad \delta(t-t_l) = \begin{cases} 1 & t = t_l \\ 0 & t \neq t_l \end{cases}, \tag{1}$$

where K is the degree[1] of node and Δ is the amount of a pheromone increase. K can be replaced by the number of source-destination pairs recorded by transmitted packets. Additionally, we define the set $T = \{0, 1, 2, 3, \cdots\}$ and the set $L \subset T$. For all $t_l \in L$, t_l satisfies $t_l < t_{l+1}$ where $l \in \{0, 1, 2, 3, \cdots\}$. For pheromone evaporation, we set an exponential function $e^{-KP_{vd}(t)}$ as an evaporation rate, which is enable to accelerate pheromone concentration speed until $KP_{vd}(t) = 1$, and then to accelerate pheromone evaporation speed with respect to node degree of node. By doing so, we can prevent excessive pheromone accumulation in a node. For pheromone concentration, after a successful packet transmission, a node increases its pheromone value by Δ. Otherwise, a pheromone value is only affected by the evaporation procedure. Based on this process, a node v forwards a packet to a neighbor node which has the highest pheromone,

$$\arg\max_{n(v) \in N(v)} P_{n(v)d}(t), \tag{2}$$

among one-hop neighbor nodes $n(v) \in N(v)$ of a node v. Here, a node v records a destination d of a packet so that it seperately manipulates pheromone $P_{vd}(t)$ corresponding to each destination. If there exists no mapping pheromone to a destination, a node v randomly selects a forwarding node $n(v)$. Algorithm 1 describes a pseudo code of HDR.

[1] The degree of node represents the number of one-hop neighbors. This value can be replaced by the number of source-destination pairs recorded through transmitted packets to reflect implicit hubs.

 ISBN: 978-84-614-6167-7

Algorithm 1

SendHelloMessage

```
//Initial Setup
For All Nodes
    Set pheromone 0
End For

//One-hop broadcast of a hello message in the air
Advertise pheromone
```

Packet Routing pseudo code

HubDetourRouting

```
If queue>0 then
  If (Neighbor Nodes for Destination) then
    ForwardingNode = GetHighestPheromone(Neighbor Nodes) by (2)
  Else
    Forwarding Node = Random(Neighbor Nodes)
  End If
  Send a packet to ForwardingNode
  delta = 1
Else
  delta = 0
Enf If
```

// Pheromone update by (1)

```
Result = Calculate_pheromone(Node)
Set pheromone Result
```

To decide Δ for stable routing, we evaluate the pheromone convergence speed with respect to Δ as shown in Figure 2(a). We adopt (3) as an error estimation metric, R.

$$R = \left| P_{vd}(t+1) - P_{vd}(t) \right|, \tag{3}$$

For $0 < \Delta \le 1$, we observe that just four iterations are sufficient to acheive $R < \varepsilon$. This property contributes to rapidly establishing a stable routing path in a dynamic network.

Assume that $L = T$ and $P_{vd}(0) = 0$. Let $Q_{vd}(t) = KP_{vd}(t)$ for some constant K. Then, $P_{vd}(t+1) = e^{-KP_{vd}(t)} P_{vd}(t) + \Delta \quad \Leftrightarrow \quad Q_{vd}(t+1) = e^{-Q_{vd}(t)} Q_{vd}(t) + K\Delta$.

 ISBN: 978-84-614-6167-7

Suppose that $K\Delta > 1$.

Lemma 1. $Q_{vd}(1) \le Q_{vd}(t+1) \le Q_{vd}(2)$ *for all* $t \in T$.

Proof. Since $Q_{vd}(0) = KP_{vd}(0) = 0$ and $e^{-Q_{vd}(t)} Q_{vd}(t) \ge 0$,
$Q_{vd}(t+1) = e^{-Q_{vd}(t)} Q_{vd}(t) + K\Delta \ge K\Delta = Q_{vd}(1)$. Since $f(x) = e^{-x}x$ is a decreasing function when $x > 1$,
$Q_{vd}(t+1) = e^{-Q_{vd}(t)} Q_{vd}(t) + K\Delta \le e^{-Q_{vd}(1)} Q_{vd}(1) + K\Delta = Q_{vd}(2)$.
Thus, $Q_{vd}(1) \le Q_{vd}(t+1) \le Q_{vd}(2)$ for all $t \in T$.

Lemma 2. $Q_{vd}(2t+1)$ *is increasing and* $Q_{vd}(2t+2)$ *is decreasing for any* $t \in T$.

Proof. From Lemma 1,
$Q_{vd}(2t+1) = e^{-Q_{vd}(2t)} Q_{vd}(2t) + K\Delta = e^{-Q_{vd}(2t)} Q_{vd}(2t) + Q_{vd}(1) \ge Q_{vd}(1)$ for any $t \in T \setminus \{0\}$. If $t = 1$, $Q_{vd}(1) \le Q_{vd}(3)$. Since $f(x) = e^{-x}x$ is a decreasing function when $x > 1$,
$Q_{vd}(4) = e^{-Q_{vd}(3)} Q_{vd}(3) + K\Delta \le e^{-Q_{vd}(1)} Q_{vd}(1) + K\Delta = Q_{vd}(2)$.
Since $Q_{vd}(4) \le Q_{vd}(2)$, $Q_{vd}(3) \le Q_{vd}(5)$.
Continuing the same process, we can easily obtain the fact that
$Q_{vd}(1) \le Q_{vd}(3) \le Q_{vd}(5) \le Q_{vd}(7) \le \cdots$ and
$Q_{vd}(2) \ge Q_{vd}(4) \ge Q_{vd}(6) \ge Q_{vd}(8) \ge \cdots$.
Therefore, $Q_{vd}(2t+1)$ is increasing and $Q_{vd}(2t+2)$ is decreasing for any $t \in T$.

Lemma 3. *Suppose* $g(x) = e^{-x}x + K\Delta$. *Then there exists a unique fixed point* ω *satisfying* $g(g(\omega)) = \omega$.

Proof. Let $h(x) = x - g(g(x))$.
$h(K\Delta) = K\Delta - (K\Delta + g(K\Delta)e^{-g(K\Delta)}) = -g(K\Delta)e^{-g(K\Delta)} < 0$ and
$h(K\Delta + \frac{1}{e}) = K\Delta + \frac{1}{e} - (K\Delta + g(K\Delta)e^{-g(K\Delta)}) = \frac{1}{e} - g(K\Delta)e^{-g(K\Delta)} > 0$.

By the intermediate value theorem [24], there exists $\omega \in [K\Delta, K\Delta + \frac{1}{e}]$ satisfying $h(\omega) = \omega - g(g(\omega)) = 0$.

Observe $h'(x) = 1 - g'(g(x))g'(x) = 1 - (e^{-g(x)} - g(x)e^{-g(x)})(e^{-x} - xe^{-x})$.

 ISBN: 978-84-614-6167-7

Since $-\frac{1}{e} < e^{-x} - xe^{-x} < 1$, $(e^{-g(x)} - g(x)e^{-g(x)})(e^{-x} - xe^{-x}) < 1$.

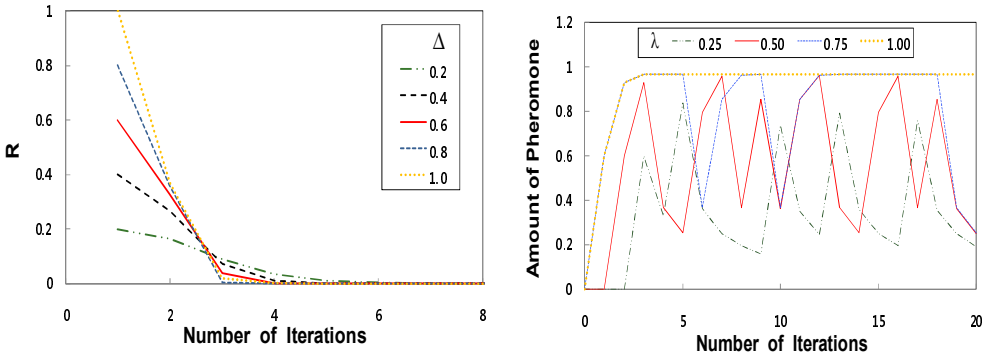
That is, $h'(x) > 0$ or $h(x)$ is an increasing function.

Hence, there exists an unique fixed point ω satisfying $g(g(\omega)) = \omega$.

Theorem 1. $P_{vd}(t)$ *is convergent where* $\Delta > \frac{1}{K}$.

Proof. From Lemma 1 , Lemma 2, and the monotone convergence theorem [25], $Q_{vd}(2t+1)$ and $Q_{vd}(2t)$ is convergent. Let $Q_{vd}(2t+1) \to \beta$ and $Q_{vd}(2t) \to \alpha$. Since $Q_{vd}(2t+1) = e^{-Q_{vd}(2t)} Q_{vd}(2t) + K\Delta$, $\beta = e^{-\alpha}\alpha + K\Delta$. For the same reason, $\alpha = e^{-\beta}\beta + K\Delta$. Let the function $g(x) = e^{-x}x + K\Delta$. Then $g(\beta) = \alpha$ and $g(\alpha) = \beta$. From Lemma 3, $\alpha = \beta$. Thus, $Q_{vd}(t)$ is convergent when $K\Delta > 1$. Hence, $P_{vd}(t)$ is convergent when $\Delta > \frac{1}{K}$.

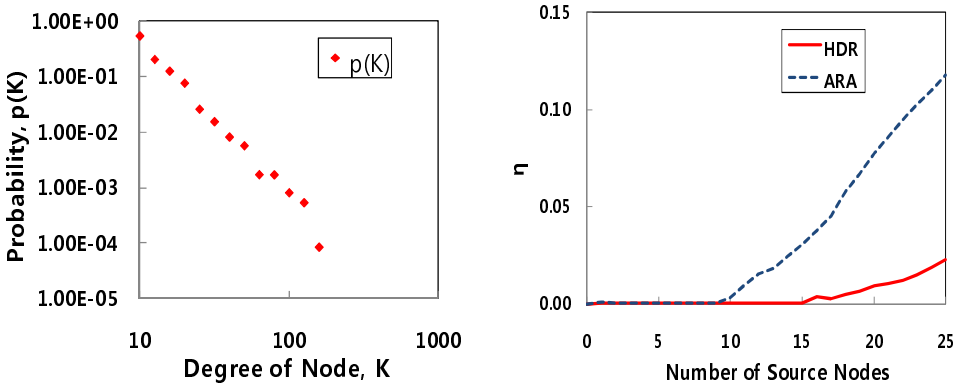
In Figure 2(b), we confirm that $P_{vd}(t)$ is convergent under various packet arrival rates λ, which is relevant to Theorem 1. Because $P_{vd}(t)$ exists in the range of $\Delta \le P_{vd}(t) \le \Delta + \frac{1}{e}$, the lower bound of $P_{vd}(t)$ increases as Δ increases. When packets arrive a node, the pheromone value of the node increases. However, the value does not exceed the pheromone value at $\lambda = 1.00$. This property helps manipute the utilization of a node; it prevents the excessive impact of a specific node to an entire network.



(a) Effect of Δ (b) Effect of packet arrival rate

Figure 2: Dynamic convergence behaviors of pheromone values.

 ISBN: 978-84-614-6167-7



(a) Node degree distribution (b) Network traffic state

Figure 3: Network scenario and comparison for load balancing performance.

4. Simulations

To observe the behaviors of HDR, we conduct simulations. For the simulations, we generate a scale-free network of 1000 nodes based on Barabási-Albert (BA) model [26]. Figure 3(a) represents distribution probability with respect to degree of node K in the network. We observe that the node degree distribution follows the power law distribution similarly to other SFNs. We compare HDR with ARA [15]. We set Δ as 0.6. For the evaporation rate of ARA, we assign 0.5 as in [13].

In the simulations, HDR autonomously constructs paths which can diversify loads detouring a hub. This characteristic increases the achievable throughput in a network. In order to measure traffic congestion in the network, we use the order parameter [27] as follows:

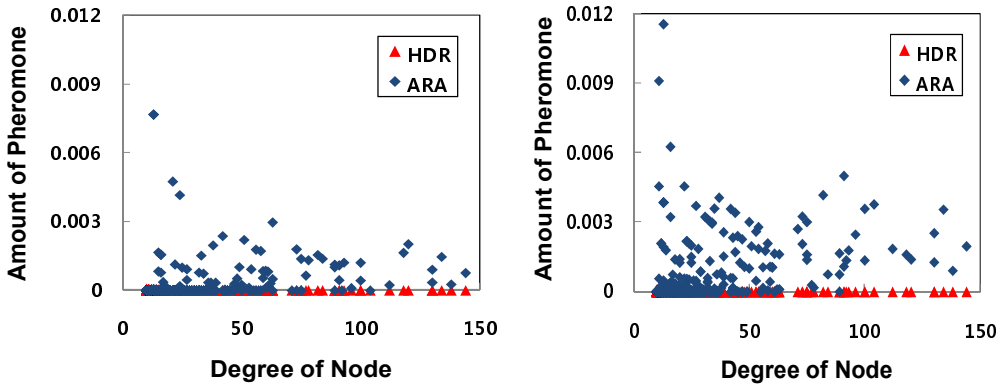
$$\eta(S) = \lim_{t\to\infty} \frac{C}{S} \frac{< \Delta Q >}{\Delta t}, \tag{4}$$

where C is the node capacity, S is the number of source nodes generating packets at each time step, $\Delta Q = Q(t+\Delta t) - Q(t)$, and $< \cdots >$ represents the average over time windows of width Δt. In Figure 3(b), the maximum overall capacity of HDR is $R_c = 13.4$ whereas that of ARA is $R_c = 8.7$. Here, R_c is the critical value at which η starts to increase from zero due the accumulated packets in the system. That is, R_c is the maximum generating rate under which the network is the stable system[2]. For $R < R_c$, the system is in the free-flow state,

[2] In the stable system, the expected lengths of all queues in the network remain bounded over all the time [28].

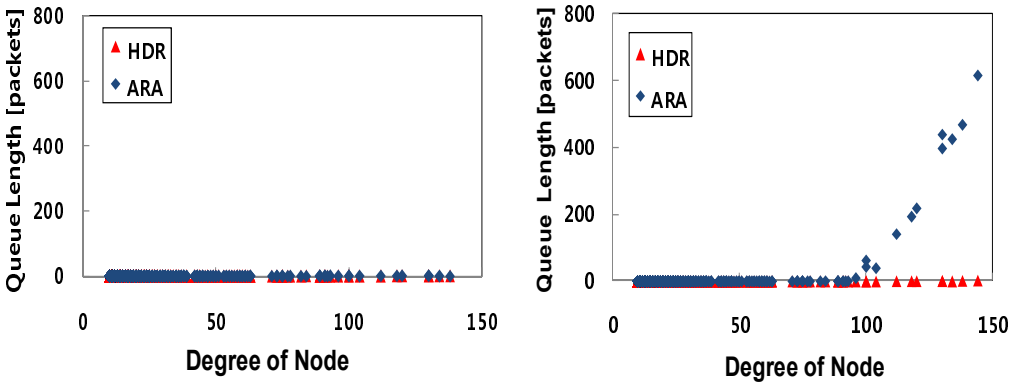
 ISBN: 978-84-614-6167-7

where all packets are serviced. While in $R \geq R_c$, there are non-serviced packets in the system due to serious traffic congestion. Therefore, we confirm that HDR achieves more network throughput capacity than ARA. According to the amount of pheromone in a node with respect to the degree of node as shown in Figure 4(a) and 4(b), HDR maintatins almost even amount of pheromone regardless of the degree of node in both lightly- and heavily-loaded situations. In ARA, the amount of pheromones in nodes with high degree (Degree of Node $\geq$ 100) tremendously increases from lightly-loaded situation to heaviliy-loaded situation. This causes the dominance of specific paths consisting of hubs (high degree nodes). Futhermore, as shown in Figure 5(a) and 5(b), all nodes of HDR and ARA maintain low queue lengths in the lightly-loaded network. However, the queue lengths of high degree nodes of ARA in the heavily-loaded situation significantly increases whereas those in HDR consistently achieve low queue lengths.



(a) Lightly-loaded case (R=10) (b) Heavily-loaded case (R=20)

Figure 4: Amount of pheromone with respect to degree of node.



(a) Lightly-loaded case (R=10) (b) Heavily-loaded case (R=20)

Figure 5: Queue length with respect to degree of node.

 ISBN: 978-84-614-6167-7

5. Conclusions

In this paper, we propose a hub detour routing scheme for future MSNs. Our scheme overcomes the limitations of traditional ant colony algorithms, where a specific node is dominantly utilized in message forwarding. With an advanced pheromone assignment function, our scheme achieves autonomous load balancing and robustness provision in scale-free MSNs. Mathematical analyses and simulations provide the implications of our scheme.

In future work, we will consider the impacts of various networking infrastructure combinations. Furthermore, we can investigate information propagation behaviors of our model and develop scaling laws of pheromone distribution in a more sophisticated manner. It is also possible to apply our work to different network structures such as machine-to-machine (M2M) communications, sensor networks, and other future Internet networking architectures.

6. References

[1] G. Lugano, *Mobile social networking in theory and practice*, First Monday **13** (2008).
[2] Foursquare, [Online], https://foursquare.com/.
[3] Gowalla, [Online], http://gowalla.com/.
[4] MyCubee, [Online], http://www.mycubee.com/.
[5] Bebo, [Online], http://www.bebo.com/.
[6] MySpace, [Online], http://www.myspace.com/.
[7] Facebook, [Online], http://www.facebook.com/.
[8] D. Boyd and N. Ellison, *Social network sites: definition, history, and scholarship*, Journal of Computer-Mediated Communication **13(1)** (2007).
[9] M. Reardon, Cisco predicts wireless-data explosion, Signal Strength - CNET News (2010).
[10] S. Eubank et al., *Modelling disease outbreaks in realistic urban social networks*, Nature **429** (2004) 180-184.
[11] M. Boguna, D. Krioukov, and K. C. Claffy, *Navigability of complex networks*, Nat Phys **5(1)** (2009) 74-80.
[12] F. Papadopoulos et al., *Greedy forwarding in dynamic scale-free networks embedded in hyperbolic metric spaces*, in Proc. of INFOCOM (2010).
[13] M. Dorigo and C. Blum, *Ant colony optimization theory: a survey*, Theoretical Computer Science **344** (2005) 243-278.
[14] G. D. Caro, F. Ducatelle, and L. M. Gambardella, *Anthocnet: an ant-based hybrid routing algorithm for mobile ad hoc networks*, LNCS **3242** (2004) 461-470.
[15] M. Gunes, U. Sorges, and I. Bouazizi, *ARA - the ant-colony based routing algorithm for MANETs*, in Proc. of IWAHN (2002).

 ISBN: 978-84-614-6167-7

[16]L. Rosati, M. Berioli, and G. Reali, *On ant routing algorithms in ad hoc networks with critical connectivity*, Ad Hoc Networks **6(6)** (2008) 827–859.
[17]J. Wang et al., *Hopnet: A hybrid ant colony optimization routing algorithm for mobile ad hoc network*, Ad Hoc Networks **7(4)** (2009) 690–705.
[18]G. Yan et al., *Efficient routing on complex networks*, Physical Review E **73** (2006) 046108.
[19]P. Hui, J. Crowcroft, and E. Yoneki, *BUBBLE Rap: social-based forwarding in delay tolerant networks*, In Proc. of ACM Mobihoc (2008).
[20]E. Daly and M. Haahr, *Social network analysis for routing in disconnected delay-tolerant MANETs*, In Proc. of ACM MobiHoc (2007).
[21]X. Wang et al., *An agenda-based routing protocol in delay tolerant mobile sensor networks*, Sensors **10(11)** (2010) 9564-9580.
[22]A. Vahdat and D. Becker, *Epidemic routing for partially connected ad hoc networks*, Technical Report CS-200006, Duke University (2000).
[23]T. Spyropoulos et al., *Spray and wait: an efficient routing scheme for intermittently connected mobile networks*, in Proc. of WDTN (2005).
[24]W. Rudin, *Pinciples of mathematical analysis* Ch. 4, 3rd ed. McGraw-Hill (2006).
[25]W. Rudin, *Pinciples of mathematical analysis* Ch. 3, 3rd ed. McGraw-Hill (2006).
[26]A.-L. Barabási and E. Bonabeau, *Scale-free networks*, Scientific American **288(5)** (2003) 60–69.
[27]A. Arenas, A. Diaz0Guilera, and R. Guimera, *Communication in networks with hierarchical branching*, Physical Review Letters **86(14)** (2001) 3196-3199.
[28]L. Tassiulas and A. Ephremides, *Stability properties of constrained queueing systems and scheduling for maximum throughput in multihop radio networks*, IEEE Transactions on Automatic Control **37(12)** (1992) 1936-1949.

 ISBN: 978-84-614-6167-7

Proceedings of the 11th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE2011
Benidorm, Spain, 26-30 June 2011

First-Principle Property Calculations for Large Molecules with Auxiliary Density Perturbation Theory

Andreas M. Köster

Departamento de Química, CINVESTAV, Avenida Instituto Politécnico Nacional 2508, A.P. 14-740 México D.F. 07000 México

email: akoster@cinvestav.mx

Abstract

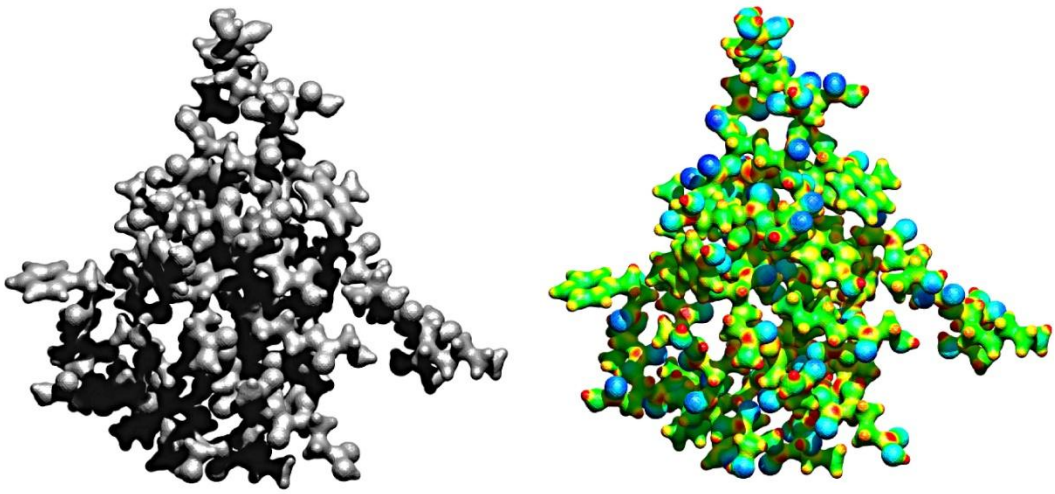
First-principle density functional theory methods with localized atomic orbital basis sets have proven very reliable for electronic, optical and magnetical property calculations. However, for large systems with hundreds or even thousands of atoms the computational demand of the usually employed coupled-perturbed Kohn-Sham (CPKS) method represents a severe obstacle. In this presentation a non-iterative alternative to the CPKS methodology in the framework of auxiliary density functional theory is described. This so-called auxiliary density perturbation theory (ADPT) approach yields CPKS results with a fraction of the computational demand.

Key words: auxiliary density functional theory, auxiliary perturbation theory, deMon2k
MSC2000: AMS Codes (optional)

 ISBN: 978-84-614-6167-7

1. Introduction

This long abstract gives a brief introduction into auxiliary density functional theory (ADFT) [1] and auxiliary density perturbation theory (ADPT) [2] as implemented in the linear combination of Gaussian-type orbital (LCGTO) density functional theory (DFT) program deMon2k [3]. With the development of ADFT first-principle calculations on systems with hundreds of atoms have become feasible. Today, geometry optimizations and transition state searches [4] can be performed for systems containing several hundred atoms [5,6,7] and the thousand-atom barrier has been broken [8], thanks mainly to the development of ADFT. Born-Oppenheimer Molecular Dynamics (BOMD) simulations [9,10,11] have now become "routine" (though sometimes costly) for systems with a hundred or so atoms for simulation times of tens or even hundreds of picoseconds. For a more detailed overview on ADFT we refer to the following review [12]. Besides structure optimizations and BOMD simulations molecular property calculations for large systems are also of outermost importance in Chemistry, Material Science and Biophysics. Whereas density expectation values, like the below depicted iso-density map and molecular electrostatic potential for insulin, are directly accessible from ADFT, response properties are less obvious to calculate.



Insulin iso-density plot (left) and molecular electrostatic potential plot on top of the iso-density surface (right).

The usually applied coupled perturbed Kohn-Sham (CPKS) methodology is computationally rather demanding due to its iterative nature and, therefore, not well suited for large systems. Recently, a non-iterative alternative to CPKS was developed in the framework of ADFT. Different to previous non-variational *ad hoc* expansions of molecular integrals in perturbation theory [13] this so-called

 ISBN: 978-84-614-6167-7

auxiliary density perturbation theory [2] is fully variational and, therefore, yields CPKS results. In ADPT the response density matrix is obtained non-iteratively by solving an inhomogeneous equation system with the dimension of the number of auxiliary functions used to expand the approximated density. Thus, it can be applied efficiently to large molecules. So far ADPT has been successfully applied to the calculation of static and dynamic polarizabilities [2,14,15], Fukui functions [16,17] and nuclear magnetic shielding constants [18]. The development of ADPT hyperpolarizabilities, second analytic derivatives and excited state calculations are currently under way in our laboratory. In the following the working equations for ADFT and ADPT are presented.

2. Auxiliary Density Functional Theory

The auxiliary density functional theory energy expression is given as:

$$E = \sum_{\mu,\nu} P_{\mu\nu}\, H_{\mu\nu} + \sum_{\mu,\nu}\sum_{\bar{k}} P_{\mu\nu}\, \langle\, \mu\, \nu \parallel \bar{k} \rangle\, x_{\bar{k}} \; - \frac{1}{2} \sum_{\bar{k},\bar{l}} \; x_{\bar{k}}\; x_{\bar{l}}\, \langle \bar{k} \parallel \bar{l} \rangle \; + \; E_{XC}[\tilde{\rho}(\boldsymbol{r})]\,.$$

Here μ and ν denote contracted atomic Gaussian-type orbitals, $\bar{k}$ and $\bar{l}$ primitive Hermite Gaussian auxiliary functions and $\parallel$ the two-electron Coulomb operator. This energy expression results from the variational fitting of the Coulomb potential [19,20] and the use of this fitted density for the calculation of the exchange-correlation energy [1,21]. The here appearing linear scaling auxiliary density is given by:

$$\tilde{\rho}(\boldsymbol{r}) = \sum_{\bar{k}} x_{\bar{k}}\; \bar{k}(\boldsymbol{r})\,.$$

The ADFT energy expression is variational. As a result the derivatives of this energy expression with respect to the density matrix elements, keeping the fitting coefficients constants, define the corresponding ADFT Kohn-Sham matrix elements:

$$K_{\mu\nu} = \frac{\partial E}{\partial P_{\mu\nu}} = H_{\mu\nu} + \sum_{\bar{k}} \langle\, \mu\, \nu \parallel \bar{k} \rangle\, x_{\bar{k}} \; + \frac{\partial E_{XC}[\tilde{\rho}(\boldsymbol{r})]}{\partial P_{\mu\nu}}\,.$$

For the partial derivative of the (local) exchange-correlation energy functional follows:

$$\frac{\partial E_{XC}[\tilde{\rho}(\boldsymbol{r})]}{\partial P_{\mu\nu}} = \int \frac{\delta E_{XC}[\tilde{\rho}(\boldsymbol{r})]}{\delta \tilde{\rho}(\boldsymbol{r})}\, \frac{\partial \tilde{\rho}(\boldsymbol{r})}{\partial P_{\mu\nu}}\; d\boldsymbol{r}\,.$$

 ISBN: 978-84-614-6167-7

As in standard Kohn-Sham methods the functional derivative defines the exchange-correlation potential, however, now calculated with the approximated density:

$$v_{XC}[\tilde{\rho}(\boldsymbol{r})] \equiv \frac{\delta E_{XC}[\tilde{\rho}(\boldsymbol{r})]}{\delta \tilde{\rho}(\boldsymbol{r})}.$$

Due to the variational nature of the approximated density the partial derivatives with respect to the density matrix elements can be found as:

$$\frac{\partial \tilde{\rho}(\boldsymbol{r})}{\partial P_{\mu\nu}} = \sum_{\bar{k}} \frac{\partial x_{\bar{k}}}{\partial P_{\mu\nu}} \, \bar{k}(\boldsymbol{r}).$$

Because the variational fitting yields for the fitting coefficients,

$$x_{\bar{k}} = \sum_{\mu,\nu} \sum_{\bar{l}} G^{-1}_{\bar{k}\bar{l}} \, \langle \bar{l} \, \| \mu\nu \rangle \, P_{\mu\nu},$$

the above derivative of the fitting coefficients is given by:

$$\frac{\partial x_{\bar{k}}}{\partial P_{\mu\nu}} = \sum_{\bar{l}} G^{-1}_{\bar{k}\bar{l}} \, \langle \bar{l} \, \| \mu\nu \rangle.$$

Thus, we find for the partial derivative of the exchange-correlation energy functional:

$$\frac{\partial E_{XC}[\tilde{\rho}(\boldsymbol{r})]}{\partial P_{\mu\nu}} = \sum_{\bar{k},\bar{l}} \langle \mu\nu \, \| \, \bar{l} \rangle \, G^{-1}_{\bar{l}\bar{k}} \int \bar{k}(\boldsymbol{r}) \, v_{XC}[\tilde{\rho}(\boldsymbol{r})] \, d\boldsymbol{r} = \sum_{\bar{k},\bar{l}} \langle \mu\nu \, \| \, \bar{l} \rangle \, G^{-1}_{\bar{l}\bar{k}} \, \langle \bar{k} | \, v_{XC}[\tilde{\rho}] \rangle.$$

In order to simplify notation we now introduce exchange-correlation fitting coefficients defined as:

$$z_{\bar{k}} \equiv \sum_{\bar{l}} G^{-1}_{\bar{k}\bar{l}} \, \langle \bar{l} | \, v_{XC}[\tilde{\rho}] \rangle.$$

With these exchange-correlation fitting coefficients we find for the ADFT Kohn-Sham matrix elements:

$$K_{\mu\nu} = \frac{\partial E}{\partial P_{\mu\nu}} = H_{\mu\nu} + \sum_{\bar{k}} \langle \, \mu\nu \, \| \, \bar{k} \rangle \, (x_{\bar{k}} + z_{\bar{k}}).$$

 ISBN: 978-84-614-6167-7

This shows immediately that the Kohn-Sham matrix elements in ADFT are independent from density matrix elements. As a result, only the approximated density (and the corresponding density derivatives in case of gradient corrected functional) are numerically calculated on a grid. Because these quantities scale linear the associated grid work is considerably reduced. Besides the efficient calculation of the three-center ERIs the simplified grid work is the main reason for the computational efficiency of the ADFT approach. In fact, the calculation of the Kohn-Sham potential is in ADFT identical to orbital free DFT approaches with the auxiliary function density as basic variable. Of course, Kohn-Sham orbitals are still used in ADFT for the calculation of the kinetic energy contribution and the orbital density which is the seed for the auxiliary density via the variational Coulomb fitting. Thus, ADFT might be seen as a combination of the basic ideas from the conventional Kohn-Sham methodology with the ones from orbital free DFT. As the above derivation of the Kohn-Sham matrix as a partial derivative of the ADFT energy expression demonstrates analytic energy derivatives can be formulated straightforward in ADFT. This is a major difference to other approaches that use a least-square fitting technique for the approximate calculation of the exchange-correlation energy and potential. Of course, this holds for analytic gradients, too. For higher energy derivatives perturbed density matrix elements are needed. Their efficient calculation in the framework of ADFT is described in the next section.

3. Auxiliary Density Perturbation Theory

Auxiliary density perturbation theory (ADPT) is derived from McWeeny's self-consistent Hartree-Fock perturbation (SCP) theory [22] employing ADFT. For simplicity of the presentation we assume perturbation independent basis and auxiliary functions. The (closed-shell) perturbed density matrix elements are then given by:

$$P_{\mu\nu}^{(\lambda)} = \frac{\partial P_{\mu\nu}}{\partial \lambda} = 2 \sum_{i}^{occ} \sum_{a}^{uno} \frac{\mathcal{K}_{ia}^{(\lambda)}}{\varepsilon_i - \varepsilon_a} \left(c_{\mu i} c_{\nu a} + c_{\mu a} c_{\nu i} \right) \quad .$$

Here λ is the perturbation parameter and $\mathcal{K}_{ia}^{(\lambda)}$ the perturbed Kohn-Sham matrix in molecular orbital representation given by:

$$\mathcal{K}_{ia}^{(\lambda)} = \sum_{\mu,\nu} c_{\mu i} c_{\nu i} K_{\mu\nu}^{(\lambda)} \quad .$$

The analytic form of $K_{\mu\nu}^{(\lambda)}$ contains in the ADFT framework only perturbed fitting coefficients:

 ISBN: 978-84-614-6167-7

$$K_{\mu\nu}^{(\lambda)} = H_{\mu\nu}^{(\lambda)} + \sum_{\bar{k}} \langle \mu\nu \parallel \bar{k} \rangle \left(x_{\bar{k}}^{(\lambda)} + z_{\bar{k}}^{(\lambda)} \right)$$

The perturbed exchange-correlation fitting coefficients can be expressed as:

$$z_{\bar{k}}^{(\lambda)} = \sum_{\bar{l},\bar{m}} \langle \bar{k} \parallel \bar{l} \rangle^{-1} \langle \bar{l} | f_{xc}[\tilde{\rho}] | \bar{m} \rangle x_{\bar{m}}^{(\lambda)}$$

The newly introduced $f_{xc}[\tilde{\rho}]$ denotes the exchange-correlation kernel of the corresponding energy functional calculated with the approximated density. Thus, the perturbed exchange-correlation fitting coefficients can be directly calculated from the perturbed Coulomb fitting coefficients $x_{\bar{m}}^{(\lambda)}$. They are obtained by the solution of the following equation system [2]:

$$(\boldsymbol{G} - 4\boldsymbol{A} - 4\boldsymbol{A}\boldsymbol{F})\boldsymbol{x}^{(\lambda)} = 4\boldsymbol{b}^{(\lambda)}$$

with

$$G_{\bar{k}\bar{l}} = \langle \bar{k} \parallel \bar{l} \rangle$$

$$A_{\bar{k}\bar{l}} = \sum_{i}^{occ} \sum_{a}^{uno} \frac{\langle \bar{k} \parallel ia \rangle \langle ia \parallel \bar{l} \rangle}{\varepsilon_i - \varepsilon_a}$$

$$F_{\bar{k}\bar{l}} = \sum_{\bar{m}} \langle \bar{k} \parallel \bar{m} \rangle^{-1} \langle \bar{m} | f_{xc}[\tilde{\rho}] | \bar{l} \rangle$$

$$b_{\bar{k}}^{(\lambda)} = \sum_{i}^{occ} \sum_{a}^{uno} \frac{\langle \bar{k} \parallel ia \rangle \langle i | r_\lambda a \rangle}{\varepsilon_i - \varepsilon_a}$$

Because the dimension of the above equation system is given by the number of auxiliary functions in the calculation a direct, non-iterative, solution is possible, even for large systems. Once the perturbed fitting coefficients are calculated the perturbed Kohn-Sham matrix and, thus, the perturbed density matrix can be obtained. Thus, ADPT yields the same solution as the corresponding CPKS calculation without employing an iterative procedure. In my presentation I will give an overview about our recent developments in ADPT and will discuss first-principle ADPT calculations of polarizabilities and hyperpolarizabilites, nuclear magnetic shielding tensors and photoabsorption spectra including molecules with more than 1,000 atoms.

 ISBN: 978-84-614-6167-7

4. References

[1] A.M. KÖSTER, J.U. REVELES, J.M. DEL CAMPO, *Calculation of the Exchange-Correlation Potential with Auxiliary Function Densities*, J. Chem. Phys. **121**, 3417 (2004).

[2] R. FLORES-MORENO, A.M. KÖSTER, *Auxiliary Density Perturbation Theory*, J. Chem. Phys. **128**, 134105 (2008).

[3] A.M. KÖSTER, P. CALAMINICI, M.E. CASIDA, R. FLORES-MORENO, G. GEUDTNER, A. GOURSOT, T. HEINE, A. IPATOV, F. JANETZKO, J.M. DEL CAMPO, S. PATCHKOVSKII, J.U. REVELES, D.R. SALAHUB, A. VELA, *deMon2k Version 3.0*, Cinvestav, Mexico-City, 2006.

[4] J.M. DEL CAMPO, A.M. KÖSTER, *A Hierarchical Transition State Search Algorithm*, J. Chem. Phys. **129**, 024107 (2008).

[5] V.D. DOMINGUEZ-SORIA, P. CALAMINICI, A. GOURSOT, *Theoretical study of the structure and properties of Na-MOR and H-MOR zeolite models*, J. Chem. Phys. **127**, 154710 (2007).

[6] P. CALAMINICI, G. GEUDTNER, A.M. KÖSTER, *First-Principle Calculations of Large Fullerenes*, J. Chem. Theor. Comput. **5**, 29 (2009).

[7] V.D. DOMINGUEZ-SORIA, P. CALAMINICI, A. GOURSOT, *Theoretical study of host - guest interactions in the large and small cavities of MOR zeolite models*, J. Phys. Chem. C **115** 6508 (2011).

[8] V.D. DOMINGUEZ-SORIA, G. GEUDTNER, J.L. MORALES, P. CALAMINICI, A.M. KÖSTER, *Robust and Efficent Density Fitting*, J. Chem. Phys. **131**, 124102 (2009).

[9] S. KRISHNAMURTY, M. STEFANO, T. MINEVA, S. BÉGU, J.M. DEVOISSELLE, A. GOURSOT, R. ZHU, D.R. SALAHUB, *Density Functional Theory-Based Conformational Analysis of a Phospholipid Molecule (Dimyristoyl Phosphatidylcholine)*, J. Phys Chem. B **112**, 13433 (2008)

[10] G.U. GAMBOA, P. CALAMINICI, G. GEUDTNER, A.M. KÖSTER, *How Important are Temperature Effects for Cluster Polarizabilities?*, J. Phys. Chem. A **112**, 11969 (2008).

[11] J.M. VASQUEZ-PEREZ, G.U. GAMBOA, A.M. KÖSTER, P. CALAMINICI, *The Discovery of unexpected Isomers in Sodium Heptamers by Born-Oppenheimer Molecular Dynamics*, J. Chem. Phys. **131**, 124126 (2009).

[12] F. JANETZKO, A. GOURSOT, T. MINEVA, P. CALAMINICI, R. FLORES-MORENO, A.M. KÖSTER, D.R. SALAHUB, *Structure Determination of Clusters: Bridging Experiment and Theory, in Nanoclusters: A Bridge across Disciplines*, Editors: P. JENA, A. CASTLEMAN, JR., Elsevier, Amsterdam, 2010.

 ISBN: 978-84-614-6167-7

[13]A. KOMORNICKI, G. FITZGERALD, *Molecular Gradients and Hessians implemented in Density Functional Theory,* J. Chem. Phys. 98, 1398 (1993).

[14]S.V. SHEDGE, J. CARMONA-ESPÍNDOLA, S. PAL, A.M. KÖSTER, *Comparison of Auxiliary Density Perturbation Theory and Non-Iterative Approximation to Coupled Perturbed Kohn-Sham Method: Case Study of Polarizabilities of Disubstituted Azoarene Molecules*, J. Phys. Chem. A **114**, 2357 (2010).

[15]J. CARMONA-ESPÍNDOLA, R. FLORES-MORENO, A.M. KÖSTER, *Time-Dependent Auxiliary Density Perturbation Theory,* J. Chem. Phys. **133**, 084102 (2010).

[16]R. FLORES-MORENO, J. MELIN, J. V. ORTIZ, G. MERINO, *Efficient evaluation of analytic Fukui functions,* J. Chem. Phys. **129**, 224105 (2008).

[17]R. FLORES-MORENO, *Symmetry Conservation in Fukui Functions*, J. Chem. Theory Comput. **6**, 48 (2010).

[18]B. ZUNIGA-GUTIERREZ, G. GEUDTNER, A.M. KÖSTER, *NMR Shielding Tensors from Auxiliary Density Functional Theory*, J. Chem. Phys. **134**, 124108 (2011).

[19]B.I. DUNLAP, J.W.D. CONNOLLY, J.R. SABIN, *Some Approximations in Applications of X-Alpha Theory,* J. Chem. Phys. **71**, 4993 (1979).

[20]J. W. MINTMIRE, J. R. SABIN, S. B. TRICKEY, *Local-Density-Functional Methods in 2-Dimensionally Periodic-Systems – Hydrogen and Beryllium Monolayers*, Phys. Rev. B **26**, 1743 (1982).

[21]B.I. DUNLAP, N. RÖSCH, S.B. TRICKEY, *Variational fitting methods for electronic structure calculations,* Mol. Phys. **108**, 3167 (2010).

[22]R. MCWEENY, *Methods of Molecular Quantum Mechanics*, Academic Press, London, 2001.

 ISBN: 978-84-614-6167-7

Proceedings of the 11th International Conference on Computational and Mathematical Methods in Science and Engineering, CMMSE2011 Benidorm, Spain, 26-30 June 2011

Kinetics of structural transformations in nano-structured intermetallics: atomistic simulations

R. Kozubski[1], A. Biborski[1], M. Kozłowski[1], Ł. Zosiak[1], P. Sowa[1], S. Brodacka[1], V. Pierron-Bohnes[2], Ch. Goyhenex[2], M. Rennhofer[3], E.V. Levchenko[4], A.V. Evteev[4], I. V. Belova[4] and G.E. Murch[4].

[1]*Interdisciplinary Centre for Materials Modelling, M. Smoluchowski Institute of Physics, Jagiellonian University in Krakow, Reymonta 4, 30-059 Krakow, Poland*

[2]*Institut de Physique et Chimie des Matériaux de Strasbourg, 23, rue du Loess, 67034 Strasbourg, France,*

[3]*AIT - Austrian Institute of Technology, Giefinggasse 2, 1210, Wien, Austria.*

[4] *The University Centre for Mass and Thermal Transport in Engineering Materials, Priority Research Centre for Geotechnical and Materials Modelling, School of Engineering, The University of Newcastle, Callaghan, NSW 2308, Australia*

email: rafal.kozubski@uj.edu.pl

Abstract

Kinetics of vacancy-mediated atomic ordering processes in nano-layered $L1_0$ and triple-defect B2 ordered intermetallics was the subject of extensive atomistic simulations. The two groups of systems differ substantially in their vacancy thermodynamics: very low and very high vacancy concentration is observed in $L1_0$ and triple-defect B2 intermetallics, respectively. Special attention was focused on the analysis of an effect of free surfaces on superstructure stability and defect concentration in the examined materials.

Two models of $L1_0$-ordered FePt intermetallic: Ising-type model with two-body interactions and a model with many-body interactions based on Analytic Bond-Order Potentials (ABOP) were simulated by the Quasi-Kinetic Monte Carlo (q-KMC) technique implemented with the vacancy mechanism of atomic migration. For the ABOP model the method was combined with Molecular Statics (MS). Simulation of "order-order" kinetics in [001]-oriented FePt nanolayers initially perfectly ordered in the c-variant $L1_0$ and modelled with two-body interactions revealed a tendency for superstructure transformation from c-variant (monoatomic planes parallel to the (001) free surface) to a- and b-variants (monoatomic planes perpendicular to the (001) free surface) (Fig.1) [1]. The transformation showed

 ISBN: 978-84-614-6167-7

complex kinetics which, except for uniform (bulk-like) disordering, involved three processes: (i) nucleation of a- and b-variant $L1_0$ domains at the surface being initially a single atomic Fe layer, (ii) slow fluctuating growth of the nucleated a- and b-variant $L1_0$ domains inward the layer and (iii) relaxation of the microstructure of the surface domains. In sufficiently thin layers, a percolation of the a- or b-variant superstructure domain nucleated at the surface through the layered sample was observed.

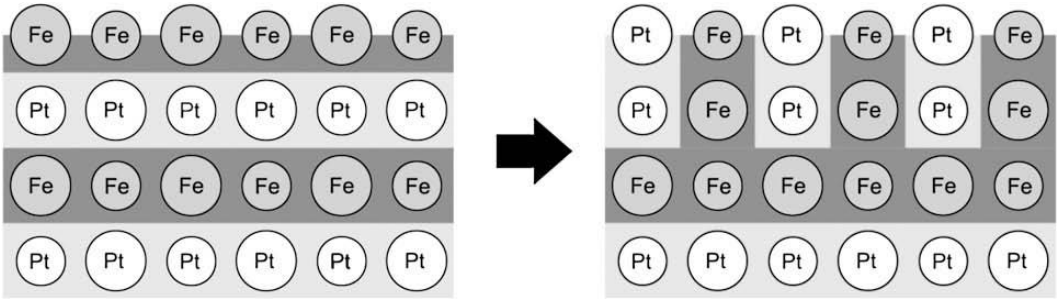


Fig.1. Scheme of an initial stage of $L1_0$ c-variant/a(b)-variant transformation in a FePt layer (atoms residing on two nn crystallographic planes are represented by small and big circles) [1].

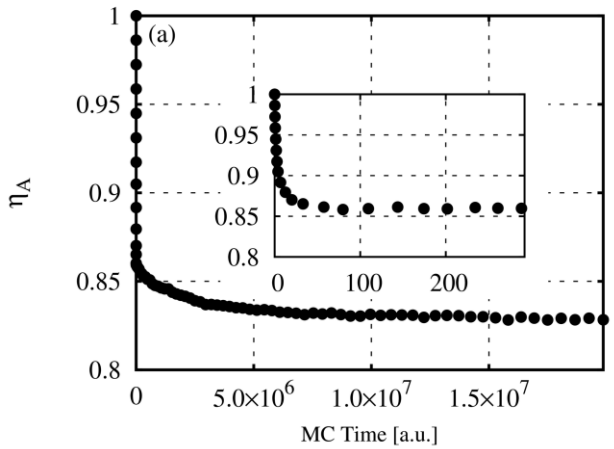
MC simulations of ABOP-based model of the same $L1_0$ FePt layers revealed strong attraction of vacancies by free surfaces. Complex atomic ordering kinetics was observed. Initially fast partial disordering of an internal part of a layer was followed by surface disordering with no $L1_0$ c-variant to a(b)-variant transformation. The transformation, however, was observed in additional Monte Carlo/Static Relaxation simulations performed with the direct exchange algorithm [2,3].
The configurational energy of (001)-oriented FePt layers $L1_0$-ordered in c- and a(b)-variants was calculated with (i) an Ising-type model, (ii) an ABOP model and (iii) a DFT-based model. In all cases, a- and b-variants of $L1_0$ resulted in being energetically stable. It was shown, however, that the models differed in the c-variant/a(b)-variant antiphase-domain boundary (APB) energy, whose height suppressed the c-variant to a(b)-variant transformation in the case of ABOP energetics.
Remarkably, the transformation was experimentally observed in FePt epitaxially deposited multilayers [4].

Triple-defect formation in B2-ordered NiAl intermetallic compound results from a strong asymmetry between the formation energies of Ni- and Al-antisite defects. Chemical disordering in the system is strictly correlated with vacancy formation, which is the reason for the very high vacancy concentration. As a consequence,

 ISBN: 978-84-614-6167-7

Kinetic Monte Carlo (KMC) simulation of ordering occurring in the system and controlled by atomic migration via the vacancy mechanism must involve complete vacancy thermodynamics – i.e. the simulated system must contain an equilibrium concentration of vacancies. NiAl was modelled with an Ising-type Hamiltonian and the temperature-dependent equilibrium concentration of vacancies was determined by means of Semi Grand Canonical Monte Carlo (SGCMC) simulations [5], which assured consistency of the entire approach. The SGCMC simulations led to the evaluation of nearest-neighbour (nn) pair-interaction energies generating the triple-defect behavior of the system. The system generated and modelled in the same way as in the SGCMC was then simulated by KMC for "order-order" kinetics. The procedure required in addition the determination of saddle-point energies assigned to particular atomic jumps to nn vacancies. Their values were estimated in relation to the nn pair-interaction energies with reference to MS simulations performed for NiAl with embedded atom method (EAM) energetics.

The procedure was comparatively applied to bulk and nano-layered B2 AB systems. In both cases, the KMC simulations were started from a configuration with no antisite defects and vacancies (whose number resulted from SGCMC) distributed at random. The results elucidated the role of triple-defect formation as the atomistic-scale origin of the experimentally observed (surprising) low rate of "order-order" kinetics in bulk NiAl. The simulated "order-order" kinetics showed two stages: (i) extremely fast generation of triple defects – i.e. creation af A-antisite defects and related shift of almost all B-vacancies to A-sublattice; the process, which, however, did not lead the system to thermodynamic equilibrium, (ii) extremely slow continuation of the process towards thermodynamic equilibrium – i.e. equilibrium concentration and configuration of antiside defects and vacancies (Fig.2).



 ISBN: 978-84-614-6167-7

Fig.2. KMC simulated "order-order" kinetics in B2-ordering triple-defect binary AB system: η_A is a Bragg-Williams-type long-range order parameter [6].

It was shown that the slow rate of the stage (ii) was due to extremely low efficiency of disordering jumps of A-atoms, which were reversed with very high probability resulting from numerous vacancies residing on A-sublattice. It is claimed that only the stage (ii) of "order-order" kinetics is observed experimentally.
In nano-layers, an additional effect of vacancy segregation on free surfaces and its influence on ordering kinetics was modelled and compared with the related Molecular Dynamics results [7].

References:

[1] M. KOZŁOWSKI, R. KOZUBSKI, CH. GOYHENEX, V. PIERRON-BOHNES, M. RENNHOFER AND S. MALINOV, *Atomic Ordering in Nano-Layered* FePt, Intermetallics **17** (2009) 907-913.

[2] R. KOZUBSKI, M. KOZLOWSKI, J. WROBEL, T. WEJRZANOWSKI, K. J. KURZYDLOWSKI, CH. GOYHENEX, V. PIERRON-BOHNES, M. RENNHOFER AND S. MALINOV, *Atomic Ordering in Nano-layered* FePt: *Multiscale Monte Carlo Simulation* , Comput. Mater. Sci. **49** (2010) S80-S84.

[3] M. KOZLOWSKI, Thesis, Jagiellonian University in Krakow, 2010

[4] M. RENNHOFER, M. KOZLOWSKI, B. LAENENS, B. SEPIOL, R. KOZUBSKI, D. SMEETS AND A. VANTOMME, *Study of reorientation processes in $L1_0$-ordered* FePt *thin films*, Intermetallics, **18** (2010) 2069-2076.

[5] A. BIBORSKI, L.ZOSIAK, R. KOZUBSKI, R.SOT AND V. PIERRON-BOHNES, *Semi-Grand Canonical Monte Carlo simulation of ternary bcc lattice-gas decomposition: Vacancy formation correlated with B2 atomic ordering in* A-B *intermetallics*, Intermetallics, **18** (2010) 2343-2352.

[6] A. BIBORSKI, Thesis, Jagiellonian University in Krakow/Université de Strasbourg, 2010.

[7] E.V. LEVCHENKO, A.V. EVTEEV, R. KOZUBSKI, I. V. BELOVA AND G.E. MURCH, *Molecular Dynamics Simulation of Surface Segregation in a (110)* B2-NiAl *Thin Film*, Phys. Chem. Chem. Phys., **13** (2011) 1214-1221.

 ISBN: 978-84-614-6167-7

Proceedings of the 11th International Conference on Computational and Mathematical Methods in Science and Engineering, CMMSE2011 Benidorm, Spain, 26-30 June 2011

Applying Analytic Hierarchy Process for the Critical Factors of Local Tourism Marketing-The case of Yanshuei District in Taiwan

Kuei-Hsien Chen[1], Chwen-Tzeng Su[2] and Ying-Tsung Cheng[3]

[1, 3] *Department of Business Administration, Nan Jeon Institute of Technology, Taiwan, Republic of China*

[2] *Department of Industrial Engineering and Management, National Yunlin University of Science and Technology, Taiwan, Republic of China*

emails: bm018@mail.njtc.edu.tw, suct@yuntech.edu.tw, yyur20@yahoo.com.tw

Abstract

The tourism is a non-chimney industry that attracts governments' attentions around the world. Both high-tech industry and tourism are the star industries for the 21st century. Global trend also affects local tourism in Taiwan, which is developing and flourishing. To attract the tourists, many local governments in Taiwan hold various festivals, as many as one hundred festivals per year. The most famous one in Taiwan is the Yanshuei Beehive Firecrackers in Yanshuei District, Tainan City. However, due to Yanshuei's remote location and insufficient local marketing resource, the festival is never sufficiently marketed. By adopting analytic hierarchy process (AHP), this research aims to discover the critical factors that affect tourism of Yanshuei. This study includes four factors: people, partnership, place, and programming. The study shows that, the factors can be ranked by importance: people, place, partnership, and programming. The most important index for the people factor is "service", "local government" for partner factor, "media" for place, and "tourist guide" for programming factor.

Key words: place marketing, tourism marketing, analytic hierarchy process

 ISBN: 978-84-614-6167-7

1. Introduction

After the oil crises, production reduced and price level increased globally; economics of cities in various western countries declined, as well as industrial and consumer demands. As a result, public is looking forward government's interference in the free market to increase the competitiveness for local economic growth. Unlike traditional government policy plans, place marketing utilizes the marketing discipline for local development to create unique identity and cultural service [5]. In the late 80s, various state and county governments in the United States faced problems such as high debt level, low employment, and fiscal deficit. As response, the United States government adopted the place marketing as a solution to those problems. Place marketing involves both social and economic activities, and the development strategy is usually executed by local authorities to attract investment, tourists, and residents [3]. Holding festivals is an important method, which boosts popularity, attracts tourists' spending, and contributes to economic developments [6].

Under such "glocalization" trend, Taiwan's place tourism industry in each region also facilitates various festivals in attracting the tourists. Among those festivals, the most famous one is the Yanshuei Beehive Firecrackers in Yanshuei District, Tainan City. Every year in the lantern festival, millions of tourists are attracted by this festival. Yanshuei possesses abundant cultural and historical heritages, as well as various tasty foods and interesting points. However, most tourists' impresses are limited to the beehive firecrackers only, and the tourism marketing is not done properly. To solve the problem, this study adopts the analytic hierarchy process (AHP) to discover the critical factors for tourism in Yanshuei District. The result may be local authorities' base for planning their strategy for place tourism marketing.

2. Research Methods

Analytic hierarchy process (AHP) is a decision making method proposed by [7]. It aims to support the decision making under uncertainty with multiple criteria, and acquire alternative selection or resource weight allocation [2,9]. AHP systematically dissects the problem and creates the hierarchy for the problem. Pairwise comparisons are adopted to acquire relative importance between factors, and alternatives are ranked as basis for selecting optimal solution [8]. Its advantages include flexibility, logical, and easy to understand, therefore it is widely used [4]. This study adopts following methods:

 ISBN: 978-84-614-6167-7

2.1 Structuring Hierarchy

First step of AHP is to structure the hierarchy. This study aims to identify the critical tourism marketing factors for Yanshuei District, and utilizes the 8P marketing mix proposed by [1]. The marketing mix includes Product, Price, Place, Promotion, People, Programming, Partnership and Packaging. Brainstorming method is also used to construct the three hierarchies: goal hierarchy, factor hierarchy, and index hierarchy. Hierarchies are shown in fig 1.

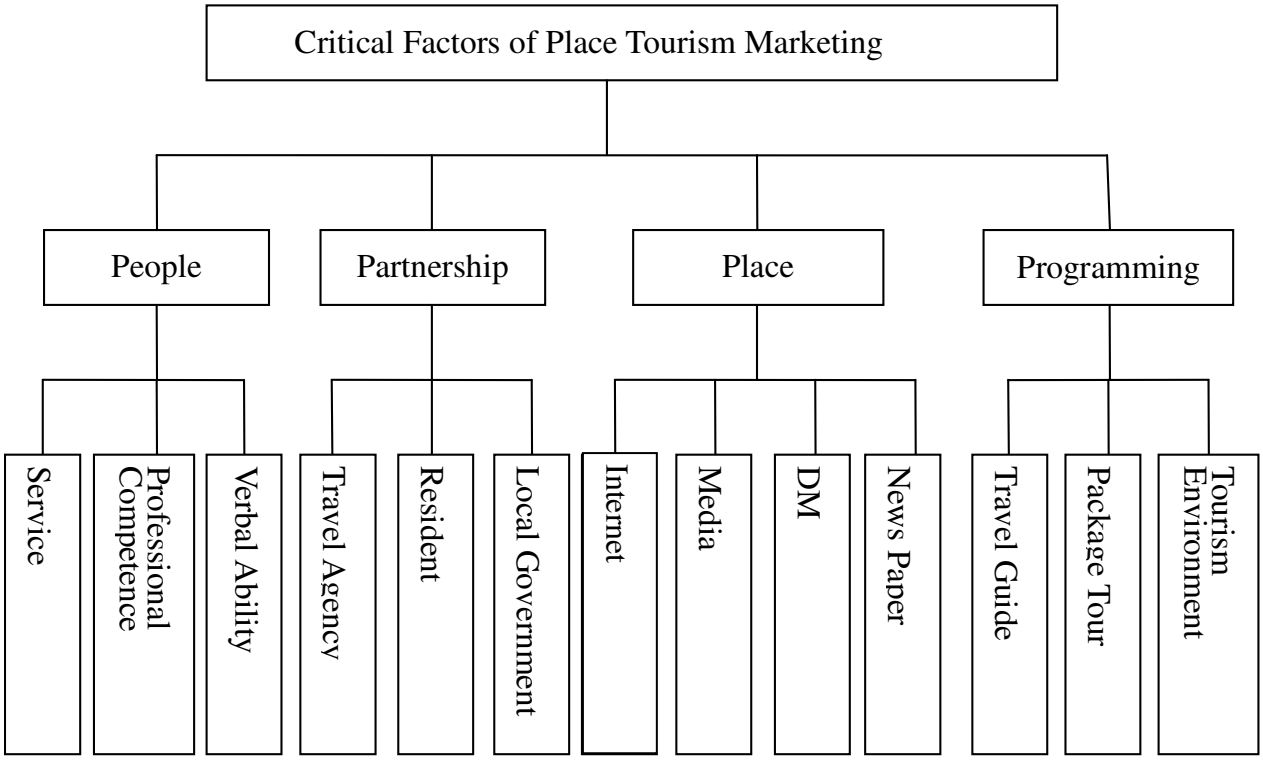


Fig 1 Hierarchy framework of the study

2.2 Pairwise comparisons matrix

Firstly, pairwise comparisons are executed for each criterion with evaluation scale. A_1, A_2, ..., A_n is a set of criteria. To quantitatively judge the relative importance of paired factor (A_i, A_j), this study adopts a nine-point scale, where 1 is for "equal importance", 3 is for "slightly superior important", 5 is for "some superiority", 7 is for "considerable superiority", and 9 is for "outright superiority". Moreover, the comparison matrix $n \times n$ is shown in equation (1):

 ISBN: 978-84-614-6167-7

$$A=[a_{ij}]=\begin{bmatrix} 1 & a_{12} & \cdots & a_{1n} \\ \frac{1}{a_{12}} & 1 & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \cdots \\ \frac{1}{a_{In}} & \frac{1}{a_{2n}} & \cdots & 1 \end{bmatrix} \tag{1}$$

where $i, j = 1,2,\ldots,n$.

After completing the pairwise comparisons matrix, the eigenvalue, which is often used in numeric analysis, is adopted to acquire the eigenvector W for pairwise comparisons matrix. It is also called the priory vector and the largest eigenvalue $\lambda_{\max}$. In practice, after multiplying the vectors and geometric mean is taken and standardized, the W is acquired as shown in equation(2). Calculation of largest eigenvalue is shown in equation (3)-(4).

$$W=\begin{bmatrix} W_1 \\ W_2 \\ \vdots \\ W_n \end{bmatrix}=\left(\prod_{j=1}^{n} a_{ij}\right)^{\frac{1}{n}} \Big/ \sum_{i=1}^{n}\left(\prod_{j=1}^{n} a\right)^{\frac{1}{n}} \qquad i,j=1,2 \ \ldots, \ n \tag{2}$$

$$A\times W = W' \tag{3}$$

$$\lambda_{\max}=\frac{1}{n}\left(\sum \frac{W'}{W}\right) \tag{4}$$

2.3 Consistence Test

After the comparisons, the consistence need to be checked for verify its credibility. The consistency ratio (CR) is shown in equation (6). C.I. stands for consistency index, and R.I. stands for random index. Calculation of CI is shown in equation (5), and RI can be referred in table 1. The overall deviation is acceptable if $CR \leqq 0.1$ and it meets the consistence criteria. If it doesn't, the relationship needs to be calculated again.

$$C.I.=\frac{\lambda_{\max}-n}{n-1} \tag{5}$$

$$C.R.=\frac{C.I.}{R.I.} \tag{6}$$

 ISBN: 978-84-614-6167-7

Table 1: Random Index

n	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
R.I	0.00	0.00	0.58	0.90	1.12	1.24	1.32	1.41	1.45	1.49	1.51	1.48	1.56	1.57	1.59

3. Model Application and Discussion

This study applies analytic hierarchy process in place tourism industry, and aims to identify the critical factors to place marketing. Yanshuei district in Taiwan is used as a case study. In addition, to understand tourists' familiarity about interesting points in Yanshuei, as well as to evaluate local authorities' competence in marketing, the questionnaire comprises of two parts. Part one includes seven famous items in Yanshuei: Yanshuei Beehive Firecrackers, Moon Port, Anes Sugar Refinery, Bridge Street, Octagon, Populace Temple, and Yanshuei Wu Temple. Part two addresses the analytic hierarchy questions as shown in fig 1.

3.1 Part I Analysis

For this study, 250 copies of questionnaire were issued to tourists in Yanshuei, 213 copies (85.2%) were filled and returned, and 198 copies (79.2%) were valid. For part one, 97% of tourists are familiar with Yanshuei Beehive Firecrackers while 3% are not; 44% of tourists are familiar with Moon Port while 56% are not; 47% of tourists are familiar with Anes Sugar Refinery while 53% are not; 50% of tourists are familiar with Bridge Street while 50% are not; 85% of tourists are familiar with Octagon while 15% are nor; 35% of tourists are familiar with Populace Temple while 65% are not; 94% of tourists are familiar with Yanshuei Wu Temple while 6% are not. Yanshuei Beehive Firecrackers is the most famous festival, and its primary location is in Yanshuei Wu Temple, therefore tourists are familiar with these two items the most. Statistics above shows that most tourists have little knowledge about other locations in Yanshuei, and still need local authorities' efforts on marketing.

3.2 Part II Analysis

This study adopts research method in chapter 2 to construct pairwise comparisons matrix for each hierarchy. By using Excel and geometric mean method, weight of each hierarchy, the largest eigenvalue, and consistency ratio are acquired. Furthermore, overall AHP validity is also calculated. Table 2-6 shows the pairwise comparisons matrix and result for each hierarchy.

Table 2 shows the second part of the study. This part includes four factors: people, partnership, place, and programming; their weights are 0.2919, 0.2713, 0.2810 and 0.1555 respectively. According to statistics above, the most important

 ISBN: 978-84-614-6167-7

factor is people, second for place, third for partnership, and the programming is the least important.

Table 2: factors for pairwise comparisons matrix

	People	Partnership	Place	Programming	Weight	Rank
People	1.0000	1.0000	1.2222	1.7143	0.2919	1
Partnership	1.0000	1.0000	1.0000	1.5714	0.2716	3
Place	0.8182	1.0000	1.0000	2.2000	0.2810	2
Programming	0.5833	0.6364	0.4545	1.0000	0.1555	4
CR=0.0084 CI=0.0076 λ_{max} =4.0228						

Table 3 is the matrix under people factor. This hierarchy includes three indexes: verbal ability, service, and professional competence; their weights are 0.2464, 0.3939 and 0.3597 respectively. The most important index is service, then professional competence, and then verbal ability.

Table 3: pairwise comparisons matrix for the people factor

	Verbal Ability	Service	Professional Competence	Weight	Rank
Verbal Ability	1.0000	0.5714	0.7500	0.2464	3
Service	1.7500	1.0000	1.0000	0.3939	1
Professional Competence	1.3333	1.0000	1.0000	0.3597	2
CR=0.0071 CI=0.0041 λ_{max} =3.0082					

Table 4 is the matrix under partnership factor. This hierarchy includes three indexes: travel agency, resident, and local government; their weights are 0.2753, 0.3255, and 0.3992 respectively. The most important index is local government, then residents, and then travel agency.

Table 4: pairwise comparisons matrix for the partnership factor

	Travel Guide	Package Tour	Tourism Environment	Weight	Rank
Travel Guide	1.0000	1.0000	1.2222	0.3559	1
Package Tour	1.0000	1.0000	1.0000	0.3328	2
Tourism Environment	0.8182	1.0000	1.0000	0.3113	3
CR=0.0039 CI=0.0022 λ_{max} =3.0045					

 ISBN: 978-84-614-6167-7

Table 5 is the matrix under place factor. This hierarchy includes four indexes: internet, DM, media and newspaper; their weights are 0.1913, 0.1889, 0.3614, and 0.2366. The most important index is media, then newspaper, internet and DM respectively.

Table 5: pairwise comparisons matrix for the place factor

	Internet	DM	Media	News paper	weight	Rank
Internet	1.0000	0.7778	0.6667	0.7500	0.1913	3
DM	1.2857	1.0000	0.6667	0.6667	0.1889	4
Media	1.5000	1.5000	1.0000	2.2000	0.3614	1
News paper	1.3333	1.5000	0.4545	1.0000	0.2366	2
CR=0.0296 CI=0.0267 λ_{max} =4.0800						

Table 6 is the matrix under programming factor. This hierarchy includes four indexes: travel guide, package tour, and tourism environment; their weights are 0.3559, 0.3328 and 0.3113 respectively. The most important index is tour guide, then package tour, and then tourism environment.

Table 6: pairwise comparisons matrix for the programming factor

	Travel Agency	Resident	Local Government	Weight	Rank
Travel Agency	1.0000	0.7778	0.7500	0.2753	3
Resident	1.2857	1.0000	0.7500	0.3255	2
Local Government	1.3333	1.3333	1.0000	0.3992	1
CR=0.0071 CI=0.0041 λ_{max} =3.0082					

According to data above, both CI and CR are less than 0.1, therefore the sample is consistent.

4. Conclusions

To boost economic activities, in recent years the governments globally strive to develop tourism industry, and the place marketing is the best tool for local development. Therefore, it's very important to execute the marketing management professionally. This study uses Yanshuei District in Taiwan as a case study, and the result shows that most people know Yanshuei only for its Yanshuei

 ISBN: 978-84-614-6167-7

Beehive Firecrackers, and have little knowledge about other locations there. It can be concluded that place marketing is not properly done there. Moreover, this study adopts analytic hierarchy process to identify critical factor for marketing Yanshuei district, and constructed four factors and 13 indexes. The study result shows that under people factor, service is the most critical index, therefore it can be concluded that tourists value the service of staff there the most. Under partnership factor, local government is the most critical index; therefore the tourism policy by local government is very important. Under place factor, the most important index is the media. Therefore it can be concluded that media is an important impression tool to improve local tourism. Under the programming factor, travel guide is the most important index, and it can be concluded that tourists value the comprehensive tour information very much.

The study result above can be referred by local governments as their basis for tourism marketing decision making. Comprehensive marketing plan will greatly improve the place marketing.

5. References

[1] A. MORRISON, *Hospitality and travel marketing (2nd ed.).* Al-bany, New York: Delmar Publishers, 1996.

[2] F. Zahedi, *The Analytic Hierarchy Process-A Survey of the Msthod and its Applications.* Interfaces. **16** (1986), 96-108.

[3] G. KEARNS and C. PHILO, *Selling places: the city as cultural capital, past and present.* London, Pergamon,1993.

[4] L. G. VARGAS, (1990). *An Overview of The Analytic Hierarchy Process and Its Applications*. European Journal of Operation Research. **48**(1990), 2-8.

[5] P. KOLTER, D. H. HAIDER, AND I. REIN, *Marketing Places: Attracting Investment, Industry, and Tourism to Cities, States, and Nations*. New York: Maxwell Macmillan Internationa, 1993.

[6] R. B. GARTRELL, *Destination Marketing for Convention and Visitor Bureaus (2nd ed)*. New York: Kendall/Hunt Publishing Company,1994.

[7] T. L. SAATY, *A Scaling Method for Priorities in Hierarchical Structure*. Journal of Mathematical Psychology, **15**(1971), 234-281.

[8] T. L. SAATY, *The Analytic Hierarchy Process*. New York: McGraw-Hill ,1980.

[9] W. MENDENHALL, D. D. WACKERLY AND R. L. SCHEAFFER, *Mathematical statistics with applications* (*4th*.). Pwskent ,Boston, 1990.

 ISBN: 978-84-614-6167-7

Proceedings of the 11th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE2011
Benidorm, Spain, 26-30 June 2011

Combination of device numerical modeling with full-wave electromagnetics

S. Labiod[1], S. Latreche[1], M. Bella[1], M.R. Beghoul[2], C. Gontrand[3]

[1]Laboratoire Hyperfréquence & Semi-conducteur (LHS), Département d'Electronique, Faculté des Sciences de l'Ingénieur, Université MENTOURI, Constantine, 25000, Algérie.
[2]Département d'Electronique, Faculté des Sciences de l'Ingénieur, Université de Jijel, Algérie.
[3]INL, 7, Bd Blaise Pascal, INSA de Lyon, 69621 Villeurbanne France.

emails: samir.labiod@gmail.com, latreche.saida@gmail.com

Abstract

In this work, we present a numerical solution of a system of partial differential equations (PDE) witch correspond to the carrier transport in a semiconductor active device submitted to an electromagnetic wave environment. This is investigated by coupling a full wave solution of Maxwell's equations to the active device model . This later corresponds to the Drift-Diffusion Model (DDM), which represent the Poisson equation and the carrier transport ones. The proposed active device is a MOS (Metal Oxide Semiconductor) transistor. Various simulations are carried out with MATLAB simulator.
The numerical model is based on a finite-difference (FD) approximation of drift-diffusion model (DDM), The discretization uses a first and second order Finite Difference scheme with up winding based on the characteristic variables in space domain. Backward Euler's scheme is used to accomplish time -domain integration. Gummel's method is used to handle iteratively of the full equations system.

 ISBN: 978-84-614-6167-7

The electromagnetic wave effect on the MOS transistor is investigated by coupling a full wave solution of Maxwell's equations using three-dimensional (3-D) FDTD scheme.
Active device results of the proposed study are in good agreement with those obtained using ISE-TCAD software.
Return loss characteristic of the MOS transistor is presented for several frequency points. Further, we present the electric field distribution in the computational domain.

Key words: PDE, Drift-diffusion model, Maxwell's equations, FTDT method, Backward Euler-implicit scheme.

MSC2000: AMS Codes (optional)

Nomenclature

q : Electronic charge.
n_i : Intrinsic carrier concentration.
n : Electron density.
p : Hole density.
V : Electrostatic potential.
U_T : Thermal voltage.
J_n : Electron current density.
J_p : Hole current density.
μ_n, μ_p : Electron and hole mobility.
D_n, D_p : Electron and hole diffusion coefficient.
τ : Dielectric relaxation time.
τ_n, τ_p : Carrier lifetimes of electron and hole.
Δt : Time step.
H : Space step.
ε : Dielectric constant.
L_D : Debye Length.
E : Electric field.
H : Magnetic field.

1. Introduction

Electromagnetic propagation effects become more and more important for the study of high-frequency circuits based on a semiconductor devices.

 ISBN: 978-84-614-6167-7

At high-frequency, many effects must be considered for the study of microwave-circuits: most notably signal reflections due to interconnecting line discontinuities, dispersion, and crosstalk phenomena.

In this paper, we propose a numerical solution for the analysis of electromagnetic effects on a MOS (Metal Oxide Semiconductor) transistor.
In a microwave circuits, the active device is typically very small in size compared to a wavelength, and it can be modeled by its equivalent current source with a very high degree of accuracy [1].
FDTD algorithm is formulated by directly "finite-differencing" Maxwell's equations. The basic theory and applications of the FDTD method are well described and can be found in [2].
With the advancement of computers, FDTD method has become a very popular tool for analysis of various electromagnetic problems, such as including active device in FDTD algorithm.
Transient simulation is described by the modeling and the simulation of the two-dimensional (2-D) MOS transistor, which achieved by solving various combinations of Poisson and continuity equations conventionally. These later are loosely coupled through successive updates of the three variables V (electrostatic potential), n (electron concentration) and p (hole concentration).
The discretization of the equation's system obtained uses a first and second order FD scheme with upwinding based on the characteristic variables in space domain.
Of course, one of the most important techniques used in the computer modeling of physical systems is the finite-difference (FD) method which represents an essential part of modern theoretical physics. It's able to generate solutions for systems that are far too intricate to be solved analytically. Backward Euler's scheme is used to accomplish time -domain integration.
The implicit matrix systems obtained is solved using an efficient and accurate algorithm based on the well-known Gummel's iterations. The three semiconductor device equations are solved in a decoupled manner. Starting by the initial solution, Poisson's equation will be solved at all grid point to update the electrostatic potential, followed by drift-diffusion equations to compute electrons and holes concentration, repeating the procedure until convergence.

Therefore, this paper presents a three- dimensional (3-D) FDTD method was originally introduced as a technique for the numerical analysis of full-wave problem [3]. We have to modify the FDTD update equations when it is employed to a simple device element such as MOS transistor. In this approach, the active device is treated as current source that interact with Maxwell's equations, exactly with Ampere's law [1]. The Berenger perfectly matched layer (PML) absorbing-boundary condition (ABC) is located to the computational domain to avoid nonphysical reflections. Return loss characteristic of the MOS transistor is extracted, through fast Fourier transform (FFT) analysis, from the transient response predicted by time-domain analysis.

 ISBN: 978-84-614-6167-7

2. Active device model

The semiconductor model used is usually obtained by applying approximations and simplification to the Boltzmann transport equation (BTE). These assumptions can result in a number of different transport models such as the drift diffusion model (DDM) [4]. This model formulates the problem using three dependent variables v, n, and p.
Poisson's equation relates the electrostatic potential to the space charge density:

$$\nabla^2 V = -\frac{q}{\varepsilon}(p(x) - n(x) + Dop(x)) \tag{1}$$

The charge conservation equations formulated for electrons and holes, respectively are:

$$\frac{\partial \mathrm{n}}{\partial \mathrm{t}} - \frac{1}{\mathrm{q}}\vec{\nabla}\vec{\mathrm{J}}_{\mathrm{n}} = -\mathrm{q}.\mathrm{r}_{\mathrm{SRH}} \tag{2}$$

$$\frac{\partial \mathrm{p}}{\partial \mathrm{t}} + \frac{1}{\mathrm{q}}\vec{\nabla}\vec{\mathrm{J}}_{\mathrm{p}} = -\mathrm{q}.\mathrm{r}_{\mathrm{SRH}} \tag{3}$$

The term Dop accounts for the net ionized impurity concentration, and r_{SRH} represents the Shockley-Read-Hall recombination, which is a general recombination process using traps in the forbidden band gap of the semiconductor.

$$\mathrm{r}_{\mathrm{SRH}} = \frac{\mathrm{n}.\mathrm{p} - \mathrm{n}_{\mathrm{i}}^2}{\tau_{\mathrm{p}}.(\mathrm{n} + \mathrm{n}_{\mathrm{i}}) + \tau_{\mathrm{n}}.(\mathrm{p} + \mathrm{n}_{\mathrm{i}})} \tag{4}$$

Current densities are expressed here through de "drift-diffusion" approximation by:

$$\vec{\mathrm{J}}_{\mathrm{n}} = -\mathrm{q}.\mathrm{n}.\mu_{\mathrm{n}}.\vec{\nabla}\mathrm{V} + \mathrm{q}.\mathrm{D}_{\mathrm{n}}.\vec{\nabla}\mathrm{n} \tag{5}$$

$$\vec{\mathrm{J}}_{\mathrm{p}} = -\mathrm{q}.\mathrm{p}.\mu_{\mathrm{p}}.\vec{\nabla}\mathrm{V} - \mathrm{q}.\mathrm{D}_{\mathrm{p}}.\vec{\nabla}\mathrm{p} \tag{6}$$

3. Numerical simulation

The discretization uses a first and second order of (1) in 2-D-finite difference mesh, leads to have

$$\mathrm{V}^{\mathrm{t}}(\mathrm{i}+1,\mathrm{j}) + \mathrm{V}^{\mathrm{t}}(\mathrm{i}-1,\mathrm{j}) + \mathrm{V}^{\mathrm{t}}(\mathrm{i},\mathrm{j}+1) + \mathrm{V}^{\mathrm{t}}(\mathrm{i},\mathrm{j}-1) - 2.\mathrm{V}^{\mathrm{t}}(\mathrm{i},\mathrm{j}) = \frac{\mathrm{q}.\mathrm{h}}{\varepsilon}\left[\mathrm{n}^{\mathrm{t}}(\mathrm{i},\mathrm{j}) - \mathrm{p}^{\mathrm{t}}(\mathrm{i},\mathrm{j}) - \mathrm{Dop}(\mathrm{i},\mathrm{j})\right] \tag{7}$$

 ISBN: 978-84-614-6167-7

The Euler implicit method seeks to approximate the derivatives in (2) and (3) with regard to the discrete solutions points defined by spatial and temporal cells [5, 6]. The electron and hole continuity equations may be discretized in implicit form as follows

$$a_1(i,j).n^{t+1}(i-1,j)+a_2(i,j).n^{t+1}(i+1,j)+a_3(i,j).n^{t+1}(i,j-1)+a_4(i,j).n^{t+1}(i,j+1)- \left[a_5(i,j)+1\right]n^{t+1}(i,j)=\frac{\tau.u_T.h^2.\mu_0}{\Delta t.L_D^2.\mu_n}\left[R(i,j)-n^t(i,j)\right] \quad (8)$$

$$a_6(i,j).p^{t+1}(i-1,j)+a_7(i,j).p^{t+1}(i+1,j)+a_8(i,j).p^{t+1}(i,j-1)+a_9(i,j).p^{t+1}(i,j+1)- \left[a_{10}(i,j)+1\right]p^{t+1}(i,j)=\frac{\tau.u_T.h^2.\mu_0}{\Delta t.L_D^2.\mu_p}\left[R(i,j)-p^t(i,j)\right] \quad (9)$$

The expression of $a_{1\text{-}10}$ can be found in [7].
In this approximation, each time step is represented by t and each spatial step by i and j.
h and Δt are limited respectively by the Debye length and the dielectric relaxation time.
The corresponding Matrix systems of (7), (8) and (9) can be written respectively as:

$$A.V^t = Y^t \quad (10)$$

$$B.n^{t+1} = n^t \quad (11)$$

$$C.p^{t+1} = p^t \quad (12)$$

A, B and C are square matrix corresponding to equations (7), (8) and (9) respectively.
In order to derive the iteration procedure, at each time step, we relate Gummel's method to successive over-relaxation (SOR) method. This is well known to converge quadratically.
The equations (1), (2) and (3) are discretized and solved in a decoupled manner. Poisson equation is solved at all grid points, followed by electron continuity equation, and then by hole's continuity equation, each equation has been solved using SOR method [8, 9].
The whole numerical procedure to calculate the final solution at each time step use Gummel's iterations between the discretized form of equations (10), (11) and (12) [10].

 ISBN: 978-84-614-6167-7

4. Electromagnetic model

The electromagnetic wave propagation can be completely characterized by solving Maxwell's equations that govern the propagation of electric field E and magnetic field H in the computational domain [11]. These equations are first-order linear, the field vectors at any point in the space at any time can be described by Maxwell's curl equations.

$$\vec{\nabla} \times \vec{E} = -\mu \frac{d\vec{H}}{dt} \tag{13}$$

$$\vec{\nabla} \times \vec{H} = \varepsilon \frac{d\vec{E}}{dt} + \vec{J}_{media} + \vec{J}_{Device} \tag{14}$$

Here, $\vec{J}_{media}$ account for the contribution of the current flowing along the distributed media and $\vec{J}_{Device}$ is an impressed current density through which the MOS transistor will be incorporated. Such a contribution comes from the time domain solution of the related device model's equations [12, 13].
A schematic of the coupling between the solvers is given in figure. 1, the active device is assumed to coincide with an E field component. The current along the element and the rate of change of the E field across the element determine the values of H field components surrounding the E field.

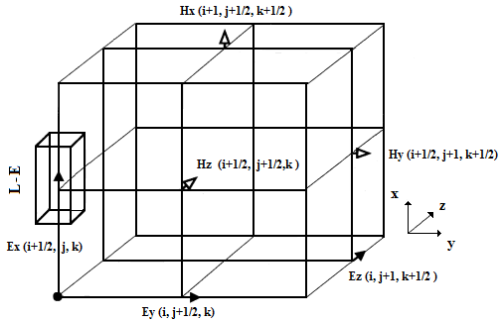


Fig.1. Discretization cell for the FDTD algorithm by incorporating lumped element

The explicit FDTD algorithm is obtained using the centered difference approximation on both the time and space first-order partial differentiations of all the components of E and H.
Using the central difference approximation in both the special and time first-order partial derivatives, we can express (13) and (14) as:

 ISBN: 978-84-614-6167-7

$$H_x^{t+1/2}(i,j,k) = H_x^{t-1/2}(i,j,k) + \frac{\Delta t}{\mu.\Delta z}\left[E_y^t(i,j,k+1/2) - E_y^t(i,j,k-1/2)\right] - \frac{\Delta t}{\mu.\Delta y}\left[E_z^t(i,j,k+1/2) - E_z^t(i,j,k-1/2)\right] \quad (15)$$

$$H_y^{t+1/2}(i,j,k) = H_y^{t-1/2}(i,j,k) + \frac{\Delta t}{\mu.\Delta x}\left[E_z^t(i+1/2,j,k) - E_y^t(i-1/2,j,k)\right] - \frac{\Delta t}{\mu.\Delta z}\left[E_z^t(i,j,k+1/2) - E_z^t(i,j,k-1/2)\right] \quad (16)$$

$$H_z^{t+1/2}(i,j,k) = H_z^{t-1/2}(i,j,k) + \frac{\Delta t}{\mu.\Delta y}\left[E_z^t(i,j+1/2,k) - E_y^t(i,j-1/2,k)\right] - \frac{\Delta t}{\mu.\Delta x}\left[E_z^t(i+1/2,j,k) - E_z^t(i-1/2,j,k)\right] \quad (17)$$

$$E_x^{t+1}(i,j,k) = E_x^t(i,j,k) + \frac{\Delta t}{\varepsilon.\Delta y}\left[H_z^{t+1/2}(i,j+1/2,k) - H_z^{t+1/2}(i,j-1/2,k)\right] - \frac{\Delta t}{\varepsilon.\Delta z}\left[H_y^{t+1/2}(i,j,k+1/2) - H_y^{t+1/2}(i,j,k-1/2)\right] - \frac{\Delta t}{\varepsilon}\left[J_{media,x}^{t+\frac{1}{2}}(i,j,k)\right] \quad (18)$$

$$E_y^{t+1}(i,j,k) = E_y^t(i,j,k) + \frac{\Delta t}{\varepsilon.\Delta z}\left[H_x^{t+1/2}(i,j,k+1/2) - H_x^{t+1/2}(i,j,k-1/2)\right] - \frac{\Delta t}{\varepsilon.\Delta x}\left[H_z^{t+1/2}(i+1/2,j,k) - H_y^{t+1/2}(i-1/2,j,k)\right] - \frac{\Delta t}{\varepsilon}\left[J_{media,y}^{t+\frac{1}{2}}(i,j,k)\right] \quad (19)$$

$$E_x^{t+1}(i+\frac{1}{2},j,k) = E_x^t(i+\frac{1}{2},j,k) + \frac{\Delta t}{\varepsilon.\Delta y}\left[H_z^{t+\frac{1}{2}}(i+\frac{1}{2},j+\frac{1}{2},k) - H_z^{t+\frac{1}{2}}(i+\frac{1}{2},j-\frac{1}{2},k)\right] - \frac{\Delta t}{\varepsilon.\Delta z}\left[H_y^{t+\frac{1}{2}}(i+\frac{1}{2},j,k+\frac{1}{2}) - H_y^{t+\frac{1}{2}}(i+\frac{1}{2},j,k-\frac{1}{2})\right] - \frac{\Delta t}{\varepsilon}\left[J_{media,x}^{t+\frac{1}{2}}(i+\frac{1}{2},j,k) + \frac{1}{2}\left(J_{lumped,x}^{t+1}(i+\frac{1}{2},j,k) + J_{lumped,x}^{t}(i+\frac{1}{2},j,k)\right)\right] \quad (20)$$

Where Δx, Δy and Δz are spatial discretization.

2. Results and discussions

The considered active device is a MOS (Metal Oxide Semiconductor) transistor with a 0.55 μm channel length.

 ISBN: 978-84-614-6167-7

Initially the device is biased to Vds=1 V and Vgs=0.8 V, and the DC distributions are obtained by solving the active device model only. The potential was obtained by the self- consistent solution of the Poisson, electron and hole continuity equations. The state of the MOS transistor under DC steady-state is represented by the distribution of the electrostatic potential, electron density and current density.

Figure 2 represents the norm of residual error for the variations in potential vector [14]. The proposed algorithm provides a good convergence; here each iteration takes approximately 0.025 s.

Figure 3.a shows the potential distribution of the device. Figure 3.b shows the computed electrons density profile of the proposed device in logarithmic scale. It can be seen that electrons has been attracted at the interface oxide-semiconductor. The figure clearly shows the onset of the pinch-off effect which indicates the lack of channel region near the drain junction.

The static I-V characteristics for the MOS transistor are finally presented in figure 4.a. The drain current was obtained using drift-diffusion model, and demonstrate a good agreement with those obtained using ISE-TCAD software (Sentaurus). To calculate I-V characteristics, we take the current density matrix for a given bias, and then integrate it over the surface at the drain contact.
Figure 4.b shows drain current variation versus time for the bias (Vds=1V, Vgs=0.8 V), the obtained result has been calculated by solving the tree implicit equations (7), (8) and (9) in a decoupled manner in the time and space. The stability of drain current appears above 80 ps and take the same value which obtained in steady-state simulation.
Notice when the channel is clearly formed, It is significant to indicate that Euler's method gives precisely the same results obtained in Stationary simulation.
Figure 5.a shows the distribution of electric field beneath the structure at 400 time steps. Figure 5.b represent return loss coefficient of the active device versus frequency. The S11-parameter has been extracted in bandwidth (100MHz to 50 GHz) using FFT (Fast Fourier Transform) of the temporal response of the considered component.

 ISBN: 978-84-614-6167-7

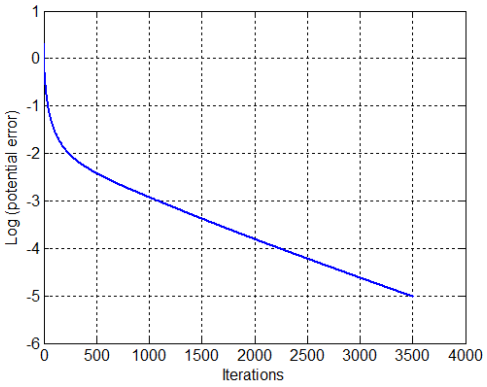


Figure 2. Potential error in logarithmic scale

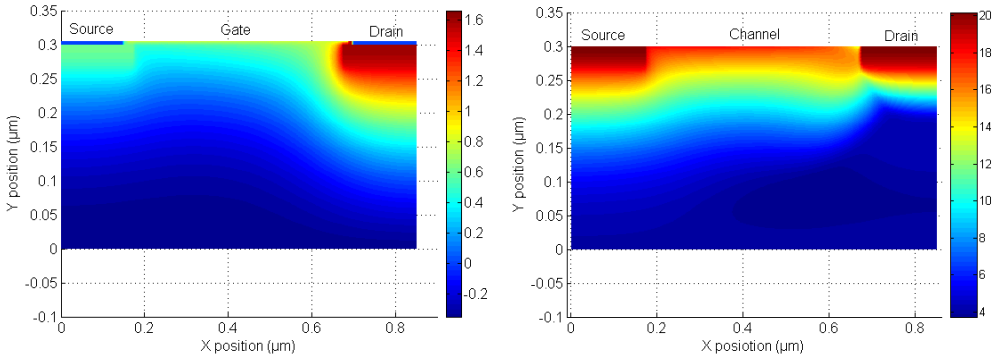


Figure 3. The Calculated steady-state results: (a) Potential distribution. (b) Electron density distribution.

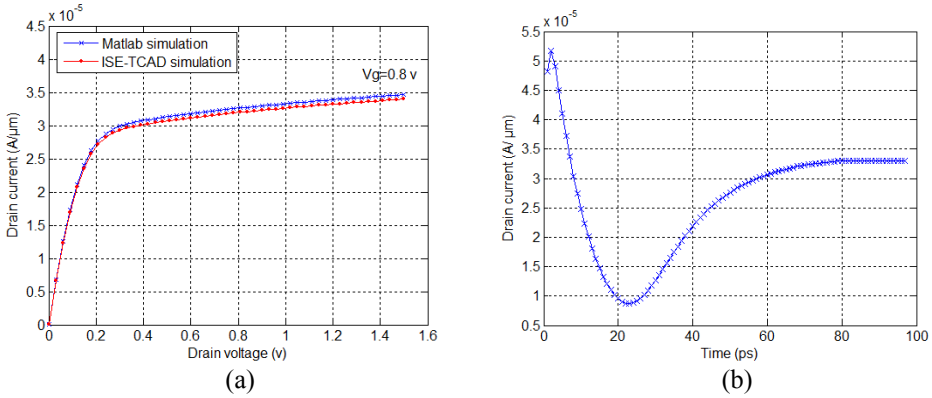
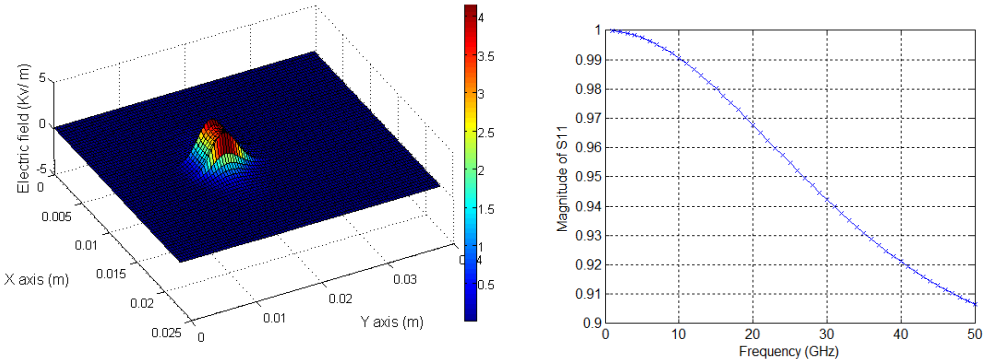


Figure 4. Calculated drain current characteristics: (a) Output characteristic. (b) Transient drain current.

 ISBN: 978-84-614-6167-7



(a) (b)

Fig.5. Electromagnetic simulations: (a) spatial distribution electric field component. (b) Return loss coefficient

3. Conclusion

The semiconductor equations of the proposed device consist of a set of nonlinear second-order partial differential equations. The descretization of the semiconductor equations is established using backward Euler scheme; the implicit system obtained is solved using the well known Gummel's scheme related with an efficient successive over-relaxation (SOR) method.
Static and transient results for active device is calculated and compared with those obtained using ISE-TCAD software (Sentaurus).
This work presented an efficient time-domain full-wave simulator for modeling and simulation of MOS transistor used for mm-wave applications.
The coupling of the drift-diffusion model with Maxwell's equations is done via the extended 3-D FDTD formalism. Next the MOS transistor model is implemented in a FDTD grid for the simulation of a simple strip line. Return lose parameter has been extracted from the FDTD algorithm by using FFT method.
The modeling results are qualitatively in agreement with theoretical concepts which confirm the validity of the proposed algorithm.

4. References

[1] C. N. KUO, R. B. WU. B. HOUSHMAND, AND T. ITOH, *Modeling of Microwave Active Devices Using the FDTD Analysis Based on the*

 ISBN: 978-84-614-6167-7

Voltage-Source Approach, IEEE MICROWAVE AND GUIDED WAVE LETTERS, VOL. 6, PP. 199-201, MAY 1996.

[2] A. TAFLOVE, *Computational electromagnetic the Finite-Difference Time-Domain Method*, ARTECH HOUSE, 1995.

[3] O. GONZALEZ, J. A. PEREDA, A. HERRERA, AND A. VEGAS, *An extension of the lumped-network FDTD method to linear two-port lumped circuits*, IEEE TRANS. MICROWAVE THEORY AND TECHNIQUES , VOL. 54, PP. 3045-3051, JULY. 2006.

[4] K. BLOTEKJAER, *Transport equations for electrons in two-valley semiconductors,* IEEE trans, Electron Devices, VOL. ED-17, NO. 1, PP. 38-47, Janvier 1970.

[5] A. QUARTERONI, R. SACCO, AND F. SALERI, *Numerical Mathematics*, SPRINGER, 2000.

[6] R. MIRZAVAND, A. ABDIPOUR, AND G. MORADI, *Full-Wave semiconductor devices simulation using ADI-FDTD method*, PROGRESS IN ELECTROMAGNETIC RESEARCH M, VOL. 11, PP. 191-202, MAY 2010.

[7] R. MIRZAVAND, A. ABDIPOUR, AND G. MORADI, W.H.A. SCHILDERS, M. MOVAHEDI, *LOD-FDTD method for physical simulation of semiconductor device*, IEEE EXPLORE. PROCEEDING ICMMT, AUGUST 2010.

[8] A. ASTE, AND R. VAHLDIECK, *Time-Domain simulation of the full hydrodynamic model*, INTERNATIONAL JOURNAL OF NUMERICAL MODELING, VOL. DOI 10, PP. 161-174, DECEMBER 2003.

[9] M. MOVAHHEDI, AND A. ABDIPOUR, *Efficient numerical methods for simulation of High-Frequency active device*, IEEE TRANS. MICROWAVE THEORY AND TECHNIQUES, VOL. 54, PP. 2636-2645, JUNE. 2006.

[10] LATRECHE S., LABIOD S. AND GONTRAND C., Electromagnetic wave effect on semicondcutor device : FDTD Method, PROCEEDINGS OF THE 2010 CMMSE, Vol. IV, pp. 1159.Ed. J. Vigo Aguiar, Spain.

[11] J. E. MARSDEN, L. SIROVICH, AND S. S. ANTMAN, *Computational Electromagnetics*, SPRINGER, 2005.

[12] P. CIAMOLINI, L. ROSELLI, AND G. STOPPONI, *Mixed-Mode circuit simulation with full-wave analysis of interconnections*, IEEE TRANS. ELECTRON DEVICES, VOL. 44, PP. 2098-2105, NOVEMBER 1997.

[13] O. GONZALEZ, J. A. PEREDA, A. HERRERA, AND A. VEGAS, *An extension of the lumped-network FDTD method to linear two-port lumped circuits*, IEEE TRANS. MICROWAVE THEORY AND TECHNIQUES , VOL. 54, PP. 3045-3051, JULY. 2006.

[14] PULLIAM, H. *Time accuracy and the use of implicite methods*. AIAA PAPER, N°. 93-3360, PP. 685-693, 1993.

 ISBN: 978-84-614-6167-7

Proceedings of the 11th International Conference on Computational and Mathematical Methods in Science and Engineering, CMMSE2011 Benidorm, Spain, 26-30 June 2011

A PERIODIC MODEL BASED ON GREEN FUNCTION AND BLOCH THEORY: DYNAMIC MODELLING OF RAILWAY TRACK.

Rachid Lassoued [1], Mostefa Lecheheb[1], Guy Bonnet [2]

[1] *Laboratoire des Matériaux et Durabilité des Constructions Département de Génie Civil Faculté des Sciences de l'Ingénieur Université Mentouri Constantine, Algérie*

[2] *Université Paris Est Marne la Vallée, Laboratoire de Modélisation et Simulation Multi Echelle, CNRS UMR 8208, 5 boulevard Descartes, 77454 Marne la Vallée Cedex, France*

emails: rachid_lassoued@yahoo.fr

Abstract

A mathematical model is presented for the propagation of structural waves on an infinitely long, periodically supported Euler beam and Timoshenko beam. The behaviour of railway platforms and the noise transmitted by the rail and the platform depend strongly on the dynamics of the rail. Usually, the rails rest on periodically located supports.The writing of dynamic equilibrium and the summation of all the force's moments lead to a system of partial differential equations for a Timoshenko beam, this differential equation of bending motion ,including the action of the external force. To analyze the vibrations according to the frequency, we carried out a Fourier transform. The work starts by calculate the Green function for a Euler beam and a Timoshenko beam without support by using the direct integration. The beam motion can be described by applying the superposition principle which states that the response from all sleepers points and from the external point force add up linearly to give the total response. The periodicity of supports is described by Bloch's theorem. The model developed gives the choice of different support types: mass support, spring support, mass/spring systems…The homogeneous system thus obtained represents a linear differential equation which governs rail response. It is initially solved in the homogeneous case and it admits a no null solution if its determinant is null. This permits the establishment of the dispersion equation to Bloch waves and wave bands. The Bloch waves and dispersion curves contain all the physics of the dynamic problem. The solution determines the Bloch propagation constant. Its real part permit to analyse the attenuation introduced by the structure and its imaginary parts, which represent the Bloch wave number, permit to analyse the free modes of vibration. The two types of support models are considered for this, Timoshenko and Euler beams. A comparison of these two cases is performed. A

 ISBN: 978-84-614-6167-7

parametric analysis for the two models is realized for the Green function.The influence of the periodic supports of the railway track on the vertical bending waves is analyzed. The supports induce passing/stopping band frequency behaviour on the waves. The stopping bands are not equally spaced, but increase with increasing frequency.

Key words: Partial differential equation, Bloch theorem, Green function, analytical method.

1. Introduction

Many components of engineering structures are constructed in a spatially periodic structure consists fundamentally of a number of identical structural components "periodic elements "which are joined together end-to-end and or Side-by-side to form the whole structure. A railway line on equi -spaced sleepers is the oldest example.
Vibration analyses of these structures are frequently required, and extensive studies of their free harmonic motions have been made over the past 25 years. Brillouin's classic work [l] on wave propagation in periodic structures has laid the foundation for much of the thinking in these studies.
The response has usually been analyzed by using the propagation constant [2],[3],[4],[5], [6] which describes the spatial relation between two bays that is to say, the response is calculated in a single bay and its phase relation with the other bays. Early investigations were primarily concerned with the pass bands and stop bands of a periodic structure, and the propagation constant was used to display these characteristics. In 1980, Mace [4-5] analyzed an infinite periodic plate reinforced by unidirectional stiffeners and excited by line or point harmonic forces. After Fourier-transforming the applied force and stiffener reactions, he found the transform of the structural response. (Munjal and Heckl, 1982) modeled the rail as an Euler beam on periodic mass supports and analyzed the propagation of vertical bending waves with a transfer matrix method. The track is modeled as a infinitely long beam (Timoshenko) supported by discrete support systems [7]. The rail is modeled as an Euler beam with flexible supports in the work of Nordborg [8]. Mead studies [9], [10], [11] of wave propagation on more general periodic engineering structures have provided valuable insight into multiple coupling of waves. Belotserkovskiy [12], the beam deflection is governed within the segment by Euler-Bernoulli partial differential equation. The Fourier transform is used to solve the problem.
The vibration motion of a periodically supported beam consists of numerous reflections combined with flexural nearfield due to the presence of elastic supports or stiffeners.
In this paper, the mathematical model is presented to predict the propagation of structural waves on an infinitely long periodically supported structures. This periodic system consists of a number of identical elements, coupled together in identical ways to form the whole system. The rail idealizes these periodic structures. The supports of the beam represent pad/sleeper/ballast system of the railway track; they are spaced regularly.
In this paper, we present the influence of the periodic supports of the railway track on the vertical bending wave propagation. The two types of models are considered for the Timoshenko beam and Euler one. A comparison of these two cases is performed. A parametric analysis for the two models is realized for the Green function.The supports

 ISBN: 978-84-614-6167-7

induce passing/stopping band frequency behaviour on the waves. The stopping bands are not equally spaced, but increase with increasing frequency.

2. Response to harmonic load on the infinite unsupported beam

➢ Euler beam

Let a thin beam with an inertia section moment I and which is subjected to a moving vertical moving load $F(x,t)$. The vertical beam displacement is the solution of the dynamic beam equation (eq. 1). Euler model is considered:

$$EI\frac{\partial^4 U(x,t)}{\partial x^4} + m\frac{\partial^2 U}{\partial t^2} - F(x,t) = 0 \quad (1)$$

Where m is the mass of the beam by unit of length and E the Young's modulus.

If the harmonic concentrated force is applied at the point xo, in the Fourier space, the movement equation becomes:

$$EI\frac{\partial^4 \tilde{U}(x,\omega)}{\partial x^4} - m\omega^2\tilde{U}(x,\omega) - \delta(x-x_0) = 0 \quad (2)$$

Where $\tilde{U}$ is the Fourier transform of$U$.

The Green function associated to this equation is the response to a unit force applied at the point xo. It is the solution of the equation:

$$\frac{d^4G(x,x_0)}{dx^4} - k_b^4 G(x,x_0) = \delta(x-x_0) \quad \text{with} \quad k_b = \sqrt[4]{\frac{m\omega^2}{EI}} \quad (3)$$

The Green function $G(x,x_0)$ can be now determined by direct integration; it is thus the solution of the equation (3) and represents the vertical vibration at the point xo of the beam rail subjected to a harmonic.

- In any point $x \neq x_0$ (in any point other than the excitation point) the system is free and :

$$\frac{d^4G(x,x_0)}{dx^4} - k_b^4 G(x,x_0) = 0 \quad x \neq x_0 \quad (4)$$

- At the point $x = x_0$., the discontinuity is assumed by the third derivative of $G(x,x_0)$. The function and its two first derived are continuous at $x = x_0$. If we integrate the equation (2) on both sides of x_0 And if we take into account the continuity of the function $G(x,x_0)$ at x_0, we obtain :

 ISBN: 978-84-614-6167-7

$$\frac{d^3}{dx^3}G(x,x_0)|_{-}^{+} = \frac{1}{EI} \tag{5}$$

- At $x = \pm\infty$ the function is limited: $|G(x,x_0)| < \infty$

The Green function is then expressed by a combination of elementary solutions of the equation (3). Taking into account the condition (at infinity), it can be written:

$$G(x,x_0) = a_1 e^{k_b(x-x_0)} + a_2 e^{ik_b(x-x_0)} \qquad x < x_0 \tag{6}$$

$$G(x,x_0) = b_1 e^{-k_b(x-x_0)} + b e^{-ik_b(x-x_0)} \quad x > x_0 \tag{7}$$

The coefficients a_1, a_2, b_1, b_2 are evaluated from the Green function and its first derived which must satisfy:

$$G(x,x_0)|_{x_{0-}}^{x_{0+}} = 0 \qquad \frac{d}{dx}G(x,x_0)\Big|_{x_{0-}}^{x_{0+}} = 0 \qquad \frac{d^2}{dx^2}G(x,x_0)\Big|_{x_{0-}}^{x_{0+}} = 0 \tag{8}$$

The solution of this problem is exposed by Heckl (Maria **A. Heckl, 2001)** in the more general case of the Timoshenko beams. We obtain in the case of the Euler beam.

$$G(x,x_0) = \frac{1}{4EIk_b^4}\left[ie^{ik_b|x-x_0|} - e^{-k_b|x-x_0|}\right] \tag{9}$$

- **Timoshenko beam**

The model of Timoshenko beam is considered for the thick beams; it is valid in a large band of frequency. It takes into account shearing and rotational inertia. It is adapted better in the case of the rail. The writing of dynamic equilibrium and the summation of all the force's moments lead to the system of the following equations.

$$\frac{\partial}{\partial x}\left[k'AG\left(\alpha - \frac{\partial v}{\partial x}\right)\right] = F - m\frac{\partial^2 v}{\partial t^2} \tag{10}$$

$$\frac{\partial}{\partial x}\left(EI\frac{\partial \alpha}{\partial x}\right) = k'AG\left(\alpha - \frac{\partial v}{\partial x}\right) + mr^2\frac{\partial^2}{\partial t^2} \tag{11}$$

Where:

α : is the rotation of the right section starting from its initial position.

v : the vertical component of the displacement

$r^2 = {}^{I}/_{A}$: the radius of section gyration.

EI : the flexion Stiffness.

m : Linear density.

 ISBN: 978-84-614-6167-7

The solutions given by this system of equations lead to the following expressions. They characterize the propagative waves and the waves of the nearfield.

$$k_p = \frac{1}{\sqrt{2}}\left[\sqrt{(k_c^2 + k_t^2)^2 + 4(k_b^4 - k_c^2 k_t^2)} + (k_c^2 + k_t^2)^2\right]^{1/2} \quad (12)$$

$$k_d = \frac{1}{\sqrt{2}}\left[\sqrt{(k_c^2 + k_t^2)^2 + 4(k_b^4 - k_c^2 k_t^2)} - (k_c^2 + k_t^2)^2\right]^{1/2} \quad (13)$$

Where:

$k_c = \omega\sqrt{\frac{\rho}{E}}$: Compression wave number. $k_t = \omega\sqrt{\frac{\rho}{Gk'}}$: Shearing Wave number.

$k_b = \omega\sqrt{\frac{\rho A\omega^2}{EI}}$: Flexion wave number for Euler beam.

k_p : Propagative waves.

$\mathrm{k_d}$: Waves at nearfield.

Then, the differential equation of the movement of the Timoshenko beam is:

$$EI\left(\frac{\partial^2}{\partial x^2} - \frac{\rho}{E}\frac{\partial^2}{\partial t^2}\right)\left(\frac{\partial^2}{\partial x^2} - \frac{\rho}{k'G}\frac{\partial^2}{\partial t^2}\right)v + \rho A\frac{\partial^2 v}{\partial t^2} = \left[1 - \frac{EI}{k'AG}\left(\frac{\partial^2}{\partial x^2} - \frac{\rho}{E}\frac{\partial^2}{\partial t^2}\right)\right]F(x,t) \quad (14)$$

To analyze the vibrations according to the frequency, we carried out a Fourier transform.

$$EI\left[\left(\frac{d^2}{dx^2} - k_c^2\right)\left(\frac{d^2}{dx^2} - k_t^2\right)v + k_b^4 v\right] = \left[1 - \frac{EI}{k'AG}\left(\frac{d^2}{dx^2} - \frac{\rho}{E}\omega^2\right)\right]F(\omega) \quad (15)$$

The Green function associated to the equation (14) is the response of a load applied at x_0 ; it is the general solution of the equation:

$$EI\left[\left(\frac{d^2}{dx^2} - k_c^2\right)\left(\frac{d^2}{dx^2} - k_t^2\right) + k_b^4\right]G(x,x_0) = \left[1 - \frac{EI}{k'AG}\left(\frac{d^2}{dx^2} - \frac{\rho}{E}\omega^2\right)\right]\delta(x,x_0) \quad (16)$$

- The function $G(x,x_0)$ is supposed to continue at $x = x_0$
- In any point $x \neq x_0$, , the system is free.

$$EI\left[\left(\frac{d^2}{dx^2} - k_c^2\right)\left(\frac{d^2}{dx^2} - k_t^2\right) + k_b^4\right]G(x,x_0) = 0 \quad (17)$$

- We determine the conditions which must check $G(x,x_0)$ derivatives at xo, by carrying out successive integrations of the equation (16).

$$G(x,x_0)\Big|_{x_{0-}}^{x_{0+}} = 0 \qquad \frac{d}{dx}G(x,x_0)\Big|_{x_{0-}}^{x_{0+}} = \frac{1}{K}$$

 ISBN: 978-84-614-6167-7

$$\frac{d^2}{dx^2}G(x,x_0)\Big|_{x_0-}^{x_0+}=0 \qquad \frac{d^3}{dx^3}G(x,x_0)\Big|_{x_0-}^{x_0+}=\frac{1}{B}\left[1+\frac{B}{K}k_c^2\right] \tag{18}$$

- The Green function is then expressed by a linear combination of the elementary solutions.

$$G(x,x_0)=a_1e^{k_d}(x-x_0)+a_2e^{ik_p}(x-x_0) \qquad \text{For} \quad x<x_0 \tag{19}$$

$$G(x,x_0)=b_1e^{-k_d}(x-x_0)+b_2e^{-ik_p}(x-x_0) \qquad \text{For} \quad x>x_0 \tag{20}$$

By imposing these conditions, we can determine the Green function $G(x,x_0)$.

$$G(x,x_0)=\frac{1}{B}\left[F_d\,e^{kd|x-x_0|}+iF_p\,e^{-ikp|x-x_0|}\right] \tag{21}$$

with: $F_d=\frac{1-\frac{B}{K}[-kp^2+kc^2]}{2kp\,(kp^2+kd^2)}$, $F_p=\frac{1-\frac{B}{K}[kp^2+kc^2]}{2kp\,(kp^2+kd^2)}$ (22)

F_d: Corresponds to the propagative waves.
F_p: Corresponds to the waves at nearfield.

3. The free motion of the periodically supported beam: Bloch waves

➢ Euler beam

We consider a beam periodically supported with N supports placed between $-\infty$ and $+\infty$. The period is given by the regular spacing l. The rigidity of the N ième discrete elastic support, corresponding to the deformation of the system (elastic support, mass support...), is Z_n , the force transmitted by the discrete support is:

$$F_n=-Z_nU(n,l) \tag{23}$$

It is considered that all the discrete supports are characterized by the same rigidity $Z_n=Z$. The global response of the structure can be described by applying the superposition principle, which consists of the linear summation of all the answers due to the various supports. Then, the displacement with the position x is given by:

$$\eta(x)=\sum_{-\infty}^{+\infty}G(x,x_n)F_n \tag{24}$$

 ISBN: 978-84-614-6167-7

Where $G(x, x_n)$ is given by the equation (9) for the Euler beam and by equation (21) for the Timoshenko one.

The relation (9) is thus valid for any point x_m.

$$\eta(x_m) + \sum_{-\infty}^{+\infty} Z\eta(x_n) G(x_m, x_n) \tag{25}$$

The application of the Floquet theorem gives the relation between two adjacent supports :

$$\eta(x_{m+1}) = \eta(x_n) e^{-\gamma l} \tag{26}$$

With: $$\gamma = \alpha + ik \tag{27}$$

γ is the Bloch constant propagation. It is generally complex. α is the wave attenuation and k is the Bloch wave number [1].

By substituting the expression of the Green function given by the equation (9) in the equation (25) and by multiplying this equation by $e^{\gamma m l}$, changing the index of summation $m - n$ by n gives the following equation which defines the relation between γ and the displacement η_0 .

$$\eta_0 + Z\eta_0 \sum_{-\infty}^{+\infty} \left[\frac{1}{4EIk_b^4} \left(i e^{ik_b|n|l} - e^{k_b|n|l} \right) e^{\gamma n l} \right] = 0 \tag{28}$$

$$\eta_0 + Z\eta_0 \sum_{-\infty}^{+\infty} \left[\frac{1}{EI} \left(g_d e^{ik_b|n|l} - i g_p e^{k_b|n|l} \right) e^{\gamma n l} \right] = 0 \tag{29}$$

In the following, only the waves not attenuated will be considered, which means that the sum equation (27) and equation (28) can be evaluated by usual formulas of the geometrical series.

For the Euler model, we can have:

$$1 + \frac{Z}{4EIk_b^3} \left[\frac{\sin(k_b l)}{\cos(k_b l) - \cos(kl)} - \frac{\sinh(k_b l)}{\cosh(k_b) - \cos(kl)} \right] = 0 \tag{30}$$

$$cos^2(kl) + cos(kl) \left[-[coshk_b l + cosk_b l] + \frac{Z}{4EIk_b^3} (-sink_b l + sinhk_b l) \right] + \frac{Z}{4EIk_b^3} [sink_b l coshk_b l - sinhk_b l cosk_b l] + cosk_b l coshk_b l = 0 \tag{31}$$

This is a second degree polynomial for $\cos(kl)$. The solution produces the values of the wave number k which is between 0 and π . It is a function of k_b which depends on the radial frequency equation (3). It leads therefore to the.

 ISBN: 978-84-614-6167-7

- **Timoshenko beam**

In order to establish the dispersion equation, we use the suitable Green function and the same reasoning is carried out in the case of to the Timoshenko beam.

$$1+\frac{Z}{EI}\left[F_p\frac{\sin(k_p l)}{\cos(k_p l)-\cos(kl)}-F_d\frac{\sinh(k_d l)}{\cosh(k_d)-\cos(kl)}\right]=0 \quad (32)$$

Then, the dispersion equation is:

$$cos^2(kl)+\left[\frac{z}{EI}\left(F_d sinh\, k_d l-F_p sin k_p l\right)-\left(cosh k_d l+cos k_p l\right)\right]cos(kl)+\frac{z}{EI}\left(F_p cos k_d l\ sin k_p l-F_d cos k_p l\ sin k_d\right)+cosh k_d l\ cos k_p l=0 \quad (33)$$

4. Results and discussions

Numerical data:

Rail :

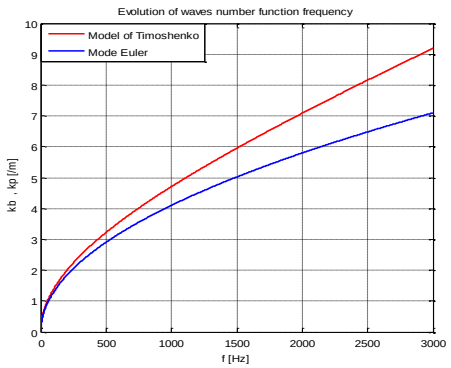
Modulus of elasticity:	$E=2.10^{11} N/m^2$
Shear Stiffness:	$G=0.77.\ 10^{11} N/m^2$
Mass per unit volume	$\rho = 8000\ kg/m^3$
Flexural Rigidity for Timoshenko beam	$E.I=6.4.\ 10^6\ Nm^2$
Flexural Rigidity for Euler beam	$E.I=4.8\ 10^6\ Nm^2$
Mass per unit length	M=60 Kg/m

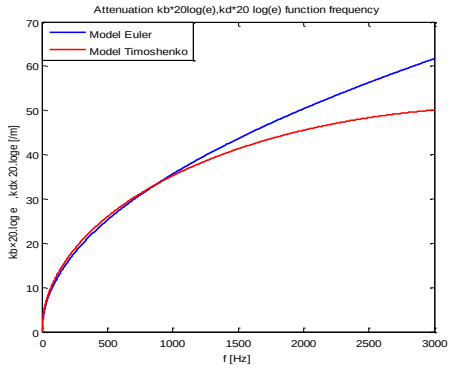
Track:

Sleeper mass	M=162 Kg/m
Rail pad stiffness	$S= 3.10^8$ N/m
Sleeper spacing	l=0.6 m

The waves numbers obtained with the two models are comparable at low frequencies, their differences is significant at 0.5 kHz figure 1. With high frequency, the Timoshenko model gives a more significant propagation with the waves numbers than the Euler model

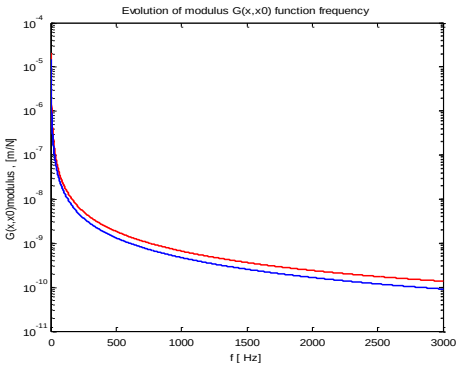
 ISBN: 978-84-614-6167-7

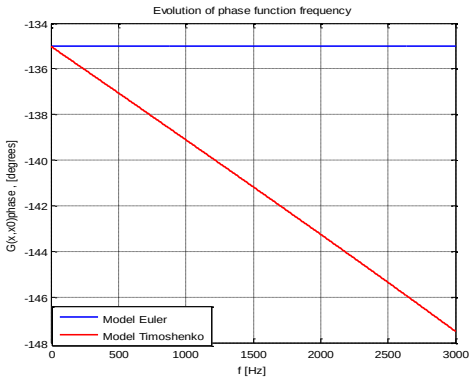




Propagative waves : Waves number m^{-1} Waves at nearfield: Attenuation dB/m

Figure 1 Comparison of the wave numbers and attenuation function frequency for two models





Modulus [N/m] Phase [degrees]

Figure 2 : Receptance G (x,x0) function frequency for tow models

Some computed results show how the receptance of the infinite periodic beam varies with frequency. This receptance evolution function of the frequency is shown figure 2.This evolution is identical for the two models.

The evolution of G(x, x0) according to x-x0 is given figure.3 for the frequency f=800 Hz. The nearfield effect is appreciable in the vicinity of the excitation point. At one meter the amplitude vibration remains constant.

 ISBN: 978-84-614-6167-7

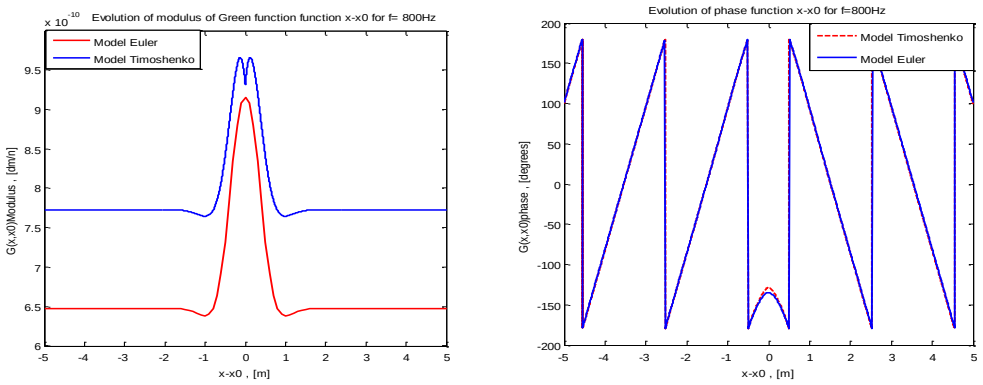


Figure 3: Evolution of G (x, x0) function x-x0 for f=800 Hz for tow models

The propagation constant γ is evaluated by the dispersion equation 30 and equation 32 respectively for the Euler model and the Timoshenko one. It is shown as a function of frequency. The analysis is carried out for the two types of supports considered: a mass periodic support and a rigid periodic support.
The attenuation spectrum shows bands of zero attenuation (passing bands) alternating with bands of positive attenuation (stopping bands).
The solution corresponding to the near fields is null (figure 1).
The other solution shows the alternation of the bands. This phenomenon of alternation was also observed by Heckl (**Maria A. Heckl, 2001)** [13] for the compression waves and torsion ones.
The curves obtained show that the flexion waves are not clearly separate. These curves highlight the behavior passing/stopping bands.
The width of the passing band increases with increasing frequency, facilitating the propagation of the waves not attenuated.

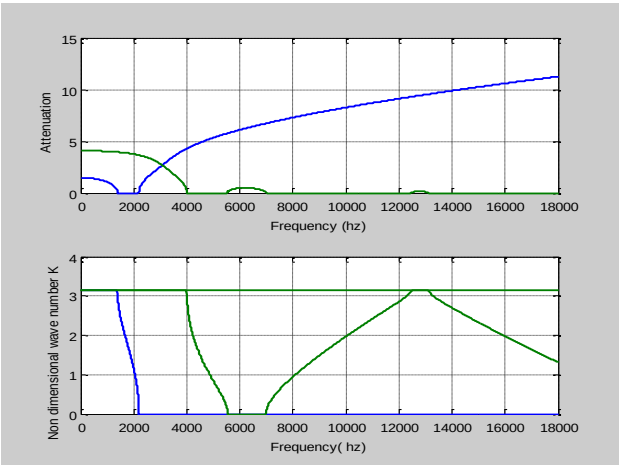


Figure 4: Vertical flexion wave for the Euler beam for an elastic support
a) Attenuation spectrum, b) Bloch wave number spectrum

 ISBN: 978-84-614-6167-7

For the mass support, at the frequency $\omega = 0$, we have a passing band; this particular value causes problems of indetermination. Some authors (**D.J Mead, 1986)** recommend that this zone can be considered as an attenuation band.

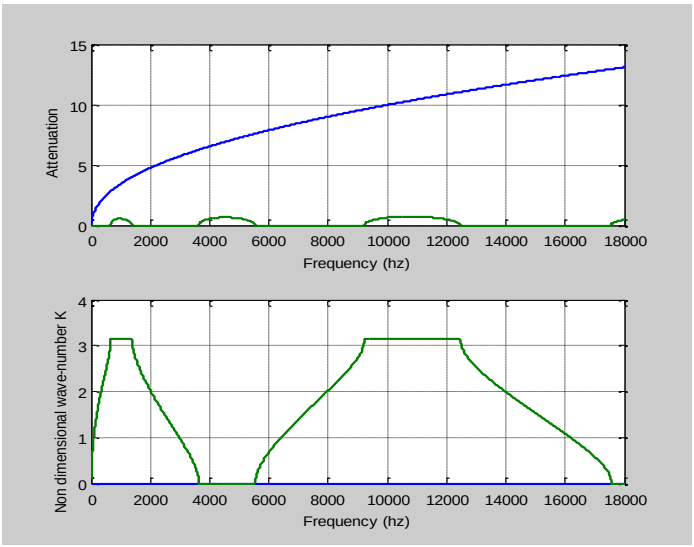
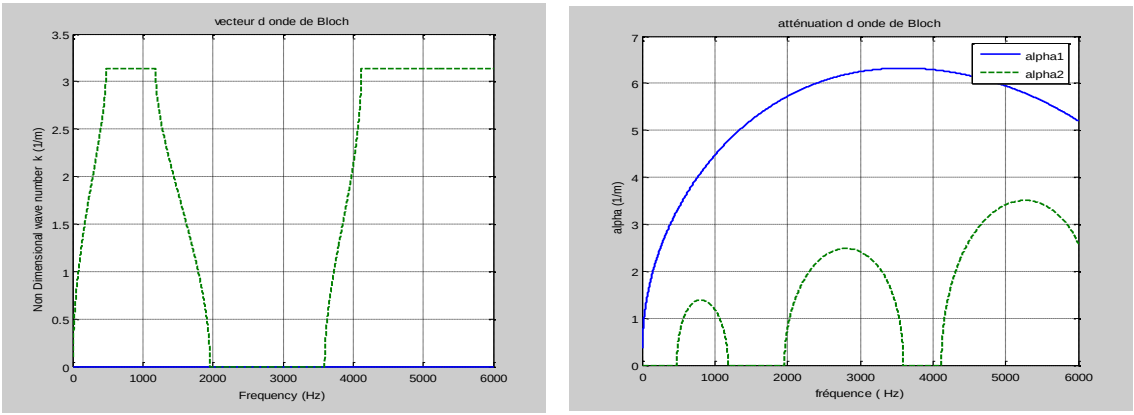


Figure 5: Vertical flexion wave for the Euler beam for a mass support
a)attenuation spectrum , b) Bloch wave number spectrum

For the Timoshenko beam on periodic supports, we can observe that the width of passing bands is smaller than the Euler beam figure 4.



a- Bloch wave number spectrum b- Attenuation spectrum

Figure 6: Vertical flexion wave for the Timoshenko beam for a mass support:

 ISBN: 978-84-614-6167-7

The stopping bands are not equally spaced, but increase with increasing frequency figure 6

5. Conclusion

We have presented an analytically approach to determine the dynamic response of the beam rail modeled by an Euler beam and Timoshenko one. These latter are infinitely long beams periodically supported. This approach is based on the Green functions, the Bloch theorem, and the superposition principle.

The established analytical model reflects the physics of the problem: infinite beam, by the use of the Green functions. It permits to describe the vibration of the structure under the influence of a unit force. These Green's functions are analyzed.

Furthermore, the Bloch theorem enabled us to analyze the free vibrations of the structure and to clarify the passing/stopping bands of Bloch waves.

In the case of vertical flexion wave, the dispersion relation spectrum showed a clear passing/stopping band behavior.

Two support types have been examined in detail: mass and elastic support.

More complex supports can be easily incorporated into the model.

This work will also permit to evaluate the dynamic response of an infinite structure on periodic supports under a moving load.

5. References

[1] BRILLOUIN *Wave propagation in periodic structures*, New York .1946 Dover Publications, Inc.

[2] M.HECKL, *Wave Propagation on Beam-Plate System*, J. Acoust.Soc. Am. **33** 5 (1960) 640–651.

[3] G. S., GUPTA, (1970), *Natural Flexural Waves and the Normal Modes of Periodically-Supported Beams and Plates*, J. Sound Vib. **13** 1 (1970) 89–101.

[4] B. R., MACE, *Periodically Stiffened Fluid-Loaded Plates, I: Response to Convected Harmonic Pressure and Free Wave Propagation*, J. Sound Vib. **74** (1980) 473–486.

[5] B. R. MACE, *Periodically Stiffened Fluid-Loaded Plates, II: Response to Line and Point Force*, J. Sound Vib. **73** 4 4 (1980) 87–504.

[6] D. J. MEAD, *Free Wave Propagation in Periodically Supported, Infinite beams*, J. Sound Vib**. 11** 2 (1970)181–197.

[7] M.J. MUNJAL, M.A. HECKL, *Vibration of a periodic rail- sleeper system excited by an oscillating stationary transverse force*, J .Sound. Vib. 81 (1982) 491-500.

[8] A.NORRDBORG. (1998). *Vertical rail vibration: point force excitation*, Acustica, 84,280-288.

 ISBN: 978-84-614-6167-7

[9] D.J. MEAD. *A general theory of harmonic wave propagation in linear periodic systems with multiple coupling*, J .Sound. Vib. 27 (1973) 235-260.
[10] D.J. MEAD. *A new method of analysing wave propagation in periodic structures: application to periodic Timoshenko beam and stiffened plates,* J. Sound. Vib. 104 (1986) 9-27.
[11] D.J.MEAD. *Wave propagation in continuous periodic structures: research contribution from Southampton, 1964-1995.* J. Sound. Vib. 190, (1996) 495-524.
[12] P.M.BELOTSERKOVSKIY. *On the oscillations of infinite periodic beams subjected to a moving concentrated force,* J. Sound.Vib. 193 3 (1996) 705-712.
[13] M. A. HECKL. *Coupled Waves on the Periodically Supported Timoshenko Beam.* J .Sound .Vib. 252 5 (2001) 849-882.

 ISBN: 978-84-614-6167-7

Proceedings of the 11th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE2011
Benidorm, Spain, 26-30 June 2011

Using statistical similarity measure and mathematical morphology for oil slick detection in Radar SAR images

Bahia Lounis[1], Grégoire Mercier[2] and Aichouche Belhadj-Aissa[1]

[1] *Université des Sciences & Technologie Houari Boumediene (USTHB), BP32 El Alia, Bab Ezzouar - Fac. d'Electronique & Informatique / LTIR 16111, Alger, Algeria.*

[2] *Institut Telecom; Telecom Bretagne- CNRS FRE 3167 lab-STICC, team CID. Technopole Brest-Iroise, CS 83818, F-29238 Brest Cedex 3, France*

emails: lounisbahia@yahoo.fr, h.belhadj@mailcity.com

Abstract

Spaceborne Synthetic Aperture Radar is well adapted for ocean pollution detection independently from daily or weather conditions. As it is sensitive to surface roughness, the presence of oil film on the sea surface decreases its backscattering resulting as dark feature patches in SAR images. In this paper, a new method for slicks detection is presented. We propose the combination of a statistical similarity measure and derivative morphological profile to isolate dark spots which are candidates to be oil slicks in SAR images. The first level of detection process is based on the measure of similarity between a local probability density function (pdf) of clean water and the local pdf of the zone to be inspected. The local distribution is estimated in the neighbourhood of each pixel and compared to a reference distribution using the Kullback-Leibler KL distance between distributions. Once spots highlighted, the second level of the process improve the detection by texture features extraction using the Derivative Morphological Profile. The algorithm has been applied to Envisat ASAR image. It yields accurate segmentation results, even for thin slicks, with a limited number of false alarms

Key words: Statistical similarity measure, Derivative Morphological Profile, SAR image analysis, Oil slicks detection, Water pollution.

 ISBN: 978-84-614-6167-7

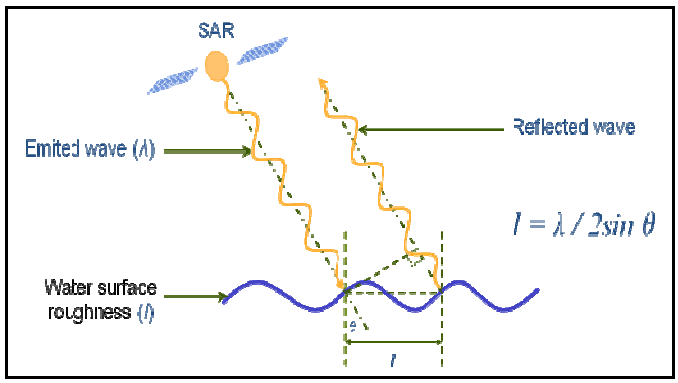
1. Introduction

Oil spills cause substantial damage to the marine environment and thus demands for necessary prevention measures. The major source of the total ocean oil pollution is caused by operational tanker oil discharges (45%) while ship accidents and oil platform accidents contribute to only 5% and 2% respectively [1]. The mapping, monitoring and statistical analysis of illicit ship discharges is thus an important component to provide reliable information to political decision makers as well as to ensure compliance with the marine protection legislation.

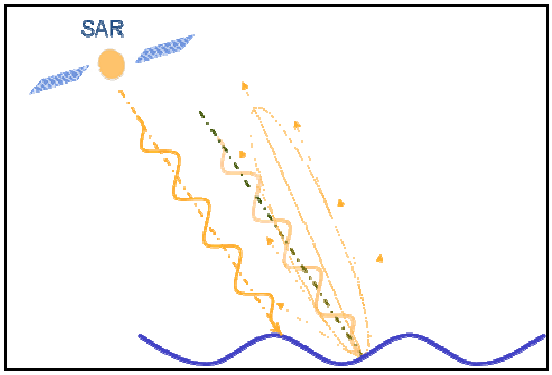
For this investigation, Synthetic Aperture Radar SAR images are widely used to monitor and therefore to detect oil pollution since they provide regularly images both day and night, even when clouds are present. As it is sensitive to surface roughness (see figure 1), the presence of oil film on the sea surface decreases its roughness and induce specular backscattering resulting as dark spots in SAR images [1,2]. However, dark areas may be also caused by other phenomena, like locally low winds, currents or natural sea slicks called "lookalikes" [2,3]. A typical example of SAR images is showed by figure 2 which presents the Prestige wreck occurred in November 2002 near the Spanish coast. An ENVISAT ASAR image was acquired on November 17, which is four days after the accident and two days before it sunk [3]. Wreckage can be seen as strong backscatters on the bottom left of the extracted image. The slick by itself is easy to detect. Nevertheless, some dark areas due to atmospheric perturbation often induce false alarms.

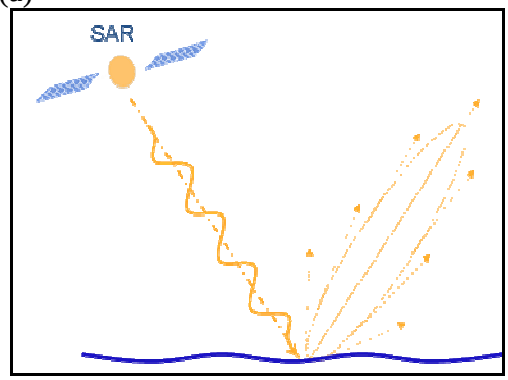
Generally, the detection process of the oil spill from SAR images consists of several successive steps: (i) detection of dark spots; (ii) spots characterization; and (iii) spots identification. The dark spot detection locates all spots which can possibly be oil slicks in the image. For each slick, a set of backscatter, textural, and geometrical features are extracted and are, then, classified into possible oil slicks and look-alikes [2].

 ISBN: 978-84-614-6167-7



(a)





(b) (c)

Figure 1: Principle of marine SAR image acquisition. (a). Bragg backscatter, (b). Volumic backscatter due to surface roughness, (c). Specular backscatter

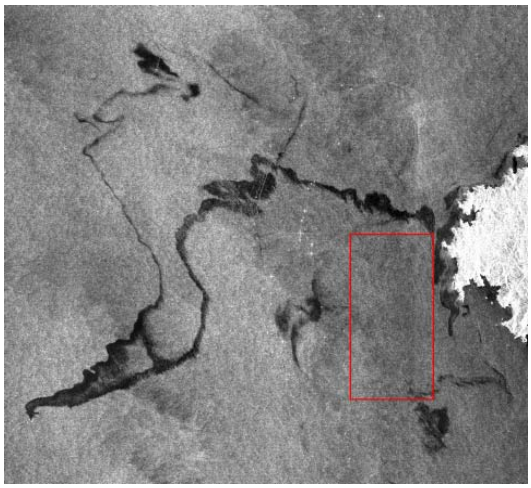


Fig. 2. Envisat ASAR image. The rectangular area characterizes the ROI that defines normal sea state taken as a reference.

 ISBN: 978-84-614-6167-7

In this paper, we are interested by the first step of this process in which the main objective is to isolate the dark patches visible in the SAR image. For this task, many techniques have been proposed in the literature. Among the most widely used, the approach based on adaptive thresholding, Fuzzy segmentation, textural analysis [1-3]. The detection results depend on the features extracted from the original image and the classification process used. In this context, we propose a new method in a semi-supervised mode that consists in three following steps. In the first step, we use a Kullback-Leibler Distance KLD [4] as similarity measure between a local probability density function (lpdf) of clean water and the local lpdf of the zone to be analysed. This step is described in Section 2. Once spots are highlighted, we focussed our interest, in the second level of detection process, to features extraction about spots geometric. In this context, we exploit the local geometrical information of each pixel using the derivative morphological profile (DMP) obtained with a granulometric approach [5], this presented in Section 3. Finally, all features extracted from the original image are integrated into a classifier process in order to discriminate between clean water and oil spill pixels and thus extract oils signatures. We chose a supervised classifier which is based on the dispersion of the data through the variance-covariance matrix extracted from the training data. This task is achieved by the likelihood classifier [6] described in Section 4. The first results obtained on Envisat ASAR image of Prestige tanker are presented and discussed in Section 5. Finally, in section 6, conclusions on the developed approach are presented in order to show the contribution of SAR images data in oil spills detections.

2. SIMILARITY MEASURE

Slick detection is performed by comparing the local pdf estimated within a sliding window and a reference pdf defined by an expert over a region of interest (ROI) characterizing a normal sea state.

2.1. Local pdf estimation

Let's consider a set of n samples $\{x_1, x_2..., x_n\}$ taken from a sliding window characterizing a random variable X. The lpdf $f_X(x)$ of X is estimated through a non parametric approach by using a cumulant-based approximation. These cumulants allowed an efficient description of lpdf shape, for example, third order k_3 is linked to the symmetry (i.e. skewness), while the fourth k_4 to the flatness (i.e. kurtosis). The density is then estimated through a series expansion. Actually, the cumulant generating function is used for such estimation. The cumulant generating function $k_X(.)$ of a random variable X is defined by:

$$k_X(w) = \ln M(w) = \sum_n k_{X;n} \frac{w^n}{n!} \quad (1)$$

 ISBN: 978-84-614-6167-7

With $M(\cdot)$ being the moment generating function:

$$M_X(w) = \int e^{wx} f_X(x)dx = \int (1 + wx + \frac{w^2}{2}x^2 + ...) f_X(x)dx \tag{2}$$

For the case of the four first order cumulants, the following expressions hold [7] :

$$\begin{aligned} k_{X;1} &= \mu_{X;1} \\ k_{X;2} &= \mu_{X;2} - \mu_{X;1}^2 \\ k_{X;3} &= \mu_{X;3} - 3\mu_{X;2}\mu_{X;1} + 2\mu_{X;1}^3 \\ k_{X;4} &= \mu_{X;4} - 4\mu_{X;3}\mu_{X;1} - 3\mu_{X;2}^2 + 12\mu_{X;2}\mu_{X;1}^2 - 6\mu_{X;1}^4 \end{aligned} \tag{3}$$

Where $\mu_{X,i}$ are the means at order i .

Let us assume that the density to be approximated is not too far from a Gaussian pdf (denoted as g_X to underline the fact that it has the same mean and variance as X). The difference between $k_X(.)$ and $k_{g_X}(.)$, gives a link between $k_X(w)$ and $k_{g_X}(.)$ through the difference of the cumulants $k_{X;n} - k_{g_{X;n}}$. By inversion, the density may be expressed by a formal Taylor-like series:

$$f_X(x) = g_X(x) + c_1 \frac{dg_X}{dx} + c_2 \frac{d^2 g_X}{dx} + ... \tag{4}$$

Since a Gaussian density is used, it comes:

$$f_X(x) = \sum_0^\infty c_r H_r(x) g_X(x) \tag{5}$$

with $H_r(x)$ known as the Chebyshev-Hermite polynomial of order r [8]. When choosing a Gaussian law so that its first and second cumulants agree with those of X , the number of terms of the series expansion is greatly reduced. This is the so-called Edgeworth series expansion. Its expression, when truncated to order 6, is the following:

$$f_X(x) = \left(1 + \frac{k_{X';3}}{6} H_3(x) + \frac{k_{X';4}}{24} H_4(x) + \frac{k_{X';5}}{120} H_5(x) + \frac{k_{X';6} + 10k_{X';3}^2}{720} H_6(x)\right) g_X(x) \tag{6}$$

Where $X' = (X_g - k_{X;1}) k_{X;2}^{-1/2}$.

 ISBN: 978-84-614-6167-7

Figure 3 shows an example of such an approximation of a histogram taken from an Envisat ASAR image, in a heterogeneous area where an oil slick is mixed up to clean water.

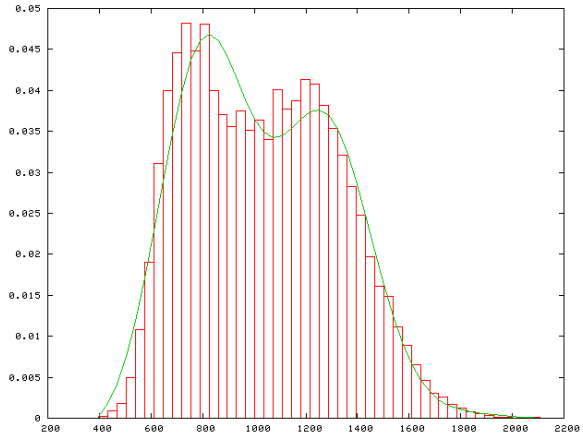


Fig. 3. Estimation of a histogram from a cumulant-based

2.2. Locals pdfs comparison

Let P_X and P_Y be two probability of the random variables X and Y. The KL divergence from Y to X, in the case where these two laws have the densities f_X and f_Y, is given by:

$$KL(Y/X) = \int \log \frac{f_X(x)}{f_Y(x)} f_X(x)dx \tag{7}$$

When the Edgeworth series expansion of the two pdfs f_X and f_Y (equation 6) is introduced into the Kullback-Leibler divergence, it yields an approximation of the Kullback-Leibler divergence by Edgeworth series, truncated at a given order. In [7], such approximation has been given up to order 4, where g_X (resp. g_Y) is a Gaussian density of same mean and variance as f_X (resp. f_Y). Then:

$$\begin{aligned} KL_{Edgeworth}(Y/X) = & \frac{1}{12}\frac{k_{X';3}^2}{k_{X;2}^2} + \frac{1}{2}\left(\log\frac{k_{Y;2}}{X;2} - 1 + \frac{1}{k_{Y;2}}(k_{X;1} - k_{Y;1} + k_{X;2}^{1/2})^2\right) \\ & - \left(k_{Y';3}\frac{a_1}{6} + k_{Y';4}\frac{a_2}{24} + k_{Y';3}^3\frac{a_3}{72}\right) - \frac{1}{2}\frac{k_{Y';3}^2}{36}\left(c_6 - 6\frac{c_4}{k_{X;2}} + 9\frac{c_2}{k_{Y;2}^2}\right) \\ & \_10\frac{k_{X';3}k_{y';3}(k_{X;1} - k_{Y,1})(k_{X;2} - k_{Y;2})}{k_{Y;2}^6} \end{aligned} \tag{8}$$

 ISBN: 978-84-614-6167-7

where the coefficients $a_1, a_2, a_3, c_2, c_3, c_4, c_6, \alpha, \beta$ are given in [8].
Since KL is not symmetric, the Kullback-Leibler distance $KLD_{Edgeworth}$ is, finally, defined by the following expression:

$$KLD_{Edgeworth}(X/Y) = K_{Edgeworth}(X/Y) + K_{Edgeworth}(Y/X) \quad (9)$$

2.3. Reference local pdf selection

The similarity measure is obtained by the comparison of the local pdf, estimated within a sliding window, and a reference pdf defined by an expert over a region of interest (ROI) characterizing a no-polluted sea surface. The KL distance is computed in a local window of increasing size. Smaller window size is appropriated to small slick detection but with the drawback of mixing up with many false alarms. On the contrary, larger window size ensures the detection but with the drawback to miss detect smaller slicks and slicks border smoothing. The rectangular area of figure 2 characterizes the ROI that defines normal sea state taken as a reference; Figure 4 presents a similarity measure between local pdf taken from a 15×15 sliding window and the pdf of a normal sea surface taken from the ROI shown on figure 2 through a KL evaluation of equation (9). The KLD vanishes only when the two pdfs (f_X and f_Y) are identical, then normal sea surfaces appear as black surfaces in similarity measure image.

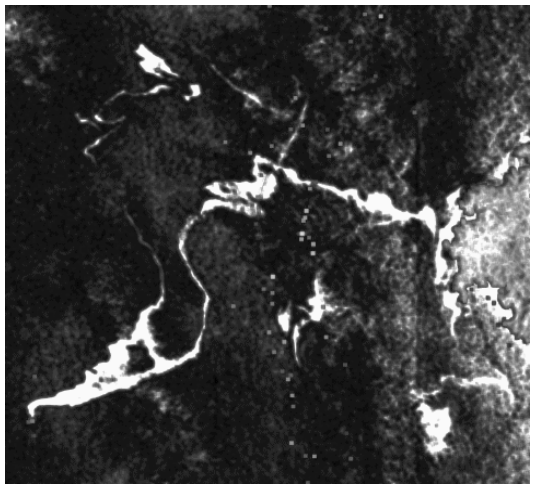


Fig. 4. The image presents similarity measure between local lpdf (taken from a 15×15 sliding window) and the lpdf of a normal sea surface (taken from the ROI shown on fig. 2) through a KL evaluation of eq. (9).

 ISBN: 978-84-614-6167-7

The visual analysing of this result shows an efficiency detection of slicks even for the small ones, but it yields some false alarms which are mainly due to the normal sea surface roughness [9]. To reduce them, we need further features about local contrast. We exploit, then, the morphological granulometry approach.

3. MORPHOLOGICAL FEATURE EXTRACTION

Granulometry is popular and powerful tool derived from the mathematical morphology theory. It has recently been introduced in remote sensing image processing for urban areas classification [5]. Besides the spectral signature, the incorporation of spatial information, achieved using Derivative Morphological Profile "DMP" in the classification process, improves well the classification results [5]. In this context, we propose to integrate this structural information in order to reduce the false alarms.

3.1. Derivative Morphological Profile DMP

The classical opening granulometry is obtained by applying morphological opening operations with structuring elements (SE s) of increasing size. The consequence is a progressive simplification of the image with a gradual disappearance of the features that are brighter than their immediate neighbourhood. Each structure is removed when it becomes smaller than the SE. Noting I as the original image and γ_S the morphological opening by reconstruction using a structuring elements SE, the morphological profile $MP_\gamma(i)$ at the scale i is defined for each pixel by:

$$\begin{aligned} &MP_\gamma(0) = I, \\ &MP_\gamma(i) = \gamma_{ES(i)}[I], \qquad i = 1,..., p. \end{aligned} \tag{10}$$

Let's note, here, that the scale designates the size of the structuring element.

Granulometric curves are often interpreted by computing their discrete derivative. After each morphological filter, the difference with the result of the previous scale is computed. For each pixel this result in the differential morphological profile DMP_γ, which is also known as the "pattern spectrum" [5], is given by:

$$DMP_\gamma(i) = MP_\gamma(i-1) - MP_\gamma(i), \qquad i = 1,..., p. \tag{11}$$

A structure is removed when the size of the SE reaches its characteristic size. Corresponding pixels are then assigned the gray-level value of the surrounding darker region. For all these pixels, this generates a large value on the DMP, a value representing the local contrast between the removed structure and its

 ISBN: 978-84-614-6167-7

surrounding region. Furthermore, the index of the iteration where the structure is removed provides an estimation of the size of the structure. Opening operations affect only structures that are brighter than their immediate neighborhood. Other structures are left unchanged and thus lead to a flat DMP, with a null value.

Similarly, a granulometry by closing is obtained using morphological closing operations, with the same set of structuring elements. In the same way, a derivative profile $DMP_{\Phi}(i)$ is obtained. Noting Φ_s the morphological closing by reconstruction, it is given by :

$$DMP_{\Phi}(i) = MP_{\Phi}(i) - MP_{\Phi}(i-1), \qquad i = 1,..., p. \tag{12}$$

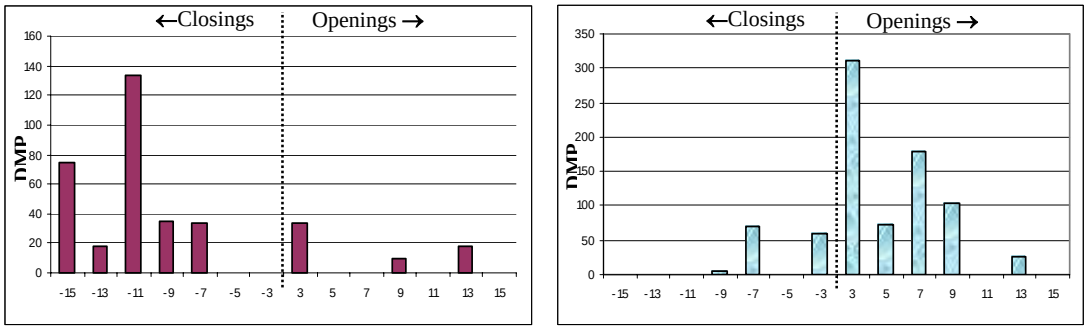
By duality to the opening operation, the closing-based profile provides information regarding the structures that are darker than their immediate surrounding and does not affect structures that are brighter than their surroundings.

Finally, in order to exploit the complementarity of both the bright and the dark structures of the image, the two $DMPs$ are concatenated as follows [5]:

$$\begin{aligned} DMP(i) &= DMP_{\gamma}(i), \qquad \text{for } i = 1,..., p \\ &= DMP_{\Phi}(-i), \qquad \text{for } i = -1,...,-p. \end{aligned} \tag{13}$$

Figure 5 presents two examples of DMP obtained from a normal sea surface water and slicks regions from the original image. One typical example is presented for each region. On each profile, the left part of scales $(i < 0)$ corresponds to the granulometries by closing, and the right part $(i > 0)$ to the granulometries by opening. In this case, the used structuring elements were increased successively 3x3, 5x5, 7x7, 9x9, 11x11, 13x13 and 15x15 in order to follow the image texture. It is noted that for each region, the DMP of most pixels has one side tendency either the “opening” side for objects brighter than their immediate surroundings or the “closing” side for objects darker than their immediate surroundings. Hence, for each pixel, the DMP provides a vector of 14-dimensional of attributes where each one characterizes a pixel by its local contrast given by DMP intensity and structure size presented by ES size. However, the most of these attributes are localised in one side of the graphs, a lot of redundancy can be seen. The information from all the features may not necessary. Thus, the DMP dimensional can be reduced by selecting the most significant ones.

 ISBN: 978-84-614-6167-7



(a) (b)

Fig. 5. Examples of typical differential morphological profiles (DMP(i)) obtained for various pixels type. (a) Oil pixel, (b) clean water pixel

3.2. Features selection based on sorting DMP indexes

In order to reduce the DMP vector dimensionality we propose to select the most discriminative ones without sacrificing the accuracy of the detection results. As previously stated, when the size of the structuring element reaches the size of one given structure, the structure is removed. This induces a noticeable change of the intensity values of the corresponding pixel and, as a consequence, a peak in the DMP. The intensity of this peak gives information of the local contrast of the pixel and the position $\arg\max_i \{DMP(i)\}$, of the greatest value within the DMP, is an estimate of the characteristic size of the structure to which the pixel belongs.

However, for capillary waves which caused textured area in SAR image, the different structures in the image do not have perfectly sharp edges. Consequently, as it is shown by figure 6, the DMP is not constituted of a single peak with a one-sample-wide support. Furthermore, the intensity of these peaks remains an information which can be used to discriminate between oil and clean water pixels. In this work, we are interested to characterize only the local contrast of sea SAR image given by the intensity of the DMP. Thus, we propose then a simple method by sorting the DMP indexes in descending order and extract only the first n ones. The goal is to find stepwise the "n indexes" which gives good classification accuracies without much computation time.

3. CLASSIFICATION

In the last step of slicks detection, all selected features extracted from the original image are integrated into a classifier process in order to discriminate between clean water and oil slicks pixels. For this task, we have used a Maximum likelihood classifier which is considered to be the most common supervised classification method used with remote sensed images data [6].

 ISBN: 978-84-614-6167-7

Besides its simplicity, this method gives satisfactory classification results. Its principle is derived from the probability theory and the Bayes' theorem, and is based on the search for the likelihood measurement between the pixel to classify and the data samples represented by their statistical parameters (mean, variance-covariance matrix) which are extracted from the training data. Generally, the maximum likelihood classification assumes that the statistics for each class in each feature are normally distributed and calculates the probability that a given pixel belongs to a specific class. Each pixel is assigned to the class that has the highest probability. This is achieved by calculating the following discriminant functions g_i for each pixel x [6]:

$$g_i(x) = \ln p(C_i) - \frac{1}{2}\ln\left|\sum_i\right| - \frac{1}{2}(x-m_i)^T.\sum_i^{-1}.(x-m_i) \tag{14}$$

with $x \in C_i \quad if \quad g_i(x) > g_j(x) \quad for \quad all \; j \neq i$

where:

i=class ($i = 1 \rightarrow C_1 = water\ class, i = 2 \rightarrow C_2 = oil\ class$)

x=N-dimensional data representing pixel intensity and its extracted features (KLD,DMP(1…n)),

$p(C_i)$=probability that class C_i occurs in the image,

$\left|\sum_i\right|$=determinant of the covariance matrix of the data in class C_i,

$\sum_i^{-1}$ =its inverse matrix,

m_i=mean vector of class C_i.

The validation of the classification process is evaluated according to two criteria: visual inspection of oil signatures and some statistics parameters computed from the confusion matrix, also called error matrix. The diagonal elements of this matrix express a pixel count of correctly classified pixels, while the non-diagonal elements represent the number of pixels that have been incorrectly classified. The reader should note that, because of the lack of ground truth map, this matrix is computed from control database.

Generally, the classification quality is evaluated by the statistic parameter OA(Overall accuracy) which designates the percentage of correctly classified pixels. In our application, in addition to OA, the classification is evaluated by the use of confusion matrix in term of the designated target O (oil slicks). In this case, the probability detection Pd is assimilated to the precision of oil signatures detection which corresponds to the oil pixels correctly classified. Therefore, the probability of false alarms reported in normal seawater W noted P_{fa} becomes equivalent to the commission error to detect the no-polluted seawater [3].

 ISBN: 978-84-614-6167-7

Let's define Mc the confusion matrix constituted of $Mc(i, j)$ elements. These statistic parameters are given by following equations.

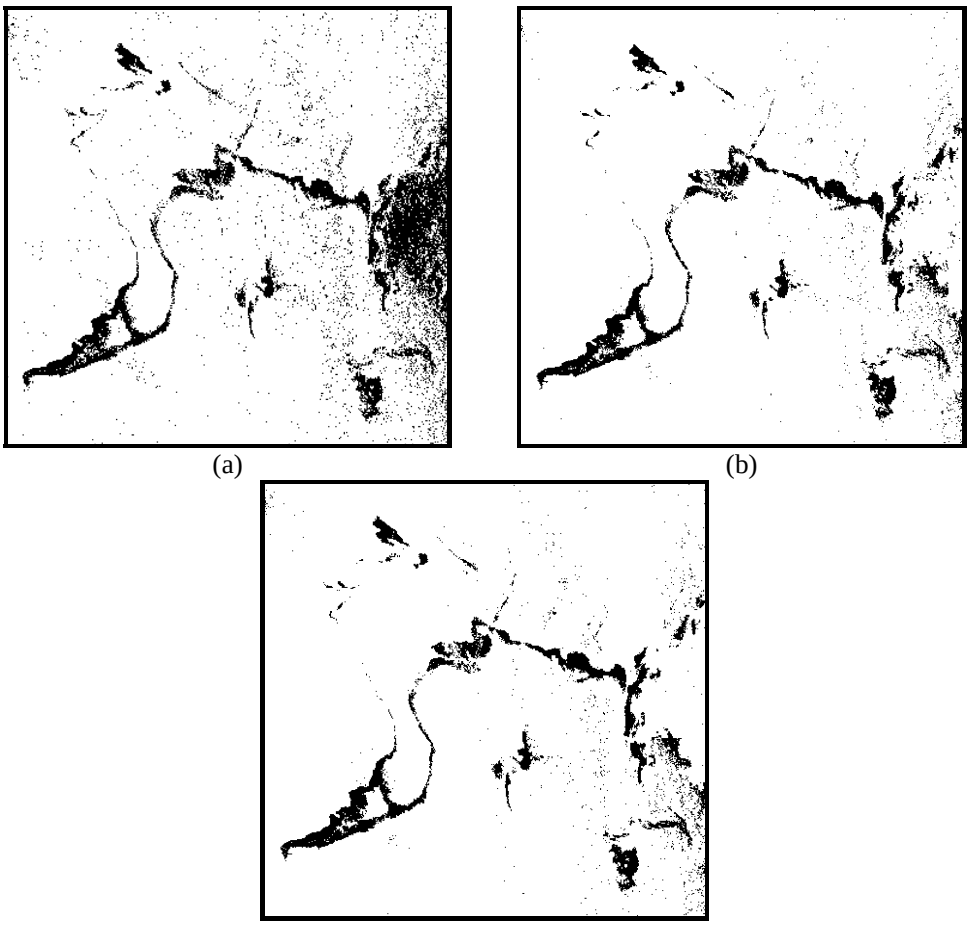
$$\begin{cases} OA = 100.\dfrac{\sum_{1}^{c} Mc(i,i)}{\sum_{i=1}^{c}\sum_{j=1}^{c} Mc(i,j)} \\ Pd = 100.\dfrac{Mc(O,O)}{\sum_{i=1}^{c} Mc(i,O)} \\ Pfa = 100.\left(1 - \dfrac{Mc(W,W)}{\sum_{i=1}^{c} Mc(i,W)}\right) \end{cases} \tag{15}$$

Where O and W designate respectively oil and Water classes and c represents the confusion matrix size.

4- RESULTS AND DISCUSSION

The proposed approach was applied to ENVISAT ASAR test image showed by figure 2. For each pixel x of this image, we constitute a vector of dimension 16 composed by its SAR intensity $I(x)$, its similarity measure $KLD(x)$ and its corresponding sorted features $DMP(x)$. Different feature combinations were tested in order to select the best one. Figure 6.a shows the detection result using the first feature only, which is the KLD measure (slicks are presented in black colour). The main dark patches are efficiently detected, in particular, the extensions of the spills to the north-west. This result is coherent with the probability detection Pd which is found to close to 95,49% (Table 1) . However, it yields some false alarms as it is noted by the probability of false alarms ($Pfa = 4{,}05$). Besides the KLD feature, the integration of the sorted DMP information improves the classification results in terms of OA, Pd and Pfa as illustrated in Table 1. It could be seen that the Pfa decreases from 4,05% obtained with ASAR and KLD feature classification to 2,55% which corresponds to the use of ASAR, KLD and all DMP features. A flattening performance is observed from the seventh feature. It corresponds to the sixth max of DMP. The visual comparison of figures 6.b and 6.c confirms that the six first DMP peaks are sufficient to detect slicks with the same accuracy as the full DMP features. Furthermore, the suppression of the non-discriminative features (from the 7^{th} to 14^{th} feature of DMP) reduces the computation time without scarifying the quality of the slicks detection (Fig. 6.b and Fig. 6.c).

 ISBN: 978-84-614-6167-7



(a) (b)

(c)

Fig. 6. Detection results. (a) Detection results using only KLD feature. (b) Detection results using ASAR with KLD and all DMP features. (c) Detection results using ASAR with KLD and reduced DMP features.

Number of features	*Pd* (%)	*Pfa* (%)	*OA* (%)
1	95,49	4,05	95,91
2	95,68	3,89	96,21
3	95,77	3,80	96,26
4	96,14	3,48	96,35
5	96,49	3,17	96,42
6	96,76	2,93	96,47
7	96,96	2,76	96,54

Table 1: Statistic classification evaluation of ASAR image for different features integration

 ISBN: 978-84-614-6167-7

5- CONCLUSION

In this paper, a new oil-slick detection approach was proposed using the combination of Kullback-Leibler (KL) distance and the Derivative Morphological Profile. Similarity measure was used for image inspection in order to highlights spots suspected to be an oil slick and is considered as the first extracted feature. Further features characterizing a local contrast of the original image are deduced using the derivative morphological profile and sorted in order to reduce DMP dimension. Finally, a supervised classifier is applied to extract slicks signatures.
Applying to ASAR image, the first results show a good detection of different slicks shapes with reducing false alarms. Furthermore, the proposed approach is strategic to detect oil slicks as it can be applied on large images, and efficient since the detection is similar to than classical tools. Actually, deeper validation has to be made by:

- Reduce human supervision by considering automatic selection of reference ROI used in KLD measure and choose an unsupervised classifier;
- Considering synergetic data derived from the same sensor (like ASAR for wind only) or any other sensor in the classifier process. Integrate such meteo-oceanic information may significantly reduce false alarms from classical look alikes.

References

[1] European Space Agency (ESA), *Oil pollution monitoring in ERS and its applications: marine*, Vol. 1, BR-128 (The Netherlands : ESA Publications Division), 1998.

[2] C.D. Clark, *Satellite remote sensing for marine pollution investigations, Marine Pollution Investigations*, Vol. 26 (1993), p 357,.

[3] G. Mercier and F. Girard-Ardhuin, *Partially Supervised Oil-Slick Detection by SAR Imagery Using Kernel Expansion*, IEEE Transactions on geoscience and remote sensing, Vol. 44, N°. 10 (2006), pp 2839-2848.

[4] J Inglada and G Mercier, *A New Statistical Similarity Measure for Change Detection in Multitemporal SAR Images and its Extension to Multiscale Change Analysis*, IEEE Trans. Geosci. Remote Sensing, vol. 45, no. 5 (2007), pp. 1432–1446.

[5] J. At. Benediktsson, M. Pesaresi and K. Arnason, 2003, *Classification and Feature Extraction for Remote Sensing Images From Urban Areas Based on Morphological Transformations*, IEEE Transactions On Geoscience And Remote Sensing, Vol. 41, N°. 9 (2003)

[6] R. Gonzalez and R. Woods, *Digital Image Processing*, 2nd ed: Prentice Hall, 2002.

 ISBN: 978-84-614-6167-7

[7] P. McCullagh, *Tensor Methods in Statistics*, Chapman and Hall, London, 1987.
[8] J. Lin, N. Saito, and R. Levine, "Edgeworth approximation of the Kullback-Leibler distance towards problems in image analysis," Tech. Rep., University of California, Davis, 1999, http://www.math.ucdavis.edu/˜saito.
[9] B. Lounis, G. Mercier & A. Belhadj aissa, Statistical similarity measure for oil slick detection in SAR image, IEEE Geoscience and Remote Sensing Society, the IGARSS 2008, Boston, Massachusetts, USA, July 6-11, 2008.

 ISBN: 978-84-614-6167-7

Proceedings of the 11th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE2011
Benidorm, Spain, 26-30 June 2011

On Techniques for Improving On-line Optimisation of Processes

M. Mansour [1]

[1] *Faculty of Electronics and Computer Science, University of Sciences and Technology, PO. BOX 32, El_Allia. Bab-Ezzouar, Algiers, Algeria*

Emails: M.Mansour@city.ac.uk,

Abstract

The solution of an on-line optimisation problem generally necessitates the calculus of derivative information. This information is needed in order to satisfy first and second order optimality conditions. There are several methods available for calculating these derivatives. In this paper methods and techniques for the estimation of the derivative information, to be used within the well-known ISOPE (Integrated System Optimization and Parameter Estimation) algorithm are investigated. Methods of Finite Difference Approximation, Dual Control Optimization, Broydon's method, Dynamic Model Identification Method, with a Nonlinear model, together with a novel neural networks scheme are presented and applied, under simulation, to a cascade Continuous Stirred Tank Reactor (CSTR) system. The results are then discussed and compared to identify the most suitable method among those used.

Key words: On-line optimization, model-based, process derivatives, ISOPE algorithm, ANN.
MSC2000: AMS Codes (optional)

1. Introduction

The requirement for processes to operate at their optimum operating condition is becoming increasingly prevalent. One such model-based algorithm that has been developed and which can achieve optimum process operation in spite of model-reality mismatch is the Integrated System Optimisation and Parameter Estimation (ISOPE) algorithm [1]. One requirement of the ISOPE algorithm, in order to

 ISBN: 978-84-614-6167-7

satisfy the necessary optimality conditions, is the need for estimates of real process derivatives. These derivatives are estimated on-line at each iteration of the algorithm. The finite difference method originally used by Roberts [1] to estimate these derivatives has proven not to be efficient in the case of large, slow and noisy processes [2]. Alternative methods have therefore been developed. The dynamic model identification technique, which is based on the identification of a dynamic model, was incorporated within the ISOPE algorithm by Zhang and Roberts [3]. Although this technique proved to be fast enough, it encountered some difficulties such as: the huge amount of data needed and the poor, inaccurate, model it produces at the beginning of the identification. After that, an algorithm with dual control effect was proposed [4]. In this algorithm the current control signal is generated to satisfy the main control goal and at the same time provide sufficient information for future identification action. The main advantage of this algorithm is that it does not need excessive set-point changes to estimate the process derivatives. However, this method encountered the same type of problems as the previous ones. Broydon's approximation method based on the well-known Broydon's family of formulas which are mainly oriented to the approximation of derivatives was also implemented [5]. Lately, a nonlinear version of the dynamic model identification was applied and implemented [2]. In this paper, a review of all these techniques together with a method based on artificial neural networks is presented. Simulations are carried out on a cascade CSTR system and a comparison is made to show the advantages and disadvantages of each method.

2. The Optimization Problem and the ISOPE algorithm

The ISOPE algorithm (or modified two steps) was proposed by Roberts [1] to solve the general optimisation problem of finding the optimum operating point of a system while it is moving one one point to another. It uses an adaptive steady-state model of the process, in which the parameters are updated periodically by comparing model outputs with those of the real process.

The general form of the algorithm is:

1. Apply the current input $\tilde{v}_k$ to the real process and obtain steady-state measurement $\tilde{y}_k^*$. Then use the mathematical model to determine the model parameters $\tilde{\alpha}_k$ to minimise the comparison index given by:

$$\begin{aligned} &\underset{\tilde{\alpha}}{Min}\ G(\tilde{u},\tilde{\alpha}) \\ &\tilde{y} = H(\tilde{v}) \\ &g(\tilde{u},\tilde{\alpha}) \le 0 \end{aligned} \tag{2-1}$$

 ISBN: 978-84-614-6167-7

where

$$G(\tilde{u},\tilde{\alpha}) = \sum_{i=1}^{r} \tilde{w}_i (\tilde{y}_i^* - h_i(\tilde{u},\tilde{\alpha}))^2 \tag{2-2}$$

and $\tilde{w}$ is a weighting vector.

2. Solve the modified model-based optimisation problem given by:

$$\begin{aligned} &\underset{\tilde{u}}{Min}\ (Q(\tilde{u}, H(\tilde{u},\tilde{\alpha})) - \lambda\tilde{u}) \\ &\tilde{y} = H(\tilde{v}) \\ &g(\tilde{u},\tilde{\alpha}) \le 0 \end{aligned} \tag{2-3}$$

In order to obtain the new candidate $\tilde{u}_{k+1}$. Where

$$\lambda = \left[\left[\frac{\partial\tilde{y}}{\partial\tilde{v}}\right]^T - \left[\frac{\partial\tilde{y}^*}{\partial\tilde{v}}\right]^T\right]\left[\frac{\partial\tilde{y}}{\partial\tilde{\alpha}}\right]^{-1}\left[\frac{\partial Q}{\partial\tilde{\alpha}}\right] \tag{2-4}$$

λ is called a modifier and is obtained following consideration that the necessary optimality conditions, of the system optimisation problem, have to be satisfied ([1]; [6]; [7]).

However, the new control $\tilde{u}_{k+1}$ is not directly applied to the system. Instead, the following relaxation scheme is used:

$$\tilde{v}_{k+1} = \tilde{v}_k + K(\tilde{u}_{k+1} - \tilde{v}_k) \tag{2-5}$$

where *K* is a relaxation gain matrix and is a tuning parameter.

These steps are repeated until convergence is reached. Convergence occurs when no further improvement is observed. In other words, when the new control is no longer a better candidate than the previous one.

From the previous cited relations, it can be seen that the requirement of the ISOPE algorithm to measure derivatives of real process outputs $\left[\partial\tilde{y}^*/\partial\tilde{v}\right]$ imposes a practical limitation to the technique. These process derivatives are calculated online, usually by applying small perturbations on the set-points and measure the resulting changes on the outputs. This process is repeated at each iteration of the algorithm. Various techniques exist and have been developed and applied for the purpose of estimating these derivatives. The Finite differences technique was originally suggested with the modified two step method [1]. Dynamic model identification (DMI) was then applied by Zhang and Roberts [3]. An algorithm for dual control effect was also suggested and implemented [4], and most recently DMI with a nonlinear model was proposed and implemented on a two CSTR system [2]. In this work, a method based on Artificial Neural Networks

 ISBN: 978-84-614-6167-7

(ANN) to estimate the real process derivatives and predict future control actions is presented. In this method, a neural network model of the real system is created, trained and adapted to the behaviour of the system. This model, imitates the behaviour of the real system within its limits. After that, this steady-state model is used to estimate the real system output derivatives with respect to the set-points in order to compute the parameter λ. All these technique are implemented and tested under simulation on a two CSTR system.

3. Estimation techniques

3.1. Finite Difference Approximation Method (FDAM)

This is the most straightforward method for calculating derivatives. It is simply:

$$D_k = \frac{\partial y}{\partial v} \approx \frac{y(v_k + \delta) - y(v_k)}{\delta} \tag{3-1}$$

where δ is a small perturbation signal applied to the system in practice in order to estimate the derivative matrix and y and v are the output and manipulated (set-point) variables respectively.
This method can give sufficient accuracy of the derivatives in acceptable time spam, in the case of small and noise-free processes which have reasonably rapid dynamics. However, it has shown to be inefficient for large and slow processes because of the huge amount of time taken for the estimation, and also because of the inaccuracy of the measurements for noise-contaminated processes.

3.2. Method for Dual control optimization

The algorithm assumes the existence of a collection of $n+1$ points $v^i, v^{i-1}, \ldots, v^{i-n}$ such that all vectors

$$\Delta v^{ik} \stackrel{def}{=} v^i - v^{i-k} \tag{3-2}$$

are linearly independent, i.e.

$$\det(S^i \stackrel{def}{=} [\Delta v^{i1}\ \Delta v^{i2} \cdots \Delta v^{in}]^T) \neq 0. \tag{3-3}$$

Directional derivatives $DF_{*_j}(v^i; S^{ik})$ of the j^{th} plant output F_{*_j} at a point v^i and in a direction $s^{ik} \stackrel{def}{=} v^i - v^{i-k}$, can be computed as:

$$DF_{*_j}(v^i; s^{ik}) = (s^{ik})^T \nabla F_{*_j}(v^i) \tag{3-4}$$

for each $k = 1, \ldots, n$, $j = 1, \ldots, m$. Therefore

 ISBN: 978-84-614-6167-7

$$S^i \nabla F_{*j}(v^i) = \begin{bmatrix} DF_{*j}(v^i; s^{i1}) \\ \vdots \\ DF_{*j}(v^i; s^{in}) \end{bmatrix}, \tag{3-5}$$

$j = 1, \ldots, m$. If the points v^{i-k} are close enough to v^i, then for every $j = 1, \ldots, m$,

$$\nabla F_{*j}(v^i) \cong (S^i)^{-1} \begin{bmatrix} F_{*j}(v^i) - F_{*j}(v^{i-1}) \\ \vdots \\ F_{*j}(v^i) - F_{*j}(v^{i-n}) \end{bmatrix}. \tag{3-6}$$

This formulation assures generation of consecutive set points v^i in such a way that the efficient estimation of the plant output derivatives using (3-6) can be applied.

3.3. Broydon's method

One way to avoid calculating derivatives is the so-called Broydon family of algorithms [5]. These kind of algorithms assure to give an estimation of the output derivatives with respect to the set-points when used within the ISOPE algorithm, following the updating scheme:

$$BR_k = BR_{k-1} + \frac{y_k - y_{k-1} - (BR_{k-1} * (u_k - u_{k-1}) * (u_k - u_{k-1})^T)}{(u_k - u_{k-1})^T * (u_k - u_{k-1})} \tag{3-7}$$

Where BR_k and BR_{k-1} are respectively the present and previous estimates of the output derivative matrix, y_k and y_{k-1} are the present and previous values of the measured outputs, while u_k and u_{k-1} are the present and previous values of the manipulated variables respectively. Equation (3-7) is called the Broydon update. In practice, The *BR* matrix is updated periodically using previous and present measurements of the output and manipulated variables and needs to be initialised at the start up.

3.4. Dynamic model identification method (DMI)

This method is based on the identification of a dynamic model that is used to approximate the real process locally at each working point for the purpose of estimating steady-state derivatives. It was first introduced into the field of optimization by [8], where it was shown to be an efficient method for identification, especially in the case of slow processes.

The key feature of the method is to approximate the real process by a dynamic model during the transient using real process information. In this case the waiting time for steady state to estimate the derivatives is avoided; these derivatives are calculated directly from the steady-state model derived from the identified dynamic model. In many cases a linear structure is assumed [9], [10], but this is

 ISBN: 978-84-614-6167-7

not always the case as general non-linear forms can also be used [8] and is described below.

It is assumed that the process is stable and can be approximated by a non-linear lumped model [8], [9]. The process model to be identified considers n different inputs $u^T=[u_1, u_2, ..., u_n]$ and m different outputs $y^T=[y_1, y_2, ..., y_m]$. A generalized second order Hammerstein model of the following form is used:

$$\begin{aligned} y_1(k) &= b_{00} + B_{l10}(q^{-1})u_1(q-d) + \ldots + B_{l11}(q^{-1})u_1^2(q-d) + \ldots + \\ & B_{lv\mu}(q^{-1})u_v(q-d)u_\mu(q-d) + \cdots + B_{lnn}(q^{-1})u_n^2(q-d) - A_{1l}^T(q^{-1})y_l(k) \end{aligned} \tag{3-8}$$

where:

$$\begin{aligned} A_{1l}(q^{-1}) &= 1 + a_{1l1}q^{-1} + \ \ldots + a_{1lm}q^{-m} \\ &= 1 + A_{1l}^T(q^{-1}) \end{aligned} \tag{3-9}$$

$$B_{lv\mu}(q^{-1}) = b_{lv\mu 1}q^{-1} + \ \ldots + b_{lv\mu m}q^{-m} \tag{3-10}$$

are polynomials of order m in the backward shift operator q^{-1}, and T denotes Transpose. More details on the method can be found in [2].

3.5 The Neural Network scheme

3.5.1 The technique

The technique is based on training a neural network to learn from the physical process itself. Once the training is finished, a steady-state neural network model, which imitates the static behaviour of the dynamical system, is obtained. This model is used, within the ISOPE algorithm, to find the system outputs to any given set-point even if these were not included in the training set. In this case, accurate system outputs are available to the ISOPE algorithm and prohibitive waiting times are avoided. These data points are used to calculate system output derivatives with respect to the set-points (Figure 1).

It has to be mentioned, that during training, switches k_2 and k_3 are closed, k_4 is open and k_1 is in position 1. This enables the algorithm to collect input/output data candidates required for the training in order to generate the identification neural network model. The states of these switches are reversed otherwise.

 ISBN: 978-84-614-6167-7

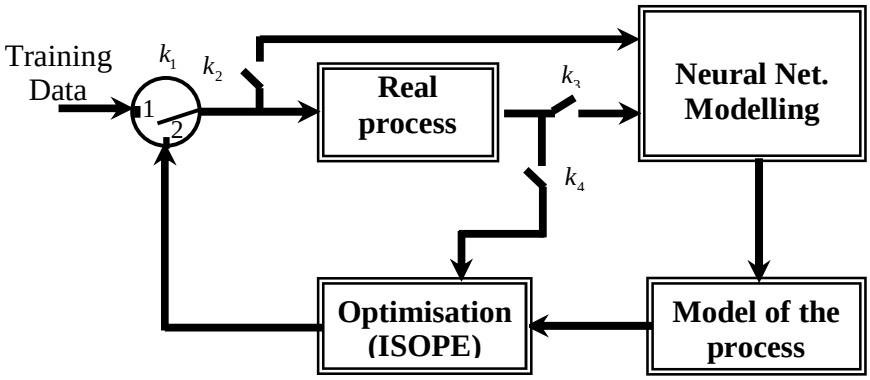


Figure 1: The neural network scheme

3.5.2 Identification

The identification problem consists of setting up a suitably parameterised model and adjusting its parameters to optimise a performance index based on the error between the plant and the identification model outputs. Every neural network model is composed of a series of weight vectors, which form, what are known as, weight matrices. These matrices are updated each time the network is trained for another input/ output data sample until no further improvement is obtained. Hence, the identification procedure consists in adjusting the parameters of the neural network in the model based on the error between the plant and the identification model outputs. Generally, the training of the network is performed once only. This takes place at the beginning of the optimisation procedure. Once a performance goal is reached, training stops, and the model and its parameters are saved to be used in the optimisation procedure. In case the system parameters change, the neural network model has to be retrained. In this case, a suitable time is given to the algorithm to perform identification and produce a new model. Once training is finished, the model parameters are updated, saved, and passed to the optimisation routine (Figure 1).

4. Simulations and Results:

In order to assess and compare the performance of all the techniques presented in this paper, a set of simulations were carried out on a two Continuous Stirred Tank Reactors (CSTR) connected in cascade ([9], [10], [11]). An exothermic autocatalytic reaction takes place in the reactors with interaction taking place in the units in both directions due to a recycle of 50% of the product stream into the first reactor (Figure 2).

The reaction is:

$$A + B \underset{k_-}{\overset{k_+}{\Leftrightarrow}} 2B \qquad (4\text{-}1)$$

 ISBN: 978-84-614-6167-7

The manipulated variables which are the set-points of the temperature controllers in both reactors are: $v=(T_1,T_2)^T$. The product concentrations associated with the second tank are outputs: $y=(Ca_2,Cb_2)^T$.

The objective function for all the simulations using this system was chosen to be linear of the measured variable C_{b2} and reflects the desire of maximising the amount of component *B* in tank 2. Thus the form of the objective function is as follow:

$$H(y,v)=-C_{b2} \tag{4-2}$$

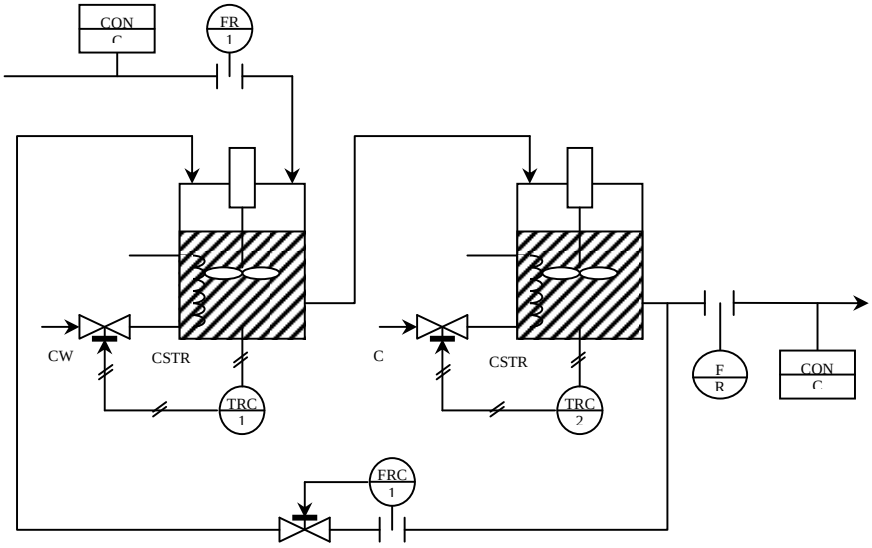


Figure 2 : The two CSTR system

The simulations were carried out using a MATLAB®/ Simulink platform, where a Simulink model of the process was created to enable periodical calls to the ISOPE algorithm (with the integrated identification technique) saved in an M-file. The starting point which is the initial steady-state condition was chosen to be: $T_1(0)=307\ K$ and $T_2(0)=302\ K$ which yields the following steady state outputs: $Ca_2=0.0141[kmol/m^3]$ and $Cb_2=0.0586[kmol/m^3]$.

In the simulations, the identification of the dynamic model was carried out during the transient, once found the updated model was used in the model-based optimization routine to produce the new process set-points.

The final converged results of the simulations for the various techniques are shown in Table 1 and figures 3 to 7. Table 2 gives the final objective function value and number of set-point changes taken to converge, obtained using all techniques mentioned above. While the figures show the trajectories taken by the outputs and manipulated variables for the FDAM, Broydon's, Dual Control method, DMI with a nonlinear model and the neural network scheme respectively.

 ISBN: 978-84-614-6167-7

We notice that all the methods converge to the correct process optimum point given by $T_1=312\ K$ and $T_2=310.2K$, with the optimum objective function value of -0.0725. This is to be expected, as all techniques satisfy the necessary system optimality conditions. From table 1 it is shown that the method using the neural network scheme converges faster than the other methods used in the simulations and gives a more accurate approximation of the derivatives. It is also seen from the same table, that the method using finite differences to estimate the derivatives takes much more time to converge (in terms of number of set-points changes), while it needs only a few iterations. This is in total agreement with what was stated in the introduction, as in the original method using finite differences needs (n+1) times the number of iterations for the estimation (n is the number of set-points); which could be prohibitive for large systems with a big number of inputs and outputs and also for slow processes.
From Figures 3 to 9, it is seen how the changes in the set-points affect the measured outputs and how they derive their values from the initial steady-state condition given by $Ca_2(0)$=0.041361 [$kmol/m^3$], $Cb_2(0)$=0.058638 [$kmol/m^3$] to the final desired solution (Ca_2=0.0275 [$kmol/m^3$], Cb_2=0.0725 [$kmol/m^3$]).

Table1: **ISOPE** algorithm with the different estimation techniques

	FDAM	**Broydon's method**	**Dual control method**	**Nonlinear dynamic model**	**Neural Network scheme**
Function value	-0.0725	-0.0725	-0.0725	-0.0725	-0.0725
Number of Set-point changes	22	12	14	10	07

5. Discussion and Conclusion

Techniques for estimating real process derivatives to be used within the ISOPE algorithm have been presented, and applied on a cascade process consisting of two Continuous Stirred Tank Reactors.
All methods, due to the satisfaction of optimality conditions, do achieve the real process optimum provided they can be implemented in a stable manner after a suitable choice of relaxation gains.
In the case of high order, slow and noisy processes, the FDAM, is not, as is well documented, a good choice. Each time a process derivative is requested, a set-point perturbation needs to be applied and a measurement time needs to be given to allow the process to settle before the derivatives are measured. Additional difficulties are observed when noise is present on the output measurement. This set-point perturbation, and the subsequent measurement time, is where the majority of time is spent in the algorithm so this is a major consideration in

 ISBN: 978-84-614-6167-7

assessing the algorithm. As can be seen from the simulation on the CSTR's system (Table 1), the FDAM, approaches twice the number of set-point changes of the various after methods and would seem not to be the perfect choice of algorithm.

The dual control method takes 14 set-point changes (Table 1) to achieve the optimum in the CSTR's simulation. This is still more than the after methods but the ability of the algorithm to estimate the derivatives without any excess in the set-point changes makes it a good choice. In this example, the most suitable method is the neural network scheme as only 07 set-point changes are needed in order to converge to the right optimum point. However, the huge amount of data needed for training the network is its major drawback.

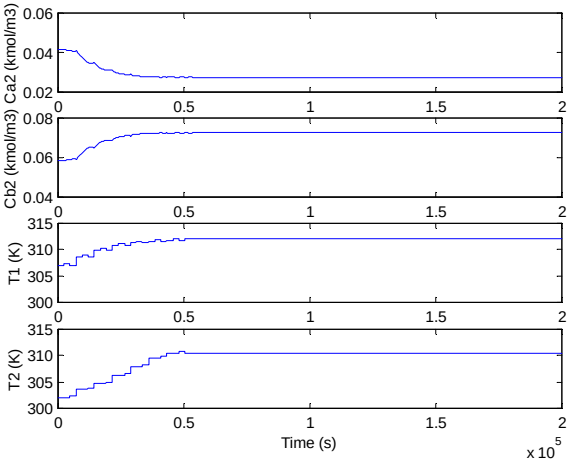


Figure (3): FDAM method.

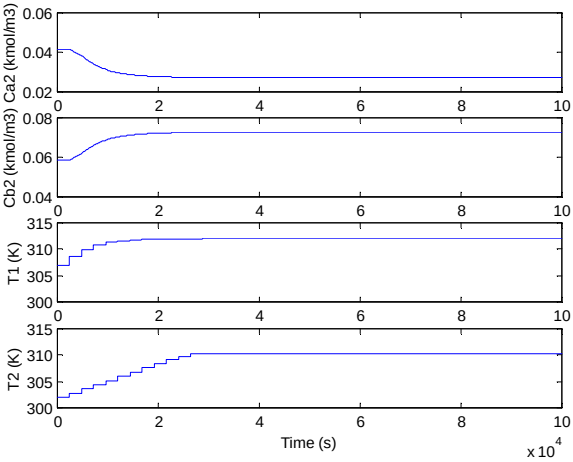


Figure (4): Broydon's method.

 ISBN: 978-84-614-6167-7

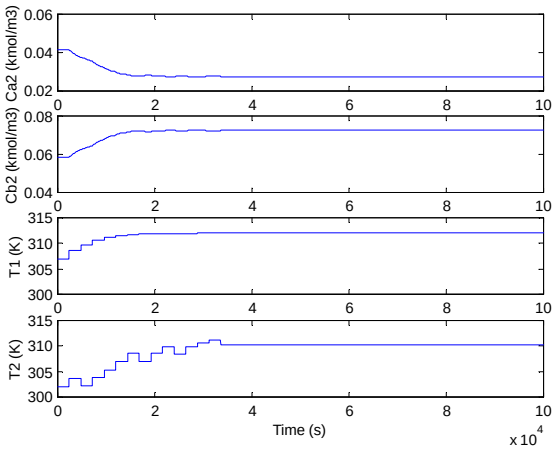


Figure (5): Dual control method.

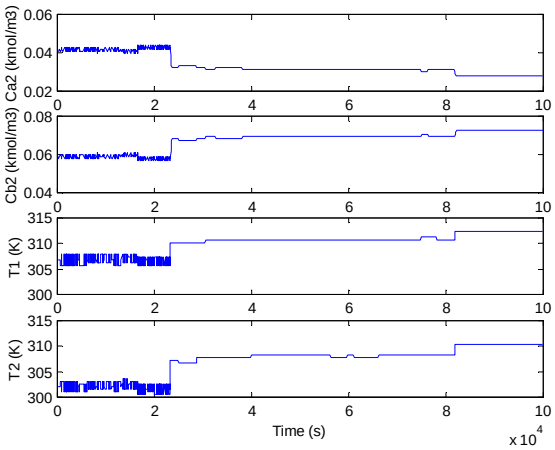


Figure (6): DMI with nonlinear model method.

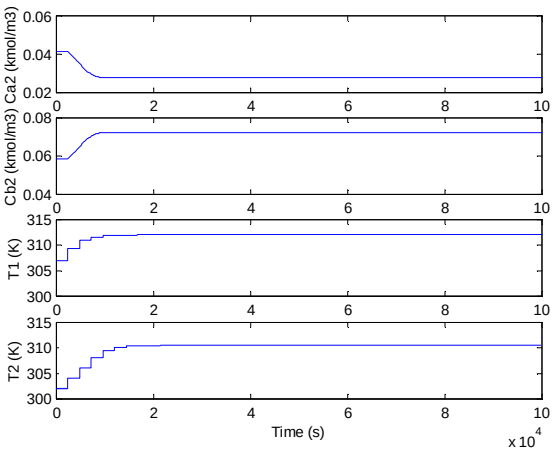


Figure (7): Neural network method.

 ISBN: 978-84-614-6167-7

6. References

[1] ROBERTS. P. D, An algorithm for steady-state system optimization and parameter estimation, *Int. J. of Systems Science*, 1979, **Vol. 10,** pp 719-734.

[2] MANSOUR, M. AND ELLIS, J. E., Comparison of methods for estimating real process derivatives in on-line optimisation, Applied Mathematical Modelling, **Vol. 27**, No , pp 275-291.

[3] ZHANG, H. AND ROBERTS, P.D., On-line steady-state optimisation of nonlinear constrained processes with slow dynamics, *Trans Ins MC*,1990, **Vol. 12,** No 5, pp 251-261.

[4] BRDYS, M. A., TATJEWSKI, P., An algorithm for steady-state optimizing dual control of uncertain plants, IFAC Workshop on new trends in design of control systems, 1992, pp 249-254.

[5] FLETCHER, R., Practical Methods of Optimization, Vol.1, A Wiley-Interscience Publication, 1980

[6] ELLIS, J. E., KAMBHAMPATI, C., SHENG, G. AND ROBERTS, P.D., Approaches to the Optimizing Control Problem, *Int. J. of Systems Science*, 1988, **Vol. 19,** No. 10, pp 1969-1985.

[7] ROBERTS. P. D, WILLIAMS. T. W. C, On an Algorithm for Combined System Optimisation and Parameter Estimation, *Automatica*, 1981, **Vol. 17,** pp 199-209.

[8] BAMBERGER, W., ISERMANN, R., Adaptive On-Line steady-State optimization of Slow dynamic Processes, *automatica*, 1978, **Vol. 14,** pp 223-230.

[9] GARCIA, C. E., MORARI, M., optimal Operation of Integrated processing Systems, *AICHE Journal*, 1981, **Vol. 27,** No. 6, pp 960-968.

[10] BECERRA, V. M., ROBERTS, P. D., GRIFFITHS, V. W., Novel developments in process optimisation predictive control, *J. Proc. Cont*, 1998, **Vol. 8,** No. 2, pp 117-138.

[11] JANG, S. S., JOSEPH, B. AND MUKAY, H., On-line optimisation of constrained multivariable chemical processes, AICHE Journal, 1987, **Vol. 33**, No1, pp 26-35.

 ISBN: 978-84-614-6167-7

Proceedings of the 11th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE2011
Benidorm, Spain, 26-30 June 2011

Application of radial basic function to predict amount of wood for production of paper pulp

Ana Martínez [1]*, Arcadio Sotto[2], and Angel Castellanos[1]

[1] *Department of Basic Sciences Applied To Forestry Engineering,*
Technical University of Madrid (UPM)
[2] *Department of Chemical and Environmental Technology, ESCET*
Universidad Rey Juan Carlos, Madrid

emails: ana.martinez@upm.es,arcadio.sotto@urjc.es,
angel.castellanos@upm.es

Abstract

In this interdisciplinary study, the proposed neural network method solves in an efficient way, how to build prediction models in engineering, in which the coefficients can explain the variable with more influence over the variable to forecast. Obtaining a good prediction and as simple as possible, i.e. with the least number of forecast variables. The systems has been applied to predict amount of wood for production of paper pulp.

Key words: Artificial neural networks, predictive model, linear regression model, wood volume

1. Introduction

The importance of paper and paper products in modern life is obvious; there is not a manufactured product, that is so important in any area of human activity. The paper provides the means to save, store and disseminate information, it is also the packaging material widely used. The applications and uses of paper and paper products are virtually unlimited and are continually developing new specialty products. Increasingly, they are adopting new technologies and methodologies, so that industry can remain competitive in the markets. The fibrous material from which to obtain the paper is called pulp. Making pasta is to transform the fibrous raw materials into fibers, so that after a series of operations can be transformed into paper. The fibrous material from which to obtain the paper is called pulp. Manufacture of pulp, aims to transform the fibrous materials, fibers so that after a

 ISBN: 978-84-614-6167-7

series of operations can be transformed into paper, cardboard, textiles, synthetic fibers, etc. The current paper pulp mainly used wood as raw material. In the manufacture of paper pulp, wood chips are boiled with appropriate chemicals in aqueous solution at elevated temperature and pressure. One of the two main methods is the Kraft process; this has come to occupy a dominant position due to its advantages in the chemical recovery and pulp strength. The Kraft process uses eucalyptus as raw material, using the boiling of wood chips in a solution (*NaOH*) and (*NaS*). Although until recently, the process is oriented towards conifers, however, there is now an increasingly marked tendency to use broad-leaved species, especially eucalyptus. Kraft pulps produce a one kind of paper very good and strong [2].

In this example uses a data file that is felling eucalyptus in Asturias, northern Spain region. The data that will work for each tree, are the diameter of "normal" in *cm* tree diameter at a height above ground of 1.30 meters (diameter as measured by the convenience), the total height of tree in *meters* (to the apex of the highest branch), age (in this case is data that is obtained because we know the age at which trees were planted, because eucalyptus is difficult, if not impossible, to count the rings growth) and finally the estimated volumes of tree bark and bark in dm^3 and the percentage of bark . In this case, as the target is for paper pulp, the difference is important because the tree bark (which is a significant proportion of the tree) is removed from the process. In pulp factories, the bark is used as fuel for energy production, a subject very fashionable at present, as source of renewable energy. Thus, in the manufacturing process, the goal is to obtain a pulp of given characteristics, a low energy consuming and getting the best performance. This study aims to predict both figures: the amount of wood that can be extracted for obtaining paper pulp, and the percentage of bark is collected and will be used as an energy source for the same process.

The example of application is considered in order to compare results using two forecasting methods: the general regression model and neural networks. In this part, the data set deals with the problem, to forecast the volume of wood. To build predictive models in engineering obtaining a good prediction with the least number of forecast variables; and in which the coefficients can explain the variable with more influence over the variable to forecast, but when among the input variables, exist some correlation (collinearity among forecast variables), in this kind of problems are really difficult to detect the importance of input variables from the coefficients of the general regression model, and neural networks have some advantages over regression model [1]. The method used is explained below and neural networks solve in an efficient way.

2. Proposed method using neural networks

This method should be applied when the importance of a variable changes from one to another to vary the range of the variable to be predicted. To search for a

 ISBN: 978-84-614-6167-7

predictive model, in these cases, implement a set of functions or hyper planes, that approximate the response variable by a function for each one of the subintervals, which has divided the range of the output variable. That is, it divides the range of the variable to predict contiguous and disjoint intervals and each interval obtained, it is implemented a network that more accurately predicted. The method is as follows:

- Normalization of the input and output variables into the interval [-1, 1] neural networks (NN) inputs and outputs (z) and $\hat{z}$ were normalized to lie in the range [-1,1] using the corresponding maximum and minimum values to preserve the interpretation of the weights and prevent numerical overflows.

- Neural network with *n* inputs and *one* output. The training algorithm considered is the backpropagation, and the activation function as sigmoid function. The neural network has been implemented with only one hidden layer, in this way; it is possible to study the weights in the neural network for the input layer [3,4].

- For this, first, the influence of independent variables X_1, X_2, ... X_n are studied over the range of the dependent variable *Y*, using the method called bisection method *(BM)* [1]. This method involves dividing the original pattern set into two subsets of patterns (values of the output variable *Y* above zero or positive (0,1] and values below zero or negative [-1,0)) and studying if the model (the weights) that the network implements, it is different or has changed for both subsets obtained. If the weights for the input variables have changed, for the two output intervals, then it will be necessary to define two functions (two neural networks one for each subset) for each one of the two output ranges obtained, for the variable *Y*, the process continues iteratively for each subset, each one divide again into two new subsets, until no changes occur in the weights of input variables over the output variable. For each one classes obtained, one neural network is trained and the value of the weights and error threshold are observed.

- Study of the highest absolute weights for the variables in each training neural network, and detect the most important variables [1].

- Verified by sensitivity analysis that the most important input variables for each subset obtained matches with those obtained in the previous step, that have been performed with neural networks and *BM*.

- Finally, for a new input pattern to know which is the network, of the first classification obtained by bisection method, that it must be use for the best

 ISBN: 978-84-614-6167-7

approximation. Now, it is necessary to use radial basis function (RBF) by cluster grouping, and to detect which of the networks obtained in the first part, it must be used to obtain the best prediction. The radial basis network is not able, in this case, give more accurate output, but it is possible to decide the cluster or class which it belongs. Radial basis neural networks detected by the algorithm *k-means* cluster potential, that exists within the data set, this allows us to detect, for a new input, which is the network to use. RBF neural networks provide a powerful alternative to multilayer perceptron to classify a pattern set. In this article, we propose to use radial basis neural networks in order to find the classification when a new input pattern appears.

So, it is possible to divide the problem in sub-problems, and to get a different function for different ranges of the variable to forecast. Each function is defined from the weights in each subset and subnet obtained, and which let us a lower error rate and also indicates the range of principal input variables; and it will be compared with the study of regression model to verify that the variables, which have been chosen to define predictive models, are right.

2.1. Predictive models

The multivariate regression model fits the data, taking into account which variables could participate in the model, and the suitable way to find the individual effects of forecasting variables over the variable to forecast, based on the extras squares addition principle [5,6]. The statistical provide estimates of the coefficients of regression model with its standard error of the estimate, a value of significance or better yet, a confidence interval, if the significance is small, the interval does not contain the value zero, it must be considered as an indication that this variable is interesting in the model. If it contains at zero (not significant) may be preferable remove the variable to simplify the model, but if the rest of the coefficients change to remove it, this is a confounding variable. Finding such variables is one of the objectives of the regression. The correlation matrix helps us to identify linear correlations between pairs of variables. Finding correlations between independent variables is a bad sign; the correlation between independent variables shows that one of the two variables should leave the model [7,8].

3. Experimental Set Up

Estimating the volume of wood for area forest, for production of paper pulp, is the example of application which is showed by comparisons the results between the linear regression model and the neural networks [9], in this example, it is estimated the volume of wood of a tree and volume of tree bark. The Data set file, are data from eucalyptus obtained from a region in the north of Spain.

 ISBN: 978-84-614-6167-7

The main aim is to detect relationships between all the variables that are in our study, and also this work seeks to estimate the wood volume for production of paper pulp, using a set of data that can be easily obtained such as: diameter, thickness bark, grow of diameter, height and age. Volume parameter is one of the most important parameters in forest research when dealing with some forest inventories. Usually, some trees are periodically cut in order to obtain such parameters using cubical proofs for each tree and for a given environment. This way, a repository is constructed to be able to compute the volume of wood for a given area or forests and for a given tree specie, in different natural environments. The method more usual to estimate volume of wood is the tree volume tables or tree volume equations. A very common equation used to estimating volume is:

$V = \eta \cdot d^b h^c$ Where *b* is next to *2* and *c* is close *1*. (1)

Where *V* denotes volume without bark, *h* denotes height, *d* denotes diameter, the factor "η" reflects more or less, as the tree is separated from the cylindrical shape.

Neural networks is a mapping $f(x_1, x_2, x_3, ...x_n) : \Re^n \to \Re$

Where $x_1 = diameter(cm)$, $x_2 = height(m)$, $x_3 = thickness\ bark(cm)$,

$x_4 = years$ $x_5 = percentage\ of\ bark$,to forecast $y = volume\ of\ wood(dm^3)$.

Firstly, the most important variables are extracted, and finally the solution is a predictive model. In a first step is necessary to normalize the pattern set in the range [-1, 1] and then apply the bisection method. Next you must carry out the study of the weights in each one of the subsets obtained. The weights show us how in this case, the data set can be divided in three different subsets by *BM*. Now, in each one of the sets of patterns obtained, the most important variable in each one of the subsets are showed by the weights using the algorithm for extraction [1] and the importance of the variables should be confirmed with the sensitivity analysis.

Table 1. The most important input variables: weights and sensitivity analysis over each output interval for volume of wood.

	Classes NN	*Weights hidden*	*Sensitivity Analysis*	*Active performance*
Diameter	1	1.05	47.36	*MSE = 0.060*
Height	1	1.17	52.63	*NMSE = 0.110*
Diameter	2	1.1	56.11	*MSE = 0.019*
Height	2	0.86	43.88	*NMSE= 0.062*
Diameter	3	0.70	41.16	*MSE = 0.027*
Height	3	0.83	58.83	*NMSE = 0.086*

 ISBN: 978-84-614-6167-7

In table 1, is showed the outputs of the three trained networks, for the three subsets obtained in the first phase *BM*, and in each class, the same results in the weights and sensitivity analysis, about the most important input variable, are reached.

Finally, a radial basis function neural network has been implemented; the method computes clusters for classification wood volume in a eucalyptus forest. It was performed an initial study using 150 patterns in training set and five input variables. All centers are stable in three points which show the three main clusters, and where the net has been possible to detect the three types of tree (see table 2). Several matrixes have been computed; where columns are input forecast variables and rows are centers (neurons). Main centers of RBF approximate real clusters in the three forest areas, following table 2 shows the real clustering. Radial basis function neural network has classified, for each one of the tree tested.

Table 2. Clusters obtained using RBF.

Variable	*Zone species*	X_1*: Diameter*	X_2*: Height*
Height	1	10.34	12.71
Diameter	2	16.42	20.01
Height	3	25.01	27.97

The hyperspace is divided into different regions or clusters starting from 16. Later, the number of clusters has been decreased till the minimum number of possible clusters is reached in order to solve the problem minimizing the mean squared error. The number of hidden neurons must be greater than the number of input variables to perform a correct learning in RBF [10,11].

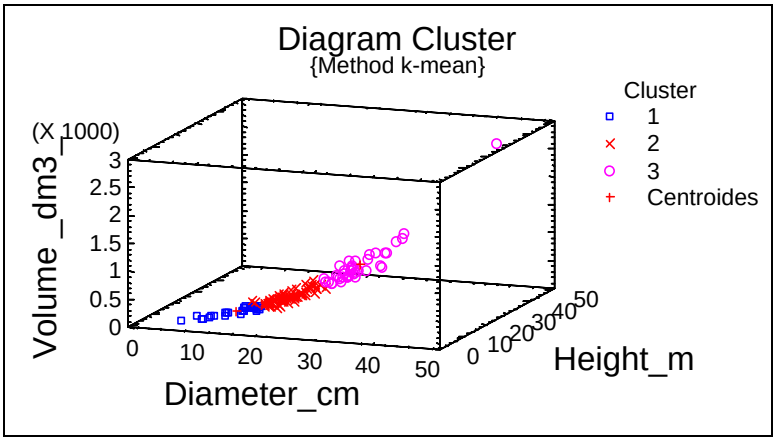


Figure 1. Clusters (three) obtained by RBF.

 ISBN: 978-84-614-6167-7

The Bisection method implemented, has divided the dataset in three subsets, indicating a change in the importance of input variable over the forecast variable. Now when the set is divided in three subsets and the weights are observed in each obtained subset or class, it is possible detect that the most important variable is changing in each class. The diameter appears as the most important variable in the second class; but in the first and third the most important variable has changed; now the principal variable is the height instead of the diameter. This information is detected in similar way by radial basis function which divided the information in three clusters. The set of eucalyptus trees have been classified into three classes, and for each one of the classes is possible to determine the input variable contributes most to the amount of paper pulp that can be obtained and predicted by a regression equation (the wood without crust can be obtained and the amount of bark tree).
The process carried out found three functions, which predict the volume of wood. After it should be check the results obtained by linear regression models, examining the likelihood ratios in the model, as well as, possible correlations between input variables. In this way, it is possible to compare the results obtained through the networks and regression model and the prediction error which is obtained with both models. The importance of input variables in the network is the same that in the regression model, as shown below.

Finally for this example of application, only two input variables are used to define the regression model, which confirms the information obtained from weights of the neural network, the sensitivity analysis and with the information provided by experts.

For each one of the three subsets obtained have been defined three regression models. They can predict more accurately, because different features have been detected on the input variable over the output variable in each subset obtained.

Formulating three regression models, one for each one sets of patterns, based upon 150 data points divided in three datasets by BM. For forecast the volume of the wood, in which height and diameter explained 65% and 23% respectively of the experimental results ($ADJR^2$ =88%) in the third class, in the same way in the second class and in the first class. The results and the prediction equations for each class are showed in table 4. The most important forecasting variable in the first class and in the third class to predict the variable *volume* is the variable *height* but in the middle class the most important input variable is the variable *diameter*. In this example it is possible observe that, the selection of the most important input variable changes, along the output range of the variable volume to predict. Three clusters have been obtained for the data set in the same way that the neural networks had obtained (see table 1 and 4)

 ISBN: 978-84-614-6167-7

Table 4. Regression estimated for each subset.

R^2	*Function*
Ht: 65% *Dn:* 23%	*V = -872.036 + 19.9774Ht + 36.3232Dn*
Ht: 19% *Dn*: 45%	*V = -169.558 + 6.08779Ht +13.5922Dn*
Ht: 60% *Dn*: 12%	*V = -97.1849 + 4.55486Ht +8.94644Dn*

Where Ht: is the height and Dn: Diameter of the tree and V: Volume of wood

The problem under study is prediction of volume of wood for production of paper pulp, and it is compared to other methods such as the formula (1) and the statistical regression analysis in order to estimate the amount of wood using typical tree variables: diameter and height. The neural network gives less estimated MSE than the standard formula (1) and regression analysis, see Table 5.

Table 5. Mean square error for three predictions.

	Error Equation (1)	Error NN	Error L. Reg.
MSE	0.05	0.003	0.01

4. Conclusions

To build predictive models in engineering, in which the coefficients can explain the variable with more influence over the variable to forecast, and to obtain a good prediction and as simple as possible, i.e. with the least number of forecast variables, be used the methods explained in this paper and where the proposed neural network method solves in an efficient way.

The implemented method is effective when the importance or characteristics of forecasting variables could change depending on the range of forecast variable, and therefore, it is necessary first apply the bisection method and divide the set in subsets, working with a neural network with only one hidden layer, and allowing the study of the weights in each subset obtained. As result of application of bisection process the proposed method showed a suitable learning rate. The importance of adding variables in a prediction model can be detected from the weights of training neural network, and also can explain which variable has the higher effect (importance) under the forecast variable. Once the first section, division, knowledge acquisition and variables identification, is finished, a control system is needed in order to choose the neural network to predict a new input

 ISBN: 978-84-614-6167-7

pattern. Therefore, it is implemented a radial basis neural network to decide the output interval and the network that should be activated.

The proposed method and the output of the net have a lower mean squared error than other prediction methods. Two problems are solved in the developed method: the suitable way to find the individual effects of forecasting variables over the variable to forecast, and the way to find a set of forecasting variables that should be included in a predictive model.

5. References

[1] A. MARTINEZ, A. CASTELLANOS, C. HERNANDEZ, F. MINGO, *Study of weight importance in NN,* Lect. notes Comput. Sci. Vol. 1611 (2004) 101-110.

[2] J. A. GARCÍA HORTAL, *Fibras Papeleras*, Edicions UPC, Barcelona, 2007.

[3] J. HEH, J.C. CHEN, *Designing a decompositional rule extraction algorithm for neural networks with bound decomposition tree,* Neural Comput. & Applic. 17(3) (2008) 297-309.

[4] J.A. MALONE, K.J. MCGARRY, C. BOWERMAN, *Rule extraction from Kohonen neural networks,* Neural Comput. & Applic. 15 (2006) 9–17.

[5] R. SETIONO, W. KHENG, J. ZURADA, *Extraccion of rules from neural networks for nonlinear regression,* IEEE Trans. Neural Network vol. 13(3) (2002) 564–577.

[6] G.C. CANAVOS, *Probabilidad y Estadística,* McGraw-Hill, Mexico, 1987.

[7] R.A. JOHNSON, D.W. WICHERN, *Applied Multivariate Statistical Analysis.* Pearson Education, New York, 2007.

[8] G.A.F. SEBER, *Linear Regression Analysis,* John Wiley, New York, 1987.

[9] M. YUAN, Y. LIN, *Model selection and estimation in regression with grouped variables,* Journal of the Royal Statistical Society: Series B 68 (2006) 49-67.

[10] A. CASTELLANOS, A. MARTINEZ, *Simultaneous control of chaotic systems using RBF networks,* I. J. Inf. Theory Appl. FOI ITHEA Vol. 2 (2008) 28-32.

[11] C. HARPHAM, W. DAWSON, *The effect of different basis functions on a radial basis function network for time series prediction: a comparative study*, Neurocomput. 69 (2006) 2161-2170.

 ISBN: 978-84-614-6167-7

Proceedings of the 11th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE2011
Benidorm, Spain, 26-30 June 2011

Videogrametry Geometry Model

Jose Martinez-Llario [1], Jose Herraez [1], Eloina Coll [1]
[1] *Instituto de Restauración del Patrimonio, Universitat Politècnica de València*
emails: jomarlla@cgf.upv.es, jherraez@cgf.upv.es, ecoll@cgf.upv.es

Abstract

Videogrammetry is a technology to measure the three-dimensional coordinates of points on an object surface. These coordinates are determined by measurements in two or more video images taken from different positions. In this work we are going to capture a video of an object with the following requirements: the video camera location is fixed and the object is turning around 360º very slowly (on a swivel base). Thousands of images are obtained and transferred to the computer. At this point we propose a model to involve the geometry of the swivel base on which the object is located and the position of the video camera. To check this model out, several points on the images are measured and the parameters which define the geometry are calculated using a least-square system. With the model proposed one can calculate in an easy way the 3D points on an object surface using a conventional video camera. We have developed a software package to test the proposed model.
Key words: Videogrammetry, Least-squares, Photogrammetry, Scanner

1. Introduction

Photogrammetry, on which is based Videogrammetry, is the science of obtaining reliable measurements from photographs [1]. The main purpose of the photogrametry is to obtain the 3D coordinates of the object which is photographed. For that, it is necessary to obtain two photographs of the object from two different positions. These two photographs are called 'photogrammetric pair' and the 3D object coordinates can be determined from them.

 ISBN: 978-84-614-6167-7

Videogrammetry is based on photogrammetry but it uses a video sequence instead of a pair of photographs. It would be very useful if we could take a video around the object, because we would have thousands of images which would cover the whole object.

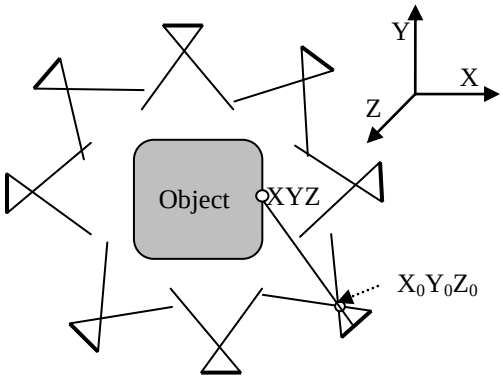


Fig. 1. Video around the object

As in photogrametry the relation between the object and each one of these pictures can be established using the collinearity condition.

The collinearity condition relates a point on the object surface (XYZ) with the image of this point in the photograph (xy) and the projection centres of the camera (X0Y0Z0). The nine terms m11...m33 correspond to the rotation matrix M(ω,φ,k) which relates the object reference system with the image (photograph) reference system. [2].

$$x = f \cdot \frac{m_{11}(X - X_0) + m_{12}(Y - Y_0) + m_{13}(Z - Z_0)}{m_{31}(X - X_0) + m_{32}(Y - Y_0) + m_{33}(Z - Z_0)}$$

$$y = f \cdot \frac{m_{21}(X - X_0) + m_{22}(Y - Y_0) + m_{23}(Z - Z_0)}{m_{31}(X - X_0) + m_{32}(Y - Y_0) + m_{33}(Z - Z_0)}$$

Equation 1. Collinearity condition

As the reader can suppose, it might be difficult to make a video around an object and to find out the rotation matrix and the projection centre for each photograph in an easy and fast way.

In this paper we want to present a model which takes into account all these variables around all the available photographs at the same time. This model must relate thousands of rotation matrices and projection centres in a single model to be solved using least-squares adjustment [3].

 ISBN: 978-84-614-6167-7

2. Instrumentation

2.1 Video camera

To take the video we have used a conventional digital video camera. The model DCR-TRV30E has been used to obtain all the video takings.

The video file is stored using a video format called mini DV [4]. This video camera stores the video in a magnetic tape using mini DV. Later, this video is transferred to the computer using an iDVconnection through a FireWire port.

2.2 Swivel base

The movement is relative; therefore it would be the same to rotate the object and to fix the video camera position instead of moving the video camera around the object. At this point, it is necessary to design swivel base on which the object is located. This base should rotate softly and without strong changes in speed. The video camera will be located in a fixed and stable tripod.

The designed swivel base is showed in fig 2. The base is made of two connected platforms using pillars. The swivel base rotates with a non constant speed (the engine speed depends on the variation of the frequency of the electricity). The speed is approximately 0.12 rad / seg., A complete turn takes around 53 seconds. Keeping in mind that the video system uses 25 frames per second, a whole turn will represent 1 325 photographs/images approximately. The platform speed has been determined in order to obtain a displacement between 0.5 and 1 pixel between consecutive photographs.

The objects used to get the 3D model are located on the bottom platform of this platform, then the video capture is carry out.

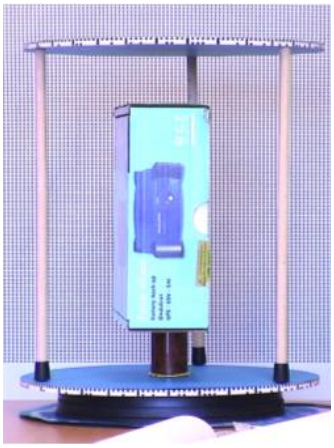


Fig. 2. The swivel base

 ISBN: 978-84-614-6167-7

3. Proposed model

Since the platform rotates through the Z axis and due to the apparent movement, the camera describes a circular trajectory. This trajectory has a constant height with the distance to the taking axis as the radio. Fig. 3 shows in a simplified way the trajectory of the camera according to the terrestrial coordinates system [5].

The coordinate system of the camera (in the terrestrial coordinate system) at a time of the video capture sequence is:

$$X = R \cdot \cos(w_0 + w \cdot t_i)$$
$$Y = R \cdot sen(w_0 + w \cdot t_i)$$
$$Z = Z_0$$

where:

Z0: Constant height of the projection centre.
w: Angular speed of platform rotation.
ti: Time i of the trajectory (t0: initial time).
w0 = w t0: Initial angle of the trajectory.

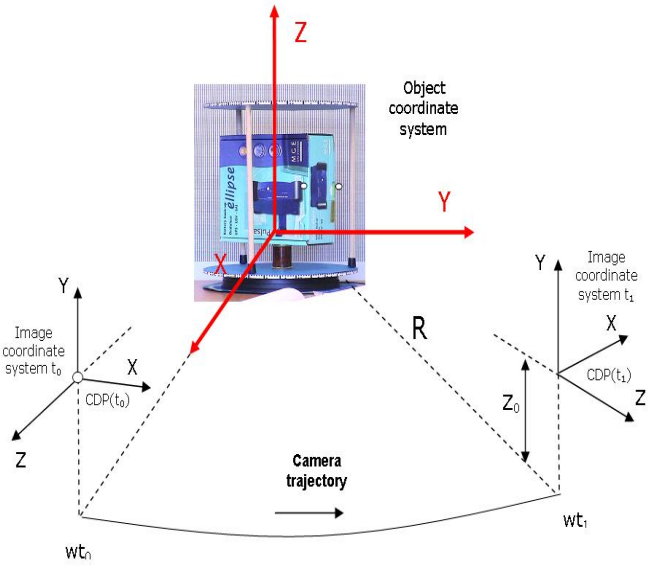


Fig. 3. Camera trajectory

4. Calculating the model

4.1 Control point coordinates

The first step to try to adjust the proposed model is to obtain the control point pixel coordinates. These control points are located on the swivel base where the

 ISBN: 978-84-614-6167-7

object to model is placed (fig. 4). To indentify accurately the control points we have designed a barcode.
The pixel coordinates are identified on some images (6 or 8 images which cover 360º). This operation is the only one the user has to carry out manually.

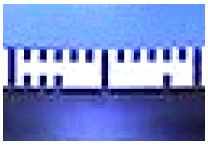


Fig. 4. Control points

4.2 Adjusting the model with the parametrized terrestrial coordinate system.

If we knew all the X0Y0Z0 coordinates of the projection centres and the control point coordinates in each frame of the video capture, then we would be able to apply the collinearity condition and to adjust it using a lest-squares adjustment [6].
The problem is obvious: we cannot measure each image coordinate of the control points and each projection centre because we are talking about more than 1 300 different spatial positions.
To simplify the problem we are going to model the platform shape, the camera trajectory, the engine speed, etc., that way we will be able to adjust the system using few variables.

The platform can be modelled with the following variables.

Variable	Description
Inferior platform (A)	
RA	Platform radio.
αA	Angle between two sequential support points.
δAx	X displacement of the platform centre.
δAy	Y displacement of the platform centre.
δAz	Z displacement of the platform centre.
μAx	Inclination of the platform in X direction.
μAy	Inclination of the platform in Y direction.
Superior platform (B)	
RB	Platform radio.
αB	Angle between two sequential support points.
δBx	X displacement of the platform centre.
δBy	Y displacement of the platform centre.
δBz	Z displacement of the platform centre.
μBx	Inclination of the platform in X direction.
μBy	Inclination of the platform in Y direction.
θB	Angle between the first control point in platform A and platform B.

Table 1. Platform model

 ISBN: 978-84-614-6167-7

To model the camera trajectory (X0Y0Z0) and being able to adjust our whole model we will use the follow variables:

Variable	Description
f	Focal distance of the camera.
d	Distance between the platform centre and the taking point.
nFrameXT	Image frame number that establishes the origin of the X axe.
AC0	Differential angle so that the image nFrameXT points out to the direction of the X axis.
W, θ, K	Angles of the rotation matrix on the X, Y, Z axes.
AK0	Initial rotation on the Z axis.
XC, YC, ZC	Terrestrial coordinates of the projection centres corresponding to each frame.
Φ	Angle on the plane XY starting from the image nFrameXT of the corresponding image.
Xi, Yi	Image coordinates of all the control points measured in the images.
nFrames	Number of total images (frames) of the video.

Table 2. Camera trajectory parameters

Next step is to put together the variables shown in table 1 and table 2 inside the collinearity condition and to apply a least-squares adjustment.

$$x = f \cdot \frac{m_{11}(X - X_0) + m_{12}(Y - Y_0) + m_{13}(Z - Z_0)}{m_{31}(X - X_0) + m_{32}(Y - Y_0) + m_{33}(Z - Z_0)}$$

$$y = f \cdot \frac{m_{21}(X - X_0) + m_{22}(Y - Y_0) + m_{23}(Z - Z_0)}{m_{31}(X - X_0) + m_{32}(Y - Y_0) + m_{33}(Z - Z_0)}$$

M = RT (w, φ, k)
f = focal distance
x, y = Pixel coordinates in the image.
X, Y, Z = Terrestrial coordinates in the object.
X0, Y0, Z0 = Projection terrestrial coordinates of the projection centre.

The values are compensated using approximate values. We obtain all the variables that define the projective conditions (collinearity), the geometry of the platform and the trajectory of the video camera.
As a result it is obtained in an accurate way the projection centres and the angles of inclination (photogrammetry external orientation) that define the trajectory of the camera.

 ISBN: 978-84-614-6167-7

5. Conclusions

The mathematical model showed so far allows us to define in an accurate way the geometry of a videogrammetry system made of a swivel base and a video camera.

The user does not have to either calculate each projection centre or control point coordinates. This model simplifies a complex videogrammetry system as if it were just a photogrammetry pair.

Once we know this whole model, the geometry object is obtained in two phases [7]. In a first step, the extraction of the object in each image of the video is obtained. The image borders are detected by using several techniques as image matching, space domain filters and the detection of image discontinuities.

In a simple way we can obtain, in an automatic way, the three-dimensional model of small objects, using a conventional device (video camera).

To test the proposed mathematical model, we have developed a software package [8]. With this software a user can obtain 3D models easily. These models are much more accurate that the ones obtained by traditional techniques and the hardware needed is much cheaper than using laser instrumentation.

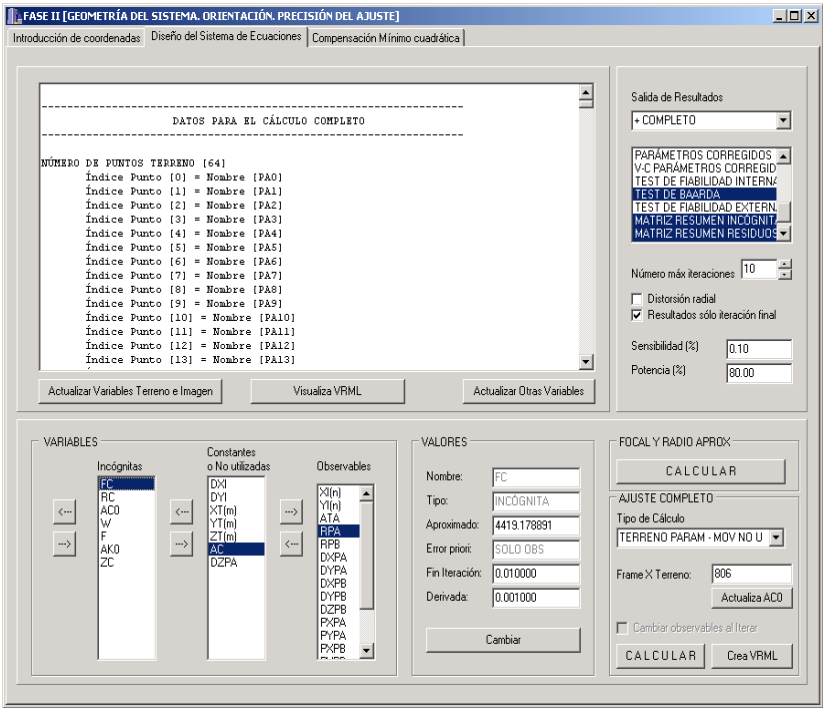


Fig. 5. Developed software for adjusting the model

 ISBN: 978-84-614-6167-7

6. References:

[1] KRAUS, K. (1997). *Photogrammetry - Volume I - fundamentals and standard processes*, Dümmler.

[2] SANJIB. K. *Fundamentals of Computational Photogrammetry*. Concept Publishing Company (2005).

[3] BASELGA, S., GARCÍA-ASENJO, L. *Global robust estimation and its application to GPS positioning*. Computers and mathematics with applications 56 (2008).

[4] HOLUB, P. *DV technology overview and video camera test [on line]*. CESNET technical report number 12/2001. http://www.cesnet.cz/doc/techzpravy/2001/12/

[5] LUHMAN, T., AT AL. *Close Range Photogrammetry: Principles, Techniques and Applications*. Wiley (2007).

[6] HERRAEZ, J., ET AL. *Epipolar Frames in a line for videogrammetry*. Advances in Signal Processing, Robotics and Communications. WSES Press. (2001) 60-63.

[7] MARTINEZ-LLARIO, J. ET AL. *Three-diemensional scanner software using a video camera*. Advances in Engineering Software. 37-7. (2006) 484-489.

[8] MARTINEZ-LLARIO, J. *Modelado fotogramétrico de objetos mediante secuencias de vídeo con imposición de condiciones de borde*. UPV. 84-688-2317-1. (2004).

 ISBN: 978-84-614-6167-7

Proceedings of the 11th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE2011
Benidorm, Spain, 26-30 June 2011

Comparing different solvers for the advection equation in the CHIMERE model.

Pedro Molina[1], Luis Gavete[1], Marta García Vivanco[2], Inmaculada Palomino[2], M. Lucía Gavete[3], Francisco Ureña[4], Juan José Benito[5]

[1] *Universidad Politécnica de Madrid, Spain*
[2] C.I.E.M.A.T., Madrid, Spain
[3] *Universidad Rey Juan Carlos, Madrid, Spain*
[4] *Universidad de Castilla-la Mancha, Ciudad Real, Spain*
[5] *U.N.E.D., Madrid, Spain*

emails: p.molina@upm.es, lu.gavete@upm.es, m.garcia@ciemat.es, inma.palomino@ciemat.es, lucia.gavete@urjc.es francisco.urena@uclm.es, jbenito@ind.uned.es

Abstract

In this paper we compare in the CHIMERE eulerian chemistry transport model four different finite volume algorithms for solving the advection equation using splitting technique. The numerical results are compared with a set of observation sites in the area of Spain and some conclusions are obtained.

Key words: advection equation, finite volume, conservative scheme, Chimère
MSC2000: AMS Codes (optional)

1. Introduction

Air pollution modeling is based on the assumption of no reciprocal effect of the chemical species on flow fields (wind velocity, turbulent diffusivity, temperature). After having pre-processed the flow fields by meteorological computations or parametrizations, the reaction-advection-diffusion PDE (Partial Differential Equation), that is, the mass continuity equation, is solved to estimate the concentrations of chemical species

 ISBN: 978-84-614-6167-7

$$\frac{\partial f}{\partial t}+\nabla\left(uf\right)=\nabla\left(k\nabla f\right)+P-L. \tag{1}$$

In this equation, characteristic of the Eulerian approach, f is a vector containing the concentrations of all model species for every grid box, u is the three dimensional wind vector, k the tensor of eddy diffusivity and P and L represent production and loss terms due to chemical reactions, emissions and deposition.

A class of conservative schemes for the advection equation has been so far proposed following the pioneering work of Godunov [1]. Rather than the piecewise constant interpolation in the original Godunov scheme, a linear interpolation function, MUSCL [2,3], and a parabolic polynomial, PPM [4,15], has been used. Other conservative schemes including a rational method has been used to solve the advection equation [5]

Our main goal of this research is to compare four different finite volume algorithms including a conservative rational method for the transport module included in the European scale Eulerian chemistry transport model CHIMERE. The results of the different methods are compared with a set of observation sites in the area of the Iberian Peninsula in Spain. Section 2 introduces the four advectives solvers that have been evaluated. In section 3 we introduce the European-scale chemistry-transport model (CHIMERE). The comparison of observed and modeled data is given in Section 4, and finally some conclusions are given in Section 5.

2. Modeling of linear advection

For simplicity of presentation we start with the scalar advection problem in one space dimension. We note f as the concentration of one typical atmospheric pollutant. The advection in one space dimension of this pollutant, during the interval $\left[0,T\right]$ is given by the following linear hyperbolic equation, called also transport equation, to which we add an initial concentration; the overall Cauchy problem is consequently the following one.

Given a field of initial concentration $f\left(x,0\right)=f_0$, a velocity wind u and a time $T>0$, we want to calculate $f(x,t)$ such that

$$\begin{cases}\dfrac{\partial f}{\partial x}+\dfrac{\partial\left(uf\right)}{\partial x}=0 \quad \forall\left(x,t\right)\in\mathbb{R}\times\left[0,T\right] \\ f\left(x,0\right)=f_0\left(x\right) \quad \forall\left(x,t\right)\in\mathbb{R}\end{cases} \tag{2}$$

 ISBN: 978-84-614-6167-7

In the following sections it will be showed the four different finite volume methods to solving it.

Let be $\left(\left[x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}\right]\right)_{i\in\mathbb{Z}}$ a partition of $[0,L]$ and $\left(\left[t^n, t^{n+1}\right]\right)n\in\mathbb{N}$ a regular partition of $[0,T]$. We define steps of time and space respectively written $\Delta t = t^{n+1} - t^n$ and $\Delta x_i = x_{i+\frac{1}{2}} - x_{i-\frac{1}{2}}$, the middle points $x_i = \frac{1}{2}\left(x_{i-\frac{1}{2}} + x_{i+\frac{1}{2}}\right)$ and the control volume $\Omega_i = \left[x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}\right]$.

By integrating (2) on $\Omega_i = \left[t^n, t^{n+1}\right]$, we obtain the integral form of the conservation law

$$\rho_i^{n+1} = \rho_i^n - \left(g^n_{i+\frac{1}{2}} - g^n_{i-\frac{1}{2}}\right) \Big/ \Delta x_i \tag{3}$$

where we define the exact flux by

$$g^n_{i+\frac{1}{2}} = \int_{t^n}^{t^{n+1}} (uf)\left(x_{i+\frac{1}{2}}, t\right) dt \tag{4}$$

and the average values of the exact solution, at the time t^n, on each cell by

$$\rho_i^n = \frac{\Delta t}{\Delta x} \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} F\left(x, t^n\right) dx \tag{5}$$

where F is the approximation of the function.

Four different approximations F_i are used in this paper corresponding to constant, linear, quadratic and rational approximations of F_i over each one of the central cells. By using time-splitting the method can be easily extended to solve advection equation in two and three dimensions. For example in two dimensions the time splitting is equivalent to do the transport of particles to the direction (*Ox*) and then according to the other direction (*Oy*).

2.1 The Upwind Method

The Godunov Method, or Upwind Method use a constant function a_i which is expressed in a generic mesh element with boundaries $x_{i-\frac{1}{2}}$, and $x_{i+\frac{1}{2}}$, and taking into account the velocity, u, the following constant approximation is used

 ISBN: 978-84-614-6167-7

$$\begin{array}{l} F_i(x) = a_i \quad \text{if } u > 0 \\ F_i(x) = a_{i-1} \quad \text{if } u < 0 \end{array} \quad \text{for } x \in \left[x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}} \right] \tag{6}$$

where a_i is the value of the function in the initial moment where this value is known

$$F_i\left(x_{i-1/2}\right) = f_{i-1/2}^n = a_i \tag{7}$$

2.2 VanLeer Method

The VanLeer method better known as MUSCL or Monotocnic Upstream-Centered Scheme for Conservation Laws, using a minmod technique, use a linear function that is expressed in a generic mesh element with boundaries $x_{i-1/2}$, and $x_{i+1/2}$, and considering the velocity $u < 0$ as

$$F_i(x) = a_i + b_i\left(x - x_{i-1/2}\right) \quad \text{for } x \in \left[x_{i-1/2}, x_{i+1/2}\right] \tag{8}$$

where a_i, and b_i are the coefficients of the interpolation function. a_i is calculated by the initial condition of each cell by

$$\begin{array}{l} F_i\left(x_{i-1/2}\right) = f_i^n \quad \text{if } u_{i+1/2}^n \geq 0 \\ a_i = f_i^n \\ F_i\left(x_{i-1/2}\right) = f_{i+1}^n \quad \text{if } u_{i+1/2}^n \leq 0 \\ a_i = f_{i+1}^n \end{array} \tag{9}$$

To calculate the slope term of the interpolation function, b_i, we need to define a slope-limiter to assume a good relation between the interpolation function with the original function as the behavior of its monotonicity. A good choice of slope is given by minmod [4].

2.3 Piecewise Parabolic Method (PPM)

The Piecewise Parabolic Method use a parabolic function, is expressed in a generic mesh element with boundaries $x_{i-1/2}$, and $x_{i+1/2}$, and considering the velocity $u < 0$ as

$$F_i(x) = a_i + b_i\bar{X}_i + c_i\bar{X}_i\left(1 - \bar{X}_i\right) \quad \text{for} \quad x \in \left[x_{i-1/2}, x_{i+1/2}\right] \tag{10}$$

where

$$\bar{X}_i = \frac{1}{\Delta x_i}\left(x - x_{i-1/2}\right)$$

 ISBN: 978-84-614-6167-7

a_i, b_i, and c_i are the coefficients of the interpolation function which are obtained by using the constraint conditions, where the constraint conditions are given in each cell by

$$\begin{aligned} F_i\left(x_{i-1/2}\right) &= f^n_{i-1/2} \\ F_i\left(x_{i+1/2}\right) &= f^n_{i+1/2} \\ \frac{1}{\Delta x_i}\int_{x_{i-1/2}}^{x_{i+1/2}} F_i(x)\,dx &= \rho_i^n \end{aligned} \tag{11}$$

where $\Delta x_i = x_{i+1/2} - x_{i-1/2}$. Then the coefficients a_i, b_i and c_i are given by

$$\begin{aligned} a_i &= F_i\left(x_{i-1/2}\right) \\ b_i &= F_i\left(x_{i+1/2}\right) - F_i\left(x_{i-1/2}\right) \\ c_i &= 6\left(f_i - \frac{1}{2}\left(F_i\left(x_{i-1/2}\right) + F_i\left(x_{i+1/2}\right)\right)\right) \end{aligned} \tag{12}$$

It is well-known that any high order interpolation tends to create spurious oscillations in numerical solutions. As a remedy for this, slope modifications were introduced in the PPM schemes. In this paper, we adopted the method of Colella and Woodward [4] for computing the interface values as,

$$f^n_{i+1/2} = \frac{1}{2}\left(\rho_i^n + \rho_{i+1}^n\right) - \frac{1}{6}\left(\overline{\delta} f_i^n - \overline{\delta} f_{i-1}^n\right) \tag{13}$$

with $\overline{\delta} f_i^n$ being the average slope in cell $\left[x_{i-1/2}, x_{i+1/2}\right]$ as follows

$$\overline{\delta} f_i = \begin{cases} 2\min\left(\delta f_i, \alpha_1\left(\rho_{i+1} - \rho_i\right), \alpha_2\left(\rho_i + \rho_{i-1}\right)\right), & \text{if } \left(\rho_{i+1} - \rho_i\right) > 0 \\ 2\max\left(\delta f_i, \alpha_1\left(\rho_{i+1} - \rho_i\right), \alpha_2\left(\rho_i + \rho_{i-1}\right)\right), & \text{if } \left(\rho_{i+1} - \rho_i\right) < 0 \end{cases} \tag{14}$$

where $\delta f_i = \left(\rho_{i+1} - \rho_{i-1}\right)/4$.

The positives α_1 and α_2 are parameters that control the average slope and affect the dispersion errors of the numerical solutions. For the piecewise parabolic method used in Chimère $\alpha_1 = \alpha_2 = 1$. To insurance good properties for the reconstruction we must to insure the parabolic monotonicity built on each cell. Then a second corrector algorithm is used to assume the monotonicity.

 ISBN: 978-84-614-6167-7

2.4 The Rational Method (RM)

Normally the interpolation function is based on polynomials. The main disadvantage of the polynomial interpolation is that can be unstable on the most common grid – equidistant grid. The rational interpolation consists of the representation of a given function as the quotient of two polynomials. The rational interpolation is an alternative for the polynomial interpolation. Its advantages are the high accuracy and absence of the problems which are typical for polynomial interpolation, such as the typical oscillations. However new difficulties can appear in the rational interpolation due to the existence of the poles.

The rational interpolation function F is expressed in each cell mesh element with boundaries $x_{i-\frac{1}{2}}$ and $x_{i+\frac{1}{2}}$ and considering the velocity $u<0$ as

$$F_i(x)=\frac{a_i+2b_i(x-x_{i-1/2})+\beta_i b_i(x-x_{i-1/2})^2}{\left(1+\beta_i(x-x_{i-1/2})^2\right)} \tag{15}$$

where a_i, b_i, and β_i are the coefficients of the rational interpolation function which are obtained by using the constraint conditions, where the constraint conditions are given in $\left[x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}\right]$ by knowing the values of the cell interfaces $f^n_{i-\frac{1}{2}}$, $f^n_{i+\frac{1}{2}}$ and the cell average value ρ_i^n for the time step n.

$$\begin{aligned} &F_i(x_{i-\frac{1}{2}})=f^n_{i-\frac{1}{2}} \\ &F_i(x_{i+\frac{1}{2}})=f^n_{i+\frac{1}{2}} \\ &\frac{1}{\Delta x_i}\int_{x_i-\frac{1}{2}}^{x_i+\frac{1}{2}} F_i(x)dx=\rho_i^n \quad \text{where} \quad \Delta x_i=x_{i+\frac{1}{2}}-x_{i-\frac{1}{2}} \end{aligned} \tag{16}$$

Then by solving the three equations (16), we obtain the coefficients of the rational interpolation function

$$\begin{aligned} &a_i=f^n_{i-\frac{1}{2}} \\ &b_i=\beta_i\rho_i^n+\frac{1}{\Delta x_i}(\rho_i^n-f^n_{i-1/2}) \\ &\beta_i=-\Delta x_i^{-1}\left[\frac{\left|f^n_{i-1/2}-\rho_i^n\right|}{\left|\rho_i^n-f^n_{i+1/2}\right|}-1\right] \end{aligned} \tag{17}$$

 ISBN: 978-84-614-6167-7

This last expression is corrected according with Xiao and Peng [6], to avoid division by zero as follows

$$\beta_i = -\Delta x_i^{-1}\left[\frac{\left|f^n_{i-1/2}-\rho^n_i\right|+10^{-20}}{\left|\rho^n_i-\rho^n_{i+1/2}\right|+10^{-20}}-1\right] \tag{18}$$

For computing the interface values we use (13) and (14) with $\alpha_1=\alpha_2=3$.

3. Model Description

CHIMERE is based on the mass continuity equation for the concentrations of chemical species in every box of a given grid:

$$\frac{\partial f}{\partial t}+\nabla(uf)=\nabla(k\nabla f)+P-L \tag{19}$$

In this equation, characteristic for the Eulerian approach, f is a vector containing the concentrations of all model species for every grid box, u is the three dimensional wind vector, k the tensor of eddy diffusivity and P and L represent production and loss terms due to chemical reactions, emissions and deposition.

In order to calculate the production and loss terms due to the chemical reactions it is necessary to solve a stiff system of ordinary differential equations, that is defined by the atmospheric chemical reactions. The numerical method for the temporal solution of the stiff system of partial differential equations is adapted from the second-order TWO-STEP algorithm originally proposed by [7] for gas phase chemistry only. It is based on the application of a Gauss-Seidel iteration scheme to the 2-step implicit backward differentiation (BDF2) formula:

$$f^{n+1}=\frac{4}{3}f^n-\frac{1}{3}f^{n-1}+\frac{2}{3}\Delta t\,R\left(f^{n+1}\right) \tag{20}$$

with f^n being the vector of chemical concentrations at time t^n, Δt the time step leading from time t^n to t^{n+1} and $R(f) = P_{(f)} - L_{(f)}$ the temporal evolution of the concentrations due to chemical production and emissions (P) and chemical loss and deposition (L). Note that L is a diagonal matrix here. After rearranging and introducing the production and loss terms this equation reads

$$f^{n+1}=\left(I+\frac{2}{3}\Delta t\,L\left(f^{n+1}\right)\right)^{-1}\left(\frac{4}{3}f^n-\frac{1}{3}f^{n-1}+\frac{2}{3}\Delta t\,P\left(f^{n+1}\right)\right) \tag{21}$$

The implicit nonlinear system obtained in this scheme can be solved pertinently with a Gauss-Seidel method [7].

In order to solve the partial differential equation (19) CHIMERE splits additively in subprocesses for which simpler PDEs exists. Then by using splitting we solve

 ISBN: 978-84-614-6167-7

the advection equation with the four different advective solvers. For modelling the diffusion CHIMERE takes into account only the vertical diffusion that is parameterized with an eddy diffusion approach. Horizontal diffusion is neglected as it is commonly done in mesoscale models. We can find a more complete description and evaluation of the Chimère model designed for seasonal simulations and real time forecasts without the use of super-computers in [8], where details about the implementation and evaluations of the modeling are given.

4. Numerical results

Simulations were carried out using the regional V2008 version of the CHIMERE model for 2008. This version calculates the concentration of 44 gaseous species and both inorganic and organic aerosols of primary and secondary origin, including primary particulate matter, mineral dust, sulfate, nitrate, ammonium, secondary organic species and water. The effect of the different numerical resolution scheme on model estimates was analyzed for a domain centred on the Iberian Peninsula in Spain (SP in Figure 1) [9]. A finer domain at a horizontal resolution of 0.1 degree covering the Iberian Peninsula was nested to a coarser European scale domain (EUR in Figure 1), ranging from 10.5W to 22.5E and from 35N to 57.5 N and a 0.2 degree horizontal resolution. A one-way nesting procedure was used: coarse-grid simulations forced the fine-grid ones at the boundaries without feedback.

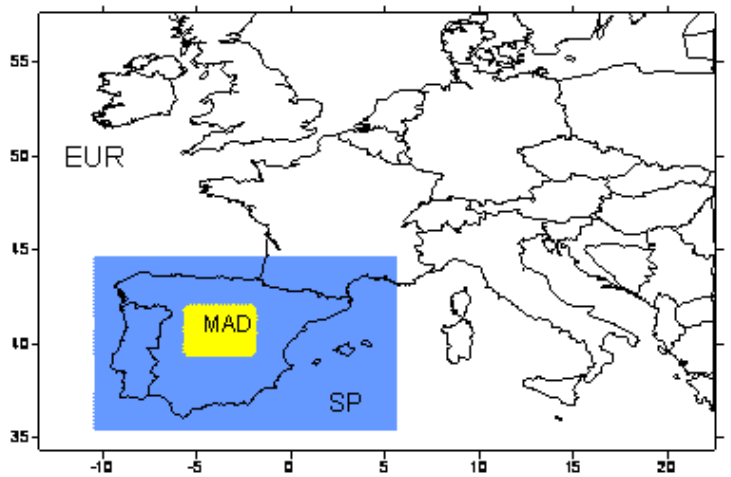


Fig.1. Location of the two domains simulated with the CHIMERE model

Boundary conditions for the coarsest domain were provided from monthly 2006 climatology from LMDz-INCA model [10] for gases concentrations and from monthly 2004 GOCART model [11] for particulate species.

Accurate emissions data in each time step are not available, then emissions for all the simulations were derived from the annual totals of the EMEP database for

 ISBN: 978-84-614-6167-7

2007 [12]. Original EMEP emissions were spatially disaggregated taking into account land use information (Global Land Cover Facility, GLCF, http://change.gsfc.nasa.gov/create.html) in order to get higher resolution emission data. For each SNAP activity sector, the total NMVOC emission was split into emissions of 227 real individual NMVOC, according to the AEAT speciation [13]. These species were then aggregated into the CHIMERE model ones.

The WRF model was used to obtain the meteorological input fields. The complete description of WRF model can be found at http://www.wrf-model.org/index.php. The simulations were carried out also for two domains, with respective horizontal resolutions of 19 Km and 10 Km. Both WRF simulations were forced by the National Centres for Environmental Prediction model (GFS) analyses.

The CHIMERE model was used with a time step of 3.3 minutes for the four methods.

The quality of model predictions obtained with the four algorithms for the transport module was analyzed by comparing model results to observations at the monitoring sites. Figure 2 shows the location of the NO_2, SO_2 and O_3, monitoring stations located inside the domain. Figures 3,4 and 5 present NO_2, O_3, SO_2 time series showing the concentration obtained with the four methods and the values registered at one of the monitoring sites (01016001 for O_3, 07010001 for NO_2, and 07032999 for SO_2 , Figure 2) for the period between June, 19th and June, 23th in 2008, as an example of the model performance.

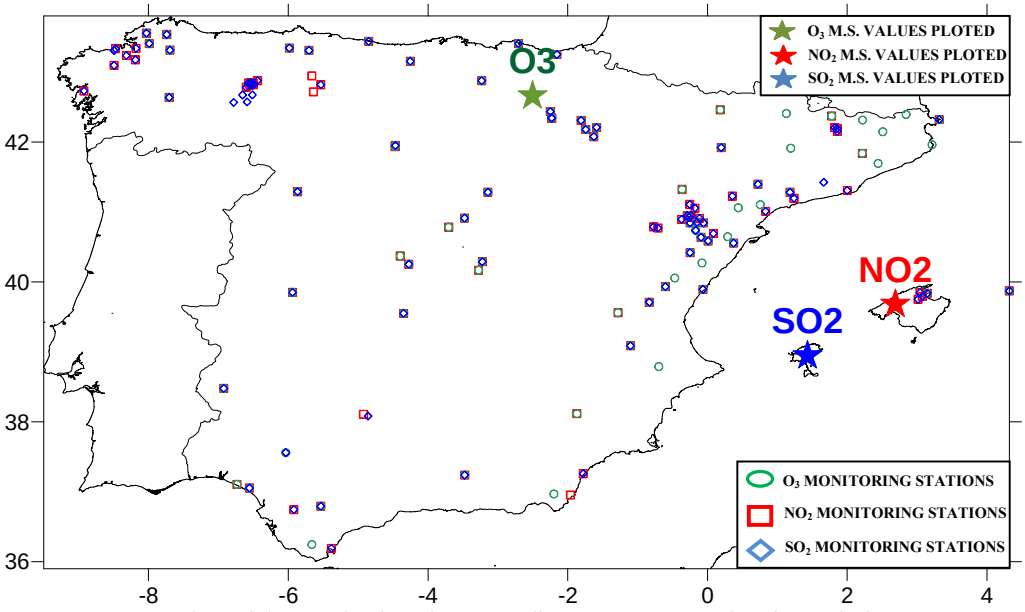


Fig. 2. Location of the monitoring sites recording NO_2, SO_2, and O_3 in Spain in 2008.

 ISBN: 978-84-614-6167-7

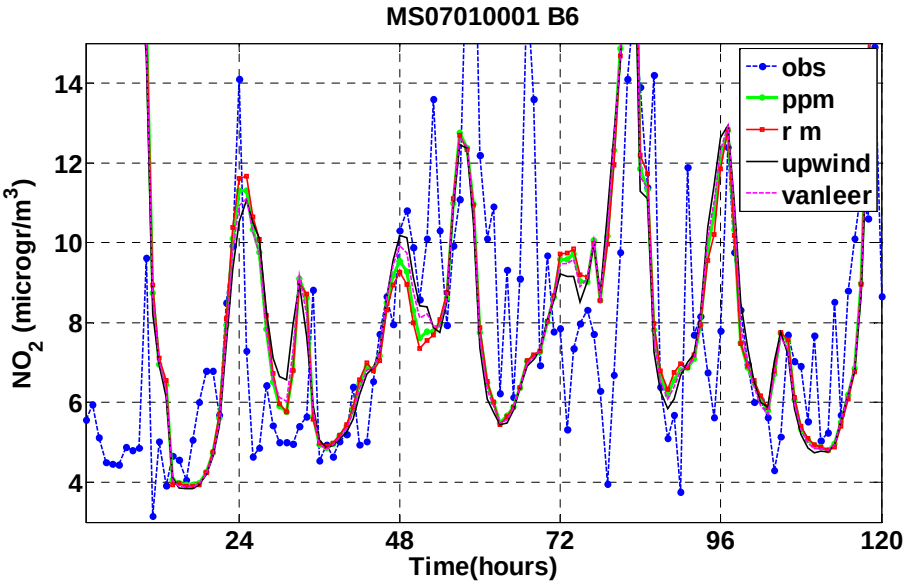


Fig. 3: Observed and simulated NO_2 concentration at 0701001 station.

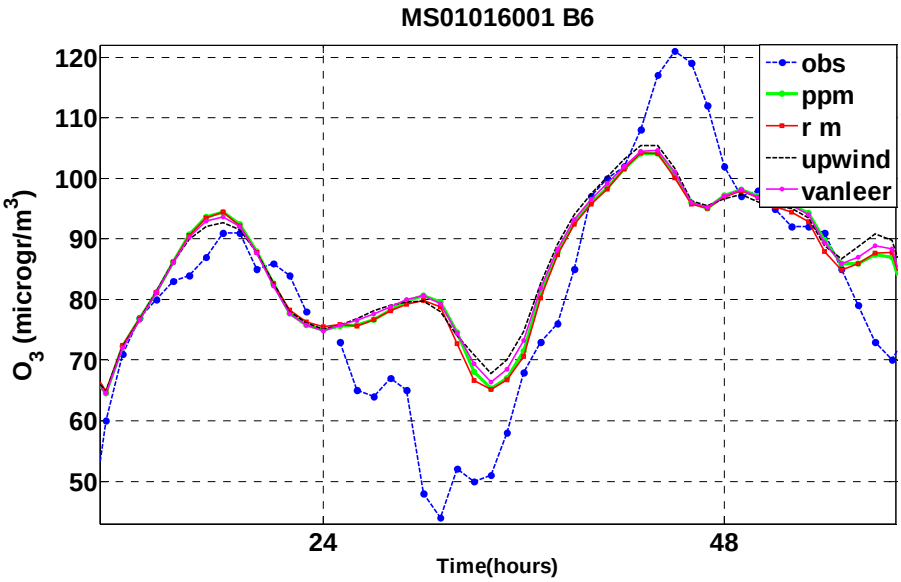


Fig. 4: Observed and simulated O_3 concentration at 01016001 station.

In order to evaluate the performance of the CHIMERE model estimates using the four different models some statistics were calculated. Table 1 presents the metrics used and their definition. Parameters such as mean bias (BMB), mean normalized bias (BMNB), mean normalized absolute error (EMNAE), root mean square error (ERMSE) and root mean normalized square error (ERMNSE) were estimated for NO_2, SO_2, and O_3.

 ISBN: 978-84-614-6167-7

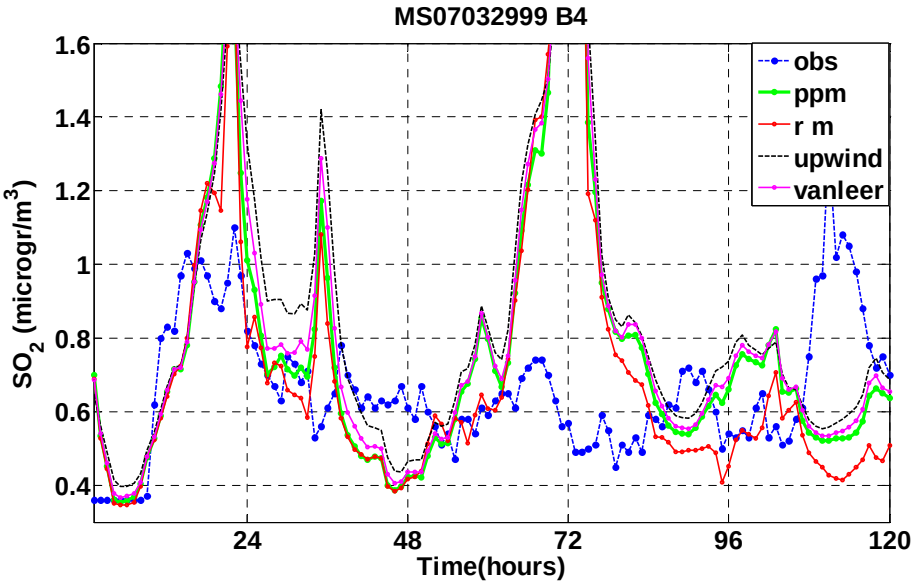


Fig. 5: Observed and simulated SO_2 concentration at 07032999 station.

Regarding ozone, only statistics for moderate-to-high ozone concentration cases (more important for human health protection) were considered by selecting predicted-observed value pairs when hourly observations were equal to or greater than the cutoff of 80 μgm^{-3}. For NO_2 and SO_2 a cutoff value of $10\mu gm^{-3}$ was used. It was taken into account 79, 103 and 76 air quality sites to estimate the statistics of O_3, NO_2, and SO_2 respectively.

Table 1. Definition of the metrics used in the evaluation of the CHIMERE model performance

Mean bias	$B_{MB} = \frac{1}{N}\sum(M_i - O_i) = \overline{M} - \overline{O}$
Mean normalized bias	$B_{MNB} = \frac{1}{N}\Sigma\left(\frac{M_i - O_i}{O_i}\right) = \left(\frac{1}{N}\Sigma\frac{M_i}{O_i} - 1\right)$
Mean normalized absolute error	$E_{MNAE} = \frac{1}{N}\Sigma\left(\frac{\lvert M_i - O_i\rvert}{O_i}\right)$
Root mean square error	$E_{RMSE} = \left[\frac{1}{N}\sum(M_i - O_i)^2\right]^{\frac{1}{2}}$
Root mean normalized square error	$E_{RMNSE} = \left[\frac{1}{N}\Sigma(\frac{M_i - O_i}{O_i})^2\right]^{\frac{1}{2}}$

N: pairs of modeled and observed concentrations *Mi* and *Oi*. The index *i* is over time series and over all the locations in the domain. * $\overline{M} = \frac{1}{N}\sum M_i \overline{O} = \frac{1}{N}\sum O_i$

 ISBN: 978-84-614-6167-7

In table 2 we show the mean observed and simulated concentration of all pollutants at 15 monitoring stations. Statistical results for all the pollutants are presented in Table 3. Mean normalized bias and mean normalized absolute error for ozone present values inside the range proposed by Tesche et al.[14] to decide on the suitability of a model.

Table 2. Mean Concentration for observed and simulated pollutant with the four different algorithms.

	NO_2					SO_2					O_3				
Station	Observed	Upwind	Vanleer	PPM	RM	Observed	Upwind	Vanleer	PPM	RM	Observed	Upwind	Vanleer	RM	PPM
7032999	2.976	6.939	6.949	6.912	6.84	101.38	116.4	116.4	116.4	116	0.7108	1.019	1.013	0.999	0.9673
12099001	6.5259	8.034	8.291	8.422	8.41	94.346	103.2	103.4	103.5	103.3	8.5632	0.733	0.729	0.737	0.7371
18189999	5.944	6.424	6.492	6.557	6.63	106.91	95.53	95.33	95.15	94.89	0.8286	0.382	0.38	0.38	0.3816
20016001	5.1868	7.215	7.025	6.887	7.02	67.498	93.95	93.42	93.2	92.51	4.3772	2.445	2.363	2.309	2.3845
28102001	5.4289	7.312	7.303	7.315	7.3	109.35	104.5	104.4	104.2	104	7.7727	0.628	0.629	0.634	0.6356
33036999	2.8702	4.957	5.05	5.073	4.97	65.891	87.95	87.11	86.58	86.46	1.1098	1.38	1.42	1.427	1.4156
11026001	3.8092	9.062	8.974	8.923	8.8	49.75	76.38	75.97	75.76	75.64	6.2433	1.658	1.744	1.794	1.7938
19061999	1.0195	4.882	4.853	4.837	4.75	95.187	94.83	94.97	94.97	94.82	0.7669	0.284	0.285	0.286	0.2846
33031029	14.112	11.42	11.35	11.3	11.5	31.98	72.23	71.34	70.77	70.13	5.5268	6.042	6.016	5.976	6.1256
27033001	2.6536	4.928	4.902	4.901	4.9	64.054	75.67	75.08	74.82	74.66	3.4286	1.726	1.77	1.819	1.8044
34023003	12.518	6.95	7.057	7.125	7.13	76.104	90.85	90.45	90.13	89.83	1.2976	0.336	0.358	0.373	0.375
25224999	1.5325	6.289	6.232	6.185	6.11	98.524	107.4	107.3	107.3	107.1	1.2632	1.042	1.027	1.017	1.0102
28016001	3.9411	6.652	6.694	6.7	6.64	110.83	102.8	102.9	102.7	102.5	6.0515	0.567	0.57	0.57	0.5643
12080007	5.3061	4.549	4.543	4.558	4.54	99.429	107.5	107.4	107.3	107	3.5015	1.016	1.039	1.047	1.0446
28102001	5.4289	7.312	7.303	7.315	7.3	109.35	104.5	104.4	104.2	104	7.7727	0.628	0.629	0.634	0.6356

Table 3. Statistics for all the pollutants evaluation

O_3	**BMB**	**BMNB**	**EMNAE**	**ERMSE**	**ERMNSE**
Upwind	2.7877	0.0410	12.80%	16.9465	16.67%
Van Leer	2.5866	0.0388	12.79%	17.0191	16.62%
PPM	2.4010	0.0369	12.80%	17.0699	16.65%
RM	2.2114	0.0351	12.78%	17.0651	16.62%
NO_2					
Upwind	-8.9937	-0.2790	59.07%	19.8850	72.60%
Van Leer	-8.8774	-0.2861	58.62%	19.7926	71.39%
PPM	-8.8055	-0.2741	59.48%	19.9736	73.63%
RM	-8.6074	-0.2623	60.29%	20.1179	75.63%
SO_2					
Upwind	-16.6580	-0.7207	76.37%	29.1502	8058.00%
Van Leer	-16.5552	-0.7047	76.25%	29.1315	80.54%
PPM	-16.4932	-0.7013	76.26%	29.1343	80.61%
RM	-16.3555	-0.6941	76.37%	29.1522	80.81%

 ISBN: 978-84-614-6167-7

5. Conclusions

In this paper we have compared the concentration of some pollutants predicted by the CHIMERE Eulerian chemistry transport model using four different methods to solve the advection equation.

The simulated concentrations for all the pollutants have been compared with a set of observation registered at some monitoring sites in Spain. There are some EPA guidelines to evaluate the accuracy of ozone model predictions (Tesche et al.[14]). The mean normalized absolute error, included in these guidelines, present, for ozone, values inside the ranges proposed inside the suggested EPA range (30-35%), (see Table 3), in order to consider an acceptable model performance. For the other pollutants, errors present higher values, as it commonly found when evaluating air quality model performance with EMEP database. The disagreement between model and observations for these pollutants is more related to accuracy of the input information, such as emissions, meteorology or land use data. For the time increment used the four advection solvers give similar results, using the 10 km resolution.

If the four advection solvers give similar results, it can be due to the very smooth functions used for emissions and also due to the very high horizontal resolution (0.1 degree). As conclusion it appears that the type of algorithm used for the advection problem is not so determinant, at least for this type of resolution. As it is important to decreases the execution time it is sufficient to use the upwind method which is the faster one.

Acknowledgements

This study was supported by Ministerio de Ciencia y Tecnología of Spain under the project CGL2008-1757/CLI.

6. References

[1] S.K. GODUNOV, *A difference scheme for numerical computation of discontinuous solutions of hydrodynamics equations*. Math. Sbornik 47, (1959) 271-306 (in Russian).

[2] B. VAN LEER, *Toward the ultimate conservative difference scheme. Part IV: A new approach to numerical convection*, J. Comput. Phys. 23 (1997) 276 -299.

[3] B. van Leer, *Toward the ultimate conservative difference scheme. Part V: A second order sequel to Godunov's method*, J. Comput. Phys. 23 (1997) 276 -299.

[4] P. COLLELA, P.R. WOODWARD, *The piecewise parabolic method (PPM) for gas-dynamical simulations*, J. Comput. Phys. 54 (1984) 174-201.

[5] L. GAVETE, M. GARCÍA VIVANCO, P. MOLINA, M. LUCÍA GAVETE, F. UREÑA, J.J. BENITO. *Implementation in Chimère of a conservative solver for the advection equation.*—cmmse10, Journal of Computational and Applied Mathematics (2011), doi:10.1016/j.cam2011.04.003.

[6] F. XIAO, X. PEN, *A convexity preserving scheme for conservative advection transport*, J. Comput. Phys. **198** (2004) 389-402.

[7] LE VEQUE, R. J., *Numerical Methods for Conservation Law*s. Birkuser Verlag (1990).

[8] J.G, VERWER, *A Gauss-Seidel iteration for stiff ODEs from chemical kinetics*, SIAM J. Scientific Computing. 15 (1994) 1243-1250

[9] M.G. VIVANCO, I. PALOMINO, R. VAUTARD, B. BESSAGNET, F. MARTÍN, L. MENUT, S. JIMÉNEZ. Multi-year assessment of photochemical air quality simulation over SPAIN, Environmental Modelling & Software, 24 (2009), 63-73.

[10] D. A. HAUGLUSTAINE, F. HOURDIN, L. JOURDAIN, M.A. FILIBERTI, S. WALTERS, J.F. LAMARQUE, , AND E. A. HOLLAND. Interactive chemistry in the Laboratoire de Meteorologie Dynamique general circulation model: Description and background tropospheric chemistry evaluation, J. Geophys. Res., 109, (2004), doi:10.1029/2003JD003957.

[11] M. CHIN, P. GINOUX, S. KINNE, B. N. HOLBEN, B. N. DUNCAN, R. V. MARTIN, J. A. LOGAN, A. HIGURASHI, AND T. NAKAJIMA. Tropospheric aerosol optical thickness fromt he GOCART model and comparisons with satellite and sunphotometer measurements, J. Atmos. Sci. 59, (2002) 461-483.

[12] V. VESTRENG, K. BREIVIK, M. ADAMS, A. WAGENER, J. GOODWIN, O. ROZOVSKKAYA, J. M. PACYNA. Inventory Review 2005, Emission Data reported to LRTAP Convention and NEC Directive, Initial review of HMs and POPs, Technical report MSC-W 1/2005, (2005) ISSN 0804-2446.

[13] N.R. PASSANT, Speciation of UK Emissions of Non-methane Volatile Organic Compounds. AEAT/ENV/R/0545, issue 1 (2000).

[14] TESCHE, T.W., GEORGOPOULOS, P., SEINFELD, J.H., CASS, G., LURMANN, F.W., ROTH, P.M., (1990). Improvement of Procedures for Evaluating Photochemical Models. Draft Final Report Prepared for California Air Resources Board. Radian, Sacramento, CA.

[15] MARTA GARCÍA VIVANCO, MAURICIO CORREA, OIER AZULA, INMACULADA PALOMINO, FERNANDO MARTÍN (2008) Influence of model resolution on ozone predictions over Madrid area (Spain). Computacional Science and Its Applications – ICSSA 2008" Lecture Notes in Computing Science LNCS 5072.". Vol. I, 165- 178. Springer ISBN-10: 3-540-69838-8

Proceedings of the 11th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE2011
Benidorm, Spain, 26-30 June 2011

A discussion on the numerical uniqueness of elastostatic problems formulated by Boussinesq potentials

Morales J.L.[1], Moreno J.A.[2] and Alhama F.[3]

[1] *Department of Structures and Construction*
[2] *Department of Mechanical Engineering*
[3]*Department of Applied Physics*
Technical University of Cartagena (UPCT)
Campus Muralla del Mar, ETSII, 30202, Cartagena, Spain

emails: joseluis.morales@upct, josea.moreno@upct.es,
paco.alhama@upct.es

Abstract

Numerical uniqueness conditions proposed by Tran-Cong for the solution of elastostatic problem formulated in terms of Boussinesq potential functions, as a particular case of Papkovich-Neuber representation, are discussed and alternative first class conditions which are more easily to implement are presented. Mathematical model is set up by two harmonic equations, uncoupled in the domain but strongly coupled at the boundary. Two applications to a hollow cylinder are studied, one related to a zero displacement field and one related to load conditions that give rise to a non-zero displacement field. Numerical solution is obtained by network method.

Key words: Boussinesq potentials, elastostatic, uniqueness

1. Background

In linear elasticity, the equilibrium equation in terms of displacements, in absence of body forces, is named Navier equation [1]:

$$\mu\nabla^2\boldsymbol{u} + (\lambda+\mu)\boldsymbol{\nabla}(\boldsymbol{\nabla}\cdot\boldsymbol{u}) = \mathbf{0} \tag{1}$$

where $\boldsymbol{u}$ is the displacements field, λ the Lamé's constant and μ the shear

 ISBN: 978-84-614-6167-7

modulus.
Equation (1) admits the general Papkovich-Neuber (PN hereinafter) displacement representations [2, 3] in the form

$$2\mu\boldsymbol{u} = \boldsymbol{\phi} - \nabla\left(\phi_0 + \frac{\boldsymbol{\phi}\cdot\boldsymbol{R}}{4(1-\nu)}\right), \quad \nabla^2\boldsymbol{\phi} = \boldsymbol{0} \quad , \quad \nabla^2\phi_0 = 0 \tag{2}$$

where $\boldsymbol{\phi}$ is a harmonic vector potential, ϕ_0 a harmonic scalar potential, $\boldsymbol{R}$ the position vector and ν the Poisson's ratio. Equation (2) is a general solution of (1) and has proved to be complete for the general case [4].

Under certain conditions, the number of potential functions of the general PN solution can be reduced without lack of completeness. Eubanks and Sternberg [5] studied the necessary conditions to delete either the scalar potential function or one of the rectangular components of the vector potential function. In addition, they considered the derived PN solution with the scalar and z-component of the vector potential for the axisymmetric case (Boussinesq potentials, $\phi_0, \boldsymbol{\phi} = \phi_z \boldsymbol{e_z}$), demonstrating that, when the meridional half-plane of the revolution solid is a simply connected domain, its completeness is guaranteed. Without a study of completeness, Boussinesq proposed his potentials 39 years before [6].

Stippes [7] was the first author that dealt with the subject of numerical uniqueness, always referring to the derived PN solution coming from delete the scalar potential. This researcher noted that Eubanks and Sternberg [5] did not make any reference to this subject, despite that in the demonstration of the completeness was implicit the uniqueness condition. In a footnote he wrote about the importance of the uniqueness conditions for the numerical solution, but with arguments only valid for the case of delete the scalar potential. Twenty five years later Tran-Cong [8] comes back to the question and, basing on the study of a null displacement field problem, he proved the necessity of additional conditions for derived PN solutions different to those studied by Stippes. In particular, he studied the case of delete the z-component of the vector function when the problem is z-convex and the Boussinesq solution. Tran-Cong [8] did not apply his conclusions to particular problems, as we do here.

In this work the Tran-Cong uniqueness conditions for the Boussinesq solution are discussed and a new alternative, more easily implemented in a numerical solution scheme, are proposed. Finally, two applications are numerically solved using the Boussinesq formulation where the potential function are assuming as primary unknowns: the first for a domain in which the displacement field is zero anywhere, and the second for a domain whose boundary supports an arbitrary load distribution that give rises to a non-zero displacement field. The numerical solution is obtained by EPSNET_10 [9], a specific software developed by the

 ISBN: 978-84-614-6167-7

network simulation research group of the UPCT [10] for elasticity applications. The network method [11] is a general proposed numerical tool whose efficiency, accurate and reliability has been demonstrated in many other fields of science and engineering. Network models are run in PSpice [12].

2. Governing equations and boundary conditions

For torsion-free axisymmetric problems in cylindrical coordinates, the displacement field for the Boussinesq solution reduce the equation (2) to

$$2\mu u_r = -\frac{\partial \phi_0}{\partial r} + -\frac{1}{4(1-\nu)} z \frac{\partial \phi_z}{\partial r} \tag{3a}$$

$$2\mu u_z = -\frac{\partial \phi_0}{\partial z} + \frac{3-4\nu}{4(1-\nu)} \phi_z - \frac{1}{4(1-\nu)} z \frac{\partial \phi_z}{\partial z} \tag{3b}$$

where the $u_\theta = \frac{\partial}{\partial \theta}$=0, and u_r= $u_r(r,z)$ and u_z= $u_z(r,z)$. The corresponding governing equations are

$$\frac{\partial^2 \phi_0}{\partial \mathrm{r}^2} + \frac{1}{\mathrm{r}} \frac{\partial \phi_0}{\partial \mathrm{r}} + \frac{\partial^2 \phi_0}{\partial \mathrm{z}^2} = 0 \tag{4b}$$

$$\frac{\partial^2 \phi_z}{\partial \mathrm{r}^2} + \frac{1}{\mathrm{r}} \frac{\partial \phi_z}{\partial \mathrm{r}} + \frac{\partial^2 \phi_z}{\partial \mathrm{z}^2} = 0 \tag{4c}$$

For the general case of mixed boundary conditions, these are can be written as

$$u_i = u_i^b \quad on \quad S_u \tag{5a}$$

$$\sigma_{ij} n_j = t_i^b \quad on \quad S_t \tag{5b}$$

S_u denotes the boundary surface points where the displacement, u_i^b, is prescribed, while S_t refers to points where the traction values, t_i^b, are imposed. σ_{ij} is the stress field and n_j the outer normal vector on S_t. Note that $S = S_t + S_u$ represents the complete boundary surface, Figure 1.

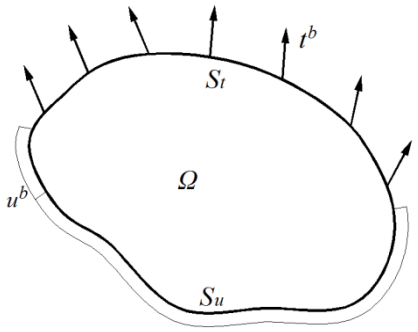


Figure 1 Scheme of the boundary conditions: tractions and displacements

 ISBN: 978-84-614-6167-7

The expressions that relate stress and potential functions, required for implementing the boundary condition (5b), are obtained from the strain-displacement equations and Hook's Law for the axisymmetric case [1]. Using equation (4) to simplify, these expressions are:

$$\sigma_{rr} = -\frac{\partial^2\phi_0}{\partial r^2} - z\frac{\partial^2\phi_z}{\partial r^2}\frac{1}{4(1-\nu)} + \frac{\partial\phi_z}{\partial z}\frac{\nu}{2(1-\nu)} \tag{6a}$$

$$\sigma_{zz} = -\frac{\partial^2\phi_0}{\partial z^2} - z\frac{\partial^2\phi_z}{\partial z^2}\frac{1}{4(1-\nu)} + \frac{\partial\phi_z}{\partial z}\frac{1}{2} \tag{6b}$$

$$\sigma_{rz} = -\frac{\partial^2\phi_0}{\partial r\partial z} - z\frac{\partial^2\phi_z}{\partial r\partial z}\frac{1}{4(1-\nu)} + \frac{\partial\phi_z}{\partial \mathrm{r}}\frac{1-2\nu}{4(1-\nu)} \tag{6c}$$

These equations can also be used as post-processing equations to derive the quantities σ_r, σ_z and σ_{rz} once the potential fields are solved. In short, the mathematical model for the Boussinesq solution is formed by equations (4) and (5) plus the expresions of displacements and stress in terms of the potential functions, equations (3) on S_u, and (6) on S_t, respectively.

According to Tran Cong [8], the Boussinesq solution (the numerical solution of the unknown potentials) is unique only with some additional conditions. These are discussed in the next section.

3. Uniqueness of the Boussinesq solution

For the following calculus, it is convenient to express the Boussinesq solution as

$$\boldsymbol{u} = \alpha\Psi\boldsymbol{e_z} - \boldsymbol{\nabla}(\varphi + z\Psi) \tag{7}$$

where $\alpha \equiv 4(1-\nu)$. The new harmonic potential, $\nabla^2\Psi = 0$ and $\nabla^2\varphi = 0$, are related with those of equation (2) by means of the expressions

$$\phi_x = \phi_y = 0,\ \phi_z = 2\mu\alpha\Psi,\ \phi_0 = 2\mu\varphi.$$

According to Tran-Cong [8], for a zero displacement field ($\boldsymbol{u} = \boldsymbol{0}$), Boussinesq solution satisfies the equation

$$\boldsymbol{0} = \alpha\Psi\boldsymbol{e_z} - \boldsymbol{\nabla}(\varphi + z\Psi) \tag{8}$$

Applying the rotational to this equation and taking into account the properties of this operator, it results $\partial\Psi/\partial\theta = \partial\Psi/\partial r = 0$, so that $\Psi = f(z)$. Also, applying the divergence operator and using the above result and the expressions $\nabla^2\Psi = 0$ and $\nabla^2\varphi = 0$, it is deduced that $\partial\Psi/\partial z = 0$. In short, in order to the Boussinesq solution represents a zero displacement field, the non-zero component of the vector potential must be constant,

$$\Psi = \mathrm{k}_1 \tag{9}$$

This result substituted in equation (8) yields

 ISBN: 978-84-614-6167-7

$$\mathbf{0} = -\frac{\partial\varphi}{\partial r}\boldsymbol{e_r} - \frac{1}{r}\frac{\partial\varphi}{\partial\theta}\boldsymbol{e_\theta} + \left[\mathrm{k}_1(\alpha-1) - \frac{\partial\varphi}{\partial z}\right]\boldsymbol{e_z} \tag{10}$$

From the two first addends of equation (10) it is obtained $\partial\varphi/\partial r = \partial\varphi/\partial\theta = 0$; this implies $\varphi = g(z)$. Substituting this result in the first addend, the condition referred to the scalar potential φ that makes the Boussinesq solution to represent a zero displacement field is derived: φ is a plane parallel to the radial axis,

$$\varphi = \mathrm{k}_1(\alpha-1)z + \mathrm{k}_2 \tag{11}$$

The arbitrary constants k_1 y k_2, define univocally the required potentials for a zero displacement field and, by extention, for any other displacement field represented by the Boussinesq solution. Figure 2 represents the equations (9) and (11); two planes defined by two constants, k1 and k2, since the slope of the φ plane is determined by the expresion $\mathrm{k}_1(\alpha-1) = \tan\beta$.

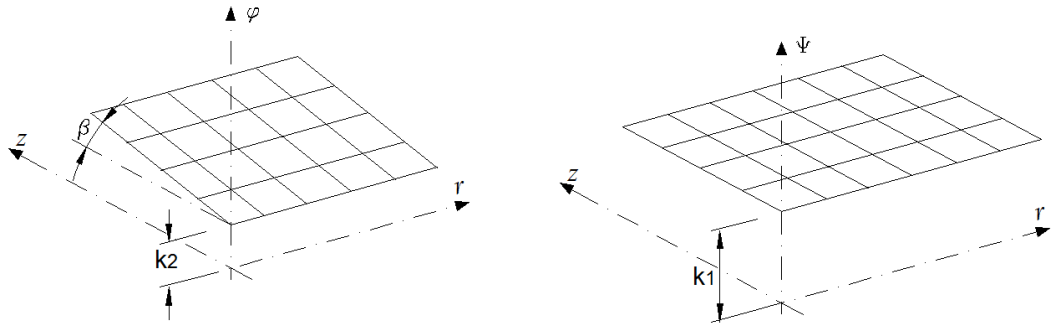


Figure 2 The Boussinesq solution for a null displacement field

Uniqueness conditions proposed by Tran-Cong [8] are:

$$\left.\begin{aligned}\varphi(\boldsymbol{a}) &= \mathrm{c}_1\\ \left.\frac{\partial\varphi}{\partial z}\right|_{\boldsymbol{a}} &= \mathrm{c}_2\end{aligned}\right\} \tag{12}$$

with $\boldsymbol{a}$ an arbitrary point of the domain, and c_1 and c_2 two constants, also arbitrarily chosen. Equation (12) is a way of fixing the constants k1 and k2, and makes unique the solutions (9) and (11): $\mathrm{k}_1 = \mathrm{c}_2/(\alpha-1)$ and $\mathrm{k}_2 = \mathrm{c}_1 - \mathrm{c}_2 a_z$.

A more simplified alternative to the above uniqueness condition (12) (in the form of Dirichlet condition) is proposed in this work:

$$\left.\begin{aligned}\varphi(\boldsymbol{a}) &= \mathrm{c}_1^*\\ \Psi(\boldsymbol{b}) &= \mathrm{c}_2^*\end{aligned}\right\} \tag{13}$$

In these equations $\boldsymbol{a}$ and $\boldsymbol{b}$ are two arbitrary points of the domain, while c_1^* and c_2^* are two arbitrarily constants. In this case, the constants k1 and k2 also define univocally the solutions (9) and (11): $\mathrm{k}_1 = \mathrm{c}_2^*$, $\mathrm{k}_2 = \mathrm{c}_1^* - \mathrm{c}_2^*(\alpha-1)a_z$.

 ISBN: 978-84-614-6167-7

About the notation of the potential functions, it is evident that the conditions (12) and (13) can be applied to the Boussinesq solution defined by both equations (3) and (7).

4. Applications

Application 1. Straight pipe with zero displacement field

The grid, 20x20, refers to the 2-D cross section of the pipe, Figure 3 (left). Rigid body movement is restricted by the vertical displacement upper boundary condition. Values of the geometry parameters are R1 = 50 mm, R2 =150 mm, H = 100 mm, while elastic constants are E (Young modulus) = 210 GPa and ν (Poisson ratio) = 0.3. Additional constants are $\phi_0 = 10$ in the cell (1,1), red circle in the Figure 3 (right), and $\phi_z = 20$ in the cell (20,1), red diamond. Figures 4 and 5 show the displacement field and potential solutions respectively

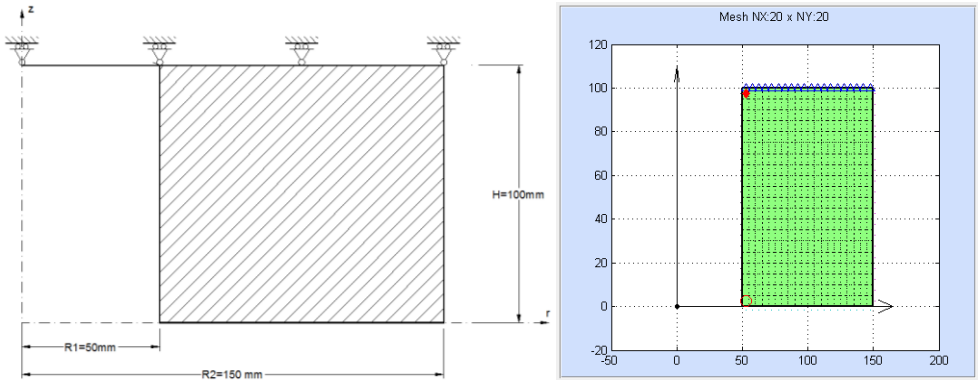


Figure 3 Physical model (left). Simple boundary conditions for a null displacement in a hollow cylinder; additional uniqueness conditions are shown in red color (right)

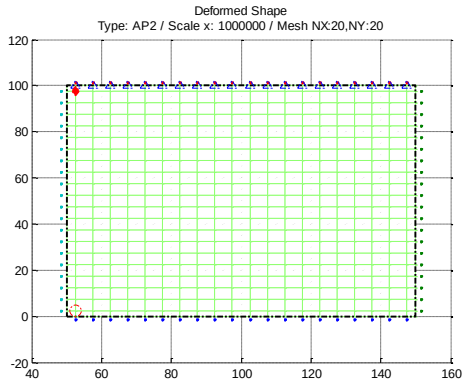


Figure 4 Solution of the displacement field by ESPNET_10

 ISBN: 978-84-614-6167-7

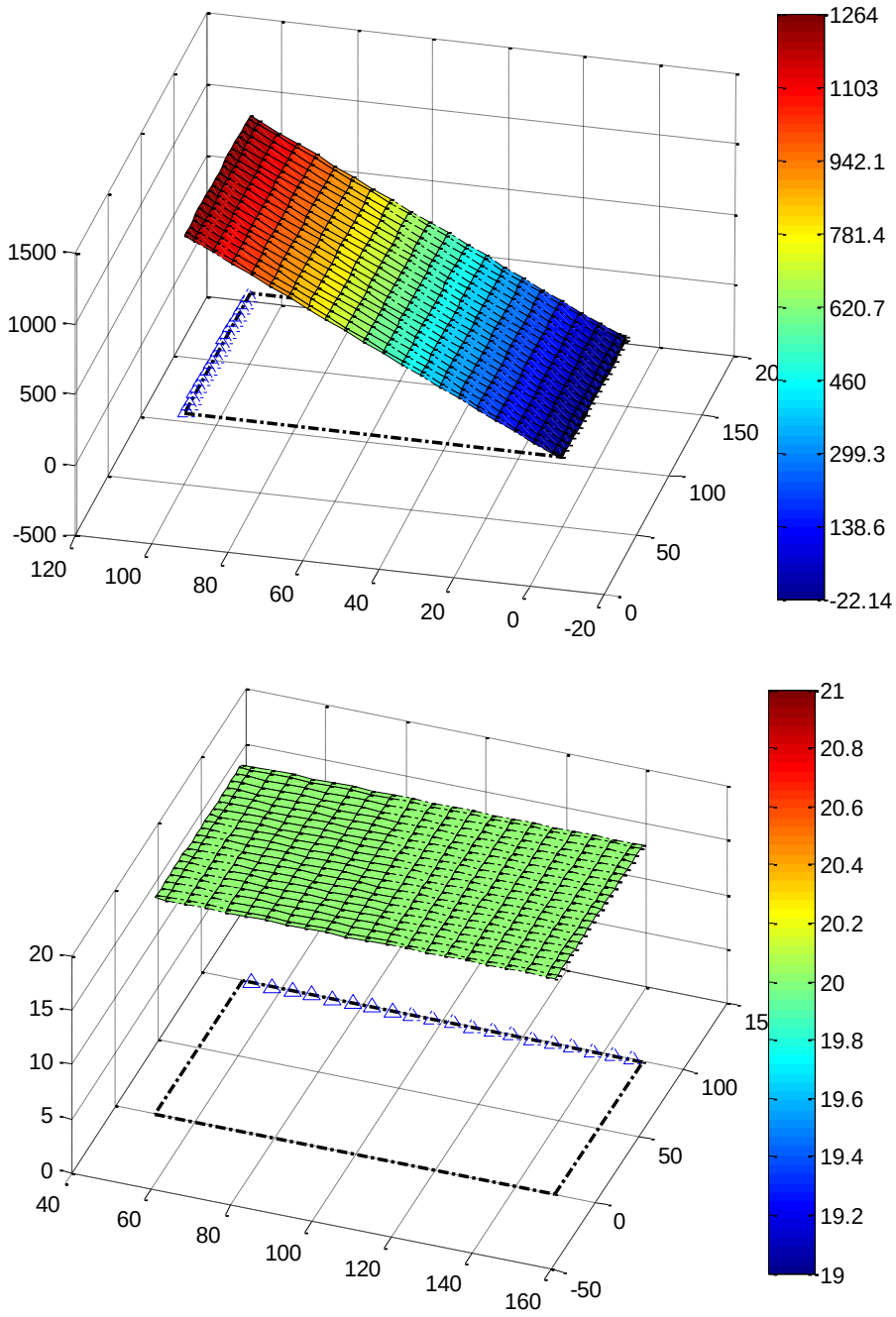


Figure 5 Potential solutions: Scalar potencial (up),vector potential (down)

Application 2. Straight pipe under arbitrary loads

The new boundary conditions referred to load and displacement are shown in Figure 5 (left). Geometric and elastic parameters of the material, grid and

 ISBN: 978-84-614-6167-7

additional conditions are de same of Application 1. Figure 6 shows the displacement field of the pipe cross section while Figure 7 depicts the potential solutions whose forms depends on the values of additional constants. These harmonic functions are smooth surfaces which curved slightly at its boundaries.

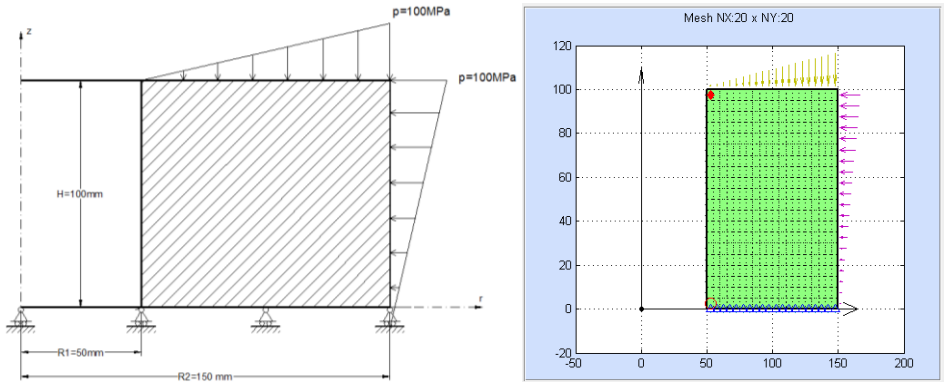


Figure 5 Physical model (left). General boundary conditions in a hollow cylinder; additional uniqueness conditions are shown in red color (right)

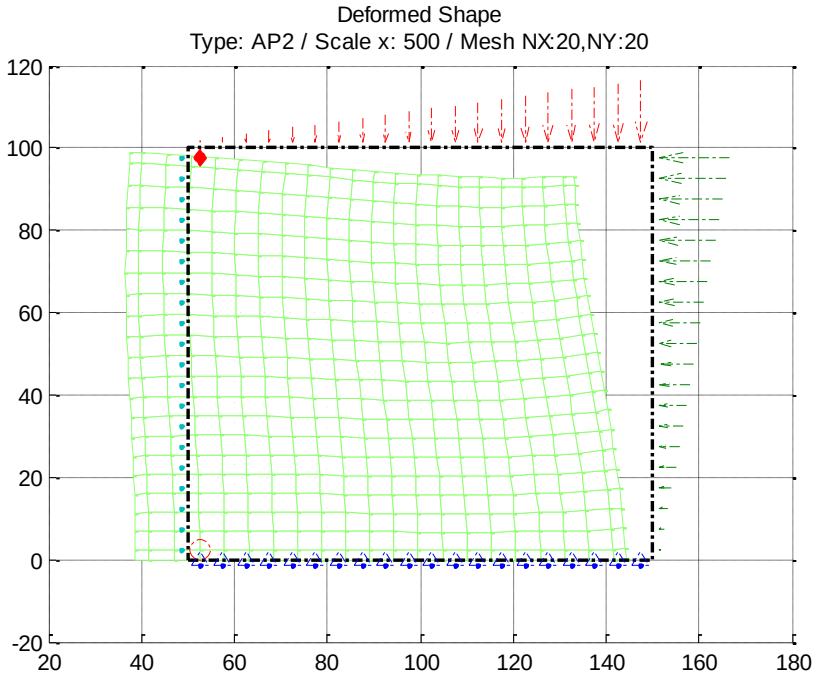


Figure 6 Solution of the displacement field by ESPNET_10

 ISBN: 978-84-614-6167-7

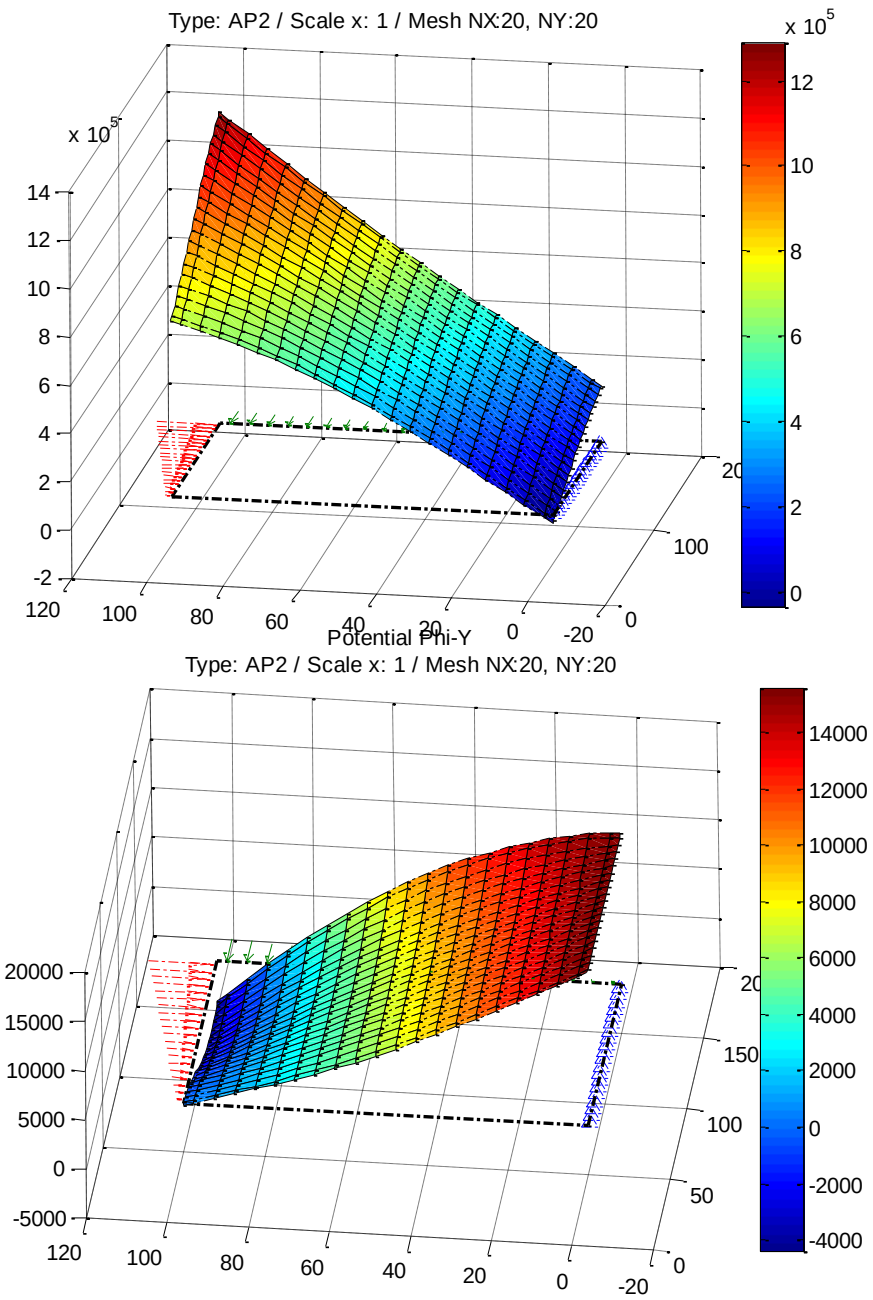


Figure 7 Potential solutions: Scalar potential (up),vector potential (down)

Finally, Figures 8 and 9 shown the contour map of the different kind of stress provide directly by code EPSNET_10 using the Equation (6a-c) and the Hooke Law for the circumferential stress component.

 ISBN: 978-84-614-6167-7

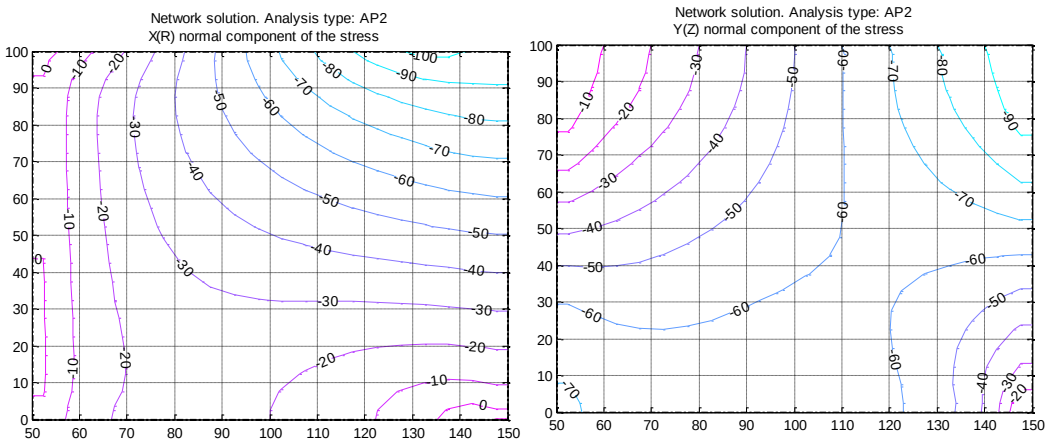


Figure 8 Stress solution: σ_{rr} (left), σ_{zz} (rigth)

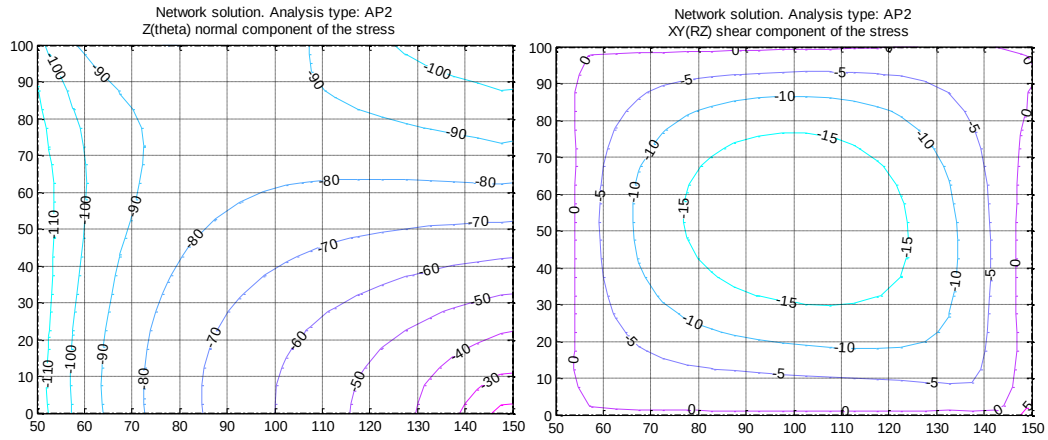


Figure 9 Stress solution: $\sigma_{\theta\theta}$ (left), σ_{rz} (rigth)

5. Conclusions

Alternative first class (Dirichlet) conditions to those proposed by Tran-Cong for the Boussinesq elasticity axisymmetric problem have been proposed and successfully applied to a hollow cylinder under a set of loads that results in a non-zero displacement field as well as for a zero displacement field (no load). These new conditions are more easily implemented in the numerical scheme. The form of the two Boussinesq potential functions in the case of zero displacement field for both Tran-Cong conditions and ours, are the same and define two planes: one horizontal for the z-component of the vector function and one parallel to the radial axis for the scalar function. The parameters that define the planes depend on the value of the numerical condition required to satisfy the uniqueness requirements.

 ISBN: 978-84-614-6167-7

6. References

[1] J.R. BARBER, *Elasticity*. Series: Solid Mechanics and its Applications, vol. 172, 3rd ed. Springer, 2010.

[2] P.F. PAPKOVICH, *An expression for a general integral of the equations of the theory of elasticity in terms of harmonic functions*, Izvest. Akad. Nauk SSSR, Ser. Matem. K. estestv. neuk, no. 10, 1932.

[3] H.NEUBER, *Ein neuer Ansatz zur Lösung räumlicher Probleme der Elastizitätstheorie*, Z. angew. Math. Mech., vol 14, p. 203, 1934.

[4] R.D. MINDLIN, *Note on the Galerkin and Papkovich stress functions*, Bulletin of American Mathematical Society, Vol. 42, pp. 373-376, 1936.

[5] R.A. EUBANKS AND E. STERNBERG, *On the completeness of the Boussinesq-Papkovich Stress Functions*, J. Rat. Mech. Anal. , Vol. 5, 735-746, 1956.

[6] J. BOUSSINESQ, *Application des potentiels a l'étude de l'équilibre et du mouvement des solides élastiques*, Gauthier-Villars, Paris, 1885.

[7] M. STIPPES, *Completeness of the Papkovich Potentials. Quart. Appl. Math*. 26, 477-483, 1969.

[8] T. TRAN-CONG, *On the Completeness and Uniqueness of the Papkovich–Neuber and the Non-axisymmetric Boussinesq, Love, and Burgatti solutions in General Cylindrical Coordinates*, J. Elast., 36, pp. 227–255, 1995.

[9] EPSNET_10 (© UPCT), Elasticity Problems Simulation Network, Network Simulation Group, Technical University of Cartagena, UPCT, 2010.

[10] http://fisica.upct.es/redes/

[11] C.F. GONZÁLEZ FERNÁNDEZ, *Network Simulation Method*, Horno Ed. Research Signpost, Trivandrum, India, 2002.

[12] PSpice, Release 6.0, Microsim Corporation, 1994.

 ISBN: 978-84-614-6167-7

Proceedings of the 11th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE2011
Benidorm, Spain, 26-30 June 2011

Numerical solution of elastostatic, axisymmetric problems using the Papkovich-Neuber potentials

Morales J.L.[1], Moreno J.A.[2] and Alhama F.[3]

[1] *Department of Structures and Construction*
[2] *Department of Mechanical Engineering*
[3]*Department of Applied Physics*
Technical University of Cartagena (UPCT)
Campus Muralla del Mar, ETSII, 30202, Cartagena, Spain

emails: joseluis.morales@upct, josea.moreno@upct.es,
paco.alhama@upct.es

Abstract

Solutions derived from the general form of Papkovich-Neuber representation, used as the governing equations of axisymmetric problems in which the unknowns are the potential functions, are numerically solve by network method. The use of each derived potentials, Boussinesq and Timpe, as well as the complete Papkovich-Neuber, leads to the same displacement and stress fields, being the potential functions clearly different in each solution.

Key words: Papkovich-Neuber representation, numerical simulation, linear elasticity, axisymmetric, network method

1. Introduction

The use of general solutions in the elasticity theory, based in stress and displacement functions, has been demonstrated to be a useful tool for the analytical study of the elastic problem. This technique allows reduce or even eliminate the strong coupling of the original differential equations of Navier or Beltrami, although the required boundary conditions are generally more complex. There are many cases in which the presence of symmetries, or the existence of infinite domains, lead to a simplification of the general solutions. This has solved the solution of problems of theoretical and practical interest, even in the presence

 ISBN: 978-84-614-6167-7

of singularities (Kelvin and Boussinesq problems). One of the most famous general solutions is due to Papkovich [1] and Neuber [2] (PN hereinafter), in which the displacement field is defined by four harmonic potential functions, grouped in a scalar potential plus a vector potential. The completeness of this solution has been demonstrated by Mindlin [3]. Eubanks and Sternberg [4] studied the uniqueness of the PN general solution in 3D rectangular coordinates domains, obtaining the conditions for which the solution could be reduced to only three harmonic potentials, maintaining the completeness.

For the axisymmetric case, the circumferential component of the vector potential disappears in the PN representation, reducing the general solution to only three harmonic potentials. In this case Boussinesq [5] proposed a general solution set up by two harmonic functions. Eubanks and Sternberg [4] noted that the Boussinesq solution refers to the PN representation in cylindrical coordinates, where the vector potential has been reduced to its axial component. These authors, demonstrated that the Boussinesq solution is valid (or complete) whenever the meridional half plane is simply connected. Another general solution for the axisymmetric case in terms of two harmonic potentials is proposed by Timpe [6] and, as the case of Boussinesq, it can be understood as a solution coming from the PN representation by reduction of the vector potential to the only radial component. The test for the completeness for Timpe solution can be carried out in a similar way that Eubanks and Sternberg made for the Boussinesq solution: meridional half plane must be, again, simply connected.

However, as regards the use of general solutions for the numerical approaches of elastostatic problems, in contrast with those based on stress functions (as Airy function [7]), displacement potentials have been scarcely applied as primary unknowns of the mathematical model.

In this work a same axisymmetric problem (cylinder under symmetric load) is numerically solved both in terms of two harmonic potentials (Boussinesq and Timpe solutions) and in terms of the complete set of three harmonic potentials (PN solution). For each case, the resulting unknown functions are markedly different each other, while the stress and displacement fields, coming from theses values, are the same.

All the simulations were carried out in EPSNET_10 [8], a specific software based on the network method [9], developed by the research group ‘simulation by networks’ of the Technical University of Cartagena. This code contains suitable subroutines to transport output data to Matlab for an optimal representation of all variables of the problem: Potential functions, displacement and stress fields. Network models are run in the circuit simulation code PSpice [10].

 ISBN: 978-84-614-6167-7

2. Papkovich-Neuber representation for axisymmetric problems. Governing equations and boundary conditions

The equilibrium equation in terms of displacements, in the absence of body forces, is the Navier equation [7]

$$\mu\nabla^2\boldsymbol{u} + (\lambda+\mu)\boldsymbol{\nabla}(\boldsymbol{\nabla}\cdot\boldsymbol{u}) = \mathbf{0} \tag{1}$$

where $\boldsymbol{u}$ is the displacements field, λ the Lamé's constant and μ the shear modulus. This equation admits the general Papkovich-Neuber displacement representations [1,2]

$$2\mu\boldsymbol{u} = \boldsymbol{\phi} - \boldsymbol{\nabla}\left(\phi_0 + \frac{\boldsymbol{\phi}\cdot\boldsymbol{R}}{4(1-\nu)}\right) \tag{2a}$$

$$\nabla^2\boldsymbol{\phi} = \mathbf{0} \quad , \quad \nabla^2\phi_0 = 0 \tag{2b}$$

where $\boldsymbol{\phi}$ is the vector potential, ϕ_0 the scalar potential, $\boldsymbol{R}$ the position vector and ν the Poisson ratio. As mention before, equation (2a) is a general solution of (1) and has proved to be complete for the general case [3].

2.1 Papkovich-Neuber solution

For cylindrical coordinates, with $\boldsymbol{\phi} \equiv \boldsymbol{\phi}(\phi_r, \phi_\theta, \phi_z)$, the components ($u_r$, u_θ, u_z) of the equation (2a) can be written as

$$2\mu u_r = -\frac{\partial\phi_0}{\partial r} + \frac{3-4\nu}{4(1-\nu)}\phi_r - \frac{1}{4(1-\nu)}\left(r\frac{\partial\phi_r}{\partial r} + z\frac{\partial\phi_z}{\partial r}\right) \tag{3a}$$

$$2\mu u_\theta = -\frac{1}{r}\frac{\partial\phi_0}{\partial\theta} + \phi_\theta - \frac{1}{4(1-\nu)}\frac{1}{r}\left(r\frac{\partial\phi_r}{r\partial\theta} + z\frac{\partial\phi_z}{r\partial\theta}\right) \tag{3b}$$

$$2\mu u_z = -\frac{\partial\phi_0}{\partial z} + \frac{3-4\nu}{4(1-\nu)}\phi_z - \frac{1}{4(1-\nu)}\left(r\frac{\partial\phi_r}{\partial z} + z\frac{\partial\phi_z}{\partial z}\right) \tag{3c}$$

The functions $\boldsymbol{\phi}$ and ϕ_0 in equation (2b) reduce to scalar and vector harmonic functions, which lead to the following four equations

$$\left.\begin{aligned}
&\frac{\partial^2\phi_0}{\partial r^2} + \frac{1}{r}\frac{\partial\phi_0}{\partial r} + \frac{1}{r^2}\frac{\partial^2\phi_0}{\partial\theta^2} + \frac{\partial^2\phi_0}{\partial z^2} = 0\\
&\frac{\partial^2\phi_r}{\partial r^2} + \frac{1}{r}\frac{\partial\phi_r}{\partial r} + \frac{1}{r^2}\frac{\partial^2\phi_r}{\partial\theta^2} + \frac{\partial^2\phi_r}{\partial z^2} - \frac{\phi_r}{r^2} - \frac{2}{r^2}\frac{\partial\phi_\theta}{\partial\theta} = 0\\
&\frac{\partial^2\phi_\theta}{\partial r^2} + \frac{1}{r}\frac{\partial\phi_\theta}{\partial r} + \frac{1}{r^2}\frac{\partial^2\phi_\theta}{\partial\theta^2} + \frac{\partial^2\phi_\theta}{\partial z^2} - \frac{\phi_\theta}{r^2} + \frac{2}{r^2}\frac{\partial\phi_r}{\partial\theta} = 0\\
&\frac{\partial^2\phi_z}{\partial r^2} + \frac{1}{r}\frac{\partial\phi_z}{\partial r} + \frac{1}{r^2}\frac{\partial^2\phi_z}{\partial\theta^2} + \frac{\partial^2\phi_z}{\partial z^2} = 0
\end{aligned}\right\} \tag{4}$$

For torsion-free axisymmetric problems the equation $u_\theta = \frac{\partial}{\partial\theta}$=0 is satisfied, and u_r and u_z depend on the variables r and z. This eliminates the second component of the potential vector, equation (3a). Thus, the Papkovich-Neuber´s solution for axisymmetric problems has the form

 ISBN: 978-84-614-6167-7

$$2\mu u_r = -\frac{\partial \phi_0}{\partial r} + \frac{3-4\nu}{4(1-\nu)}\phi_r - \frac{1}{4(1-\nu)}\left(r\frac{\partial \phi_r}{\partial r} + z\frac{\partial \phi_z}{\partial r}\right) \quad (5a)$$

$$2\mu u_z = -\frac{\partial \phi_0}{\partial z} + \frac{3-4\nu}{4(1-\nu)}\phi_z - \frac{1}{4(1-\nu)}\left(r\frac{\partial \phi_r}{\partial z} + z\frac{\partial \phi_z}{\partial z}\right) \quad (5b)$$

$$\frac{\partial^2 \phi_0}{\partial \mathrm{r}^2} + \frac{1}{\mathrm{r}}\frac{\partial \phi_0}{\partial \mathrm{r}} + \frac{\partial^2 \phi_0}{\partial \mathrm{z}^2} = 0 \quad (6b)$$

$$\frac{\partial^2 \phi_r}{\partial \mathrm{r}^2} + \frac{1}{\mathrm{r}}\frac{\partial \phi_r}{\partial \mathrm{r}} + \frac{\partial^2 \phi_r}{\partial \mathrm{z}^2} - \frac{\phi_r}{\mathrm{r}^2} = 0 \quad (6b)$$

$$\frac{\partial^2 \phi_z}{\partial \mathrm{r}^2} + \frac{1}{\mathrm{r}}\frac{\partial \phi_z}{\partial \mathrm{r}} + \frac{\partial^2 \phi_z}{\partial \mathrm{z}^2} = 0 \quad (6c)$$

To complete the mathematical model, boundary equations are required. For the mixed boundary conditions, these are

$$u_i = u_i^b \quad on \quad S_u \quad (7a)$$

$$\sigma_{ij} n_j = t_i^b \quad on \quad S_t \quad (7b)$$

S_u denotes boundary surface points where the displacement, u_i^b, is prescribed, while S_t refers to points where the traction values, t_i^b, are imposed. σ_{ij} is the stress field and n_j the outer normal vector on S_t. Note that $S = S_t + S_u$ represents the complete boundary of the middle section, Figure 2.

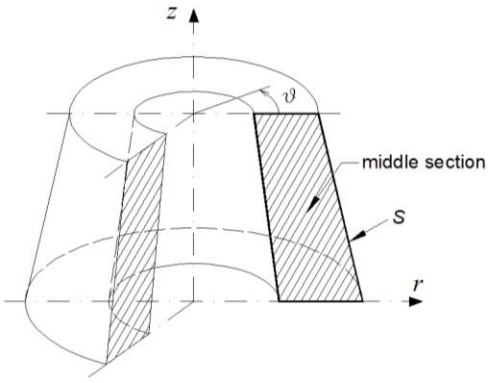


Figure 1. Axissymetric body

The expressions that relate stress and potential functions, required for implementing the boundary condition (7b), are obtained from the strain-displacement equations and Hook's Law for the axisymmetric case [7]. Using equation (6) to simplify, these expressions are:

$$\sigma_{rr} = -\frac{\partial^2 \phi_0}{\partial r^2} - r\frac{\partial^2 \phi_r}{\partial r^2}\frac{1}{4(1-\nu)} + \frac{\partial \phi_r}{\partial \mathrm{r}}\frac{1}{2} + \frac{\phi_r}{r}\frac{\nu}{2(1-\nu)} - z\frac{\partial^2 \phi_z}{\partial r^2}\frac{1}{4(1-\nu)} + \frac{\partial \phi_z}{\partial z}\frac{\nu}{2(1-\nu)} \quad (8a)$$

$$\sigma_{zz} = -\frac{\partial^2 \phi_0}{\partial z^2} - r\frac{\partial^2 \phi_r}{\partial z^2}\frac{1}{4(1-\nu)} + \frac{\partial \phi_r}{\partial \mathrm{r}}\frac{\nu}{2(1-\nu)} + \frac{\phi_r}{r}\frac{\nu}{2(1-\nu)} - z\frac{\partial^2 \phi_z}{\partial z^2}\frac{1}{4(1-\nu)} + \frac{\partial \phi_z}{\partial z}\frac{1}{2} \quad (8b)$$

$$\sigma_{rz} = -\frac{\partial^2 \phi_0}{\partial r \partial z} - r\frac{\partial^2 \phi_r}{\partial r \partial z}\frac{1}{4(1-\nu)} - z\frac{\partial^2 \phi_z}{\partial r \partial z}\frac{1}{4(1-\nu)} + \frac{\partial \phi_z}{\partial \mathrm{r}}\frac{1-2\nu}{4(1-\nu)} + \frac{\partial \phi_r}{\partial z}\frac{1-2\nu}{4(1-\nu)} \quad (8c)$$

These formulas can also be used as post-processing equations to derive the

 ISBN: 978-84-614-6167-7

quantities σ_{rr}, σ_{zz} and σ_{rz} once the potential fields are known. In short, the mathematical model for the Papkovich-Neuber´s solution is formed by equations (6) and (7) plus the expresions of displacements and stress in terms of the potential functions, equations (5) on S_u, and (8) on S_t, respectively.

2.2 Boussinesq's and Timpe´s solutions

Now, we will consider two known solutions that emerge from the Papkovich-Neuber representation. These result by dropping ϕ_r or ϕ_z, Boussinesq's and Timpe's solution, respectively. The first, $\phi_r = 0$, is setting by equations

$$2\mu u_r = -\frac{\partial \phi_0}{\partial r} - \frac{1}{4(1-\nu)} z \frac{\partial \phi_z}{\partial r} \tag{9a}$$

$$2\mu u_z = -\frac{\partial \phi_0}{\partial z} + \frac{3-4\nu}{4(1-\nu)} \phi_z - \frac{1}{4(1-\nu)} z \frac{\partial \phi_z}{\partial z} \tag{9b}$$

$$\frac{\partial^2 \phi_0}{\partial \mathrm{r}^2} + \frac{1}{\mathrm{r}} \frac{\partial \phi_0}{\partial \mathrm{r}} + \frac{\partial^2 \phi_0}{\partial \mathrm{z}^2} = 0 \tag{10a}$$

$$\frac{\partial^2 \phi_z}{\partial \mathrm{r}^2} + \frac{1}{\mathrm{r}} \frac{\partial \phi_z}{\partial \mathrm{r}} + \frac{\partial^2 \phi_z}{\partial \mathrm{z}^2} = 0 \tag{10b}$$

$$\sigma_{rr} = -\frac{\partial^2 \phi_0}{\partial r^2} - z \frac{\partial^2 \phi_z}{\partial r^2} \frac{1}{4(1-\nu)} + \frac{\partial \phi_z}{\partial z} \frac{\nu}{2(1-\nu)} \tag{11a}$$

$$\sigma_{zz} = -\frac{\partial^2 \phi_0}{\partial z^2} - z \frac{\partial^2 \phi_z}{\partial z^2} \frac{1}{4(1-\nu)} + \frac{\partial \phi_z}{\partial z} \frac{1}{2} \tag{11b}$$

$$\sigma_{rz} = -\frac{\partial^2 \phi_0}{\partial r \partial z} - z \frac{\partial^2 \phi_z}{\partial r \partial z} \frac{1}{4(1-\nu)} + \frac{\partial \phi_z}{\partial \mathrm{r}} \frac{1-2\nu}{4(1-\nu)} \tag{11c}$$

Timpe's solution, $\phi_z = 0$, leads to the governing equations:

$$2\mu u_r = -\frac{\partial \phi_0}{\partial r} + \frac{3-4\nu}{4(1-\nu)} \phi_r - \frac{1}{4(1-\nu)} r \frac{\partial \phi_r}{\partial r} \tag{12a}$$

$$2\mu u_z = -\frac{\partial \phi_0}{\partial z} - \frac{1}{4(1-\nu)} r \frac{\partial \phi_r}{\partial z} \tag{12b}$$

$$\frac{\partial^2 \phi_0}{\partial \mathrm{r}^2} + \frac{1}{\mathrm{r}} \frac{\partial \phi_0}{\partial \mathrm{r}} + \frac{\partial^2 \phi_0}{\partial \mathrm{z}^2} = 0 \tag{13a}$$

$$\frac{\partial^2 \phi_r}{\partial \mathrm{r}^2} + \frac{1}{\mathrm{r}} \frac{\partial \phi_r}{\partial \mathrm{r}} + \frac{\partial^2 \phi_r}{\partial \mathrm{z}^2} - \frac{\phi_r}{\mathrm{r}^2} = 0 \tag{13b}$$

$$\sigma_{rr} = -\frac{\partial^2 \phi_0}{\partial r^2} - r \frac{\partial^2 \phi_r}{\partial r^2} \frac{1}{4(1-\nu)} + \frac{\partial \phi_r}{\partial \mathrm{r}} \frac{1}{2} + \frac{\phi_r}{r} \frac{\nu}{2(1-\nu)} \tag{14a}$$

$$\sigma_{zz} = -\frac{\partial^2 \phi_0}{\partial z^2} - r \frac{\partial^2 \phi_r}{\partial z^2} \frac{1}{4(1-\nu)} + \frac{\partial \phi_r}{\partial \mathrm{r}} \frac{\nu}{2(1-\nu)} + \frac{\phi_r}{r} \frac{\nu}{2(1-\nu)} \tag{14b}$$

$$\sigma_{rz} = -\frac{\partial^2 \phi_0}{\partial r \partial z} - r \frac{\partial^2 \phi_r}{\partial r \partial z} \frac{1}{4(1-\nu)} + \frac{\partial \phi_r}{\partial \mathrm{z}} \frac{1-2\nu}{4(1-\nu)} \tag{14c}$$

3. The network model

It is designed from the finite differential equations derived from Boussinesq´s or Timpe´s solutions, equations (10) and (13), respectively. Using the nomenclature of Figure 2, these equations are (15) and (16), respectively.

 ISBN: 978-84-614-6167-7

$$\left.\begin{aligned}\frac{\phi^0_{k,0}-\phi^0_{k,2}}{\left(\Delta r^2/_2\right)}+\frac{\phi^0_{k,0}-\phi^0_{k,4}}{\left(\Delta r^2/_2\right)}+\frac{\phi^0_{k,0}-\phi^0_{k,3}}{\left(\Delta z^2/_2\right)}+\frac{\phi^0_{k,0}-\phi^0_{k,1}}{\left(\Delta z^2/_2\right)}-\left[\frac{1}{r_{k,0}}\frac{\phi^0_{k,2}-\phi^0_{k,4}}{\Delta r}\right]=0\\ \frac{\phi^z_{k,0}-\phi^z_{k,2}}{\left(\Delta r^2/_2\right)}+\frac{\phi^z_{k,0}-\phi^z_{k,4}}{\left(\Delta r^2/_2\right)}+\frac{\phi^z_{k,0}-\phi^z_{k,3}}{\left(\Delta z^2/_2\right)}+\frac{\phi^z_{k,0}-\phi^z_{k,1}}{\left(\Delta z^2/_2\right)}-\left[\frac{1}{r_{k,0}}\frac{\phi^z_{k,2}-\phi^z_{k,4}}{\Delta r}\right]=0\end{aligned}\right\} \quad (15)$$

$$\left.\begin{aligned}\frac{\phi^0_{k,0}-\phi^0_{k,2}}{\left(\Delta r^2/_2\right)}+\frac{\phi^0_{k,0}-\phi^0_{k,4}}{\left(\Delta r^2/_2\right)}+\frac{\phi^0_{k,0}-\phi^0_{k,3}}{\left(\Delta z^2/_2\right)}+\frac{\phi^0_{k,0}-\phi^0_{k,1}}{\left(\Delta z^2/_2\right)}-\left[\frac{1}{r_{k,0}}\frac{\phi^0_{k,2}-\phi^0_{k,4}}{\Delta r}\right]=0\\ \frac{\phi^r_{k,0}-\phi^r_{k,2}}{\left(\Delta r^2/_2\right)}+\frac{\phi^r_{k,0}-\phi^r_{k,4}}{\left(\Delta r^2/_2\right)}+\frac{\phi^r_{k,0}-\phi^r_{k,3}}{\left(\Delta z^2/_2\right)}+\frac{\phi^r_{k,0}-\phi^r_{k,1}}{\left(\Delta z^2/_2\right)}-\left[\frac{1}{r_{k,0}}\frac{\phi^r_{k,2}-\phi^r_{k,4}}{\Delta r}-\frac{\phi^r_{k,0}}{r^2_{k,0}}\right]=0\end{aligned}\right\} \quad (16)$$

Following the steps of the design [9], the four first addends of equations (15) and (16) are implemented by simple resistors while the last addends between brackets (coupled terms) are implemented by controlled current sources.

To extend the model to the whole domain, $N_x \times N_y$ networks must be electrically connected each other along axes x and y, Figure 3. Finally, displacement boundary conditions are directly implemented by means of constant voltage generators, while traction conditions are implemented by controlled voltage sources.

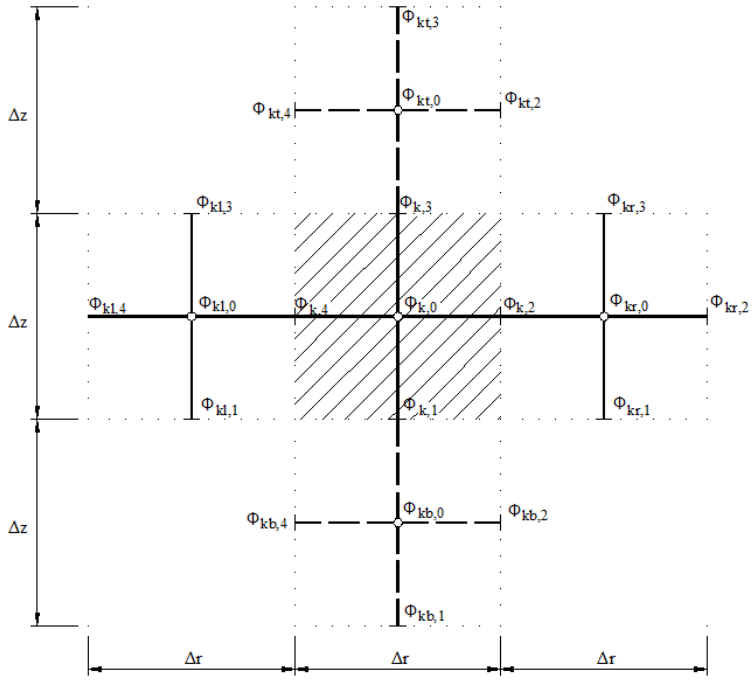


Figure 2 Nomenclature of the nodes. Note that there is one figure for the scalar potential (ϕ_0) and another for the i-component of the vector potential (ϕ_i)

 ISBN: 978-84-614-6167-7

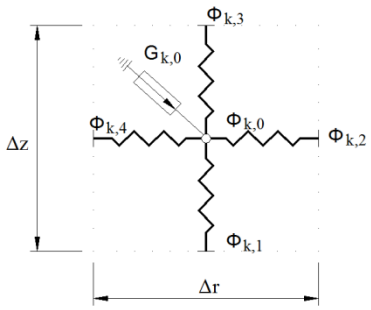


Figure 3 Network model of the volume element

4. The code EPSNET_10 (© UPCT)

This program, developed in a window ambience of MATLAB as programming tool makes the following: i) implements routines for data input, ii) designs the network model, iii) starts up PSpice for the numerical solution of the model and, iv) accesses to MATLAB for the post processing of output data. In its educational version, EPSNET_10 can be applied for solving 2-D problems in rectangular and axisymmetric coordinates. The types of analysis and formulation, as well as the list of quantities that can be drawn are listed in Table 1.

Type of analysis	Quantities to be graphed
Navier Plane Stress	DISPLACEMENT SURFACES
Navier Plane Strain	Deformed Shape
Navier Axisymmetric	X/R Component
Potential Axisymmetric Phi0, PhiR	Y/Z Component
Potential Axisymmetric Phi0, PhiZ	STRESS SURFACES
Potential Axisymmetric Phi0, PhiR, PhiZ	X/R Normal component
Potential Plane Stress Phi0, PhiX	Y/Z Normal component
Potential Plane Stress Phi0, PhiY	.
Potential Plane Stress Phi0, PhiX, PhiY	STRESS CONTOUR
Potential Plane Strain Phi0, PhiX	.
Potential Plane Strain Phi0, PhiY	.
Potential Plane Strain Phi0, PhiX, PhiY	POTENTIAL SURFACES

Table 1 Analysis and formulation options (left) and quantities to be drawn (rigth) provided by EPSNET_10

5. Applications

The physical model is a solid cylinder with the restricted displacement and load distribution depicted in Figure 4 (left). The grid, 20x20, refers to the 2-D section, Figure 4 (right). Values of the geometry and elastic parameters are R = 100 mm, H = 100 mm, E (Young modulus) = 210 GPa and ν (Poisson ratio) = 0.3. Additional conditions are: i) PN representation, $\phi_0=\phi_r=0$ in the cell (20,20), $\phi_r=0$ in the cell (1,20); ii) Timpe representation, $\phi_0=0$ in the cell (1,20); iii) Boussinesq representation, $\phi_0=0$ in the cell (1,1).

 ISBN: 978-84-614-6167-7

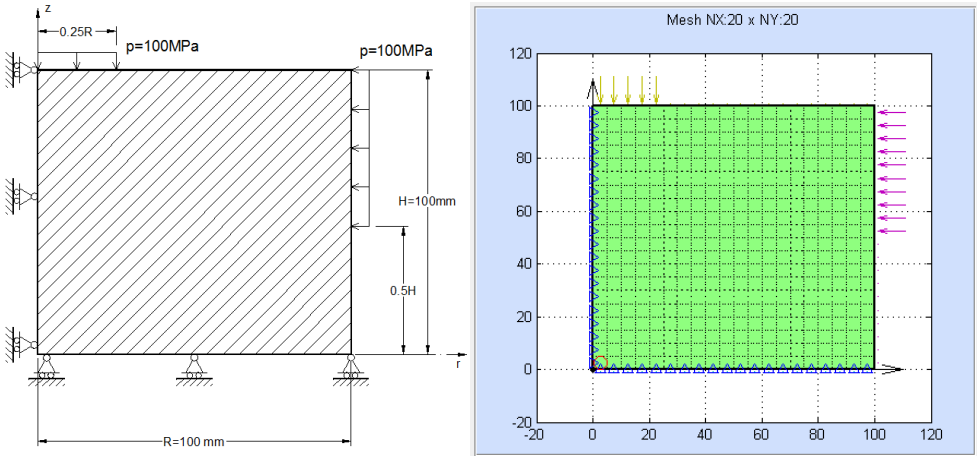


Figure 4 Physical model (left). Mesh grid (right)

On the one hand, Figure 5 shows the von Mises stress and the deformed section of the solid cylinder for the three solutions studied. Both stress and displacements are quite similar as expected. On the other hand, Figures 6, 7 and 8 show the spatial representation of the potential functions relatives to the Timpe, Boussinesq and PN solutions, respectively. The perspective of these figures has been changed for a better representation of the spatial surfaces. It is interesting to see that the unknown potential functions are quite different for each solution; however, these provide the same elastic solution trough the equations (14), (11) and (8), for Timpe, Boussinesq and PN potential, respectively. In addition, it is worthy to note that the potential solutions are always strongly dependent on the additional conditions, but always providing the same elastic solution.

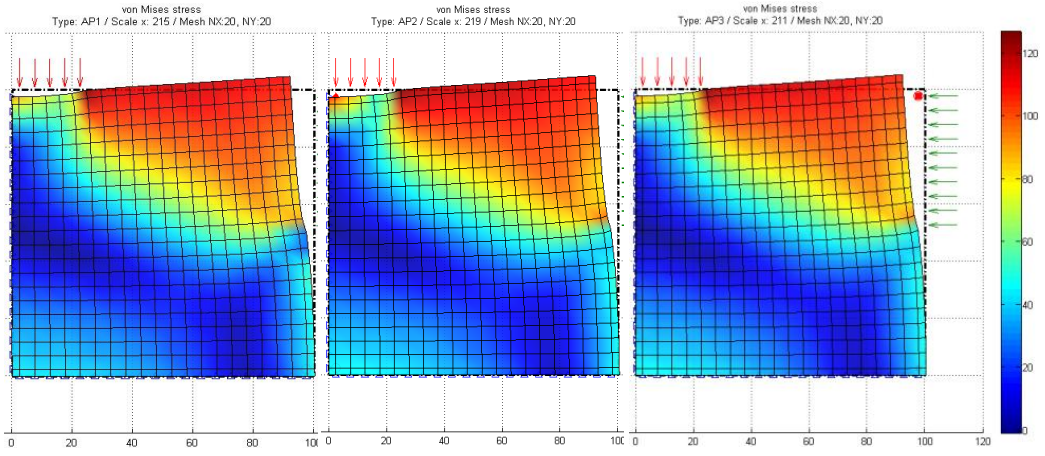


Figure 5 Von Mises stress on the deformed shape. Timpe solution (left), Boussinesq solution (center) and PN solution (right)

 ISBN: 978-84-614-6167-7

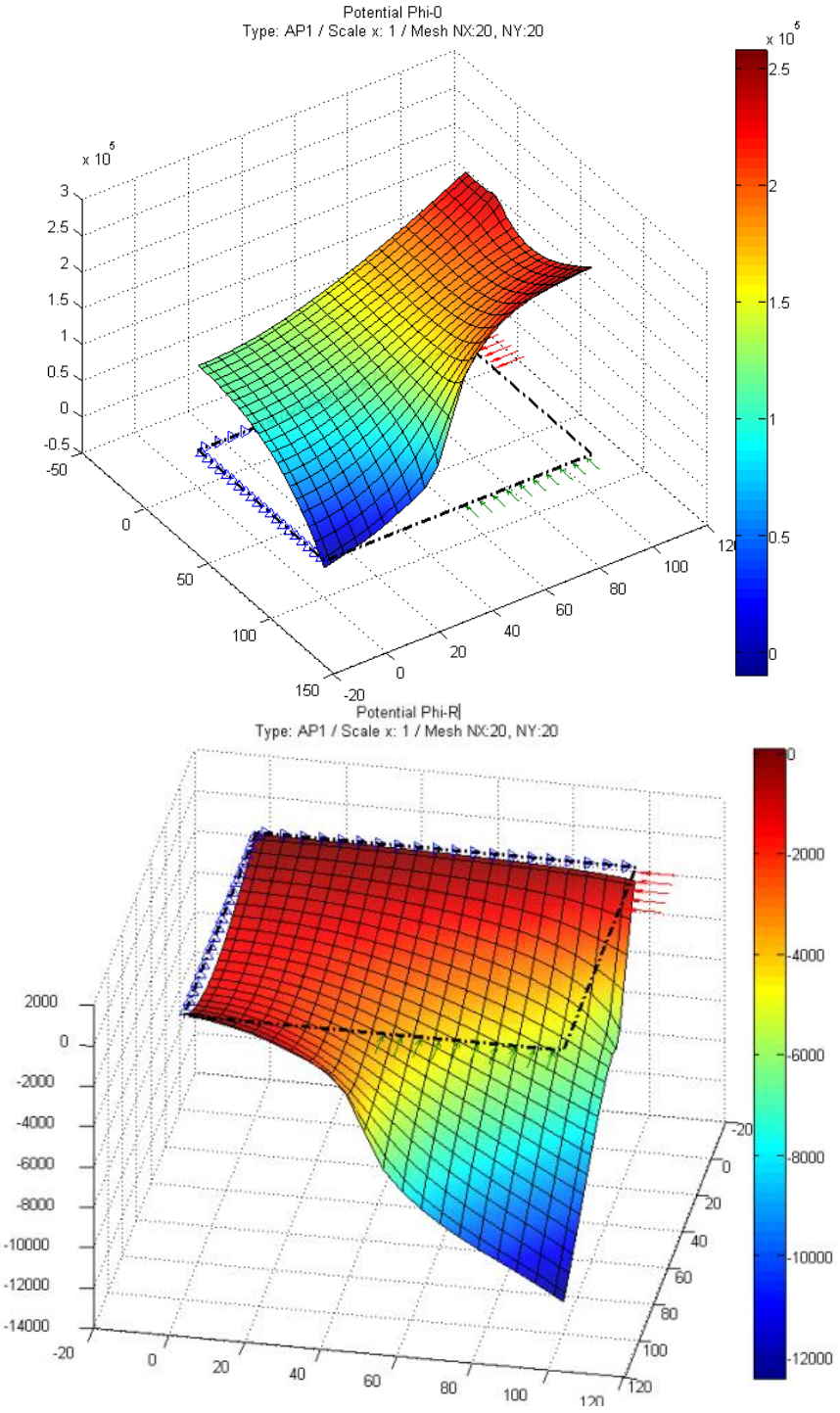


Figure 6 Timpe's solution: ϕ_0 (up) and ϕ_r (down)

 ISBN: 978-84-614-6167-7

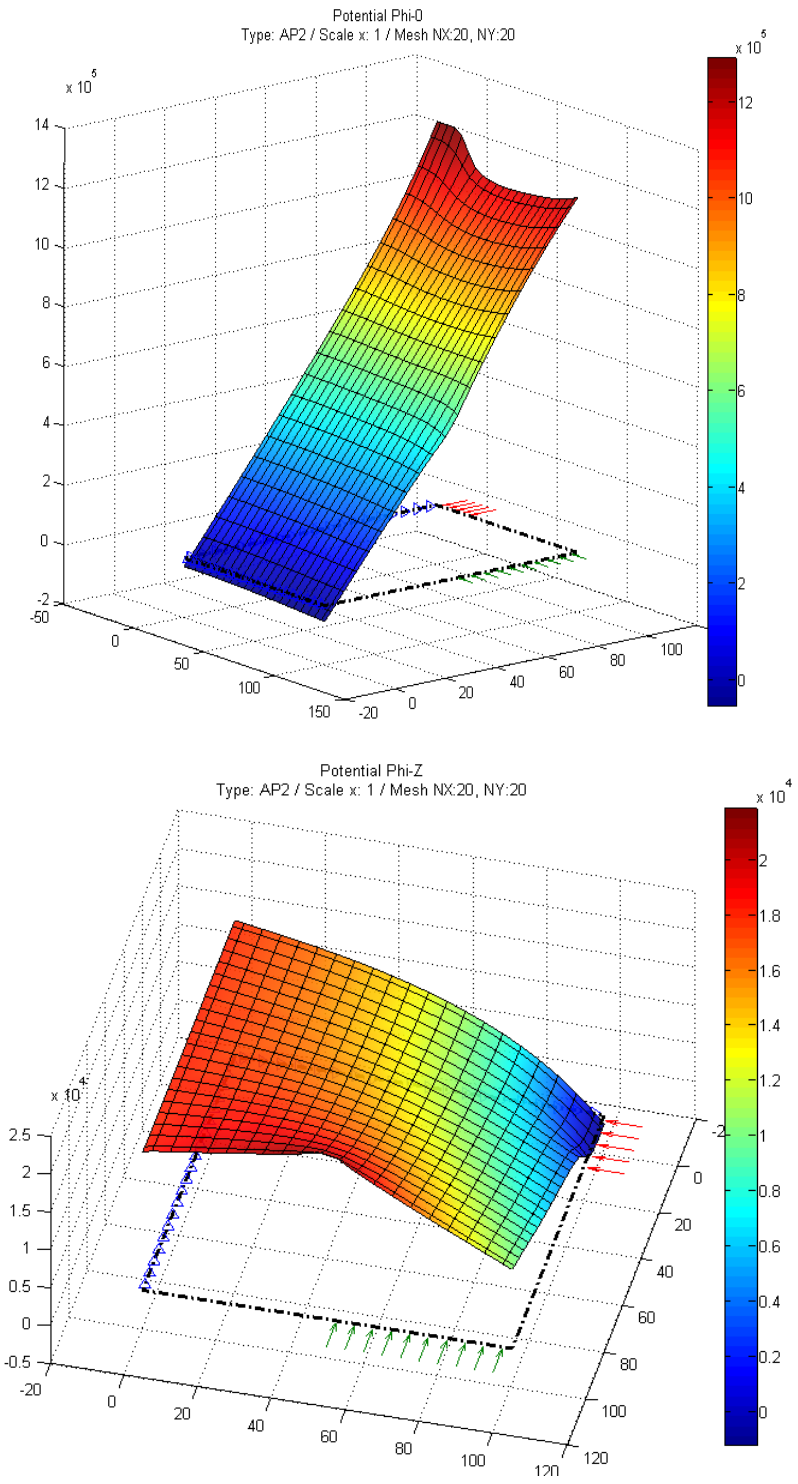


Figure 7 Boussinesq's solution: ϕ_0 (up) and ϕ_z (down)

 ISBN: 978-84-614-6167-7

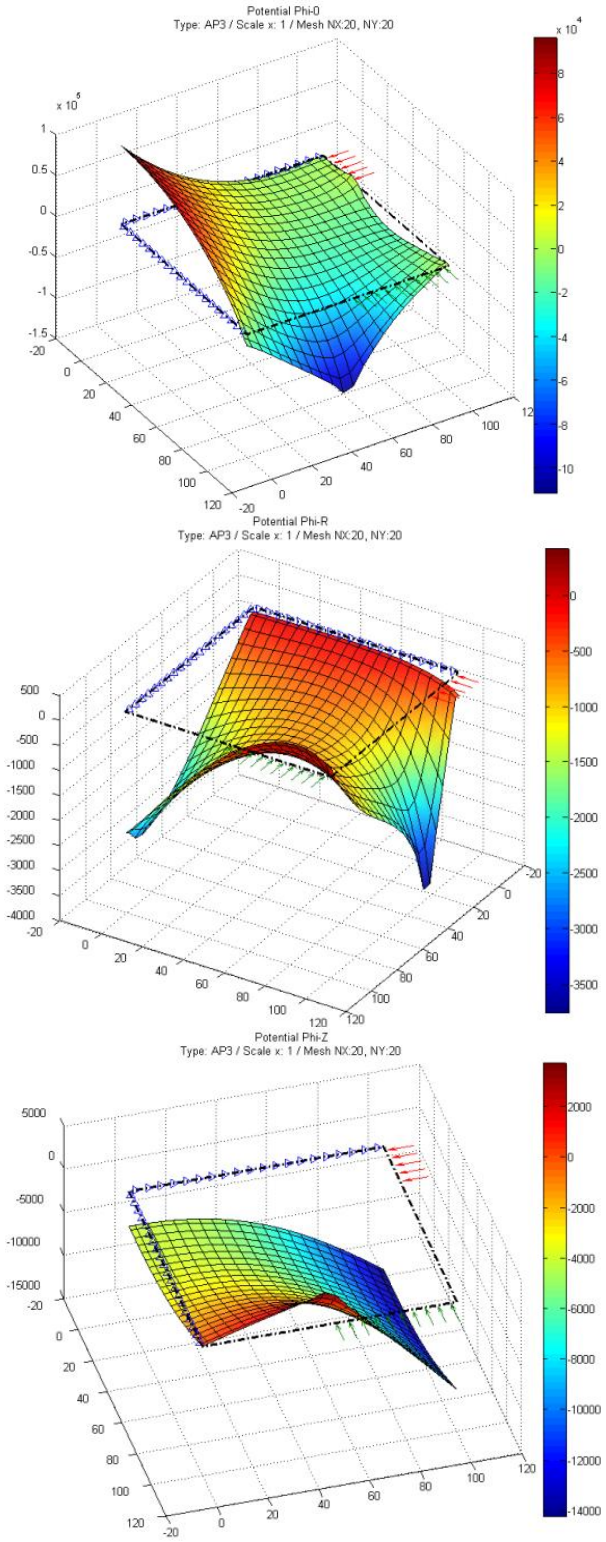


Figure 8 PN's solution: ϕ_0 (up), ϕ_r (center) and ϕ_z (down)

 ISBN: 978-84-614-6167-7

6. Conclusions

The axisymmetric elastic problem formulated by harmonic potential functions has been numerically solved using as the unknown variables the primary potentials: ϕ_0 and ϕ_r for Timpe's, ϕ_0 and ϕ_z for Boussinesq's and ϕ_0, ϕ_r and ϕ_z for Papkovich-Neuber's representation. Simulations are carried out by the program EPSNET_10 whose numerical scheme is based in the network simulation method. The proposed application for the three cases (a solid cylinder under particular displacement and load boundary conditions) provides the same elastic solution for the three representations, as expected, while potential results are strongly different for each formulation.

7. References

[1] P.F. PAPKOVICH, *An expression for a general integral of the equations of the theory of elasticity in terms of harmonic functions*, Izvest. Akad. Nauk SSSR, Ser. Matem. K. estestv. neuk, no. 10, 1932.

[2] H. NEUBER, *Ein neuer Ansatz zur Lösung räumlicher Probleme der Elastizitätstheorie*, Z. angew. Math. Mech., vol 14, p. 203, 1934.

[3] R.D. MINDLIN, *Note on the Galerkin and Papkovich stress functions, Bulletin of American Mathematical Society*, Vol. 42,pp. 373-376, 1936.

[4] R.A. EUBANKS AND E. STERNBERG, *On the completeness of the Boussinesq-Papkovich Stress Functions*, J. Rat. Mech. Anal. , Vol. 5, 735-746, 1956.

[5] J. BOUSSINESQ, *Application des potentiels a l'étude de l'équilibre et du mouvement des solides élastiques*, Gauthier-Villars, Paris, 1885.

[6] A. TIMPE, *Achensymmetrische Deformation von Under Hungsorpern*, Z. Angew. Math. Mech., 4, pp. 361–376, 1924.

[7] J.R. BARBER, *Elasticity*. Series: Solid Mechanics and its Applications, vol. 172, 3rd ed. Springer, 2010.

[8] EPSNET_10, Elasticity Problems Simulation Network, Network Simulation Group, Technical University of Cartagena, UPCT, 2010.

[9] C.F. GONZÁLEZ FERNÁNDEZ, *Network Simulation Method*, Horno Ed. Research Signpost, Trivandrum, India, 2002.

[10] PSpice, Release 6.0, Microsim Corporation, 1994.

[11] T. TRAN-CONG, *On the Completeness and Uniqueness of the Papkovich–Neuber and the Non-axisymmetric Boussinesq, Love, and Burgatti solutions in General Cylindrical Coordinates*, J. Elast., 36, pp. 227–255, 1995.

 ISBN: 978-84-614-6167-7

Proceedings of the 11th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE2011
Benidorm, Spain, 26-30 June 2011

Complete modal representation with discrete Zernike polynomials. Critical sampling in non redundant grids

Rafael Navarro[1], and Justo Arines[2]

[1]*ICMA, Consejo Superior de Investigaciones Científicas & Universidad de Zaragoza*
[2] *Departamento de Física Aplicada, Universidade de Santiago de Compostela*

emails: rafaelnb@unizar.es, justo.arines@usc.es

Abstract

An invertible discrete Zernike transform, DZT is proposed and implemented. Three types of non-redundant samplings, random, hybrid (perturbed deterministic) and deterministic (spiral) are shown to provide completeness of the resulting sampled Zernike polynomial expansion. When completeness is guaranteed, then we can obtain an orthonormal basis, and hence the inversion only requires transposition of the matrix formed by the basis vectors (modes). These types of nonredundant sampling patterns are also shown to guarantee completeness of the basis formed by the sampled partial derivatives of Zernike polynomials, commonly used to reconstruct the wavefront from its slopes (wavefront sensing). In the ideal noise-free case, this enables one to recover double the number of modes J than sampling points I (critical sampling J ¼ 2I). With real data, noise amplification makes the optimal number of modes lower I < J <2I. Our computer simulations show that optimized nonredundant sampling provides a significant improvement of wavefront reconstructions, with the number of modes recovered about 2.5 higher than with standard sampling patterns.
Key words: Aberration expansions; Wavefront sensing

1. Introduction

Zernike polynomials (ZPs) are continuous functions that form an orthogonal complete basis on a circle of unit radius. This property makes them extremely useful in many areas of science and technology, especially in Optics. They are extensively applied in optical design and testing, wavefront sensing, adaptive optics, wavefront shaping, interferometry, surface metrology (profilometry, topography) or atmospheric optics among others. In addition to these applications,

 ISBN: 978-84-614-6167-7

ZPs are particularly useful in optical computing, modeling, simulations, inverse problems, etc.

However, the continuous orthogonal basis becomes incomplete and non-orthogonal when the continuous circle is discretized by commonly used sampling grids (square, hexagonal, polar, etc.) The lack of these two essential properties (completeness and orthogonality) strongly limits the assumed theoretical advantages of a modal representation by ZPs in real applications. On the other hand, an important set of applications (wavefront sensing, ray tracing) deal with the measurement or computation of wavefront slopes. In that case, the basis functions are partial derivatives of ZPs. Again discrete versions can only be used in practice. The degree of discretization depends on the device and on the application, ranging for a number of samples below fifty to more than a million. Most typical values are between one and a few hundreds. Until recently, physical sampling grids where built in monolithic block and could be hardly modified. This also affected to the computational schemes which tended to mimic physical implementations. Nowadays, spatial light modulators or laser ray tracing systems permit a high flexibility so that the sampling patterns can be modified almost in real time.

In this work we overview our research [1][2], including novel unpublished latest results, on non redundant sampling patterns which allow to keep the most fundamental property of continuous ZPs that is completeness of the representation.

2. Methods

Completeness is demonstrated empirically, through computer simulations, both for basis: ZPs and their gradients (two partial derivatives). Completeness is a necessary condition to guarantee critical sampling, which means that the number of samples equals the number of modes. In the case of wavefronts this means that the number of points in the grid I equals the number of recovered modes J: $J = I$. In the case of gradient, since one has two measurements at each point it means that $J = 2 I$. This possibility was not explored before. Furthermore, the lack of completeness and orthogonality was the reason way most implementations (including numerical ray tracing and numerical simulations) considered a strong oversampling with $J << I$, which involves a great expense of resources.

Once completeness in the discrete domain is achieved it is possible to work with critical sampling which in turns warrantee the reversibility of the modal representation (Zernike transform); that is I (or 2 I) samples can be recovered from the J modes and vice versa. However this only happens in the ideal noise-free case. The computation of the Zernike representation involves the inversion of a typically large matrix. The inversion is ill-posed due to the lack of orthogonality. This means that noise (if present) is amplified unless the basis is orthogonal. We have studied the effects of noise amplification by realistic numerical simulations, finding the optimal number of modes which provides the

 ISBN: 978-84-614-6167-7

best reconstruction for different noise levels (signal-to-noise-ratios). In addition, we explored different strategies to improve orthogonality. One possible solution is to apply the QR matrix factorization (Q is orthogonal) which permits to find orthogonal modes in the discrete domain. All processes are invertible and well-conditioned, as far as one operates between discrete domains. The ill-posed problem now appears when one tries to interpolate in order to recover the continuous wavefront from discrete samples. A significant improvement towards orthogonality can also be obtained by using a quadratic increase of sampling density towards the periphery. Figure 1 shows a non-redundant spiral with such inhomogeneous density of samples, which provided a high performance in our simulations.

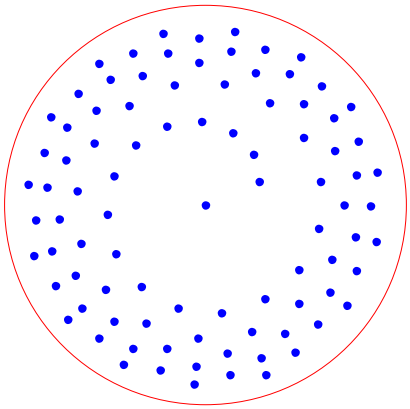


Fig. 1. Spiral sampling pattern (I= 91 samples) with quadratic increase of density towards periphery.

3. Results

As an illustrative example, Figure 2 shows the results of a realistic computer simulation of wavefront sensing of ocular aberrations, in which we compared different non-redundant sampling patterns with one standard hexagonal (redundant). The number of sampling points was always 91, and the measurements signal-to-noise ratio was 30. The horizontal axis represents the number J of Zernike modes reconstructed and the vertical axis shows the reconstruction error (RMS) of the wavefront. The optimal number of modes Jopt is the value for the minimum error. For J > Jopt noise amplification increases the reconstruction error. The worst result corresponds to standard sampling patterns (hexagonal) Jopt $< 60 <$ I, whereas the best result corresponds to the inhomogeneous spiral of Fig. 1 where Jopt $> 120 >$ I exceeds the number of sampling points even in presence of noise (typical experimental values). As a result the reconstruction is improved with a significantly lower RMS error.

 ISBN: 978-84-614-6167-7

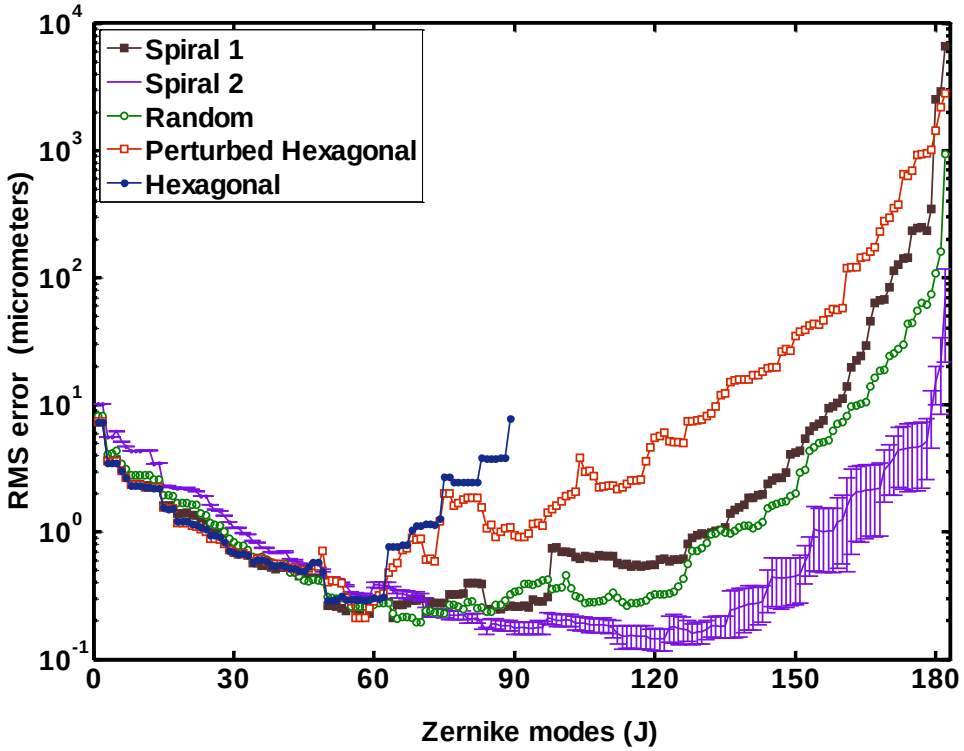


Fig. 2. Summary of results of a comparative study of wavefront sensing with different sampling patterns

4. Conclusion

In conclusion, a set of new families of non-redundant patterns, especially inhomogeneous spirals, is proposed to sample the circle, as they permit to keep completeness, and improve orthogonality of the resulting discrete Zernike polynomials. This opens many possibilities in different applications since one can increase (by a factor of 2 typically) the number of modes recovered, which provides a better reconstruction. This could provide important savings in experimental devices, or actuators in adaptive optics setups. In the field of optical design and computing, the invertibility of the transform can play an essential role in optimization procedures or iterative processes. Within the discrete domain (that is when the estimation of the continuous signal is not necessary), then it is possible to guarantee orthogonality, thus totally avoiding noise amplification, which guarantee critical sampling and full recovery of the samples.

 ISBN: 978-84-614-6167-7

5. References

[1] R. NAVARRO, J. ARINES, AND R. RIVERA, *Direct and Inverse Discrete Zernike Transform*, Opt. Express **17,** (2009) 24269-24281.
[2] R. NAVARRO, J. ARINES, R. RIVERA, *Wavefront sensing with critical sampling*, Opt. Letters, **36** (2011) 433-435.

 ISBN: 978-84-614-6167-7

Proceedings of the 11th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE2011
Benidorm, Spain, 26-30 June 2011

Electronic Structure Computations in Molecular Architectures Based on Heteroborane Clusters

Josep M. Oliva[1]

[1] *Instituto de Química-Física "Rocasolano", Consejo Superior de Investigaciones Científicas, Serrano, 119, 28006 Madrid, Spain*

J.M.Oliva@iqfr.csic.es

Abstract

We present electronic structure computations on cyclic structures based on icosahedral carborane $CB_{11}H_{12}^{\bullet}$ radicals as building units, where the dot "$^{\bullet}$" represents an unpaired electron. The building units may be directly connected through covalent cage C-B bonds or using acetylene bridge units, resulting in tricyclic structures. The combination to a total spin S in the cyclic structures allows a corresponding range of magnetic modifications. We also report low-lying excited state energies using a broken-symmetry approach combining Hartree-Fock and Density Functional Theory.

Key words: Schrödinger equation, boron clusters, polyradical, broken-symmetry solution, computational chemistry, molecular architecture
MSC2000: 81V55, 81V70, 92E10

1. Introduction

The synthesis of stable radical structures – a molecule with one unpaired electron, total spin S = 1/2 – containing boron dates back to the 1920s [1], when the radical anion $[BPh_3]^{\bullet-}$ was detected. A few decades later, using electron spin resonance (ESR) experiments, it was shown that the unpaired electron is mainly located at boron [2]. One of the most interesting features of boron chemistry is the formation of very stable polyhedral clusters, such as the well-known icosahedral dianion *closo*-$B_{12}H_{12}^{2-}$ [3] (from now on simply $B_{12}H_{12}^{2-}$). Along the next five decades after its synthesis, we have witnessed an explosion of synthetic achievements using derivatives of this icosahedral cage leading to a very rich variety of crystal structures and molecular architectures. There are plenty of derivatives from

 ISBN: 978-84-614-6167-7

$B_{12}H_{12}^{2-}$ and a comprehensive literature citation is not possible here; thus, one can find derivatives by substitution of (i) cage atoms – heteroboranes – (ii) *exo* hydrogens atoms, (iii) addition of transition-metal atoms [4], and even (iv) the construction of linear finite 1D chains connected through covalent bonds [5,6].

Our case in point here is precisely the existence of stable radicals with formula $[closo\text{-}(C_nB_{12-n})Me_{12}]^{\bullet(n-1)}$, with $n = 0$ and $n = 1$ for icosahedral anionic borane [7] and neutral carborane [8] radicals, respectively. These radicals have been characterized, among other means, by X-ray crystallography. More recently, even two such carborane radicals (and their corresponding anions) have been connected through acetylene and ethylene bridges, leading to the dianion, radical anion and biradical species [9], as shown in Figure 1.

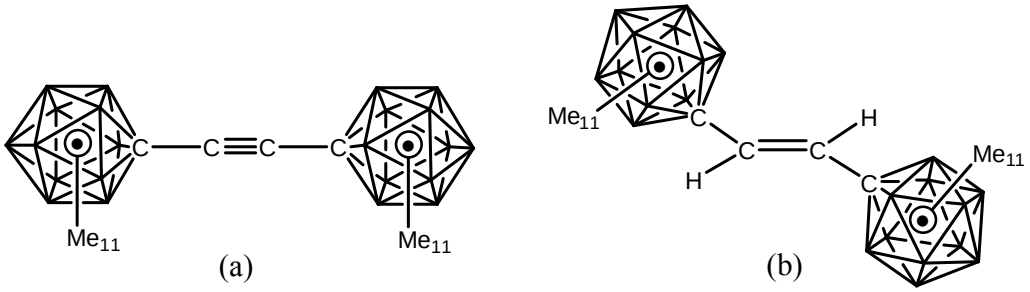


FIG. 1. Icosahedral carborane molecule $[CB_{11}Me_{12}]^{\bullet}$ connected through (a) acetylene and (b) ethylene bridges. The dot inside the cage refers to an unpaired electron. Each non-connected vertex corresponds to a B-H moiety.

In a recent work, the electronic structure of the dimer structure - dianion, radical anion and biradical - from Fig. 1a was studied using high-level quantum chemical computations [10], with methyl groups being substituted by hydrogen atoms. The conclusion was that the groundstate biradical is of singlet nature – total spin S = 0 – with the triplet state – total spin S = 1 – only (adiabatically) 0.005 eV higher in energy, very close to $k_B \cdot T$ at room temperature. This energy difference lies within the microwave region of the electromagnetic spectrum and therefore, one could in principle populate the triplet state using photons of this energy. Extension to longer polyradical 1D chains has also been studied recently, using the same building units [11]. For the particular case of the trimer triradical shown in Figure 2, it was found that the high-spin state – S = 3/2 – has lower energy than the low-spin state – S = 1/2 – with an energy difference of $\Delta E \sim 0.013$ eV, within the far-IR region of the electromagnetic spectrum. Care should be taken for the energy difference for longer polyradical chains since the size of the system prevents from performing high-level quantum chemical *ab initio* computations, such as CASPT2 – second order perturbation theory applied to a multiconfigurational wave function. For instance, in the trimer structure from Figure 2, we used the broken-

 ISBN: 978-84-614-6167-7

symmetry (BS) method with an unrestricted hybrid Hartree-Fock / Density functional Theory model chemistry.

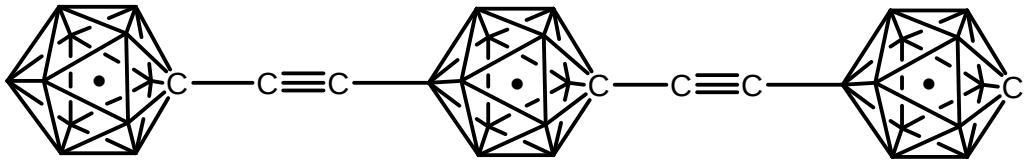


FIG. 2. Structure of one of the four trimer triradicals that can be built using two orientations of the carborane cage – see Fig. 5 from Ref [11] – with the carbon atom on the right (A) or left (B) of the cage. This trimer corresponds to A-CC-A-CC-A. The other three trimers are A-CC-B-CC-A, A-CC-B-CC-B, and B-CC-B-CC-A. Each non-connected vertex corresponds to a B-H moiety (C-H moiety for the non-connected vertex labeled with "C" on the top most right). The dot inside the cage refers to an unpaired electron.

DFT and hybrid Hartree-Fock/DFT methodologies fail in the proper description of singlet biradicals, such as those derived from Figure 1. We previously showed in carborane monomers biradicals how a standard DFT representation is approximately valid in computing triplet states but is totally inadequate for spin contaminated singlet biradicals [12]. Therefore, for low-lying spin states one usually turns to so-called BS DFT or spin correction for DFT computations [13,14], calibrated with high-level CASPT2 computations [15]. The basis of this method is described in Refs. [13,14] and in Section 2 below.

2. Three-fold cyclic structures derived from $CB_{11}H_{12}^{\bullet}$

We now turn to cyclic structures with one unpaired electron per cage, using as building unit the carborane radical $CB_{11}H_{12}^{\bullet}$. The simplest such structure is a triangle of $CB_{11}H_{12}^{\bullet}$ radicals connected directly through C-B covalent bonds. Given the fact that experimentally, one needs bridge units for connecting the carborane cages [9], we will also take into account the structure with acetylene bridges. Both such triangular structures are displayed in Figure 3. There are many different ways of connecting the $CB_{11}H_{12}^{\bullet}$ radical to form a triangle; we have chosen the most symmetrical one for simplicity reasons, with a $\hat{C}_3$ axis perpendicular to both structures – Figure 3. The dot inside each cage represents again an unpaired electron.

 ISBN: 978-84-614-6167-7

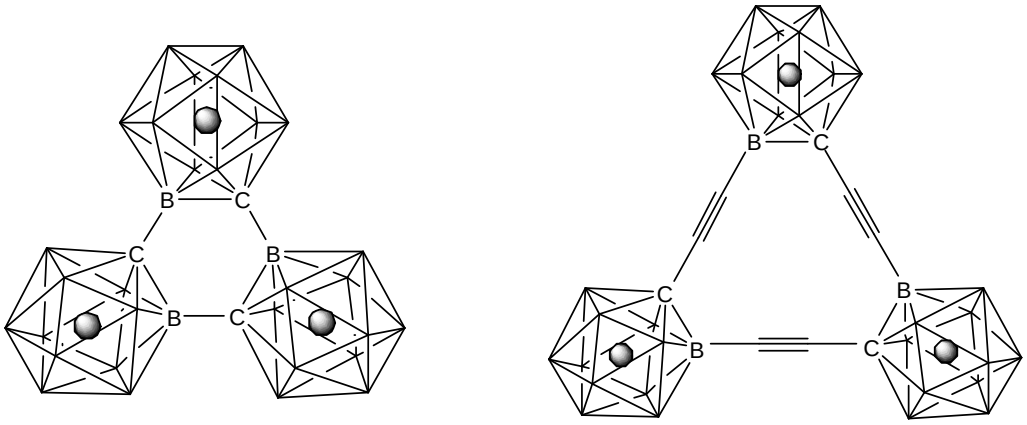


FIG. 3. Three-fold cyclic structures built by joining three $CB_{11}H_{12}^{\bullet}$ radicals with (a) direct C-B covalent bond → (**1**), and (b) direct C-B bond and an acetylene bridge → (**2**). The small circle inside every cage represents an unpaired electron.

The three unpaired electrons from the cyclic structures of Figure 3 can couple to a high-spin state with total spin S = 3/2 or to a low-spin state with S = ½. Again, DFT methods can reproduce high-spin state energies fairly well. However, for the low-spin state we should rely again onto BS DFT methods. As in the case of the linear structure from Figure 2, we used the BS DFT method to determine the high-spin and low-spin states of the three-fold cyclic structures, without and with acetylene bridge connections. The results are gathered in Table 1. For simplicity the three-fold cyclic structureswithout and with acetylene bridges are labelled as (**1**) and (**2**) respectively. The energy difference between the quartet and doublet states – ΔE_{DQ} – is calculated using the spin-projected formula within the broken-symmetry formalism [11]:

$$\Delta E_{DQ} = E_{1/2} - E_{3/2} = \frac{E_{unr,S=1/2} - E_{3/2}}{1-b^2} = \frac{12\cdot(E_{unr,S=1/2} - E_{3/2})}{15 - 4\cdot\left\langle \hat{S}^2 \right\rangle_{unr,S=1/2}}. \qquad (1)$$

This equation is determined as follows. Let us suppose that we have a three-electron wave function which has a doublet and a quartet component:

 ISBN: 978-84-614-6167-7

$$\Psi_{unr,S=1/2} = a\Psi_{1/2} + b\Psi_{3/2}, \quad (2)$$

with $a^2 + b^2 = 1$. Now, the expectation value of the square of spin for this wave function is

$$\left\langle \Psi_{unr,S=1/2} \left| \hat{S}^2 \right| \Psi_{unr,S=1/2} \right\rangle = \frac{3}{4}a^2 + \frac{15}{4}b^2, \quad (3)$$

and the corresponding energy expectation value

$$E_{unr,S} = \left\langle \Psi_{unr,S=1/2} \left| \hat{H} \right| \Psi_{unr,S=1/2} \right\rangle = a^2 \left\langle \Psi_{1/2} \left| \hat{H} \right| \Psi_{1/2} \right\rangle + b^2 \left\langle \Psi_{3/2} \left| \hat{H} \right| \Psi_{3/2} \right\rangle. \quad (4)$$

We know that DFT reproduces fairly well the energy of high-spin states, in this case $\Psi_{3/2}$, and we can calculate the expectation value of the square of spin for the wave function from Eq.(2), $\Psi_{unr,S=1/2}$. Therefore, by using the BS formalism, Eq.(1) follows immediately by combining Eq.(3) and Eq.(4).

TABLE 1. Energies (in atomic units), $<\hat{S}^2>$ expectation values, and energy differences ΔE_{DQ} (eV) – Equation (1) – between high-spin (quartet) and low-spin (doublet) states for the three-fold cyclic structures (**1**) and (**2**) - Figure 3a and Figure 3b respectively. We used the UB3LYP/6-31G* method for high-spin states and broken-symmetry UB3LYP/6-31G* method for the low-spin states. See Refs. [11, +elíseo] for more details. BS = broken-symmetry. ΔE computed with Eq.(1).

Structure, Spin	Energy (au)	$<\hat{S}^2>$	ΔE (eV)
(**1**), S = 3/2	-952.829852	3.7612	---
(**1**), BS	-952.829762	1.7555	0.0037
Structure, Spin	Energy (au)	$<\hat{S}^2>$	ΔE (eV)
(**2**), S = 3/2	-1181.351576	3.7581	---
(**2**), BS	-1181.350845	1.7493	0.0298

Table 1 gathers the energy difference between high-spin and low-spin states using the BS approximation, Eq.(1). The geometries for every system were optimized by using analytical gradients of the energy versus nuclear displacements. Spin densities for the different spin states of trimers (**1**) and (**2**) are displayed in Figure 4 and Figure 5 respectively.

 ISBN: 978-84-614-6167-7

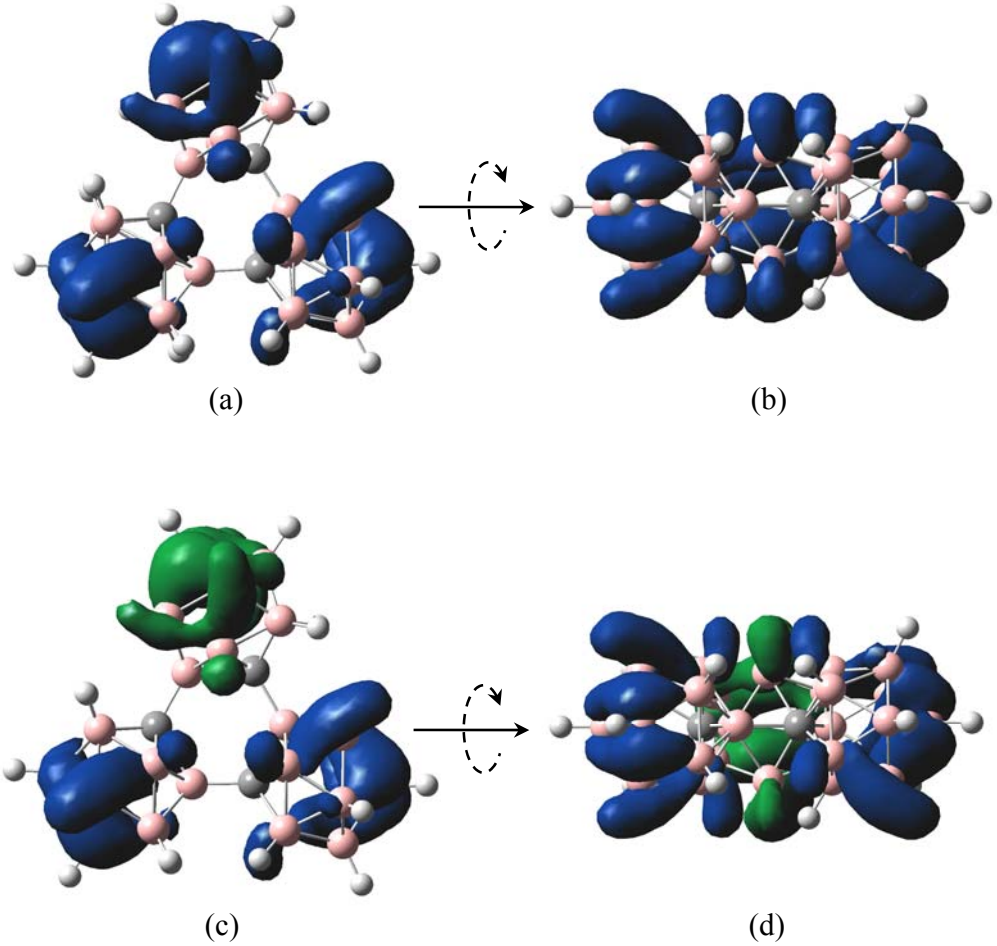


FIG. 4. Spin density for trimer (**1**). High-spin state with S = 3/2: (a) top view, (b) side view. Low-spin state with S = ½ using a broken-symmetry solution: (c) top view, (d) side view. Spin density isovalue $|\rho_s|$ = 0.002. UB3LYP/6-31G* computations.

 ISBN: 978-84-614-6167-7

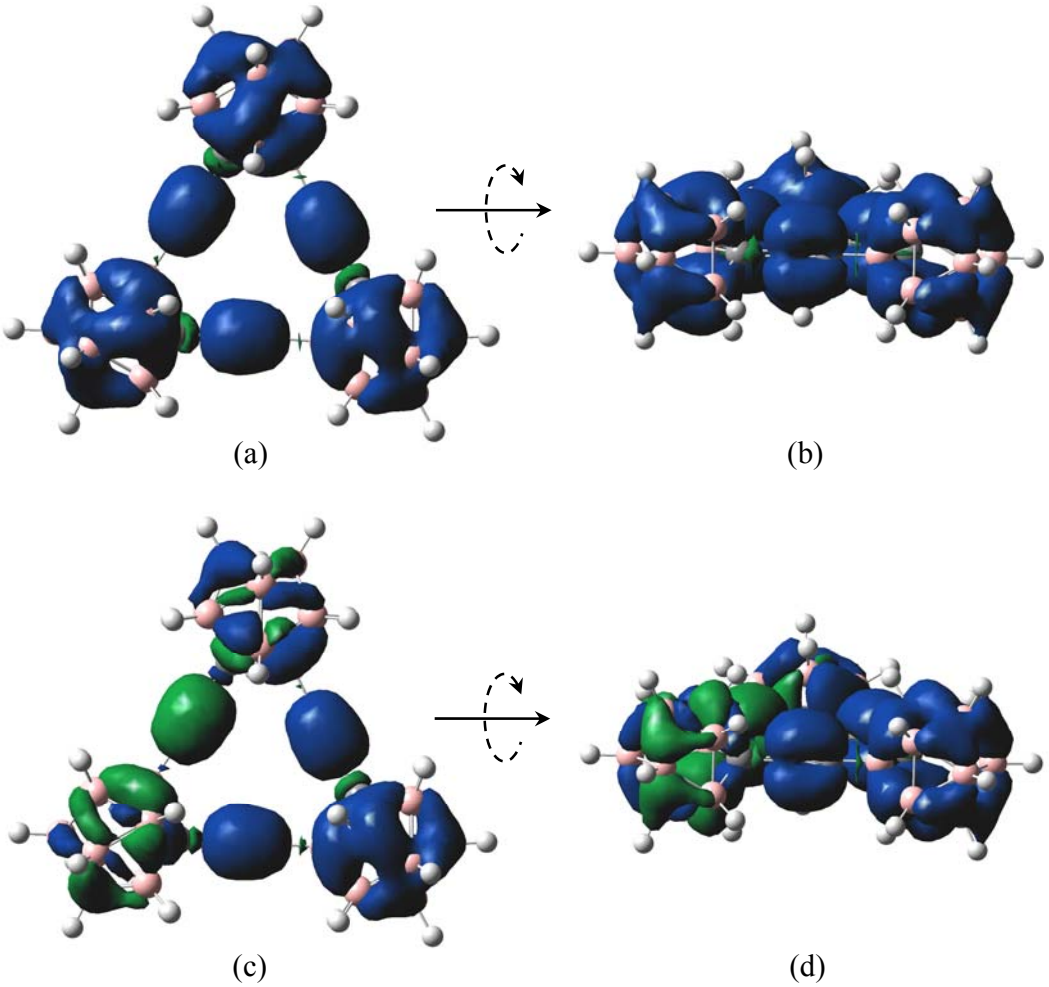


FIG. 5. Spin density for trimer (**2**). High-spin state with S = 3/2: (a) top view, (b) side view showing the π spin density along C≡C triple bonds. Low-spin state with S = ½ using a broken-symmetry solution: (c) top view, (d) side view showing the π spin density along C≡C triple bonds. Spin density isovalue $|\rho_s|$ = 0.001. UB3LYP/6-31G* computations.

 ISBN: 978-84-614-6167-7

As shown in Figure 4, the spin density is mainly located in the carborane cages. For the low-spin state, two cages display a positive spin density and the remaining cage a negative spin density as it should be for an S = ½ system and three electrons. When the cages are connected with an acetylene bridge unit {–C≡C–} - Figure 5 - the low-spin state is described as positive spin density in one cage and mixture of positive and negative spin densities in the other two cages. The two connecting {–C≡C–} units to the cage with positive spin density display also positive π spin density; the ramining bridge unit connecting the other two cages has negative π spin density. As for the high-spin state, the picture is very similar with all regions having positive spin-density with small regions of negative spin density near the carbon atoms.

The energy differences gathered in Table 1 give an idea of the photon energies that would be necessary in order to populate low-spin states, corresponding to the far-IR region of the electromagnetic spectrum.

3. Conclusions

When the number of "active" electrons, namely, electrons taken into account explicitly in the electronic structure calculations, exceeds one dozen, the low-lying excited states, particularly low-spin states in polyradicals, must be approximated. Post Hartree-Fock methodologies, like CASPT2 [15] and methods within Valence-Bond (VB) theory have restrictions [16] due to computational demand. One then has to rely to approaches such as broken-symmetry methods, based on hybrid Hartree-Fock / DFT functionals [14]. The question then is to what extent the later methodology would allow for an estimation of an energy gap of an infinite 1D chain – Figure 2 – or cycle – Figure 3. Energy spectra in a 1D Heisenberg chain may be predictable with the Bethe ansatz [17]. Another question is whether the cycles from Figure 3 can be synthesized. Indeed, as mentioned above, 1D finite chains have been synthesized, using the $CB_{11}(Me)_{11}$ unit as building block, and connecting it through acetylene and ethylene bridges [9]. The main goal of this research line is the prediction of energy spectra for real systems derived from radical units $CB_{11}H_{12}^{\bullet}$ connected in different dimensions. This radical unit can be considered as a model system for a one-electron site in the lattice models that physicists use for derivation of model Hamiltonians [18,19]. A few steps fowards have been taken for the particular case of heteroborane cages and more effort is needed. Density-matrix renormalization group (DMRG) methods [20] are promising tools for tackling the challenging problems that we have put foward here in the particular case of polyradical architectures.

 ISBN: 978-84-614-6167-7

The author is indebted to Professor Eliseo Ruiz (Barcelona), Professor Ignacio Cirac and Michael Lubasch (Germany) for illuminating discussions. This work was supported by the Dirección General de Universidades e Investigación de la Comunidad de Madrid under Grant No. S2009/ESP/1691 and Program MODELICO-CM and by the European Union through projects FP7-ICT-2009-4-248909-LIMA and FP7-ICT-2009-4-248855-N4E.

4. References

[1] E. KRAUSE AND H. POLACK, *Reindarstellung des "Triphenylboryl-natriums" und der Verbindungen des Bortriphenyls mit den übrigen Alkalimetallen*, Ber. Dtsch. Chem. Ges. **59** (1926) 777-785.

[2] J. E. LEFFLER, G. B. WATTS, T. TANIGAKI, E. DOLAN AND D. S. MILLER, *Triarylboron anion radicals and the reductive cleavage of boron compounds*, J. Am. Chem. Soc. **92** (1970) 6825-6830.

[3] M. F. HAWTHORNE AND A. R. PITOCHELLI, *The isolation of the icosahedral $B_{12}H_{12}^{2-}$ ion*, J. Am. Chem. Soc. **82** (1960) 3228-3229.

[4] I. B. SIVAEV, V. I. BREGADZE AND S. SJÖBERG, *Chemistry of closo-dodecaborate anion $[B_{12}H_{12}]^{2-}$: A review*, Collect. Czech. Chem. Commun. **67** (2002) 679-727.

[5] X. YANG, W. JIANG, C. B. KNOBLER AND M. F. HAWTHORNE, *Rigid-rod molecules: carborods. Synthesis of tetrameric p-carboranes and the crystal structure of bis(tri-n-butylsilyl)tetra-p-carborane*, J. Am. Chem. Soc. **114** (1992) 9719–9721.

[6] J. MÜLLER, K. BAŠE, T. F. MAGNERA AND J. MICHL, *Rigid-rod oligo-p-carboranes for molecular tinkertoys. An inorganic Langmuir-Blodgett film with a functionalized outer surface*,

[7] T. PEYMANN, C. B. KNOBLER AND M. F. HAWTHORNE, *An unpaired electron incarcerated within an icosahedral borane cage: Synthesis and crystal structure of the blue, air-stable [closo-$B_{12}(CH_3)_{12}$] $^{\bullet-}$ radical*, Chem. Commun. (1999) 2039-2040.

[8] B.T. KING, B.C. NOLL, A.J. MCKINLEY AND J. MICHL, *Dodecamethylcarba-closo-dodecaboranyl $[CB_{11}Me_{12}]^{\bullet}$, a stable free radical*, J. Am. Chem. Soc. **118** (1996) 10902-10903.

[9] L. ERIKSSON, K. VYAKARANAM, J. LUDVÍK AND J. MICHL, J. Org. Chem. **72** (2007) 2351-2356.

[10] J. M. OLIVA, L. SERRANO-ANDRÉS, Z. HAVLAS AND J. MICHL, *On the electronic structure of a dianion, a radical anion, and a neutral biradical $(HB)_{11}C\text{-}C{\equiv}C(BH)_{11}$ carborane dimer*, J. Mol Struct. : THEOCHEM **912** (2009) 13-20.

 ISBN: 978-84-614-6167-7

[11] J. M. OLIVA, *Energy landscapes in boron chemistry: bottom-top approach towards design of novel molecular architectures*, Adv. Quant. Chem., accepted for publication.

[12] L. SERRANO-ANDRÉS, D. J. KLEIN, P. v. R. SCHLEYER AND J. M. OLIVA, *What electronic structures and geometries of carborane mono- and ortho-, meta-, and para-diradicals are preferred?*, J. Chem. Theory Comput. **4** (2008) 1338-1347.

[13] A. A. OVCHINNIKOV AND J. K. LABANOWSKI, *Simple spin correction of unrestricted density-functional calculation*, Phys. Rev. A **53** (1996) 3946-3952.

[14] E. RUIZ, A. RODRIGUEZ-FORTEA, J. CANO, S. ALVAREZ AND P. ALEMANY, *About the calculation of exchange coupling constants in polynuclear transition metal complexes*, J. Comp. Chem. **24** (2003) 982-989.

[15] B.O. ROOS, K. ANDERSSON, M.P. FÜLSCHER, P.-Å. MALMQVIST, L. SERRANO-ANDRÉS, K. PIERLOOT AND M. MERCHÁN, *Multiconfigurational perturbation theory: Applications in electronic spectroscopy*, Adv. Chem. Phys. **93** (1996) 219-331.

[16] T. THORSTEINSSON AND D.L. COOPER, *An overview of the CASVB approach to modern valence bond calculations*, Quantum Systems in Chemistry and Physics, Volume 1: Basic problems and models systems ed. A. HERNÁNDEZ-LAGUNA, J. MARUANI, R. MCWEENY AND S. WILSON Kluwer, Dordrecht, 2000. pp 303-326.

[17] H. BETHE, *Zur Theorie der Metalle. I. Eigenwerte und Eigenfunktionen der linearen Atomkette*, Zeitschrift für Physik A **71** (1931) 205-226.

[18] J. HUBBARD, Proc. R. Soc. London A, *Electron correlations in narrow energy bands*, **276** (1963) 238-257; J. HUBBARD, Proc. R. Soc. London A, *Electron correlations in narrow energy bands. III. An improved solution* **281** (1964) 401-419.

[19] F. MEZZACAPO AND J.I. CIRAC, *Ground-state properties of the spin-1/2 antiferromagnetic Heisenberg model on the triangular lattice: A variational study based on entangled-plaquette states*, New J. Phys. 12 (2010) 103039, 1-8.

[20] G. CHAN AND S. SHARMA, *The density matrix renormalization group in quantum chemistry*, Annu. Rev. Chem. **62** (2011) 465-481.

 ISBN: 978-84-614-6167-7

Proceedings of the 11th International Conference on Computational and Mathematical Methods in Science and Engineering, CMMSE2011 Benidorm, Spain, 26-30 June 2011

Comparison of different classification algorithms for terrestrial laser scanner segmentation

C. Ordóñez[1], J. Martínez[2], F.J. de Cos[3], F. Sánchez-Lasheras[4]

[1,3] *Department of Mining Exploitation, University of Oviedo (Spain)*

[2] *Centro Universitario de la Defensa, Academia General Militar. Zaragoza (Spain)*

[4] *Research Department Techniproject Ltd, Oviedo (Spain)*

emails: ordonezcelestino@uniovi.es, jmtorres@unizar.es, fjcos@uniovi,es, fsancheztecniproject.com

Abstract

This article compares different classification techniques for distinguishing between construction materials using information captured by a terrestrial laser scanner from an historical building. This information consists of data on colour and reflected signal intensity.

Of the three techniques used (k-means, classification trees and multilayer perceptron neural networks), the classification trees produced the best results when both classification capacity and ease of interpretation were considered.

Key words: terrestrial laser scanner, segmentation algorithms, signal intensity

1. Introduction

Terrestrial laser scanners (TLS) are frequently used to geometrically document historical buildings. These devices rapidly capture clouds of points that are defined by coordinates in a reference system. They also provide information on colour and on the intensity of the wave reflected at each point. This kind of information can be used to automatically detect data on specific elements in the measured object, in a mathematical process called segmentation [2], [4].

Our aim was to test and compare different classification techniques, both supervised and non-supervised, used to segment a point cloud obtained using a TLS. Also studied was the discriminatory capacity of colour versus reflected signal intensity.

 ISBN: 978-84-614-6167-7

2. Materials and methods

2.1 Data collection

The equipment used for this segmentation problem was a Riegl LMS-Z390i 3-D time-of-flight laser scanner. This scanner emits an infrared pulse and measures distances in a range of 1.5 to 400 m, with a nominal accuracy of 6 mm to 50 m in normal conditions.

It also provides, for each point in the measured object, colour values measured according to the RGB (red, green and blue) scale and values for reflected wave intensity. For our research, these values were captured and recorded for the facade of an historic building with five types of materials: stone, metal, glass, cement and vegetation.

2.2 Classification algorithms

2.2.1 Cluster analysis

The k-means algorithm [7], one of the simplest non-supervised algorithms for resolving a clustering problem, is based on an iterative partitioning process. It divides a collection of n vectors into c groups (clusters G_i , $i=1,...,c$) and locates centroids for each cluster by minimizing an inequality (or distance) function formulated as follows [1]:

$$J = \sum_{i=1}^{c} \sum_{\mathbf{x}_j \in G_i} d^2(\mathbf{x}_j - \mathbf{k}_i)$$

where $\mathbf{k}_i$ is the cluster centroid i; $d(\mathbf{x}_j - \mathbf{k}_i)$ is the distance between the i-th centroid ($\mathbf{k}_i$) and the j-th element in the dataset.

2.2.2 Classification trees

Decision trees are one of the most widely used supervised learning models, are they are simple classification models that are graphic and easily understood.

Decision trees are constructed recursively using an induction process and following a top-down strategy that starts with general concepts and ends with particular examples. This is known as Top-Down Induction on Decision Trees (TDIDT) [8].

A crucial issue is dividing each classification tree node, done by defining a function [5] that compares the heterogeneity or impurity of the parent node with the sum of the impurities of the offspring nodes generated by means of a division D:

$$\varphi(D,t) = I(t) - \sum_{r=1}^{l} I(t_r)\, p_r$$

 ISBN: 978-84-614-6167-7

where $I(t)$ represents the impurity measure for the node t, l the number of offspring for the same node and p_r the proportion for the node majority class r.

2.2.3 Multilayer perceptron neural networks

Neural network models were developed from biological neural models in the 1940s, based on work by a psychiatrist and a mathematician, McCulloch and Pitts, respectively. [6].

In the classification problem, the network implements a function $\mathbf{f} : \mathbf{X} \subset \mathrm{R}^d \rightarrow \mathbf{Y} \subset \mathrm{R}^c$, where c represents the number of classes. The function implemented by the multilayer perceptron (MLP) is as follows:

$$f(\mathbf{x}, \boldsymbol{\theta}) = \sum_{j=1}^{h} c_j \psi(\mathbf{w}_j^T \mathbf{x} + w_{j0}) + c_0$$

where $\psi_j(\mathbf{x}) = \psi(\mathbf{w}_j^T \mathbf{x} + \mathbf{w}_{j0})$ is the activation function for the hidden layer units, $\mathbf{w}_j \in \mathrm{R}^d$ is the vector of parameters for the units and $w_{j0} \in \mathrm{R}$ is the threshold value. The function ψ may be sigmoid, logistic or hyperbolic tangent. The algorithm typically used to train the MLP is the back-propagation algorithm [3].

3. Results

The data captured by the TLS formed a sample $\mathbf{x}_i, y_i \ _{i=1}^{n}$ with $n = 4775$, with $\mathbf{x}_i$ as the vector of RGB values and I (reflected intensity) and with y_i as the value (1 to 5) representing the type of material.

The fit for each technique was measured using the error rate:

$$TE = \frac{1}{m} \sum_{i=1}^{m} 1_{\ y_i \neq g(\mathbf{x}_i)}$$

where m represents the size of the test set (in this case, 10% of the entire sample) and where $g(\mathbf{x}_i)$ is the value of the class obtained for each vector $\mathbf{x}_i$ of data.

Table 1 shows the error rates for each method, including or excluding reflected signal intensity. The classification trees, as well as having a low error rate, have the advantage that they show the class definition process as a graph.

 ISBN: 978-84-614-6167-7

Table 1. **Error rates (ER) for each technique including and excluding reflected signal intensity (RSI).**

Technique	ER including RSI	ER excluding RSI
k-MEANS	32%	28%
CART	2%	1%
MLP	1%	0.5%

4. Conclusions

Of the classification techniques used to segment the point cloud captured by the TLS, the non-supervised technique shows a high error rate. Of the supervised techniques, the error rate for the MLP was half that obtained with the classification trees. The latter also have the advantage that they allow graphs to be constructed that facilitate interpretation of the classification criteria used.

Including the reflected signal intensity reduces error by 50% for the supervised techniques, indicating that this variable is important for segmentation purposes.

References

[1] S. ALBAYRAK AND F. AMASYALI, *Fuzzy c-means clustering on medical diagnostic systems*, International XII. Turkish Symposium on Artificial Intelligence and Neural Networks, 2003, Turkey.

[2] J. ARMESTO, B. RIVEIRO-RODRÍGUEZ, D. GONZÁLEZ-AGUILERA, M.T. RIVAS-BREA, *Terrestrial laser scanning intensity data applied to damage detection for historical Buildings*, Journal of Archaeological Science **37** (2010) 3037-3047.

[3] R. BATTITI, *First and second-order methods for learning: between steepest descent and Newton´s method*. Neural computation, **4** (1992) 141-166.

[4] J. M. BIOSCA AND J.L. LERMA, *Unsupervised robust planar segmentation of terrestrial laser scanner point clouds based on fuzzy clustering methods,* Photogrammetry and Remote Sensing **63** (2008), 84-98.

[5] V. CHERKASSKY AND F. MULIER, *Learning from data: concepts, theory and method,* John Wiley & Sons, Inc, 1998.

[6] W. S. MACCULLOCH AND W.S. PITTS, *A logical calculus of the ideas inmanent in nervous activity*, Bulletin of Mathematical Biophysics **5** (1943), 115-133.

[7] J.B. MACQUEEN, *Some Methods for Classification and Analysis of Multivariate Observations*, Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, University of California Press **1** (1967), 281-297.

[8] J.R. QUINLAN, *Induction on Decision Trees*. Machine Learning **1** (1986), 81-106.

 ISBN: 978-84-614-6167-7

Proceedings of the 11th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE2011
Benidorm, Spain, 26-30 June 2011

Computational Fluid Dynamics in Root Canal Procedures

M. Patrício[1], J. M. Santos[2], P. Oliveira[3] and F. Patrício[4]

[1] *CMUC - School of Technology and Management of the Polytechnic Institute of Leiria*
[2] *Dentistry, Faculty of Medicine, University of Coimbra*
[3] *CMUC - Department of Mathematics, University of Coimbra*
[4] *CMUC*

email: miguel.patricio@estg.ipleiria.pt, jmmdossantos@gmail.com, poliveir@mat.uc.pt, mfsp@mat.uc.pt

Abstract

Success of root canal treatment is dependent upon the effective removal of microorganisms and their by-products. This is usually achieved by mechanical instrumentation and irrigation of the canal with an antiseptic solution. However, the efficacy of this method is difficult to measure, due to the small scales involved. The computational simulation of irrigation provides useful tools for the study of this phenomenon.

In this work we resort to computational fluid dynamics to determine the velocity of the irrigant over the root canal, for different needle tip shapes. A stagnation curve will be computed for each case. An algorithm is proposed to evaluate particle removal from the canal.
Key words: computational fluid dynamics, flow rate, irrigation, needle.

1. Introduction

The aim of endodontic and restorative dentistry is the conservation of natural tooth structure. In the past, it was quite usual to extract teeth with extensive structural and biologic compromise. Nowadays, with the scientific progression in the field of dentistry and increasing conservative demands by the population, it is sought to preserve severely destroyed and previously infected teeth. One common course of treatment to accomplish this purpose is root canal therapy, which consists in removing the pulp tissue, debris and microorganisms harboured inside

 ISBN: 978-84-614-6167-7

the canal space. Operative protocol for this procedure involves coronal access to the pulp chamber followed by removal of the damaged pulp tissue, shaping the canal with mechanical instruments and cleaning with irrigating solutions. As a consequence of the action of shaping instruments in the canal walls, dentin chips are produced and accumulated along the root canal, mainly in the apical area. To achieve success these debris must be removed, together with tissue remnants and microorganisms. Tooth disinfection and the flushing of the debris are usually accomplished by irrigating the canal with an antiseptic solution. However, it has been indicated that standard procedure is not effective in removing all dentin chips and bacteria, compromising the objectives of the treatment and jeopardizing the outcome, cf. [5]. Moreover, the efficacy of the procedure is paramount for the long-term success of the treatment and depends upon numerous factors such as the system of delivery, the geometry of the canal, the size, position and shape of the needle or the flow velocity, cf. for example [1, 6]. In particular, the flow velocity should be weighted carefully, as high velocities may cause the irrigant to extrude towards the periapical area, which may lead to tissue damage. At the same time it should be high enough to promote debris and bacteria removal.

In this paper we aim to look at how different needle geometries influence irrigant flow on a root canal, as well as assert how much of the bacteria present in the canal are successfully flushed out. It is difficult to perform *in vivo* measurements in root canals, due to their microscopic size. An alternative approach, which we adopt, is provided by computational fluid dynamics (CFD), cf. [2, 3] for example. This allows for the numerical simulation of irrigation in root canals in real physical conditions. On the other hand, it provides sufficient information for tracking the bacteria by means of an algorithm that we propose.

Section 2 is devoted to properly formulating the setting of the problem and establishes the related mathematical model. The latter is expressed by the Navier-Stokes equations, defined over an appropriate domain and completed with boundary conditions. Both the geometries of the root canal and needle are included in the model. Different needle types will be considered, allowing for a discussion of their merits.

In section 3 it is discussed how CFD may be employed in order to compute the fluid velocity field in a root canal. In particular, the existence of a stagnation curve for each needle is addressed. The motion of particles of negligible weight is tracked, as a function of time. The flushing region of the computational domain is determined for each needle. This corresponds to the portion of the domain where the irrigation procedure ensures that the particles initially present in the root canal are successfully flushed out.

Finally a brief conclusion is drawn up in section 4. The performances of the needles are compared and future work is proposed.

 ISBN: 978-84-614-6167-7

2. Problem formulation

Several models that represent root canals may be found in the literature. These models differ in shape, dimensions or other features such as the possibility of inclusion of an apical region, cf. for example [1, 3, 4]. Also, there is a discussion on the types of needles that are used for fluid delivery, see [2].

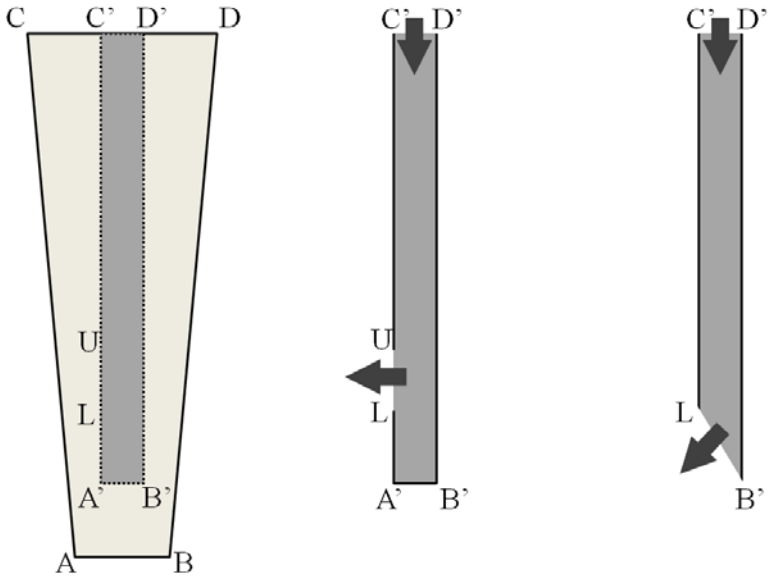


Figure 1

a) Root canal. The darker shaded rectangle represents the region where the needle is to be inserted; b) CE-SSO needle (the arrows represent the direction of the flow); c) BOE needle.

In this paper, it is assumed that the root canal is shaped such that its 2D section is the outer polygon with endpoints **A**, **B**, **C**, **D** represented in Figure 1a). The existence of an apical region is neglected. Besides the root canal, it is important to model the needle which will deliver the irrigation fluid. Two problems are considered in the present work: the simulation of flows associated with a close-end single side-opening needle (CE-SSO) and with a bevelled open-ended needle (BOE). The first of these needles is represented in Figure 1b) by a polygon of endpoints **A'**, **B'**, **C'**, **D'**. The line segment **LU** is its side opening. As for the BOE needle, the endpoints of the respective polygon are **L**, **B'**, **C'**, **D'**, see Figure 1c). The segment **LB'** is open. The horizontal and vertical coordinates for each point $\mathbf{Q}=(Q_x, Q_y)$ in the model geometries are displayed in Table 1.

 ISBN: 978-84-614-6167-7

Table 1: Relevant coordinates in model geometries.

Point Q	A	B	C	D	A'	B'	C'	D'	L	U
Q_x	-1.61E-4	1.61E-4	-7,85E-4	7.85E-4	-1.61E-4	1.61E-4	-1.61E-4	1.61E-4	-1.61E-4	-1.61E-4
Q_y	0	0	1.8E-2	1.8E-2	3E-3	3E-3	1.8E-2	1.8E-2	4E-3	5E-3

In both problems we are considering that the fluid is delivered by the corresponding needle into the root canal such that at the inlet, which is the line segment C'D', the vertical component of the velocity is a linear function of time: $u_y(x, y, t) = -u_0 t$, t in $[t_0, t_f]$. Both line segments CC' and D'D are outlets through which the fluid is flushed out. The pressure is prescribed to be equal to p_0 at these outlets. The remaining boundaries are assumed to be solid walls at which a no-slip condition is imposed. We assume that initially the root canal is filled with the injection fluid. In our models, the walls of the needles have small thickness and are not part of the computational domains Ω_1 and Ω_2, related to the CE-SSO needle and the BOE needle respectively.

Mathematically, the physical behaviour of the fluid may be modelled by the Navier-Stokes equations for incompressible flow. These equations relate the fluid velocity vector field $\mathbf{u}=(u_x, u_y)$ and the pressure p over the computational domains Ω_i:

$$\rho \frac{\partial \mathbf{u}}{\partial t} + \rho(\mathbf{u}\cdot\nabla)\mathbf{u} = \nabla\cdot[-\rho\mathbf{I} + \mu(\nabla\mathbf{u} + (\nabla\mathbf{u})^{T})] + \mathbf{F}, \tag{2.1}$$

$$\nabla\cdot\mathbf{u} = \mathbf{0}, \tag{2.2}$$

for i=1, 2, the latter equation being the continuity equation. Here, **F**, μ and ρ represent the volume force, the dynamic viscosity and the density, respectively. Equations (2.1)-(2.2) are completed with appropriate boundary conditions:

$$u_x(x, y, t)=0;\ u_y(x, y, t)= -u_0 t, \qquad \text{for } (x, y) \in \Gamma_i,\ t \in [t_0, t_f], \tag{2.3}$$

$$u_x(x, y, t)=0;\ u_y(x, y, t)=0, \qquad \text{for } (x, y) \in \Gamma_w,\ t \in [t_0, t_f], \tag{2.4}$$

$$p(x, y, t)= p_0, \qquad \text{for } (x, y) \in \Gamma_o,\ t \in [t_0, t_f]. \tag{2.5}$$

Here, Γ_i, Γ_w and Γ_o represent the inlet, wall and outlet portions of the boundary of the domain, respectively. Moreover, the velocity is set to be equal to zero in the initial time instant:

$$u_x(x, y, 0)=0;\ u_y(x, y, 0)=0, \qquad \text{for } (x, y) \in \Omega_i, \tag{2.6}$$

The values used for the parameters included in the models are displayed in Table 2.

 ISBN: 978-84-614-6167-7

Table 2: Relevant parameters.

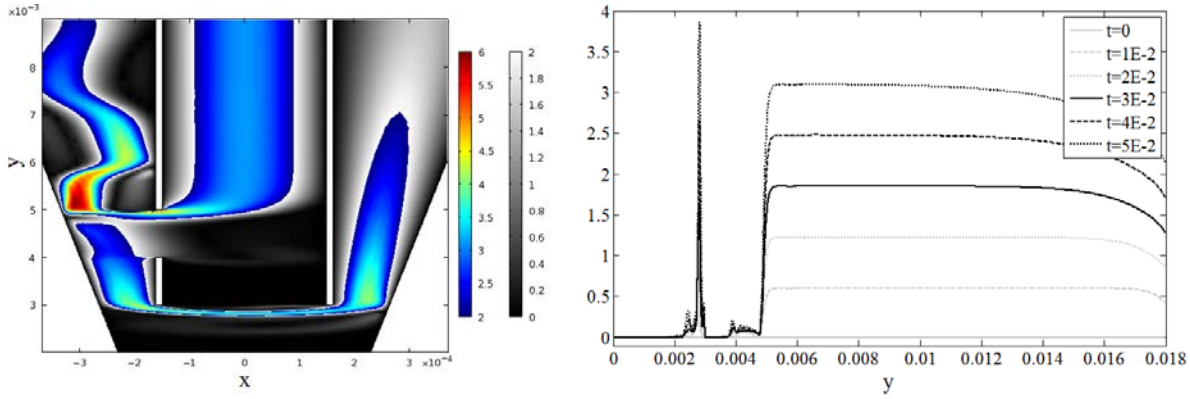
Parameter	t_0	t_f	u_0	η	ρ	p_0
Value	0	0.05	43	0.001	998	0

3. Fluid velocity and stagnation curve

The velocity that an irrigant will attain along the root canal is critical to the cleaning efficacy of the irrigation protocol and this is determinant for the success of the endodontic treatment. In this section we aim to compute the velocity vector fields for the problems associated with the CE-SSO needle and the BOE needle.

The values for both the horizontal and vertical components of the velocity are not easy to predict or measure in real physical situations. For each of our models it is assumed that equations (2.1) - (2.6) accurately describe the behaviour of the irrigant fluid. As a closed form solution for these equations is rather difficult to obtain we resort to a CFD approach. This involves the discretisation of the computational domain in a fine mesh, allowing for the retrieval of a numerical approximation to the solution of the problem. Here we take, for the case of each needle, a triangular mesh with maximum element size 4E-5, such that about 1E6 elements were considered. For that purpose, linear finite elements in space are used, coupled with the M.O.L. approach. The time dependent solver BDF – variable step variable methods in the time domain $t \in [t_0, t_f]$ with relative tolerance 1E-2 and absolute tolerance 1E-3 is employed.

Figure 2a) depicts the velocity field magnitude for $t=t_f$, over a portion of the root canal, when the needle is the CE-SSO. Also, the velocity field magnitude along $x=0$ is plotted in Figure 2b) as a function of y and for several time instants. Likewise, corresponding results for the BOE needle are displayed in Figure 3.



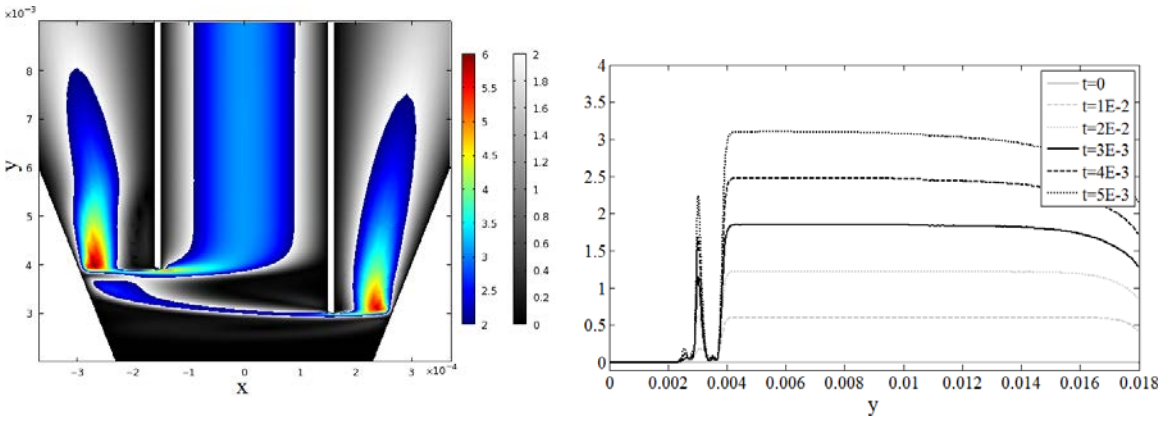
a) b)

Figure 2 CE-SSO needle:
a) Magnitude of the velocity field for $t=t_f$; b) Magnitude of the velocity field, along $x=0$, computed at different time instants.

 ISBN: 978-84-614-6167-7

It can be observed in Figure 2a) that the magnitude of the velocity of the irrigant is larger in a region near the opening of the needle, as expected. The fluid velocity attains a high magnitude along the corresponding side of the root canal. Note the fluid circulation below the needle and onto the opposite side of the root canal which is displayed in the figure. This is also important to ensure disinfection and debriment throughout.

As the velocity of the fluid at the inlet grows with time, the velocity magnitude along the root canal and the needle also grows globally. This is illustrated in Figure 2b). Note that the upper side of the needle outlet is located at y=5E-3. This offers an explanation for the fact that the velocity magnitude, whichever the time instant, drops suddenly for values of y shortly smaller than that value. Also, there is a peak of the velocity magnitude when y is slightly inferior to 3E-3, which is caused by circulation of irrigant fluid bellow the needle.



a) b)

Figure 3 BOE needle:
a) Magnitude of the velocity field for $t=t_f$; b) Magnitude of the velocity field, along x=0, computed at different time instants.

The results displayed in Figure 3, corresponding to the BOE needle, are similar to those of the CE-SSO needle. The peak of the velocity magnitude corresponding to the fluid circulation bellow the needle end is now attained for a greater value of y, as the needle is not closed-ended at the bottom. It is also smaller in value, illustrating that the CE-SSO needle is capable of irrigating further in the direction of the apical region. For both needles, it is observed that the velocity magnitude declines sharply, after reaching its peak, as y decreases from 3E-3 to zero. This holds true not only for x=0, but rather for all values of x, suggesting the existence of a stagnation curve, which we define as the furthest set of points located under the needle lower end at which the fluid attains a predefined critical velocity

 ISBN: 978-84-614-6167-7

magnitude v_c. It is assumed that below this velocity the penetration of fresh irrigant is no longer relevant. In this paper we take v_c=0.1. We plot the stagnations curves for both needles and for $t=t_f$ in Figure 4.

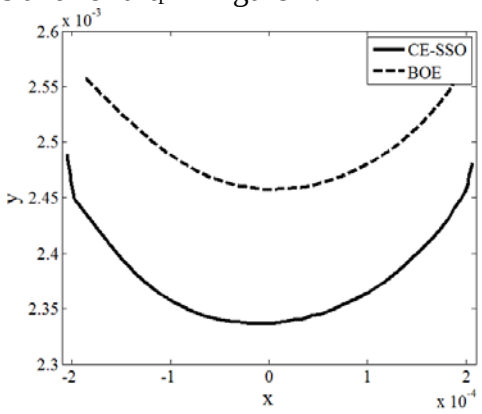


Figure 4: Stagnation curves.

The stagnation curve for the CE-SSO needle lies clearly below that of the BOE needle. This illustrates the fact that the former allows for a broader irrigation. Observe that the notion of stagnation plane that has been reported in the literature translates in our analysis into the horizontal line tangent to the lowest point of the stagnation curve, cf. [4].

4. Tracing particles

The magnitude of the velocity of the irrigant fluid provides an insight to the effectiveness of the irrigation procedure. More information can be obtained by ascertaining the flushing region of the root canal. For the problems related to each needle, this may be defined as the portion of the root canal over which the particles that are initially present are flushed out.

In this section we focus only on particles with negligible mass that are assumed to be dragged by the fluid, offering no resistance. Note that the velocity vector field has been determined, for the case of each needle, as a discrete function of both the spatial coordinates and time. By resorting to an appropriate fitting method, a good approximation to the velocity vector of any particle present in the fluid, at a given time instant, is available. An algorithm may be set up to track the path that will be followed by a particle P which is located at the point of coordinates $(x^0, y^0) \in \Omega_i$, of the computational domain, in the initial time instant $t=t_0$. The algorithm consists of an iterative procedure which at the n^{th} step determines the coordinates (x^n, y^n) of the particle in the time instant $t_n=t_0+\Delta t\times n$. Here, Δt is a prescribed time step. We denote by (u_x^n, u_y^n) the corresponding velocity of the particle when $t= t_n$. The final goal of the algorithm is to determine whether P is flushed out, which is the same as to say that $y^n>y_{max}$, for some value of n, where y_{max}=1.8E-2 is the

 ISBN: 978-84-614-6167-7

maximum height of the root canal. As an output, the Boolean variable P_{out} assumes the value 1 in the case in which P is successfully flushed out and 0 otherwise. The algorithm reads:

Algorithm

Set n = 0. Given t_0, t_f, $(x^0, y^0) \in \Omega_i$, y_{max}, Δt.

1- While $(x^n, y^n) \in \Omega_i$ and $(t_0+\Delta t \times n)<t_f$,
   a) Compute
   $$(u_x^n, u_y^n)$$
   and
   $$x^{n+1} = x^n + \Delta t\, u_x^n,$$
   $$y^{n+1} = y^n + \Delta t\, u_y^n.$$
   b) Increment n:
   $$n := n+1.$$
   End while.
2- If $y^n > y_{max}$, $P_{out}=1$. Otherwise $P_{out}=0$.

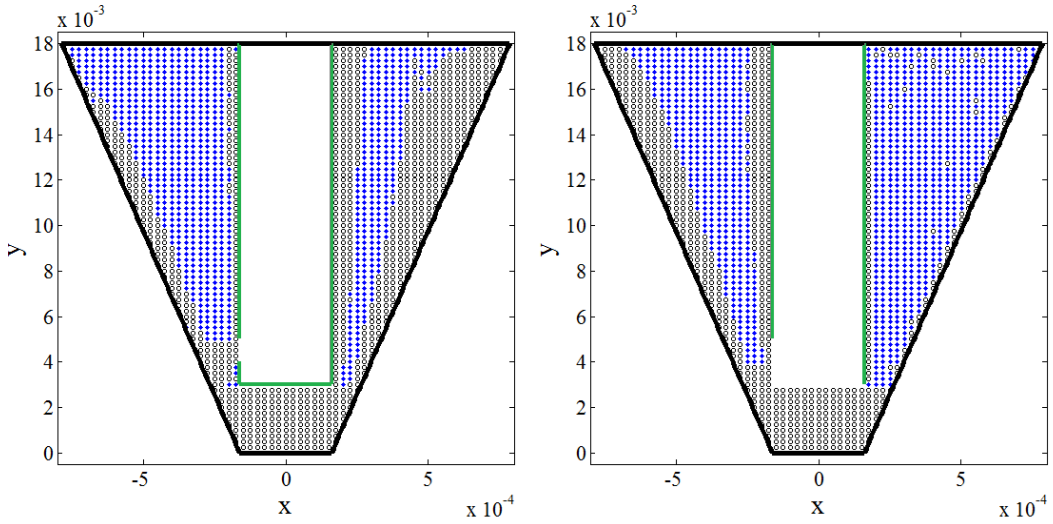


Figure 5

a) Particles flushed out using the CE-SSO needle; b) Particles flushed out using the BOE needle.

This algorithm may be applied to illustrate the behaviour of the particles when the fluid is delivered using either needle (i=1, 2). For that purpose, a rectangular mesh of points is considered over the root canal, the mesh width being h_x=2.5E-5 and h_y=2.5E-4 in the x-direction and the y-direction, respectively. Particles are considered to be located at each of these mesh points at the initial time instant $t=t_0$. The application of the algorithm allows determining which of these particles

 ISBN: 978-84-614-6167-7

are flushed out of the canal after the irrigation procedure, at $t=t_f$. The results are plotted in Figure 5. There, each particle is represented by a circle. Those which are successfully flushed out are depicted as a blue filled circle. Note that if the delivery of the fluid were to be continued, more particles would be flushed out, namely some of those below the needle.

As can be observed, the CE-SSO needle is more effective in flushing out the particles that are, in the figure, to the left of the needle opening. The BOE needle, on the other hand, flushes out more particles on the opposite side. This is consistent with the results plotted in Figures 2 and 3. Note that this is a 2D analysis of a 3D phenomenon, but in any case seems to indicate that better irrigation results might be obtained if the needle is rotated during the procedure of delivery, which is in fact advocated in practice by some clinicians.

5. Conclusions and future work

The present study evaluates how the flow patterns of irrigant delivered by two types of needles differ. The CE-SSO needle achieves a greater penetration depth in the sense that its stagnation curve lies further away from the lowest end of the needle. On the other hand, while this needle achieves better results at flushing out particles located between the needle and the wall facing its outlet, it is outperformed by the BOE needle in the region between the needle and the opposite wall. This finding underlines the importance of rotating the needle for adequate irrigant replacement.

In the future the analysis may be extended to 3D, include the apical region, more needle types and different flow rates, as well as account for the shear stress on the root canal walls. The algorithm should be improved to take into account the interactions of the particles with the walls.

6. References

[1] C. BOUTSIOUKIS, B. VERHAAGEN, M. VERSLUIS, E. KASTRINAKIS, ET AL., *Irrigant Flow in the Root Canal: Experimental Validation of an Unsteady Computational Fluid Dynamics Model using High-speed Imaging,* Int. Endod. J., 43 (2010), 393-403.

[2] C. BOUTSIOUKIS, B. VERHAAGEN, M. VERSLUIS, E. KASTRINAKIS, ET AL., *Evaluation of Irrigant Flow in the Root Canal Using Different Needle Types by an Unsteady Computational Fluid Dynamics Model,* J. Endod., 36 (2010), 875-979.

[3] Y. GAO,Y, M. HAAPASALO,Y. SHEN, H. WU, ET AL., *Development and Validation of a Three-dimensional Computational Fluid Dynamics Model of Root Canal Irrigation,* J. Endod., 35 (2009), 1282-1287.

[4] K GULABIVALA, Y-L NG, M GILBERTSON AND I EAMES, *The Fluid Mechanics of Root Canal Irrigation,* Physiol. Meas., 31 (2010), 49–84.

 ISBN: 978-84-614-6167-7

[5] T. KOCHARIAN, *Root Canal Irrigation - an Engineering Analysis Using Computational Fluid Dynamics*, Master of Engineering project reports, Department of Mechanical and Industrial Engineering, Faculty of Applied Science and Engineering, University of Toronto (2010).

[6] Y. SHEN, Y. GAO, W. QIAN, N. RUSE, ET AL., *Three-dimensional Numeric Simulation of Root Canal Irrigant Flow with Different Irrigation Needles*, J. Endod., 36 (2010), 884-889.

 ISBN: 978-84-614-6167-7

Proceedings of the 11th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE2011
Benidorm, Spain, 26-30 June 2011

Rainy fields motion computation using optical flow

O. Raaf[1], A.Adane[1]

[1] *Laboratory of Image Processing and Radiation, Department of Telecommunications, Faculty of Electronics and Computer Science, University of Science and Technology Houari Boumediene (U.S.T.H.B.)*

emails: rf_ouarda@yahoo.fr, tirgez@yahoo.fr

Abstract

The climatic changes observed from the beginning of the third millennium till to-day, have made necessary the weather forecast in real time. Appropriate meteorological previsions are then crucial to prevent the populations against possible disasters. Thanks to remote sensing tools like satellite or radar, these phenomena can be easily observed in space and time. In the case of radar observations, precipitation fields are detected with a good resolution and the motion of clouds can be followed at any time. The goal of this work is the detection and extraction of velocity fields of precipitation from radar images by solving the equation of optical flow. Two artificial 'cube' and 'taxi 'sequence of images and, series of PPI filtered images 512 x 512 pixels were tested. The motion of rainy cells was analyzed by using global motion model parameters. Applying the differential method of 'Horn & Schunck a cell isolated from a rainy radar image allowed us to detect the overall sense of moving objects.

Key words: Velocity fields, Rainy clouds, Optical Flow, Radar image, Horn & Schunck

1. Introduction

The study and analysis of cloud progress has been the subject of several scientific works in recent years. To account for these phenomena, most meteorological networks are now equipped with automatic measurement systems and remote sensing tools [1]. In particular, different geostationary satellites intended to meteorological observations, orbit the Earth in the equatorial plane for about thirty years. Radars, Lidars and Sodars are other remote sensing systems used to collect meteorological data characterising the air motions and cloud formations in

 ISBN: 978-84-614-6167-7

the atmosphere [2-3]. Many meteorological stations are equipped with classical radars. For a great part, these radars are pulsed, non Doppler and especially designed to detect rainy clouds. Their antenna usually scans a circular earth surface of about two hundred km of radius and the resulting PPI (Plan Position Indicator) images of precipitation fields are collected at time intervals tuneable from five to fifteen minutes. In comparison with meteorological satellites, the classical radars operate at smaller scale. However, they can advantageously observe the cloud motions in real-time and more efficiently account for the cloud changes over the time. The radar equipment represents rain cells with: position, shape, intensity, dimensions and displacement. The data radar can be used to detect other severs atmospheric phenomena's such as hail storms and tornados.

Motion estimation from an image sequence is used in several applications. In robotics, it can identify and anticipate changes in the position of objects. In video compression, it allows the fullest possible understanding where the temporal redundancy of the sequence and information describing an image using the surrounding images. In meteorology, it allows the detection and forecasting of rainy clouds including those which are dangerous and predicting their motion.

The movement, in a sequence of images is visible through changes in spatial distribution of a photometric variable between two successive images, such as luminance, brightness, or reflectivity which is the variable used in weather images. With the use of non-linear algorithms, signal processing and modern mathematics have been opened to the study of such a random phenomenon. So there are several large families of methods for motion estimation according to the different choices (differential methods, frequency, methods based on the phase or the energy,…)

In this study, we are interested to solving the optical flow equation by using global method of Horn & Schunck in the detection and extraction of the velocity of rain fields observed in radar images. Thus, first a formulation of the optical flow equation is given. After that, an application of the method on images from the databank is illustrated. Finally, the results are discussed and interpreted.

2. Optical flow equation

Optical flow is a method that is based on the calculation of apparent motion in 2D images. It is therefore possible to estimate the motion of a point by superposing two consecutive images using the method of optical flow [4-5]. The optical flow uses the assumption of conservation of brightness *I(x,y)*. It is represented by the following equation:

$$\mathrm{I}(p,t) = I\ (p\ + V(p)\,\delta t, t + \delta t) \tag{1}$$

 ISBN: 978-84-614-6167-7

With $v(p) = v(u, v)$: velocity vector at point p at time t. u and v are respectively the velocity along x and y.

Since the movement does not vary too much from one image to another, employing differential equations needed:

$$I(x + dx, y + dy, t + dt) = I(x, y, t) + \frac{\delta I}{\delta t} dx + \frac{\delta I}{\delta t} dy + \frac{\delta I}{\delta t} dt + o(\varepsilon) \quad (2)$$

With ε contains higher order terms of Taylor expansion of first order and goes to 0 when dt tends to 0. Considering that the quantities dx, dy and dt is sufficiently small and that equation (2) is satisfied, then:

$$\nabla I(p_i, t) V(p_i, t) + I_t = 0 \quad (3)$$

I_x, I_y, I_t : spatio-temporal derivative of intensity at point $p\ (x, y)$ at time t.
$\nabla I = [\, I_x \;\; I_y]$: Gradient Space

Several methods have been used to solve the equation of optical flow. Differential methods belong to the techniques commonly used for calculating the optical flow in image sequences. Their gains lie in reducing the complexity of calculations commonly used in matching methods while increasing the range of measurable displacements. They can be classified into global methods such as the technique of Horn and Schunck and Local as the Lucas-Kanade approach and that based on Gabor filters.

In other studies we have experienced global method of Lukas and Kanade and that based on Gabor filter to determine the velocity fields of moving rain cells [6]. In this study we will apply the method of global differential Horne and Schunck with rain clouds seen on weather radar images.

3. Horn & Schunck Algorithm

It is considered among one of the first methods of calculating optical flow. It was developed in 1981 by researchers G. Horn and B. Schunck[7].
His calculation program is designed to avoid transparency effects and shadowing. These are costly in terms of computation. For this, Horn et al assume that the image is 2D[8]. Subsequently, they assume that the luminance is uniformly distributed over the surface. In this case, the brightness of a point of the image corresponds to a point of the scanned object. They also assume that the luminance varies little from one image to another.
In this method, we introduce a constraint named constraint or smoothing term

 ISBN: 978-84-614-6167-7

regulation α which postulates that the desired velocity field is regular. The motion estimation is then done by minimizing a cost function E which is on the whole domain of the image (global approach), with:

$$E = \sum[(\nabla \mathrm{I}(\mathrm{p_i}, \mathrm{t})\mathrm{V}(\mathrm{p_i}, \mathrm{t}) + \mathrm{I_t}(\mathrm{p_i}, \mathrm{t})]^2 + \alpha(\nabla \mathrm{I}(\mathrm{p_i}, \mathrm{t})^2) \quad (4)$$

The minimization is solved numerically using an iterative Gauss-Seidel type, which deducts the vector components u and v. This amounts to solve the Euler-Lagrange equivalent. The result is given iteratively by:

$$\begin{cases} u_n - \bar{u}_n = -I_x \dfrac{I_x u_n + I_y v_n + I_t}{\alpha^2 + I_x^2 + I_y^2} \\ v_n - \bar{v}_n = -I_y \dfrac{I_x u_n + I_y v_n + I_t}{\alpha^2 + I_x^2 + I_y^2} \end{cases} \quad (5)$$

$\bar{u}_n, \bar{v}_n$: Main values of u_n *and* v_n respectively

4. RESULTS AND INTERPRETATION

- This algorithm was tested on three image sequences:
  The artificial sequence 'cube' consisting of a cube placed on a turntable, the image size is 256 × 240 pixels.
- The artificial sequence ‘taxi’: is a shot of a street intersection, and is composed of three cars and a pedestrian in motion. The two vehicles move down in the opposite direction. The taxi to the center turns into the corner, the image size is 256 × 190 pixels.
- A real sequence of radar images 512x 512 pixels with 64 levels of reflectivity taken on August 4th, 1996 from 16:45 until to 17h 15. These images are collected by non-coherent radars every 5 minutes installed at the meteorological stations of Bordeaux (France). In order to view the evolution of the clouds movements, we have presented on Fig.1 a small portion of the radar images were the clouds are concentrated

All images used in this application were reduced according to their size in order to see the movement.

 ISBN: 978-84-614-6167-7

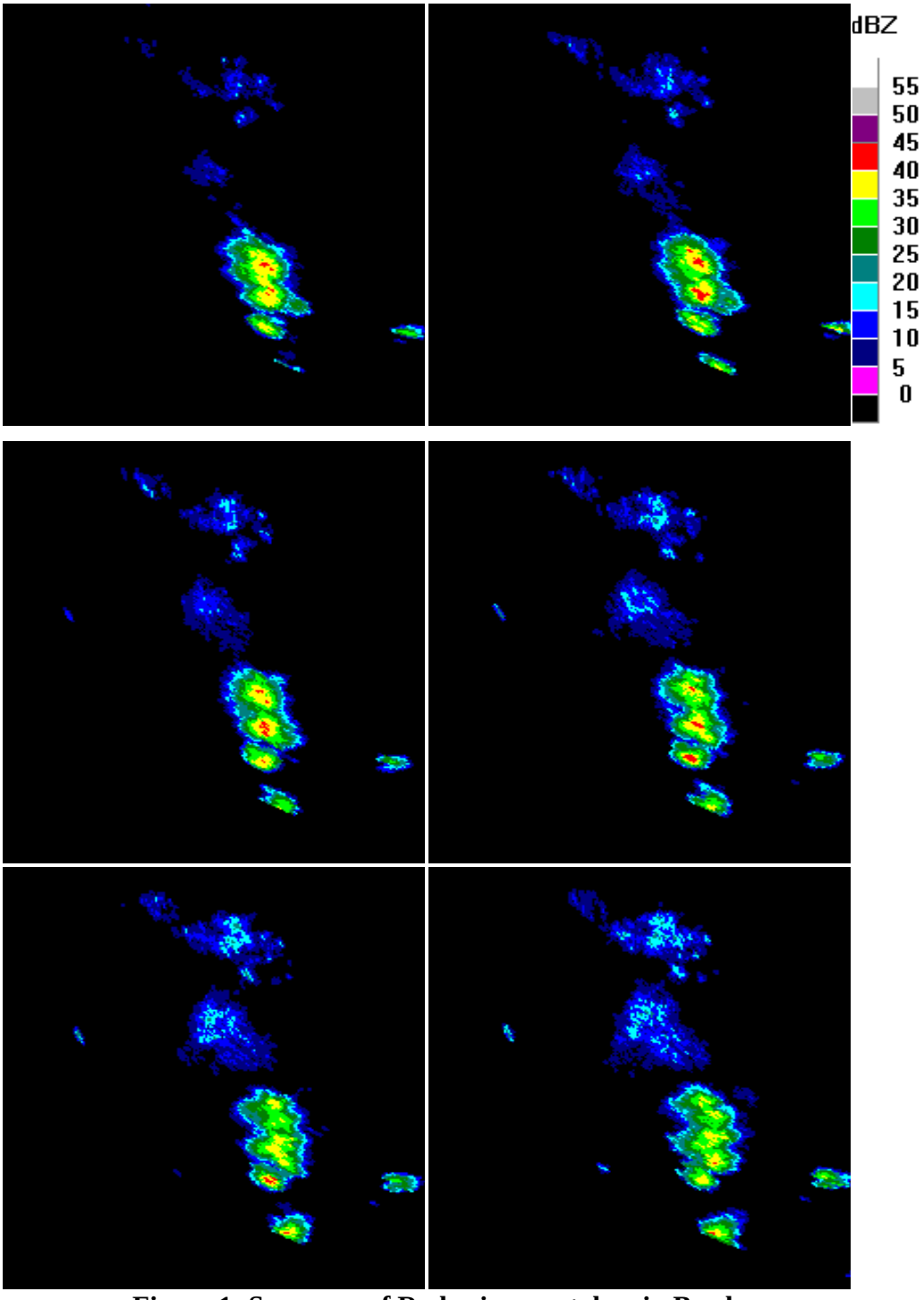


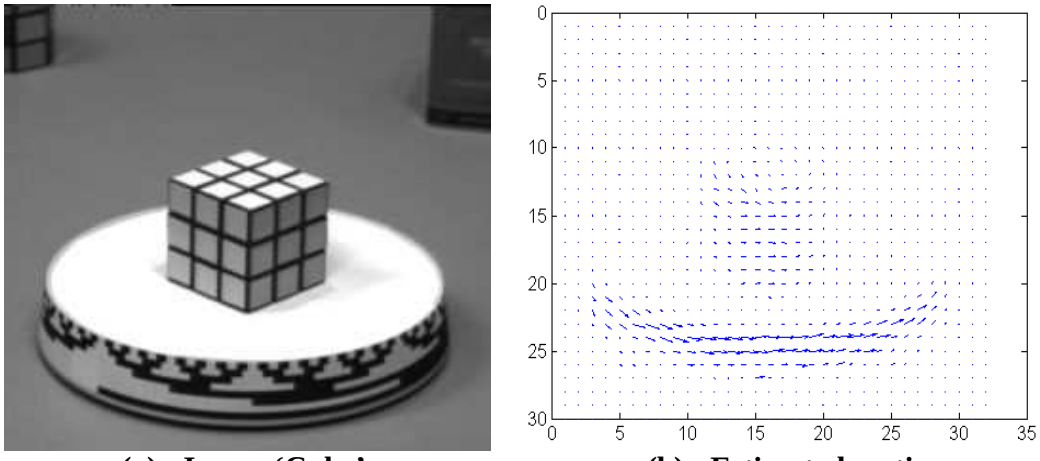
Figure 1: Sequence of Radar images taken in Bordeaux (04/08/1996-16h54:17h15)

 ISBN: 978-84-614-6167-7

After several tests on the parameterization constants involved in the computation of optical flow for the three sequences used, best results were obtained under the following conditions:

- Cube: K = 3, 60 iterations, alpha=50: Figure 2 shows velocity vectors corresponding to the rotation of the plate and that of the cube below it.
- Taxi: K = 3, 60 iterations, alpha=20: Figure 3 shows three moving objects that are both vehicles of low gray moving in opposite directions and the white taxi in the center that turns the corner. The movement of the pedestrian is not detected given its small size.
- Image radar: K = 2, 10 iterations and alpha=20: In Figure 4 appears vectors speeds almost headed north-east which indicates the direction of the overall development of the precipitation. We also note that the estimated motion is null in the center of the big cells and increases gradually in the vicinity of the contour, which seems logical because the cloud changes little in 5 minutes and that mainly the edges of the cell that are most subject to the constraint of deformation given the time gap reduced. The movement of small rain cell is not detected. Outside the rain cell the estimated field is practically null

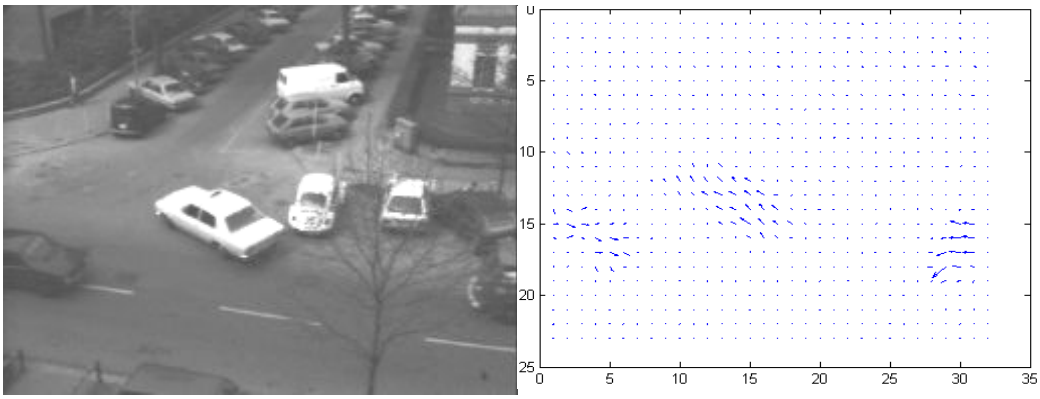
On setting the parameters for obtaining good results, there is a tradeoff between alpha and the number of iterations since the greater the number of iteration is big plus there are better movement area boundaries but instead was removed small displacements as in the case of pedestrian movement was not detected in the sequence 'Taxi' (Figure 3). On the other hand, whenever alpha increases the result is noisy month to achieve optimum value from which the result begins to deform



(a) : Image 'Cube' **(b) : Estimated motion**

Fig.2 : Optical flow estimate for the sequence 'cube'

 ISBN: 978-84-614-6167-7



(a): Image 'Taxi' **(b) : Estimated motion**

Fig.3 : Optical flow estimate for the sequence 'taxi'

5. Conclusion

We have seen in this study the global method of Horn and Schunck and the influence of parameters choices on the results. In all three cases: taxi,cube and radar images, the estimated optical flow shows the main objects that are moving in the sequences. We also deduce that a higher number of iteration eliminates the small movements and retain the main movements. Applying this method to the radar cells gave us motion fields for the edge pixels those move or those change intensity value. Also, setting the parameters estimation depends mainly on the size, shape and evolution of the cell. Thus, the application of this method to weather radar images allowed us to detect the global direction of movement of the storm.

6. References

[1] D. Renaut, Les satellites meteorologiques (Meteorological satellites), La Meteorologie. 45 (2004) 33-37.

[2] A.Meischner, Weather radar: principles and advanced applications. Springer, 2005.

[3] H. Sauvageot , Radar Meteorology. Artech House, Boston, 1992.

[4] Baron J.L., Fleet DJ. and Beauchemin SS: Performance of optical flow techniques. *In International Journal on Computer Vision*, vol. 12, pp. 43-77(1994).

[5] E. Mémin, Estimation du flot optique-contribution et panorama des différents approches ; Habilitation à diriger des recherches. Université de Rennes 1,Rennes(2003).

[6] *O. Raaf and A. Adane, . Nonlinear Dynamics of Complex Systems Book'' edited by Albert C.J., José Antonio Tenreiro Machado, and Dumitru*

 ISBN: 978-84-614-6167-7

Baleanu qui sera publié par Springer Science+Business To be publeshed in Septembre 2011

[7] D. Lucas, Takeo Kanade. An Iterative Image Registration Technique with an Application to Stereo Vision. From *Proceedings of Imaging Understanding Workshop*, pp. 121-130 (1981).

[8] A. Bruhn and J. Weickert, Lucas/Kanade Meets Horn/Schunck: Combining Local and Global Optic Flow Methods. International Journal of Computer Vision 61(3), 211–231, 2005.

 ISBN: 978-84-614-6167-7

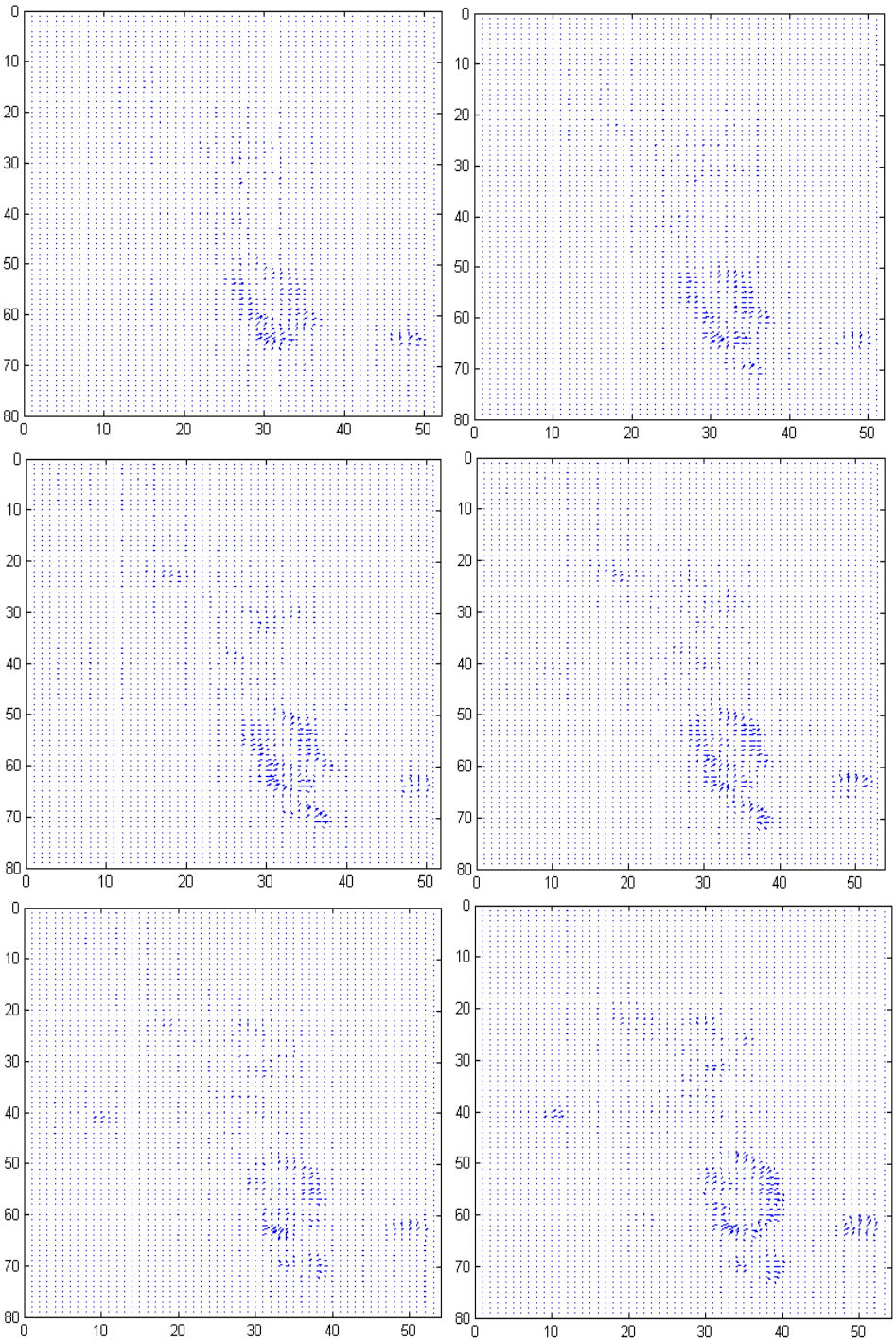


Fig.4 : Motion estimation for the sequence of radar images.

 ISBN: 978-84-614-6167-7

Proceedings of the 11th International Conference on Computational and Mathematical Methods in Science and Engineering, CMMSE2011 Alicante, Spain, 26-30 June 2011

Impulsive Biological Pest Control of the Sugarcane Borer

Marat Rafikov[1], Alfredo Del Sole Lordelo[1] and Elvira Rafikova[2]

[1] *Centro de Engenharia, Modelagem e Ciências Sociais Aplicadas, Universidade Federal do ABC, Santo André, SP, Brazil*
[2] *Departamento de Engenharia Mecânica,UNICAMP, Campinas, SP, Brazil*

emails: marat9119@yahoo.com.br, alfredo.lordelo@ufabc.edu.br, elviraelya@yahoo.com.br

Abstract

In this paper, we propose an impulsive biological pest control of the sugarcane borer *(Diatraea saccharalis)* by its egg parasitoid *Trichogramma galloi* based on a mathematical model in which the sugarcane borer is represented by the egg e larval stages, and the parasitoid is considered in term of the parasitized eggs. By using the Floquet theory and the small amplitude perturbation method, we show that there exists an asymptotically stable pest-eradication periodic solution when some conditions hold. The numerical simulations show that the impulsive release of parasitoids provides reliable strategies of the biological pest control of the sugarcane borer.

Key words: mathematical modelling, biological control, sugarcane borer, egg parasitoid
MSC2000: AMS 92D25, 34H05, 49N90

1. Introduction

One of challenges of the improvements in the farming and harvesting of cane is the biological pest control. Biological control is defined as the reduction of pest populations by using their natural enemies: predators, parasitoids and pathogens [3]. Parasitoids are species which develop within or on the host and ultimately kill it. Thus, parasitoids are commonly reared in laboratories and periodically released in high-density populations as biological control agents of crop pests [4].

The sugarcane borer *Diatraea sacharalis* is reported to be the most important sugarcane pest in south-east region of Brazil [1]. The sugarcane borer builds

 ISBN: 978-84-614-6167-7

internal galleries in the sugarcane plants causing direct damages that result in apical bud death, weight loss and atrophy. Indirect damages occur when there is contamination by yeasts that cause red rot in the stalks, either causing contamination or inverting the sugar, increasing yield loss in both sugar and alcohol [2].

Cotesia flavipes is important wasp parasitoid of the sugarcane borer larvae in Brazil [1]. In spite of the biological control of *Diatraea saccharalis* by *Cotesia flavipes* is considered successful in Brazil, there are some areas where *Cotesia flavipes* has not the good control. The using of the egg parasitoid *Trichogramma galloi* is considered an interesting option in this case [5].

Mathematical modeling is an important tool used in studying agricultural problems. Thus, a good strategy of biological pest control, based on mathematical modeling, can increase the ethanol production. The application of host-parasitoid models for biological control were reviewed in [6].

In [7] was proposed a mathematical model of interaction between the sugarcane borer *(Diatraea saccharalis)* and its egg parasitoid *Trichogramma galloi* which consist of three differential equations:

$$\begin{aligned} \frac{dx_1}{dt} &= r(1-\frac{x_1}{K})x_1 - m_1x_1 - n_1x_1 - \beta x_1x_2 \\ \frac{dx_2}{dt} &= \beta x_1x_2 - m_2x_2 - n_2x_2 \\ \frac{dx_3}{dt} &= n_1x_1 - m_3x_3 - n_3x_3 \end{aligned} \tag{1}$$

were x_1 is the egg density of the sugarcane borer, x_2 is the density of eggs parasitized by *Trichogramma galloi* and x_3 is the larvae density of the sugarcane borer; r is the net reproduction rate; K is the carrying capacity the environment; m_1, m_2 and m_3 are mortality rates of the egg, parasitized egg and larvae populations; n_1 is the fraction of the eggs from which the larvae emerge at time t; n_2 is the fraction of the parasitized eggs from which the adult parasitoids emerge at time t; n_3 is the fraction of the larvae population which moults into pupal stage at time t; β is the rate of parasitism.

Recently, many authors have investigated the different population models concerning the impulsive pest control [8]–[13]. The impulsive pest control strategies based on prey-predator models were presented in [8], [11], [13]. The host-parasitoid model with impulsive control was considered in [10]. Impulsive strategies of a pest management for SI epidemic models were proposed in [9], [12].

In this paper, we suggest an impulsive differential equations [14] to model the process of the biological pest control of the sugarcane borer. So we develop (1) introducing a periodic releasing the parasitoids at fixed times

 ISBN: 978-84-614-6167-7

$$\begin{cases} \left.\begin{aligned} \frac{dx_1}{dt} &= r(1-\frac{x_1}{K})x_1 - m_1x_1 - n_1x_1 - \beta x_1x_2 \\ \frac{dx_2}{dt} &= \beta x_1x_2 - m_2x_2 - n_2x_2 \\ \frac{dx_3}{dt} &= n_1x_1 - m_3x_3 - n_3x_3 \end{aligned}\right\} t \neq n\tau,\ n \in Z_+, \\ \left.\begin{aligned} \Delta x_1(t) &= 0 \\ \Delta x_2(t) &= p \\ \Delta x_3(t) &= 0 \end{aligned}\right\} t = n\tau,\ n \in Z_+, \end{cases} \tag{2}$$

where p is the release amount of the parasitized eggs at $t = n\tau,\ n \in Z_+, Z_+ = \{0,1,2,...,\},\ \tau$ is the period of the impulsive effect. $\Delta x_i = x_i(t^+) - x_i(t),\ x_i(t^+) = \lim_{t \to t^+} x_i(t),\ i = 1,2,3.$ That is, we can use releasing parasitized eggs to eradicate pests or keep the pest population below the economic damage level.

2. Preliminary

In this section, we will give some definitions, notations and some lemmas which will be useful for our main results.

Let $R_+ = [0,\infty), R_+^3 = \{x \in R^3 : x > 0\}$. Denote $f = (f_1, f_2, f_3)^T$, the map defined by the right hand side of the first three equations of the system (2). Let $V_0 = \{V : R_+ \times R_+^3 \mapsto R_+\}$, continuous on $(n\tau,\ (n+1)\tau] \times R_+^3$, $\lim_{(t,y)\to(n\tau^+,x)} V(t,y) = V(n\tau^+,x)$ exist and V is locally Lipscitzian in x.

Definition 2.1. $V \in V_0$, then for $(t,x) \in (n\tau,\ (n+1)\tau] \times R_+^3$, the upper right derivative of *V(t,x)* with respect to the impulsive differential system (2) is defined as

$$D^+V(t,x) = \lim_{h \to 0} \sup \frac{1}{h}[V(t+h, x+hf(t,x)) - V(t,x)] \ .$$

The solution of system (2), denoted by $x(t) : R_+ \mapsto R_+^3$, is continuously differentiable on $(n\tau,\ (n+1)\tau] \times R_+^3$. Obviously, the global existence and uniqueness of solution of system (2) is guaranteed by the smoothness properties of *f*, for details see [14].

We will use a basic comparison results from impulsive differential equations.

Lemma 2.1 [14]. Let $V \in V_0$, assume that

 ISBN: 978-84-614-6167-7

$$\begin{cases} D^+V(t,x) \le g(t,V(t,x)),\ t \ne n\tau, \\ V(t,x(t^+)) \le \psi_n(V(t,x(t))),\ t = n\tau, \end{cases} \tag{3}$$

where $g: R_+ \times R_+^3 \mapsto R_+$ is continuous on $(n\tau,(n+1)\tau] \times R_+$ and $\psi_n : R_+ \mapsto R_+$ is nondecreasing. Let *R(t)* be the maximal solution of the scalar impulsive differential equation

$$\begin{cases} \dot{u}(t,x) = g(t,u(t)),\ t \ne n\tau, \\ u(t^+) = \psi_n(u(t)),\ t = n\tau, \\ u(0^+) = u_0 \end{cases} \tag{4}$$

existing on $[0,\infty)$. Then $V(0^+,x_0) \le u_0$ implies that $V(t,x(t)) \le R(t), t \ge 0$, where *x(t)* is any solution of (2), similar results can be obtained when all the directions of the inequalities in the lemma are reversed and ψ_n is nonincreasing. Note that if we have some smoothness conditions of *g* to guarantee the existence and uniqueness of solutions for (4), then *R(t)* is exactly the unique solution of (4).

Next, we consider the following sub-system of system (2)

$$\begin{cases} \dfrac{dx_2}{dt} = -m_2x_2 - n_2x_2, & t \ne n\tau \\ \Delta x_2(t) = p, & t = n\tau \\ x_2(0^+) = x_{20} \ge 0 \end{cases} \tag{5}$$

Lemma 2.2. System (5) has a unique positive periodic solution $\tilde{x}_2(t)$ with period τ and for every solution $x_2(t)$ of (5) $|x_2(t) - \tilde{x}_2(t)| \to 0$ as $t \to \infty$, where

$$\tilde{x}_2(t) = \frac{pe^{-(m_2+n_2)(t-n\tau)}}{1-e^{-(m_2+n_2)\tau}},\ t \in (n\tau,(n+1)\tau],\ n \in Z_+ \tag{6}$$

$$\tilde{x}_2(0^+) = \frac{p}{1-e^{-(m_2+n_2)\tau}} \tag{7}$$

Proof. Integrating and solving the first equation of (5) between pulses, we get

$$x_2(t) = x_2(n\tau^+)e^{-(m_2+n_2)(t-n\tau)},\ t \in (n\tau,(n+1)\tau]. \tag{8}$$

After each successive pulse, we can deduce the following map of system (8)

$$x_2((n+1)\tau^+) = x_2(n\tau^+)e^{-(m_2+n_2)\tau} + p,\ t \in (n\tau,(n+1)\tau] \tag{9}$$

Equation (9) has a unique fixed point $x_2^* = \dfrac{p}{1-e^{-(m_2+n_2)\tau}}$, it corresponds to the unique positive periodic solution $\tilde{x}_2(t)$ of system (5), the initial value $\tilde{x}_2(0^+) = \dfrac{p}{1-e^{-(m_2+n_2)\tau}}$. The fixed point x_2^* of map (9) implies that there is a

 ISBN: 978-84-614-6167-7

corresponding cycle of period τ in $x_2(t)$, that is $\widetilde{x}_2(t)=\dfrac{pe^{-(m_2+n_2)(t-n\tau)}}{1-e^{-(m_2+n_2)\tau}},\ t\in(n\tau,(n+1)\tau],\ n\in Z_+$. From (9) we obtain

$$x_2(n\tau^+)=x_2(0^+)e^{-n(m_2+n_2)\tau}+\frac{p(1-e^{-n(m_2+n_2)\tau})}{1-e^{-(m_2+n_2)\tau}},$$

thus, $x_2(n\tau^+)\to x_2^*$ as $t\to\infty$, so $\widetilde{x}_2(t)$ is globally asymptotically stable. From (6) and (8) we have $x_2(t)=(x_2(0^+)-\widetilde{x}_2(0^+))e^{-(m_2+n_2)(t-n\tau)}$.

Consequently, $x_2(t)\to\widetilde{x}_2(t)$ as $t\to\infty$, that is $|x_2(t)-\widetilde{x}_2(t)|\to 0$ as $t\to\infty$.□

Therefore, system (2) has a pest-eradication periodic solution $(0,\widetilde{x}_2(t),0)$, where $\widetilde{x}_2(t)$ is defined by (6).

To study the stability of the pest-eradication periodic solution of (2) we present the Floquet theory for a linear τ periodic impulsive equation

$$\begin{cases}\dfrac{dx}{dt}=A(t)x, & t\neq\tau_k,\ t\in R\\ x(t^+)=x(t)+B_kx(t), & t=\tau_k,\ k\in Z_+\end{cases}\tag{10}$$

Then we introduce the following conditions:

(H_1) $A(\cdot)\in PC(R,\ C^{n\times n})$ and $A(t+\tau)=A(t)$ $(t\in R)$, where $PC(R,\ C^{n\times n})$ is the set of all piecewise continuous matrix functions which is left continuous at $t=\tau_k$, and $C^{n\times n}$ is the set of all $n\times n$ matrices.

(H_2) $B_k\in C^{n\times n}$, $\det(E+B_k)\neq 0$, $\tau_k<\tau_{k+1}$ $(k\in Z_+)$.

(H_3) There exist a $h\in Z_+$, such that $B_{k+h}=B_k$, $\tau_{k+h}=\tau_k+\tau$ $(k\in Z_+)$.

Let $\Phi(t)$ be the fundamental matrix of (10), then there exists a unique nonsingular matrix $M\in C^{n\times n}$ such that

$$\Phi(t+\tau)=\Phi(t)M.\tag{11}$$

By equality (11) there correspondents to the fundamental matrix $\Phi(t)$ the constant matrix M which is called monodromy matrix of (10). All monodromy matrices of (10) are similar and have the same eigenvalues. The eigenvalues $\lambda_1,\lambda_2,...,\lambda_n$ of the monodromy matrices are called the Floquet multipliers of (12).

Lemma 2.3 [14]. (Floquet theory) Let conditions (H_1-H_3) hold. Then the linear τ periodic impulsive system (10) is

a) stable if and only if all multipliers λ_i (i = 1,2, …, n) of equation (10) satisfy the inequality $|\lambda_i|\le 1$.

b) asymptotically stable if and only if all multipliers λ_i (i = 1,2, …, n) of equation (10) satisfy the inequality $|\lambda_i|<1$.

 ISBN: 978-84-614-6167-7

c) unstable if $|\lambda_i| > 1$ for some i = 1,2, …, n.

3. Stability of the pest-eradication periodic solution

In this section, we study the stability of the pest-eradication periodic solution $(0, \tilde{x}_2(t), 0)$ of the system (2). Next, we present an important result, concerning a condition that guarantees the local stability of this solution.

Theorem 3.1. The pest-eradication periodic solution $(0, \tilde{x}_2(t), 0)$ of the system (2) is locally asymptotically stable provided that inequality

$$p > \frac{(r - m_1 - n_1)(m_2 + n_2)\tau}{\beta} \tag{12}$$

holds.

Proof. The local stability of a periodic solution $(0, \tilde{x}_2(t), 0)$ of system (2) may be determined by considering the behavior of small-amplitude perturbations $(y_1(t), y_2(t), y_3(t))$ of the solution.

Define

$$x_1(t) = y_1(t), \quad x_2(t) = \tilde{x}_2(t) + y_2(t), \quad x_3(t) = y_3(t), \tag{13}$$

where $y_1(t), y_2(t), y_3(t)$ are small perturbations.

Linearizing the system (2), we have the following linear τ periodic impulsive system

$$\begin{cases} \left.\begin{aligned} \frac{dy_1}{dt} &= r\,y_1 - m_1 y_1 - n_1 y_1 - \beta \tilde{x}_2 y_1 \\ \frac{dy_2}{dt} &= \beta \tilde{x}_2 y_1 - m_2 y_2 - n_2 y_2 \\ \frac{dy_3}{dt} &= n_1 y_1 - m_3 y_3 - n_3 y_3 \end{aligned}\right\} t \neq n\tau,\ n \in Z_+, \\ \left.\begin{aligned} y_1(t^+) &= y_1(t) \\ y_2(t^+) &= y_2(t) \\ y_3(t^+) &= y_3(t) \end{aligned}\right\} t = n\tau,\ n \in Z_+, \end{cases} \tag{14}$$

Let $\Phi(t)$ be the fundamental matrix of (14). Then $\Phi(t)$ must satisfy the following equation

$$\frac{d\,\Phi(t)}{d\,t} = \begin{bmatrix} r - m_1 - n_1 - \beta \tilde{x}_2 & 0 & 0 \\ \beta \tilde{x}_2 & -m_2 - n_2 & 0 \\ n_1 & 0 & -m_3 - n_3 \end{bmatrix} \Phi(t) \tag{15}$$

 ISBN: 978-84-614-6167-7

and initial condition

$\Phi(t) = I$

where I is the identity matrix.

The solution of (16) is

$$\Phi(t) = \begin{bmatrix} \exp(\int_0^t (r - m_1 - n_1 - \beta \tilde{x}_2(s))ds) & 0 & 0 \\ * & \exp(-(m_2 + n_2)\tau) & 0 \\ * & * & \exp(-(m_3 + n_3)\tau) \end{bmatrix}, \quad (16)$$

there is no need to calculate the exact form of (*) as it is not required in the analysis that follows. The resetting impulsive condition of (14) become:

$$\begin{bmatrix} x_1(n\tau^+) \\ x_2(n\tau^+) \\ x_3(n\tau^+) \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_1(n\tau) \\ x_2(n\tau) \\ x_3(n\tau) \end{bmatrix} \quad (17)$$

Hence, if absolute values of all eigenvalues of

$$M = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \Phi(\tau) = \Phi(\tau) \quad (18)$$

are less than one, the τ periodic solution is locally stable. Then the eigenvalues of M are the following

$$\begin{aligned} \lambda_1 &= \exp(\int_0^t (r - m_1 - n_1 - \beta \tilde{x}_2(s))ds) \\ \lambda_2 &= \exp(-(m_2 + n_2)\tau) < 1 \\ \lambda_3 &= \exp(-(m_3 + n_3)\tau) < 1 \end{aligned} \quad (19)$$

From (19) one can see that $|\lambda_1| < 1$ if and only if condition (12) holds true. According to Lemma (2.3), the pest-eradication periodic solution $(0, \tilde{x}_2(t), 0)$ is locally asymptotically stable. □

4. Numerical simulations of the impulsive biological control

For numerical simulations of interactions between the sugarcane borer and its parasitoid were used the following values of model coefficients: $n_1 = 0.1$, $n_2 = 0.1$, $n_3 = 0.02439$, $m_1 = 0.03566$, $m_2 = 0.03566$, $m_3 = 0.00256$, $K = 25000$. These values were obtained based on data published about the use of the egg parasitoid *Trichogramma galloi* against the sugarcane borer *Diatraea saccharalis*these [1], [5], [7]. Fig. 1 shows the population oscillations for

 ISBN: 978-84-614-6167-7

$r = 0.1908$ and $\beta = 0.0001723$ without control. One can see from this figure that the sugarcane borer larvae density x_3 takes on values more than the pest density threshold level $x_E = 2500$ numbers/ha [1]. Densities above this level cause economic damages the sugarcane crops. In this case, it is necessary to apply the biological control.

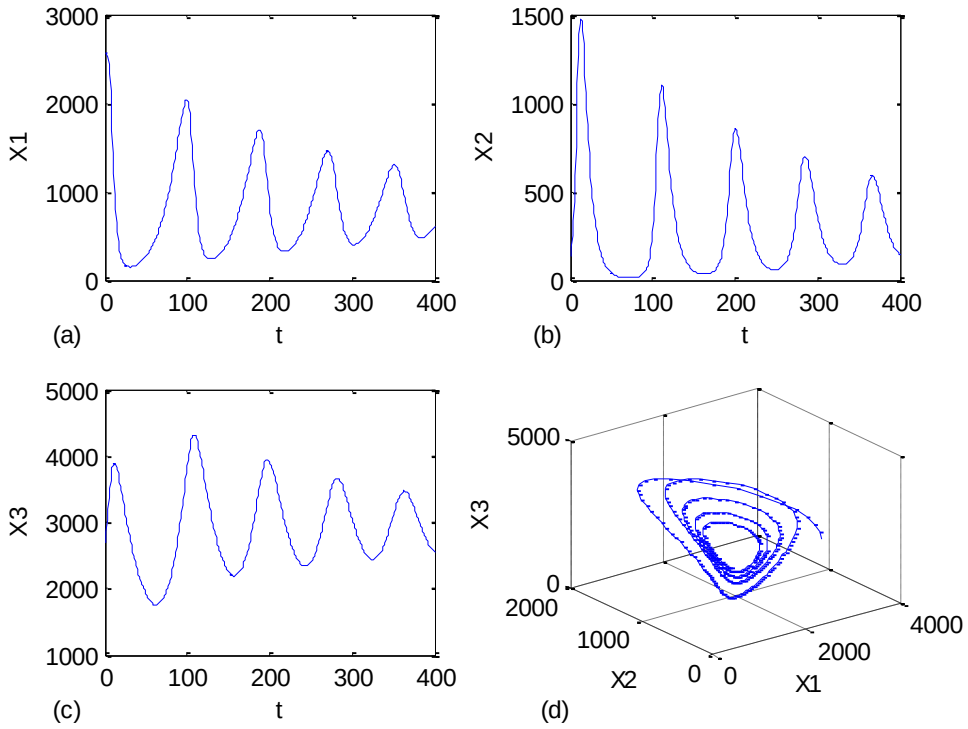


Fig.1. Evolution of the egg (a), parasitized egg (b), larvae populations (c) and phase portraits (d) of system (2) without control

From Theorem 3.1, we have shown that the pest-eradication periodic solution $(0, \tilde{x}_2(t), 0)$ of the system (2) is locally asymptotically stable if the condition (12) holds:

$$p > p_{\min} = \frac{(r - m_1 - n_1)(m_2 + n_2)\tau}{\beta} \qquad (20)$$

Choosing $\tau = 70$ days from (20) we derive that when $p > p_{\min} = 3027$ parasitoids/ha the pest-eradication periodic solution of the host – parasitoid system is asymptotically stable. Dynamical behaviour of the system with impulsive control $p = 3500$ parasitoids/ha is shown in Fig. 2. We can conclude that this control strategy is seemed successful because the larvae population of the sugarcane borer goes to extinction. But the aim of the biological control is not to eliminate all larvae population. The aim of the biological control

 ISBN: 978-84-614-6167-7

of the sugarcane borer is to keep the larvae population at an acceptable low level (below the economic injury level) that indicates the pest densities at which applied biological control is economically justified. It is known that the economic injury level is $x_E = 2500$ numbers/ha [1].

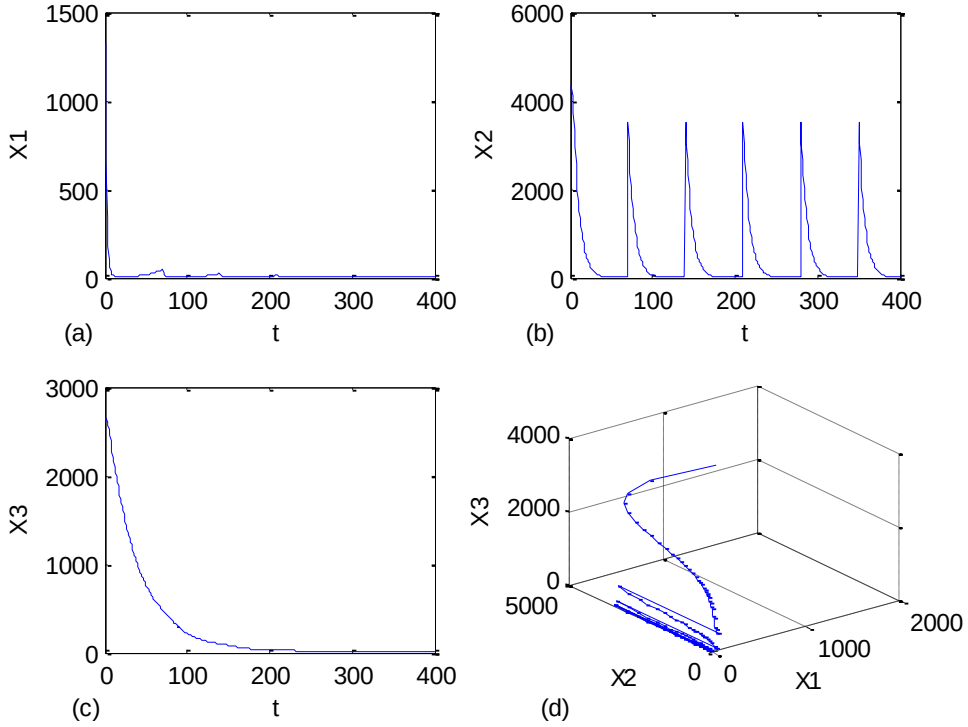


Fig. 2. Evolution of the egg (a), parasitized egg (b), larvae populations (c) and phase portraits (d) of system (2) for $p = 3500$ parasitoids/ha

Choosing the release amount $p = 1500$ parasitoids/ha, we can control the larvae population below the economic injury level (see Fig. 3). It is obvious that the cost of the control strategy $p = 1500$ is less than the cost of $p = 3500$.

Applying the control strategy $p = 1000$, we can see that the number of larvae individuals exceed x_E at some time (see Fig. 4). For this control the system experiences chaotic behavior which is showed in Fig. 4 (d).

7. Conclusion

In this paper, we suggest a system of impulsive differential equations to model the process of the biological control of the sugarcane borer by periodically releasing of parasitoids. By using the Floquet theory and small amplitude perturbation method, we have proved that for any fixed period τ there exists a locally

 ISBN: 978-84-614-6167-7

asymptotically stable pest-eradication periodic solution $(0, \tilde{x}_2(t), 0)$ of the system (2) if the number of the parasitoids in periodic releases is greater than some critical value $p_{\min}$.

When the stability of the pest-eradication periodic solution is lost, the numerical results show that the system (2) has rich dynamics, including chaotic behavior.

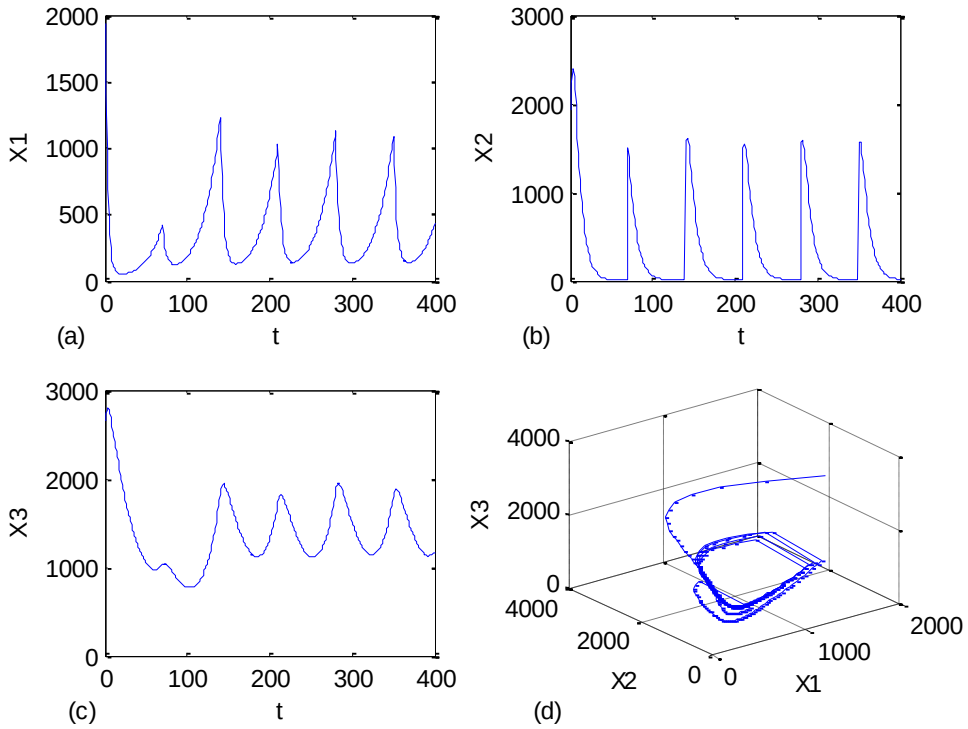


Fig. 3. Evolution of the egg (a), parasitized egg (b), larvae populations (c) and phase portraits (d) of system (2) for $p = 1500$ parasitoids/ha

If we choose the biological control strategy by periodical releases of the constant amount of parasitoids, the results of Theorem 3.1 can help in design of the control strategy by informing decisions on the timing of parasitoid releases. In this case, from (12) we have:

$$\tau < \tau_{\max} = \frac{\beta p}{(r - m_1 - n_1)(m_2 + n_2)} \tag{21}$$

From (21) we can conclude that there exists a locally asymptotically stable pest-eradication periodic solution $(0, \tilde{x}_2(t), 0)$ of the system (2) if the impulsive period is less than some critical value $\tau_{\max}$.

 ISBN: 978-84-614-6167-7

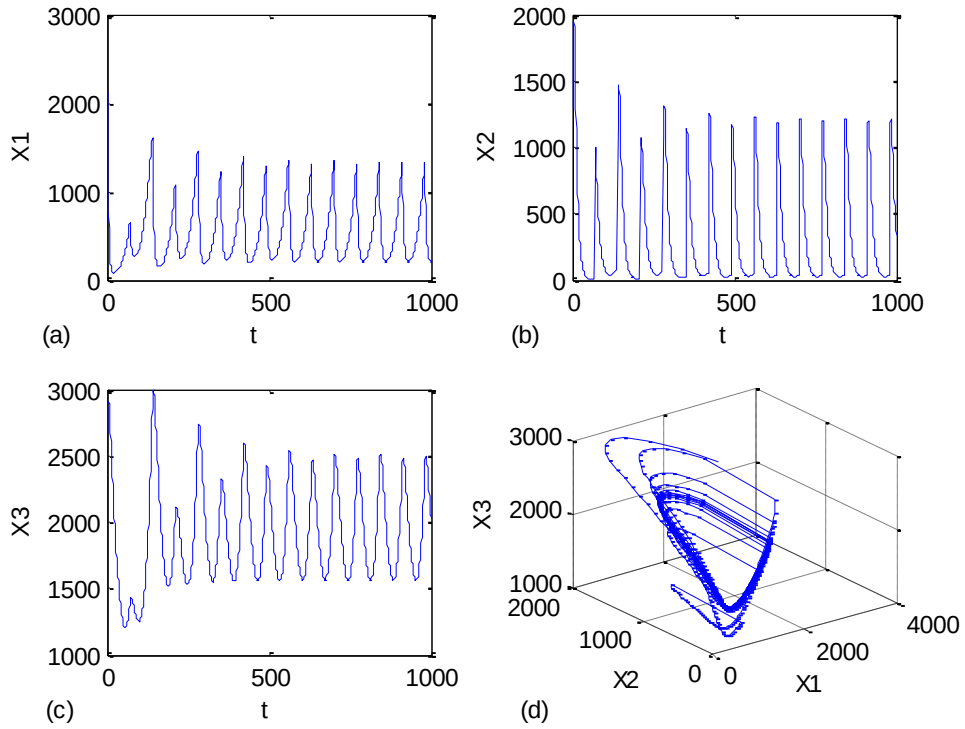


Fig. 4. Evolution of the egg (a), parasitized egg (b), larvae populations (c) and phase portraits (d) of system (2) for $p = 1000$ parasitoids/ha

Thus, the results of the present study show that the impulsive release of parasitoids provides reliable strategies of the biological pest control of the sugarcane borer.

Acknowledgments

The authors thank Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) and Conselho Nacional de Pesquisas (CNPq) for the financial supports on this research.

References

[1] J.R.P. PARRA, P.S.M. BOTELHO, B.S.C. FERREIRA, J.M.S. BENTO, *Controle Biológico no Brasil: Parasitóides e Predadores*, São Paulo, Editora Manole, 2002.

[2] N. MACEDO AND P.S.M. BOTELHO, *Controle integrado da broca da cana-de-açúcar, Diatraea saccharalis (Fabr., 1794)(Lepidoptera:Pyralidae)*, Brasil Açucar. **106** (1988) 2-14.

 ISBN: 978-84-614-6167-7

[3] P. DEBACH, *Biological control by natural enemies,* Cambridge, Cambridge University Press, 1974.

[4] H.J. BARCLAY, I. S. OTVOS AND A.J. THOMSON, *Models of periodic inundation of parasitoids for pest control*, Can. Ent. **117**(1970)705-716.

[5] P.S.M. BOTELHO, J.R.P. PARRA, J.F. CHAGAS NETO, C.P.B OLIVEIRA, *Associação do parasitóide de ovos Trichogramma galloi Zucchi (Himenóptera: Trichogrammatidae) e do parasitóide larval Cotesia flavipes (Cam.)(Himenóptera: Brancconidae) no controle de Diatraea saccharalis (Fabr., 1794)(Lepidoptera:Pyralidae) em cana-de-açúcar*, Anais da Sociedade Entomologica do Brasil, 28(3)(1970)491-496.

[6] N.J. MILLS AND W.M. GETZ, *Modeling the biological control of insect pest: review of host–parasitoid models*, Ecological Modelling **93** (1996) 121–143.

[7] M. RAFIKOV, AND E.H. LIMEIRA, *Mathematical Modeling of the Biological Pest Control of the Sugarcane Borer*, Proceedings of the 10th International Conference on Computational and Mathematical Methods in Science and Engineering, CMMSE2010, Almeria, Spain, 27-30 June 2010, pp. 1228-1242.

[8] Z. LU, X. CHI AND L. CHEN, *Impulsive control strategies in biological control of pesticide*, Theoretical Population Biology **64** (2003) 39-47

[9] P. GEORGESCU AND G. MOROSANU, *Pest regulation by means of impulsive control*, Applied Mathematics and Computation, **190** (2007) 790-803.

[10] S. TANG AND R.A. CHEKE, *Models for integrated pest control and their biological implications*, Mathematical Biosciences **215** (2008) 115-125.

[11] H.K. BAEK, S.D. KIM AND P. KIM, *Permanence and stability of an Ivlev-type predator-prey system with impulsive control strategies*, Mathematical and Computer Modelling **50** (2009) 1385-1393.

[12] J. JIAO, L. CHEN AND S. CAI, *Impulsive control strategy of pest management SI model with nonlinear incidence rate*, Applied Mathematical Modelling **33** (2009) 555-563.

[13] S. NUNDLOLL, L. MAILLERET AND F. GROGNARD, *Two models of interfering predators in impulsive biological control*, Journal of Biological Dynamics **4(1)** (2010) 102-114.

[14] D. BAINOV AND P. SIMEONOV, *Impulsive differential equations: periodic solutions and applications*, vol., 66, Longman, England,1993.

 ISBN: 978-84-614-6167-7

Proceedings of the 11th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE2011
Beenidorm, Spain, 26-30 June 2011

Solving Nonlinear Equations by a Tabu Search Strategy

Gisela C.V. Ramadas[1] and Edite M.G.P. Fernandes[2]

[1] *Department of Mathematics, Polytechnic Institute of Porto, 4200-072 Porto, Portugal*
[2] *Algoritmi R&D Center, University of Minho, 4710-057 Braga, Portugal*
emails: gcv@isep.ipp.pt, emgpf@dps.uminho.pt

Abstract

Solving systems of nonlinear equations is a problem of particular importance since they emerge through the mathematical modeling of real problems that arise naturally in many branches of engineering and in the physical sciences. The problem can be naturally reformulated as a global optimization problem. In this paper, we show that a metaheuristic, called Directed Tabu Search (DTS) [16], is able to converge to the solutions of a set of problems for which the *fsolve* function of MATLAB® failed to converge. We also show the effect of the dimension of the problem in the performance of the DTS.

Key words: nonlinear equations, metaheuristic, tabu search

1 Introduction

The numerical solution of some problems in engineering, chemistry, physics, medicine and economic areas, aims at determining the roots of a nonlinear system of equations. The modeling of these problems can lead to simple and almost linear systems, but mostly large, complex and difficult to solve systems of nonlinear equations arise. In this paper we consider solving the nonlinear system of equations

$$F(x)=0 \tag{1}$$

where

$$F:\Omega\subset\Re^n\rightarrow\Re^n,\quad \Omega\equiv[l,u]=\left\{x\in\Re^n : l_i\le x_i\le u_i, i=1,...,n\right\},$$
$$F(x)=(f_1(x),...,f_n(x))^T$$

and the functions

$$f_i(x),\quad i=1,...,n$$

 ISBN: 978-84-614-6167-7

are continuously differentiable, using a metaheuristic. Probably the most famous techniques are based on Newton's method [2,6,8,12,24,29]. They require analytical or numerical first derivative information. Newton's method is the most widely used algorithm for solving nonlinear systems of equations. It is computationally expensive, in particular if n is large, since a system of linear equations is required to be solved at each iteration. The Quasi-Newton methods use less expensive iterations than Newton, but their convergence properties are not very different. In general, Quasi-Newton methods avoid either the necessity of computing derivatives, or the necessity of solving a full linear system per iteration or both tasks [25]. In [13], a new technique for solving systems of nonlinear equations reshaping the system as a multiobjective optimization problem in proposed. The authors applied a technique of evolutionary computation to solve the problem obtained after the change. In [14], the authors propose techniques for computing all the multiple solutions in nonlinear systems. Another technique to solve systems of nonlinear equations is presented in [18], where a heuristic continuous global optimization GRASP is applied. A genetic algorithm is proposed in [4].

The problem of solving a nonlinear system of equations can be naturally formulated as a global optimization problem. Problem (1) is equivalent, in the sense that it has the same solution, to finding the globally smallest value of the l_2-norm error function, related to solving the system of equations (1), defined by

$$\min_{x\in\Omega\subset\Re^n} \Psi(x) \equiv \left\| F(x) \right\|_2 . \tag{2}$$

Here, the global minimum, and not just a local minimum, of the objective function $\Psi(x)$, in the set Ω, is to be found. The classical local search methods, like Newton-type methods, have some disadvantages, when compared to global search methods. In particular

i) the final solution is heavily dependent on the the starting point of the iterative process;
ii) they can be trapped in local minima;
iii) they require differentiable properties of all the equations of nonlinear system.

We use the Example 1 below to show this local trap behavior.

Example 1: Consider the following system of nonlinear equations

$$f_1(x_1, x_2) = x_1 - \sin(2x_1 + 3x_2) - \cos(3x_1 - 5x_2) = 0$$
$$f_2(x_1, x_2) = x_2 - \sin(x_1 - 2x_2) + \cos(x_1 + 3x_2) = 0.$$

Figure 1 shows the graphical representation of the l_2-norm error function $\Psi(x)$. The multi-modal nature of $\Psi(x)$ makes the process of detecting a global minimum a difficult one. Nine different starting points were used to solve Example 1 by *fsolve* from MATLAB®. In this MATLAB function, the default trust-region dogleg algorithm with no analytical Jacobian is used. Although all

 ISBN: 978-84-614-6167-7

the nine starting points are in the neighborhood of the solution, the method converges to the required solution only twice.

Figure 1. Graphical representation of $\Psi(x)$ for Example 1.

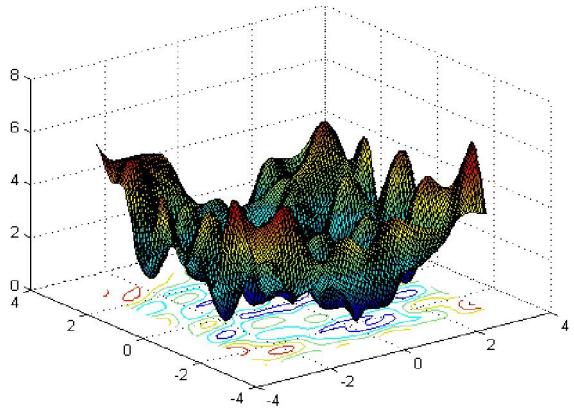


Table 1 shows the results obtained from MATLAB. The first column in the table presents the tested starting points, $x^{(0)}$, as well as the value of the output parameter "exitflag" of MATLAB. The value "1" means that the method converged to a root where the first-order optimality measure is less than a pre-specified tolerance, and "-2" means that it converged to a point which is not a root where the sum-of-squares of the function values is greater than or equal to a pre-specified tolerance.

Table 1. Solutions obtained by *fsolve* from MATLAB for different starting points.

Starting point $x^{(0)}$ /"exitflag"	(f_1, f_2) at solution	n. of iterations	n. of function evaluations
(0, 0) / "1"	(-0.41e-12, -0.24e-12)	5	18
(1, 1) / "1"	(0.36e-8, 0.23e-8)	6	21
(0, 1) / "-2"	(-0.03, 0.94)	27	60
(2, 2) / "-2"	(0.11, 0.88)	31	70
(-1, 1) / "-2"	(-0.55, 0.65)	55	130
(1, -1) / "-2"	(0.01, -0.04)	24	57
(-1, -1) / "-2"	(-0.52, 0.10)	31	74
(2, -2) / "-2"	(-0.16, -0.57)	32	71
(-2, -2) / "-2"	(-0.17, -1.53)	34	77

Thus, to be able to converge to a global solution, a global search strategy is required. The most important global search techniques invoke exploration and exploitation search procedures aiming at:

i) diversifying the search in all the search space;
ii) intensifying the search in promising areas of the search space.

A well-known class of global search techniques, the metaheuristics, use random procedures that invoke artificial intelligence tools and simulate nature

 ISBN: 978-84-614-6167-7

behaviors. The word "metaheuristics" is used to describe all heuristics methods that are able to achieve a good quality solution in a reasonable time. Due to their random features, metaheuristics have, in general, slow convergence since they may fail to detect promising search directions in the neighborhood of a global minimum.

There are two classes of metaheuristics. A population-based heuristic defines and maintains throughout the iterative process a set of solutions. The most known population-based heuristic is the Genetic Algorithm [10]. A point-to-point heuristic defines just one solution at the end of each iteration which will be used to start the next iteration. Simulated Annealing (SA) [15] and Tabu Search [9] are two examples of point-to-point methods.

The Tabu Search (TS) is a metaheuristic developed primary for solving combinatorial problems [9]. The TS introduced by Cvijović and Klinowski [5] for continuous optimization guides the local search out of local optima and has the ability to explore new regions. It is an iterative procedure that maintains a list of the movements most recently made, avoiding in subsequent iterations the execution of movements that lead to solutions already known to have been visited. Usually, the slow convergence of TS is overcame by incorporating a classical local search strategy into the main algorithm. In general, this type of hybridization occurs in the final stage of the iterative process when the solution is in the vicinity of the global solution. An example of such method is presented in [16]. The therein proposed method, called Directed Tabu Search (DTS), uses strategies, like the Nelder-Mead method [28] and the Adaptive Pattern Search [15], to direct a tabu search.

This paper aims at assessing the performance of a tabu search method when solving a system of nonlinear equation (1), using the function $\Psi(x)$ as a measure of the progress of the algorithm towards the solution. According to the formulation (2), this means that the fitness of each trial solution x is assessed by evaluating the function Ψ at x. And a solution $\tilde{x}$ is better than x if $\Psi(\tilde{x}) < \Psi(x)$. In this paper, and due to the reported success when solving global optimization problems of the form (2), the DTS variant of the tabu search is extended to be able to solve nonlinear systems of equations. In particular, we aim at analyzing the behavior of the extended version of the DTS method when solving some difficult problems that are not solved by Newton-type methods.

The organization of the paper is as follows. Section 2 describes the Directed Tabu Search method and the Section 3 reports the computational experiments. Finally, we summarize our conclusions in Section 4.

2 The Metaheuristic Tabu Search

2.1 Basic Tabu Search

TS is an iterative process which operates in the following way. The algorithm starts with a randomly generated initial solution, x, and by applying pre-defined

 ISBN: 978-84-614-6167-7

moves in its neighborhood it generates a set Y of solutions. The objective function to be minimized is evaluated at all solutions in Y, and the best of all, y, becomes the current solution, $x \leftarrow y$ (even if it is worse than x). Accepting uphill moves, the algorithm avoids to get trapped in a local mínimum. The previous procedure is repeated until a given stopping condition is reached. Further, the algorithm also stops when the solution does not improve for *Nmax* iterations. To avoid cycling, since a point already visited may be generated again, a set of points already visited are stored in a list, called Tabu List (TL). The solutions in Y that belong to the TL are eliminated. This TS structure is called short-term memory TS. The use of this type of flexible memory is of great advantage in contrast to the rigid structures of large memory, like those present in Branch-and-Bound methods, or the lack of memory that exists in the Simulated Annealing method [3]. To improve performance, long-term memory TS structures have been proposed to record important attributes like elite and frequently visited solutions. The Directed Tabu Search [16] implemented in this paper for solving nonlinear systems of equations contains long-term memory structures. This is important since the method is able to keep diversity, like population-based methods.

2.2 Directed Tabu Search

The Directed Tabu Search method of Hedar and Fukushima [16] uses direct search methods in order to stabilize the search especially in the vicinity of a local minimum. Two variants of the DTS are therein proposed: one is based on the Nelder-Mead (NM) method, as a local search, inside the exploration step of the algorithm, and the other uses the Adaptive Pattern Search (APS) strategy in the exploration step. Furthermore, the Kelley´s modification of the NM method [19] is still used in the therein called intensification search in the final stage of the process. We note that the DTS method can be classified as a multi-start method. The multi-start methods are designed to build powerful search procedures and guided by a global exploration and local search. These multi-start methods have been successfully applied either in nonlinear global optimization problems or in combinatorial optimization.

The DTS method is based on three procedures: exploration, diversification and intensification search procedures. The structure of the DTS is shown below in Algorithm 1.

Algorithm 1 (DTS method)

Step 1 Randomly generate an initial solution
Step 2 Exploration search procedure
  Step 2.1 Neighborhood search
  Step 2.2 Local search
  Step 2.3 Solution update
  Step 2.4 If *improvement criteria* are reached, repeat from *Step 2.1*
Step 3 Diversification search procedure
Step 4 If *stopping criteria* are not reached, repeat from *Step 2*
Step 5 Intensification search procedure

 ISBN: 978-84-614-6167-7

The main loop (outer cycle) of the DTS method, consisting of the exploration and diversification search procedures, begins with an initial solution. A brief description of the procedures follows. Other mathematical details can be found in [16].

2.2.1 Exploration Search

The exploration search aims to explore the solution space. It uses direct search methods as neighborhood search and local search strategies to generate trial points. These may be based on either the simplex method of NM or on the APS strategy. Here, we use the APS variant. This variant of the DTS mainly focuses its strategy on the definition of an approximate descent direction (ADD), v, for the fitness Ψ, as proposed in [15]. Thus, based on a set of n trial points, y_i, $i=1,...,n$, using the standard pattern directions, in the neighborhood of the current solution x, the descent direction is computed as follows:

$$v=\sum_{i=1}^{n} w_i u_i \quad \text{where} \quad w_i=\frac{\Psi(y_i)-\Psi(x)}{\sum_{j=1}^{n}\left|\Psi(y_i)-\Psi(x)\right|} \quad \text{and} \quad u_i=-\frac{(y_i-x)}{\|y_i-x\|},\ i=1,...,n.$$

Briefly, pattern directions are constructed parallel to the coordinate axes, and a set of points are generated in the neighborhood of the current solution x, along these directions with appropriate step length.

Moreover, anti-cycling is prevented not only with the standard Tabu List but also with the inclusion of novel Tabu Regions (TR). The DTS method implements four new TS memory elements.

- The multi-ranked Tabu List (r-TL) is a set of some visited solutions that are ranked and saved according to two features separately, namely: i) ranked in ascending order according to their recency, ii) ranked in ascending order according to their Ψ values.
- Further, two types of regions, around each solution saved in the r-TL are defined:
  - the Tabu Region (TR) in which no new trial point is allowed to be generated,
  - the Semi-Tabu Region (STR), which is a surrounding region around TR, with a radius from its center greater than the radius of the TR, that aims to allow the generation of neighboring trial points when the trial solution lies inside STR.
- Finally, the other memory element is the Visited Region List (VRL) that contains information concerned with the centers of the visited regions and the frequency of visiting these regions. This information is crucial to explore the space outside these visited regions.

These new long-term memory structures are very important since they allow the method in the diversification search and intensification search procedures to behave as an intelligent search technique.

 ISBN: 978-84-614-6167-7

The exploration search procedure is repeatedly applied in order to generate n trial solutions using search strategies in the neighborhood of the current solution. If a better move between these raised solutions is found, the current solution is updated and the algorithm continues to the next iteration of this inner cycle (exploration search procedure). Otherwise, the current solution is still better than all the raised solutions in the neighborhood, and the local search strategy generates new local trial points and the current solution is updated with the best trial point generated. The multi-ranked Tabu List is updated and if a new region is reached, the Visited Regions List is updated with information about this region. When the number of iterations in the inner cycle exceeds a pre-specified value or improvement has not been obtained in some consecutive iterations (the *improvement criteria* in *Step 2.4* of Algorithm 1), the diversification search procedure is applied to locate a new starting point, from which the exploration search procedure is repeated.

2.2.2 Diversification Search

A diversification procedure aims to generate a new initial trial point outside the visited regions. The information stored in the VRL is used to direct the search towards new regions. The VRL serves as a diversification tool in the search, with the aim of diversifying the search for areas that have not been visited in the solution space.

2.2.3 Intensification Search

When one of the best obtained trial solutions is sufficiently close to a global minimum, or its value has not been changed for *Nmax* iterations (*stopping criteria* in the *Step 4* of Algorithm 1), then the intensification search procedure is applied at the final stage to refine the best solution visited so far. In this case, a local direct search method, the Kelley´s modification of the Nelder-Mead method [19] and [20], is used. A solution still closer to the global minimum is then obtained.

3 Computational Experiments

We selected, and coded in MATLAB®, 99 test problems from the literature and solved them using *fsolve* from MATLAB, the extended version of the DTS method [16], and a SA version for comparative purposes. The problems in our database are referred to as P1, P2, …, P99. They represent systems of nonlinear equations of different sizes and complexity. The results of the numerical experiments were obtained in a personal computer with an AMD Turion 2.20 GHz processor and 3 GB of memory. Due to the stochastic nature of DTS and SA algorithms, each problem was run 30 times and the best of the 30 solutions, as well as the average of the 30 obtained solutions were registered. The DTS algorithm starts with an initial solution, which is randomly generated inside the range $[l,u]$. Although not all problems defined in the literature have a registered

 ISBN: 978-84-614-6167-7

$[l,u]$, we selected a specific range for each problem, as shown in Table 4 in the Appendix. The lower and upper limits are the same for all the components of the solution. The parameters defined for the *stopping criteria* of the algorithm are $Nmax = 10\,n$ and $\varepsilon = 10^{-8}$ (in *Step 4* of Algorithm 1), where the condition that defines the closeness of the best solution to the global minimum is the following

$$\Psi(x_k) \leq \varepsilon .$$

The goal of this paper is twofold:

i) to evaluate the effectiveness of the extended DTS to solve systems of nonlinear equations that are not solved by Newton´s method;

ii) to analyze the effect of the dimension of the problem in the performance of the extended DTS algorithm.

Table 2. Comparison of *fsolve*, DTS and SA

		fsolve			DTS			SA		
	n	$\Psi(x)$	*nfe*	*nit*	Ψ_{avg}	*a-nfe*	Ψ_{min}	Ψ_{avg}	*a-nfe*	Ψ_{min}
P6	2	**0.0025**	57	20	0.5747	327	0.019	0.6677	265	0.0715
P9	2	6.9989	49	20	3.2664	242	1.8e-04	4.1994	243	**2.9e-05**
P25	10	13.830	1001	90	0.1227	5101	0.0142	0.4112	1014	**0.0049**
P26	3	0.6642	106	30	0.4195	482	**9.5e-06**	0.2428	303	1.3e-05
P27	3	0.1008	109	30	0.0563	667	**0.0011**	0.0669	339	0.0015
P44	33	2.1965	1170	47	1.5666	75246	**0.6737**	3.5248	3302	0.7079
P45	33	**0.3983**	3306	104	5.7368	48664	3.5176	6.1427	3302	4.2613
P50	2	**0.2476**	47	20	0.24	247	**0.2476**	0.2476	240	**0.2476**
P52	5	0.0022	261	50	4.0e-02	1556	5.0e-04	8.2e-02	507	**2.0e-04**
P53	6	2.8074	427	60	0.0015	1958	**0.0001**	0.0019	604	**0.0001**
P55	6	138.42	385	60	4.6e+06	3634	**122.84**	8.0e+06	631	129.65
P56	8	0.4919	609	80	1.1e+01	4355	**0.4286**	1.1439	802	0.4308
P62	6	**0.0467**	379	60	5.4e+10	4512	45.353	5.1e+11	616	44.31
P63	6	0.0062	427	60	0.0020	2030	**5.4e-05**	0.0177	602	0.0002
P64	8	3.2592	657	80	0.4907	3216	**5.1e-05**	3.1581	887	0.0003
P70	10	25.782	981	100	53.5830	5870	**9.0034**	47.1696	1002	**9.0034**
P72	51	2.7307	1989	50	1.9574	151108	**1.1470**	2.0317	5102	1.4808
P77	2	4.5632	45	20	0.8533	243	**0.7377**	1.0068	211	**0.7377**
P80	3	0.0630	124	30	2.0e-05	516	5.2e-06	0.1139	325	**3.9e-06**
P84	3	0.0005	124	30	3.1e-08	482	1.4e-09	5.0e-08	144	**1.2e-09**
P88	10	0.7460	66	5	8.6e-03	4773	1.4e-03	8.8e-03	1002	**7.7e-04**
P96	2	2.8098	47	20	0.1040	258	**1.7e-05**	0.4617	265	1.8e-05
P99	2				0.0600	252	**4.0e-06**	9.6e-02	292	1.1e-05

To resume the main achievements of our numerical experiments, we show in Table 2 the results of the 23 problems (from the database of 99 problems) that

 ISBN: 978-84-614-6167-7

were selected because *fsolve* was not able to converge to the required solution with a tolerance of $\varepsilon = 10^{-8}$ and within $10\,n$ iterations ("exitflag" is "-2" for P44 and P72 and is "0" for all the others). Here, we use the initial approximation provided in the literature. Problem P99 of our list is the Example 1 described in Section 1.

The characteristics of the selected problems are listed in the Appendix of the paper (Table 4). Moreover, the results of a selection of other problems from our database with varied dimensions are shown in Table 3.

Table 3. Comparison based on problems with varied dimensions

		fsolve			DTS			SA		
	n	$\Psi(x)$	*nfe*	*nit*	Ψ_{avg}	*a-nfe*	Ψ_{min}	Ψ_{avg}	*a-nfe*	Ψ_{min}
P14	2	**2.8e-08**	19	6	0.0092	251	4.4e-06	4.6e-05	287	2.3e-06
	10	**3.7e-08**	55	4	0.93	4918	0.0023	0.8545	1022	0.0031
	30	**8.6e-14**	155	4	5.8	50073	1.9910	5.2442	3002	1.3550
P15	2	**0**	15	4	4.9e-05	246	5.6e-07	3.5e-05	258	1.6e-06
	10	**7.2e-07**	101	10	0.7422	4731	4.7e-05	20.1567	1040	2.7e-05
	30	**4.1e-08**	311	10	2.220	39478	0.0023	1.0e+07	3257	26.4957
P16	2	**3.5e-07**	9	2	4.1e-05	250	8.8e-06	3.2e-05	232	3.8e-06
	10	**3.1e-08**	33	2	0.006	4915	0.0006	0.005	1002	0.0009
	30	**5.1e-05**	62	1	0.25	44379	0.1201	0.2868	3002	0.0934
P17	2	**3.5e-07**	9	2	3.0e-05	249	3.7e-06	1.6e-05	246	1.9e-06
	10	**3.7e-07**	33	2	4.8e-04	3978	8.3e-05	0.0003	1002	7.3e-05
	30	**6e-07**	93	2	0.015	36010	0.0012	0.0045	3002	0.0009
P18	2	**5e-09**	15	4	2.6e-02	252	1.4e-06	3.8e-05	223	1.1e-06
	10	**1.1e-09**	55	4	1.00	5067	0.6301	1.0404	1002	0.9375
	30	**4.2e-14**	186	5	1.2	40639	1.0569	1.322	3002	1.0568
P19	2	**2.6e-08**	18	5	4.3e-05	259	6.8e-06	7.9e-05	212	1.0e-05
	10	**1.5e-08**	66	5	1.6	5108	1.7e-05	1.6027	1002	0.0073
	30	**5e-12**	217	6	1.9	50845	0.0026	1.753	3002	1.753
P21	4	**8.1e-15**	41	8	4.4e-04	962	1.7e-04	6.7e-04	426	6.8e-05
	52	**1.7e-14**	424	7	2.8884	156373	0.5985	6.8724	5202	3.4877
P22	2	**0.0001**	6	1	8.4128	296	0.0722	18023.3	314	0.0793
	50	**0.0005**	102	1	6.5e+07	98182	1.1e+04	3.3e+05	5002	7.5e+03
P26	3	**0.6642**	106	30	0.4195	482	9.5e-06	0.2428	303	1.3e-05
	33	**2.1965**	1170	47	1.5666	75146	0.6737	3.5248	3302	0.7079

Tables 2 and 3 summarize the numerical results, where n represents the dimension (number of variables = number of functions in the system), $\Psi(x)$ is the value of the l_2-norm error function at the found solution, *nfe* is the number of function evaluations and *nit* the number of iterations, all provided by MATLAB; Ψ_{avg} represents the average of the obtained solutions over the 30 runs, *a-nfe*

 ISBN: 978-84-614-6167-7

gives the average number of function evaluations computed over the 30 runs (rounded to the nearest integer) and Ψ_{min} is the best solution obtained during the 30 runs. The best solutions found in these comparisons are printed in "bold". We compare $\Psi(x)$ of *fsolve* with Ψ_{min} of DTS and SA.

4 Conclusions

In this paper we show that nonlinear systems of equations can be effectively solved by implementing a global optimization method to a fitness function, which represents the l_2-norm error function related to the solving of the system of equations. The application of an extended version of the metaheuristic Directed Tabu Search, proposed in [16], for solving complex and difficult nonlinear systems of equations has been analyzed and tested. From Table 2, we may conclude that DTS is mostly able to converge to the solutions of the selected systems that are not solved by a Newton-type method. However, the results of Table 3 are not so promising. As expected, when a Newton-type method converges, the accuracy of the found solution is higher than that obtained by a metaheuristic. Furthermore, the performance of both tested metaheuristics (DTS and SA) is greatly affected by the dimension of the problem. This suggest that a combination of global- and local-type search procedures, carried out in separate iterations, depending on the need for an exploration of the search space or an exploitation of a promising region in the search space, will improve performance. For the local-type iteration, a derivative local search method seems crucial, as long as analytical or numerical derivatives could be used. This issue will be addressed in the near future.

Appendix – Problems used in numerical experiments

Table 4. Characteristics of the problems

	n	[*l*,*u*]	reference	Problem name in cited paper
P6	2	[-10,10]	[27]	P3-*Powell badly scaled function*
P9	2	[-10,15]	[27]	P2-*Freudenstein and Roth function*
P14	2,10,30	[-10,10]	[27]	P26-*Trignometric function*
P15	2,10,30	[-10,10]	[27]	P27-*Brown almost-linear function*
P16	2,10,30	[-10,10]	[27]	P28-*Discrete boundary value function*
P17	2,10,30	[-10,10]	[27]	P29-*Discrete integral equation function*
P18	2,10,30	[-10,10]	[27]	P30-*Broyden tridiagonal function*
P19	2,10,30	[-10,10]	[27]	P31-*Broyden banded function*
P21	4,52	[-5,5]	[8]	D2-*Augmented Rosenbrock*
P22	2,50	[0,370]	[8]	D3-*Powell badly scaled*
P25	10	[10,50]	[8]	D6-*Shifted and augmented trigonometric function with an Euclidian sphere*
P26/P44	3/33	[-10,10]	[8]	D7-*Diagonal of three variables premultiplied by a quasi-orthogonal matrix*
P27/P45	3/33	[-10,10]	[8]	D8-*Diagonal of three variables premultiplied by an orthogonal matrix, combined with inverse trigonometric function*
P50	2	[-10,10]	[18]	pp. 2004
P52	5	[-20,20]	[26]	*Combustion of propane chemical equilibrium equations*
P53	6	[-3,3]	[27]	14-*Wood function*
P55	6	[-10,10]	[29]	*Semicoductor boundary condition*
P56	8	[-10,30]	[1]	2.3-*The human heart dipole*

 ISBN: 978-84-614-6167-7

Table 4. Characteristics of the problems (cont.)

	n	$[l,u]$	reference	Problem name in cited paper
P62	6	[0,60]	[17]	*Problem 2*
P63	6	[-10,10]	[23]	*Example 2*
P64	8	[-10,15]	[7]	*Equation 3.1*
P70	10	[-100,100]	[30]	*Example 4.1-Nonlinear resistive circuit*
P72	51	[-5,5]	[8]	D7-*Diagonal of three variables premultiplied by a quasi-orthogonal matrix*
P77	2	[0,3.5]	[4]	*Example 1*
P80	3	[0,4]	[27]	5-*Beale function*
P84	3	[0,1]	[22]	*Example 6.2*
P88	10	[-10,10]	[21]	*Example 2-The Beam problem*
P96	2	[-10,10]	[12]	pp. 498 (*Problem N4*)

Acknowledgement The first author is supported by the CIDEM - Centre for Research & Developement in Mechanical Engineering, from Portugal.

References

[1] B.M. AVERICK, R.G. CARTER AND J.J. MORÉ, *The MINPACK-2 test problem collection* (*Preliminary version*), Technical Memorandum n. 150, Mathematics and Computer Science Division, Argonne National Laboratory, Argonne, IL, 1991.

[2] E. BODON, A. DEL POPOLO, L. LUKŠAN AND E. SPEDICATO, *Numerical performance of ABS codes for systems of nonlinear equations*, Technical Report DMSIA 01/2001, Universitá degli Studi di Bergamo, Bergamo, Italy, 2001.

[3] L. CAVIQUE, C. REGO AND I. THEMIDO, *Estruturas de vizinhança e procura local no problema da clique máxima*, Investigação Operacional, **22**, (2002) 1-18.

[4] C.H. CHEN, *Finding roots by genetic algorithms*, In: 2003 Joint Conference on AI, Fuzzy System and Gray System, Taipei, Taiwan, (2003) 4-6.

[5] D. CVIJOVIĆ AND J. KLINOWSKI, *Taboo search: an approach to the multiple minima problem,* Science, **267**, (1995) 664-666.

[6] J.E. DENNIS AND R.B. SCHNABEL, *Numerical methods for unconstrained optimization and nonlinear equations*, Prentice-Hall, Inc., 1983.

[7] M. DRĂGAN, *On solving large sparse systems of nonlinear equationsusing threads*, In: Proceedings of NMCM2002, An Euro Conference on Numerical Methods and Computational Mechanics, Miskolc, Hungary, (2002) 49-56.

[8] A. FRIEDLANDER, M.A. GOMES-RUGGIERO, D.N. KOZAKEVICH, J.M. MARTINEZ AND S.A. SANTOS, *Solving nonlinear systems of equations by means of Quasi-Newton methods with a nonmonotone strategy*, Optimization Methods and Software, **8**, (1997) 25-51.

[9] F. GLOVER, *Future paths for integer programming and links to artificial intelligence*, Computers and Operations Research **13**, n. 5, (1986) 533-549.

[10] F. GLOVER AND M. LAGUNA, *Tabu Search,* Kluver Academic Publishers, 1997.

[11] D.E. GOLDBERG, *Genetic Algorithms in Search, Optimization \& Machine Learning,* Addison-Wesley (1989).

[12] M.D. GONZÁLEZ-LIMA AND F.M. OCA, *A Newton-like method for nonlinear system of equations*, Numer. Algor. **52** (2009) 479-506.

[13] C. GROSAN AND A. ABRAHAM, *A new aproach for solving nonlinear equations systems*, IEEE Transactions on systems, man, and cybernetics – part A: systems and humans, **38**, n. 3, (2008) 698-714.

 ISBN: 978-84-614-6167-7

[14] C. GROSAN AND A. ABRAHAM, *Multiple solutions for a system of nonlinear equations*, International Journal of Innovative Computing, Information and Control, (2008).
[15] A. HEDAR AND M. FUKUSHIMA, *Heuristic Pattern Search and its Hybridization with Simulated Annealing for Nonlinear Global Optimization*, Optimization Methods and Software, **19**, (2004) 291-308.
[16] A. HEDAR AND M. FUKUSHIMA, *Tabu search direct by direct search methods for nonlinear global optimization*, European Journal of Operational Research, **170**, (2006) 329-349.
[17] K.L. HIEBERT, *An evaluation of mathematical software that solves systems of nonlinear equations*, ACM Transactions on Mathematical Software, **8**, n.1, (1982) 5-20.
[18] M.L. HIRSCH, P.M. PARDALOS AND M. RESENDE, *Solving systems of nonlinear equations with continuous GRASP*, Nonlinear Analysis: Real World Applications, **10,** (2009) 2000-2006.
[19] C.T. KELLEY, *Detection and remediation of stagnation in the Nelder-Mead algorithm using a sufficient decrease condition*, SIAM Journal on Optimization, **10**, (1999) 43-55.
[20] C.T. KELLEY, *Iterative Methods for Optimization*, Frontiers Appl. Math., Vol. 18, SIAM, Philadelphia, PA, 1999.
[21] C.T. KELLEY, L. QI, X. TONG AND H. YIN, *Finding a stable solution of a system of nonlinear equations arising from dynamic systems,* Journal of Industrial and Management Optimization, **7**, n. 2, (2011) 497-521.
[22] L.N. KOLEV, *Interval methods for circuit analysis*, World Scientific, Singapore, 1993.
[23] L.V. KOLEV, *An improved interval linearization for solving nonlinear problems*, Numerical Algorithms, **13**, n. 1-4, (2004) 213-224.
[24] J.M. MARTINEZ, *Algorithms for solving nonlinear systems of equations,* In Continuous Optimization: the state of art, edited by E. Spedicato. Kluwer Academic Publishers, (1994) 81-108.
[25] J.M. MARTINEZ, *Practical quasi-Newton methods for solving nonlinear systems*, Journal of Computational and Applied Mathematics, **124**, (2000) 97-122.
[26] K. MEINTJES AND A.P. MORGAN, *Chemical equilibrium systems as numerical test problems*, ACM Transaction on Mathematical Software, **16**, n. 2 (1990) 143-151.
[27] J.J. MORÉ, B.S. GARBOW AND K.E. HILLSTROM, *Testing unconstrained optimization software*, ACM Transactions on Mathematical Software, **7**, n. 1, (1981) 17-41.
[28] J.A. NELDER AND R. MEAD, *A simplex method for function minimization*, Computing Journal, **7**, (1965) 308-313.
[29] U. NOWAK AND L. WEIMANN, *A family of Newton codes for systems of highly nonlinear equations*, Technical Report. TR-91-10, 1991.
[30] K. YAMAMURA, H. KAWATA AND A. TOKUE, *Interval solution of nonlinear equations using linear programming*, BIT – Numerical Mathematics, **38**, n. 1, (1998) 186-199.

 ISBN: 978-84-614-6167-7

Proceedings of the 11th International Conference on Computational and Mathematical Methods in Science and Engineering, CMMSE2011 Benidorm, Spain, 26-30 June 2011

H.264/AVC Full-pixel Motion Estimation for GPU Platforms

R. Rodríguez-Sánchez[1], J. L. Martínez[2], G. Fernandez-Escribano[1], J. M. Claver[3] and J. L. Sánchez[1]

[1] *Instituto de Investigación en Informática de Albacete (I3A), Universidad de Castilla-La Mancha*
[2] *ArTeCs Group, Universidad Complutense de Madrid*
[3] *Departamento de Informática, Universidad de Valencia*

emails: rrsanchez@dsi.uclm.es, joseluis.martinez@fdi.ucm.es, gerardo@dsi.uclm.es, jclaver@uv.es, jsanchez@dsi.uclm.es

Abstract

H.264/MPEG-4 part 10 is the latest standard for video compression and promises a significant advance compared with the commercial standards currently most in use. H.264/AVC aims at providing good video quality with lower bit rate than the previous coding standards such as MPEG-2 or MPEG-4. This improvement in performance occurs at the expense of increasing computing needs. Recently, the progress of GPUs has attracted a lot of attention. The hardware design includes multiple cores, bigger memory sizes and high bandwidth memory, which offer practical and acceptable solutions for speeding both graphic and non-graphic applications. In this paper we present an implementation of the H.264/AVC full-pixel Motion Estimation, which is able to reduce the execution time up to 98.12% on average and it reduces the energy consumption by a factor of 9.83 compared with the reference implementation, while there is a negligible drop in rate distortion for the encoded video.

Key words: Heterogeneous computing, Hardware accelerators, H.264/AVC, Motion Estimation, Inter Prediction

 ISBN: 978-84-614-6167-7

1. Introduction

H.264/AVC is a block-oriented motion-compensation-based codec standard that outperforms other previously existing video codecs. It aims at providing good video quality with lower bit rate than previous codecs, at the cost of an increased encoder [1]. By adopting new coding techniques, H.264/AVC can generate a high coding efficiency. However, real time encoding is very difficult to achieve due to its high computational complexity. Moreover, these techniques improve the coding gain but produce a high computational cost and large system memory bandwidth requirements. Therefore, it is necessary to reduce the encoding time as much as possible.

In this sense, in the past few years new heterogeneous architectures have been introduced in high-performance computing [2]. Examples of such architectures include Graphics Processing Units (GPUs). GPUs are accelerator devices with hundreds of similar processing cores which are designed and organized with the goal of achieving higher performance. Although GPUs can be used for general purposes, they come primarily from multimedia and computer or console gaming.

The GPU uses an unusual programming model, so effective programming is not an easy task. Fortunately, the main GPU manufacturers have developed their own tools for transparent programming. For example, CUDA (Compute Unified Device Architecture)[3], which is a powerful GPU architecture that enables dramatic increases in computing performance by harnessing the power of the GPU. CUDA abstracts both SIMD and task parallelism into thousands of simultaneous threads. These modifications largely improve the flexibility and programmability of GPUs.

This paper proposes an implementation of the H.264/AVC full-pixel Motion Estimation (ME) algorithm in a GPU as a coprocessor to assist the CPU. The proposed method parallelizes the inter prediction in H.264/AVC, which is the most time consuming task. Our ME algorithm is optimized for CUDA architecture by using a large number of threads that can execute on GPU in parallel and can make an efficient use of the available resources. Starting with the smaller MacroBlocks (MBs) sub-partitions, the proposed algorithm is able to build the entire tree structured motion compensation algorithm from bottom to top, and also, it reaches a 53x speedup on average and it consumes less energy than the baseline algorithm running on a CPU.

The rest of the paper is organized as follows: Section 2 contains a brief overview of H.264/AVC and GPU programming. In Section 3 some related proposals are presented. Section 4 describes the approach presented in this paper. In Section 5 the proposal presented is evaluated. Finally, conclusions are given in Section 6.

2. Background

H.264 [4] or MPEG-4 part 10 Advanced Video Coding (AVC) [5] is a compression video standard developed jointly by the ITU-T Video Coding Experts Group (ITU-T VCEG) and the ISO/IEC Moving Picture Experts Group (MPEG). H.264/AVC promises a significant advance in terms of quality and distortion compared with the commercial standards most in use such as MPEG-2 or MPEG-4 [1]. H.264/AVC can be used thus in a variety of applications such as DVD, video-streaming or HDTV.

For the inter frame coding, the H.264/AVC standard adopts many video coding techniques, such as multiple reference frame, weighted prediction, de-blocking filter, variable block size and quarter sample accurate motion compensation, among others. In

 ISBN: 978-84-614-6167-7

particular, the process of variable block size ME can search for the optimal matching block and it is able to eliminate the temporal redundancy between adjacent frames. This procedure is known as the tree structure motion compensation algorithm.

In a nutshell, the inter prediction in H.264/AVC supports motion compensation block sizes ranging from 16x16, 16x8, 8x16 to 8x8; where each of the sub-divided regions is a MB partition. If the 8x8 mode is chosen, each of the four 8x8 block partitions within the MB may be further split in 4 ways: 8x8, 8x4, 4x8 or 4x4, which are known as sub-MB partitions. Moreover, H.264/AVC also allows intra predicted modes, and a skipped mode in inter frames for referring to the 16x16 mode where no motion and residual information is encoded. Therefore, H.264/AVC allows not only the use of the MBs in which the images are decomposed, but also allows the use of smaller partitions from dividing the MBs in different ways. Fig. 1.a shows the different block sizes in which an MB can be divided, and Fig. 1.b shows the MB sub-partitions in which 8x8 partitions can be further divided.

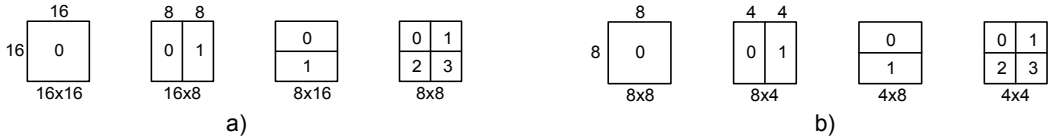


Fig. 1 H.264/AVC block sizes, a) MB partitions, b) 8x8 sub-partitions.

For each of these different MB sizes, the ME procedure is carried out, and a separate motion vector is required for each partition/sub-partition. Encoding a motion vector (MV) for each partition can take a significant number of bits, especially if small partition sizes are chosen. Moreover, MVs neighboring partitions are often highly correlated, so each MV is predicted from vectors of nearby previously coded partitions. Candidate MBs include the left MB, above MB, and above-right MB against the current MB. The predicted MV of the current MB is computed as the median of the candidate MVs. The Motion Vector prediction (MVp) depends on the motion compensation partition size and on the availability of nearby vectors. This MVp, which is generated from the neighboring MB, is added to the current MV.

On the other hand, modern graphics cards now incorporate a processor chip, which is commonly referred to as Graphics Processing Unit, or GPU. Today, GPUs are SIMD processors which are capable of performing arbitrary programmable operations. Recent GPUs are able to achieve 177 GB/s of memory transfer rate and 2 TFLOPS working in simple precision. GPUs have recently moved from being exclusively used in graphics applications to being used in what is now called General Purpose Computing on GPU (GPGPU) [6].

The main feature of these devices is a large number of processing elements integrated into a single chip at the expense of a significant reduction in cache memory. For instance, the architecture of the NVIDIA ® GPUs consists of a set of SIMD multiprocessors called Stream Multiprocessors (SM). Each SM has 8, 32 or 48 processing elements called cores and a set of resources shared by all cores: 32 bits registers, local shared memory, a cache for texture GPU memory and a cache for constant GPU memory. Each multiprocessor executes some thread blocks in time slots and a concrete thread block is always executed by the same multiprocessor.

GPU architecture offers a new challenge for engineering since the programming model must be adapted to the available hardware in order to obtain good performance and

 ISBN: 978-84-614-6167-7

exploit the full potential of the GPU. This problem has been solved by GPU manufacturers, such as NVIDIA or AMD-ATI, by proposing new languages or even extensions for the most commonly-used high-level programming languages.

In this respect, NVIDIA proposes the CUDA [3] parallel computing architecture, which is a software platform for programming massively parallel high-performance computing applications on their company powerful GPUs. CUDA development tools work alongside a conventional C/C++ compiler, so programmers can mix GPU code with general purpose code for the host CPU. At this point, the programmers do not write explicitly threaded code, a hardware thread manager handles threading automatically, which is an important feature of CUDA.

Thus, the inter prediction algorithm proposed in the H.264/AVC encoding algorithm fits well in the GPU philosophy, and offers a new challenge for the GPUs. The main issue is how to efficiently distribute all the computations over the GPU.

3. Related work

This section gathers relevant work focused on video processing using GPU-based frameworks. In the framework of video processing using GPU, one of the pioneer approaches was developed by Kelly and Kokaram in 2004 [7]. In this work, the authors propose using computer graphics hardware for fast image interpolation. Basically, the authors implemented the well-known full search block matching ME algorithm by using the OPENGL API. The results show a Speedup up to four times faster. However, the ME algorithm is only a part of the current video coding standards. In 2006, Ho et al in [8] presented a ME algorithm for the H.264/AVC using GPUs based on block-by-block basis providing a mechanism which is able to adjust the arithmetic intensity to maximize the performance on different GPUs. In 2007, Lee et al. in [9] showed a multi-pass and frame parallel algorithms to accelerate the H.264/AVC ME using a GPU. They unroll and rearrange the multiple nested loops involved in the ME algorithm by using the multi-pass method over the GPU. In 2008, Chen and Hang in [10] proposed an implementation of the H.264/AVC ME algorithm using CUDA. The algorithm is based on an efficient block-level parallel algorithm for the variable block size ME in H.264/AVC. They decompose the H.264/AVC ME algorithm into 5 steps so that they can achieve highly parallel computation. However, they do not show any results about Rate-Distortion (RD) performance.

At this point, this paper presents improvements made to the previously proposed algorithm [11] showing results in terms of time and RD distortion as well as including a GPU energy consumption analysis of the presented approach.

4. Inter Prediction GPU-based Implementation

In this section, we describe our H.264/AVC full-pixel ME GPU-based implementation. As mentioned before, it is the most time consuming task. We made some previous analysis in order to obtain this conclusion. As example, the full-pixel ME can take 89.29% of the total time in average, considering 18 VGA (640x480 pixels) sequences encoded with different QPs and 32 as search range.

Reference full-pixel ME sequentially obtains the Sum of Absolute Differences (SAD) cost for all positions checked inside the search area for all possible partitions/sub-

 ISBN: 978-84-614-6167-7

partitions defined by the standard, for each MB in a frame. Our main idea is to generate all the motion information at the beginning of each frame, dividing full-pixel ME into three tasks: the first one obtains the SAD costs exclusively for the 4x4 sub-partitions, the second one obtains the motion information for the other higher partitions/sub-partitions by using the data generated by the first task, and the third one performs a reduction to obtain the best match, that is, the MV with the lowest SAD cost. The ME is performed into three GPU kernels, the first GPU kernel performs the first task, the second GPU kernel performs the second task and a partial reduction of the generated data and the third GPU kernel performs the final reduction.

In order to work with a GPU as a coprocessor, the required data must be explicitly transferred into its DRAM memory. Thus, at the beginning of the encoding process some data are transferred from main memory associated to the CPU to GPU DRAM memory. The data that does not change during the encoding process (frame sizes, search area dimensions, search area distribution) are moved into the GPU constant memory at the beginning of the encoding process. On the other hand, at the beginning of coding each frame, the frame itself and the reference frame are transferred to the GPU DRAM memory.

4.1 Kernel 1

The goal of the first kernel is to obtain the required 4x4 SAD costs. Fig. 2.a shows the search area distribution defined by the reference H.264/AVC encoder for a given MB, it contains (2 x Search range)2 positions. The search area distribution follows a spiral pattern, position 0 corresponding to the center of the search area, MV (0,0).

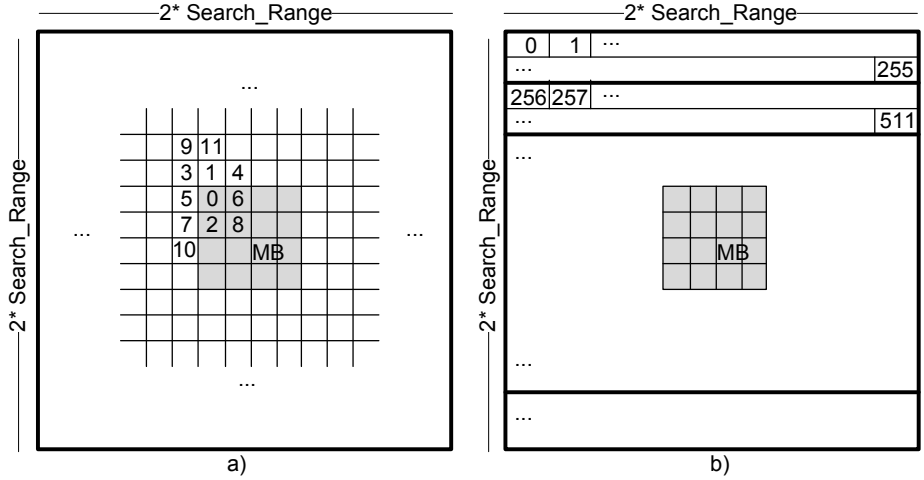


Fig. 2 Search area distribution, a) spiral pattern, b) custom pattern.

In this search area distribution, consecutive positions do not correspond to adjacent memory positions and locality cannot be exploited. Notice that pixels of the search area are accessed several times because each one is used to calculate more than one 4x4 block SAD cost. Therefore, to exploit locality and benefit of GPU texture memory cache or multiprocessors shared memory a new search area distribution is

 ISBN: 978-84-614-6167-7

defined. Fig. 2.b shows the new search area distribution where position 0 corresponds to the top-left corner of the search area and the positions are distributed by rows.

In the reference implementation one independent search area is defined per any of the partitions/sub-partitions available for the full-pixel ME predicted as explained in Section 2. But our implementation, uses the motion data from 4x4 sub-partitions to build the other partitions/sub-partitions motion data. So, we define a common search area for all partitions/sub-partitions, and the predictors for this search area are the half of the 16x16 MV of the MB sited in the same position in the previous frame.

One thread per position inside the search area is generated and 256 threads are grouped into a GPU thread block, computing N complete search area rows in the same thread block, where N depends on the dimensions of the search area. The SAD calculation is carried out in 4x4 blocks, and therefore each MB is divided into sixteen 4x4 blocks for each search area position. Each GPU thread calculates the 16 4x4 SAD costs of its associated position. These sixteen SAD costs are the basic data to build the structured motion tree in the next step.

The GPU used have a 16 KB local shared memory per multiprocessor, which is the fastest memory available in the GPU. It is known that an access to shared memory without bank conflicts is as fast as a register access. Therefore, we decided to use this kind of memory. As a consequence, of the new search area distribution, the required search area portion for a concrete thread block is contiguous. Hence, at the beginning of this kernel all threads cooperate to copy in shared memory the MB and the portion of search area assigned.

On the other hand, when using the same data type as reference software (unsigned short, i.e. 2 bytes) for MB and search area data, bank conflicts occur because of accessing to two contiguous position implies multiple access to the same bank at the same time, and the accesses must be serialized. We consider changing the memory data type, up-casting from unsigned short to integer since shared memory banks are organized in words of 4 bytes.

With this implementation, the GPU occupancy factor is 50% due to the register usage is very high (21 registers). New thread block sizes such as 192 or 384 can improve this index to 75%, but we decided to maintain 256 threads as the thread block size since, from experience, it gives the best performance, and allows applying the last improvement.

4.2 Kernel 2

The main purpose of the second kernel is to build the structured motion tree, obtaining in this way the SAD costs for all partitions/sub-partitions. This kernel also makes a first reduction of the generated data. As input, this kernel takes the motion information of 16 4x4 blocks of a MB for 64 positions. It generates the motion information for all partition/sub-partitions combinations, and reduces the amount of information, obtaining the best MV for each partition/sub-partition of the 64 positions. For this purpose, 64 GPU threads are grouped into a thread block, each of them building the SAD costs of a concrete position for all partitions/sub-partitions. Intermediate results are stored in local shared memory.

Fig. 3 shows how to build the motion information for all partitions/sub-partitions from the 4x4 SAD costs generated in the first kernel for a concrete position Pst. In order

 ISBN: 978-84-614-6167-7

to obtain the motion information for the 4x8 and 8x4 sub-partitions it is only necessary to add 2 4x4 SAD costs, for the 8x8 partitions it is only necessary to add 2 4x8 SAD and so on. Also, in order to obtain conflict-free accesses to shared memory, intermediate results for all partitions/sub-partitions are stored in the local shared memory, and are grouped using a structure composed of an unsigned short with the SAD cost and an unsigned short indicating its associated position (4 bytes).

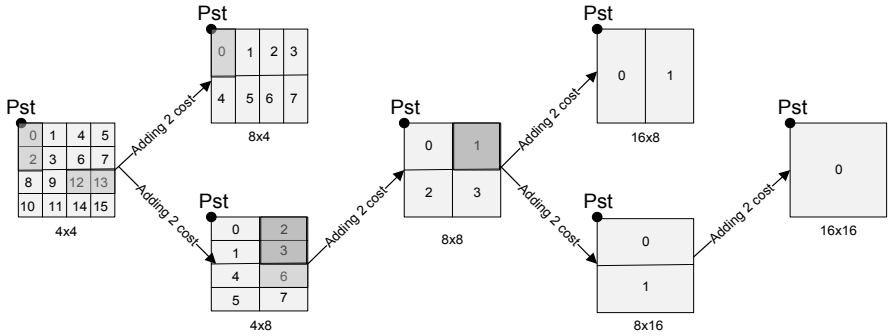


Fig. 3 Building the SAD partitions/sub-partitions.

To perform the reductions from local shared memory we propose an algorithm in 6 steps. We define a structure composed by 8 * 64 elements which is filled in with the motion information and later the data is reduced. For the 4x4 sub-partition 2 iterations are needed; 4x8 and 8x4 sub-partitions are performed in 1 iteration per each; 8x8, 8x16 and 16x8 partitions are performed together in 1 iteration; and finally, 16x16 partition is performed in the last iteration. Considering that 64 positions require 8*64*4 bytes of local shared memory and 21 registers per thread, results in a 38% GPU occupancy.

Local shared memory usage is 100% in all iterations except in the last one, in which it is 12.5%. But in this case, it is possible to allocate an extra row in the data structure (9*64 elements) to include the last iteration in the previous one, thus maintaining GPU usage. So, the new implementation is performed in just 5 iterations.

Once the SAD costs for every partition/sub-partition are generated, this information is used to compute the final cost. This calculation adds a penalization for large MVs. Penalization is computed using the following equation:

$$newSAD\_cost = SAD\_cost + k * vector\_bits \quad (1)$$

Where *newSAD_cost* is the penalized SAD cost, *SAD_cost* is the old SAD cost without penalization, *vector_bits* is the MV length associated to each position, and *k* is a constant which depends on the QP parameter used to encode the sequence (*k* is defined by the JM reference software).

The summing of costs for new partitions/sub-partitions has to be performed before adding the penalty but experimental results show that it is better to calculate the penalization in advance, update the shared memory with the penalized SAD costs, reduce the amount of information and recalculate the SAD cost from texture memory, although this means more GPU texture memory accesses to calculate the new partition/sub-partition SAD costs.

Fig. 4 shows a generic binary reduction in which each thread involved in the reduction process (t0 to tm-1) performs a reduction for each of the 8 rows in local shared

 ISBN: 978-84-614-6167-7

memory (b_0 to b_{N-1}). Note that m is equal to half of the remaining positions in any of the six iterations needed to reduce from 64 positions to 1. In order to complete the reduction process, six iterations are needed, starting with 64 (2^6) SAD costs per sub-partition and finishing with one SAD cost per sub-partition, where the SAD-matrix size is reduced by half in each iteration. The reductions are performed with SAD-matrix sizes of 64, 32, 16, 8, 4, and 2 using T strides, such that T=S/2, to obtain the best SAD cost per block. These strides are chosen to avoid local shared memory bank conflicts. The code for the six iterations is unrolled to avoid unnecessary loop climbs. Intermediate results are allocated to local shared memory.

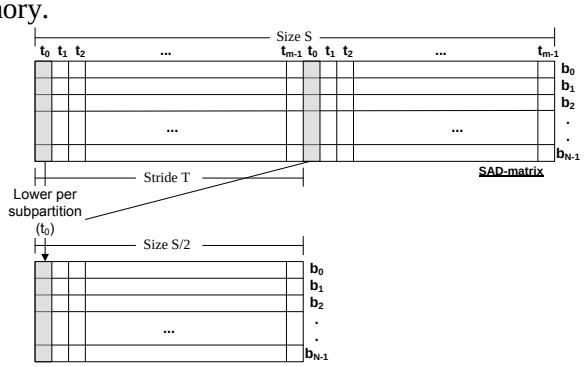


Fig. 4 Binary reduction scheme.

In the first iteration of the reduction 64 threads perform 4 comparisons each, in the second one 64 threads perform 2 comparisons, and in the third one 64 threads perform 1 comparison. After that, in the fourth, fifth and sixth iteration, 32, 16 and 8 threads perform 1 comparison respectively, because there are less comparisons remaining as available threads. Hence, the number of used threads is constant and equal to the maximum (64), until the number of comparisons is lower than the number of threads.

4.3 Kernel 3

Last kernel follows the same reduction procedure carried out in kernel 2 but using different data, so no more explanation is needed. This kernel obtains the best SAD cost for each one of the MB partitions/sub-partitions of each MB in a frame.

5. Performance Evaluation

In this section, we evaluate the behaviour of our parallel implementation of the H.264/AVC encoder. Firstly, we describe the parameters used in the H.264/AVC encoder configuration and then ran the parallel code considering several video sequences.

5.1 System

In order to show the performance of the H.264/AVC ME implementation proposed in this paper, it was implemented in the H.264/AVC JM 15.1 reference software encoder [12]. The parameters used in the H.264/AVC encoder configuration file were those included in the baseline profile of the mentioned reference software. Only six parameters were changed in the configuration file: The number of reference frames was set to 1 in order to keep the complexity as low as possible because higher values imply excessive time

 ISBN: 978-84-614-6167-7

consumption; RD-Optimization was disabled for the same reason as the *NumberReferenceFrames* parameter; The GOP pattern was fixed to I(11)P; The tests were carried out with popular sequences in VGA format (640 x 480), and therefore the *SourceWidth* and *SourceHeight* parameters were changed accordingly; The parameter *FramesToBeEncoded* was adjusted according to the sequence, in order to encode the full sequence; The Quantization Parameter (QP) called *QPISlice* and *QPPSlice* was varied among 28, 32, 36 and 40 according to [13].

The host machine used was an Intel® Core™ 2 Quad CPU Q9300 running at 2.50 GHz with 4GB of DDR2 memory. It included a GPU NVIDIA GTX285, with NVIDIA driver and CUDA support (190.18). The operating system was Linux Ubuntu 10.4 x64 with GCC 4.4.

5.2 Metrics

We show results of execution time, quality of the resulting encoded video sequence and energy consumption for the whole system. For applications such as the H.264/AVC encoder, it is very important to reduce their response time. Therefore, most of the optimizations are aimed at reducing the execution time. However, as a consequence of some of the implementations developed to increase the H.264/AVC encoder performance, such as the redistribution of the search area and data reductions, the resulting H.264/AVC bit stream was changed. The bit stream changes affect the quality and size of the output video sequence; hence it is necessary to study these metrics. Note that the changes in the bit streams do not modify the format of the encoded video sequence and the output video sequence can be decoded normally by any H.264/AVC decoder.

On the other hand, current GPUs suffer from higher power consumption requirements. Consequently, power and energy consumption become essential metrics in this kind of studies. Due to lack of space, we have not included in this paper the metrics description. However, a deeply description of them is available in [14].

5.3 Results

Table 1 shows the Speedup, and its associated Time reduction (TR), focusing on the parallelized process of the ME (full-pixel ME) on the GPU compared with the H.264/AVC JM 15.1 reference software encoder, depending on the sequence encoded. The results show significant improvements, obtaining a Speedup up to 71x and 53.13x in average, which means a time reduction up to 98.59% and 98.12 in average. Also, Table 1 shows the results in terms of ΔPSNR and ΔBitrate for the full encoder process. This should provide a good analysis of the performance of the proposed encoder for all different kinds of video content. Compared with the reference encoder, the proposed approach has a PSNR drop of, at most, 0.152 dB for a given bitrate, and a bitrate increase of, at most, 5.07% for a given PSNR in VGA format. This drop in RD performance is negligible if the computational savings are taken into account (the encoding time is a key feature in the design of real-time H.264/AVC encoders).

Only Tempete sequence was chosen for testing the energy consumption of the implementations because the results respect to this metric are independent of the sequence, so we do not consider necessary to show results for more sequences.

 ISBN: 978-84-614-6167-7

Table 1 Time and RD performance of the proposed GPU-based algorithm.

Sequence, Number of frames and QPs			Time Reduction (mean) from H.264/AVC 15.1 reference encoder			
			TR%	***Speedup***	***ΔBitrate***	***ΔPSNR***
Canoe	220	(28,32,36,40)	98.42	63.37	4.46	0.152
Fun fair	260	(28,32,36,40)	98.06	51.64	3.44	0.116
Harp	220	(28,32,36,40)	98.05	51.36	3.74	0.117
Mobile	260	(28,32,36,40)	98.30	58.86	3.76	0.149
Parade	220	(28,32,36,40)	98.33	59.79	2.95	0.108
Sgi-ant	220	(28,32,36,40)	97.02	33.52	1.52	0.052
Softfootball	220	(28,32,36,40)	98.59	71.00	3.57	0.114
Tempete	260	(28,32,36,40)	98.27	57.91	3.97	0.139
Waterfall	260	(28,32,36,40)	98.01	50.31	5.07	0.147
mean			***98.12***	***53.13***	***3.61***	***0.121***

Fig. 5 shows the power consumption obtained for coding one GOP (12 frames) using the reference H.264/AVC encoder and for coding GPU implementation presented in this paper. Compared with the reference encoder, the GPU implementation consumes more power but for a shorter period of time. The reference encoder consumes 127.89 Watts in 64.26 seconds, but for the GPU implementation the time is shorter, 5 seconds. In the figure, 11 consumption peaks can be seen for the GPU implementation (1 GOP is composed of 1 intra frame and 11 P frames where the GPU code is executed). Each GPU peak can be further divided into two components; the first one corresponds to the CPU-GPU memory transfer, consuming around 225 Watts, and the execution of the GPU kernels consuming around 200 Watts. Note that the power consumption for CPU code in the GPU implementation is around 150 Watt, which is higher than for the reference execution because the GPU is always active, waiting for new kernels.

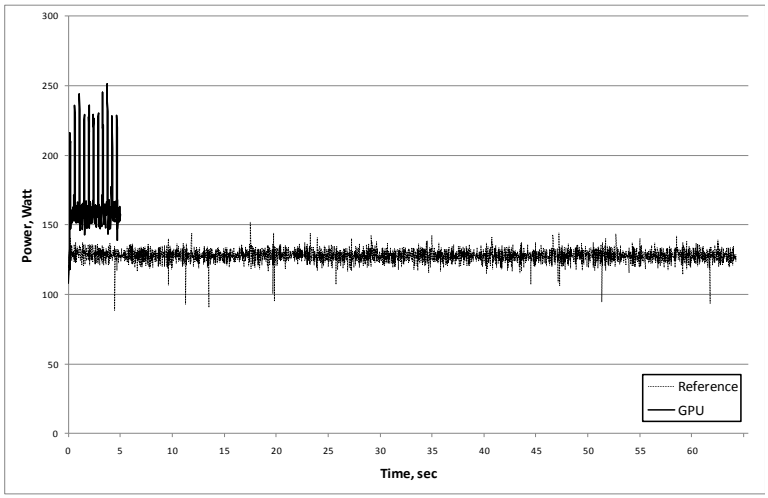


Fig. 5 Detailed power consumption comparison for coding a GOP.

Table 2 shows the average power, the time consumed and the energy consumed when coding one GOP for the reference H.264/AVC encoder and for the GPU implementation. The energy consumption for the GPU implementation is 9.83 times better than for the reference implementation.

Despite the improvement, GPU usage is not efficient either from the performance point of view, in terms of throughput, nor from the power consumption perspective, if we

 ISBN: 978-84-614-6167-7

consider watts per work performed. GPU is working only about 16% (0.801s) of the total time, so the resources usage can be improved a lot.

Table 2 One sequence energy consumption for coding a GOP.

Implementation	Power (Watt)	Time (seconds)	Energy (Jules)
Reference	127.89	64.26	8,218.21
GPU version	167.28	5.00	836.40

As our computing platform has four cores, four sequences can be simultaneously coded, sharing GPU resources and using different CPU cores. Thus, GPU usage can be increased and consequently total throughput will also increase.

Table 3 shows the average power, the time consumed and the energy consumed in coding one GOP of four instances of Tempete sequence for the reference H.264/AVC encoder and for the GPU implementation. The energy consumption for the GPU implementation is 8.49 times better than for the reference implementation.

Table 3 Four sequences and four CPUs energy consumption.

Implementation	Power (Watt)	Time (seconds)	Energy (Jules)
Reference	158.34	69.68	11,033.33
GPU version	212.46	6.12	1,300.26

Finally, with the execution of the four instances of Tempete sequence, the GPU is working 0.782s, 0.782s, 0.801s and 0.777s for each of them, which means a 51% (3.142s) of the total time, raising the GPU throughput while maintains the energy efficiency of the proposed algorithm.

6. Conclusions

This paper presents in detail an implementation of the H.264/AVC full-pixel ME in a GPU-based platform. This paper evaluates performance in terms of time, but also in terms of energy consumption and GPU throughput, which are very important GPU-platform features.

In this way, the proposed implementation was tested using a large and varied set of video sequences with the aim of evaluating their performance. As a result, the encoder performed very well with all of them, achieving a reduction in computational time of up to 98.12% for the H.264/AVC full-pixel ME procedure with negligible rate distortion loss, and the energy consumption was up to 9.83 times better than for the reference implementation.

Finally, coding four sequence simultaneously using 4 CPU cores allows improving the system resources utilization while maintains the energy efficiency of the proposed algorithm.

Acknowledgments.

This work was supported by the Spanish MEC and MICINN, as well as European Comission FEDER funds, under Grants CSD2006-00046 and TIN2009-14475-C04. It was also supported by The Council of Science and Technology of Castilla-La Mancha under Grants PEII09-0037-2328, PII2I09-0045-9916, and PCC08-0078-9856.

7. References

[1] Wiegand, T., Sullivan, G.J., Bjontegaard, G., Luthra, A.: Overview of the H.264/AVC Video Coding Standard. In: IEEE Trans. Circuits Syst. Video Technol. 13(7), 560 - 576 (2003).

[2] Feng, W.-c., Manocha, D.: High-performance computing using accelerators. In: Parallel Computing, vol 33, n. 10-11, pp 645-647, November 2007.

[3] NVIDA.: NVIDIA CUDA Compute Unified Device Architecture-Programming Guide, Version 2.3. August 2009.

[4] ISO/IEC International Standard 14496-10:2003.: Information Technology – Coding of Audio - Visual Objects – Part 10: Advanced Video Coding.

[5] ISO/IEC International Standard 14496-10:2005.: Information technology – Coding of audio-visual objects – Part 10: Advanced Video Coding.

[6] GPGPU (2007), General-purpose computation using graphics hardware, http://www.gpgpu.org.

[7] Kelly, F., Kokaram, A.: Fast Image Interpolation for Motion Estimation using Graphics Hardware. In IS&T/SPIE Electronic Imaging - Real-Time Imaging VIII, vol. 5297, pp. 184-194, San Jose (California), USA, January 2004.

[8] Ho, C.-W., Au, O.C., Gary Chan, S.-H., Yip, S.-K., Wong, H.-M.: Motion Estimation for H.264/AVC using Programmable Graphics Hardware. In Proceedings Of IEEE International Conference on Multimedia and Expo, ICME, pp. 2049-2052, Toronto, Canada, July 2006.

[9] Lee, C.-Y., Lin, Y.-C., Wu, C.-L., Chang, C.-H., Tsao, Y.-M., Chien, S.-Y.: Multi-Pass and Frame Parallel Algorithms of Motion Estimation in H.264/AVC for Generic GPU. In Proc. of IEEE International Conference on Multimedia and Expo, ICME, pp. 1603-1606, Beijing, China, July 2007.

[10] Chen, W.-N., Hang, H.-M.: H.264/AVC motion estimation implementation on Compute Unified Device Architecture (CUDA). In Proceedings of IEEE International Conference on Multimedia and Expo, ICME, pp. 679-700, June 2008, Hannover, Germany.

[11] Rodriguez, R., Martínez, J. L., Fernández-Escribano, G., Claver, J. M., Sánchez, J. L. Accelerating H.264 Inter Prediction in a GPU using CUDA. In International Conference on Consumer Electronics, ICCE 2010, 10.4-2, pp. 463-464, Las Vegas, USA.

[12] Joint Video Team (JVT) of ISO/IEC MPEG and ITU-T VCEG. Reference Software to Committee Draft. JVT-F100 JM15.1, 2009, http://iphome.hhi.de/suehring/tml/.

[13] Sullivan, G., Bjøntegaard, G.: Recommended Simulation Common Conditions for H.26L Coding Efficiency Experiments on Low-Resolution Progressive-Scan Source Material. In ITU-T VCEG, Doc. VCEG-N81. September 2001.

[14] Rodríguez-Sánchez, R., Martínez, J. L., Fernández-Escribano, G., Sánchez, J. L., Claver, J. M., Díaz, P.: Technical Report DIAB-11-01-2. Departamento de Sistemas Informáticos Universidad de Castilla-La Mancha, January 2011.

 ISBN: 978-84-614-6167-7

Proceedings of the 11th International Conference on Computational and Mathematical Methods in Science and Engineering, CMMSE2011 Benidorm, Spain, 26-30 June 2011

Thermal Stress Wave Propagation Study of Functionally Graded Thick Hollow Cylinder

Ali Safari-Kahnaki, Mohammad Mohammadi-Aghdam and Mohammad Reza Eslami

Thermoelasticity Center of Excellency, Mechanical Engineering Department, Amirkabir University of Technology, Tehran, Iran

Email: alisafarik@aut.ac.ir

Abstract

This paper presents an analytical method to study the dynamic behavior of stress field in a finite length functionally graded (FG) thick hollow cylinder under thermal loading. The thermomechanical properties are assumed to vary continuously through the radial direction as a nonlinear power function. Using Laplace transform and series solution, the thermoelastic Navier equations as well as heat conduction equation are analytically solved. The solution of displacement field in FG cylinder is obtained in the Laplace domain. The fast Laplace inverse transform (FLIT) method is employed to transfer the results from Laplace domain to time domain. The effects of thermal loading on dynamic characteristics of FG thick hollow cylinder are investigated in various points across thickness of cylinder for various grading patterns of FGMs. Good agreement can be seen between presented predictions with other result in open literature.

Key words: Thermal Stresses, Functionally Graded Materials, Finite Length Hollow Cylinder, Wave Propagation Analysis, Analytical Method.

1. Introduction

The applications of functionally graded materials (FGMs) are increasing because of their capability to control the thermomechanical stresses in structures under

 ISBN: 978-84-614-6167-7

thermal and mechanical loadings. FGMs are new kind of composite materials whose thermomechanical properties vary continuously along certain directions. In the recent years, a number of analytical solutions have been obtained by some researchers to determine the stress distribution in structures especially in cylindrical pressure vessels.

Jabbari et al. (2006 and 2009) studied the thermoelasticity analysis in steady-state conditions for infinite and finite length functionally graded hollow cylinders subjected to thermomechanical loading. They used the Bessel functions in their method for obtaining the general solution of cylinder. Shao et al. (2007) developed an analytical method based on series solution for analysis of transient thermomechanical stresses in a FG thick hollow circular cylinder with infinite length. The inertia term in equation of motion was not considered by them and they considered pseudo dynamic conditions for their problem. Hosseini et al. (2007) studied the vibration and dynamic analysis of functionally graded thick hollow cylinders with axisymmetry and plane strain conditions. The mean velocity of radial stress wave propagation, natural frequency, and dynamic behavior of FG cylinder were presented in their work using Galerkin finite element with linear functionally graded elements and Newmark's method. Shariyat et al. (2009) investigated stress wave propagation characteristics in FG thick hollow cylinders with temperature-dependent material properties subjected to dynamic thermomechanical loading. They used a second-order point-collocation method to analyze nonlinear transient thermal stresses and elastic wave propagation in FG cylinders. Safari-Kahnaki et al. (2010 and 2011) studied the thermal stress field and radial stress wave propagation in FG thick hollow cylinder with infinite and finite length under thermal shock loading. They did not solved heat conduction equation in their studies and took temperature loading as a simple form which was modeled as suddenly rising of uniform temperature of cylinder body.

This paper presents an effective analytical method to study the thermoelastic stress field in a functionally graded thick hollow cylinder with finite length induced by axisymmetric thermal loading. The governing equations of the problem are considered based on transient thermoelasticity conditions (considering inertia term in equations of motion). All equations are transferred to the Laplace domain and analytically solved using series solution technique. To study the time history of two-dimensional thermoelastic stresses, the stress field in FG cylinder is obtained in time domain using the Fast Laplace Inverse Transform (FLIT) method. The effects of various grading patterns of FGMs on distribution of thermal stresses are graphically shown for several points across thickness of FG cylinder.

 ISBN: 978-84-614-6167-7

2. Mathematical Equations

A functionally graded thick hollow cylinder with inner and outer radii a and b and finite length l is considered. The axisymmetry and plane strain conditions are assumed for the problem. Consequently, strain-displacement relations can be written as

$$\varepsilon_r(r,z,t)=\frac{du(r,z,t)}{dr},\quad \varepsilon_\theta(r,z,t)=\frac{u(r,z,t)}{r},$$
$$\varepsilon_z(r,z,t)=\frac{dw(r,z,t)}{dz},\quad \varepsilon_{zr}(r,z,t)=\frac{dw(r,z,t)}{dr}+\frac{du(r,z,t)}{dz}. \tag{1}$$

where $u(r,z,t)$ and $w(r,z,t)$ are radial and axial displacements and $\varepsilon_r(r,z,t)$, $\varepsilon_\theta(r,z,t)$, $\varepsilon_z(r,z,t)$, and $\varepsilon_{zr}(r,z,t)$ are radial, hoop, axial, and shear strains, respectively. Subscripts r, θ and z denote radial, circumferential, and axial directions, respectively and Also, the terms r, z and t show radius, length, and time.

The dynamic thermo-elastic stresses are given by

$$\sigma_r(r,z,t)=\frac{E(r)}{(1+\nu)(1-2\nu)}\left[(1-\nu)\varepsilon_r(r,z,t)+\nu\varepsilon_\theta(r,z,t)+\nu\varepsilon_z(r,z,t)\right]-\frac{\alpha(r)E(r)}{(1-2\nu)}T(r,z,t)$$
$$\sigma_\theta(r,z,t)=\frac{E(r)}{(1+\nu)(1-2\nu)}\left[\nu\varepsilon_r(r,z,t)+(1-\nu)\varepsilon_\theta(r,z,t)+\nu\varepsilon_z(r,z,t)\right]-\frac{\alpha(r)E(r)}{(1-2\nu)}T(r,z,t)$$
$$\sigma_z(r,z,t)=\frac{E(r)}{(1+\nu)(1-2\nu)}\left[\nu\varepsilon_r(r,z,t)+\nu\varepsilon_\theta(r,z,t)+(1-\nu)\varepsilon_z(r,z,t)\right]-\frac{\alpha(r)E(r)}{(1-2\nu)}T(r,z,t) \tag{2}$$
$$\sigma_{zr}(r,z,t)=\frac{E(r)}{2(1+\nu)}\varepsilon_{zr}(r,z,t)$$

where $\sigma_r(r,z,t)$, $\sigma_\theta(r,z,t)$, $\sigma_z(r,z,t)$, and $\sigma_{zr}(r,z,t)$ are radial, hoop, axial, and shear stresses, respectively. Also, the terms ν, $E(r)$, $\alpha(r)$, and $T(r,z,t)$ are Poisson's ratio, elastic modulus, thermal expansion coefficient, and temperature of FG cylinder, respectively. The heat conduction equation and equations of motion considering the inertia terms for finite length cylinder become

$$\frac{\partial T}{\partial r^2}+\left(\frac{1}{r}+\frac{\lambda'(r)}{\lambda(r)}\right)\frac{\partial T}{\partial r}+\frac{\partial T}{\partial z^2}=\frac{1}{k(r)}\frac{\partial T}{\partial t}$$
$$\frac{\partial\sigma_r}{\partial r}+\frac{\partial\sigma_{zr}}{\partial z}+\frac{\sigma_r-\sigma_\theta}{r}=\rho\frac{\partial^2 u}{\partial t^2},\qquad \frac{\partial\sigma_{zr}}{\partial r}+\frac{\partial\sigma_{zz}}{\partial z}+\frac{\sigma_{zr}}{r}=\rho\frac{\partial^2 w}{\partial t^2}. \tag{3}$$

where λ and k are thermal conductivity coefficient and thermal diffusivity of FG cylinder and ρ is the mass density. As it is mentioned in introduction section, in preceding researches in this field, the pseudo dynamic thermo-elastic problem (considering equation of motion without inertia terms) was solved by analytical method.

 ISBN: 978-84-614-6167-7

For considered finite length FG cylinder with simply-supported ends, the thermal and mechanical boundary conditions can be expressed as,

$$\begin{aligned}
&T(a,z,t)=H(t)\sin(\frac{\pi z}{l}) && T(a,z,t)=0\\
&u(r,0,t)=\sigma_z(r,0,t)=0 && \text{at}\quad z=0\\
&u(r,l,t)=\sigma_z(r,l,t)=0 && \text{at}\quad z=l\\
&\sigma_r(a,z,t)=\sigma_{zr}(a,z,t)=0 && \text{at}\quad r=a\\
&\sigma_r(b,z,t)=\sigma_{zr}(b,z,t)=0 && \text{at}\quad r=b
\end{aligned} \tag{4}$$

where $H(t)$ is the Heaviside step function.

Furthermore, the initial conditions are considered as follows

$$T(r,z,0)=0,\qquad u(r,z,0)=\frac{\partial u(r,z,0)}{\partial t}=0,\qquad w(r,z,0)=\frac{\partial w(r,z,0)}{\partial t}=0 \tag{5}$$

For the sake of simplicity, some dimensionless parameters can be introduced as (Shao 2007)

$$\begin{aligned}
&\xi=\frac{r}{\bar{r}},\ \xi_1=\frac{a}{\bar{r}},\ \xi_2=\frac{b}{\bar{r}},\ \bar{r}=\frac{a+b}{2},\ Z=\frac{z}{\bar{r}},\ L=\frac{l}{\bar{r}},\ \bar{T}(\xi,Z,\bar{t})=\frac{T(r,z,t)}{T_0}\\
&t^*=\left(\frac{C_V}{\bar{r}}\right)t,\ C_V=\sqrt{\frac{E}{\rho}},\ \bar{t}=\left(\frac{k_0}{\bar{r}^2}\right)t,\ \bar{\lambda}(\xi)=\frac{\lambda(r)}{\lambda_0},\ \bar{k}(\xi)=\frac{k(r)}{k_0},\ \bar{E}(\xi)=\frac{E(r)}{E_0},\\
&\bar{\alpha}(\xi)=\frac{\alpha(r)}{\alpha_0},\ \bar{\rho}(\xi)=\frac{\rho(r)}{\rho_0},\ \bar{u}(\xi,Z,t^*)=\frac{u(r,z,t)}{\alpha_0 T_0\bar{r}},\ \bar{w}(\xi,Z,t^*)=\frac{w(r,z,t)}{\alpha_0 T_0\bar{r}}\\
&,\ \bar{\sigma}_r(\xi,Z,t^*)=\frac{\sigma_r(r,z,t)}{\alpha_0 T_0 E_0},\ \bar{\sigma}_\theta(\xi,Z,t^*)=\frac{\sigma_\theta(r,z,t)}{\alpha_0 T_0 E_0},\\
&\bar{\sigma}_z(\xi,Z,t^*)=\frac{\sigma_z(r,z,t)}{\alpha_0 T_0 E_0},\quad \bar{\sigma}_{rz}(\xi,Z,t^*)=\frac{\sigma_{rz}(r,z,t)}{\alpha_0 T_0 E_0}.
\end{aligned} \tag{6}$$

where λ_0, k_0, C_V, α_0, T_0, E_0 and ρ_0 are reference values of thermal conductivity coefficient, thermal diffusivity, stress wave propagation speed, thermal expansion coefficient, temperature, elastic modulus and density, respectively.

The heat conduction equation and governing equations of motion can be written in dimensionless form using defined dimensionless parameters as

$$\frac{\partial\bar{T}}{\partial\xi^2}+\left(\frac{1}{\xi}+\frac{\bar{\lambda}'(\xi)}{\bar{\lambda}(\xi)}\right)\frac{\partial\bar{T}}{\partial\xi}+\frac{\partial\bar{T}}{\partial Z^2}=\frac{1}{\bar{k}(\xi)}\frac{\partial\bar{T}}{\partial\bar{t}} \tag{7}$$

 ISBN: 978-84-614-6167-7

$$\frac{\partial^2\overline{u}}{\partial\xi^2}+\left(\frac{1}{\xi}+\frac{\overline{E}'}{\overline{E}}\right)\frac{\partial\overline{u}}{\partial\xi}+\frac{\overline{u}}{\xi}\left(\frac{\nu}{1-\nu}\frac{\overline{E}'}{\overline{E}}-\frac{1}{\xi}\right)+\frac{1-2\nu}{2-2\nu}\frac{\partial^2\overline{u}}{\partial Z^2}+\left(\frac{1}{2-2\nu}\right)\frac{\partial^2\overline{w}}{\partial\xi\partial Z}+\left(\frac{\nu}{1-\nu}\right)\frac{\overline{E}'}{\overline{E}}\frac{\partial\overline{w}}{\partial Z}$$
$$-\left(\frac{1+\nu}{1-\nu}\right)\overline{\alpha}\frac{\partial\overline{T}}{\partial\xi}\ -\left(\frac{1+\nu}{1-\nu}\right)\frac{[\overline{E}\overline{\alpha}]'}{\overline{E}}\overline{T}=\frac{(1+\nu)(1-2\nu)}{(1-\nu)}\frac{\overline{\rho}}{\overline{E}}\frac{\partial^2\overline{u}}{\partial t^{*2}}$$
$$\frac{1}{2-2\nu}\frac{\partial^2\overline{u}}{\partial\xi\partial Z}+\left(\frac{1}{2-2\nu}\frac{1}{\xi}+\frac{\overline{E}'}{\overline{E}}\right)\frac{\partial\overline{u}}{\partial Z}+\left(\frac{1-2\nu}{2-2\nu}\right)\left[\frac{\partial^2\overline{w}}{\partial\xi^2}+\left(\frac{1}{\xi}+\frac{\overline{E}'}{\overline{E}}\right)\frac{\partial\overline{w}}{\partial\xi}\right.$$
$$\left.+\left(\frac{2-2\nu}{1-2\nu}\right)\frac{\partial^2\overline{w}}{\partial Z^2}\right]-\left(\frac{1+\nu}{1-\nu}\right)\overline{\alpha}\ \frac{\partial\overline{T}}{\partial Z}=\frac{(1+\nu)(1-2\nu)}{(1-\nu)}\frac{\overline{\rho}}{\overline{E}}\frac{\partial^2\overline{w}}{\partial t^{*2}}$$

The boundary conditions at the inner and outer surfaces can be expressed as

$$\overline{T}\left(\xi_1,Z,\overline{t}\right)=H\left(\overline{t}\right)\sin(K_nZ),\qquad \overline{T}\left(\xi_2,Z,t^*\right)=0$$
$$(1-\nu)\frac{\partial\overline{u}\left(\xi_1,Z,t^*\right)}{\partial\xi}+\nu\frac{\overline{u}\left(\xi_1,Z,t^*\right)}{\xi_1}+\nu\frac{\partial\overline{w}\left(\xi_1,Z,t^*\right)}{\partial Z}=(1-\nu)\overline{\alpha}\left(\xi_1\right)\overline{T}\left(\xi_1,Z,t^*\right)$$
$$(1-\nu)\frac{\partial\overline{u}\left(\xi_2,Z,t^*\right)}{\partial\xi}+\nu\frac{\overline{u}\left(\xi_2,Z,t^*\right)}{\xi_2}+\nu\frac{\partial\overline{w}\left(\xi_2,Z,t^*\right)}{\partial Z}=(1-\nu)\overline{\alpha}\left(\xi_2\right)\overline{T}\left(\xi_2,Z,t^*\right)\qquad(8)$$
$$\frac{\partial\overline{w}\left(\xi_1,Z,t^*\right)}{\partial\xi}+\frac{\partial\overline{u}\left(\xi_1,Z,t^*\right)}{\partial Z}=0\quad,\qquad\frac{\partial\overline{w}\left(\xi_2,Z,t^*\right)}{\partial\xi}+\frac{\partial\overline{u}\left(\xi_2,Z,t^*\right)}{\partial Z}=0$$

3. Solution Procedure

One solution for the problem that satisfies both Eq. (3) and boundary conditions (4) can be considered as

$$\overline{T}\left(\xi,Z,\overline{t}\right)=\sum_{n=o}^{\infty}\overline{T}_n\left(\xi,\overline{t}\right)\sin(K_nZ)\qquad\overline{u}\left(\xi,Z,t^*\right)=\sum_{n=o}^{\infty}\overline{u}_n\left(\xi,t^*\right)\sin(K_nZ)$$
$$\overline{w}\left(\xi,Z,t^*\right)=\sum_{n=o}^{\infty}\overline{w}_n\left(\xi,t^*\right)\cos(K_nZ)\qquad K_n=\frac{n\pi}{L},\quad n=0,1,2,3,\ldots.\qquad(9)$$

Substituting Eqs. (9) into Eq. (7) together with boundary conditions (8), leads to:

$$\frac{\partial\overline{T}_n\left(\xi,\overline{t}\right)}{\partial\xi^2}+\left(\frac{1}{\xi}+\frac{\overline{\lambda}'(\xi)}{\overline{\lambda}(\xi)}\right)\frac{\partial\overline{T}_n\left(\xi,\overline{t}\right)}{\partial\xi}-{K_n}^2\overline{T}_n\left(\xi,\overline{t}\right)=\frac{1}{\overline{k}(\xi)}\frac{\partial\overline{T}_n\left(\xi,\overline{t}\right)}{\partial\overline{t}}$$
$$\frac{\partial^2\overline{u}_n\left(\xi,t^*\right)}{\partial\xi^2}+\left(\frac{1}{\xi}+\frac{\overline{E}'}{\overline{E}}\right)\frac{\partial\overline{u}_n\left(\xi,t^*\right)}{\partial\xi}+\left(\frac{\nu}{1-\nu}\frac{\overline{E}'}{\overline{E}}-\frac{1}{\xi}\right)\frac{\overline{u}_n\left(\xi,t^*\right)}{\xi}-\left(\frac{1-2\nu}{2-2\nu}\right){K_n}^2\overline{u}_n\left(\xi,t^*\right)$$
$$-K_n\left(\frac{1}{2-2\nu}\right)\frac{\partial\overline{w}_n\left(\xi,t^*\right)}{\partial\xi}-K_n\left(\frac{\nu}{1-\nu}\right)\frac{\overline{E}'}{\overline{E}}\overline{w}_n\left(\xi,t^*\right)-\left(\frac{1+\nu}{1-\nu}\right)\frac{[\overline{E}\overline{\alpha}]'}{\overline{E}}\overline{T}_n\left(\xi,t^*\right)\qquad(10)$$
$$-\left(\frac{1+\nu}{1-\nu}\right)\overline{\alpha}\frac{\partial\overline{T}_n\left(\xi,t^*\right)}{\partial\xi}=\frac{(1+\nu)(1-2\nu)}{(1-\nu)}\frac{\overline{\rho}}{\overline{E}}\frac{\partial^2\overline{u}_n\left(\xi,t^*\right)}{\partial t^{*2}}$$

 ISBN: 978-84-614-6167-7

$$K_n\left(\frac{1}{1-2\nu}\right)\frac{\partial \overline{u}_n\left(\xi,t^*\right)}{\partial \xi}+\left(\frac{1}{2-2\nu}\frac{1}{\xi}+\frac{1-2\nu}{2-2\nu}\frac{\overline{E}'}{\overline{E}}\right)\overline{u}_n\left(\xi,t^*\right)+\left(\frac{1-2\nu}{2-2\nu}\right)\left[\frac{\partial^2 \overline{w}_n\left(\xi,t^*\right)}{\partial \xi^2}\right.$$
$$\left.+\left(\frac{1}{\xi}+\frac{\overline{E}'}{\overline{E}}\right)\frac{\partial \overline{w}_n\left(\xi,t^*\right)}{\partial \xi}-K_n{}^2\left(\frac{2-2\nu}{1-2\nu}\right)\overline{w}_n\left(\xi,t^*\right)\right]-\left(\frac{1+\nu}{1-\nu}\right)\left(K_n\overline{\alpha}\right)\overline{T}_n\left(\xi,t^*\right)$$
$$=\frac{(1+\nu)(1-2\nu)}{(1-\nu)}\frac{\overline{\rho}}{\overline{E}}\frac{\partial^2 \overline{w}_n\left(\xi,t^*\right)}{\partial t^{*2}}$$

and

$$\overline{T}_n\left(\xi_1,\overline{t}\right)=H\left(\overline{t}\right),\qquad \overline{T}\left(\xi_2,\overline{t}\right)=0$$
$$(1-\nu)\frac{\partial \overline{u}_n\left(\xi_1,t^*\right)}{\partial \xi}+\nu\frac{\overline{u}_n\left(\xi_1,t^*\right)}{\xi_1}-K_n\,\nu\,\overline{w}_n\left(\xi_1,t^*\right)=(1+\nu)\overline{\alpha}\left(\xi_1\right)\overline{T}_n\left(\xi_1,t^*\right)$$
$$(1-\nu)\frac{\partial \overline{u}_n\left(\xi_2,t^*\right)}{\partial \xi}+\nu\frac{\overline{u}_n\left(\xi_2,t^*\right)}{\xi_2}-K_n\,\nu\,\overline{w}_n\left(\xi_2,t^*\right)=(1+\nu)\overline{\alpha}\left(\xi_2\right)\overline{T}_n\left(\xi_2,t^*\right) \tag{11}$$
$$\frac{\partial \overline{w}_n\left(\xi_1,t^*\right)}{\partial \xi}+K_n\overline{u}_n\left(\xi_1,t^*\right)=0,\qquad \frac{\partial \overline{w}_n\left(\xi_2,t^*\right)}{\partial \xi}+K_n\overline{u}_n\left(\xi_2,t^*\right)=0.$$

Applying the Laplace transform with initial conditions (5) to Eqs. (10) and (11), the following equations are derived:

$$\frac{\partial^2 \tilde{T}_n(\xi,s)}{\partial \xi^2}+\left(\frac{1}{\xi}+\frac{1}{\overline{\lambda}(\xi)}\frac{d\overline{\lambda}(\xi)}{d\xi}\right)\frac{\partial \tilde{T}_n(\xi,s)}{\partial \xi}-K_n{}^2\tilde{T}_n\left(\xi,\overline{t}\right)=\frac{s}{\overline{k}(\xi)}\tilde{T}_n(\xi,s)$$
$$\frac{\partial^2 \tilde{u}_n(\xi,s)}{\partial \xi^2}+\left[\frac{1}{\xi}+\frac{\overline{E}'(\xi)}{\overline{E}(\xi)}\right]\frac{\partial \tilde{u}_n(\xi,s)}{\partial \xi}+\left[\frac{\nu}{1-\nu}\frac{1}{\xi}\frac{\overline{E}'(\xi)}{\overline{E}(\xi)}-\frac{1}{\xi^2}-K_n{}^2\left(\frac{1-2\nu}{2-2\nu}\right)\right.$$
$$\left.-\frac{(1+\nu)(1-2\nu)}{(1-\nu)}\frac{\overline{\rho}(\xi)}{\overline{E}(\xi)}s^2\right]\tilde{u}_n(\xi,s)-K_n\left(\frac{1}{2-2\nu}\right)\frac{\partial \tilde{w}_n(\xi,s)}{\partial \xi}-K_n\left(\frac{\nu}{1-\nu}\right)\frac{\overline{E}'(\xi)}{\overline{E}(\xi)}\tilde{w}_n(\xi,s)$$
$$=-\left(\frac{1+\nu}{1-\nu}\right)\frac{[\overline{E}(\xi)\overline{\alpha}(\xi)]'}{\overline{E}(\xi)}\tilde{T}_n(\xi,s)-\left(\frac{1+\nu}{1-\nu}\right)\overline{\alpha}\frac{\partial \tilde{T}_n(\xi,s)}{\partial \xi} \tag{12}$$
$$K_n\left(\frac{1}{2-2\nu}\right)\frac{\partial \tilde{u}_n(\xi,s)}{\partial \xi}+K_n\left[\frac{1}{2-2\nu}\frac{1}{\xi}+\frac{1-2\nu}{2-2\nu}\frac{\overline{E}'(\xi)}{\overline{E}(\xi)}\right]\tilde{u}_n(\xi,s)+\left(\frac{1-2\nu}{2-2\nu}\right)\left[\frac{\partial^2 \tilde{w}_n(\xi,s)}{\partial \xi^2}\right.$$
$$\left.+\left(\frac{1}{\xi}+\frac{\overline{E}'(\xi)}{\overline{E}(\xi)}\right)\frac{\partial \tilde{w}_n(\xi,s)}{\partial \xi}\right]-\left(K_n{}^2+\frac{(1+\nu)(1-2\nu)}{(1-\nu)}\frac{\overline{\rho}(\xi)}{\overline{E}(\xi)}s^2\right)\tilde{w}_n(\xi,s)$$
$$-\left(\frac{1+\nu}{1-\nu}\right)\left(K_n\overline{\alpha}\right)\tilde{T}_n\left(\xi,s\right)=0$$
$$\tilde{T}_n\left(\xi_1,s\right)=\frac{1}{s},\qquad \tilde{T}_n\left(\xi_2,s\right)=0$$

 ISBN: 978-84-614-6167-7

$$(1-\nu)\frac{\partial \widetilde{u}_n(\xi_1,s)}{\partial \xi}+\nu\frac{\widetilde{u}_n(\xi_1,s)}{\xi_1}-K_n\nu\,\widetilde{w}_n(\xi_1,s)=(1+\nu)\overline{\alpha}(\xi_1)\widetilde{T}_n(\xi_1,s) \tag{13}$$

$$(1-\nu)\frac{\partial \widetilde{u}_n(\xi_2,s)}{\partial \xi}+\nu\frac{\widetilde{u}_n(\xi_2,s)}{\xi_2}-K_n\nu\,\widetilde{w}_n(\xi_2,s)=(1+\nu)\overline{\alpha}(\xi_2)\widetilde{T}_n(\xi_2,s)$$

$$\frac{\partial \widetilde{w}_n(\xi_1,s)}{\partial \xi}+K_n\widetilde{u}_n(\xi_1,s)=0\quad,\quad\frac{\partial \widetilde{w}_n(\xi_2,s)}{\partial \xi}+K_n\widetilde{u}_n(\xi_2,s)=0\,.$$

where

$$\widetilde{T}_n(\xi,s)=\pounds\left[\overline{T}_n\left(\xi,t^*\right)\right],\qquad \widetilde{u}_n(\xi,s)=\pounds\left[\overline{u}_n\left(\xi,t^*\right)\right],\qquad \widetilde{w}_n(\xi,s)=\pounds\left[\overline{w}_n\left(\xi,t^*\right)\right].$$

The term £ is the Laplace transform operator. According to the series method for ordinary differential equation, if the coefficients $\overline{E}'(\xi)/\overline{E}(\xi)$, $\left[\overline{E}(\xi)\overline{\alpha}(\xi)\right]'/\overline{E}(\xi)$ and $\overline{\rho}(\xi)/\overline{E}(\xi)$ are analytical at $\xi=1$ and can express as Taylor's series in terms of $(\xi-1)$ as follows

$$\Gamma_1(\xi)=\frac{\overline{E}'(\xi)}{\overline{E}(\xi)}=\frac{1}{\overline{E}(\xi)}\frac{d\,\overline{E}(\xi)}{d\,\xi}=\sum_{n=0}^{\infty}\Gamma_{1,n}(\xi-1)^n\,,\Gamma_3(\xi)=\frac{\overline{\rho}(\xi)}{\overline{E}(\xi)}=\sum_{n=0}^{\infty}\Gamma_{3,n}(\xi-1)^n$$

$$\Gamma_2(\xi)=\frac{1}{\overline{E}(\xi)}\frac{d\left[\overline{E}(\xi)\overline{\alpha}(\xi)\right]}{d\,\xi}=\sum_{n=0}^{\infty}\Gamma_{2,n}(\xi-1)^n\,,\ \Gamma_4(\xi)=\overline{\alpha}(\xi)=\sum_{n=0}^{\infty}\Gamma_{4,n}(\xi-1)^n \tag{14}$$

$$\Gamma_5(\xi)=\frac{\overline{\lambda}'(\xi)}{\overline{\lambda}(\xi)}=\frac{1}{\overline{\lambda}(\xi)}\frac{d\,\overline{\lambda}(\xi)}{d\xi}=\sum_{n=0}^{\infty}\Gamma_{5,n}(\xi-1)^n\,,\Gamma_6(\xi)=\frac{1}{\overline{k}(\xi)}=\sum_{n=0}^{\infty}\Gamma_{6,n}(\xi-1)^n$$

where

$$\Gamma_{1,n}=\frac{1}{n!}\Gamma_1^{(n)}(1),\ \ \Gamma_{2,n}=\frac{1}{n!}\Gamma_2^{(n)}(1),\ \ \Gamma_{3,n}=\frac{1}{n!}\Gamma_3^{(n)}(1),\ \ \Gamma_{4,n}(\xi)=\frac{1}{n!}\Gamma_4^{(n)}(1),$$

$$\Gamma_{5,n}=\frac{1}{n!}\Gamma_5^{(n)}(1),\ \ \Gamma_{6,n}(\xi)=\frac{1}{n!}\Gamma_6^{(n)}(1).$$

then the solution of Eqs. (12) can be expressed as Taylor's series at $\xi=1$ as:

$$\widetilde{T}_n(\xi,s)=\sum_{k=0}^{\infty}A_k(s)(\xi-1)^k\,,\widetilde{u}_n(\xi,s)=\sum_{k=0}^{\infty}B_k(s)(\xi-1)^k\,,\widetilde{w}_n(\xi,s)=\sum_{k=0}^{\infty}D_k(s)(\xi-1)^k\,. \tag{15}$$

Substituting Eqs. (15) into Eq. (12) and employing series properties in mathematics, one can obtain the following recurrence relations

$$(k+1)(k+2)A_{k+2}=s\sum_{j=0}^{k}\left[A_{j-1}+A_j\right]\Gamma_{6,k-j}-(k+1)^2A_{k+1}-\sum_{j=0}^{k}\left[jA_j+(j+1)A_{j+1}\right]\Gamma_{5,k-j}$$
$$+(a^2-k^2)A_k+2a^2A_{k-1}+a^2A_{k-2}$$

 ISBN: 978-84-614-6167-7

$$
\begin{aligned}
-(k+1)(k+2)B_{k+2} &= (k+1)(2k+1)B_{k+1} + \left[k^2 - K_n{}^2\left(\frac{1-2\nu}{2-2\nu}\right) - 1\right]B_k \\
&- K_n{}^2\left(\frac{1-2\nu}{2-2\nu}\right)\left[2B_{k-1} + B_{k-2}\right] + \sum_{j=0}^{k}\left[\Gamma_{1,j-2} + 2\,\Gamma_{1,j-1} + \Gamma_{1,j}\right](k+1-j)B_{k+1-j} \\
&+ \left(\frac{\nu}{1-\nu}\right)\sum_{j=0}^{k}\left[\Gamma_{1,j-1} + \Gamma_{1,j}\right]B_{k-j} - s^2\left(\frac{(1+\nu)(1-2\nu)}{1-\nu}\right)\sum_{j=0}^{k}\left[\Gamma_{3,j-2} + 2\,\Gamma_{3,j-1} + \Gamma_{3,j}\right]B_{k-j} \\
&- K_n\left(\frac{1}{2-2\nu}\right)\left[(k+1)D_{k+1} + 2kD_k + (k-1)D_{k-1}\right] \\
&- K_n\left(\frac{\nu}{1-\nu}\right)\sum_{j=0}^{k}\left[\Gamma_{1,j-2} + 2\Gamma_{1,j-1} + \Gamma_{1,j}\right]D_{k-j} - \left(\frac{1+\nu}{1-\nu}\right)\sum_{j=0}^{k}\left[\Gamma_{2,j-2} + 2\,\Gamma_{2,j-1} + \Gamma_{2,j}\right]A_{k-j} \\
&- \left(\frac{1+\nu}{1-\nu}\right)\sum_{j=0}^{k}\left[\Gamma_{4,j-2} + 2\,\Gamma_{4,j-1} + \Gamma_{4,j}\right](k+1-j)A_{k+1-j} \\
-(k+1)(k+2)D_{k+2} &= (k+1)(2k+1)D_k + 1 + k^2 D_k - K_n{}^2\left(D_k + 2D_{k-1} + D_{k-2}\right) \\
&+ K_n\left(\frac{1}{1-2\nu}\right)\left[(k+1)B_{k+1} + (2k+1)B_k + kB_{k-1}\right] \\
&+ \sum_{j=0}^{k}\left[\Gamma_{1,j-2} + 2\,\Gamma_{1,j-1} + \Gamma_{1,j}\right](k+1-j)D_{k+1-j} - 2(1+\nu)s^2\sum_{j=0}^{k}\left[\Gamma_{3,j-2} + 2\,\Gamma_{3,j-1} + \Gamma_{3,j}\right]D_{k-j} \\
&+ K_n\sum_{j=0}^{k}\left[\Gamma_{1,j-2} + 2\Gamma_{1,j-1} + \Gamma_{1,j}\right]B_{k-j} - \left(\frac{2+2\nu}{1-2\nu}\right)K_n\sum_{j=0}^{k}\left[\Gamma_{4,j-2} + 2\,\Gamma_{4,j-1} + \Gamma_{4,j}\right]A_{k-j}
\end{aligned}
\tag{16}
$$

From Eqs. (16), all coefficients in Eqs. (15) can be obtained as follows

$$
\begin{aligned}
A_k(s) &= A_0 \mathrm{X}_{1k}(s) + A_1 \mathrm{X}_{2k}(s) \\
B_k(s) &= B_0 \Psi_{1k}(s) + B_1 \Psi_{2k}(s) + D_0 \Psi_{3k}(s) + D_1 \Psi_{4k}(s) + \Psi_{5k}(s) \\
D_k(s) &= B_0\,\Omega_{1k}(s) + B_1 \Omega_{2k}(s) + D_0 \Omega_{3k}(s) + D_1 \Omega_{4k}(s) + \Omega_{5k}(s).
\end{aligned}
\tag{17}
$$

where $\mathrm{X}(s)$, $\Psi(s)$ and $\Omega(s)$ can be derived from the corresponding recurrence expressions (16) and A_0, A_1, B_0, B_1, D_0 and D_1 are unknown constants, which can be determined by using the boundary conditions (13).

To determine the temperature, radial and axial displacements in time domain, the present work uses the fast Laplace inverse transform (FLIT) method (Durbin 1974).

4. Numerical Results and Discussion

Consider a FG hollow cylinder with finite length *l* and inner radius *a* and outer radius *b*. The thermomechanical properties of FG cylinder are considered to vary through the radial direction. The distribution of thermomechanical properties is modeled by a nonlinear power function in terms of volume fraction as follows

 ISBN: 978-84-614-6167-7

$$p = p_m V_m + p_c V_c = p_c + (p_m - p_c) V_m \quad , \quad V_m = \left(\frac{r-a}{b-a} \right)^{\beta} \tag{18}$$

and p is the effective thermomechanical property of FGM, V_m and V_c are volume fractions of metal and ceramic, and subscripts c and m stand for ceramic and metal, respectively. The term β is a non-negative volume fraction exponent that governs the distribution of the constituent materials through the thickness of FG cylinder.

In order to study the validity of the presented method, the results have been compared with the results obtained by YING and WANG (2010) for isotropic thick hollow cylinder assuming β to be zero. Figures 1a-d show the comparison of the obtained results using presented analytical method and method used by YING and WANG (2010).

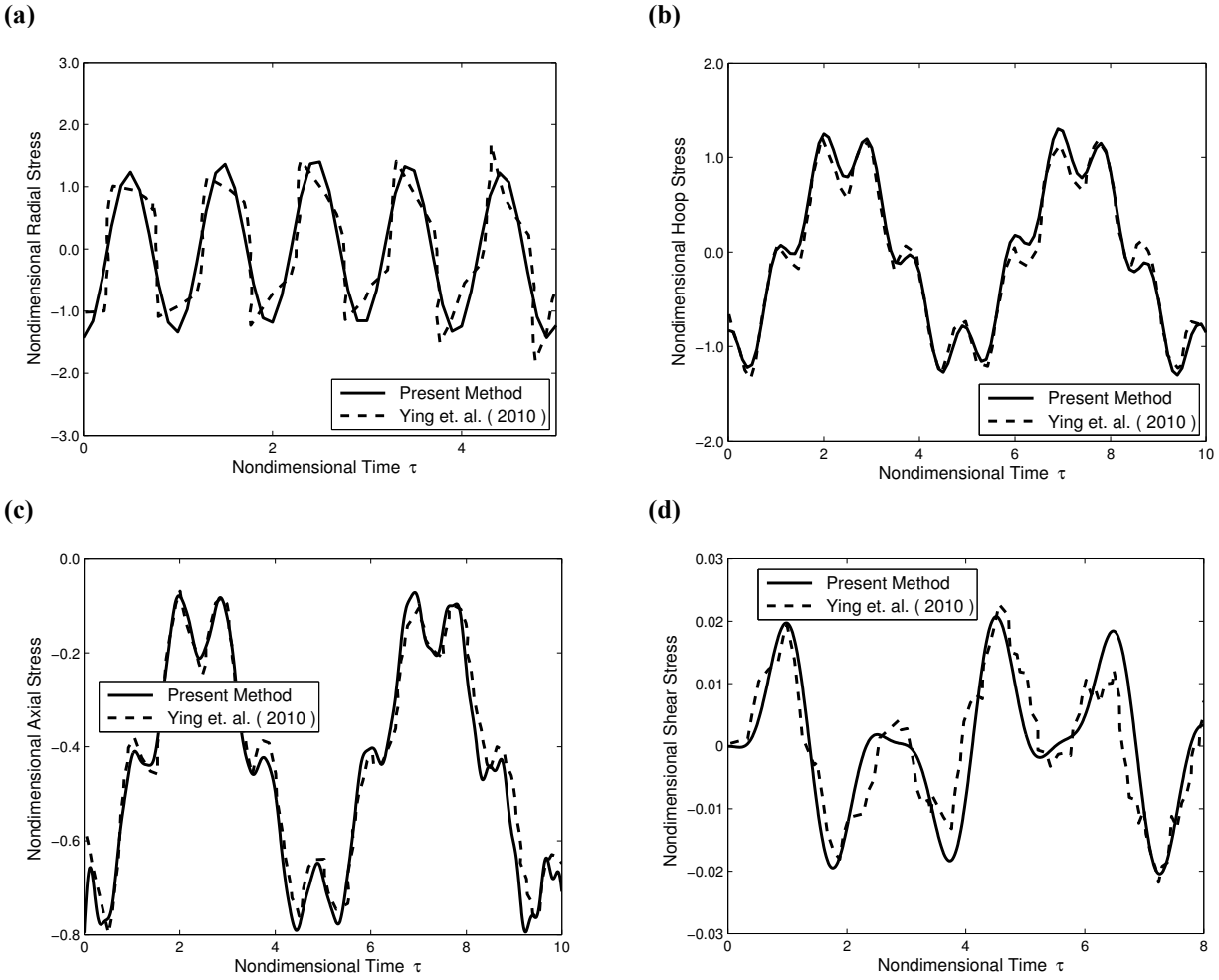


Fig. 1 Time histories of dynamic stresses for isotropic hollow cylinder at middle point of thickness, **(a)** radial stress $\overline{\sigma}_r$, **(b)** hoop stress $\overline{\sigma}_\theta$, **(c)** axial stress $\overline{\sigma}_z$, **(d)** shear stress $\overline{\sigma}_{zr}$

 ISBN: 978-84-614-6167-7

To study the effects of thermal loading on dynamic behavior of short FG cylinder, the bounding surfaces of the cylinder is assumed to be traction free and supposed that the FG cylinder is subjected to the thermal loading as Eq. (13) at inner and outer surfaces. The results are obtained for a ceramic-metal functionally graded finite length cylinder with alumina and aluminum as the ceramic and metal constituents, respectively. The material properties given in Table 1 are used in computing the numerical results.

Table 1. Thermo-mechanical properties of FG cylinder (OOTAO 2009)

Material	E (GPa)	ρ (kg/m^3)	α (1/K)	λ (W/m.K)	ν	k (m^2/s)
Alumina(Inner Surface)	343	3880	8.0×10^{-6}	36	0.33	11.9×10^{-6}
Aluminum Alloy(Outer Surface)	70	2688	23.6×10^{-6}	222	0.33	90.6×10^{-6}

Effect of gradient parameter β on the dynamic response of radial stress at middle point of thickness and $Z = L/2$ is shown in Fig. 2. The amplitude of variation and the peak values of radial stresses are increasing when the value of β is increased. Increasing gradient parameter β enables us to have a hollow cylinder with more ceramic volume fraction. It can be concluded from Fig. 2 when the value of β is increased, the radial stress wave propagation speed is also increased. The stress wave propagation speed C_v in a pure ceramic cylinder is larger than speed in a pure metal cylinder because the elasticity modulus of the alumina is greater than the aluminum elasticity modulus. This phenomenon can be concluded from Fig. 2. When the distance between two peak points in each diagram is decreased, it means that the wave propagation velocity is decreased. Figure 3 depicts the effects of gradient parameter β on time history of hoop stress at the inner surface of FG cylinder for $Z = L/2$, respectively. We can see that by increasing the value of gradient parameter β, the peak values of hoop stresses are decreased. Also the effects of gradient parameter β on dynamic responses of axial stress at various points on thickness of FG cylinder for $Z = L/3$ are illustrated in Fig. 4. To assess the dynamic behavior of shear stress in finite length FG cylinder, the time histories of shear stress are drawn in Fig. 5. Figure 5 shows that the amplitude of variation in shear stress diagrams are increased when the values of gradient parameter β are decreased. It means that the shear stresses converge to big values when the FG cylinder converges to full metal cylinder.

5. Concluding Remarks

In this article, an analytical method based on Laplace transform and series solution is developed to study the two-dimensional thermoelastic stresses in a finite length functionally graded thick hollow cylinder under thermal loading at

 ISBN: 978-84-614-6167-7

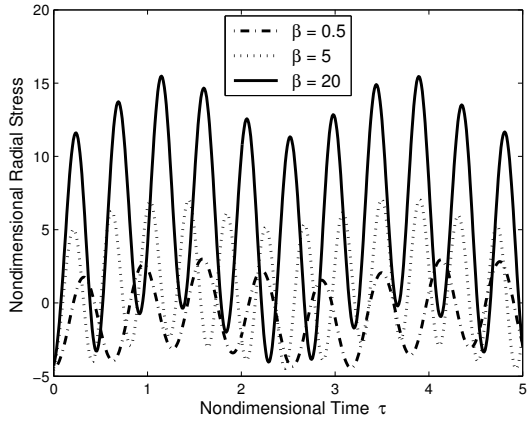


Fig. 2 Time history of radial stress $\overline{\sigma}_r$ at middle point of thickness and $Z = L/2$ for various β

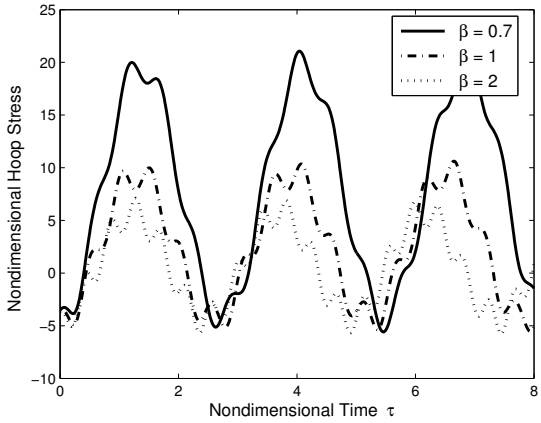


Fig. 3 Time history of hoop stress $\overline{\sigma}_\theta$ at inner surface for $Z = L/2$ and various β

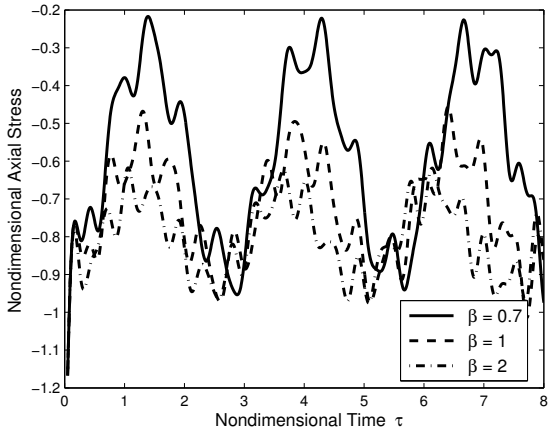


Fig. 4 Time history of axial stress $\overline{\sigma}_z$ for $Z = L/2$ and various β at outer surface

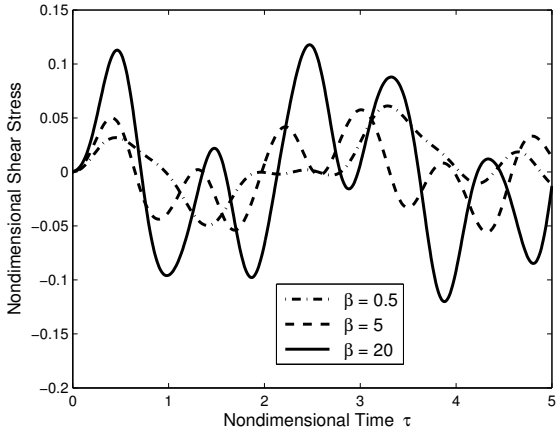


Fig. 5 Time history of shear stress $\overline{\sigma}_{zr}$ at middle point of thickness for $Z = L/2$ and various β

boundaries. The cylinder is considered with simply-supported ends and traction free on bounding surfaces. Thermomechanical properties of the FG cylinder are assumed to be temperature independent and vary continuously and smoothly through the radial direction as a nonlinear power function in terms of volume fraction. The dynamic behaviors of thermoelastic stresses are obtained for various grading patterns FGMs in some points across the thickness of FG cylinder. The main results of this article can be outlined as follows:

– By increasing the value of gradient parameter, the radial stress wave propagation speed increases, also, the peak values of hoop and axial stresses decrease, but the peak values of radial and shear stresses increase.

 ISBN: 978-84-614-6167-7

- From engineering perspective, the presented data and results furnish a ground for optimum design of FG cylinder. For example in the studied case of paper, the values of dynamic stresses are calculated in various conditions, which can be used for practical design of FG pressure vessels.
- The presented analytical method has a high capability to use in wave propagation analysis of FG structures.

6. References

[1] M. JABBARI, A. BAHTUI, M.R. ESLAMI, *Axisymmetric mechanical and thermal stresses in thick long FGM cylinders*, J. of Thermal Stresses. **29** (2006) 643-663.

[2] M. JABBARI, A. BAHTUI, M.R. ESLAMI, *Axisymmetric mechanical and thermal stresses in thick short length FGM cylinders,* J. Press. Vessels Pip. **86** (2009) 296-306.

[3] Z.S. SHAO, T.J. WANG, K. ANG, *Transient thermomechanical analysis of functionally graded hollow circular cylinderS,* J. of Thermal Stresses. **30** (2007) 81-104.

[4] S.M. HOSSEINI, M. AKHLAGHI, M. SHAKERI, *Dynamic response and radial wave propagation velocity in thick hollow cylinder made of functionally graded materials*. Eng. Comput. **24** (2007) 288-303.

[5] M. SHARIYAT, S.M.H. LAVASANI, M. KHAGHANI, *Nonlinear transient thermal stress and elastic wave propagation analyses of thick temperature-dependent FGM cylinders, using a second-order point-collocation method,* Appl. Math. Modell. DOI:10.1016/j.apm.2009.07.007.

[6] M. TAHANI, S.M. HOSSEINI, A. SAFARI-KAHNAKI, A. TODOROKI, *Transient and dynamic stress analysis of functionally graded thick hollow cylinder subjected to thermal shock loading using an analytical method*, JSME Int. J. Ser. A. **4** (2010) 1-14.

[7] A. SAFARI-KAHNAKI, M. TAHANI, S.M. HOSSEINI, *Two-dimensional dynamic analysis of thermal stresses in a finite-length FG thick hollow cylinder subjected to thermal shock loading using an analytical method*, Acta Mech. (2011) DOI 10.1007/s00707-011-0478-y.

[8] F. DURBIN, *Numerical inversion of laplace transforms: an efficient improvement to Dubner and Abate's method,* Computer Journal. **17** (1974) 371–376.

[9] J. YING, H. M. WANG, *Axisymmetric themoelastic analysis in a finite hollow cylinder due to nonuniform thermal shock*, International Journal of Pressure Vessels and Piping. **87** (2010) 714-720.

[10] Y. OOTAO, *Transient thermoelastic and piezothermoelastic problems of functionally graded materials,* J. of Thermal Stresses. **32** (2009) 656-697.

 ISBN: 978-84-614-6167-7

Proceedings of the 11th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE2011
Benidorm, Spain, 26-30 June 2011

Computational Modelling of Some Problems of Elasticity and Viscoelasticity and Non-Fickian Viscoelastic Diffusion

Simon Shaw, M.K. Warby and J.R. Whiteman
BICOM, Mathematical Sciences, Brunel University, Uxbridge UB8 3PH, UK

Emails: simon.shaw@brunel.ac.uk, mike.warby@brunel.ac.uk,
john.whiteman@brunel.ac.uk

Abstract

The reliability of computational models of physical processes has received much attention and involves issues such as the validity of the mathematical models being used, the error in any data that the models need, and the accuracy of the numerical schemes being used. These issues are considered in the context of elastic and hyperelastic deformation, when finite element approximations are applied. Goal oriented techniques using specific quantities of interest (QoI) are used for estimating discretisation and modelling errors.

The computational modelling of the rapid large inflation of hyperelastic circular sheets modelled as axisymmetric membranes is treated, with the aim of estimating engineering QoI and their errors. Fine (involving inertia terms) and coarse (quasi-static) models of the inflation are considered. The techniques are applied to thermoforming processes where sheets are inflated into moulds to form thin-walled structures.

The case of a penetrant diffusing into a polymer body, where the diffusion is non-Fickian and induces a steep travelling wave, is then considered and references to finite element results are given.

Key words: elasticity, viscoelasticity, hyperelasticity, non-Fickian diffusion, finite element modelling, goal oriented methods, thermoforming

1. Introduction

The process of computational modelling for problems of continuum mechanics consists of two main phases. The mathematical model of the physics (reality) has first to be defined, after which a numerical approximation of the model has to be

 ISBN: 978-84-614-6167-7

derived and solved to give a numerical solution in terms of quantities of interest (QoI). As each of these phases introduces error, in addition to any error in the data of the problem, the *reliability* of the process is acknowledged to be of great importance. The process of assessment of the error in the mathematical model, modelling error, is called *validation*, whilst that of the error in the numerical approximation is *verification.* Reliability is directly related to *validation* and *verification* (V & V) and is increasingly being studied; see e.g. Babuška et al. [1] and Babuška et al. [2].

In this short review paper we consider computational modelling of problems of elasticity, hyperelasticity and viscoelasticity using finite element methods. Thinking first of *verification* we present various *a priori* error analyses and *a posteriori* error estimators, with references to papers where these have been derived. These are followed by brief descriptions of a number of applications, with numerical results for the QoI. The *validation* of the models in the context of some of these applications is then addressed using goal oriented techniques as proposed by Oden and Prudhomme [3] and applied by Shaw et al. in [4].

In order to lead up to computational models for these, in the next section we proceed first with a framework for describing deformation and defining our notation, and then progress to hyperelastic (large) deformation. We then discuss problems of viscoelasticity and finally of non-Fickian diffusion.

2. Mathematical models, weak formulations and finite element methods

2.1 Solid Mechanics Framework (Small Displacement Case)

Let $\mathcal{G}$ be a compressible solid body with mass density ρ which in its undeformed state occupies the open bounded domain $\Omega \subset \mathbb{R}^n, n = 2,3$ with polygonal/polyhedral boundary $\partial\Omega$. A point in $\bar{\Omega} \equiv \Omega \cup \partial\Omega$ is denoted by $\boldsymbol{x} \equiv (x_i)_{i=1}^3$, when n=3. The boundary $\partial\Omega$ is partitioned into disjoint subsets Γ_D and Γ_N such that $\partial\Omega \equiv \Gamma_D \cup \Gamma_N$, $\Gamma_D \cap \Gamma_N = \varnothing$ and meas $(\Gamma_D) > 0$. Suppose that, for time $t \in I \equiv (0,T]$, $T > 0$, the body $\mathcal{G}$ is acted upon by body forces $\boldsymbol{f}(x,t) \equiv (\boldsymbol{f}_i(\boldsymbol{x},t))_{i=1}^3$ for $\boldsymbol{x} \in \Omega$ and surface tractions $\boldsymbol{g}(\boldsymbol{x},t) \equiv (g_i(\boldsymbol{x},t))_{i=1}^3$ for $\boldsymbol{x} \in \Gamma_N$. The displacement at a point $\boldsymbol{x}$ under the action of the forces $\boldsymbol{f}$ and $\boldsymbol{g}$ is $\boldsymbol{u} \equiv (u_i(x,t))_{i=1}^3$, $\boldsymbol{x} \in \Omega$, $t \in I$, and with a small displacement assumption $\boldsymbol{x} + \boldsymbol{u} \approx \boldsymbol{x}$, so that we do not need to distinguish

 ISBN: 978-84-614-6167-7

between the deformed and undeformed domains in most terms. Let $\underline{\boldsymbol{\sigma}} \equiv \left(\sigma_{ij}\right)_{i,j=1}^{3} \equiv \left(\sigma_{ij}\left(\boldsymbol{x},t\right)\right)_{i,j=1}^{3}$ denote the stress resulting from the deformation.

Applying Newton's second law of motion, relating force to the rate of change of linear momentum, to this configuration we obtain the momentum equations

$$\rho\left(\boldsymbol{x}\right)\ddot{u}_i\left(\boldsymbol{x},t\right) - \sigma_{ij,j}\left(\boldsymbol{x},t\right) = f_i\left(\boldsymbol{x},t\right), \quad i = 1,2,3 \text{ in } \Omega \times I \tag{1}$$

and these together with the boundary and initial conditions

$$u_i\left(\boldsymbol{x},t\right) = 0 \text{ in } \Gamma_D \times \bar{I} \tag{2}$$

$$\sigma_{ij}\hat{n}_j = g_i\left(\boldsymbol{x},t\right), \text{ in } \Gamma_N \times \bar{I} \tag{3}$$

$$u_i\left(\boldsymbol{x},0\right) = u_i^0\left(\boldsymbol{x}\right), \boldsymbol{x} \in \Omega, \tag{4}$$

$$\dot{u}_i\left(\boldsymbol{x},0\right) = u_i^1\left(\boldsymbol{x}\right), \boldsymbol{x} \in \Omega, \tag{5}$$

define the *dynamic* deformation problem, where $\hat{\boldsymbol{n}} \equiv \left(\hat{n}_i\right)_{i=1}^{n}$ is the unit outward normal to Γ_N, the Einstein convention has been used, and $v_{,j} \equiv \partial v / \partial x_j$.

If the inertia terms can be neglected in the deformation and assuming that $\boldsymbol{u} = 0 \quad \forall t < 0$, we obtain the quasistatic problem, where $i, j = 1,2,3$,

$$-\sigma_{ij,j}\left(\boldsymbol{x},t\right) = f_i\left(\boldsymbol{x},t\right), \text{ in } \Omega \times I \tag{6}$$

$$u_i\left(\boldsymbol{x},\boldsymbol{t}\right) = 0, \text{ in } \Gamma_D \times \bar{I} \tag{7}$$

$$\sigma_{ij}\hat{n}_j = g_i\left(\boldsymbol{x},t\right), \text{ in } \Gamma_N \times \bar{I}, \tag{8}$$

In order to complete the definitions of the dynamic and quasistatic problems it is necessary to have a constitutive relationship connecting the stress to the displacement and its derivatives. The constitutive relationship reflects the behaviour of the material of the body $\mathcal{G}$.

2.2 Linear elasticity and weak formulation

In the case of small displacement gradients the strain is described by the infinitesimal strain tensor $\underline{\boldsymbol{\varepsilon}}\left(\boldsymbol{u}\right) \equiv \left(\varepsilon_{ij}\left(\boldsymbol{u}\right)\right)_{i,j=1}^{n}$ as

$$\varepsilon_{ij}\left(\boldsymbol{u}\right) = \frac{1}{2}\left(\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i}\right), \; i, j = 1,2,3. \tag{9}$$

 ISBN: 978-84-614-6167-7

For an isotropic linear elastic material Hooke's law connects stress to strain (i.e. to the derivatives of the displacement) and we have

$$\sigma_{ij} = \lambda \nabla . \boldsymbol{u} \delta_{ij} + \mu \varepsilon_{ij}(\boldsymbol{u}), \tag{10}$$

where δ_{ij} is the Kronecker delta, and λ and μ are the Lamé coefficients of the material. More generally the relation for a linear elastic material can be written in the form

$$\underline{\boldsymbol{\sigma}} = \underline{\boldsymbol{D}}\,\underline{\boldsymbol{\varepsilon}}. \tag{11}$$

We note that elasticity is a time independent phenomenon, so that the mathematical model for linear elasticity is based on equations (6), (7), (8) and (10) with $\boldsymbol{u}(\boldsymbol{x},t)$ depending only on quantities at time *t*.

In order to obtain a weak form from these equations we introduce the usual Sobolev spaces $H^r(\Omega), r = 0,1,\ldots,$ and for $V_1, V_2, \ldots, V_n \subset H^r(\Omega)$ we define the space *V*, such that

$$\begin{aligned} V &\equiv V_1 \times V_2 \times \ldots \times V_n \\ &\equiv \left\{ \boldsymbol{v} \in \left(H^1(\Omega) \right)^n : \boldsymbol{v} = 0 \text{ on } \Gamma_D \right\}. \end{aligned} \tag{12}$$

Multiplying (6) by a test function $\boldsymbol{v} \in V$, and integrating by parts over Ω we obtain

$$\int_\Omega \sigma_{ij}(\boldsymbol{u}) \varepsilon_{ij}(\boldsymbol{v}) d\Omega = (\boldsymbol{f}, \boldsymbol{v})_\Omega + (\boldsymbol{g}, \boldsymbol{v})_{\Gamma_N}, \tag{13}$$

where the (.,.) are inner products. Applying Hooke's law (10) we obtain the weak form of the isotropic linear elasticity problem: find $\boldsymbol{u} \in V$ such that

where

$$\begin{aligned} &a(\boldsymbol{u}, \boldsymbol{v}) = L(\boldsymbol{v}) \; \forall \boldsymbol{v} \in V, \\ &a(\boldsymbol{v}, \boldsymbol{w}) \equiv \int_\Omega \lambda \nabla . \boldsymbol{v} \nabla . \boldsymbol{w} + \mu \varepsilon_{ij}(\boldsymbol{v}) \varepsilon_{ij}(\boldsymbol{w}) d\Omega, \\ &L(\boldsymbol{v}) \equiv \int_\Omega f_i v_i \partial\Omega + \int_{\Gamma_N} g_i v_i d\Gamma, \;\; i, j = 1,2,3. \end{aligned} \tag{14}$$

In order to apply the finite element method to problem (14), we first partition Ω into a set of elements $\left\{ \Omega_i^h \right\}_{i=1}^{N_E}$, where $\bar{\Omega}^h \subset \bar{\Omega} = \mathrm{U}_i \Omega_i^h$ and $\partial\Omega^h \equiv \partial\Omega$, each with diameter h_i and define $h \equiv \max\limits_{1 \le i \le N_E} h_i$ We construct finite dimensional spaces

$V_i^h \equiv \text{span} \left\{ \Phi_i(\boldsymbol{x}) \right\}_{i=1}^{N_N} \subset V_i$, for $1 \le i \le n$ with each $\Phi_i \in \mathbb{P}^r$, a piecewise

 ISBN: 978-84-614-6167-7

polynomial of degree r over the partition, where the $\Phi_i(\boldsymbol{x})$ are basis functions associated with the N_N nodes of the partition. Finally we define

$$V^h \equiv V_1^h \times V_2^h \times \ldots \times V_n^h \subset V$$

The finite element problem is: find $\boldsymbol{u_h} \in V^h$ such that

$$a(\boldsymbol{u}_h, \boldsymbol{v}_h) = L(\boldsymbol{v}_h) \quad \forall \boldsymbol{v}_h \in V^h. \tag{15}$$

There is a vast literature associated with the derivation of *a priori* estimates for the error $\boldsymbol{e}_h \equiv (\boldsymbol{u} - \boldsymbol{u}_h)$ of the form

$$\|\boldsymbol{e}_h\|_{\alpha,\Omega} \leqq \mathcal{C}\, h^{\beta(\alpha,r)} |\boldsymbol{u}|_{r,\Omega}, \tag{16}$$

where $\| \ \|_{q,\Omega}$ is the norm on $H^q(\Omega)$; see e.g. Ciarlet [5], Oden and Reddy [6] and Whiteman [7]. In (16) the function $\beta(\alpha, r)$ depends on the regularity of the solution $\boldsymbol{u}$ of (14) and $\mathcal{C}$ is a constant that depends on α but is independent of $\boldsymbol{u}$ and the mesh. Estimates of this type provide rates of convergence of $\boldsymbol{u_h}$ to $\boldsymbol{u}$ with decreasing mesh size h.

Similarly many *a posteriori* error estimators, (i.e. calculable error estimators involving the calculated solution $\boldsymbol{u}_h$) and based for example on residuals $R(\boldsymbol{v})$ of the type

$$R(\boldsymbol{v}_h) = \sum_{i=1}^{N_E} L(\boldsymbol{u}_h) - a(\boldsymbol{u}_h, \boldsymbol{v}_h), \tag{17}$$

have now been derived, see e.g. Oden and Ainsworth [8] and Babuska et al [2], and again the performance of these depends on the regularity of $\boldsymbol{u}$. The process of verification is made possible by the use of estimates of the type of (16) and (17).

2.3 Large Deformation Elasticity and Application to Thermoforming Processes

Motivated by the problem of the large deformation of a thin polymer sheet that will be considered later, we now describe the large elastic deformation under the action of applied pressure loading for the case of an elastic sheet, $\mathcal{B}$, using a Lagrangian description. Again $\boldsymbol{x}$ denotes a point in the body which in the deformation undergoes a displacement $\boldsymbol{u}$ so that $\boldsymbol{x} \to \boldsymbol{x} + \boldsymbol{u} \equiv \boldsymbol{w}$. In the large deformation case $\boldsymbol{u}$ and the displacement gradient are no longer small so that care is needed to distinguish between the undeformed and the deformed states. An outcome of this in the description of the deformation is to introduce the nominal stress $\underline{\Pi} \equiv (\det \underline{\boldsymbol{F}})\, \underline{\boldsymbol{F}}^{-1} \underline{\boldsymbol{\sigma}}$, where $\underline{\boldsymbol{F}} \equiv (\partial w_i / \partial x_j)$, $j = 1, 2, 3$ is the deformation gradient and, as before, $\boldsymbol{\sigma}$ is the Cauchy stress.

 ISBN: 978-84-614-6167-7

The equations of equilibrium of a body undergoing large elastic deformation, corresponding to (6) for small deformation, can for the three-dimensional case be written as

$$-\sum_{j=1}^{3}\frac{\partial \Pi_{ij}}{\partial x_j} = f_i \qquad i = 1,2,3. \tag{18}$$

The problem that we shall consider involves the large deformation of a thin sheet, with mid-surface Ω, which is clamped on the boundary Γ of Ω and which in its undeformed state has thickness h_0. The sheet occupies the region

$$\mathcal{B} \equiv \left\{ \left(x_1, x_2, x_3\right) : \boldsymbol{x} = \left(x_1, x_2\right)^T \in \Omega, \left|x_3\right| < h_0 / 2 \right\}. \tag{19}$$

Now $x_3 = 0$ on the mid-surface Ω, which deforms as $\left(x_1, x_2, 0\right) \to \left(x_1 + u_1, x_2 + u_2, u_3\right)$, and, assuming that normals to Ω remain normal, we obtain a two-dimensional description of the sheet with $\boldsymbol{u} = \boldsymbol{u}\left(\boldsymbol{x}\right)$. The sheet is modelled as a membrane, thus being unable to support bending so that $\underline{\boldsymbol{\sigma}}.\boldsymbol{n} = 0$, where $\boldsymbol{n}$ is the unit normal to the deformed mid-surface Ω.

In this context of a membrane approximation to the general three-dimensional case, the two-dimensional equations for the problem, when there is a pressure loading P and assuming that the body forces $\boldsymbol{f}$ are zero, lead to the weak form of (18): find $\boldsymbol{u} \in V$ such that

$$a\left(\boldsymbol{u};\boldsymbol{v}\right) = 0 \quad \forall \boldsymbol{v} \in V, \quad a\left(\boldsymbol{u};\boldsymbol{v}\right) = a_1\left(\boldsymbol{u};\boldsymbol{v}\right) - P a_2\left(\boldsymbol{u};\boldsymbol{v}\right), \tag{20}$$

where

$$a_1\left(\boldsymbol{u};\boldsymbol{v}\right) \equiv \int_{\Omega} h_0 \left(\Pi^T : \nabla \boldsymbol{v}\right) d\Omega, \tag{21}$$

$$a_2\left(\boldsymbol{v};\boldsymbol{w}\right) \equiv \int \boldsymbol{v}.\left(\frac{\partial \boldsymbol{w}}{\partial x_1} \times \frac{\partial \boldsymbol{w}}{\partial x_2}\right) d\Omega, \tag{22}$$

and now the space $V \equiv V_1 \times V_2$ is such that

$$V \equiv \left\{ \boldsymbol{v} \in \left(H^1\left(\Omega\right)\right)^2 : \boldsymbol{v} = 0 \text{ on } \Gamma \right\}. \tag{23}$$

The finite element method is applied to obtain an approximation $\boldsymbol{u}_h$ to $\boldsymbol{u}$ the solution of (20) for the case of incremental loading of the sheet. As we have a Lagrangean description of the deformation, the spatial mesh is defined on the reference configuration $\Omega \subset ¡^2$ and, for each load increment P_j, the nonlinear system

$$a_1\left(\boldsymbol{u}_h;\boldsymbol{v}_h\right) - P_j a_2\left(\boldsymbol{u}_h;\boldsymbol{v}_h\right) = 0 \quad \forall \boldsymbol{v}_h \in V^h, \tag{24}$$

 ISBN: 978-84-614-6167-7

is solved for $\boldsymbol{u}_h \in V^h$ using Newton's method, where $V^h \equiv V_1^h \times V_2^h \subset V$ and the $V_i^h, i = 1,2$ are spaces of piecewise polynomial functions defined over the partition of Ω, see e.g. Karamanou et al. [8].

For the problem (20), noting that $a(\boldsymbol{u};\boldsymbol{v})$ is a semilinear form (i.e. linear in arguments to the right of the semi-colon), we suppose that we wish to approximate the quantity of interest $J(\boldsymbol{u}), \boldsymbol{u} \in V$ with $J(\boldsymbol{u}_h), \boldsymbol{u}_h \in V^h$.

If $a'(.;.)$ and $J'(.;.)$ are Gateaux derivatives of $a(.;.)$ and $J(.;.)$ respectively then, if $\boldsymbol{e}_h = \boldsymbol{u} - \boldsymbol{u}_h$,

$$J(\boldsymbol{u}) - J(\boldsymbol{u}_h) = \int_0^1 J'(\boldsymbol{u}_h + s\boldsymbol{e}_h;\boldsymbol{e}_h)ds \tag{25}$$

and

$$\begin{aligned} -a(\boldsymbol{u}_h;\boldsymbol{z}) &= a(\boldsymbol{u};\boldsymbol{z}) - a(\boldsymbol{u}_h;\boldsymbol{z}) \\ &= \int_0^1 a'(\boldsymbol{u}_h + s\boldsymbol{e}_h;\boldsymbol{e}_h,\boldsymbol{z})ds. \end{aligned} \tag{26}$$

If we now consider the (dual) linear problem find $\boldsymbol{z} \in V^h$ such that

$$\int_0^1 a'(\boldsymbol{u}_h + s\boldsymbol{e}_h;\boldsymbol{v}_h,\boldsymbol{z})ds = \int_0^1 J'(\boldsymbol{u}_h + s\boldsymbol{e}_h;\boldsymbol{v}_h)ds \quad \forall \boldsymbol{v}_h \in V^h, \tag{27}$$

then we have a representation of the error as

$$J(\boldsymbol{u}) - J(\boldsymbol{u}_h) = -a(\boldsymbol{u}_h;\boldsymbol{z}). \tag{28}$$

But $\boldsymbol{z}$ depends on $\boldsymbol{u}$ so that (27) cannot be solved as it stands and some form of approximation has to be adopted. One strategy for this is to apply the left hand rule for the integration giving

$$a'(\boldsymbol{u}_h;\boldsymbol{v}_h,\hat{\boldsymbol{z}}) = J'(\boldsymbol{u}_h;\boldsymbol{v}_h), \quad \forall \boldsymbol{v}_h \in \hat{V}^h, \tag{29}$$

where $\hat{V}^h \subset \hat{V}^h$, giving the estimate

$$J(\boldsymbol{u}) - J(\boldsymbol{u}_h) \approx -a(\boldsymbol{u}_h,\hat{\boldsymbol{z}}). \tag{30}$$

This "machinery" for estimating the discretisation error will be referenced for a problem of free inflation of a thin polymer sheet in a later section.

We have so far treated only discretisation error. In order to consider modelling error we introduce the concept of *fine* and *coarse* problems in the context of the deformation of the sheet. For example the problem (20) which is quasi-static could be taken as a coarse problem and the fine problem could be similar, but with the inclusion of inertia terms in (18). In many practical situations it is not clear whether inertia terms are important to the modelling. Suppose therefore that the fine problem has the weak form

 ISBN: 978-84-614-6167-7

$$A(\boldsymbol{U};\boldsymbol{v}) = 0 \quad \forall v \in V, \tag{31}$$

where the semilinear form $A(.;.)$ contains the $a_1(.;.)$ and $a_2(.;.)$ of (20) and the inertia terms. The dual approximating problem corresponding to (29) is now

$$A'(\boldsymbol{u}_h;\boldsymbol{v}_h,\hat{\boldsymbol{z}}) = J'(\boldsymbol{u}_h;\boldsymbol{v}_h) \quad \forall \boldsymbol{v}_h \in \tilde{V}^0, \tag{32}$$

where the $\boldsymbol{u}_h$ is the solution of the coarse problem and now $\tilde{V}^0 \subset V$ is an appropriate finite dimensional space. We now have the estimate

$$J(\boldsymbol{U}) - J(\boldsymbol{u}_h) \approx -A(\boldsymbol{u}_{h_j}\hat{\boldsymbol{z}})$$

for the combined modelling and discretisation errors. More details of these goal oriented techniques can be found in Shaw et al. [4].

The large hyperelastic deformation of thin polymeric sheets into moulds as in thermoforming processes has been considered in [8] and [14]. The sheets are clamped at the edges and are acted on by pressure. The deformation has two stages; the first is free inflation prior to contact with mould, the second is inflation after part of the sheet has made contact with the mould. Considering first the free inflation phase, the goal oriented techniques of this section have been applied to the deformation of a circular sheet, modelled as described taking (20) without inertia terms as the coarse model and (31), including inertia terms, as the fine model. Estimates of both the discretisation error for the finite element approximation of the coarse model and of the combined modelling and discretisation errors for the fine model have been obtained; see [4], and these demonstrate that the method is robust for this free inflation problem under the given conditions.

The second phase of the inflation has in recent years been modelled extensively under the assumptions that the deformation is quasistatic and that perfect sticking of the sheet to the mould takes place on contact, see e.g. Karamanou et al. [8] and Warby et al. [14]. To our knowledge no error estimates have been obtained for this case.

2.4 Viscoelasticity

The formation and use of non-metallic materials has been one of the great advances of science, engineering, medicine and manufacturing of recent years. A feature of the deformation of polymeric solid materials is that when they are subjected to sustained loading, in addition to an elastic response, they can exhibit time dependent creep. For example a polymer test specimen subjected to an instantaneously applied and sustained tensile loading will undergo an initial elastic (solid) deformation, followed over time by creep during which the specimen will continue to stretch. Creep is a viscous fluid effect and, due to the

 ISBN: 978-84-614-6167-7

dual elastic and viscous responses, materials that exhibit this type of behaviour are said to be **viscoelastic.** If the loading is removed from the solid it will experience an instantaneous elastic recovery followed by a reverse time dependent recovery in which the solid returns to its original state. For this reason viscoelastic solids are said to possess **memory.**

Returning again to the case of small displacements and small strains, we recall that in the case of linear elasticity the constitutive relation was $\underline{\boldsymbol{\sigma}} = \underline{\boldsymbol{D}}\underline{\boldsymbol{\varepsilon}}$ as in (11). Turning now to viscoelastic deformation of the body $\mathcal{G}$, and assuming this to be both quasistatic and small, the deformation $\boldsymbol{u}(\boldsymbol{x},t)$ is governed by (6) – (8) for $(\boldsymbol{x},t) \in \Omega \times I$ and the strain $\underline{\boldsymbol{\varepsilon}}$ is as in (9). For linear elasticity the constitutive relation was Hooke's law (11), but in the case of linear viscoelasticity where the materials possess memory, i.e. the current stress depends on the history of the deformation, it is necessary to introduce time dependence and to augment Hooke's law with a memory term. In this way the stress can now be expressed as a linear functional of the strain, so that

$$\underline{\sigma}(\boldsymbol{x},t) = \underline{\boldsymbol{D}}(\boldsymbol{x})\underline{\boldsymbol{\varepsilon}}\left(\boldsymbol{u}(\boldsymbol{x},t)\right) - \int_0^t \frac{\partial \underline{\boldsymbol{D}}}{\partial s}(\boldsymbol{x},t-s)\underline{\boldsymbol{\varepsilon}}\left(u(\boldsymbol{x},s)\right)ds, \tag{33}$$

where $\underline{\boldsymbol{D}} \equiv \left(D_{ijkl}(\boldsymbol{x})\right)_{i,j,k,l=1}^{n}$ is a fourth order tensor of relaxation functions with components which are assumed to be $C^1(I)$ functions of *t.* At $t = 0$ it is assumed that $\underline{\boldsymbol{\varepsilon}} = 0$**.**

In order to define a weak formulation of this quasistatic problem we need a test space of admissible functions on $\Omega \times I$ and for this we proceed in two stages. We first multiply by a space only test function and integrate over Ω and then extend the test space and integrate over *I.* Multiplication of (6) by a (space only) function $\boldsymbol{v} \in V$**,** see (12), produces (13) for any $t \in I$**,** which on use of (31) leads to the problem: find $\boldsymbol{u} \in L_\infty(I;V)$ such that

$$a(\boldsymbol{u}(t),\boldsymbol{v}) = L(t;\boldsymbol{v}) + \int_0^t b(t,s;\boldsymbol{u}(s),\boldsymbol{v})ds \quad \forall \boldsymbol{v} \in V, \tag{34}$$

where

$$a(\boldsymbol{w},\boldsymbol{v}) \equiv \int_\Omega D_{ijkl}(0)\varepsilon_{kl}(\boldsymbol{w})\varepsilon_{ij}(\boldsymbol{v})d\Omega, \tag{35}$$

$$b(t;s;\boldsymbol{w},\boldsymbol{v}) \equiv \int_\Omega \frac{\partial D_{ijkl}}{\partial s}(t-s)\varepsilon_{kl}(\boldsymbol{w})\varepsilon_{ij}(\boldsymbol{v})d\Omega \tag{36}$$

for all $\boldsymbol{w},\boldsymbol{v} \in V$ and $L: I \times V \to \mathbb{R}$ is a time dependent linear form as in (14). As (34) contains no time derivative we seek the solution $\boldsymbol{u} \in L_\infty(I;V)$ by solving the "fully weak" problem: find $\boldsymbol{u} \in L_\infty(I;V)$ such that

 ISBN: 978-84-614-6167-7

$$a(\boldsymbol{u},\boldsymbol{v}) = L(v) \quad \forall v \in L_1(I;V), \tag{37}$$

where

$$a(\boldsymbol{u},\boldsymbol{v}) = \int_0^T \mathcal{A}(\boldsymbol{u}(t),\boldsymbol{v}(t))dt - \int_0^T \int_0^t \mathcal{B}(t,s;u(s),v(t))dsdt, \tag{38}$$

$$L(\boldsymbol{v}) = \int_0^t L(t;v(t))dt. \tag{39}$$

In order to apply the finite element method to (37) we first split the prismatic domain $\Omega \times I$ into M time slabs $\Omega_i \equiv \Omega \times I_i$ where $I_i = \{(t_{i-1},t_i)\}_{i=1}^M$ and partition each of these into M_i elements Ω_{ij} and define for each Ω_i the space

$$H_i \equiv \left\{ \boldsymbol{v} \in V \cap \left(C(\bar{\Omega})\right)^n, \text{ where } v \text{ is linear on } \Omega_{ij} \text{ for each } j = 1,...M_i \right\}$$

and hence the space-time finite element spaces $V^{r,h}$, where

$$V^{r,h} \equiv \left\{ \boldsymbol{v} \in L_\infty(I;V) : v\big|_{I_i} \in \mathrm{P}(I_i;H_i) \forall i = 1,2,3 \right\} \tag{40}$$

Functions in $V^{r,h}$ are continuous in space but usually discontinuous in time.
Many *a priori* error estimates have been derived for this type of finite element discretisation and take the form

$$\|\boldsymbol{u} - \boldsymbol{U}_h\|_{L_\infty(I;V)} \leqq \mathcal{C}(T)\left(\Pi_h \left\| hD^2\boldsymbol{u} \right\|_{L_\infty(I;L_2(\Omega))} + \Pi_k \left\| \frac{k^{r+1}\partial^{r+1}\boldsymbol{u}}{\partial t^{r+1}} \right\|_{L_\infty(I;V)} \right), \tag{41}$$

where $\boldsymbol{U}_h$ denotes the finite element approximation, $\mathcal{C}$(T) is a stability constant, the Π's are positive constants and *h* and *k* are maximum values of the space and time mesh lengths respectively. Note that the right hand side of (39) contains, as might be expected, a space and a time term. Further details can be found in Shaw and Whiteman [9] and [10] and Riviére et al. [11].

As viscoelastic materials display characteristics of both elastic solids and viscous fluids, many models involving combinations of springs and dashpots have been proposed, for example the Maxwell solid model, see e.g. Ferry [12]. For these cases the stress relaxation functions are represented using Prony series of decaying exponential functions so that the stress in (33) can be expressed in terms of internal variables of the model, for example internal stresses. These internal variables each satisfy ordinary differential equations in time; it is by integrating these ODE's that (33) can be obtained. Thus an alternative approach to the above history integral formulation of the linear viscoelastic problem is to solve a coupled system of PDE's consisting at each time step of an elastic problem of the type as in (14), but with internal variables contained in the right hand side, together with a system of ordinary differential equations in time for the internal variables; see e.g. Shaw et al. [13] where finite element models and error estimates are presented.

 ISBN: 978-84-614-6167-7

The formulation using internal variables has been extended to the case of finite viscoelastic deformation of a thin sheet, motivated by the large elastic deformation model described above using the nominal stress. Thus in this case we have a finite elasticity equation of the type (20) but now involving additional internal variable terms on the right hand side, which is solved coupled to a system of now nonlinear ODE's for the internal variables, see Karamanou [8].

2.5 Non-Fickian Diffusion

We finally consider the case of a penetrant diffusing into a polymer body Ω, over the time interval *I*. Thomas and Windle [16] demonstrated that this type of diffusion is non-Fickian, and that the diffusing penetrant induces differential stress with the result that a steep travelling wave front develops.

Cohen et al. [17] proposed a model for this diffusion in which the (scalar) function $u(\boldsymbol{x},t)$, the concentration of penetrant, and a scalar analogue $\sigma(\boldsymbol{x},t)$ of the stress satisfy the coupled system

$$\dot{u}(\boldsymbol{x},t) - \nabla^2 u(\boldsymbol{x},t) = f(\boldsymbol{x},t) + \nabla^2 \sigma(\boldsymbol{x},t) \text{ in } \Omega \times I, \tag{42}$$

$$\dot{\sigma}(\boldsymbol{x},t) + \gamma(u)\sigma(\boldsymbol{x},t) = u(\boldsymbol{x},t) \text{ in } \Omega \times I, \tag{43}$$

$$u(\boldsymbol{x},t) = 0 \text{ on } \Gamma_D \times \bar{I}, \tag{44}$$

$$\left(\nabla u(\boldsymbol{x},t) + \nabla \sigma(\boldsymbol{x},t)\right).\hat{\boldsymbol{n}} = g \text{ on } \Gamma_N \times \bar{I}, \tag{45}$$

$$u(\boldsymbol{x},0) = u_o(\boldsymbol{x}), \tag{46}$$

$$\sigma(\boldsymbol{x},0) = \sigma_o(\boldsymbol{x}), \tag{47}$$

where γ^{-1} is the relaxation time of the material and f, u_o, σ_o are given functions and constants.

If it is assumed that σ_o is zero, then solving the ODE (43) for σ and substitution into (42) produces at time *t* the nonlinear parabolic Volterra equation

$$\dot{u}(\boldsymbol{x},t) - \nabla^2 u(\boldsymbol{x},t) = f(\boldsymbol{x},t) + \nabla^2 \int_o^t e^{-\int_s^t \gamma(u(\boldsymbol{x},\varsigma))d\varsigma} u(\boldsymbol{x},s)\,ds. \tag{48}$$

Thus the non-Fickian diffusion problem can be considered either in the context of (42) – (47) or via (48).

 ISBN: 978-84-614-6167-7

Bauermeister and Shaw [18] have considered this problem extensively using (42) – (47) using Galerkin finite element methods in space together with Crank-Nicolson schemes in time, and produced *a priori* error estimates together with convincing demonstrations of the presence of the travelling wave front in their numerical results.

References

[1] BABUŠKA I, NOBILE F AND TEMPONE R *Reliability in computational science*.
Numerical Methods for Partial Differential Equations **23**, (2007) 753 – 784.
[2] BABUŠKA I, WHITEMAN J.R. AND STROUBOULIS T. Finite Elements: *An*
Introduction and a posterior Error Estimation. Oxford University Press, (2010)
[3] ODEN J.T. AND PRUDHOMME S. *Goal-oriented error estimation and adaptivity*
for the finite element method. Computers and Mathematics with Applications **41**,
(2001) 735 – 756.
[4] SHAW S, WARBY M.K. AND WHITEMAN J.R. *Discretization and Modelling*
error in the context of the rapid inflation of hyperelastic membranes. IMA Journal
of Numerical Analysis, (2010) 302-333.
[5] CIARLET P.G. *The Finite Element Method for Elliptic Problems*. North-Holland,
Amsterdam. (1978).
[6] ODEN J.T. AND REDDY J.N. *An Introduction to the Mathematical Theory of Finite Elements*. Wiley, New York. (1976)
[7] WHITEMAN J.R. (ed.) *The Mathematics of Finite elements and applications*. Academic Press, London, (1973).
[8] KARAMANOU M, WARBY M.K. AND WHITEMAN J.R. *Computational*
modelling of thermoforming processes in the case of finite viscoelastic materials.
Computer Methods in Applied Mechanics and Engineering **195,** (2006) 5220-5238
[9] SHAW S AND WHITEMAN J.R. *Numerical solution of linear quasistatic*
hereditary viscoelastic problems. SIAM Journal of Numerical Analysis 38, (2000)
80-97.
[10] SHAW S AND WHITEMAN J.R. *Towards robust adaptive finite element*
methods for partial differential Volterra equation problems arising in

 ISBN: 978-84-614-6167-7

viscoelasticity. In J.R. Whiteman (ed.) The Mathematics of Finite Elements and
Applications. Wiley, Chichester, 56 – 80, (1997).
[11] RIVIERE B, SHAW S, WHEELER MF. AND WHITEMAN J.R. *Discontinuous*
Galerkin finite element method for linear elasticity and quasistatic linear viscoelasticity. Numerische Mathematik **95,** (2003) 347-376.
[12] FERRY J.D. *Viscoelastic Properties of Polymers*. Wiley, New York, (1970)
[13] SHAW S., WARBY M.K. AND WHITEMAN J.R. *A comparison of hereditary*
integral and internal variable approaches to numerical solid viscoelasticity. In
Proc. XIII Polish Conf. on Computer Methods in Mechanics. Posnan, (1997).
[14] WARBY M.K., WHITEMAN J.R., JIANG W-G, WARWICK P, WRIGHT T.
Finite element simulation of thermoforming processes for polymer sheets, Mathematics and Computers in Simulation, **2105**, (2002) 1-10.
[15] SZEGDA D, SONG J, WARBY M.K. AND WHITEMAN J.R. *Computational*
Modelling of a Thermoforming Process for Thermoplastic Starch. American Institute of Physics, Proceedings **908**, (2007) 35 – 47.
[16] THOMAS N. and WINDLE A.H., *A theory of Case II diffusion.* Polymer (1982),
529 - 542.
[17] COHEN D.S., WHITE A.B. and WITELSKI T.P., *Shock formation in a multidimensional viscoelastic diffusive system.* SIAM J. Appl. Math. **55** (1995)
348 - 368.
[18] BAUERMEISTER N. and SHAW S., *Finite element approximation of non-Fickian polymer diffusion.* IMA J. Numer. Anal. **30** (2010) 702 – 730.

 ISBN: 978-84-614-6167-7

Proceedings of the 11th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE2011
Benidorm, Spain, 26-30 June 2011

Modeling Polymer Degradation and Erosion for Biodegradable Biomedical Implant Design

João S. Soares
Center for Mathematics and its Applications (CEMAT)
Department of Mathematics, Instituto Superior Técnico/UTL
Av. Rovisco Pais, 1049-001 Lisboa, Portugal

joao.soares@math.ist.utl.pt

Abstract

Biodegradable implants show great potential in many areas of medicine, and have already demonstrated success in simple applications such as sutures. For more complex devices, there are considerable challenges associated with the use of biodegradable materials. Here, we summarize previous efforts at implementing biodegradable technology to several areas of medicine, discuss the specific challenges involved, and present our recent modeling frameworks that can benefit this field.

1. Introduction

Over the last 50 years, biodegradable materials have found a wide variety of applications in the medical field ranging from biodegradable sutures, pins and screws for orthopaedic surgery, implants for local drug delivery, tissue engineering scaffolds, and biodegradable endovascular and urethral stents. The ability to predict the evolution of biodegradable polymers over the course of degradation would enhance the biodegradable implant design process and our main research objective is to endow biomedical implant designers with models that describe the response and evolution of the biodegradable polymers as they soften, degrade, and erode during the service life of the implant. Currently, the complex and challenging design process is largely based on a combination of intuition and guesswork accompanied with resource-expensive trial-and-error approaches that do not allow for systematic optimization and often fail [1].

 ISBN: 978-84-614-6167-7

2. Biodegradable Stents, Drug Delivery Implants, and Tissue Engineering Scaffolds

Biodegradable stents offer the potential to improve long-term patency rates by providing support just long enough for the artery to heal. The underlying hypothesis is that the need for mechanical support provided by employing a stent is temporary and limited to the intervention and shortly thereafter. While there are no conclusive data, it is generally thought that support should be provided for at least six months – beyond that, no utility or advantage for permanent stents has been demonstrated [2]. On the other hand, the advantages of biodegradable stents include: (i) if a stent degrades and is absorbed by the body, it will not be an obstacle for future treatments and will not be a permanent potential nidus for infection; (ii) the gradual softening of the material would not only permit a smooth transfer of the load from the stent to the healing artery, but also would contravene the permanent imposition of extremely high stresses on the stented artery; and finally (iii) a polymeric stent is a reservoir of an appreciable size for the incorporation of drugs, and polymer degradation and erosion may enhance drug release kinetics [3].

Drug delivery therapies have been enormously impacted by polymer-based systems [4]. The techniques employed today differ in concept but all share one particular feature, the mechanism of diffusion of certain species, the drug, through a matrix, usually polymeric in nature. In one approach, the drug is physically entrapped in an injectable or implantable capsule of solid polymer. Early forms of these systems involved non-degradable polymers in membrane-controlled diffusion such as silicone rubber, which released low molecular weight drug for extremely long times. Physically embedding drug in polymers at concentrations high enough to create a series of interconnecting pores through which the drug can afterward slowly diffuse from the matrix system was another early alternative. The next step in drug delivery implant technology was the employment of biodegradable polymers as carriers, where the combination of diffusion through pores as well as polymeric matrix degradation and erosion allowed better controllable release rates [5].

Tissue engineering scaffolds is another emerging medical where the application of biodegradable polymers has been touted for great success [6], and is currently hampered by the lack of rational tools to describe and predict the evolution of the response of the scaffolding material. Biodegradable scaffold properties influence all stages of tissue engineering and stringent requirements are placed at design stage. A tissue engineer needs to control degradation, erosion, and absorption of the polymeric matrix, not only in vitro during the cell seeding phase as cells proliferate and secrete extracellular matrix, but also in vivo as host tissue grows and remodels around and into the transplant.

 ISBN: 978-84-614-6167-7

3. Polymer Degradation and Erosion

Polymer degradation is the irreversible chain scission process that breaks down polymeric chains down to oligomers and finally monomers and results in a decrease of molecular weight. The prevailing mechanism of biological degradation of synthetic biodegradable aliphatic polyesters (the most common class of polymers employed in the medical field, such as polylactic acid and polyglycolic acid) is scission of the hydrolytically unstable backbone chain by passive hydrolysis, i.e. when in an aqueous environment, their ester linkages are cleaved with absorbed water. Extensive degradation leads to erosion, which is the process of dissolution and washing away of a polymeric matrix and results in material loss from the polymeric bulk. Lost material can either be monomers, oligomers, parts of the polymer backbone, or even parts of the polymer bulk [7-9]. Water uptake and hydrolytic scission compete in the process of erosion and modulate the behavior of the polymeric matrix: (i) if degradation is fast, the diffusing water is absorbed quickly by hydrolysis and hindered from penetrating deep into the polymer bulk, and consequently, degradation and erosion are confined to the surface of the matrix in a highly heterogeneous process referred to as surface erosion; on the other hand, (ii) if degradation is considerably slower than water uptake (as in common aliphatic polyesters), the polymer degrades and erodes through its entire swollen bulk in a process which is termed as homogeneous or bulk erosion [10].

4. Mechanical Modeling of Biodegradable Polymers

We have proposed a thermodynamically consistent constitutive model to describe the evolution of mechanical response of polymers which undergo strain-induced degradation. We introduce a degradation parameter that quantifies the local extent of scission and affects directly the material properties, lessening the ability of the material to store energy. In such way, material properties appearing in the constitutive specification of the material are material functions of degradation instead of simply being material constants, and naturally confer a sound physical interpretation to the internal variable of the material [11]. Furthermore, a kinetic equation governing the evolution of the degradation parameter results from the maximization of the rate of dissipation with the Clausius-Duhem inequality as a restriction for allowable processes [12]. We have studied the behavior of this class of materials either in simplified geometries [13], as well as with real stent geometries [14], and performed preliminary in vitro experiments to investigate the influence of applied loads on degradation [15].

 ISBN: 978-84-614-6167-7

5. Multiscale Modeling of Polymer Degradation and Erosion

We have introduced a general class of multiscale mixture models suitable to describe water-dependent degradation, both bulk and surface erosion modes in a unified sense, and in conjunction with drug release [16]. The multiscale model is borne within the framework of mixture theory by treating the polymer system as a mixture of a finite number of constituents describing the polydisperse polymeric system, each representing chains of an average size, and two additional constituents, water and drug. Hydrolytic scission of individual chains occurs at the molecular level and is described with a kinetic model, whereas constituents diffuse individually accordingly to Fick's 1st law at the bulk level – such analysis confers a multiscale aspect to the resulting reaction-diffusion system. A shift between two different types of behavior, each identified to surface of bulk erosion, is observed with the variation of a single non-dimensional parameter relating time scales of reaction (degradation) and of diffusion (water uptake and erosion). Mass loss follows a sigmoid decrease in bulk eroding polymers, whereas decreases linearly when an erosion front develops in surface eroding polymers. Drug release from biodegradable matrices is a process that is intrinsically coupled to degradation and erosion: drug release is mostly diffusion-controlled for slowly degrading polymers, whereas with unstable surface eroding matrices, drug release is dramatically enhanced in an erosion-controlled process.

6. Acknowledgements

The author acknowledges the support of the Fundação para a Ciência e Tecnologia (SFRH/BD/17060/2004 and SFRH/BPD/63119/2009) and the National Institute of Health (R01 EB000115). The author thanks MOX – Politecnico di Milano and CEMAT – Instituto Superior Técnico for their continued support. James E. Moore, Jr., Kumbakonam R. Rajagopal, and Paolo Zunino are essential collaborators and acknowledgements should also be extended to Adélia Sequeira and Melissa A. Grunlan.

7. References

[1] J.E. MOORE, J.S. SOARES, K.R. RAJAGOPAL, *Biodegradable stents: Biomechanical modeling challenges and opportunities*, Cardiovasc. Eng. Technol. **1** (2010) 52-65.

[2] A. COLOMBO, E. KARVOUNI, *Biodegradable stents : "Fulfilling the mission and stepping away"*, Circulation **102** (2000) 371-373.

[3] J.S. SOARES, *Bioabsorbable polymeric drug-eluting endovascular stents: A clinical review*, Minerva Biotecnol. **21** (2009) 217-230.

[4] R. LANGER, *Drug delivery and targeting*, Nature **392** (1998) 5-10.

 ISBN: 978-84-614-6167-7

[5] J.S. SOARES, *Diffusion of a fluid through a spherical elastic solid undergoing large deformations*, Int. J. Eng. Sci. **47** (2009) 50-63.
[6] S. LEVENBERG, R. LANGER, *Advances in tissue engineering*, in: G.P. SCHATTEN, (Ed), *Current topics in developmental biology, vol. 61*, Elsevier Academic Press Inc, San Diego, 2004, 113-134.
[7] W. SCHNABEL, *Polymer degradation*, Macmillan Publishing, New York, 1981.
[8] W.L. HAWKINS, *Polymer degradation*, 1st, Springer-Verlag, Berlin, 1984.
[9] A. GOPFERICH, *Mechanisms of polymer degradation and elimination*, in: A.J. DOMB, J. KOST, D.M. WISEMAN, (Eds), *Handbook of biodegradable polymers*, Harwood Academic Publishers, Australia, 1997, 451-471.
[10] F.V. BURKERSRODA, L. SCHEDL, A. GOPFERICH, *Why degradable polymers undergo surface erosion or bulk erosion*, Biomater. **23** (2002) 4221-4231.
[11] J.S. SOARES, J.E. MOORE, K.R. RAJAGOPAL, *Constitutive framework for biodegradable polymers with applications to biodegradable stents*, ASAIO J. **54** (2008) 295-301.
[12] K.R. RAJAGOPAL, A.R. SRINIVASA, A.S. WINEMAN, *On the shear and bending of a degrading polymer beam*, Int. J. Plast. **23** (2007) 1618-1636.
[13] J.S. SOARES, K.R. RAJAGOPAL, J.E. MOORE, JR., *Deformation-induced hydrolysis of a degradable polymeric cylindrical annulus*, Biomech. Model. Mechanobiol. **9** (2010) 177-186.
[14] J.S. SOARES, J.E. MOORE, K.R. RAJAGOPAL, *Modeling of deformation-accelerated breakdown of polylactic acid biodegradable stents*, J. Med. Dev. **4** (2010).
[15] J.S. SOARES, *Constitutive modeling of biodegradable polymers for application in endovascular stents*, PhD Dissertation, Texas A&M University, College Station, TX, 2008,
[16] J.S. SOARES, P. ZUNINO, *A mixture model for water uptake, degradation, erosion, and drug release from polydisperse polymeric networks*, Biomater. **31** (2010) 3032-3042.

 ISBN: 978-84-614-6167-7

Proceedings of the 11th International Conference on Computational and Mathematical Methods in Science and Engineering, CMMSE2011 Benidorm, Spain, 26-30 June 2011

New silicon materials built from the assembly of $Ti@Si_{16}$ and $Sc@Si_{16}K$ super-atom units.

M. B. Torres[1] and L. C. Balbás[2]

[1] *Departamento de Matemáticas y Computación, Universidad de Burgos, 09006, Burgos*

[2] *Departamento de Física Teórica, Atómica y Óptica, Universidad de Valladolid, E=47011, Valladolid.*

emails: begonia@ubu.es, balbas@fta.uva.es

Abstract

We study, from first-principles calculations, the structural and electronic properties of several low lying energy equilibrium structures of isoelectronic Si_nM clusters (M = Sc−, Ti) for n=14-18. The main result is that those clusters with n = 16 are more stable than their neighbours, in agreement with recent time of flight mass spectra and photoelectron spectroscopy experiments. As a second step we consider the nearly spherical endohedral $M@Si_{16}$ cage-like clusters (M = Sc−, Ti) with D_{4d} symmetry as the basis unit to study the formation of $[Ti@Si_{16}]_n$, and $[Sc@Si_{16}K]_n$ aggregates and how their properties evolves with increasing size *(n* ≤ 9*)*. We identify especially stable linear, planar, and three-dimensional patterns, which can serve as seeds to grow low-dimensional infinite systems. Calculations of BCC, FCC, and Single Cubic crystal meta-stable phases, having the $Ti@Si_{16}$ superatom as basic unit, as well as cubic NaCl and CsCl structures of bulk $Sc@Si_{16}K$ are performed. Moreover, the structure and cohesive energy of a few periodic linear chains and wires have been optimized. $[Ti@Si_{16}]_{2n}$ ($[Ti@Si_{16}]_{3n}$) nanowires formed by stacking $Ti@Si_{16}$ dimers (triangular trimers) along the vertical axis, and rotated 90º (60º) each other have been characterized. Finally, preliminary results for the H_2 adsorption on $[Ti@Si_{16}]_{2n}$ and $[Ti@Si_{16}]_{3n}$ finite wires are presented.

Key words: superatoms; cluster-assembled materials;nanoparticles;clusters; nanocrystals

 ISBN: 978-84-614-6167-7

1. Introduction

Interest in the study of aggregates built from small atomic clusters is motivated by their potential use as building blocks for new functional materials and devices at the nanoscale [1]. To achieve this goal, it is important to investigate how the system geometry depends on the interparticle coupling and how it affects the physical properties of the systems. Chemically stable building blocks should have a closed electron configuration with a large energy gap between the highest occupied molecular orbital (HOMO) and the lowest unoccupied molecular orbital (LUMO). The other important factor determining the cluster stability is the atomic geometry. Thus, the cooperative effects between electronic and geometrical factors can provide a guiding principle for designing stable building-block clusters.

The construction of new optoelectronic materials, via the assembly of molecules of the type $Sc@Si_{16}Y$ where Y is an alkali atom, was suggested by Nakajima and co-workers [2], assuming that the large HOMO-LUMO gap of the ionic superatom can be maintained in the range of ~2 eV when the supermolecule is formed, and a solid phase can eventually be grown from them. On the other hand, Reis et al. [3] have shown the possible survival under room conditions of a metastable crystal with hcp structure formed from the Frank-Kasper isomer (T_d symmetry) of $M@Si_{16}$ (M = Ti, Zr, or Hf). Gueorguiev et al. [4] have studied finite $[M@Si_{12}]_n$ nanowires with a hexagonal cross section for $n \leq 7$ and M = Fe, Ni, Co, Ti, V, and Cu, finding that the HOMO-LUMO gap decreases gradually toward metallic behavior. Pentagonal and hexagonal core-shell silicon nanowires with various core compositions, including 3d transition metal atoms, have been investigated by Berkdemir and Gülseren [5] who also revealed a metallic behavior in all the cases, being the most favorable cross section hexagonal or pentagonal depending on the encapsulated atom. Leitsmann et al. [6] have investigated the influence of substitutional and interstitial transition metal doping on the electronic and magnetic properties of silicon nanocrystals.

Kumar and co-workers [7] have shown that adsorption of atomic hydrogen on cage-like doped silicon clusters lead to additional stabilization of these cage-like structures. However, to the best of our knowledge, the adsorption of molecular H_2 on these silicon systems, which could be interesting for hydrogen storage porpoises, has not been considered previously.

In a recent work [8] we have reported the geometrical and electronic structure of several low lying energy isomers of $M@Si_n$ clusters (M = Sc−, Ti, V^+) in the range n = 14–18. We have obtained good agreement with the experimental results obtained by Nakajima and coworkers [9] for the endohedral character, extra stability, HOMO- LUMO gap, and electron affinity of $M@Si_{16}$ clusters. As a first

 ISBN: 978-84-614-6167-7

step to explore bulk phases of materials composed of neutral $M@Si_{16}X$ entities, we have also studied [10] the equilibrium structures and electronic properties of aggregates, finite and infinite (periodic) wires of $[Ti@Si_{16}]_n$ and $[Sc@Si_{16}K]_n$ systems formed by assembling $M@Si_{16}$ units with D_{4d} symmetry. In this work, we will review same of those results, report new calculations of some bulk phases, and study the adsorption of H_2 in some of these aggregates and wires.

In section 2 is described our computational approach. In section 3 we review the results for $Ti@Si_{16}$ and $Sc@Si_{16}K$, and in section 4 the results for aggregates $[Ti@Si_{16}]_n$ and $[Sc@Si_{16}K]_n$. In section 5 are shown some calculated meta-stable structures of periodic crystals and nanowires formed by these silicon doped aggregates as cell units. Preliminary results for the adsorption of H_2 are presented in section 6. Finally, in section 7 we summarize the results.

2. Computational procedure

We have used the density functional theory (DFT) code SIESTA [11] within the generalized gradient approximation as parameterized by Perdew et al. [9] for exchange-correlation effects. Specifically, we have used norm conserving scalar relativistic pseudopotentials [12] in their fully nonlocal form, generated from the atomic valence configuration $3s^2 3p^2$ for Si (with core radii 1.9 a. u. for *s* and *p* orbitals), and the semi-core valence configuration $4s^2 3p^6 3d^n$ for Sc (n = 1*)* and Ti (n = 2). For K is used the valence configuration $4s^1$ with core radius 3.64 a.u. for all *s, p,* and *d* valence orbitals. For the present calculations is taken a double-ζ basis *s, p* (for Si, K,) and *s, p, d* (for M), with single polarization *d* (for Si, K) and *p* (for M). The grid fineness is controlled by the energy cut off of the plane waves that can be represented in it without aliasing (120 Ry for aggregates and 150 Ry for wires and bulk in this work). The equilibrium geometries result from unconstrained conjugate-gradient structural relaxation using the DFT forces. We try out several initial structures for each cluster (typically more than 20) until the force on each atom was smaller than 0.010 eV/Å. The integration in *k*-space for the bulk calculations was performed using a 4 × 4 × 4 Monkhorst pack grid. For wires in the *z* direction we use a 1×1×12 Monkhorst pack grid, and similarly for wires grown in the *x* or *y* directions.

3. $Ti@Si_{16}$ and $Sc@Si_{16}K$ superatoms.

We have determined [8] from first-principle calculations the geometrical and electronic structure of several low-lying energy isomers of $M@Si_n$ clusters (M) Sc-, Ti) with n = 14-18, obtaining a good agreement with the experimental results of Nakajima and co-workers [2, 9]. In Figure 1 is shown their structure of a few low lying energy isomers. The ground state geometry of these 68 valence electron clusters is a distorted Frank- Kasper C_{3v} structure. However, for the ground state

 ISBN: 978-84-614-6167-7

of $[M@Si_{16}]_n$ aggregates, the favoured geometry contains the isomers with D_{4d} symmetry, which was called fullerene-like (f-like) by Kumar and co-workers [14]. This fact has important consequences for our study.

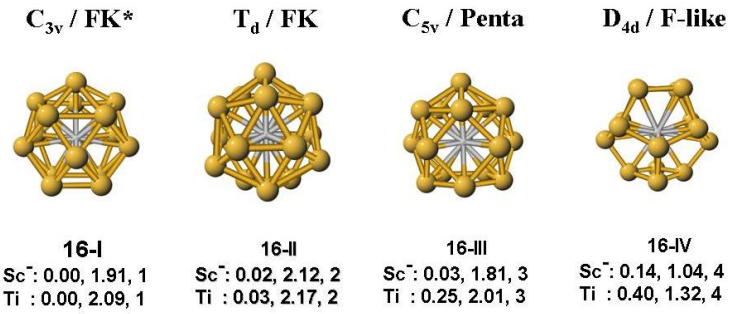


Figure 1. Structure and symmetry of several low-lying energy isomers of $M@Si_{16}$ clusters: FK* = distorted Frank-Kasper; FK = Frank- Kasper; Penta = pentagonal; F-like = fullerene-like. The total energy difference, in eV, with respect to the lowest-energy configuration, as well as the HOMO-LUMO gap (eV), and ordinal number of the isomer are shown.

In reference [8] we provided an interpretation of the electronic structure and orbital projected density of states (PDOS) of $M@Si_{16}$ clusters in the context of the spherical shell model perturbed by the crystalline field of the underlying ionic geometry. The covalent bonding in $M@Si_{16}$ clusters results from the hybridization of the Si empty-cage states and the valence states of the endohedral atom having equal angular momentum *l*, which is of *d* type for the HOMO of the low lying energy isomers. This picture has been confirmed by the recent angle-dependent photoelectron spectroscopy experiments of Lau and co-workers [15].

4. $[Ti@Si_{16}]_n$ and $[Sc@Si_{16}K]_n$ aggregates.

In this section is presented a selection of $[Ti@Si_{16}]_n$ and $[Sc@Si_{16}K]_n$ isomers resulting from the optimization of a large set of initial arrangements. Firstly, we have investigated systematically the $n = 2$ configurations to select those that can serve as seeds in the construction of larger aggregates following well-defined patterns. As a second step, we studied n = 3-9 aggregates constructed according to these patterns, particularly certain types of configurations that can be used to grow infinite wires and nanotubes. The binding energy of a $[M@Si_{16}X]_n$ aggregate is defined as the difference in the aggregate energy between it and that of n identical $M@Si_{16}X$ separated units: $E_b(n) = n\ E(M@Si_{16}X) - E([M@Si_{16}X]n)$. Several $[Ti@Si_{16}]_n$ configurations are shown in Figure 2, together with the corresponding binding energies, HOMO-LUMO gaps, electric dipole moments, and the types of bonding between the two D_{4d} cages.

Although the ground state of $Ti@Si_{16}$ ($Sc@Si_{16}K$) has the Frank-Kasper *Td* structure (C_{5v} structure with the K atom on the pentagonal face), the $[Ti@Si_{16}]_2$ ($[Sc@Si_{16}K]_2$) dimer is preferably formed with two f-like D_{4d} units. That D_{4d} cage

 ISBN: 978-84-614-6167-7

is composed of two square faces and eight pentagons and contains only two types of non-equivalent Si atoms: those that are in the vertices of the square faces (Si_1) and those that are not (Si_2). The most favourable configurations for both dimer results from two Si_1-Si_1 bonds and three Si_2-Si_2 bonds between the Si atoms of opposite pentagonal faces of two D_{4d} M@Si_{16} cages. The K atom is not involved directly in the bonding. That dimer configuration, called here PTP-2 arc 2 (PTP = pentagon to pentagon), is used to grow planar aggregates having arc geometry or ring geometry. Nearly degenerate $[Sc@Si_{16}K]_n$ ring isomers result differing in the K bonding site, which lead to very different dipole moments. By stacking these ring units along the vertical axis could be grown meta-stable nanotubes with response to an external electrical field varying over a broad range of values.

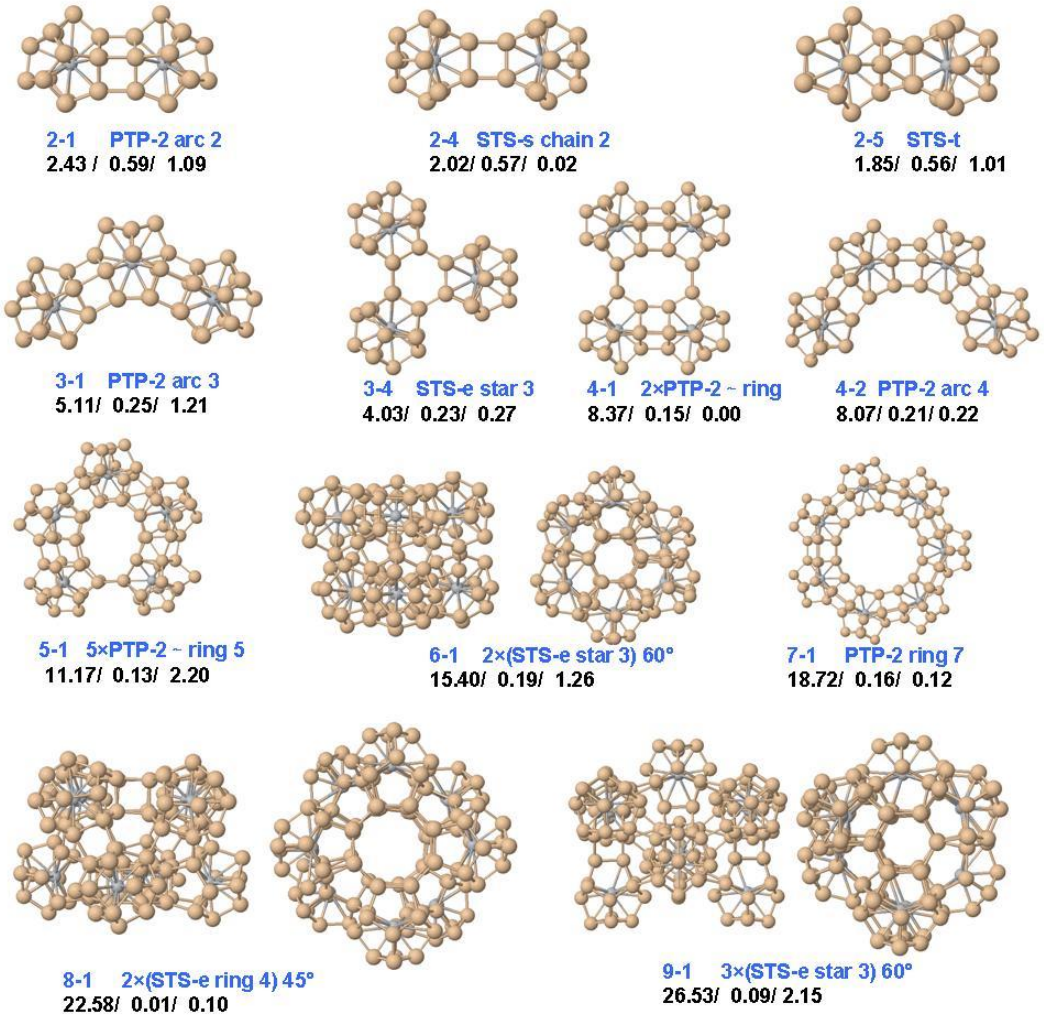


Figure 2. Several low-lying energy isomers of $[Ti@Si_{16}]_n$ aggregates formed from $Ti@Si_{16}$ units with D_{4d} symmetry. Binding Energies (E_b), HOMO-LUMO Gaps, and Electric Dipole Moments are shown.

 ISBN: 978-84-614-6167-7

For n =3, the arc 3 trimer configuration arising from PTP-2 bonding is more favourable than the star 3 one provided by the STS-e type of bonding, with the M atoms forming an equilateral triangle. However, the stacking of these regular star 3 trimer units along the vertical axis, with a rotation of 60° between consecutive units, yields very stable $[M@Si_{16}X]_{3m}$ aggregates, particularly those doped with Ti. For n = 4, the bonding of two PTP-2 arc 2 dimers by four STS-e Si1-Si1 bonds leads to a specular double arc 2 form that is the first isomer for $[Ti@Si_{16}]_4$ and the second isomer for $[Sc@Si_{16}K]_4$ aggregates. These ~ring and arc 4 configurations have comparable binding energies for n = 4.

The first isomer quoted for n = 5 aggregates is an irregular PTP-2 ring 5 for $[Ti@Si_{16}]_5$ and a PTP-2 arc 5 configuration for $[Sc@Si_{16}K]_5$. The more favourable isomer of n = 6 aggregates with Ti dopant is a 3D structure formed by the stacking of two star 3 structures with a rotation of 60° between both units. For $[Ti@Si_{16}]_7$, the arc structure converges to the ring one. Several planar ring and arc structures are reported for $[Sc@Si_{16}K]_7$. Nearly degenerate planar ring isomers of $[Sc@Si_{16}K]_n$ (n = 6-9) aggregates show dramatic change in their dipolar moments when the K atoms undergo slight changes in their bonding sites. The vertical stacking of two star 4 structures rotated 45° is the first isomer of $[Ti@Si_{16}]_8$. $[Ti@Si_{16}]_9$ structure, formed by the vertical stacking of three star 3 isomers, rotated 60° consecutively, having high binding energy, points to interesting nanowires.

5. Crystals and wires from Silicon-Doped Aggregates as Cell Units

We have calculated several crystal phases having the $Ti@Si_{16}$ superatom as basic unit. For the BCC case, we obtained that the FK isomer, with Td symmetry, reaches a bulk meta stable minimum with 0.96 eV cohesive energy for 9.75 Å lattice constant, whereas the f-like D_{4d} isomer leads to a deeper minimum with ~5.72 eV cohesive energy for ~8.75 Å lattice constant. The orientation of the cluster in the cell has a controllable effect. For the Single Cubic (SC) and FCC phases, the cohesive energy is smaller than for the BCC one. Similar calculations have been performed for SC, BCC, FCC, NaCl, and CsCl crystals taking the $Sc@Si_{16}K$ supermolecule as the basic unit cell. Both the FK and f-like isomers for the $Sc@Si_{16}$ component of the supermolecule were considered. In all these cases, we obtained local minima, being the NaCl meta-stable bulk phase with the f-like isomer of $Sc@Si_{16}$ the one with the largest cohesive energy (5.58 eV).

 ISBN: 978-84-614-6167-7

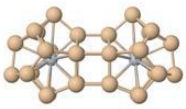
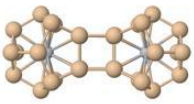
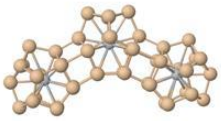
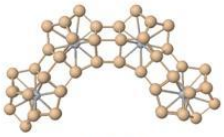
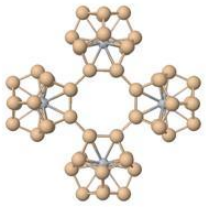
Isomer in single cubic cell		Bond type and wire direction	Cohesive energy (eV)	Lattice constant (Å)
Ti-2.1		PTP-2 arc2 wire along *z*	1.81	8.00
Ti-2.2		STS-s chain2 along *x*	2.50	15.00
Ti-3.1		PTP-2 arc3 wire along *z*	2.70	8.00
Ti-3.3		STS-e star3 wire along *z*	4.28	7.75
Ti-4.2		PTP-2 arc4 wire along *z*	3.42	8.00
Ti-4.4		STS-e ring4 wire along *z*	5.70	7.75

Figure 3. Cohesive energy (*E*coh) per unit cell and lattice constant (*l*) for infinite wires formed from single cubic unit cells composed of [the Ti@Si16]*n* aggregates showed Figure 2, which are specified in the first column. The grown direction, given in the second column as *x* , *y*, or *z*, correspond to left–right, down–up, and perpendicular to the plane of the figure, respectively.

Figure 3 shows the cohesive energy per unit cell and the lattice constant of infinite wires composed of some of the $[Ti@Si_{16}]_n$ aggregates. The $n > 4$ aggregates with ring structure which can lead to nanotubes. To compare the $[Ti@Si_{16}]_n$ wires formed from PTP- 2 arc-*n* isomers with those formed from STS-e ring-*n* isomers we take the cohesive energy per cell unit and D_{4d} unit. Thus, we have obtained 0.91 eV, 0.90 eV, and 0.86 eV for arc-*n* wires with $n = 2$, 3, and 4, respectively, whereas for STS-e we have obtained 1.43 eV and 1.42 eV for star3 and ring4 wires, respectively. These results indicate that (i) although arc3 and arc4 aggregates are more stable than star3 and ring4 ones, respectively, STS-e wires

 ISBN: 978-84-614-6167-7

(star3 and ring4) have more cohesive energy per D_{4d} superatom than the PTP-2 arc-*n* wires, and (ii) the binding energy per superatom decreases with the size of the aggregate in the unit cell. From the Ti-4.4, rig4 aggregate can be grown a planar sheet resulting a very stable structure with 1.75 eV per D_{4d} superatom unit, which is larger than for the Ti-4.4 ring4 wire.

Moreover, $[Ti@Si_{16}]_{2n}$ and $[Ti@Si_{16}]_{3n}$ periodic nanowires formed by stacking STS-e dimers and Ti-3.3 trimers along the vertical axis and rotated 90 and 60 each other have recently been obtained. The geometry of the corresponding cell units is shown in Figure 4.

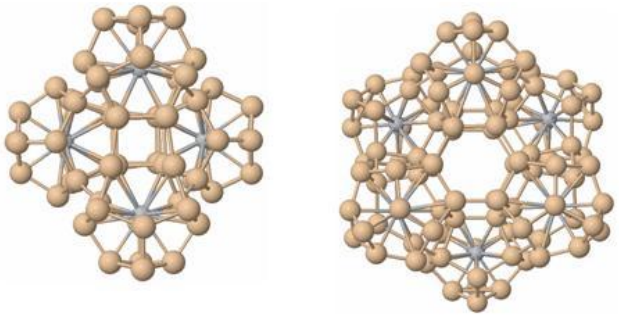


Figure 4. Geometries of the cell units of $[Ti@Si_{16}]_{2n}$ and $[Ti@Si_{16}]_{3n}$ periodic nanowires. Two views are given: top and upper.

6. H_2 adsorption by $[Ti@Si_{16}]_n$ aggregates and wires.

We have studied the adsorption de H_2 on the aggregates $(Ti@Si_{16})_n$ (n = 4, 6) formed by stacking two STE-e dimer and two Ti-3.3 trimer which are twisted 90° and 60°, respectively. Moreover, we will present preliminary results on the adsorption of H_2 on $[Ti@Si_{16}]_{2n}$ and $[Ti@Si_{16}]_{3n}$ nanowires.
For the $[Ti@Si_{16}]_2$ unit the H_2 molecule binds preferably on top of the square face between the two Ti@Si16 units, with binding energy 0.12 eV. The question is: how many H_2 molecules can be adsorbed per $[Ti@Si_{16}]_2$ unit ?

7. Summary and Outlook

In conclusion, we have optimized several $[Ti@Si_{16}]_n$ and $[Sc@Si_{16}K]_n$, structures for $n \leq 9$ with D_{4d} symmetry for the $M@Si_{16}$ unit. Particularly interesting aggregates, to explore wires obtained from them, are: (i) planar rings for $n \geq 6$, which can be grown into nanotubes; (ii) 1D aggregates formed by stacking dimers with a mutually orthogonal M-M axis; (iii) aggregates formed by stacking the star3 trimer, rotated 60° with respect to each other; and (iv) aggregates formed by

 ISBN: 978-84-614-6167-7

stacking star4 $n = 4$ aggregates along the vertical axis that are rotated 45° with respect to each other. In all these cases are found nearly degenerate isomers whose electric moments depend dramatically on M dopant and K bonding sites.

The BCC crystal from f-like $Ti@Si_{16}$ superatom has larger cohesive energy than other crystals with the FK $Ti@Si_{16}$ superatom in the unit cell. For the $Sc@Si_{16}K$ superatom, the NaCl bulk phase with the f-like isomer is the one with the largest cohesive energy. A planar sheet from that star4 isomer of $Ti@Si_{16}$ results particularly stable. Moreover, we have characterized interesting $[Ti@Si_{16}]_{2n}$ and $[Ti@Si_{16}]_{3n}$ nanowires formed by stacked along the vertical axis and twisted 90 and 60 degrees, respectively.

The study of adsorption of H_2 in some of this aggregates and nanowires is in progress, particularly the number of H_2 molecules (saturation) which can be adsorbed in a unit cell with a reasonable binding energy (larger than 0.15 eV per H_2)

References

[1] S. A. Claridge et al., ACS Nano **3** (2009) 244.
[2] K. Koyasu et al., J. Phys Chem A **111** (2007) 42.
[3] Reis et al., PRB 76, 233406 (2007) ; J Phys. Cond. Matter **22** (2010) 035501.
[4] Gueorguiev et al., Chem. Phys. Lett. **458** (2008) 170.
[5] C. Berkdemir et al., PRB **80** (2009)115334.
[6] Leitsmann et al., PRB **80** (2009) 104412.
[7] V. Kumar et al., PRB **75** (2007) 155425.
[8] M. B. Torres et al., PRB **75** (2007) 205425-1-12.
[9] Nakajima et al., JACS **127** (2005) 4998; J. Chem. Phys. **129** (2008) 214301.
[10] M. B. Torres et al., Int. J. Quantum Phys 111, 444 (2011); J. Phys. Chem. C. 115, 335 (2011)
[11] Soler et al., J. Phys. Cond. Matter **14**, (2002) 2745.
[12] J. P. Perdew eta al., PRB **77** (1996) 3865.
[13] N. Troullier and J. Martins, PRB **43** (1991) 1993.
[14] Kumar et al., PRL **87** (2001) 045503; Chem. Phys. Lett. **363**, (2002) 319.
[15] T. Lau et al., PRB **79** (2009) 053201; JCP **134** (2011) 041102.

 ISBN: 978-84-614-6167-7

Proceedings of the 11th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE2011
Benidorm, Spain, 26-30 June 2011

Finite-difference schemes for a two-dimensional problem of femtosecond pulse interaction with semiconductor.

Vyacheslav A. Trofimov[1], Maria M. Loginova[1]

[1] *Faculty of Computational Mathematics and Cybernetics,*
Lomonosov Moscow State University
emails: vatro@cs.msu.ru, mloginova@cs.msu.ru

Abstract

In this report we proposed two finite-difference schemes for 2D problem of femtosecond laser pulse interaction with semiconductor. The first finite-difference scheme is a conservative one, which is based on the integral-interpolation method at its construction. Another one is additive finite-difference scheme, constructed on the base of split-step method. We compare results of computer simulation which are obtained by using both finite-difference schemes. We demonstrate that the split-step method is ineffective for computation of the laser pulse interaction with semiconductor at the strong changing of solution during small time interval.

Key words: femtosecond pulse, semiconductor, laser pulse, finite-difference schemes, conservative scheme.
MSC2000:

1. Introduction

The interaction of laser pulse with semiconductor is a problem of interest during many years. In the past few years this interest is firstly caused by the possibility to investigate the motion of free-charged particles in semiconductor due to action of the THz radiation [1-6]. Secondly, this interest has appeared due to an opportunity of creation of the all-optical bistable device with femtosecond time of switching at action of femtosecond laser pulse on semiconductor. In both cases the nonlinear response of semiconductor take place. To provide a computer simulation of such kind of problems it is necessary to develop a finite-difference scheme which possesses very good properties: high accuracy of computation, stability and asymptotic stability and conservatism.

At the present time there are two approaches for developing the finite-difference schemes: conservative finite-difference scheme and ones on the base of split-step method. Hence, it is very important to compare the efficiency of application of both approaches to construction of the fintite-difference scheme for computer simulation of the laser pulse interaction with semiconductor in the 2D problem. It should be noted that we have

 ISBN: 978-84-614-6167-7

proposed early the conservative finite-difference schemes for the 1D problem. Previous results have demonstrated many advantages of developed finite-difference schemes.

2. Problem Statement

As it is known, an interaction of high-intensive femtosecond laser pulse or the THz pulse with semiconductor is accompanied by various nonlinear effects [7-9]. For example, there are modes of oscillating the semiconductor characteristics and phenomenon of optical bistability. This process is described by the following set of two-dimensional differential equations:

$$\begin{aligned}
&\frac{\partial^2\varphi}{\partial x^2}+\frac{\partial^2\varphi}{\partial y^2}=\gamma(n-N),\ \ \frac{\partial I}{\partial x}+\delta_0 I\delta(n,N,\varphi)=0,\\
&\frac{\partial n}{\partial t}=D_x\frac{\partial}{\partial x}(\frac{\partial n}{\partial x}-\mu_x n\frac{\partial\varphi}{\partial x})+D_y\frac{\partial}{\partial y}(\frac{\partial n}{\partial y}-\mu_y n\frac{\partial\varphi}{\partial y})+G(n,N,\varphi)-R(n,N),\\
&\frac{\partial N}{\partial t}=G(n,N,\varphi)-R(n,N),\ R=\frac{nN-n_0^2}{\tau_p},\ G=q_0 I\delta(n,N,\varphi),\ \ q_0>0,\\
&0<x<L_x,\ \ 0<y<L_y,\ \ t>0.
\end{aligned} \tag{1}$$

$$\begin{aligned}
&\left.\frac{\partial\varphi}{\partial x}\right|_{x=0,L_x}=\left.\frac{\partial\varphi}{\partial y}\right|_{y=0,L_y}=0,\ \left.\frac{\partial n}{\partial x}\right|_{x=0,L_x}=\left.\frac{\partial n}{\partial y}\right|_{y=0,L_y}=0,\ I\big|_{x=0}=I_0(y)I_1(t),\\
&n\big|_{t=0}=N\big|_{t=0}=n_0,\ I\big|_{t=0}=0.
\end{aligned} \tag{2}$$

In the set of equations (1) – (2) the following variables are introduced: *x, y*-dimensionless spatial coordinates, *t* - time, *n* - concentration of free electrons in conductivity zone of semiconductor, *N* – concentration of ionized donors. Function φ- dimensionless electric field potential, *I* – intensity of laser radiation propagating along the *x* axis. Coefficients of electrons diffusion D_x, D_y and coefficients of electrons mobility μ_x, μ_y are non-negative constants. Parameter γ depends, in particular, on the maximal achieving concentration of free charge carriers, n_0 - equilibrium value of free electrons concentration and ionized donors one, τ_p characterizes a recombination time. Functions *G* and *R*, describe generation and recombination of semiconductor free charges particles correspondingly. Light energy absorption coefficient $\delta(n,N,\varphi)$ can be approximated by different ways. For example, in present work we consider the following approximations

$$\begin{aligned}
&\delta_\varphi(n,N,\varphi)=(1-N)e^{(\beta|\varphi|)},\ \ \beta>0,\\
&\delta_n(n,N,\varphi)=(1-N)e^{-\psi(1-\xi n)},\ \ \psi>0,\ \ \xi>0.
\end{aligned}$$

 ISBN: 978-84-614-6167-7

The boundary and initial conditions (2) correspond to absence of an external electric field and current through the semiconductor surface. It is supposed, that at the initial moment of time the semiconductor is electrically neutral.
It is necessary to notice that for construction and for testing the finite - difference schemes for a problem (1) - (2) we take $\delta_0 = 0$. So, we didn't consider change of laser pulse intensity along axis *x*. In this case the function of free electrons generation takes in the following way:

$$G = q_0 q(x,y,t)\delta(n,N,\varphi),\ \ q(x,y,t) = e^{-\left(\frac{x-0.5L_x}{0.1L_x}\right)} e^{-\left(\frac{y-0.5L_y}{0.1L_y}\right)} \left(1 - e^{-10t}\right).$$

The function *q(x,y,t)* describes the optical radiation intensity profile with maximal value q_0. Thus, we consider the generation of free charge carriers by the Gaussian laser beam in both spatial directions. The centre of beam coincides with centre of spatial domain. Such approach allows estimating the accuracy of computer simulation results obtained on the base of various finite- difference schemes. Because the problem (1) has the symmetric solution on both spatial coordinates then the solution of the finite-difference scheme must be also symmetric for $\delta_0 = 0$ In this case we can also estimate the efficiency of computer simulation results for 2D case by comparing them with the corresponding results for 1D case (approximation of optical thin layer) problems [10], [11].

3. Construction of finite-difference schemes

To construct the finite-difference scheme we introduce in the domain $\overline{G} = \{0 \le x \le L_x\} \times \{0 \le y \le L_y\} \times \{0 \le t \le L_t\}$
the uniform time and space grids $\Omega = \omega_x \times \omega_y \times \omega_t$, where

$$\omega_x = \left\{ x_i = ih_x,\ \ i = \overline{0, N_x},\ \ h = {L_x}/{N_x} \right\},$$

$$\omega_y = \left\{ y_j = jh_y,\ \ j = \overline{0, N_y},\ \ h_y = {L_y}/{N_y} \right\},$$

$$\omega_t = \left\{ t_k = k\tau,\ \ k = \overline{0, N_t},\ \ \tau = {L_t}/{N_t} \right\}.$$

We use the following notations for any grid function f_h defined on Ω:

$$f = f_h(x_i, y_j, t_k),\ \ f_{x\pm1} = f_h(x_{i\pm1}, y_j, t_k),\ \ f_{y\pm1} = f_h(x_i, y_{j\pm1}, t_k),$$

$$f_{x\pm0.5} = 0.5(f_h(x_i, y_j, t_k) + f_h(x_{i\pm1}, y_j, t_k)),$$

$$f_{y\pm0.5} = 0.5(f_h(x_i, y_j, t_k) + f_h(x_i, y_{j\pm1}, t_k)),$$

$$\hat{f} = f_h(x_i, y_j, t_{k+1}),\ \ \tilde{f} = f_h(x_i, y_j, t_{k+0.5}),\ \ \overset{0.5}{f} = 0.5(f + \hat{f}).$$

 ISBN: 978-84-614-6167-7

Below we omit the subscript *h* on mesh functions. Nevertheless, we preserve the above-introduced notations. The first and the second differential derivatives are defined as usual and notated as follows: $f_x,\ \ f_{\bar{x}},\ \ f_{\bar{x}x},\ \ f_y,\ \ f_{\bar{y}},\ \ f_{\bar{y}y}, f_t.$

At the finite-difference schemes construction we should achieve, that the difference analogs of conservation laws for the problem (1-2) are valid [12]. As it is well known, for system of equations (1) – (2) the law of charge preservation takes place. Thus for our problem the following conservation law (invariant) takes place [13]:

$$Q(t) = \int_0^{L_y}\int_0^{L_x} \big(n(x,y,t) - N(x,y,t)\big)dxdy = 0. \tag{3}$$

So, for the problem (1) - (2) a property of finite-difference scheme conservatism consists in the preservation of invariant *Q(t)*.

Scheme 1.

On the base of integro-interpolation method [12] we have constructed conservative finite - difference scheme for problem (1) - (2) (see as well [14]). To solve the obtained set of nonlinear difference equations we used iterative process, which is constructed by taking into account of increasing the computation accuracy and validity of the conservatism property at iterations:

$$\hat{\varphi}^{s+1}_{\bar{x}x} = \gamma(\hat{n}^{s+1} - \hat{N}^{s+1}),\ 1 < i_x < N_x - 1,\ 1 < i_y < N_y - 1,\ \ k > 0,$$

$$\frac{\hat{n}^{s+1} - n}{\tau} = 0.5D_x\left[\left(\hat{n}^{s+1}_{\bar{x}x} + n_{\bar{x}x}\right) - \mu_x\left((\hat{n}^{s}_{x_{+0.5}}\hat{\varphi}^{s}_{x})_{\bar{x}} + (n_{x_{+0.5}}\varphi_x)_{\bar{x}}\right)\right] +$$

$$+ 0.5D_y\left[\left(\hat{n}^{s+1}_{\bar{y}y} + n_{\bar{y}y}\right) - \mu_y\left((\hat{n}^{s}_{y_{+0.5}}\hat{\varphi}^{s}_{y})_{\bar{y}} + (n_{y_{+0.5}}\varphi_y)_{\bar{y}}\right)\right] + \overset{s,s+1}{G}_{0.5} - \overset{s,s+1}{R}_{0.5}, \tag{4}$$

$$\frac{\hat{N}^{s+1} - N}{\tau} = \overset{s,s+1}{G}_{0.5} - \overset{s,s+1}{R}_{0.5}, \quad \overset{s,s+1}{G}_{0.5} = 0.5q_0 q\left(\delta(\hat{n}^{s}, \hat{N}^{s+1}, \hat{\varphi}^{s}) + \delta(n, N, \varphi)\right),$$

$$\overset{s,s+1}{R}_{0.5} = 0.5\left(\frac{\hat{n}^{s}\hat{N}^{s+1} - n_0^2}{\tau_p} + \frac{nN - n_0^2}{\tau_p}\right).$$

The finite-difference scheme (4) approximates the initial set of differential equations with the accuracy $O(h_x^2 + h_y^2 + \tau^2)$. The boundary conditions are approximated with the first order in spatial coordinates. It is caused by necessity of the realization of conservatism property.

 ISBN: 978-84-614-6167-7

$$\hat{n}^{s+1}_{x,0} = \hat{n}^{s+1}_{\bar{x},N_x} = 0, \quad \hat{n}^{s+1}_{y,0} = \hat{n}^{s+1}_{\bar{y},N_x} = 0, \quad \hat{\varphi}^{s+1}_{x,0} = \hat{\varphi}^{s+1}_{\bar{x},N_x} = 0, \quad \hat{\varphi}^{s+1}_{y,0} = \hat{\varphi}^{s+1}_{\bar{y},N_x} = 0,$$
$$n\big|_{k=0} = N\big|_{k=0} = n_0, \quad \varphi\big|_{k=0} = \varphi_x\big|_{k=0} = \varphi_y\big|_{k=0} = 0. \tag{5}$$

As the initial approximation for the iterative process, we take the values of functions obtained on the previous time layer: $\hat{f}^{s=0} = f$, and the termination criterion is given by the inequality

$$\left|\hat{f}^{s+1} - \hat{f}^{s}\right| \le \varepsilon_1 \left|\hat{f}^{s}\right| + \varepsilon_2, \quad \varepsilon_1 > 0, \quad \varepsilon_1 > 0,$$

where f is one of the functions n, N, φ. This inequality must be valid for all functions simultaneously.

For solution of the written set of the equations it is possible to use various methods. In particular, for solution of the Poisson equation concerning the electric field potential and Helmholtz equation (in the case of D_x=D_y) concerning the free electrons concentration it is suitable to use the Fast Fourier Transform (FFT) method because it allows to find the solution with high accuracy. This method also can be used for computation of the first and second difference derivatives entering into the equations (4). It is important for increasing the computational accuracy because in some cases the solution of the problem (4) – (5) can lose the symmetry at computation of the difference derivatives at their usual definition. At computer simulation in the given paper we used FFT from Intel MKL [15].

Scheme 2.

At the construction of finite-difference scheme for the problem (1) - (2) it is also possible to use the method of total approximation [12], [16]:

$$\frac{\tilde{n} - n}{\tau} = \frac{1}{2}(D_x \tilde{n}_{\bar{x}x} - D_x \mu_x \left[(\tilde{n}_{x_{+0.5}} \tilde{\varphi}_x)_{\bar{x}}\right]) + \frac{1}{2}(D_y n_{\bar{y}y} - D_y \mu_y \left[(n_{y_{+0.5}} \varphi_y)_{\bar{y}}\right]) + \frac{1}{2}(\tilde{G} - \tilde{R}),$$
$$\frac{\tilde{N} - N}{\tau} = \tilde{G} - \tilde{R}, \quad \tilde{\varphi}_{\bar{x}x} = \gamma(\tilde{n} - \tilde{N}),$$
$$G = q_0 q \delta(n, N, \varphi), \quad R = \frac{nN - n_0^2}{\tau_p}, \quad \tilde{G} = q_0 q \delta(\tilde{n}, \tilde{N}, \tilde{\varphi}), \quad \tilde{R} = \frac{\tilde{n}\tilde{N} - n_0^2}{\tau_p}, \tag{6}$$

 ISBN: 978-84-614-6167-7

$$\frac{\hat{n}-\tilde{n}}{\tau}=\frac{1}{2}(D_x\tilde{n}_{\bar{x}x}-D_x\mu_x\left[(\tilde{n}_{x_{+0.5}}\tilde{\varphi}_x)_{\bar{x}}\right])+\frac{1}{2}(D_y\hat{n}_{\bar{y}y}-D_y\mu_y\left[(\hat{n}_{y_{+0.5}}\hat{\varphi}_y)_{\bar{y}}\right])+\frac{1}{2}(\hat{G}-\hat{R}),$$

$$\frac{\hat{N}-N}{\tau}=\hat{G}-\hat{R},\qquad \hat{\varphi}_{\bar{x}x}=\gamma(\hat{n}-\hat{N}),$$

$$\hat{G}=q_0q(x,y,t)\delta(\hat{n},\hat{N},\hat{\varphi}),\quad \hat{R}=\frac{\hat{n}\hat{N}-n_0^2}{\tau_p},$$

$$1<i_x<N_x-1,\ 1<i_y<N_y-1,\ k>0.$$

Thus, the computation is made in two stages – at first one we compute values of $\tilde{n}$ at a semi-layer *k+0.5* of time. Then by means of found values we compute $\hat{n}$ on time layer *k+1*. Obviously, written equations are nonlinear ones. Hence it is necessary to construct the iteration process which looks as follows:

$$\frac{\overset{s+1}{\tilde{n}}-n}{\tau}=\frac{1}{2}(D_x\overset{s+1}{\tilde{n}}_{\bar{x}x}-D_x\mu_x\left[(\overset{s}{\tilde{n}}_{x_{+0.5}}\overset{s}{\tilde{\varphi}}_x)_{\bar{x}}\right])+$$

$$+\frac{1}{2}(D_yn_{\bar{y}y}-D_y\mu_y\left[(n_{y_{+0.5}}\varphi_y)_{\bar{y}}\right])+\frac{1}{2}(\overset{s,s+1}{\tilde{G}}-\overset{s,s+1}{\tilde{R}}),$$

$$\frac{\overset{s+1}{\tilde{N}}-N}{\tau}=\overset{s,s+1}{\tilde{G}}-\overset{s,s+1}{\tilde{R}},\quad \overset{s+1}{\tilde{\varphi}}_{\bar{x}x}=\gamma(\overset{s+1}{\tilde{n}}-\overset{s+1}{\tilde{N}}),$$

$$\overset{s,s+1}{\tilde{G}}=q_0q\delta(\overset{s}{\tilde{n}},\overset{s+1}{\tilde{N}},\overset{s}{\tilde{\varphi}}),\quad \overset{s,s+1}{\tilde{R}}=\frac{\overset{s}{\tilde{n}}\overset{s+1}{\tilde{N}}-n_0^2}{\tau_p},$$

$$\frac{\overset{s+1}{\hat{n}}-\tilde{n}}{\tau}=\frac{1}{2}(D_x\tilde{n}_{\bar{x}x}-D_x\mu_x\left[(\tilde{n}_{x_{+0.5}}\tilde{\varphi}_x)_{\bar{x}}\right])+ \tag{7}$$

$$+\frac{1}{2}(D_y\overset{s+1}{\hat{n}}_{\bar{y}y}-D_y\mu_y\left[(\overset{s}{\hat{n}}_{y_{+0.5}}\overset{s}{\hat{\varphi}}_y)_{\bar{y}}\right])+\frac{1}{2}(\overset{s,s+1}{\hat{G}}-\overset{s,s+1}{\hat{R}}),$$

$$\frac{\overset{s+1}{\hat{N}}-N}{\tau}=\overset{s,s+1}{\hat{G}}-\overset{s,s+1}{\hat{R}},\quad \overset{s+1}{\hat{\varphi}}_{\bar{x}x}=\gamma(\overset{s+1}{\hat{n}}-\overset{s+1}{\hat{N}}),$$

$$\overset{s,s+1}{\hat{G}}=q_0q(x,y,t)\delta(\overset{s}{\hat{n}},\overset{s+1}{\hat{N}},\overset{s}{\hat{\varphi}}),\quad \overset{s,s+1}{\hat{R}}=\frac{\overset{s}{\hat{n}}\overset{s,s+1.}{\hat{N}}-n_0^2}{\tau_p},$$

$$1<i_x<N_x-1,\ 1<i_y<N_y-1,\ k>0.$$

For the solution of the equations concerning the free electrons concentration we used tridiagonal algorithm or 1D FFT method. Derivatives entering into the right side of these equations were computed by using the FFT method.

 ISBN: 978-84-614-6167-7

4. Computer simulation results.

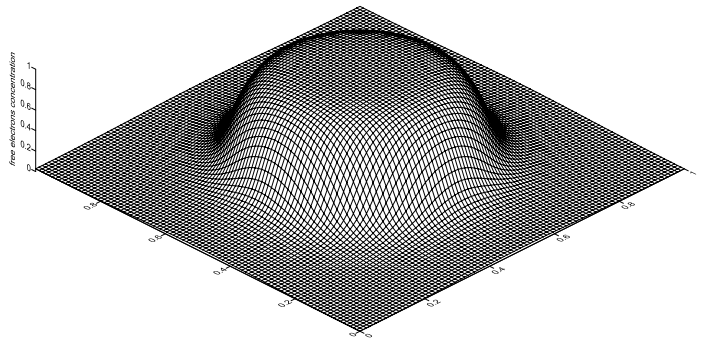
Analyzing the computer simulation results for various sets of parameters for the problem (1) – (2) it is possible to assert that the conservative finite-difference scheme 1 constructed above is an effective tool for solving this problem. This finite-difference scheme allows to investigate the various nonlinear processes occurring in the semiconductor under the action of high-intensity laser pulse: formation of domain with high concentration of ionized donors (optical bistability mode) (fig. 1), or the mode of free electron concentration oscillation (fig. 2). It is necessary to stress that using the conservative finite-difference scheme in a combination with FFT method allows to make a computation with high accuracy.

As it follows from Fig. 1, Fig. 2 the solution keeps symmetry of the concentration distribution even for calculation with enough large mesh steps and during long time interval. It is also necessary to note there is the high accuracy of the invariant $Q(t)$ preservation – $Q(100)=3.181042^{-10}$ for the case corresponding Fig. 1 and $Q(1000)= 3.972298^{-13}$ for the case corresponding Fig. 2. Thus, it is possible to conclude that this finite-difference scheme also will be effective for the computation of evolution of semiconductor characteristics with taking into account of nonlinear dependence of electrons mobility coefficient and of the absorption coefficient from light induced electric field.

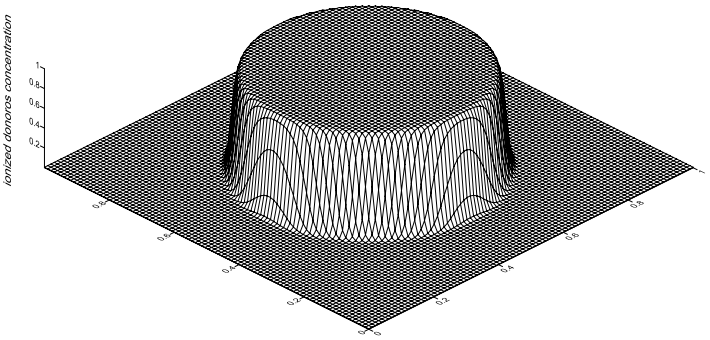
With respect to the finite-difference scheme 2 it is necessary to notice that for the problem (1) – (2) its application field is limited only by computation of smooth mode of changing the semiconductor characteristics when the distribution of free electrons concentration and ionized donors concentration are the functions like Gaussian ones (Fig. 3). For other modes of optical radiation interaction with semiconductor the computation provided on this finite-difference scheme interrupted because of the accumulation of rounding error. Reducing the grids steps does not result in substantial improvement of the situation.

 ISBN: 978-84-614-6167-7

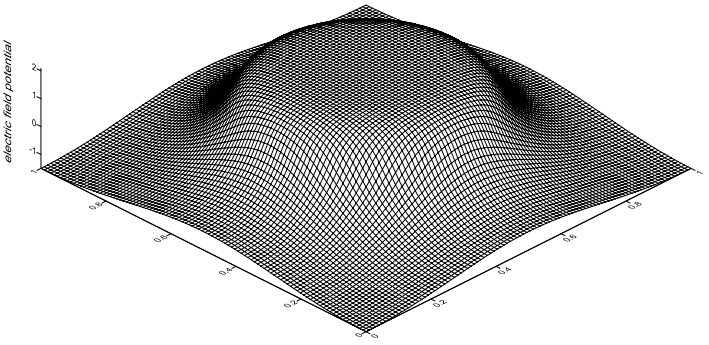
Fig. 1. Calculations by scheme 1. Distributions of free electrons concentration n (a), ionized donors concentration N (b), electric field potential φ (c) realized at time moment t=100 for the absorption coefficient δ_φ, grids steps $h_x = h_y = 10^{-2}, \tau = 10^{-3}$ and parameters values $\beta = 8, \mu_x = \mu_y = 1, D_x = D_y = 10^{-3}, \gamma = 10^3, n_0 = 0.01, \tau_p = 1, q_0 = 1.$



a)



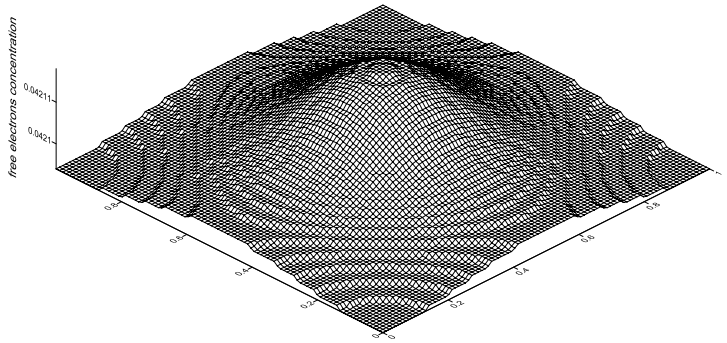
b)



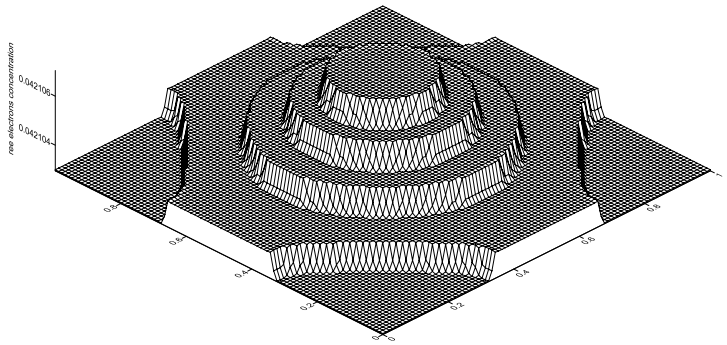
c)

 ISBN: 978-84-614-6167-7

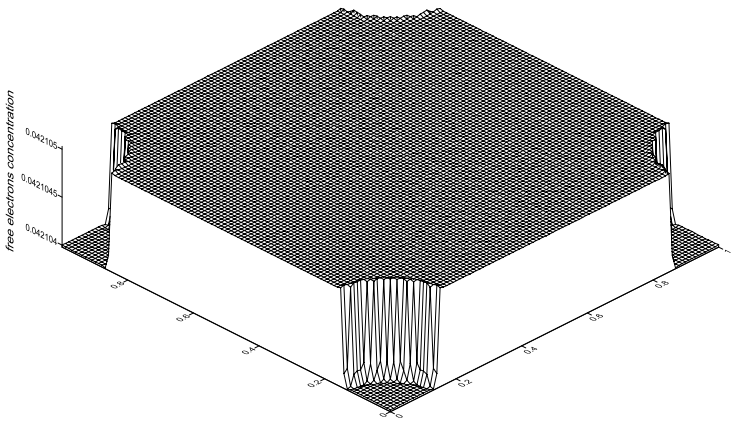
Fig. 2. Calculations by scheme 1. Distributions of free electrons concentration n realized at time moments t=500 (a), 600 (b), 700 (c) for the absorption coefficient δ_n, grids steps $h_x = h_y = 10^{-2}, \tau = 10^{-2}$ and parameters values $\psi = 2.553, \xi = 3, \mu_x = \mu_y = 0, D_x = D_y = 10^{-3}, \gamma = 10^3, n_0 = 0.01, \tau_p = 1, q_0 = 1.$



a)



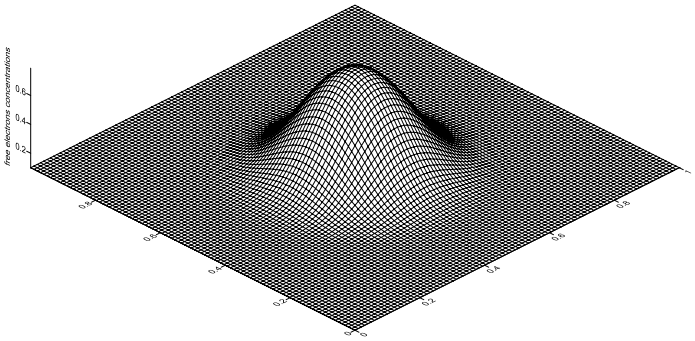
b)



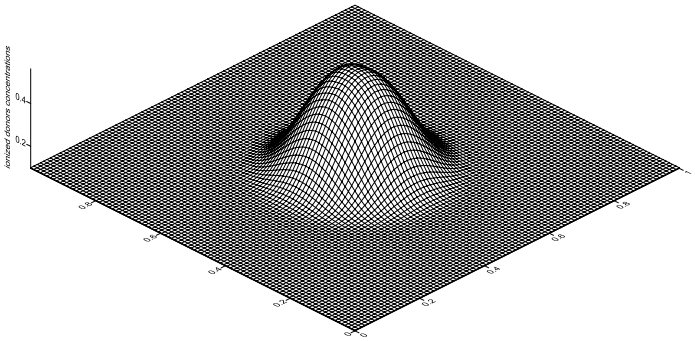
c)

 ISBN: 978-84-614-6167-7

Fig. 3. Calculations by scheme 2. Distributions of free electrons concentration *n* (a), ionized donors concentration *N* (b) realized at time moment t=10 for the absorption coefficient δ_φ, grids steps $h_x = h_y = 10^{-2}, \tau = 10^{-3}$ and parameters values $\beta = 1, \mu_x = \mu_y = 1, D_x = D_y = 10^{-4}, \gamma = 10^3, n_0 = 0.1, \tau_p = 1, q_0 = 1.$



a)



b)

Acknowledgements
This paper was partly financially supported by Russian Foundation for Basic Research (grant number 10-07-91152 GFEN-a).

5. References

[1] R. P. PRASANKUMAR, A. SCOPATZ, D. J. HILTON, A. J. TAYLOR, R. D. AVERITT, J. M. ZIDE AND A. C. GOSSARD, *Carrier dynamics in self-assembled ErAs nanoislands embedded in GaAs measured by optical-pump terahertz-probe spectroscopy*, Appl. Phys. Lett. 86(20), 201107 (2005)

[2] R. D. AVERITT AND A. J. TAYLOR, *Ultrafast optical and far-infrared quasiparticle dynamics in correlated electron materials*, J. Phys.-Condens. Matter 14, R1357-R1390 (2002).
[3] J. LLOYD-HUGHES, S. K. E. MERCHANT, L. FU, H. H. TAN, C. JAGADISH, E. CASTRO-CAMUS AND M. B. JOHNSTON, *Influence of surface passivation on ultrafast carrier dynamics and terahertz radiation generation in GaAs*, Appl. Phys. Lett. 89 (23), 232102-1-3 (2006).
[4] M. C. BEARD, G. M. TURNER, AND C. A. SCHMUTTENMAER, *Subpicosecond carrier dynamics in low-temperature grown GaAs as measured by time-resolved terahertz spectroscopy*, J. Appl. Phys. 90 (12), 5915-5923 (2001).
[5] D. G. COOKE, F. A. HEGMANN, YU. I. MAZUR, ZH. M. WANG, W. BLACK, H. WEN, G. J. SALAMO, T. D. MISHIMA, G. D. LIAN, AND M. B. JOHNSON, *Ultrafast carrier capture dynamics in InGaAs/GaAs quantum wires*, J. Appl. Phys. 103 (2), 023710 (2008).
[6] JAN LINNROS, *Carrier lifetime measurements using free carrier absorption transients*, J. Appl. Phys. 84 (1), 275 (1998).
[7] N.B. DELONE, V. P. KRAYNOV, *Nonlinear ionization of atoms by laser radiation*, Moscow: Phismatlit, 2001 (in Russian).
[8] V.L. BONCH-BRUEVICH, S.G. KALASHNIKOV, *Physics of semiconductors*, Moscow: Nauka, 1990 (in Russian).
[9] R. SMITH, *"Semiconductors,"* Cambridge University Press, 1978.
[10] V.A. TROFIMOV, M.M. LOGINOVA, "On the Efficiency of Some Difference Schemes for Problems of Nonlinear Interaction of Optical Radiation with a Semiconductor," *J. Differential Equations*, vol. 42, N. 8, pp. 1189–1198, 2006.
[11] V.A. TROFIMOV, M.M. LOGINOVA, "Comparison of some difference schemes for the problem of femtosecond pulse interaction with semiconductor in the case of nonlinear mobility coefficient," *In Book "Lecture Notes in Computer Science"*, Editors Z.Li et al. Springer-Verlag GmbH, vol. 3401, pp. 535-542, 2005.
[12] A.A. SAMARSKII, *Theory of Difference Schemes*, Moscow: Nauka, 1989.
[13] V.P. IL'IN, *Difference methods for electrophysics problems*, Novosibirsk: Nauka, 1985 (in Russian).
[14] V.A. TROFIMOV, M.M. LOGINOVA. Conservative Finite-Difference Scheme for a Two-Dimensional Problem of Plasma Generation in a Semiconductor under a Laser Pulse Action. // *Proceedings of "10th International Conference on Laser&Fiber-Optical Networks Modeling"*.Sevastopol, Ukraine. 12-14 September, 2010. P.186-188.
[15] http://software.intel.com/en-us/articles/intel-math-kernel-library-documentation/
[16] G.I. MARCHUK, *Methods of computational mathematics*, Moscow: Nauka, 1977 (in Russian).

 ISBN: 978-84-614-6167-7

Proceedings of the 11th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE2011
Benidorm, Spain, 26-30 June 2011

A Modified Ant Colony Optimization for the Replenishment Policy of the Supply Chain under Asymmetric Deterioration Rate

Jui-Tsung Wong[1] Kuei-Hsien Chen[2] and Chwen-Tzeng Su[3]

[1]Department of International Business Management, Shih Chien University Kaohsiung Campus, Taiwan, Republic of China.
[2]Department of Business Administration, Nan Jeon Institute of Technology, Taiwan, Republic of China.
[3]Department of Industrial Engineering and Management, National Yunlin University of Science and Technology, Taiwan, Republic of China.
emails: wongjt@mail.kh.usc.edu.tw, bm018@mail.njtc.edu.tw, suct@yuntech.edu.tw

Abstract

This paper proposes a dynamic lot-sizing problem with asymmetric deteriorating commodity (DLSPADC). DLSPADC is proved to be an NP-hard problem. Of the property of the DLSPADC optimal solution, the demand may be split, and the possible split point may be a real number value. Under such circumstances, the number of the possible replenishment planning of each period is enormous. As a result, to more efficiently solve the problem, the possible partition points of the optimal replenishment policy have to be lessened. To solve the DLSPADC problem, this paper discovers several properties and theorem. Ant colony optimization (ACO) is developed from the theorem and 0-1 binary code to solve the problem. The modified ACO algorithms of history literature were tested. The results show the ACO algorithms with pheromone intensity bounds do not perform as well in DLSPADC problems.

Keyword: dynamic lot-sizing problem, modified ant colony optimization, asymmetric deteriorating commodity
MSC2000: 68W15

1. Introduction

Globalization is the dominant force driving the market today. It allows upstream industries and downstream industries to locate in different geographical regions. To survive in this competitive environment, the supply chain has now become part of the operational model. Under vendor managed inventory (VMI), which makes good use of the supply chain, the supplier is responsible for managing the buyer's inventory, which shortens the leading time for imbibing the demand information on the supplier's side. This paper proposes a dynamic lot-sizing problem based on the VMI supply chain.

Since Wagner and Whitin [25] first used dynamic programming to solve a

 ISBN: 978-84-614-6167-7

one-stage dynamic lot-sizing problem, lot-sizing problems have been important topics. Multistage lot-sizing problems have also been discussed ([28], [6]). Also, Lee *et al*. [17] used dynamic programming to solve a two-stage dynamic lot-sizing problem with transport cost. Jaruphongsa *et al*., [15] further used dynamic programming to solve a two-stage dynamic lot-sizing problem with delivery time window while transport costs were also under consideration. However, since the serviceable amount of deteriorating commodities such as perishable foods, electronic parts, drugs and bloods decrease with time, lot-sizing problems taking deterioration into account are often discussed in continual time-varying situations. Gupta and Gerchak [7] studied a lot-sizing problem for a deteriorating commodity when the demand is constant. In this problem, deteriorating commodities are considered for exponential remaining lifetimes. Hyun Kim and Hong [10] studied a lot-sizing problem with deteriorating production process and rework when the demand is constant. Abad [1] studied a lot-sizing problem with an exponential deterioration rate. In this problem, backordering is allowed. Literatures on other lot-sizing problems under continual time-varying models with deteriorating commodities are: Warrier and Shah [26], Hwang and Shinn [11], Jalan and Chaudhuri [16], Misra [18], Sarker *et al*. [20], Papachristos and Skouri [19], Balkhi [2], and Teng *et al*. [23]. Next, this paper introduces the lot-sizing problems considering deterioration in a discrete time-varying scenario. Smith [21] proposed a one-stage dynamic lot-sizing problem in which the commodity deteriorates after a constant period. The goal was to find the maximum profit by deciding the replenishment schedule and selling prices. Friedman and Hoch [5] proposed a one-stage dynamic lot-sizing problem with non-increasing deterioration rate. Hsu [12] proposed a one-stage dynamic lot-sizing problem considering perishable inventory. He proposed dynamic programming to solve this problem, which assumed that supply is unlimited and backordering is not allowed. Hsu [13] further explored a one-stage dynamic lot-sizing problem for perishable commodity when backordering is allowed. In these related literatures, the assumptions of the dynamic lot-sizing problem were the supply is unlimited and that only the deterioration of serviceable commodity is considered. In reality, because of resources constraints and seasonal variations, the supply of material is limited and cannot be controlled. In addition, both the raw material and the commodity deteriorate; yet, the raw material may be processed into a serviceable commodity (for example, by adding preservatives to foods). This means the deterioration rates for the raw material and the commodity differ, therefore it is called *asymmetric deterioration*. Such industries include fishing, agriculture, recycling of secondary materials. Also, such issues are similar to lot-sizing problems with inventory limits. Bitran and Yanasse [3] probe into the computational complexity and properties of the capacitated lot size problem under a particular cost structure. Gutiérrez *et al*. [9] proposed new properties for non-backlogging lot-sizing problems whose productivity is constants. They assume the replenishment amount is a non-negative integral. Different from the problems in this paper, the bound of inventory limit is not amassed. This paper proposes a dynamic lot-sizing problem for a similar industry.

The remaining sections of the paper are as follows. Section 2 formulates

 ISBN: 978-84-614-6167-7

the DLSPADC, and discusses the character and complexity of the problem. Section 3 introduces the DLSPACD solution procedure of the algorithm based on ant optimization. Section 4 compares the performance of each algorithm. Section 5 provides a summary and some closing remarks.

2. Modeling for a supply chain

The dynamic lot-sizing problem in this paper is a replenishment planning problem that takes the asymmetric deteriorating commodity into account. In this model, the final customer demand and the supplied material quantity are deterministic. The DLSPADC mainly involves deciding the replenishment period (that is, if during period *t*, supply is being provided to downstream businesses, then *t* is called a *replenishment period*) and replenishment amount for a simple supply chain where there are a manufacturer and a wholesaler, as shown in figure 1.

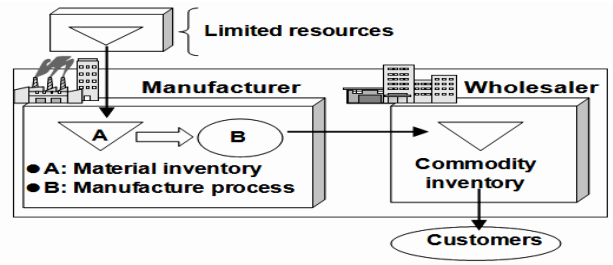


Figure 1. Graphical representation of the proposed problem

2.1 Assumptions and problem formula

Let *t* and *T* represent the period index and the planning horizon respectively, for $t = 1,\ldots, T$. The symbols that make the problem are defined as figure 2.

The proposed problem involves finding the optimal replenishment policy in each replenishment cycle length (i.e., the length of the planning horizon). In a deterministic problem, the focus would be to explore the replenishment planning when the time is in the frozen period of the time fences; that is, when the quantity of material and demand are deterministic. This paper solves a dynamic lot-sizing problem. In this problem, the demand d_t and procurement quantity of material b_t are both deterministic.

This section constructs the proposed dynamic lot-sizing problem. The main assumptions are as follows:

(1) $h_{2t} \le h_{1t}$, for $t = 1, 2, \ldots, T$.

(2) $K_{1t} \ge K_{1,t+1}$, $p_{2t} \ge p_{2,t+1}$, $p_{1t} \ge p_{1,t+1}$, $s_{1t} \ge h_{1t}$, for $t = 1, 2,\ldots, T$.

(3) No deterioration cost.

(4) Calculate serviceable materials at the beginning of the period.

(5) d_t and b_t are deterministic and uncontrollable, for $t = 1, 2,\ldots, T$.

(6) Backordering is not allowed at the manufacturer, but is allowed at the wholesaler.

(7) $\alpha_2 \ge \alpha_1$.

(8) d_t is a non-negative integer.

(9)The supply chain adopts the VMI policy; in other words, the manufacturer decides the inventory level at the wholesaler.

 ISBN: 978-84-614-6167-7

d_t	denotes the demand at the wholesaler in period t
b_t	denotes the procurement quantity of material at the manufacturer in period t
y_t	denotes the production and transportation (i.e., replenishment) quantity of commodity in period t
I_{2t}	denotes the inventory level of material at the manufacturer at the end of period t
I_{1t}	denotes the inventory level of commodity at the wholesaler at the end of period t
K_{1t}	denotes the fixed cost of production and transportation to the wholesaler in period t
p_{2t}	denotes the unit procurement cost of material at the manufacturer in period t
p_{1t}	denotes the unit cost of production and transportation to the wholesaler in period t
h_{2t}	denotes the unit holding cost of material at the manufacturer in period t
h_{1t}	denotes the unit holding cost of commodity at the wholesaler in period t
s_{1t}	denotes the unit shortage cost of commodity at the wholesaler in period t
α_2	denotes the inventory deterioration rate of material at the manufacturer
α_1	denotes the inventory deterioration rate of commodity at the wholesaler

Figure 2. The notations of DLSPADC

The objective is to minimize the total cost of adopting the VMI policy. The total cost include the manufacturer's cost, cost of providing replenishment for the wholesaler, and wholesaler's cost, as shown in equation (1). The formulation of the DLSPADC model is as follows:

Minimize

$$TC(y_t) = \sum_{t=1}^{T}\left[p_{2t}b_t + h_{2t}I_{2t} + K_{1t}\delta(y_t) + p_{1t}y_t + h_{1t}\left(I_{1t}\right)^{+} + s_{1t}\left(I_{1t}\right)^{-} \right] \quad (1)$$

Subject to:

$$y_t = (1-\alpha_2)I_{2,t-1} + b_t - I_{2t} \qquad t = 1,2,...,T \quad (2)$$

$$d_t = (1-\alpha_1)\left(I_{1,t-1}\right)^{+} - \left(I_{1,t-1}\right)^{-} + y_t - I_{1t} \qquad t = 1,2,...,T \quad (3)$$

$$I_{2,0} = I_{1,0} = 0 \qquad t = 1,2,...,T \quad (4)$$

$$y_t, I_{2t} \geq 0 \qquad t = 1,2,...,T \quad (5)$$

where $\delta(x)=1$ if $x>0$ and 0 otherwise. $(x)^{+} = \max\{x, 0\}$. $(x)^{-} = -\min\{x, 0\}$. Equation (2) and (3) are the inventory balance constraints at the manufacturer and at the wholesaler respectively. Equation (4) shows the assumption of this paper where the inventory levels at the manufacturer and at the wholesaler are set at the level of the beginning of the initial period. Equation (5) is a non-negative integer constraint.

Note that y_t and I_{2t} are non-negative numbers, which means y_t is set to be equal to or less than $(1-\alpha_2)I_{2,t-1} + b_t$.

2.2 The property and theorem of the problem

Note that when $j < i$, this paper defines $\sum_{v=i}^{j} x_v = 0$. Moreover, the replenishment quantity must satisfy the constraint in supply quantity; that is, $y_V \leq (1-\alpha_2)\, I_{2,V-1} + b_V$.

Property 1.1. There exists an optimal solution under $\alpha_2 > \alpha_1$ so:

 ISBN: 978-84-614-6167-7

$I_{1,t-1}y_t\left[(1-\alpha_2)I_{2,t-1}+b_t-y_t\right]=0$, for $t=1,2,...,T$.

Proof. Suppose that $I_{1,j-1}>0$, $y_j>0$, $(1-\alpha_2)I_{2,j-1}+b_j-y_j>0$, there then exists a previous replenishment period i which makes $I_{1,j-1}>0$. Here, on the condition the quantity of supply is not affected (i.e. the value of $I_{1,j}$ remains the same), Q can be added to the replenishment amount in period j, which makes $I_{1,j-1}$ equal to 0. When y_j equals $(1-\alpha_2)I_{2,j-1}+b_j$, $I_{1,j-1}$ cannot be equal to 0. □

In figure 3 (a), the $I_{1,j}$ is point e, no matter in the replenishment planning {a_1, b_1, c_1, d_1, e} before the adjustment, or in the replenishment planning {a_2, b_2, c_2, d_2, e} after the adjustment. Thus, the satisfaction of the demands is unaffected, and the holding cost can be lessened. In figure 3 (b), because $I_{2,j}$ is equal to 0, y_i cannot be adjusted to make $I_{1,j-1}$ equals to 0. Property 1.1 shows there exists an optimal solution so when $I_{1,t-1}>0$, then $y_t \in \{0, (1-\alpha_2)I_{2,t-1}+b_t\}$. This shows the properties of the proposed problem. The above shows the properties of the proposed problem. The way to solve this problem is through deciding the replenishment period and the partition points.

Property 1.2. There exists an optimal solution under $\alpha_2=\alpha_1$ so:

$I_{1,t-1}y_t=0$, for $t=1,2,...,T$.

Proof. The proving method is similar to Property 1.1. The only difference is the reduced amount of $I_{1,j-1}$ resulted from $\alpha_2=\alpha_1$ must be equal to the increased amount of $I_{2,j-1}$. Under such a circumstance, the replenishment amount Q must be able to be increased in period j to make $I_{1,j-1}$ equal to 0.

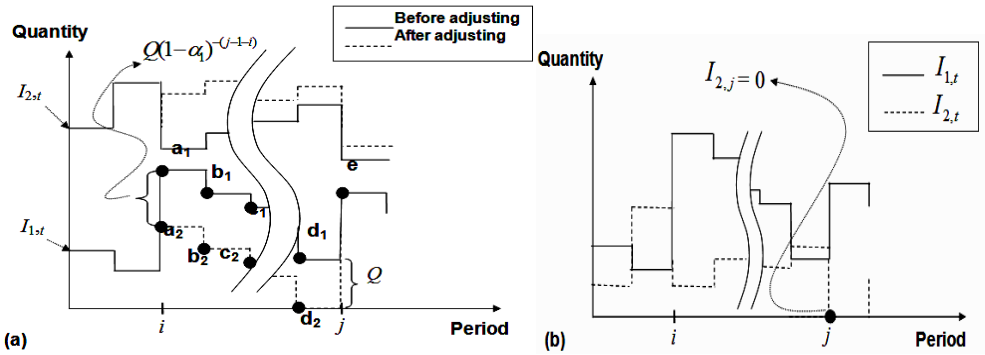


Figure 3. Illustration of Property 1.1: (a) When $I_{1,j-1}$ can be adjusted to 0; (b) When $I_{1,j-1}$ cannot be adjusted to 0

Lemma 1. There exists an optimal solution so the demands in those periods, excluding the first and the last period, cannot be split when the demand in a sequence of periods is possibly met by a replenishment period.

Proof. Within a particular time span in a replenishment period during which the demand can be met, this program can be considered without a supply constraint in replenishment planning. Under such circumstances, the basic ideas in Wagner and Whitin's [25] theorem 2 can be used to produce the result of

 ISBN: 978-84-614-6167-7

Lemma 1 in this paper. □

Theorem 1. There exists an ideal solution so in replenishment period *j*, the value of y_j must satisfy:

$$y_j = \left\{ \min \left\{ (1-\alpha_2) I_{2,j-1} + b_j, \left\{ (d_i - u_i), (d_i - u_i) + \sum_{v=i+1}^{i+1} d_v \right. \right. \right. \\ \left. \left. \left. , (d_i - u_i) + \sum_{v=i+1}^{i+2} d_v, \ldots, (d_i - u_i) + \sum_{v=i+1}^{T} d_v \right\} \right\} \right\} , \quad (6)$$

where d_i represents the demand in period *i*; period *i* is where the position of the split point of the previous replenishment period is found; u_i represents the amount of the previous replenishment period that satisfies the demand of period *i*.

Proof. Based on the results of Lemma 1 and Property 1.1.

Therefore, Theorem 1 shows the replenishment amount for a certain demand within a feasible supply range should not be split in its consideration (i.e. $y_t \in \{0, \sum_{j=l_1}^{l_2} d_j \mid 1 \le l_1 \le l_2 \le T\}$).

3. Implementing the ACO approach

Ant colony optimization (ACO) was first introduced in Dorigo's [4] doctoral thesis. It is a meta-heuristic technique that mimics the behavior of real ants. Ants communicate and optimize the path between the nest and the food source through pheromone. An artificial ant builds a solution through probability function, whose generation hinges on the pheromone information. When the ant completes forming a solution, the amount of pheromone trail on the path would affect the solution quality, that is, the better the solution, the more pheromone trail is left on the path. This paper develops the ACO algorithms in binary code to solve the proposed DLSPADC.

3.1 Solution construction

The 0-1 binary-coded system (Wong and Chen, 2010) is applied to this problem. In performing the ACO algorithm, the ant finds the optimal replenishment period and the partition point in demand by deciding whether the bit on the code is 0 or 1. In solving the DLSPADC, there are two phases in the ant's construction of a solution. The first phase decides the replenishment period, whereas the second phase decides the partition point in demand given the already determined replenishment period. The partition point must be feasible (i.e., satisfy the constraint in supply). The transition probability is the main function used by the ant in its decision-making. An ant completes a tour by deciding the value of all bits through the pheromone trail, and prefers choosing the bits with the greater pheromone trail to form the solution.

The transition probability of the ant in phase 1 is as follows:

$$P^1(t) = \begin{cases} \dfrac{\tau_1(t)}{\tau_1(t) + \tau_0(t)} & \text{demands have not been satisfied} \\ 0 & \text{otherwise} \end{cases} , \quad (7)$$

where $P^1(t)$ is the probability of the ant choosing 1 in period *t* (i.e., the probability of period *t* being a replenishment period). $\tau_j(t)$ is the amount of

 ISBN: 978-84-614-6167-7

pheromone trail in the position j during period t, $j \in \{0, 1\}$.

The transition probability of the ant in phase 2 is as follows:

$$P^{2k}_{s^1(\tilde{d}'_t,\tilde{i})s^2}(d'_t,i)=\frac{\tau_1(d'_t,i)}{\sum_{d''}\sum_{i''}\tau_1(d'',i'')} \quad (d'',i'') \in N^k_{s^1(\tilde{d}'_t,\tilde{i})s^2} \tag{8}$$

where $P^{2k}_{s^1(\tilde{d}'_t,\tilde{i})s^2}(d'_t,i)$ is the probability of ant k choosing the partition point in demand (d'_t,i) during the current replenishment period s^2 when the partition point in the previous replenishment period is s^1 and the chosen partition point is $(\tilde{d}'_t,\tilde{i})$. $\tau_1(d'_t,i)$ is the amount of pheromone trail in the element i of the demand d'_t in the sub-vector. $N^k_{s^1(\tilde{d}'_t,\tilde{i})s^2}$ is set of the feasible *applicable* partition point of demand of ant k in the current replenishment period s^2 when the previous replenishment period is s^1 and the partition point is $(\tilde{d}'_t,\tilde{i})$.

The procedure of an ant constructing a solution is as follows.

Step 1. R_t is determined through equation (7).

Step 2. Let $t = 1$.

Step 3. If $R_t = 1$, then perform the decision of the partition point (to step 3.1).

Step 3.1. With the concept of equation (6), determine $N^k_{s^1(\tilde{d}'_t,\tilde{i})s^2}$, for $s^2 = t$. Here, if the limit of the feasible partition point is a demand split, than the generation of a sub-vector is needed.

Step 3.2. Use equation (8) to determine the partition point of the current replenishment period s^2.

Step 4. If $t = T$, end. Otherwise, next step.

Step 5. $t = t + 1$ and go to the step 3.

3.3 Update of pheromone trails

In the ACO algorithm, the pheromone information is crucial for the ant to find the optimal solution. Traditional ACO is applied to DLSPADC, and ant k sets pheromone in the path of a completed tour in vector $\boldsymbol{R}$ as follows:

$$\Delta\tau^k_j(t)=\begin{cases}\dfrac{Q}{f_k} & \text{if } j \in T^{1k}\\ 0 & \text{otherwise}\end{cases} \tag{9}$$

where Q is an adjustable parameter. f_k is the objective function value of ant k. T^{1k} is the path of a completed tour by ant k in vector $\boldsymbol{R}$. The traditional pheromone placement method is used in the ACO algorithm of Hiroyasu *et al*. [14].Update the amount of pheromone trail $\tau_j(t)$, as shown in equation (10):

$$\tau_j(t)=\rho\tau_j(t)+\Delta\tau_j(t) \tag{10}$$

where ρ is a parameter between 0 and 1. $\Delta\tau_j(t)=\sum_{k=1}^{U}\Delta\tau^k_j(t)$.

Besides, ant k sets pheromone in the path of a completed tour in vector $\boldsymbol{E}$ as the following equation:

$$\Delta\tau^k(d'_t,i)=\begin{cases}\dfrac{Q}{f_k} & \text{if } (d'_t,i) \in T^{2k}\\ 0 & \text{otherwise}\end{cases} \tag{11}$$

 ISBN: 978-84-614-6167-7

where T^{2k} is the path of a completed tour by ant k in vector $\boldsymbol{E}$.

Update the amount of pheromone trail $\tau_1(d_t', i)$ as follows:

$$\tau(d_t', i) = \rho\tau(d_t', i) + \Delta\tau(d_t', i) \tag{12}$$

where $\Delta\tau(d_t', i) = \sum_{k=1}^{U} \Delta\tau^k(d_t', i)$.

Through the modified pheromone placement method, this paper tries to make the solution quality more influential on the pheromone amount placed. In the modified method, ant k sets pheromone in the path of a completed tour in vector $\boldsymbol{R}$ as follows:

$$\Delta\tau_j^k(t) = \begin{cases} \dfrac{Q}{f_k} & \text{if } f_\vartheta - f_k \geq 0 \text{ and } j \in T^{1k} \\ \dfrac{1}{\log_{10}\left(\max\left\{10, \left(f_k - f_\vartheta\right)\right\}\right)} \dfrac{Q}{f_k} & \text{if } f_\vartheta - f_k < 0 \text{ and } j \in T^{1k} \\ 0 & \text{otherwise} \end{cases} \tag{13}$$

where Q is an adjustable parameter. f_k is the objective function value of ant k. f_ϑ is the objective function value of the current best solution. T^{1k} is the path of a completed tour by ant k in vector $\boldsymbol{R}$.

Ant k sets pheromone in the path of a completed tour in vector $\boldsymbol{E}$ as the following equation:

$$\Delta\tau^k(d_t', i) = \begin{cases} \dfrac{Q}{f_k} & \text{if } f_\vartheta - f_k \geq 0 \text{ and } (d_t', i) \in T^{2k} \\ \dfrac{1}{\log_{10}\left(\max\left\{10, \left(f_k - f_\vartheta\right)\right\}\right)} \dfrac{Q}{f_k} & \text{if } f_\vartheta - f_k < 0 \text{ and } (d_t', i) \in T^{2k} \\ 0 & \text{otherwise} \end{cases} \tag{14}$$

where T^{2k} is the path of a completed tour by ant k in vector $\boldsymbol{E}$.

Property 2. There exists an optimal solution so when the partition point of the demand in the previous replenishment period S falls in the last period of the planning horizon, the amount of the replenishment period V must satisfy:

$$y_V = \min\left\{\left(1-\alpha_2\right) I_{2,t-1} + b_t, \left(d_T - u_2\right)\left(1-\alpha_2\right)^{-(T-V)}\right\},$$

where u_2 is the amount of the previous replenishment period S that satisfies term T.

Proof. Following Theorem 1, when the demand partition point of the previous replenishment period S falls in the last period of the planning horizon and there are demands remaining unsatisfied, the partition point of replenishment period V should either satisfy the remaining demands or be the maximum of the workable partition point. Otherwise, it equals a violation of Theorem 1.

3.4 Mutation operator

The mutation in this algorithm mainly changes the R_t chosen by the ant; that is, the replenishment period would be changed. This paper uses the two-point

 ISBN: 978-84-614-6167-7

mutation method.

3.5 Procedure of the algorithms

This paper decides the partition point and the replenishment period through the ACO algorithm to solve the DLSPADC. The algorithm performance of the paper is shown through the comparison with the ACO algorithm of history literature. The details of each algorithm are as follows:

(1) Touring ant colony optimization (TACO) [14] is traditional ACO algorithm without visibility.

(2) ACO+MO+MP is the ACO algorithm with mutation operator and modified pheromone trail placement.

In these algorithms, there are three phases in an ant's construction of solution. The first phase is deciding the R_t. In the second phase, if the condition of mutation is being met, then R_t undergoes mutation operator. A dynamic mutation rate is being applied in the mutation process; that is, the original mutation rate ϕ is being multiplied by a function. This function value decreases as the number of iterations increases. In the third phase, $E_{d_t'}$ is determined.

The procedure of TACO is shown in figure 4(a). The method proposed by Hiroyasu *et al*. [14] is used to better understand whether the mutation operator and the modified pheromone placement can efficiently strengthen the solution quality. The procedure of ACO+MO+MP is shown in figure 4(b). As shown in the figure, this algorithm has a mutation operator while TACO does not, and the pheromone placements are different as well.

/*Initialization*/
Input: ρ, Q, U, and M;
Generate U initial solutions;
Calculate objective function value of each solution to obtain $\boldsymbol{f}$ = $\{f_1, f_2, ..., f_U\}$;
Set ϑ be the current best solution and the corresponding f_ϑ;
Initialize pheromone trails;
/*Main loop*/
For $m = 2$ **to** M **do**
  For $k = 1$ **to** U **do**
    Determine R_t by equation (7);
    Determine E_q by equation (8);
  End for
  Obtain $\boldsymbol{f}$ by calculating objective function;
  Update pheromones by equations (8)-(11);
  Update ϑ and the corresponding f_ϑ;
End for
Return ϑ;

/*Initialization*/
Input: ρ, Q, ϕ, U, and M;
Generate U initial solutions;
Calculate objective function value of each solution to obtain $\boldsymbol{f}$ = $\{f_1, f_2, ..., f_U\}$;
Set ϑ be the current best solution and the corresponding f_ϑ;
Initialize pheromone trails;
/*Main loop*/
For $m = 2$ **to** M **do**
  For $k = 1$ **to** U **do**
    Determine R_t by equation (7);
    $r \leftarrow$ random [0, 1];
    If $r < \phi \exp\left[-M/(M-m+1)\right]$ **then**
      Perform the mutation operators;
    End if
    Determine E_q by equation (8);
  End for
  Obtain $\boldsymbol{f}$ by calculating objective function;
  Update pheromones by equations (10), (12), (13), and (14);
  Update ϑ and the corresponding f_ϑ;
End for
Return ϑ;

Figure 4. Pseudo-code for (a) the TACO; (b) the ACO+MO+MP

4. Numerical analysis

It is shown in this paper that the computational complexity theory of DLSPADC is an NP-hard problem. Thus, to lower the computational cost, this paper adopts the heuristic algorithm to solve the problems. The best algorithm for DLSPADC problems is found through the comparisons of TACO and ACO+MO+MP.

4.1 Parameter test

ACO+MO+MP differs from TACO in that it has a mutation operator and the modified pheromone trails placement. During the execution of the algorithm,

 ISBN: 978-84-614-6167-7

the mutation rate decreases as the iteration increases. Whether the different initial mutation rate is sensitive to the solution quality is also an issue. Further, trail persistence ρ is another important parameter. Therefore, the influence of the parameters on the solution quality is first discussed. In the experiment, the DLSPADC parameters are: $h_{2t} = 0.05$, $h_{1t} = 0.2$, $K_{1t} = 20$, $p_{2t} = 1$, $p_{1t} = 0.6$, and $s_{1t} = 0.8$, for $t = 1, 2, \ldots, 30$. $\alpha_2 = 0.2$ and $\alpha_1 = 0.1$. b_t stands for the discrete uniform distribution between 19 and 21, and d_t between 2 and 4. The two levels of parameter ρ are 0.6 and 0.85, and 0.1 and 0.4 for ϕ. The other parameters of the algorithm are: $Q = 1$, $U = 70$, and $M = 400$. The results of 30 runs of each level in table 1 show that parameters ρ and ϕ are not significant to the solution quality. This means in the policy of dynamic mutation rate, different starting mutation rates are not sensitive to the solution quality.

Table 1. Analysis of parameters ρ and ϕ

Factor	Level				*F*-ratio	*P*-value
ρ	0.6	0.6	0.85	0.85	0.214	0.645
ϕ	0.1	0.4	0.1	0.4	1.074	0.302
Avg.	925.682	924.851	925.236	924.679		
$\rho \times \phi$					0.042	0.838

4.2 Algorithm comparison

The DLSPADC parameters are: $h_{2t} = 0.05, h_{1t} = 0.2, K_{1t} = 100, p_{2t} = 1, p_{1t} = 0.6, s_{1t} = 0.8, \alpha_2 = 0.2, \alpha_1 = 0.1$, and $b_t = 20$. d_t is the discrete uniform distribution between 2 and 4. The algorithm parameters are: $\rho = 0.85, \phi = 0.4, Q = 1$, $U = 70$, and $M = 800$. Table 2 shows the test results of each algorithm with a higher fixed cost. The results are surprising in that ACO+MO+MP has the best performance.

Table 2. The test results of each algorithm with higher fixed cost

T		TACO	ACO+MO+MP
80	Min	3179.7460	3121.8736
	Avg.	3218.7845	3140.4370
	Max	3269.2726	3190.5560
100	Min	4089.7652	4083.3551
	Avg.	4162.4637	4142.5192
	Max	4231.2016	4204.4172

5. Conclusions

Advancing IT technology globalizes businesses and brings opportunities for multi-national cooperation. Thus, the policymaker stresses much more than ever the timely replenishment with proper amounts from the upstream businesses in the supply chains to the downstream businesses. Vender managed inventory is an important and commonly used strategy to shorten the lead time in the supply chain. This paper proposes a dynamic lot-sizing problem with asymmetric deteriorating commodity in the VMI supply chain. The problem is divided into a deterministic model. In the deterministic model, the properties and theorem of the problem are discussed. The DLSPADC performance of the two ACO-based algorithms is compared. In the comparison of TACO and ACO+MO+MP, ACO+MO+MP has the better performance. This paper

 ISBN: 978-84-614-6167-7

suggests two directions for future researches. First, discover more properties of DLSPADC to minimize the decision points or consider the DLSPADC from multiple manufacturers. Second, Lee *et al*. [17] proposed a two-stage dynamic lot-sizing problem with transport, but the deterioration rate of the commodity was not considered. Thus, future research may include the transport cost of the proposed problem.

6. References

[1] P.L. ABAD, *Optimal lot size for a perishable good under conditions of finite production and partial backordering and lost sale*, Computers and Industrial Engineering. 38 (2000) 457-465.

[2] Z.T. BALKHI, *The effects of learning on the optimal production lot size for deteriorating and partially backordered items with time varying demand and deterioration rates*, Applied Mathematical Modelling. 27 (2003) 763-779.

[3] G.R. BITRAN AND H.H. YANASSE, *Computational complexity of the capacitated lot size problem*, Management Science. 28 (1982) 1174-1186.

[4] M. DORIGO, *Optimization, learning and natural algorithms*, PhD thesis, Politecnico di Milano. Italy. 1992.

[5] Y. FRIEDMAN AND Y. HOCH, *A dynamic lot-size model with inventory deterioration*, INFOR. 16 (1978) 183-188.

[6] C. GENCER, S. EROL AND Y. EROL, *A decision network algorithm for multi-stage dynamic lot sizing problems*, International Journal of Production Economics. 62 (1999) 281-285.

[7] D. GUPTA AND Y. GERCHAK, *Joint product durability and lot sizing models*, European Journal of Operational Research. 84 (1995) 371-384.

[8] M.R. GAREY AND D.S. JOHNSON, *Computer and Intractability-A Guide to the Theory of NP-Completeness*, W. H. Freeman and Company. New York. 1979.

[9] J. GUTIÉRREZ, A. SEDEÑO-NODA, M. Colebrook and J. Sicilia, *A new characterization for the dynamic lot size problem with bounded inventory*, Computers and Operations Research. 30 (2002) 383-395.

[10] C. HYUN KIM AND Y. HONG, *An optimal production run length in deteriorating production processes*, International Journal of Production Economics. 58 (1999) 183-189.

[11] H. HWANG AND S.W. SHINN, *Retailer's pricing and lot sizing policy for exponentially deteriorating products under the condition of permissible delay in payments*, Computers & Operations Research. 24 (1997) 539-547.

[12] V.N. HSU, *Dynamic economic lot size model with perishable inventory*, Management Science. 46 (2000) 1159-1169.

[13] V.N. HSU, *An economic lot size model for perishable products with age-dependent inventory and backorder costs*, IIE Transactions. 35 (2003) 775-780.

[14] T. HIROYASU, M. MIKI, Y. ONO AND Y. MINAMI, *Ant colony for continuous functions*, The Science and Engineering. Doshisha University. Japan. 2000.

[15] W. JARUPHONGSA, S. CETINKAYA AND C.Y. LEE, *Warehouse space capacity and delivery time window considerations in dynamic lot-sizing for a*

 ISBN: 978-84-614-6167-7

simple supply chain, International Journal of Production Economics. 92 (2004) 169-180.

[16] A.K. JALAN AND K.S. CHAUDHURI, *Structural properties of an inventory system with deterioration and trended demand*, International Journal of Systems Science. 30 (1999) 627-633.

[17] C.Y. LEE, S. CETINKAYA AND W. JARUPHONGSA, *A dynamic model for inventory lot sizing and outbound shipment scheduling at a third-party warehouse*, Operations Research. 51 (2003) 735-747.

[18] R.B. MISRA, *Optimum production lot size model for a system with deteriorating inventory*, International Journal of Production Research. 13 (1975) 495-505.

[19] S. PAPACHRISTOS AND K. SKOURI, *An optimal replenishment policy for deteriorating items with time-varying demand and partial – exponential type – backlogging*, Operations Research Letters. 27 (2000) 175-184.

[20] B.R. SARKER, S. MUKHERJEE AND C.V. BALAN, *An order-level lot size inventory model with inventory-level dependent demand and deterioration*, International Journal of Production Economics. 48 (1997) 227-236.

[21] L.A. SMITH, *Simultaneous inventory and pricing decisions for perishable commodities with price fluctuation constraints*, INFOR. 13 (1975) 82-87.

[22] C.T. SU AND J.T. WONG, *Design of a replenishment system for a stochastic dynamic production/forecast lot-sizing problem under bullwhip effect*, Expert Systems with Applications. 34 (2008) 173-180.

[23] J.T. TENG, C.T. CHANG AND S.K. GOYAL, *Optimal pricing and ordering policy under permissible delay in payments*, International Journal of Production Economics. 97 (2005) 121-129.

[24] V. T'KINDT, N. MONMARCHÉ, F. TERCINET AND D. LAÜGT, *An ant colony optimization algorithm to solve a 2-machine bicriteria flowshop scheduling problem*, European Journal of Operational Research. 142 (2002) 250-257.

[25] H.M. WAGNER AND T.M. WHITIN, *Dynamic version of the economic lot size model*, Management Science. 5 (1958) 89-96.

[26] T.V. WARRIER AND N.H. SHAH, *A lot-size model with partial backlogging when the amount received is uncertain for deteriorating items*, International Journal of Systems Science. 30 (1999) 205-210.

[27] J.T. WONG AND K.H. CHEN, *A framework for solving stochastic lot-sizing problem by simulation and artificial intelligence*, The 2010 International Conference on Innovation and Management. 7- 10 July 2010, Penang, Malaysia.

[28] W.I. ZANGWILL, *A backlogging model and multi-echelon model of a dynamic economic lot size production system–a network approach*, Management Science. 15 (1969) 506-527.

 ISBN: 978-84-614-6167-7

Proceedings of the 11th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE2011
Benidorm, Spain, 26-30 June 2011

Analysis of natural and post-LASIK cornea deformation by 2D FEM simulation

A. Zarzo[1], P. Schäfer[1] and L. Casasús[1]

[1] *Departamento de Matemática Aplicada, E.T.S. Ingenieros Indutriales, Universidad Politécnica de Madrid, Spain.*

emails: alejandro.zarzo@upm.es, aphilipp_schaefer@aecor.de, lcasasus@etsii.upm.es

Abstract

A two dimensional finite element model (FEM for short) of the cornea is implemented. Various patterns of corneal geometry and decomposition of corneal tissue are simulated to obtain the resulting curvatures. In addition, LASIK incisions are calculated on models of normal and keratoconus affected corneas to observe the occurring deformations.

Key words: template, instructions
MSC2000: AMS Codes (optional)

1. Introduction

Good understanding of the mechanics of the human cornea is essential for the correction of refractive errors by means of surgery as well as for the recognition and treatment of diseases like e.g. keratoconus. In order to investigate the corneal behavior and to try to predict the outcome of operations and diseases researchers have created a great number of eye models in the recent two decades.

It is commonly accepted that the finite element method (FEM) is the best way to analyze the mechanics of the human eye. Simulations have been carried out not only for the cornea, but also for other parts of the eye such as the lens. For the sake of simplicity the first FEM-models of the cornea contained several assumptions and abstractions. Using two-dimensional models, just implementing a linear elastic material model and including only the cornea into the model can be an appropriate approximation for several purposes. However, the more realistic the model is the more authentic are the results of the analysis. Therefore, scientists went over to designing 3D models, employing hyperelastic and

 ISBN: 978-84-614-6167-7

anisotropic material behavior and modeling the cornea in conjunction with the limbus and the sclera.

The aims of the simulations have been to investigate amongst others the astigmatic change in shape of the cornea after making incisions which were representative for various types of operations such as e.g. cataract surgery [8] or arcuate keratotomy [16]. Also the resulting corneal curvature after LASIK surgery has been examined, for example by Cabrera Fernández et al. [6] and Argento, Cosentino, Darchuk and Elvira [2]. Other authors focused their work on the emergence of keratoconus. Pandolfi and Manganiello [18] as well as Carvalho et al. [7] modeled the formation of keratoconic ectasia by locally changing the mechanical properties of the corneal tissue. In both works anisotropic material behavior of the corneal tissue was implemented in 3D models of the cornea. Lanchares Sancho [15] additionally included the limbus and the sclera in her model. This listing of studies is not all-encompassing, but intends to give a short overview.

This paper aims to be a first step to drive forward the investigation of the keratoconus disease in its initial stage. Detecting subclinical keratoconus or keratoconus in early stages is of great interest because it allows us to start treatment of the disease in its early stages. Furthermore, detection of keratoconus is important and valuable information when it comes to consider a refractive surgery. In this case it is also interesting to predict what would happen to a cornea with initial keratoconus if a surgery would be performed. Although keratoconus in an advanced stage can be certainty recognized, it is still complicated to diagnose the disease in very early stages when patients do not yet suffer from any symptoms. Moreover, there are no clear threshold criteria to distinguish between normal eyes and eyes with subclinical keratoconus [4] being nowadays a relevant case study. Against this background, this work is a preliminary step which intends to improve the understanding of the biomechanical mechanisms of the process of formation of keratoconus and other cornea deformations.

On taking this step in this work a 2D-FEM model of the cornea is presented in order to investigate both the natural and the post-LASIK deformation of the cornea. For this, among other things, a study of different boundary conditions at the limbus is done to investigate how they affect to the mechanical behavior of the cornea. Furthermore, the influence of the initial form of the cornea on the outcome of the simulations is also considered, since there are distinct approximations of the corneal surface in literature.

 ISBN: 978-84-614-6167-7

2. Two dimensional FEM model of the cornea.

Material characteristics.

For the simulations of the cornea the finite element analysis software *Abaqus/CAE Student Edition 6.9-2* (Dassault Systèmes Simulia Corp.) was used in which an isotropic hyperelastic material model of the stroma was implemented. Although the true the stroma is anisotropic, the assumption of isotropy was made to simplify the modeling process in this initial step we present here. The implementation of the relastic material model was done by using the Odgen strain energy potential [9]. In order to enable Abaqus to calculate the coefficients test data had to be provided. For this the stress-strain function of Cabrera Fernández et al. [6] has been used. Furthermore, the assumption of almost incompressibility was made and several material characteristics have been used to simulate different stages of deterioration.

Geometry of the mocel.

The model of the cornea here implemented for the simulations is, of course a great simplification of the real human cornea. However, it should serve to have a first glance on the emergence of keratoconus and the impact of refractive surgery. Although these procedures in reality are three-dimensional, their fundamental mechanisms can be analyzed in 2D as well.

On the other hand, since the cornea is the element of interest in this work, only the cornea and limbus are included in the model. This implies that a set of numerical experiments has to be performed to choose the boundary conditions on the limbus in such manner that they have the same influence on the cornea as it would have the sclera (if the whole 3D eye is considered). This matter will be discussed in section 3.4.

Furthermore, it had to be considered which layers of the cornea should be integrated into the model. In an initial study a model containing the three biggest layers, which are epithelium, stroma and endothelium, was implemented. A linear elastic behavior with Young's modulus 0.02 MPa and Poisson's ratio 0.49 was assigned to epithelium and endothelium. However, as compared with a model which only comprised the stroma, there was nearly no change in the mechanical behavior. The difference in the elevation of the corneal apex when applying the intraocular pressure has been checked to be of the order of magnitude of 1 μm. This is equal to only 0.136% of the total displacement. Therefore, to perform the final simulations here we consider here models containing only the stroma. This approach is common in literature, too.

 ISBN: 978-84-614-6167-7

The physiological conditions of the cornea result from applying the intraocular pressure to its interior surface. The *in vivo* shape of the anterior corneal surface is described for example by Lanchares [15]. In order to obtain this shape after the application of the pressure in a first step, the initial stress-free *ex vivo* form (less "curved" than the real one) has to be chosen in such a way that the application of intraocular pressure leads to a realistic *in-vivo* form. By carrying out a complex iterative analysis process the *ex vivo* shape could be determined exactly. In this work the *ex vivo* form of the model which corresponds to an *in vivo* shape which is very similar to the one mentioned by Lanchares was obtained by means of a trial-and-error process. The boundary conditions applied hereby were those of constraining the limbus in displacement Figure 3.1 depicts the two states of the model.

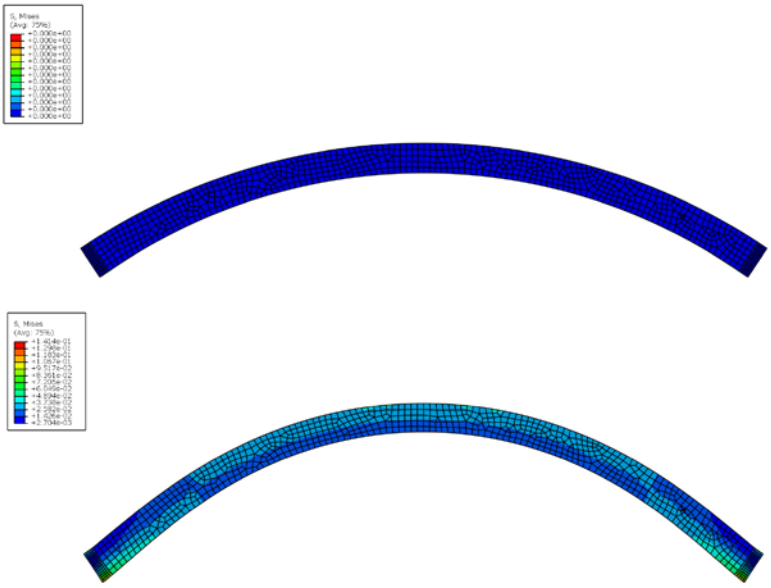


Figure 2.1: **top:** Model in the *ex vivo* state;
bottom: Model in the *in vivo* state

 ISBN: 978-84-614-6167-7

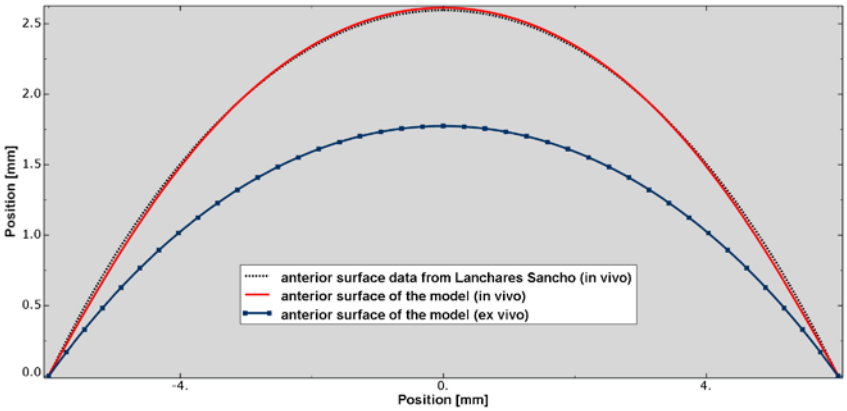


Figure 3.3: Cornea anterior surface of the model compared to the Lanchares model [15].

Figure 2.2 shows the *ex-vivo* form of anterior surface of the cornea in our model and a comparison between the *in-vivo* forms of Lanchares [15] and the one produced by our model, being apparent that both are almost indistinguishable.

Boundary conditions.

The finite element model contains the cornea and the limbus which are fastened together by a TIE-constraint. This type of constraint fuses together surfaces of two distinct regions of a model even though they have different meshes on the connected surfaces [10]. The sclera, however, is not included in the model. Hence, before applying the intraocular pressure, the model has to be fixed in a certain way by applying boundary conditions to the surface of the limbus. Ideally, the boundary conditions should have the same influence on the limbus and the cornea as it would have the sclera in a complete model of the eyeball.

The sclera is much stiffer than the cornea. The limbus forms the part between these two elements in which the stiffness increases. In order to account of this mechanical property five consecutive partitions with distinct stiffness were defined on the limbus as depicted in figure 2.3. In the absence of concrete data about the rigidity of the sclera the following linear elastic material behavior was chosen: Poisson's ratio is 0.49 for all the layers, while distinct values of Young's modulus were assigned to the layers. Each of these regions has a length of 0.040 mm and a width of 0.620 mm.

 ISBN: 978-84-614-6167-7

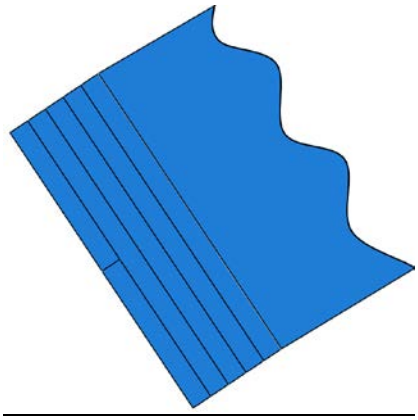


Figure 2.3: Picture detail of the model showing the partitions with distinct stiffness on the limbus: The value increases from the cornea to the sclera.

Up to now, many researchers have either decided to fix the border of their models (containing limbus *and* cornea or only the cornea) by constraining the displacement of the nodes on the boarder to zero. Or they avoided the problem of choosing adequate boundary conditions on the limbus by modeling the whole ocular globe. The latter way is clearly the better one, but of course much more complicated.

In this work several approaches regarding the boundary conditions on the limbus were realized, compared and implemented in the simulations. Firstly, the limbus was fixed by prohibiting displacement of the surface nodes as mentioned above. This is the simplest possible type of boundary conditions and it was used for the trial-and-error-process to adjust the pressure-free *ex vivo* shape of the model. The reason for choosing this type of boundary conditions for the adjustment is that authors like Lanchares [15] or Pandolfi and Manganiello [18] had already proven their suitability for this purpose.

As an alternative type of boundary conditions we also considered the possibility of giving the limbus movement more degrees of freedom in order to account for the fact that the sclera, although being much more rigid than the limbus and the cornea, does have certain flexibility and is also “inflated” by the intraocular pressure, too. That means that the sclera has also an *ex vivo* and an *in vivo* state as well, even though the difference between them might be very small. Thus, to connect the limbus to mechanical spring elements in the simulations to give it more degrees of freedom were implemented. Figure 2.4 illustrates this concept. Since this approach was completely new, some reasonable assumptions had to be made for the definition of the springs. For the implementation of the springs the instances of the limbus in the model were connected to the ground by connectors of the type *cartesian* with linear elastic properties.

 ISBN: 978-84-614-6167-7

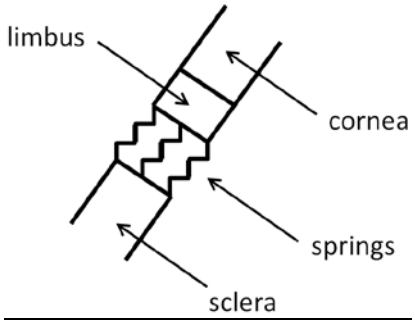


Figure 2.4: Connection of the limbus to the sclera modeled by spring elements

The connectors have two degrees of freedom which means that their ends connected to the limbus can move in all directions in the plane of the 2D simulation. Two uncoupled spring stiffness coefficients for different behavior in perpendicular directions were defined. For this the coefficient in the direction tangential to the stroma (and to the not modeled sclera) was considered to be 5 times bigger than the coefficient in the radial direction which is perpendicular to the stroma because the sclera mechanically is a thin shell whose in-plane stiffness is supposed to be significantly bigger than its outof-plane stiffness. Various models with 1, 3 and 5 springs on each instance of the limbus were created.

After defining the boundary conditions, one shoul apply ("load") the intraocular pressure for the simulations. Its value is assumed to be 15 mmHg, which is equal to 0.002 MPa. As the intraocular pressure normally ranges from 10 to 21 mmHg [15], this value is a good average and is commonly used for simulations in literature, too. Before carrying out the simulations of stroma degradation and LASIK, some initial analyses were realized to investigate the influences of variations of various parameters on the designed model.

• *Spring(s) vs. fixed limbus:* Connecting the limbus to springs instead of constraining all the nodes on its radial surface in displacement gives the limbus and the cornea the possibility to rotate. Thus, the *in vivo* shape of the cornea is a little bit more bellied at its sides and shows more displacement along the optical axis. This effect is most pronounced with only 1 spring on each instance of the limbus. The difference between 3 and 5 springs is very small. In both cases the *in vivo* shape is something in between the cases of a fixed limbus and 1 spring (see figure 2.5).

 ISBN: 978-84-614-6167-7

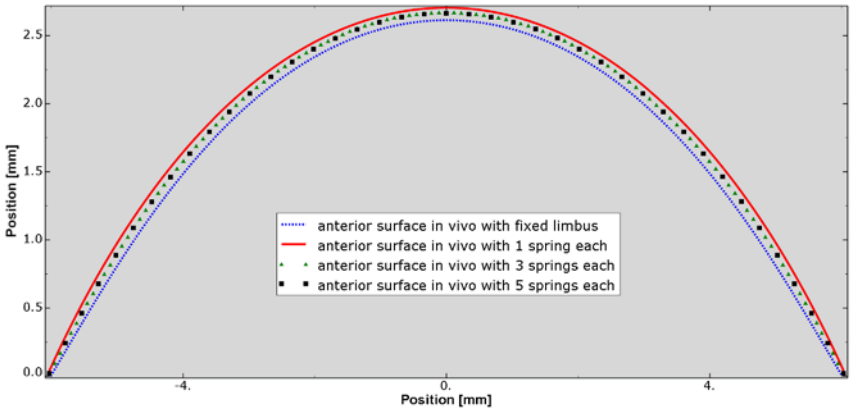


Figure 2.5: Influence of various boundary conditions on the *in vivo* state of the model (The effects on the posterior surface of the model are very similar.)

• *Spring coefficients:* Varying the radial coefficient(s) of the spring(s) has no influence on the *in vivo* state of the cornea. An alteration of the coefficient(s) for the tangential direction influences the displacement along the optical axis.

• *Value of the intraocular pressure:* For all types of boundary conditions on the limbus the resulting form of the *in vivo* cornea remains fairly similar. Even if the intraocular pressure is doubled to 30 mmHg there is no difference in this trend.

• *Epithelium and endothelium:* There is no significant difference in the corneal behavior if the epithelium and the endothelium are removed. Of course, the stroma is the most important element.
• *Ex vivo geometry:* As it could be expected, small changes in the initial form only lead to similar small changes of the deformed corneal shape, too. There are no "surprises". In order to analyze the order of magnitude of the influence of the above mentioned alterations of several parameters of the model a comparison of the values of the displacement of the corneal apex was performed. In the first part the thickness of the model was altered. In the second part the boundary conditions on the limbus were varied. The value of the thickness alteration was assumed to 5% because the thickness assumptions in literature vary in approximately this range. See for example [15] and [18]. What has bveen observed is that the effects of changes of the boundary conditions are significantly bigger than the effects of changes of the stromal-thickness. Therefore, the boundary conditions should be chosen carefully depending on the modeled situation.

3. Simulation of stroma degradation.

In this work the stroma degradation was modeled as a local loss of stiffness in certain regions of the stroma. For this, the position and the size of the affected region and additionally the grade of deterioration of the stromal tissue was altered

 ISBN: 978-84-614-6167-7

in various simulations to examine the dependency of the cornea deformations on these parameters.

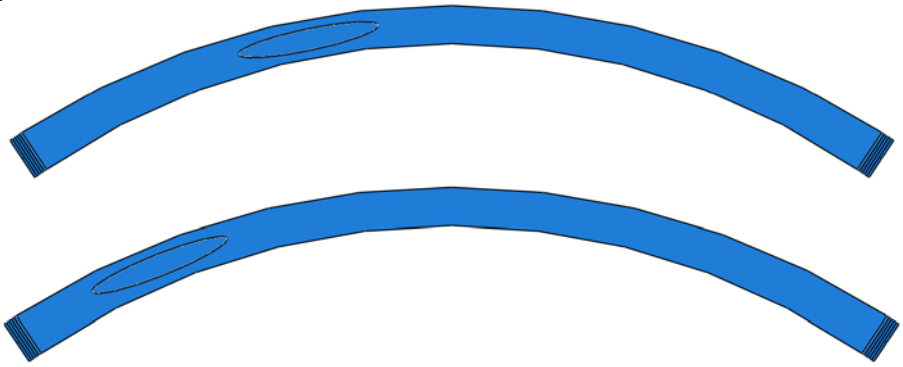


Figure 3.1: Different positions of the degradation regions on the model used in distinct simulations (size of the regions: medium)

The protrusion of keratoconus type deformations usually emerges outside of the corneal center. Therefore, two distinct positions of the degradation region on the side of the stroma were selected for the analyses. These positions are depicted in figure 3.1. As can be seen the degradation regions were modeled by elliptic partitions.

To account for various grades of deterioration of the affected material three different material types has been considered, corresponding to stiffness which are equal to 50%, 20% and 10% of the surrounding healthy stroma, respectively.
In order to simulate the formation of keratoconus like deformations in the most realistic way various steps were included in the numerical experiments. The aim was to establish the *in vivo* state of the healthy cornea model at first. Subsequently, the deterioration of the stroma in the affected region was realized, just like it happens in nature as well. To do this in *Abaqus* for every distinct simulation of stroma degradation an elliptic hole was created in the stroma corresponding to the dimensions of the desired region of deteriorated tissue. In addition, two new elliptic parts of the same dimensions were generated. The material properties of the healthy tissue were assigned to one of them, the other part got the properties of the deteriorated tissue.

One instance of both parts was inserted into the assembly of the model at the position of the hole in the stroma so that they were lying one on top of the other. Of course, during the simulation never both of these two instances were activated at the same time, but it was switched between them. Each of the two instances was tied to the stroma by using a TIE-constraint. To define the sequence of the simulation properly actually four steps were needed:

• *Initial Step:* The boundary conditions on the limbus instances are defined.

 ISBN: 978-84-614-6167-7

• *Step 1:* The elliptic part with the deteriorated material is excluded from the analysis and then the intraocular pressure is applied. The result of this step is the *in vivo* state of the healthy cornea.

• *Step 2:* The elliptic healthy part is removed, too.

• *Step 3:* The elliptic deteriorated part is added again. This gives the final result of the analysis.

The result of this simulation procedure for a model with a big degradation region which has a 10% deterioration is shown in figure 3.2. It can be seen that a slight protrusion of both the anterior and the posterior surface of the stroma is formed under the influence of the intraocular pressure.

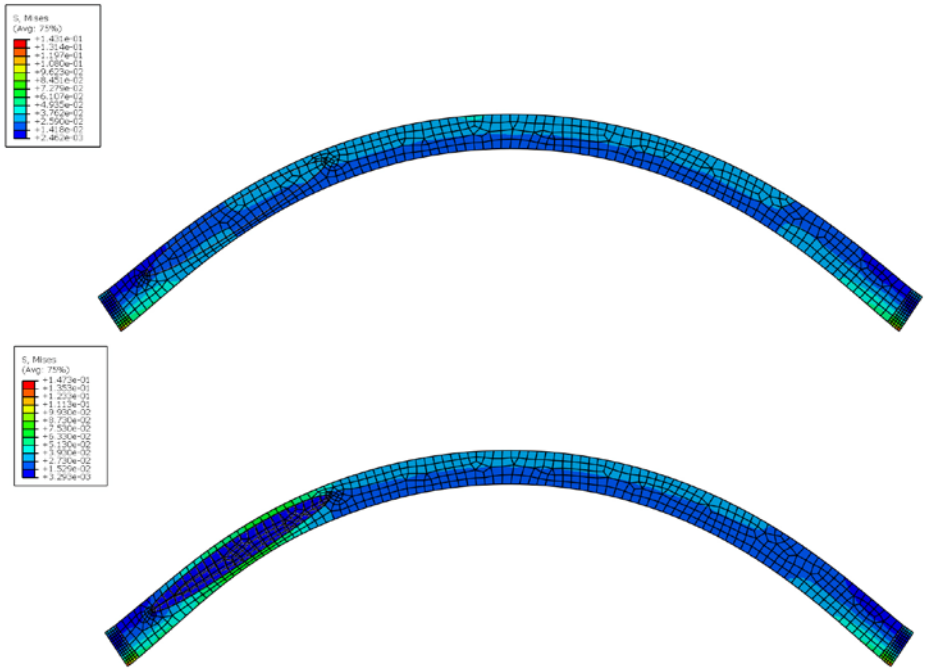


Figure 3.2: Simulation of a big degradation region with a 10% deterioration;
top: Healthy cornea in the *in vivo* state;
bottom: Deteriorated cornea in the *in vivo* state (color: von Mises stress)

An alternative to this complex procedure is to define just one single degradation partition on the stroma right at the beginning and to apply the intraocular pressure directly to the final assembly. This is *not* what happens in patients wiht keratoconus type deformations, however, it leads to a very similar result of the analysis. It turned out that the difference in displacement between the results of the two distinct ways of simulating is only of the order of magnitude of 10^–5 mm even in the worst case of a big degradation region and a deterioration. Thus, it can be concluded that it is not worthy to implement the more complex simulation procedure.

 ISBN: 978-84-614-6167-7

4. Simulation of lASIK surgery.

For the investigation of the effects of a LASIK surgery on the cornea various simulations of surgical correction of myopia were carried out. Tissue has to be ablated in the center of the cornea for this purpose. Several values of the ablation depth, *t0*, has been considered, corresponding to levels of myopia of −2 D, −4 D, −6 D, −8 D and −10 D, respectively. The diameter of ablation and the refractive index were assumed to be 6 mm and 1.377. In the simulations actually not the operation itself (creating the flap, ablating the tissue and repositioning the flap) but the changes on the corneal tissue which are caused by the LASIK operation were modeled. In other words, the transformation of the cornea froma pre-surgery state into the post-surgery state was simulated by realizing a change in its geometry (thickness). For this it was assumed that the ablation of a layer with the thickness *t*0, which in reality is carried out on the tissue under the flap, would result in a post-surgery cornea whose total thickness has decreased by *t*0. This assumption consequently indicates the form of the residual anterior corneal surface. For a realistic simulation of LASIK surgery with *Abaqus* the following proceeding was carried out. On the basic model of the healthy *ex vivo* cornea an additional partition was defined (see figure 4.1). This partition had the same material characteristics as the rest of the cornea and represented the material which had to be removed later on during the surgery. Similar to the course of action in the stroma degradation simulations, several steps of simulation were defined for the LASIK surgery as well. This served to initially establish the *in vivo* state of the complete cornea before realizing the ablation under physiological conditions. After removing part of its structure the cornea deformed again and ended up in its final post-LASIK state.

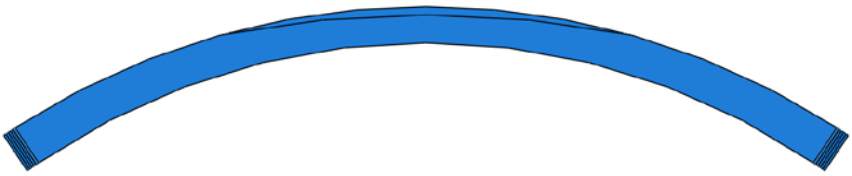


Figure 3.12: Partition for the ablation of corneal tissue in a simulation of LASIK surgery (refractive correction: 10 D)

The distinct steps of the simulation are:
• *Initial Step:* The boundary conditions on the limbus instances are defined.

• *Step 1:* The intraocular pressure is applied to the entire model. The result of this step is the *in vivo* state of the cornea.

• *Step 2:* The predefined partition representing the tissue to be ablated is excluded from the simulation. This causes further deformation and leads to the final result.

 ISBN: 978-84-614-6167-7

As described above, the partition which determined the ablation area and thus also the residual new surface of the cornea was defined already on the un-deformed *ex vivo* shape. There is no other way to do this in *Abaqus*. It was decided to use an ellipse for this purpose whose parameters were adapted to the depth of ablation $t0$ in each distinct case. However, in order to define the ellipse properly several restrictions were needed. Firstly, two points which had to be included in the ellipse were already given: the central apex point, which is positioned in the distance $t0$ under the former corneal apex, and the intersection point of the former anterior corneal surface with the 3 mm radius of the ablation region (S = 6 mm). Secondly, mirror symmetry of the partition with respect to the vertical axis was demanded. Yet, one more condition was still missing. Therefore, two different approaches were made to overcome this problem. One approach was to choose arbitrarily just any ellipse which fulfilled the given conditions. The other approach was to demand that the transition from the ablation region to the rest of the anterior corneal surface should be smooth, that means that the first derivatives of the curves should be equal at the intersection point. By means of this additional restriction the corresponding ellipse was perfectly defined, its semi-axes could be calculated and the partition line could be constructed in *Abaqus*.

After carrying out the first LASIK simulations it was concentrated on the correction level of 10 D using the two distinct forms of the ablation because it became apparent that the disorder induced into the stromal structure by a LASIK surgery had to be relatively big to provoke the emergence of keratoconus like deformations, if this should be possible to do with the model at all.

The already mentioned distinct possibilities to design the simulation process are generally valid for the simulations of LASIK, too. That is, the difference between carrying out all the three steps of the analysis mentioned above in order to realize the ablation on the *in vivo* cornea and, on the other hand, realizing the ablation already on the *ex vivo* cornea before applying the intraocular pressure is negligible. Therefore, it can be recommended for further investigations to follow the simpler way of designing the simulation procedure by cutting the model already in its *ex vivo* state before applying the intraocular pressure.

Acknowledgements.

AZ agknowledges partial fiancial support from Ministerio de Educación y Ciencia of Spain under grants MTM2008-06689 and MTM2009-14668-C02-02 and from Consejería de Innovación, Ciencia y Empresa de la Junta de Andalucía, Spain. Also, AZ has been partially funded by UPM under some contracts. The three authors specially acknowledge financial support from Departamento de Matemática Aplicada, ETSI Industriales de la UPM.

5. References

[1] K. Anderson, A. El-Sheikh, and T. Newson. Application of structural analysis to the mechanical behaviour of the cornea. *J. R. Soc. Interface*, 1:3–15, 2004.

[2] C. Argento, M. J. Cosentino, V. Darchuk, and G. Elvira. Sensibilidad estructural de la cornea ante diferentes configuraciones de corrección óptica por medio de ablacionado láser (LASIK). *4_ Jornadas de Desarrollo e Innovación*, November 2002.

[3] Augenärzte Dr. Jaschke & Kollegen. Crosslinking. Website. Accessed May 7, 2011. http://www.jaksche-kollegen.de/showpage.php?LEISTUNGSSPEKTRUM/Operationen/8226_UV_Crosslinking_bei_Keratokonus&SiteID=41.

[4] J. Bühren, D. Kook, G. Yoon, and T. Kohnen. Detection of subclinical keratoconus by using corneal anterior and posterior surface aberrations and thickness spatial profiles. *Investigative Ophtahalmology & Visual Science*, Vol. 51, No. 7, July 2010.

[5] I. N. Bronštejn, editor. *Handbook of mathematics*. Knovel library. Springer, Berlin, 4. ed. edition, 2004.

[6] D. Cabrera Fernández, A. M. Niazy, R. M. Kurtz, G. P. Djotyan, and T. Juhasz. Fnite element analysis applied to cornea reshaping. *Journal of Biomedical Optics*, vol. 10(6), November/December 2005.

[7] L. A. Carvalho, M. Prado, R. H. Cunha, A. C. Neto, A. P. Jr., P. Schor, and W. Chamon. Keratoconus prediction using a finite element model of the cornea with local biomechanical properties. *Arq Bras Oftalmol.*, 72(2):139–145, 2009.

[8] J. R. Crouch, J. C. Merriam, and E. R. Crouch. Finite element model of cornea deformation. *MICCAI 2005, LNCS*, 3750:591–598, 2005.

[9] Dassault Systèmes. Abaqus theory manual. Abaqus 6.9 Online Documentation, 2009.

[10] Dassault Systèmes. Abaqus user's manual. Abaqus 6.9 Online Documentation, 2009.

[11] Eye Clinic, P.C. LASIK. Website. Accessed May 1, 2011. http://www.eyeclinicpc.com/lasik/lasik.htm.

 ISBN: 978-84-614-6167-7

[12] N. Garijo Millán. Simulación de la queratoplastia y del proceso de cicatrización de heridas en la córnea. Master's thesis, Universidad de Zaragoza, 2010.

[13] G. A. Holzapfel, T. C. Gasser, and R. W. Odgen. A new constitutive framework for arterial wall mechanics and a comparative study of material models. *Journal of Elasticity*, 61:1–48, 2000.

[14] D. Köhler. Keratokonus. Website. Accessed April 22, 2011. http://www.prosehen. de/html/keratokonus.html.

[15] E. Lanchares Sancho. *Modelado biomecánico de los componentes refractivos del ojo humano y tratamientos refractivos asociados.* PhD thesis, Universidad de Zaragoza, 2010.

[16] R. Navarro, F. Palos, E. Lanchares, B. Calvo, and J. A. Cristóbal. Lower- and higherorder aberrations predicted by an optomechanical model of arcuate keratotomy for astigmatism. *Journal of Cataract & Refractive Surgery*, 35(1):158–165, 2009.

[17] OFTALVIST clínicas oftalmológicas. Lasik: láser in situ queratomileusis. Website. Accessed May 1, 2011. http://www.oftalvist.es/lasik-laser-in-situ-queratomileusisp43. aspx.

[18] A. Pandolfi and F. Manganiello. A model for the human cornea: constitutive formulation and numerical analysis. *Biomechanics and Modeling in Mechanobiology*, 5:237–246, 2006.

[19] PYCOMALL. eye eyeball iris pupil :: 3D Models. Website. Accessed May 10, 2011. http://www.pycomall.com/product.php?productid=16280.

[20] P. Rodríguez Ausín and A. Villarrubia Cuadrado. Queratocono forma «fruste». *Studium Ophthalmologicum*, 25(1):27–35, 2007.

[21] A. Saad and D. Gatinel. Topographic and tomographic properties of forme fruste keratoconus corneas. *Investigative Ophthalmology & Visual Science*, Vol. 51, No. 11:5546–5555, 2010.

[22] Science Photo Library. Side view of a woman's healthy green/grey eye. Website. Accessed April 23, 2011. http://www.sciencephoto.com/images/download_lo_res.html?id=804200421.

[23] South Bay Ophthalmology. Refractive Surgery & Laser Vision Correction. Website. Accessed May 7, 2011. http://www.southbayophthalmology.com/laser-visioncorrection/ lasik-laser-assisted-in-situ-keratomileusis.

 ISBN: 978-84-614-6167-7

Proceedings of the 11th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2011
26–30 June 2011.

The MWF Method for Kinetic Models: An Overview and Research Perspective

Carlo Bianca[1], Marzio Pennisi[2] and Santo Motta[2]

[1] *Dipartimento di Matematica, Politecnico di Torino*

[2] *Dipartimento di Matematica & Informatica, Università di Catania*

emails: carlo.bianca@polito.it, mpennisi@dmi.unict.it, motta@dmi.unict.it

Abstract

Many physical or biological phenomena deal with the dynamics of interacting particles (of inert matter or living being). These classes of phenomena are well described in physics using a kinetic approach based on Boltzmann equation and in biology with a generalized kinetic theory (kinetic theory for active particles). In general, the analytical solutions of the related models are missing thus become extremely relevant the development of numerical approaches. The particle method are a class of numerical methods used to find a numerical solution of Boltzmann equations. The MWF-method for kinetic equations was firstly proposed by S. Motta and J. Wick in 1992 and recently generalized for the equations system case.

The aim of this talk is to overview the method and its applications in biology, physics and astronomy.

Key words: Numerical Methods, Particles, Linear Collision Kernel
MSC 2000: 65Z05, 65M50, 65M75

1 Introduction

Both analytical and numerical methods are frequently proposed to approximate solutions of the Boltzmann equation, and more in general kinetic equations [1, 2]. In the last three decades, particle methods represented a class of numerical methods widely used for Vlasov equation [3, 4]. Deterministic particle methods use particles as quadrature nodes for computing an approximate solution of the collision integral. In classical deterministic particles methods, particles are kept fixed in the velocity space and the evolution is reflected in changing their weights in time.
A new approach was presented by Motta and Wick in the MWF-method [5] and a new formulation, oriented to implementation purpose, was presented by Motta [6, 7]. The basic idea of the method consists in rewriting the collision kernel as divergence of a flux

 ISBN: 978-84-614-6167-7

and formally transform the kinetic equation with collision into a system of a collisionless kinetic equation (Vlasov equation) and the divergence equation for the flux with appropriate boundary conditions. At each time step the flux is computed at a finite set of particle points which are the quadrature points. Then the collision induced *velocity vector* is computed and added to the Vlasov equation which is solved numerically with an upwind scheme. For this reason the method is referred as the MWF-method (**M**otta-**W**ick **F**lux method). When the density function is approximated with a finite set of points with equal weight (moving particles) using Dirac-δ functions, and these points are chosen as quadrature nodes, then the Vlasov equation can be used to compute the particle motion and, consequently, the evolution of the density distribution. In this form, the method belongs to the class of the meshfree methods [8], which have been widely used mostly in fluid-dynamics and solid mechanics. The method was tested in different simple scenarios: for semiconductor kernels in 2D and 3D [9, 10]. Comparison with other particle methods was presented by Wick [11].
The MWF-method was recently extented to kinetic equation system [12] to provide a numerical tool for scenarios, like those occurring in biology, were many equations may be required [13]. A formal convergence proof of the method for the homogeneous one-dimensional case is published in [14].

2 The Method

Let $x = (x_1, x_2, x_3) \in \Omega_x \subset \mathbb{R}^3$ be a typical point in space, $v = (v_1, v_2, x_3) \in \Omega_v \subset \mathbb{R}^3$ a typical point in velocity, $t \geq 0$ a typical time and $\Omega = \Omega_x \times \Omega_v$. The function space $\mathcal{M}(\Omega)$ defined as

$$\mathcal{M}(\Omega) = \left\{ f(x,v,t) : \Omega \times [0,\infty) \to \mathbb{R}^+, \int_\Omega f(x,v,t)\, dx\, dv = 1 \right\}, \tag{1}$$

and equipped with the Chebichev norm

$$\|f\|_\infty = \sup_{(x,v)\in\Omega} |f(x,v,t)|,$$

is a normed space. Assume $f \in \mathcal{M}(\Omega)$ is a solution of the following first-order, semilinear partial differential kinetic equation

$$\partial_t f + div_{(x,v)}(f\,(u,F)) = Q(f), \tag{2}$$

where

$$u(v) : \Omega_v \to \mathbb{R}, \quad F(x,t) : \Omega_x \times [0,\infty] \to \mathbb{R},$$

and the inhomogeneity $Q(f) : \mathcal{M}(\Omega) \to L^1(\Omega_v)$ is a collisional operator which describes short range interactions and satisfies the conservation hypothesis

$$\int_{\Omega_v} Q(f)\, dv = 0. \tag{3}$$

 ISBN: 978-84-614-6167-7

We associate to the equation (2) an initial condition $f_0(x,v) \in \mathcal{M}(\Omega)$ such that

$$f(x,v,0) = f_0(x,v).$$

The latter choice and the conservation property (3) guarantees that the solution f belongs to the space $\mathcal{M}(\Omega)$ for all times. Then $f(\cdot,t)$ can be interpreted as the density of the probability measure $\mu(t)$.
The MWF method consists in rewriting the equation (2) in a conservation law in divergence form redefining the collisions as a flux. To do that one rewrites the collision term $Q(f)$ as divergence of a flux $\psi(v,t) = (\psi_1(v,t), \psi_2(v,t), \psi_3(v,t)) : \Omega_v \times [0,\infty) \to \mathbb{R}^3$

$$div_v \psi = -Q(f), \tag{4}$$
$$\psi \cdot \mathbf{n}_{\Omega_v} = 0, \quad v \in \Omega_v. \tag{5}$$

with $\psi_i(v,t) : \Omega_v \times [0,\infty) \to \mathbb{R}$, $i = 1,2,3$, and formally transform the problem in a collisionless one. Moreover the boundary condition guarantees the conservative property of the system.

The definition of ψ is not unique. Indeed if ψ satisfies (4), for every vector field χ, the vector field

$$\varphi = \psi + \nabla \times \chi,$$

will satisfy (4) as well. This is a trivial consequence of the fact that $\nabla \cdot (\nabla \times \chi) = 0$. Nevertheless, the results obtained using the MWF method do not depend on the choice of a particular gauge in definition (4). The associated "velocity vector" g is given according to

$$\int_B f\, g\, dv = \int_B \psi\, dv, \tag{6}$$

for all Borel sets $B \subset \Omega_v$. In particular, if we consider

$$I^i_{[\alpha_i,\beta_i]} = \{v \in \Omega_v : \alpha_i \le v_i \le \beta_i\}, \quad i = 1,2,3,$$

we obtain

$$\int_{I^i_{[\alpha_i,\beta_i]}} \psi_i\, dv = -\int_{I^i_{[\alpha_i,\beta_i]}} \Big(\int_{\alpha_i}^{v_i} Q(f)\, dv_i' \Big) dv.$$

The left hand side of (6) can be evaluated by using a suitable integration formula in each interval $I^i_{[\alpha_i,\beta_i]}$. Taking the above considerations into account, the equation defined in (2) thus reads

$$\partial_t f + div_{(x,v)}(f\,(u,F)) + div_v \psi = 0, \tag{7}$$

or

$$\partial_t f + div_{(x,v)}(f\,(u,F+g)) = 0. \tag{8}$$

The last equation is formally identical to a Vaslov equation. Since the element of g is added to the given vector field, only the computation of g is needed.

 ISBN: 978-84-614-6167-7

3 The Algorithm

The MW-method can be summarized as follows:

1. Construct the initial distribution;
2. Identify the particles belonging to each spatial subdomain;
3. For each spatial subdomain:

   a) compute the quantity

   $$\Psi_i = -\int_{I^i_{[\alpha_i,\beta_i]}} \Big(\int_{\alpha_i}^{v_i} Q(f)\, dv_i' \Big)\, dv;$$

   b) compute the induced force field g on the particles;
4. Compute the force F acting on the particles;
5. Compute the new particles positions in the phase space with

   $$\mathbf{x}^{n+1} = \mathbf{x}^n + \mathbf{v}^n \Delta t \tag{9}$$
   $$\mathbf{v}^{n+1} = \mathbf{v}^n + (F^n + g^n)\Delta t \tag{10}$$

   and, if time is less than t_{final}, go back to STEP 2, else stop.

It is worth precising that STEP 1 and 4 can be performed using different strategies and algorithm which are well note in the literature [15].

Acknowledgements

CB was partially supported by the FIRB project-RBID08PP3J-Metodi matematici e relativi strumenti per la modellizzazione e la simulazione della formazione di tumori, competizione con il sistema immunitario, e conseguenti suggerimenti terapeutici.

References

[1] C. Ringhofer, *Dissipative discretization methods for approximations to the Boltzmann equation*, Math. Models Methods Appl. Sci., **12** (2001) 133-148.

[2] C. Ringhofer, *An entropy-based finite difference method for the energy transport system*, Math. Models Methods Appl. Sci., **11** (2001) 769–795.

[3] H. Neunzert and J. Struckmeier, *Particle methods for Boltzmann equation*, Acta Numerica **4** (1995) 417-457.

[4] P.A. Raviart, An Analysis of Particle Methods, in *Numerical Methods in Fluid Dinamic* (edited by F. Brezzi), Springer Lecture Notes in Math., **1127** (1985) 243–324.

 ISBN: 978-84-614-6167-7

[5] S. Motta, J. Wick, *A new numerical methods for kinetic equations in several dimensions*, Computing, **46** (1991) 223–232.

[6] S. Motta, *A new formulation and gauge invariance of the MW-CRF method for kinetic equations*, Mathematical and Computer Modelling, **36** (2002) 403–410.

[7] S. Motta, *Energy conservation property of the MW-CRF deterministic particle method*, Applied Mathematics Letters, **16** (2003) 287–292.

[8] T. Belytschko, J. S. Chen, *Meshfree and particle methods*, John Wiley and Sons Ltd, (2007).

[9] C. Barone, S. Motta, *The CRF-Method for semiconductors' intravalley collision kernels: I - The 2D case*, Le Matematiche, **XLVII** (I) (1992) 163–175.

[10] C. Barone, S. Motta, *The CRF-method for semiconductors' intravalley collision kernels: II - The 3D case*, Le Matematiche, **XLVIII** (I) (1993) 109-122.

[11] J. Wick, *Numerical approaches to the kinetic semiconductor equation*, Computing, **52** (1994) 39–49.

[12] C. Bianca, F. Pappalardo, and S. Motta, *The MWF method for kinetic equations system*, Comp. & Math. Appl, **57** (2009) 831–840.

[13] C. Bianca, N. Bellomo, *Towards a Mathematical Theory of Complex Biological Systems*, World Scientific Publishing, (2011).

[14] C. Bianca, S. Motta, *The MWF method: a convergence theorem for homogenous one-dimensional case*, Comp. & Math. Appl, **58** (2009) 579–588.

[15] R. W. Hockney, J. W. Eastwood, *Computer Simulations Using Particles*, Adam Hilger, (1988).

 ISBN: 978-84-614-6167-7

Proceedings of the 11th International Conference on Computational and Mathematical Methods in Science and Engineering, CMMSE 2011 26–30 June 2011.

Numerical analysis of a mixed kinetic-diffusion surfactant model for the Henry isotherm

José R. Fernández[1], Maria del Carmen Muñiz[2] and Cristina Núñez[2]

[1] *Departamento de Matemática Aplicada I, University of Vigo*

[2] *Departamento de Matemática Aplicada, University of Santiago de Compostela*

emails: jose.fernandez@uvigo.es, mcarmen.muniz@usc.es, cristina.nunez.garcia@usc.es

Abstract

This paper deals with the numerical analysis of surfactants behavior at the air-water interface, taking into account the mixed kinetic-diffusion model evolving to the Henry isotherm. The existence and uniqueness of a weak solution is recalled. Then, fully discrete approximations are obtained by using a finite element method and the backward Euler scheme. Error estimates are stated from which, under adequate additional regularity conditions, the linear convergence of the algorithm is deduced. Finally, a numerical simulation is presented in order to demonstrate the behavior of the solution for a commercially available surfactant.

Key words: Mixed kinetic-diffusion model, surface tension, surfactant, finite element approximation, numerical simulations.

1 Introduction

The study of surfactant adsorption dynamics at the air-water interface has been revealed as a determinant issue for its application in areas such as biochemistry, medicine, agrochemistry, metallurgy, food processing and so on (see [1, 2, 8, 10, 13]).

The process consists in the incorporation of surfactant molecules to the new formed surface in a surfactant solution, reducing drastically its surface tension, and it is mathematically modeled by the partial differential equation of diffusion in one spatial dimension, considering either infinite or finite diffusion length, coupled with the corresponding adsorption model by means of a suitable boundary condition at the subsurface, the unknowns being both bulk and surface concentration. The work of Ward and Tordai (see [13]) pioneered a mathematical research concerned in achieving analytical solutions for the diffusion controlled model considering infinite diffusion length and then obtaining approximations for long and short times (see [7, 11]). Regarding the finite

 ISBN: 978-84-614-6167-7

diffusion length and the nonlinear isotherms, numerical methods have been used to approximate their solutions (see, for instance, [9]) but, to our knowledge, their numerical analysis is nowadays an open problem. In this paper, we deal with the numerical analysis of the diffusion problem with finite diffusion length for the linear mixed kinetic-diffusion model, evolving into the so-called Henry isotherm at equilibrium (the diffusion-controlled model for this isotherm has been recently studied in [4]).

The outline of this paper is as follows. In Section 2, we briefly describe the mathematical model and we introduce the variational formulation of the problem, for which an existence and uniqueness result is recalled. Fully discrete approximations are introduced in Section 3 by using a finite element method and the implicit Euler scheme for the spatial and time discretizations, respectively. An error estimate result is stated from which the linear convergence is deduced under suitable regularity assumptions. Finally, in Section 4 a numerical example is shown to demonstrate the behavior of the solution for a commercially available surfactant.

2 Statement of the problem. Mathematical model and variational formulation

In order to introduce the whole dynamic process, it is important to take into account the boundary called *subsurface* (see [1, 2]), which is located a few molecular diameters below the air-water interface and splits the domain where only diffusion takes place and the region in which only adsorption-desorption occurs.

Let us denote by x the distance from the interface and $c(x,t)$ the concentration of surfactant at point $x \in [0,l]$ and time $t \in [0,T]$. The boundary $x = 0$ of the spatial interval corresponds to the location of the subsurface. Denoting by $\Gamma(t)$ the time-dependent surface concentration and taking into account the Fick's law, we consider the diffusion partial differential equation:

$$\frac{\partial c}{\partial t}(x,t) - D\,\frac{\partial^2 c}{\partial x^2}(x,t) \;=\; 0, \quad x \in (0,l), \;\; t > 0, \tag{1}$$

together with the boundary conditions (see [1, 10]):

$$D\,\frac{\partial c}{\partial x}(0,t) = \frac{d\Gamma}{dt}(t), \quad t > 0, \tag{2}$$

$$c(l,t) = c_b, \quad t > 0, \tag{3}$$

and the initial conditions:

$$c(x,0) = c_0(x), \;\; x \in (0,l), \tag{4}$$

$$\Gamma(0) = \Gamma_0. \tag{5}$$

In equations (1)-(3), D is the diffusion coefficient and the positive constant c_b is the bulk concentration. Besides, $c_0(x)$ is a function defined in $[0,l]$ and being equal to c_b

 ISBN: 978-84-614-6167-7

on $x = l$. We remark that the surface concentration, Γ, actually becomes an unknown of the system and then an additional condition must be given in order to close the problem. Hereafter, we consider the simplest linear kinetic expression modeling the mass transfer between the surface and subsurface at low concentrations, which leads to the following ordinary differential equation (see [1, 12]):

$$\frac{d\,\Gamma}{dt}(t) = k_H^a\, c(0,t) - k_H^d\, \Gamma(t), \tag{6}$$

where k_H^a and k_H^d are the adsorption and desorption constants, respectively. At equilibrium or steady-state, $d\,\Gamma/dt = 0$ and, from equation (6), the classical Henry isotherm is recovered.

Moreover, assuming regularity, the previous ODE together with the initial condition (5) can be straightforwardly integrated and boundary condition (2) reads

$$D\,\frac{\partial c}{\partial x}(0,t) = k_H^a\, c(0,t) - \phi(t, c(0,\cdot)), \tag{7}$$

where

$$\phi(t,\zeta) = k_H^d\,\Gamma_0\, e^{-k_H^d\, t} + k_H^d\, k_H^a\, e^{-k_H^d\, t} \int_0^t e^{k_H^d \tau} \zeta(\tau) d\,\tau. \tag{8}$$

We are now concerned in analyzing problem (1), (3) and (4), together with the new boundary condition (7). Moreover, for the sake of clarity in the presentation, and in order to simplify the presentation of the next section, we assume that c_b equals zero and so a homogeneous boundary condition is imposed on the right end of the spatial interval.

Multiplying equation (1) by a smooth function z defined in $[0,l]$ such that $z(l) = 0$, integrating in $(0,l)$ and using the integration by parts formula and equation (7), we obtain, for a.e. $t \in [0,T]$,

$$\int_0^l \frac{\partial c}{\partial t}(x,t) z(x) dx + \int_0^l D\,\frac{\partial c}{\partial x}(x,t)\,\frac{\partial z}{\partial x}(x) dx + k_H^a\, c(0,t)\, z(0) = \phi(t, c(0,\cdot))\, z(0).$$

Let V be the Hilbert space

$$V = \{v \in H^1(0,l);\, v(l) = 0\},$$

endowed with the inner product

$$((v,w)) = \int_0^l \frac{\partial v}{\partial x}\,\frac{\partial w}{\partial x} dx,$$

and the associated norm $\|v\|_V = ((v,v))^{1/2}$. We denote by $\gamma_0 : H^1(0,l) \to \mathbb{R}$ the trace operator on $x = 0$. Furthermore, we recall the inner product in $H = L^2(0,l)$ given by

$$(v,w)_H = \int_0^l v(x)\; w(x) dx,$$

 ISBN: 978-84-614-6167-7

with associated norm $\|v\|_H = (v,v)_H^{1/2}$. Moreover, we consider the Hilbert space $\mathcal{V} = L^2(0,T;V)$ with dual space $\mathcal{V}' = L^2(0,T;V')$ together with

$$W_2(0,T;V) = \{v \in L^2(0,T;V);\ \dot{v} \in L^2(0,T;V')\},$$

where we denote the time derivative by a dot above. We now state the weak formulation of problem (1), (3), (4) and (7).

Problem P. For a given $c_0 \in H$, find a function $c \in W_2(0,T;V)$ such that

$$\langle \dot{c}(t), v\rangle_{V'\times V} + D\ ((c(t),v)) + k_H^a\, \gamma_0(c(t))\, \gamma_0(v) = \phi(t,\gamma_0(c))\, \gamma_0(v), \quad \text{for a.e. } t \in (0,T), \quad \forall v \in V, \tag{9}$$

$$c(0) = c_0. \tag{10}$$

The existence and the uniqueness of solution to Problem P is given in the next theorem. Its proof is based on classical results for linear parabolic equations and fixed-point techniques (see [5] for details).

Theorem 2.1 *Let k_H^a, k_H^d and D be positive constants. If $c_0 \in H$ then there exists a unique solution $c \in W_2(0,T;V)$ to Problem P.*

3 Fully discrete approximations: numerical analysis

In this section, we now consider a fully discrete approximation of problem (9)-(10), taking into account a finite-dimensional space $V^h \subset V$ to approximate the space V, obtained, for instance, by using a finite element method. Here, $h > 0$ denotes the spatial discretization parameter. Besides, we consider a partition of the time interval $[0,T]$, denoted by $0 = t_0 < t_1 < \cdots < t_N = T$. In this case, we use a uniform partition of the time interval $[0,T]$ with step size $k = T/N$ and nodes $t_n = n\,k$ for $n = 0,1,\ldots,N$. For a continuous function $z(t)$, we use the notation $z_n = z(t_n)$ and, for the sequence $\{z_n\}_{n=0}^N$, we denote by $\delta z_n = (z_n - z_{n-1})/k$ its corresponding divided differences.

Therefore, using the backward Euler scheme, the fully discrete approximations are considered as follows.

Problem $\mathbf{P}^{hk}$. Find $c^{hk} = \{c_n^{hk}\}_{n=0}^N \subset V^h$ such that

$$c_0^{hk} = c_0^h, \tag{11}$$

and, for $n = 1,\ldots,N$ and for all $v^h \in V^h$,

$$(\delta c_n^{hk}, v^h)_H + D\ ((c_n^{hk}, v^h)) + k_H^a\ \gamma_0(c_n^{hk})\, \gamma_0(v^h) = \phi_{n-1}^{hk}\, \gamma_0(v^h), \tag{12}$$

where $c_0^h \in V^h$ is an appropriate approximation of the initial condition c_0 and

$$\phi_{n-1}^{hk} = k_H^d\, \Gamma_0\, e^{-k_H^d\, t_n} + k_H^d\, k_H^a\, k \sum_{j=0}^{n-1} e^{k_H^d (t_j - t_n)} \gamma_0(c_j^{hk}). \tag{13}$$

 ISBN: 978-84-614-6167-7

Under the assumptions of Theorem 2.1 and using Lax-Milgram theorem, we easily deduce the existence of a unique discrete solution to Problem $\mathbf{P}^{hk}$.

In the sequel, we present some error estimates for the difference $c_n - c_n^{hk}$ assuming the following additional regularity:

$$c \in C([0,T];V) \cap C^1([0,T];H). \tag{14}$$

Applying a discrete version of Gronwall's inequality (see [6]), after some tedious algebraic manipulations (see [5] for details) we have the following result which states some a priori error estimates on the approximate solutions.

Theorem 3.1 *Under the assumptions of Theorem 2.1 and assuming that regularity condition (14) holds, there exists a positive constant $\beta > 0$, independent of the discretization parameters h and k, such that the following error estimates are satisfied for all $\{v_n^h\}_{n=1}^N \subset V^h$,*

$$\begin{aligned}\max_{0\le n\le N} \|c_n - c_n^{hk}\|_H^2 + k\sum_{j=0}^{N}\Big[D\|c_j - c_j^{hk}\|_V^2 + \alpha|\gamma_0(c_j - c_j^{hk})|^2\Big] \\ \le \beta\Big[\|c_0 - c_0^h\|^2 + \max_{1\le n\le N}\{\|\dot c_n - \delta c_n\|_H^2 + \|c_n - v_n^h\|_V^2 + I_n^2\} \\ + \sum_{j=1}^{N-1}\frac{1}{k}\|c_j - v_j^h - (c_{j+1} - v_{j+1}^h)\|_H^2\Big],\end{aligned} \tag{15}$$

where $\delta c_n = (c_n - c_{n-1})/k$ and the integration error I_n is given by

$$I_n = k_H^a k_H^d\, e^{-k_H^d t_n}\Big|\int_0^{t_n} e^{k_H^d \tau}\gamma_0(c(\tau))\,d\tau - \sum_{j=0}^{n-1} k\, e^{k_H^d t_j}\gamma_0(c(t_j))\Big|.$$

Estimates (15) are the basis for the convergence analysis. From now on and in order to approximate the space V, we consider the finite element space V^h defined in the following form:

$$\begin{aligned}V^h = \{v^h \in C([0,l])\,;\, v^h_{|[a_{i-1},a_i]} \in P_1([a_{i-1},a_i]),\ \text{for } i = 1,\dots,M,\\ v^h(l) = 0\},\end{aligned} \tag{16}$$

where the spatial discretization of the interval $[0,l]$ is given by $0 = a_0 < a_1 < \ldots < a_M = l$ and $h = l/M$. Moreover, $P_1([a_{i-1},a_i])$ denotes the set of polynomials of degree less or equal to one in the interval $[a_{i-1},a_i]$, $i = 1,\dots,M$, and let us assume further regularity conditions on the solution to the continuous problem:

$$c \in \mathcal{C}([0,T];H^2(0,l)), \quad \dot c \in L^2(0,T;V), \quad \ddot c \in \mathcal{C}([0,T];H). \tag{17}$$

Corollary 3.2 *Under the assumptions of Theorem 3.1 and the additional regularity conditions (17), the linear convergence of the algorithm is obtained; i.e. there exists a positive constant $\beta > 0$, independent of h and k, such that*

$$\max_{0\le n\le N}\|c_n - c_n^{hk}\|_H \le \beta\,(h+k).$$

 ISBN: 978-84-614-6167-7

4 Numerical results

In this section, we first describe the numerical scheme implemented in MATLAB in order to obtain the numerical approximations of Problem P^{hk} and then, we present some numerical results to exhibit its behavior in the simulation of a commercially available surfactant.

Considering the finite element space defined in (16), for $n = 1, 2, \ldots, N$ and given $c_{n-1}^{hk} \in V^h$, the discrete concentration at time $t = t_n$ of surfactant, c_n^{hk}, is then obtained from equation (12) solving the problem:

$$(c_n^{hk}, v^h)_H + D\ k\,((c_n^{hk}, v^h)) + k_H^a\, k\, \gamma_0(c_n^{hk})\,\gamma_0(v^h) \\ = (c_{n-1}^{hk}, v^h)_H + k\,\phi_{n-1}^{hk}\,\gamma_0(v^h), \quad \forall v^h \in V^h,$$

where value ϕ_{n-1}^{hk} is given in (13). This leads to a linear system which is solved by using classical Cholesky's method. This numerical scheme was implemented on a 3.2 Ghz PC using MATLAB, and a typical 1D run ($h = k = 0.01$) took about 0.6 seconds of CPU time.

4.1 Simulation of hexanol

As an example, we consider a dilute solution of the commercial alcohol hexanol, using the data from references [1] and [12], namely:

$$c_b = 3.44\,\mathrm{mol/m^3}, \quad D = 7.16 \times 10^{-10}\mathrm{m^2/s}, \quad k_H^a = 1.73 \times 10^{-4}\,\mathrm{m/s}, \\ k_H^d = 157\,\mathrm{s^{-1}}, \quad l = 10^{-6}\,\mathrm{m}, \quad T = 0.5\,\mathrm{s}, \quad \Gamma_0 = 0\,\mathrm{mol/m^2}.$$

Moreover, the initial condition c_0 is defined as $c_0(x) = c_b$ for all $x \in [0, 10^{-6}]$.

Using the discretization parameters $h = 10^{-8}$ and $k = 10^{-4}$, the concentration at final time and the evolution in time of the subsurface concentration are shown in Fig. 1.

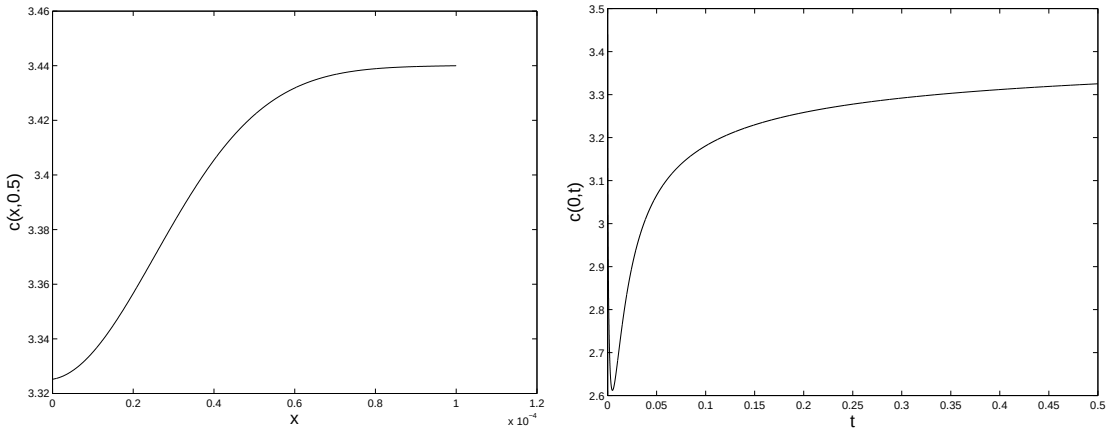


Figure 1: Concentration at final time (left) and evolution in time of the subsurface concentration.

 ISBN: 978-84-614-6167-7

Finally, the surface equation of state, relating the surface tension σ with the surface concentration Γ, is given by

$$\sigma(t) = \sigma_0 - n\,R\,T\,\Gamma(t),$$

where $\sigma_0 = 0.072\,\mathrm{N/m}$ denotes here the surface tension of pure water, $T = 293.71\,\mathrm{K}$ is the temperature, $R = 8.31\,\mathrm{J/(K\,mol)}$ represents the gas constant and n is a constant which is equal to one for a non-ionic surfactant. In Fig. 2 the evolution in time of the surface tension obtained with our algorithm is shown (left) and also the one obtained modeling the whole problem with Comsol Multiphysics (right), stating the agreement of both results. Finally, we remark that these numerical calculations are in good agreement with the experimental dynamic surface tensions of the hexanol solution reported in Fig. 6 of [12].

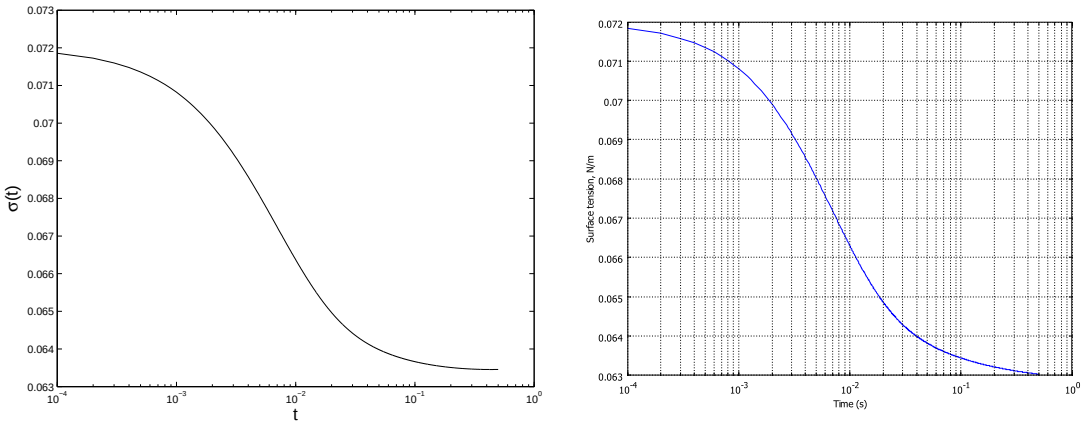


Figure 2: Comparison between the surface tension obtained with our algorithm (left) and COMSOL (right), semi-log scale.

References

[1] C. H. Chang and E. I. Franses, *Adsorption dynamics of surfactants at the air/water interface: a critical review of mathematical models, data and mechanisms*, Colloids and Surfaces **100** (1995) 1–45.

[2] J. Eastoe and J. S. Dalton, *Dynamic surface tension and adsorption mechanisms of surfactants at the air/water interface*, Adv. in Colloid Interface Sci. **85** (2000) 103–144.

[3] I. Egry, E. Ricci, R. Novakovic and S. Ozawa, *Surface tension of liquid metals and alloys Recent developments*, Adv. Colloid Interface Sci. **159** (2010) 198-212.

[4] J. R. Fernández and M. C. Muñiz, *Numerical analysis of surfactant dynamics at air-water interface using the Henry isotherm*, J. Math. Chem. (in press, DOI: 10.1007/s10910-011-9847-y).

 ISBN: 978-84-614-6167-7

[5] J. R. Fernández, M. C. Muñiz and C. Núñez, *A mixed kinetic-diffusion surfactant model for the Henry isotherm*, Preprint (2011).

[6] W. Han and M. Sofonea, *Quasistatic contact problems in viscoelasticity and viscoplasticity*, Stud. Adv. Math., vol. 30, American Mathematical Society - International Press, Providence, 2002.

[7] R. S. Hansen, *Diffusion and the kinetics of adsorption of aliphatic acids and alcohols at the water-air interface*, J. Colloid Sci. **16** (1961) 549–560.

[8] V.N. Kazakov, A.F. Vozianov, O.V. Sinyachenko, D.V. Trukhin, V.I. Kovalchuk and U. Pison, *Studies on the application of dynamic surface tensiometry of serum and cerebrospinal liquid for diagnostics and monitoring of treatment in patients who have rheumatic, neurological or oncological diseases*, Adv. Colloid Interface Sci. **86** (2000) 1–38.

[9] R. Miller, *On the solution of diffusion controlled adsorption kinetics for any adsorption isotherms*, Colloid & Polymer Sci. **259** (1981) 375–381.

[10] R. Miller, P. Joos and V. B. Fainerman, *Dynamic surface and interfacial tensions of surfactant and polymer solutions*, Adv. Colloid Interface Sci. **49** (1994) 249–302.

[11] K. L. Sutherland, *The kinetics of adsorption at liquid surfaces*, Australian J. Sci. Research, Series A: Physical Sciences **5** (1952) 683–696 (1952).

[12] C. Tsonopoulos, J. Newman and J. M. Prausnitz, *Rapid aging and dynamic surface tension of dilute aqueous solutions*, Chem. Engrg. Sci. **26** (1971) 817–827.

[13] A. F. H. Ward and L. Tordai, *Time-dependence of boundary tensions of solutions*, J. Chem. Physics **14**(7) (1946) 453–461.

 ISBN: 978-84-614-6167-7

Proceedings of the 11th International Conference on Computational and Mathematical Methods in Science and Engineering, CMMSE2011 Benidorm, Spain, 26-30 June 2011

QNANO: computational platform for electronic properties of semiconductor and graphene nanostructures

M.Korkusinski[1], M.Zielinski[1,3], E.Kadantsev[1,2], O.Voznyy[1], A.D.Guclu[1], P.Potasz[1,4], A. Trojnar[1,2] and P. Hawrylak[1,2]

[1] *Quantum Theory Group,Institute for Microstructural Sciences, National Research Council of Canada, Ottawa, Canada K1A0R6*
[2]*Department of Physics, University of Ottawa, Ottawa, Canada*
[3]*Institut of Physics, Copernicus University, Torun, Poland*
[4]*Institut of Physics, Wroclaw University of Technology, Wroclaw, Poland*

email: pawel.hawrylak@nrc-cnrc.gc.ca

Abstract

QNANO: computational platform enabling steps toward predictive calculations of electronic and optical properties of million atom semiconductor and graphene nanostructures is reviewed.

Key words: semiconductor nanostructures, graphene, computational nanoscience

1. Introduction

The size of semiconductor and carbon-based nanostructures such as self-assembled quantum dots, nanowires, nanocrystals and graphene nanostructures involving millions of atoms precludes calculation of their electronic properties using ab-initio methods, such as, e.g., GW-BSE approach. We discuss here one of the approximate methods, VFF-tb-CI, implemented in QNANO computational platform. QNANO combines six steps steps [1] (a) determination of all atomic constituents , (b) calculation of equilibrium position of atoms using valence force field model (VFF) (c) ab-initio calculation of electronic structure for strained bulk materials and determination of appropriate effective mass, k*p, or tight-binding Hamiltonian , (d) calculation of quasi-electron and quasi-hole states (equivalent to the GW step) using a linear combination of sp3d5s* atomic orbitals approach in a tight binding approximation (tb), (e) inclusion of the effect of final state interactions by defining an effective Hamiltonian of interacting excited quasi-particles, solved using the configuration interaction method (CI), and (f)

 ISBN: 978-84-614-6167-7

calculation of electronic and/or optical properties. This methodology has been applied to strained InAs/GaAs and InAs/InP quantum dots, CdTe quantum dots containing magnetic ions, CdSe and PbSe nanocrystals and graphene nanostructures.

2. Self-assembled quantum dots and nanocrystals

For strained quantum dots the problem is a multi-scale problem, with strain extending over micrometers and electronic states confined to nanometers. The VFF calculations have to be carried for up to 10^9 atoms using the Keating model with material parameters derived from bulk elastic constants c_{ij}. The tb parameters for unstrained InAs and GaAs are obtained by fitting of the tb bulk band edges and effective masses to those obtained in experiment or by ab-initio calculations, with the valence band offset (VBO) built into the parameter set. The dependence of band edges on lattice deformation computed using DFT [2] is used to find strain corrections to tb parameters. The Coulomb matrix elements for CI are obtained with tb wave functions involving ~10^8 orbitals, with onsite and nearest-neighbor terms computed by approximating the tb basis with Slater orbitals. The interactions are screened by a distance-dependent dielectric function and, typically ~10^4 configurations are used as a basis for multiexciton complexes.

The method is illustrated by computing the electronic and optical properties of a lens-shaped and disk-shaped InAs/GaAs[1],As/InP[2]self–assembled quantum dots and CdSe nanocrystals[3].

3. Graphene nanostructures

Using a combination of ab-initio and tb-HF-CI methods we determine electronic, magnetic and optical properties of gate controlled graphene quantum dots[4]. The dependence of the energy gap on shape, size and edge for graphene quantum dots with up to a million atoms is predicted. We show that triangular graphene quantum dots with zigzag edges combine magnetism with optical transitions simultaneously in the THz, visible and UV spectral ranges, determined by strong electron-electron and excitonic interactions. The relationship between optical properties and finite magnetic moment and charge density controlled by an external gate is discussed[4].

4. References

[1] M.ZIELINSKI , M. KORKUSINSKI, AND P. HAWRYLAK , PHYS. REV. B 81, 085301 (2010); M. KORKUSINSKI ET AL., J.APPL.PHYS. **105**, 122406 (2009);W. SHENG, ET AL., PHYS. REV. B71, 035316 (2005).

[2] EUGENE S. KADANTSEV, MICHAL ZIELINSKI, MAREK KORKUSINSKI, AND PAWEL HAWRYLAK , J. APPL. PHYS. **107**, 104315 (2010). EUGENE S. KADANTSEV AND PAWEL HAWRYLAK, APPL.PHYS.LETT. **98**, 023108 (2011).

[3] M.KORKUSINSKI, O. VOZNYY, P. HAWRYLAK,PHYS. REV. **B 82**, 245304 (2010).

[4] A.D.GUCLU, P.POTASZ, AND P. HAWRYLAK , PHYS. REV. **B 82**, 155445 (2010). A.D. GUCLU, P. POTASZ, O. VOZNYY, M. KORKUSINSKI, P. HAWRYLAK, PHYS.REV.LETTERS, **103**, 246805 (2009).

 ISBN: 978-84-614-6167-7

Proceedings of the 11th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE2011
Benidorm, Spain, 26-30 June 2011

Free Helical Gold Nanowires: A Density of States Analysis

Xiao-Jing Liu and I.P. Hamilton

Department of Chemistry, Wilfrid Laurier University, Waterloo,
N2L3C5, Ontario, Canada

emails: xxliu@wlu.ca, ihamilton@wlu.ca

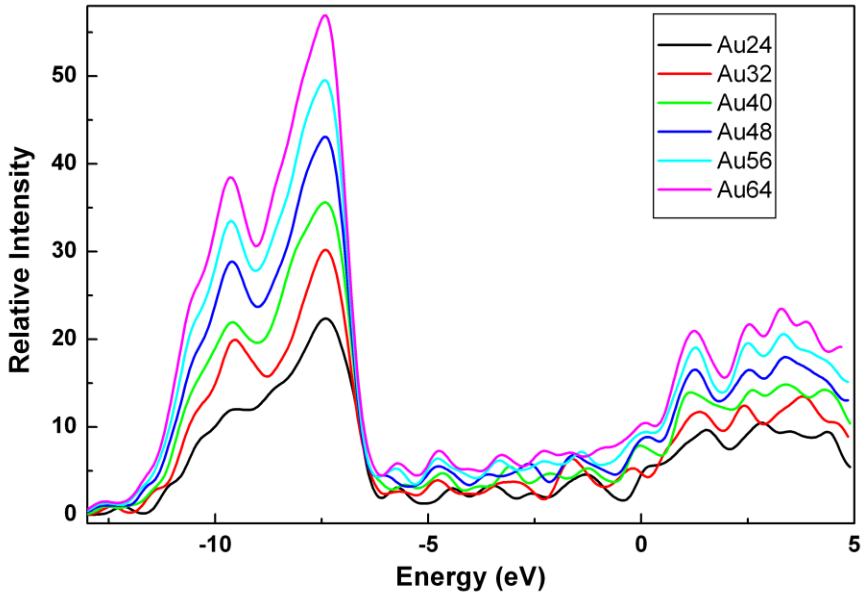
Nano-sized metal clusters represent a form of matter that has yet to be fully explored. They have been shown to exhibit size-related properties that differ significantly from both small clusters and the bulk. A close relationship exists between the properties of metal nanoclusters and their geometries but it is currently difficult to elucidate this connection by experimental techniques alone. In this regard, quantum chemical calculations can give detailed insight into the nature of these species. Density functional theory has become an increasingly important tool for these quantum chemical calculations since the effects of electron correlation, which are typically large for strongly correlated systems such as metals, can be included at a moderate computational cost and relativistic effects can be efficiently included via an effective core (or pseudo) potential. Relativistic effects are very significant for gold, resulting in a large contraction of the 6s orbital and a small 6s/5d energy gap, and gold exhibits a wide range of bonding quite distinct from any other metal. [1] Properties of bare gold clusters have been studied intensively by both experimental and theoretical methods and the global minimum structures of neutral gold clusters are now well established up to about Au_{20}. Larger clusters such as Au_{42}, Au_{55} and Au_{72} have also been examined and found to exhibit a wide variety of compact and cage-like structures.

Nanowires are nanoclusters of very large extension along one direction only and, as they are generally not equilibrium states, they typically require support. Supported gold nanowires are an essential component of interconnections in electronic nanodevices, which could find application in sensors, waveguides, photonics, and piezoelectronics. Single-strand gold nanowires have been observed and multi-shell helical gold nanowires were observed in the experimental study of Kondo and Takayanagi. [2] These helical gold nanowires consist of at least one

 ISBN: 978-84-614-6167-7

coaxial tube of gold atoms and there may be a central strand of gold atoms. The central strand of gold atoms has a linear structure while the coaxial tube of gold atoms has a helical structure, which can be pictured as a triangular gold sheet, which has been folded cylindrically onto itself. For the simplest of the helical gold nanowires there is one coaxial cylindrical tube with seven gold atoms for each atom of the central strand and this is therefore termed a 7-1 structure. We have recently shown that, for Au_{24}, Au_{32} and Au_{40} (which have 3, 4, and 5 gold atoms in the central strand) these structures are stable as free helical gold nanowires. [3] To better understand the nature of these species, we extend our relativistic density functional theory study to include Au_{48}, Au_{56} and Au_{64}, and we perform a detailed density of states analysis. We partition these species into fragments and then analyze the charge transfer and the electronic polarization among these fragments using the charge decomposition analysis scheme. We also analyze the fragment orbital contributions to the HOMO and LUMO to gain a better understanding of the reactivity of these free helical gold nanowires.

Density of states (relative intensity) as a function of the energy:



References

[1] P. PYYKKÖ, *Theoretical Chemistry of Gold III*, Chem. Soc. Rev. **37** (2008) 1967-1997.

[2] Y. KONDO AND K. TAKAYANAGI, *Synthesis and Characterization of Helical Multi-Shell Gold Nanowires*, Science **289** (2000) 606-608.

[3] X. LIU, I.P. HAMILTON, R.P. KRAWCZYK AND P. SCHWERDTFEGER. *Free Helical Gold Nanowires: A Relativistic Density Functional Study*, in preparation.

 ISBN: 978-84-614-6167-7

Proceedings of the 11th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2011
26–30 June 2011.

QM/MM simulations of protein immobilization on surfaces via metallic clusters

C.F. Sanz-Navarro[1], P. Ordejón[1] and R. E. Palmer[2]

[1] *Centro de Investigación en Nanociencia y Nanotecnología - CIN2 (CSIC-ICN), Campus UAB, E-08193 Bellaterra, Spain,*

[2] *Nanoscale Physics Research Laboratory, School of Physics and Astronomy, The University of Birmingham, Edgbaston, Birmingham B15 2TT, UK*

emails: csanz@cin2.es, ,

Abstract

We present a new QM/MM methodology [1] that combines the SIESTA approach to density-functional theory (DFT) with an AMBER force field. We apply our QM/MM methodology to the atomistic simulation of the immobilization of human oncostatin M (OSM) on a graphite surface. The stable immobilization of an individual protein, with controlled orientation, on a surface is a promising technique to build bioelectronic devices. However, the lack of a basic understanding about how proteins bind a surface at the atomistic level makes challenging the development of technological applications. Therefore, simulations are often very interesting to investigate the protein-surface binding. In our simulations, OSM is immobilized by binding it to a gold cluster previously pinned on the graphite substrate. The region close to the protein-cluster binding site is treated with DFT, while the rest is modelled with a classical force field. We address the nature of the interaction between the protein and the gold clusters pinned on the graphite support.

References

[1] C. F. Sanz-Navarro, R. Grima, A. García, E. A. Bea, A. Soba, J. M. Cela and P. Ordejón *Theor. Chem. Acc.* **128** 825 (2011).

 ISBN: 978-84-614-6167-7

Proceedings of the 11th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE2011
Benidorm, Spain, 26-30 June 2011

Astronomical causes of anomalous hot summers

Nikolay Sidorenkov

Hydrometcentre of Russia, Moscow, 123242, Russia; sidorenkov@mecom.ru

In the summer of 2010, an unprecedented heat wave of record duration and intensity was observed over European Russia. The heat wave took hundred thousands of lives, led to fires that destroyed hundreds of villages and about one million hectares of forest, and cost the Russian economy hundred millions of dollars.

It is believed that daily temperature anomalies vary randomly with time. However, the spectral analysis of a long-time series of daily air temperature anomalies revealed pronounced components with a lunar year period of 355 days, a lunar evection half-period of 206 days, a lunar quarter year period of 87 days, and a lunar sidereal month period of 27 days (See Figure 12.3 [1]).

The tides influence cloud amount. The amplitude and phase of lunar tides affect the cloud cover at an observation site. In clear skies, the atmosphere is heated by solar radiation during the long daytime in summer but is cooled due to escaping infrared radiation during the long nighttime in winter. As a result, in clear skies the daily mean air temperature (T) in summer exhibits positive anomalies, while its negative anomalies are observed in winter. In cloudy skies, the air temperature anomalies have opposite signs. In the spring and autumn T does not depend on the cloud amount because the day is approximately equal in length to the night. In this way, the interaction of the lunar gravitational effects with the atmospheric radiation conditions creates oscillations of T with lunar periods and amplitudes depending on the physical and geographical conditions of the site.

In Moscow the amplitude of the "solar" 365-day oscillations of temperature of T is about 15°, and the basic 355-day "lunar" oscillation of T is about 5°. The "solar" temperatures oscillations interfere with the "lunar" temperature oscillations to form beats of T. A beat is characterized by a periodic variation in the amplitude of the resulting oscillation. When the phases of the interfering oscillations coincide, their partial amplitudes add together so that the resulting amplitude of T becomes maximal (15°+5°=20°). As a result, hot summers and cold winters are observed during these periods. Then the phases of the oscillations gradually diverge, and the amplitude of the resulting oscillation of T decreases to become minimal at a phase difference of 180°, when the

 ISBN: 978-84-614-6167-7

amplitudes of the interfering oscillations are subtracted (15°–5°=10°). As a result, cool summers and warm winters are observed in these cases.

The beat frequency is equal to the half-difference between the frequencies of the interfering oscillations. The interference of 365-day and 355-day oscillations gives rise to a period beat of 35.2 years, which is known in climatology as the Brückner cycle.

The addition of the "solar" semiannual period of oscillations of air temperature and its 206-day "lunar" period generates beats of T with a period of 4.4 years. As a result, the 35-year cycle of the annual oscillation amplitude of T becomes hidden so that the extrema of T seem to vary randomly (there appear "clones" like the extrema in 1936 and 1938 and in 2002 and 2010).

The sequence of hot summer seasons over European Russia in 1901, 1936/1938, 1972, and 2002/2010 is associated primarily with the 35-year beats of air temperature. In 2010 their influence was supplemented with the effects of some eclipse cycles: a 19-year doubled Metonic cycle (analogous to that in 1972), an 8-year octaeteris subcycle (peat and forest fires also occurred over European Russia in August and September, 2002), a 29-year inex cycle (the 1981 summer was hot and dry), and other less significant lunar cycles.

Clearly, the 2010 heat wave over European Russia resulted from beats of not only temperature but also all the other hydrometeorological characteristics, i.e., pressure, wind, humidity, etc. These conditions correspond to nearly stationary blocking highs persisting for a long time. What are the forces that cause them to persist? There are strong reasons to believe that these are the anomalous gravitational forces produced by slow variations in the relative positions of the Moon, Earth, and Sun, by the rotation of their major axes (apses), by the motion of the nodes of their orbits, and by variations in their orbital parameters [1,2]. The variations in the mutual configurations in the Earth–Moon–Sun system generate gravitational perturbations that slowly propagate in near-Earth space and induce atmospheric baric waves (blocking anticyclones and depressions), which move over the Earth's surface together with the gravitational perturbations. Blocking highs cause anomalous frosts in winter and anomalous heat waves in summer.

Heat waves over European Russia were observed during the summers of 1972, 2002, and 2010. One year earlier (in 1971, 2001, and 2009, respectively) a summer heat wave occurred in Western Siberia, and a similar phenomenon was observed in Western Europe during the summers of 1973 and 2003. These facts suggest that the summer locations of the centers of nearly stationary blocking highs move from east to west at a velocity of about 40° per year. Therefore, a heat wave can be expected in Western Europe during the summer of 2011. The anomalously cold December of 2010 in Western Europe agrees with this prediction, since a cold winter in the air temperature beats is followed by a hot summer. Over European Russia, there will be a depression with cool and wet weather in the summer of 2011.

 ISBN: 978-84-614-6167-7

References

1. Sidorenkov, N.S. (2009), The interac tion between Earth's rotation and geophysical processes. W ILEY-VCH Verlag Gm bH & Co. KGaA, Weinheim, 2009. 317 pp.
2. Wilson I., Sidorenkov N. Catastrophic h eat-waves in Southern Australia and extremes in the Solar/Lunar tides - are they linked? Geophysical Research Abstracts. 12, EGU2010-10157-1, (2010).

 ISBN: 978-84-614-6167-7

Proceedings of the 11th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE2011
Benidorm, Spain, 26-30 June 2011

Sinchronizations of the geophysical processes and asymmetries in the solar motion about the Solar System's barycentre

Nikolay Sidorenkov [1], Ian R.G. Wilson [2], Anatoly I. Kchlystov [3]
[1] *Hydrometcentre of Russia, Moscow, 123242, Russia;* sidorenkov@mecom.ru
[2] *Queensland Department of Education, Training and the Arts Toowoomba, Australia;* irgeo@ozemail.com.au
[3] *Sternberg Astronomical Institute, Moscow, Russia;* ai42@sai.msu.ru

It is well known that many geophysical processes vary on inter-annual to decadal timescales. These variations are usually attributed to terrestrial causes that include: the Earth's core-mantle coupling; the effects of internal driven stochastic oscillations in the climatic system; the effects of the global conveyer belt upon ocean surface temperatures etc. However, we contend that the empirical evidences and facts demand that this generally accepted assumption should be revised and modified.

We find that the observed changes in the specific mass of the Antarctic and Greenland ice sheets closely correspond to the specific mass variations that are needed to explain the "decadal-long" fluctuations in LOD (Sidorenkov, 2009). Since the mass of the Antarctic and Greenland ice sheets depend on long-term climate variations, it is reasonable to assume that the decadal fluctuations in the Earth's rotation may also correlate with the variations in the major climatic indices. Following this line of reasoning, we have found that the atmospheric circulation regimes and the ten-year running mean of the Northern Hemisphere air temperature anomalies are well correlated with the changes in the Earth's rotation rate. In addition, Stanislav Perov and Nikolay Sidorenkov (2009), have found a significant correlation between fluctuations in the Earth's rotational rate and activity of the India monsoon through Rupa Kumar's data (2004). This correlation is supported by the relationship that Ian Wilson has found between the deviation of the Earth's LOD from its long– term trend and the Pacific Decadal Oscillation (PDO). Wilson finds that whenever there is a large deviation in the Earth's LOD from its long--term trend, the PDO index transitions to its positive phase. It is important to note that the observed changes in the LOD precede those in the anomalies of the precipitation in Indian monsoon and in the PDO by about eight years.

Ian Wilson has found that the times when Solar/Lunar tides had their greatest impact upon the Earth are closely synchronized with the times of greatest

 ISBN: 978-84-614-6167-7

asymmetry in the Solar Inertial Motion (S IM). Over the last 800 years, the Earth has experience exceptionally strong tidal forces in following sequence of years: 1247, 1433, 1610, 1787 and 1974 (Keeling and Whor f, 1997). Wilson shows that these exceptionally strong tidal forces closely correspond in tim e to the first peak in the asymmetry of the SIM that occurs just after a period low asymmetry. These first peaks in asymmetry in the SIM occur in following sequence of years 1251, 1432, 1611, 1791, and 1971, closely correspond the years of peak tidal force.

Thus, there appear to be periodic a lignments between th e lunar apsides, syzygies and lunar nodes that occur at almost exactly the same times that the SIM becomes most asymmetric for the first time after a period of low asymmetry in the SIM. It m eans that p recession and s tretching of the Lunar orbit (i.e. the factors that control the long-term variation of the lunar tides that are experienced here on Earth) are almost perfectly synchronized with the SIM.

If the Solar system just consisted of Jupiter and the Sun, the barycentre of the Solar System would move in an alm ost circular orbit located just ab ove the surface of the Sun (i.e. about 1.08 solar radi i), called the sub-Jupiter point. Hence, the actual m otion of the barycentre about the centre of the S un (or equivalently the Sun abo ut the barycentre) can b e considered as a com bination of the smooth symmetrical motion produced by Jupiter, combined with an additional, of ten asymmetric motion, caused by the other thr ee Jovan planets (principally, Saturn and Neptune). The distance of the centre-of-mass from the Sub-Jupiter point is an excellent indicator of the level of asymmetry of the Sun's orbital m otion, at any given time.

A.I.Khlystov et al., (1995) showed that the Earth and the other planets move in ellipses with the Sun at one foci, while at the same time they (effectively) share in the m otion of the Sun around barycenter of the solar system (Figure 7). From the above results two important conclusions follows:

1) The m ovement of each planet is transferred to the Sun, and then b ack from it to all of the other planets. In other words, the Sun acts as a re-transmitter of gravitational motion over all of the solar system.

2) Activization of sim ilar physical processes shou ld take p lace simultaneously on all bodies in the solar system.

Support for the last conclusion comes from the investigation of link between the most severe droughts on the Eart h and powerful dust storm s on Mars (Khlystov, 1995).

Ian Wilson et al. (2008) presented eviden ce that claimed that changes in the Sun's equatorial rotation rate are synchr onized with changes in the Sun' s orbital motion about the barycentre of the Solar System . This pa per showed that the recent maximum asymmetries in the Solar motion about the barycentre have occurred in the years 1865, 1900, 1934, 197 0 and 2007. These years closely match the points of inflection in the Earth's LOD.

Furthermore one-three years after these years in the European part of Russia v ery high hot weather in the summer was observed. This high hot weather in the summer occur in the years 1901, 1936 and 1938, 1972, and 2010, closely

 ISBN: 978-84-614-6167-7

correspond the years of m aximum asymmetries in the Solar m otion about the barycentre.

In addition, Ian W ilson (Sidorenkov and W ilson 2009) shows that, fro m 1700 to 2000 A.D., on every occasion where the Sun has experienced a maximum in the asymmetry of its motion about the centre-of- mass of the Solar System, the Earth has also experien ced a significant de viation in its rotation ra te (i.e. LOD) from that expected from the long-term trends.

The long-term trends in Earth's rotati on is attributed the action of tidal friction and the Barkin's mechanism of th e displacements of the Earth's lay ers (Barkin, 2000).

Thus decadal variation s of many ge odynamic, climatic and even weather processes are synchronized with a phenomenon that is linked to the changes in the solar motion about the barycentre of the Solar System.

Thus from the empirical data, we argue that there is compelling evidence to support the idea that these correlations ar e due to the shared m otion of the Sun and Earth about the barycentre of the Sola r System. We show that asymmetries in this shared motion lead to the de cadal fluctuations in the climatologically and geophysical processes, including long term changes in the Earth's rotation rate.

References

1. D'Arrigo, R., Villalba, R ., and W iles, G. (2001), ``Tree-ring estimates of Pacific decadal climate variability", *Clim. Dyn.*, 18, pp. 219--224.
2. Khlystov A.I., V.P.Dolgachyev, and L.M.Domozhilova (1995). *Trudy of the Sternberg Astronomical Institute, Moscow*, vol. 64, part 1, 91-102.
3. Keeling C.D. and Whorf T.P. (1997). Possible forcing of global temperature by the oceanic tid es. *Proc. Natl. Acad. Sci. USA* Vol. 94, pp. 8321–8328.
4. Rupa Kumar K, Pant G.B., Beig G., Srinivasan G. (2004). Climate Change Science: A Scoping S tudy for In dia Sponsored by the British High Commission in India. *Indian Institute of Tropical Meteorology, Pune.*
5. Perov S.P., Sidorenkov N.S.(2009). M.A. Petrosyanz and problems of the meteorology and climatology. All Russian Conference. Moscow State University. Moscow. MAKC Press, 2009. P.67.
6. Sidorenkov, N.S. (2009), The interac tion between Earth's rotation and geophysical processes. W ILEY-VCH Verlag Gm bH & Co. KGaA, Weinheim, 2009. 317 pp.
7. Wilson, I. R. G., Carter, B. D., and White, I. A., (2008), ``Does a Spin-Orbit Coupling Between th e Sun and the Jovian Pl anets Govern the Solar Cycle?", *Publications of the Astronomical Society of Australia*, 25, pp. 85--93.
8. N.S. Sidorenkov, Ian W ilson. The decadal fluctuations in the Earth's rotation and in the climate characteristics. In: Proceedings of the "Journees

 ISBN: 978-84-614-6167-7

2008 Systemes de reference spatio-temporels", M. Soffel and N. Capitaine (eds.), Lohrmann-Observatorium and Observatoire de Paris. 2009, pp. 174-177.

9. Barkin Yu. V. A mechanism of variations of the Earth rotation at different timescales. //In Polar Motion: Historical and Scientific Problems. ASP Conference Series. - V. 208. - 2000. S. Dick, D. McCarthy and B. Luzum eds. – P. 373-379.

 ISBN: 978-84-614-6167-7

Proceedings of the 11th International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE2011
Benidorm, Spain, 26-30 June 2011

High-throughput peptide structure prediction with distributed volunteer computing networks

T. Strunk, M. Wolf, W. Wenzel
Institute of Nanotechnology
Karlsruhe Institute of Technology
PO Box 3640, 76021 Karlsruhe, Germany

Abstract

Many short peptides are involved in important biological processed in the cell. Recent investigations have focused on the use of artificial peptides as antimicrobial drugs and antibiotics that differentially target bacterial and eucariotic cells. Because the number of possible peptide sequences is very large functional peptide design necessitates automated synthesis and screening of number of peptide sequences. Protein structure prediction methods can aid in this design process by providing structure function relationships for the interaction of the peptide with the membrane.
Here we show the applicability of our novel de-novo peptide prediction method, which allowed prediction of the structure of 10 peptides with a resultion of 2.4 Angström all-atom RMSD to the respective experimental structure. We employ massively parallel simulated annealing simulations to sample a sizable fraction of the peptide's conformational space on our volunteer computing network POEM@HOME. Using our free-energy function PFF02, we were able to select the native conformation as the global minimum of the protein free energy for peptides of both helical and sheet topologies. Our prediction protocol could allow the automated screening of large peptide databases for their structural features and by that enable the rapid prototyping of peptides for novel peptide design.

 ISBN: 978-84-614-6167-7